1 Cumulative distribution functions

A cumulative distribution function is an alternative way of describing the probability distribution of a random variable. When you have completed this chapter, you should

- know the meaning of a cumulative distribution function and how to find it for both continuous and discrete random variables
- be able to find the probability density from a cumulative distribution function.

1.1 Finding the cumulative distribution from the probability density

When you have statistical data about a continuous random variable, you can represent it either with a histogram or with a cumulative frequency diagram. A similar choice can be made for graphs representing theoretical probability models.

One possibility is to use a probability density function $f(x)$. This is conventionally defined over the complete set $\mathbb{R}$ of real numbers, although often there are intervals $]-\infty, a]$ or $[b, \infty[$ for which $f(x) = 0$.

Probability density functions have the following properties (see Higher Level Book 2 Section 10.2).

- $f(x) \geq 0$ for all $x \in \mathbb{R}$.
- The area under the probability density graph is equal to 1. That is,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

- The probability that the random variable X lies in the interval $[c, d]$ is equal to the area under the probability density graph over this interval. That is,

$$\mathrm{P}(c \leq X \leq d) = \int_{c}^{d} f(x)\,\mathrm{d}x.$$

Another possibility is to use a **cumulative distribution function** $F(x)$, which is defined as the probability that the random variable X is less than or equal to x. That is,

$$F(x) = \mathrm{P}(X \leq x).$$

For a continuous random variable it is not important whether you define the function as $\mathrm{P}(X \leq x)$ or $\mathrm{P}(X < x)$, since the probability that X is equal to any particular single value x is 0. But the distinction is important if you extend the definition to include discrete random variables.

Example 1.1.1

Fig. 1.1 shows the graph of a probability density function

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

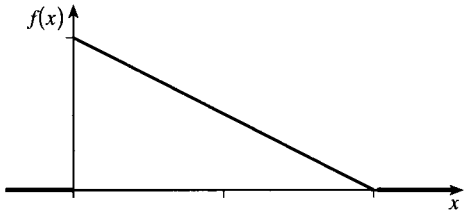

Fig. 1.1

Find the cumulative distribution function $F(x)$, and draw its graph.

When x is negative the probability density is 0, so $F(x) = 0$ for $x \le 0$. For $0 \le x \le 2$, the cumulative probability is equal to the area under the line in the interval $[0, x]$, which is shown by the shaded region in Fig. 1.2. This is a trapezium with parallel sides of lengths 1 and $1 - \frac{1}{2}x$, at a distance x apart. Therefore

$$F(x) = \tfrac{1}{2}\left(1 + \left(1 - \tfrac{1}{2}x\right)\right) \times x = x - \tfrac{1}{4}x^2.$$

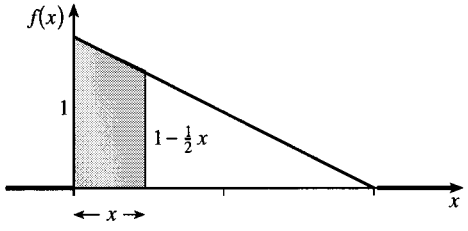

Fig. 1.2

This gives $F(2) = 2 - \frac{1}{4} \times 2^2 = 2 - 1 = 1$; you can check that this is the area of the triangle. Beyond $x = 2$ the probability density is again 0, so that there is no further increase in the cumulative probability. Therefore $F(x) = 1$ for $x > 2$.

The graph of $F(x)$ is shown in Fig. 1.3.

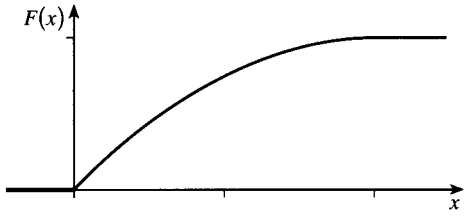

Fig. 1.3

In most cases finding $F(x)$ for a given $f(x)$ involves integration, but this presents a small problem of notation. If you write the probability $\mathrm{P}(X \le x)$ as $\int f(x)\,\mathrm{d}x$ evaluated over the interval $]-\infty, x]$, then the letter x appears twice, once inside the integral and the other as one of the limits of integration. Fortunately, in finding a *definite* integral it doesn't matter what letter you use inside the integral: the value of $\int_c^d f(x)\,\mathrm{d}x$ is exactly the same as $\int_c^d f(t)\,\mathrm{d}t$ or $\int_c^d f(u)\,\mathrm{d}u$. So it is better to avoid the problem by using a different letter inside the integral, writing the cumulative probability for example as

$$\mathrm{P}(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

If a continuous random variable has probability density $f(x)$, the cumulative distribution function $F(x) = \mathrm{P}(X \le x)$ is given by

$$\int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Example 1.1.2

A probability density function is defined by the equation

$$f(x) = \begin{cases} 0 & \text{for } x < 1, \\ \dfrac{1}{x^2} & \text{for } x \geq 1. \end{cases}$$

(a) Check that this is a valid probability density function.

(b) Find an expression for the cumulative distribution function $F(x)$.

(a) It is obvious that $f(x) \geq 0$ for all x, so you just have to prove that $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{1} 0\,\mathrm{d}x + \int_{1}^{\infty} \frac{1}{x^2}\,\mathrm{d}x.$$

The first integral on the right is clearly 0, and the second is the limit as $v \to \infty$ of

$$\int_{1}^{v} \frac{1}{x^2}\,\mathrm{d}x = \left[-\frac{1}{x}\right]_1^v = 1 - \frac{1}{v}.$$

Since $\lim_{v\to\infty}\left(1 - \frac{1}{v}\right) = 1 - 0 = 1$, it follows that

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 0 + 1 = 1.$$

(b) If $x < 1$, then clearly $F(x) = 0$.
If $x \geq 1$ then

$$\begin{aligned} F(x) &= \mathrm{P}(X \leq x) \\ &= \int_{-\infty}^{1} 0\,\mathrm{d}t + \int_{1}^{x} \frac{1}{t^2}\,\mathrm{d}t \\ &= 0 + \left[-\frac{1}{t}\right]_1^x \\ &= 1 - \frac{1}{x}. \end{aligned}$$

The graph of $f(x)$ is shown in Fig. 1.4, and that of $F(x)$ in Fig. 1.5.

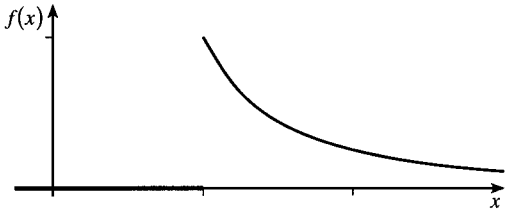

Fig. 1.4

Fig. 1.5

Sometimes to describe $f(x)$ you need different non-zero expressions over different intervals of the domain. This is illustrated in the next example.

Example 1.1.3

In a population of a particular breed of dog the age distribution is modelled by the probability density function shown in Fig. 1.6, with equation

$$f(x) = \begin{cases} \frac{1}{10} & \text{for } 0 < x \leq 6, \\ \frac{1}{30}x - \frac{1}{360}x^2 & \text{for } 6 < x \leq 12, \\ 0 & \text{otherwise,} \end{cases}$$

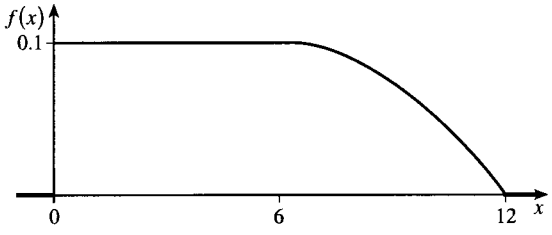

Fig. 1.6

where x denotes the age in years. Find expressions for the cumulative distribution function. Hence find the median and quartile ages of the dogs in the population.

Clearly $F(x) = 0$ for $x \leq 0$. For $0 < x \leq 6$ the region under the probability density graph is a rectangle with width x and height $\frac{1}{10}$, so that $F(x) = \frac{1}{10}x$. For $6 < x \leq 12$ the region under the graph is a rectangle of width 6 and height $\frac{1}{10}$ and a region under a curve (see the shaded area in Fig. 1.7) so

$$\begin{aligned} F(x) &= \frac{6}{10} + \int_6^x \left(\frac{1}{30}t - \frac{1}{360}t^2\right) \mathrm{d}t \\ &= \frac{6}{10} + \left[\frac{1}{60}t^2 - \frac{1}{1080}t^3\right]_6^x \\ &= \frac{6}{10} + \left(\frac{1}{60}x^2 - \frac{1}{1080}x^3\right) - \left(\frac{36}{60} - \frac{216}{1080}\right) \\ &= \frac{1}{1080}\left(216 + 18x^2 - x^3\right). \end{aligned}$$

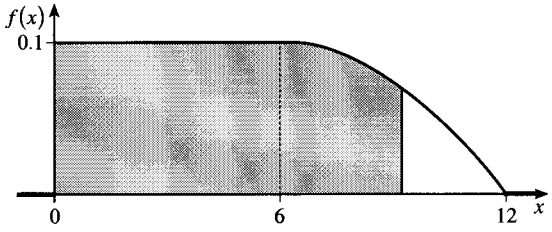

Fig. 1.7

For $x > 12$ the probability density is 0, so the value of $F(x)$ remains constant and equal to

$$F(12) = \frac{1}{1080}(216 + 2592 - 1728) = \frac{1}{1080} \times 1080 = 1.$$

Putting all this together,

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{1}{10}x & \text{for } 0 < x \leq 6, \\ \frac{1}{1080}\left(216 + 18x^2 - x^3\right) & \text{for } 6 < x \leq 12, \\ 1 & \text{for } x > 12. \end{cases}$$

The graph of $F(x)$ is shown in Fig. 1.8.

From Higher Level Book 2 Section 10.4 the median M is the value of x such that $\mathrm{P}(X \leq M) = \frac{1}{2}$, that is $F(M) = \frac{1}{2}$. Since $F(6) = \frac{6}{10} > \frac{1}{2}$, M lies in the interval $]0,6]$, so that $F(M) = \frac{1}{10}M$. Therefore $\frac{1}{10}M = \frac{1}{2}$, so $M = 5$.

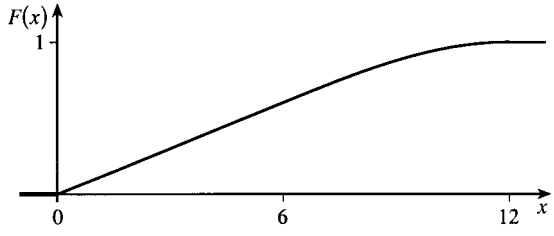

Fig. 1.8

Similarly the lower quartile Q_1 lies in the interval $]0,6]$. Therefore $\frac{1}{10}Q_1 = \frac{1}{4}$, so $Q_1 = 2.5$.

However, the upper quartile Q_3 lies in the interval $]6,12]$, so that

$$F(Q_3) = \tfrac{1}{1080}\left(216 + 18{Q_3}^2 - {Q_3}^3\right).$$

Since $F(Q_3)$ has to equal $\frac{3}{4}$, $x = Q_3$ has to satisfy the equation

$$\tfrac{1}{1080}\left(216 + 18x^2 - x^3\right) = \tfrac{3}{4}, \quad \text{which is} \quad x^3 - 18x^2 + 594 = 0.$$

Using a calculator to solve this equation, the root in the interval $]6,12]$ is 7.533… .

So the upper quartile $Q_3 = 7.53$, correct to 3 significant figures.

1.2 Finding the probability density from the cumulative distribution

Since the cumulative distribution is found from the probability density by integrating, you can find the probability density from the cumulative distribution by differentiating. For example, in Example 1.1.1 the cumulative distribution function over the interval $[0,2]$ is $F(x) = x - \frac{1}{4}x^2$. Differentiating this gives $F'(x) = 1 - \frac{1}{2}x$, which is the equation for the line segment joining $(0,1)$ to $(2,0)$ in the probability density graph.

If a continuous random variable has cumulative distribution function $F(x)$, the probability density function $f(x)$ is given by

$$f(x) = F'(x).$$

There is one small complication in applying this rule. Although the graph of $F(x)$ is always a continuous line, it may contain isolated points where there is a sudden change of direction, such as $(0,0)$ in Fig. 1.3 and $(1,0)$ in Fig. 1.5. You can't differentiate $F(x)$ at these points, so the equation $f(x) = F'(x)$ doesn't apply there. This doesn't matter, since $f(x)$ is only used to calculate probabilities over an interval, not at isolated points. The usual convention is to define $f(x)$ at these points so that the intervals used to describe $f(x)$ are the same as those used to describe $F(x)$.

In some applications it is easier to begin by finding the cumulative distribution and then to find the probability density by differentiation.

Example 1.2.1

A bad darts player is equally likely to hit any point on the board. (He is allowed to ignore any throws which miss the board completely.) The radius of the board is a. Find the cumulative distribution and the probability density for the random variable R, the distance from the centre at which a dart hits the board.

The distance of any hit from the centre must be between 0 and a, so the probability density is zero in the intervals $r < 0$ and $r > a$.

If r is between 0 and a, the hits for which $R \leq r$ lie inside (or on the circumference of) a circle of radius r, shown shaded in Fig. 1.9.

Since the hits are uniformly distributed over the surface of the board, the probability that $R \leq r$ is equal to the ratio of the area of this circle to the area of the whole board. That is,

$$F(r) = \mathrm{P}(R \leq r) = \frac{\pi r^2}{\pi a^2} = \frac{r^2}{a^2}.$$

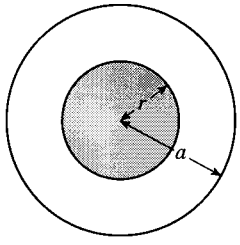

Fig. 1.9

So, over the interval $0 \leq r < a$, the probability density is given by

$$f(r) = F'(r) = \frac{2r}{a^2}.$$

The equations for the cumulative distribution and the probability density are then

$$F(r) = \begin{cases} 0 & \text{for } r < 0, \\ \dfrac{r^2}{a^2} & \text{for } 0 \leq r \leq a, \\ 1 & \text{for } r > a \end{cases} \quad \text{and} \quad f(r) = \begin{cases} 0 & \text{for } r < 0, \\ \dfrac{2r}{a^2} & \text{for } 0 \leq r \leq a, \\ 0 & \text{for } r > a. \end{cases}$$

1.3 Extension to discrete probability distributions

The definition of the cumulative distribution as $\mathrm{P}(X \leq x)$ is still valid if the random variable X is discrete, but the value of this probability remains constant in the intervals between successive elements of the sample space. The graph of the cumulative distribution function therefore consists of a set of line segments parallel to the x-axis.

Example 1.3.1

Draw graphs of the probability distribution and of the cumulative distribution function for the Poisson probability $\mathrm{Po}(2.5)$.

Your calculator probably has a program to calculate cumulative Poisson probabilities, and you should make sure that you know how to use this. However, you may only be able to use it to find the probability that $X \leq n$ for a particular value of n, and to draw the graph you want a table of values of $\mathrm{P}(X \leq n)$ for $n = 0, 1, 2, 3, \ldots$. So it may be more convenient to generate these as a sequence from the Poisson formula

$$\mathrm{P}(X = n) = \mathrm{e}^{-m} \frac{m^n}{n!}$$

with $m = 2.5$.

This is very simple, because the factor e^{-m} doesn't involve n, and

$$\frac{m^n}{n!} = \frac{m \times m^{n-1}}{n \times (n-1)!} = \frac{m}{n} \times \frac{m^{n-1}}{(n-1)!}.$$

So, if you use u_n to denote the probability $\mathrm{P}(X = n)$, values of u_n can be found from the inductive definition

$$u_0 = \mathrm{e}^{-m}, \qquad u_n = \frac{m}{n} \times u_{n-1} \text{ for } n = 1, 2, 3, \ldots.$$

Now in this example you want not only values of $P(X = n)$ but also those of $P(X \leq n)$, which is the sum sequence of u_n (see Higher Level Book 1 Section 30.4). So if you denote $P(X \leq n)$ by v_n, values of v_n can be found from the inductive definition

$$v_0 = u_0 = e^{-m}, \qquad v_n = v_{n-1} + u_n = v_{n-1} + \frac{m}{n} \times u_{n-1} \text{ for } n = 1, 2, 3, \ldots .$$

With $m = 2.5$, these equations give the values (to 3 significant figures) for u_n and v_n in Table 1.10.

n	0	1	2	3	4	5	6	...
$u_n = P(X = n)$	0.082	0.205	0.257	0.214	0.134	0.067	0.028	...
$v_n = P(X \leq n)$	0.082	0.287	0.544	0.758	0.891	0.958	0.986	...

Table 1.10

To draw the cumulative distribution graph, you need to define $F(x) = P(X \leq x)$ not just when x is a natural number 0, 1, 2, ... but when x is any real number. Using the values of $P(X \leq n)$ given in Table 1.10, it follows that

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.082 & \text{for } 0 \leq x < 1, \\ 0.287 & \text{for } 1 \leq x < 2, \\ 0.544 & \text{for } 2 \leq x < 3, \end{cases} \qquad \text{and so on.}$$

The graph of the probability distribution u_n is given in Fig. 1.11, and that of the cumulative distribution is given in Fig. 1.12. In the latter graph, a small dot has been added at the left end of each line segment to make clear that the jump occurs immediately *before* the integer value of x.

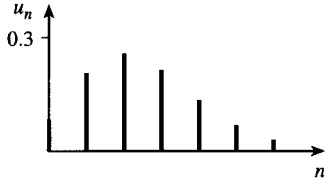

Fig. 1.11

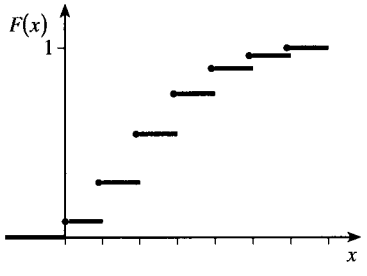

Fig. 1.12

Exercise 1

1 The probability density function of a random variable, X, is

$$f(x) = \begin{cases} \frac{1}{8}x & 0 \leq x \leq 4, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function of X. (b) Find the median.

2 The probability density function of a continuous random variable, X, is

$$f(x) = \begin{cases} \frac{2}{3} & 0 \le x < 1, \\ \frac{4}{3} - \frac{2}{3}x & 1 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $F(x)$. (b) Find the lower quartile. (c) Find the 85th percentile.

3 The probability density function of a continuous random variable, X, is

$$f(x) = \begin{cases} \frac{3}{14}x & 0 \le x < 2, \\ \frac{3}{28}x(4-x) & 2 \le x \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $F(x)$. (b) Find the median. (c) Find $\mathrm{P}(2.5 \le X \le 3)$.

4 Daily sales of petrol, X, at a service station (in tens of thousands of litres) are distributed with probability density function

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

(a) Find $F(x)$. (b) Find the median in litres.

5 The cumulative distribution of a random variable, X, is

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{1}{108}\left(9x^2 - x^3\right) & 0 \le x \le 6, \\ 1 & x > 6. \end{cases}$$

Find the probability density function of X.

6 The cumulative distribution function of a random variable, X, is

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{1}{18}x^2 & 0 \le x < 3, \\ \frac{2}{3}x - \frac{1}{18}x^2 - 1 & 3 \le x \le 6, \\ 1 & x > 6. \end{cases}$$

Find the probability density function of X.

7 The random variable X metres represents the side of a square. The area of the square is represented by the random variable Y metre^2.

(a) If X has uniform probability density 0.2 over the interval $]0,5[$, find the cumulative distribution function $F(x) = \mathrm{P}(X \le x)$.

(b) Deduce the cumulative distribution function for the area, $G(y) = \mathrm{P}(Y \le y)$.

(c) Hence find the probability density function $g(y)$ for the area of the square.

8 The random variable X metres represents the base of a rectangle of area 1 metre^2. The height is represented by the random variable Y metres.

(a) If X has uniform probability density 1 over the interval $]0,1[$, find the cumulative distribution function $F(x) = \mathrm{P}(X \leq x)$.

(b) Deduce the cumulative distribution function for the height, $G(y) = \mathrm{P}(Y \leq y)$.

(c) Hence find the probability density function $g(y)$ for the height of the rectangle.

9* In Example 1.1.3, suppose that the probability density function for the age distribution of the dog population remains the same over time, and that the total number of dogs in the population remains constant. It can then be shown that, if the random variable Y represents the age in years to which a dog lives, then the probability density function for Y is

$$g(y) = -yf'(y),$$

where f is the probability density function defined in Example 1.1.3. Use this to find the median age to which a dog of this breed lives.

10 A random variable X has binomial probability $\mathrm{B}(3,0.4)$. Find the values of the cumulative distribution function in the intervals $]-\infty,0[$, $[0,1[$, $[1,2[$, $[2,3[$ and $[3,\infty[$.

11 Draw the graph of the cumulative distribution function for Poisson probability $\mathrm{Po}(1.7)$.

2 Geometric and exponential probability

The two probability models described in this chapter both arise from the same question: how long must you wait until a certain event occurs? When you have completed this chapter, you should

- be able to calculate and apply geometric and exponential probability
- know expressions for the expectation and variance of these probability models
- understand the connection between exponential and Poisson probability.

2.1 The geometric distribution

Consider the following three examples.

1 A dice is rolled until a six is scored. Let X be the number of rolls up to and including the roll on which the first six occurs.

2 A card is selected, with replacement, from a standard pack of cards until an ace is drawn. Let Y be the number of selections up to and including the one on which the first ace occurs.

3 A person has a 1 in 17 chance of winning a prize in a lottery. She keeps playing the lottery once each week. Let W be the number of weeks up to and including the week in which she first wins a prize.

These three random variables have certain similarities.

In the first example,

$$\mathrm{P}(X = 1) = \mathrm{P}(\text{a six occurs on the first throw}) = \frac{1}{6}.$$

Let s represent the event that a six occurs on a trial and let f represent the event that a six does not occur. Then $X = 2$ corresponds to the sequence fs. Similarly the event $X = 3$ corresponds to the sequence ffs. Notice that there is only one possible sequence for each value of X.

You can now calculate the probability distribution for X:

$$\mathrm{P}(X = 2) = \mathrm{P}(fs) = \left(\frac{5}{6}\right) \times \left(\frac{1}{6}\right) = \frac{5}{36},$$

$$\mathrm{P}(X = 3) = \mathrm{P}(ffs) = \left(\frac{5}{6}\right)^2 \times \frac{1}{6} = \frac{25}{216},$$

$$\mathrm{P}(X = 4) = \mathrm{P}(fffs) = \left(\frac{5}{6}\right)^3 \times \frac{1}{6} = \frac{125}{1296},$$

and so on.

You can generalise this as

$$\mathrm{P}(X = x) = \left(\frac{5}{6}\right)^{x-1} \times \frac{1}{6} \quad \text{for } x = 1, 2, 3, \ldots .$$

In the second example the probability distribution formula for Y will be $\mathrm{P}(Y = y) = \left(\frac{12}{13}\right)^{y-1} \times \frac{1}{13}$ for values of $y = 1, 2, 3, \ldots$ since