# 7.5 Modelling random behaviour: random variables and probability distributions

In the rest of this chapter you will apply your knowledge and understanding of combined events to model random behaviour in various processes that occur in real life.

Recall from Chapter 3 that **discrete** data is either data that can be **counted**, for example the number of cars in a car park, or data that can only take specific values, for example shoe size.

**Continuous** data can be **measured**, for example height, weight and time.

For example if a café has four coffee percolators, the **number** of percolators that fail on a given day, $X$, is a quantity that changes randomly according to the durability and reliability of the machinery. In contrast, the **weight** of a bag of coffee, $Y$, varies randomly according to the weighing and packaging processes in the factory. These are both examples of **random variables**: the value of these variables changes according to chance. The value of $Y$ is a **measurement** that can take **any** real number in an interval.

We use the following terminology for random variables that are countable.

**Example:** Let $X$ represent the number of percolators failing on a given day.

| Terminology | Explanation |
|---|---|
| $X$ is a **discrete random variable.** | Discrete: $X$ can be found by **counting**. |
| | Random: $X$ is the result of a **random** process. |
| | Variable: $X$ can take **any value** in the domain $\{0, 1, 2, 3, 4\}$. |

A first step in acquiring knowledge about $X$ is to fill in a table:

| **Number of failing percolators ($x$)** | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **$P(X = x)$** | $P(X = 0)$ | $P(X = 1)$ | $P(X = 2)$ | $P(X = 3)$ | $P(X = 4)$ |

The five probabilities must add to one, and they must each satisfy $0 \leq P(X = x) \leq 1$.

Knowing them establishes the **probability distribution** of $X$ since the table then shows how the entire probability of 1 is distributed to each value of the random variable in its domain.

A **discrete probability distribution** is the set of all possible values of a discrete random variable (a subset of $\mathbb{Z}$) together with their corresponding probabilities.

**TOK**

A model is used to represent a mathematical situation in this case. In what ways may models help or hinder the search for knowledge?

## Investigation 13

**Finding and representing discrete probability distributions.**

One fair tetrahedral die with faces numbered 1, 2, 3 and 4 is thrown. Let $A$ denote the value of the number thrown.

**1** Complete the table to show the probability distribution of $A$:

| $a$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\mathrm{P}(A=a)$ | | | | |

**2** Draw a bar chart to represent the probability distribution of $A$ and make a general statement for $\mathrm{P}(A=a)$.

Two fair tetrahedral dice with faces numbered 1, 2, 3 and 4 are thrown. Let $B$ denote the sum of the two values thrown.

**3** Use combined events to complete the table to show the probability distribution of $B$:

| $b$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $\mathrm{P}(B=b)$ | | | | | | | |

**4** Draw a bar chart to represent the probability distribution of $B$ and make a general statement for $\mathrm{P}(B=b)$ using a piecewise function.

**5** Compare and contrast the shapes of the bar charts for $A$ and for $B$.

**6** Compare and contrast the probability distributions of $A$ and $B$.

**7** **Factual** How could you use your answer to **3** to predict the number of times $B$ is at least 6 in 200 trials?

**8** **Conceptual** How can you represent a discrete probability distribution?

**9** **Conceptual** How can you find a discrete probability distribution?

In the following investigation you will discover and explore an example of a discrete probability distribution with useful and surprising applications in real life.

## Investigation 14

**Discovering a discrete random distribution**

Imagine a list giving the populations of 267 countries. Let $Z$ be the first digit in the number of people in a randomly chosen country.

**1** **Factual** Make your own subjective judgment of the probability distribution of $Z$. What would the domain of $Z$ and its shape be? Share this judgment in a pair and then with your class.

Now consider the data giving the population of 267 countries found here: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html

Collaborate with others to complete a frequency table as follows for the 267 populations:

| First digit | 1 | 2 | ... | 9 |
|---|---|---|---|---|
| Frequency | | | | |

Continued on next page

Hence construct a bar chart to show experimental probabilities for $Z$.

**2** **Conceptual** How does your bar chart compare to your answer for **1**?

Let $Y$ be the first digit of number in the measure of the **area** of a randomly chosen country.

Repeat **1** and **2** with the area of each country found here https://www.cia.gov/library/publications/the-world-factbook/rankorder/2147rank.html

**3** **Factual** Use your knowledge of functions to propose a model for these sets of data.

**4** **Conceptual** How may a discrete probability distribution be determined, apart from the method you gave in the previous investigation?

A **discrete probability distribution function** $f(t)$ assigns to each value of the random variable its corresponding probability: $f(t) = \mathrm{P}(T = t)$.

$f(t)$ is commonly referred to by the abbreviation "pdf".

## Example 17

**a** Show that the function $f(t) = \frac{t-2}{15}$, $t \in \mathbb{Z}$, $3 \leq t \leq 7$, defines a discrete probability distribution by constructing a table of values.

**b** If $\mathrm{P}(T = t) = f(t)$, represent the distribution as a bar chart.

**a**

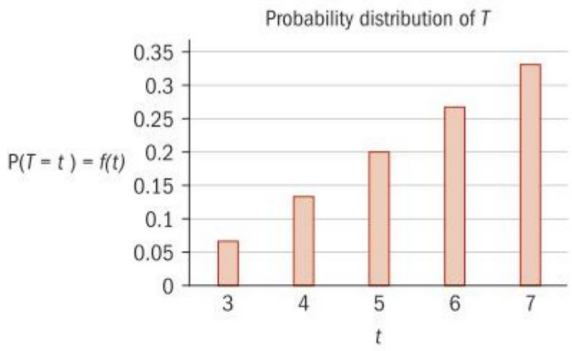
| $t$ | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| $f(t)$ | $\frac{1}{15}$ | $\frac{2}{15}$ | $\frac{3}{15}$ | $\frac{4}{15}$ | $\frac{5}{15}$ |

Substitute each value of $t$ from the domain given.

Since $\frac{1+2+3+4+5}{15} = 1$, $f(t) = \frac{t-2}{15}$ defines a probability distribution on the domain given.

Show that the values of the function add to 1.

**b**




Label the graph completely in order to communicate what it represents.

## Example 18

A fair cubical die and a fair tetrahedral die are thrown. The discrete random variable $S$ is defined as the sum of the numbers on the two dice.

**a** Construct the probability distribution of $S$ as:

**i** a table of values **ii** a bar chart **iii** a piecewise function.

**b** Hence find the probabilities:

**i** $P(S > 2)$ **ii** $P(S$ is at most $6)$ **iii** $P(S \leq 6 | S > 2)$.

**a**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1** | 2 | 3 | 4 | 5 | 6 | 7 |
| **2** | 3 | 4 | 5 | 6 | 7 | 8 |
| **3** | 4 | 5 | 6 | 7 | 8 | 9 |
| **4** | 5 | 6 | 7 | 8 | 9 | 10 |

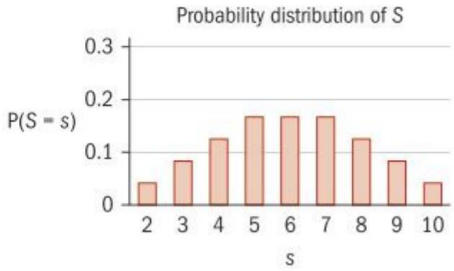
Draw a sample space diagram.

**i**

| $s$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{P(S=s)}$ | $\frac{1}{24}$ | $\frac{2}{24}$ | $\frac{3}{24}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{3}{24}$ | $\frac{2}{24}$ | $\frac{1}{24}$ |

Read off the probability of each event $P(S = s)$ in turn from the sample space diagram.

**ii**




Do not forget to label both axes.

**iii**

$$f(s) = \begin{cases} \frac{s-1}{24} & 2 \leq s \leq 4 \\ \frac{1}{6} & 5 \leq s \leq 7 \\ \frac{11-s}{24} & 8 \leq s \leq 10 \end{cases} \quad \text{where } s \in \mathbb{N}.$$

Making a general statement helps represent the probability distribution function concisely.

**b i** $P(S > 2) = 1 - \frac{1}{24} = \frac{23}{24}$

**ii** $P(S \leq 6) = \frac{1}{24} + \frac{1}{12} + \frac{1}{8} + \frac{1}{6} + \frac{1}{6} = \frac{7}{12}$

Take care to interpret "at most 6" correctly.

**iii** $P(S \leq 6 | S > 2) = \frac{P(S \leq 6 \cap S > 2)}{P(S > 2)}$

$= \frac{P(3 \leq S \leq 6)}{P(S > 2)} = \frac{\frac{13}{24}}{\frac{23}{24}} = \frac{13}{23}$

Use the formula for conditional probability and find the intersection of the two sets.

Continued on next page

Or

| $s$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $P(S = s)$ | $\frac{1}{24}$ | $\frac{2}{24}$ | $\frac{3}{24}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{3}{24}$ | $\frac{2}{24}$ | $\frac{1}{24}$ |

$P(3 \leq S \leq 6)$ can be found as $\frac{2}{24} + \frac{3}{24} + \frac{1}{6} + \frac{1}{6} = \frac{13}{24}$ from the diagram. The whole sample space of $P(S \leq 6 | S > 2)$ is $P(S > 2) = \frac{23}{24}$, hence

$$P(S \leq 6 | S > 2) = \frac{P(3 \leq S \leq 6)}{P(S > 2)} = \frac{\frac{13}{24}}{\frac{23}{24}} = \frac{13}{23}$$

You are not obliged to use the formula: a diagram can be used to give an equivalent solution. Many students find this method easier.

## Exercise 7I

**1** Consider these three tables. State which table(s) could not represent a discrete probability distribution, giving reasons.

**a**

| $a$ | 1 | 2 | 3 | 7 |
|---|---|---|---|---|
| $\mathbf{P}(A=a)$ | 0.1 | 0.2 | 0.03 | 0.4 |

**b**

| $b$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\mathbf{P}(B=b)$ | 0.1 | −0.2 | 0.4 | 0.4 |

**c**

| $c$ | 4.5 | 3 | 1 | 0 |
|---|---|---|---|---|
| $\mathbf{P}(C=c)$ | 0.2 | 0.2 | 0.5 | 0.1 |

**2** Show that the function $f(t) = \frac{t-4}{21}$, $t \in \mathbb{Z}$, $5 \leq t \leq 10$, defines a discrete probability distribution by constructing a table of values.

**3** $f(x) = \frac{x}{19}$, $x \in \{1, 5, 7, k\}$, defines a discrete probability distribution. Find the value of $k$.

**4** Sarah researches multiple births in a clinic, where she keeps records over a period of years of the genders of triplets born there. There are eight possible sequences of genders in a set of triplets, for example MFM.

**a** Construct the sample space of all possible sequences.

**b** Assuming P(Male) = P(Female) = 0.5, construct the probability distribution table of the random variable $F$ = the number of females born in a set of triplets.

**5** The probability distribution of a discrete random variable $A$ is defined by this table:

| $a$ | 5 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| $\mathbf{P}(A=a)$ | 0.5 | 0.05 | 0.04 | 0.1 | 0.2 | $P(A=12)$ |

Find:

**a** $P(A = 12)$ **b** $P(8 < A \leq 10)$

**c** $P(A$ is no more than $9)$

**d** $P(A$ is at least $10)$ **e** $P(A > 8 | A \leq 11)$.

**6** Nico and Artem are designing a card game. In their pack of cards, each card has 2, 1, 0 or 5 printed on it. A card is selected from the pack and the number on the card defines a random variable $X$.

Nico states that the probability distribution of the game is as follows:

| $x$ | 2 | 1 | 0 | 5 |
|---|---|---|---|---|
| $\mathbf{P}(X=x)$ | 0.3 | 0.27 | 0.25 | 0.1 |

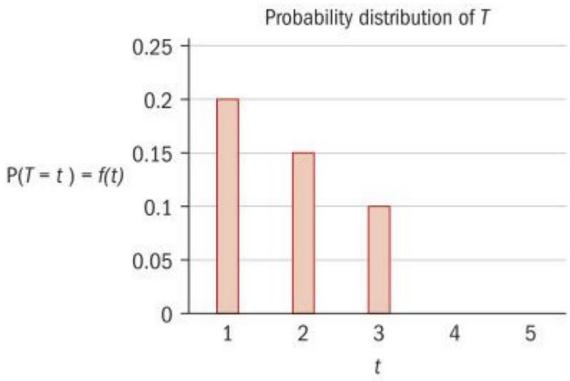
**a** Explain why Nico is wrong.

**b** Artem correctly states that the probability distribution is

| $x$ | 2 | 1 | 0 | 5 |
|---|---|---|---|---|
| $\mathbf{P}(X=x)$ | 0.28 | 0.2 | $p$ | $3p$ |

Find the value of $p$.

**7** Part of the discrete probability distribution of the discrete random variable $T$ with domain $\{1, 2, 3, 4, 5\}$ is shown in the bar chart.




Given that $P(T = 4) = 4P(T = 5)$, construct the probability distribution table for $T$.

## Example 19

The probability distribution of a discrete random variable $U$ is defined by

$P(U = u) = k(u - 3)(8 - u)$, $u \in \{4, 5, 6, 7\}$.

**a** Find the value of $k$ and hence draw the table of the probability distribution of $U$.

**b** In 100 trials, calculate the expected value of each possible outcome of $U$.

**c** Find the mean of the values of $U$ found in these 100 trials assuming the frequencies of each outcome of $U$ are given by your expected values in **b**.

**d** Interpret your answer.

**a**

| $u$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| $\mathbf{P}(U=u)$ | $4k$ | $6k$ | $6k$ | $4k$ |

Represent the probability distribution in a table.

$4k + 6k + 6k + 4k = 20k = 1 \Rightarrow k = \frac{1}{20}$,

hence

| $u$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| $\mathbf{P}(U=u)$ | $\frac{1}{5}$ | $\frac{3}{10}$ | $\frac{3}{10}$ | $\frac{1}{5}$ |

Use the fact that the probabilities must add to 1 on the domain of the pdf.

**b** The expected number of occurrences of $u = 4$ is $100 \times \frac{1}{5} = 20$. Similarly, the expected number of occurrences of 5, 6 and 7 are 30, 30 and 20 respectively.

Use the formula: expected number of occurrences of $A = n\,P(A)$.

Continued on next page

**c** A grouped frequency table for these expected number of occurrences is

| $u$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| **Expected frequency of $u$** | 20 | 30 | 30 | 20 |

Hence the mean value of $U$ with these frequencies is

$$\frac{20\times4+30\times5+30\times6+20\times7}{100}=\frac{550}{100}=5.5$$

Each number of occurrences is a frequency.

Use the formula for the population mean:

$$\mu=\frac{\sum_{i=1}^{k}f_ix_i}{n} \text{ where } n=\sum_{i=1}^{k}f_i$$

**d** 5.5 is not a number in the domain of the probability distribution function. Nevertheless, it models the central value of $U$ expected in a data set of 100 trials.

Recall the definition of the mean as a measure of central tendency.

You can express the calculation in part **c** of the previous exercise more briefly as:

$$\frac{20\times4+30\times5+30\times6+20\times7}{100}=4\times\frac{20}{100}+5\times\frac{30}{100}+6\times\frac{30}{100}+7\times\frac{20}{100}$$

$$=4\times\frac{1}{5}+5\times\frac{3}{5}+6\times\frac{3}{5}+7\times\frac{1}{5}$$

Notice that this is the sum of the product of each value of the random variable with its corresponding probability, or $\sum_u uP(U=u)$. This leads to a further generalization for all discrete random variables:

The **expected value** of a discrete random variable $X$ is $E(x)=\mu=\sum_x xP(X=x)$

## Example 20

A newsagent in Oxford takes delivery of six copies of a Scottish newspaper each Sunday. The newsagent has a regular order from her customers for three newspapers but sales vary according to current events, sport etc. The newsagent has collected data over several years to help predict her sales, creating a probability distribution table for the random variable $S$, the number of Scottish newspapers sold on Sunday.

| $s$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $P(S=s)$ | 0.05 | 0.39 | 0.29 | 0.22 | 0.05 |

Hence find the expected number of newspapers sold and interpret its meaning in context.

| | |
|---|---|
| $E(S) = \sum_s sP(S = s) =$<br>$2\times0.05+3\times0.39+4\times0.29+5\times0.22+6\times0.05$<br>$= 3.83$<br>On average, the newsagent should expect to sell more than 3 newspapers although 3 is the most likely outcome.<br>There is more than a 50% chance that she will sell more than her regular order. | Apply the formula for the expected value.<br><br>Draw conclusions from the given information. |

## Investigation 15

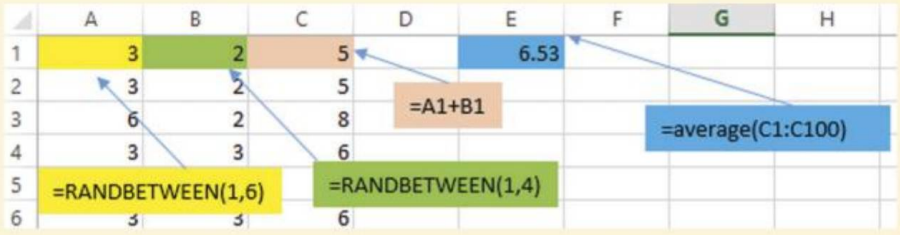
Consider the distribution in Example 18: "A fair cubical die and a fair tetrahedral die are thrown. The discrete random variable $S$ is defined as the sum of the numbers on the two dice." Consider an experiment in which 100 values of $S$ are calculated in 100 trials.

Predict the average of these 100 values of $S$ by:

**a** deducing from the shape of the bar chart representation

**b** application of the formula for the expected value of a discrete random variable

**c** constructing a data set of 100 values of $S$ with a spreadsheet and finding its mean by entering these formulae and dragging A1, B1 and C1 down to the 100th row.




1 **Factual** What are the strengths and weaknesses of each approach? Discuss.

2 **Conceptual** What does the expected value of a discrete random variable predict about the outcomes of a number of trials?

Many countries organize national lotteries in which adults buy a ticket giving them a chance to win one of a range of cash prizes. Profits are often invested in "good causes". For example, the UK National Lottery has distributed over UK£37 billion through thousands of grants to good causes including sport, art and health projects since 1994.

It is possible to use the expected value formula to manage the prize structure of a lottery in order to maintain profitability.

| Prize | Probability | Cash value per winner (UK £) |
|---|---|---|
| 1st (Jackpot) | $\frac{1}{45\,057\,474}$ | 5 421 027 |
| 2nd | $\frac{1}{7\,509\,579}$ | 44 503 |
| 3rd | $\frac{1}{144\,415}$ | 1018 |
| 4th | $\frac{1}{2180}$ | 84 |
| 5th | $\frac{1}{97}$ | 25 |
| 6th | $\frac{1}{10.3}$ | Free ticket |

Looking at the cash prizes only for simplicity, we can find the expected winnings as follows:

Expected cash winnings =

$$5\,421\,027 \times \frac{1}{45\,057\,474} + 44\,503 \times \frac{1}{7\,509\,579} + 1018 \times \frac{1}{144\,415}$$
$$+ 84 \times \frac{1}{2180} + 25 \times \frac{1}{97} \approx 0.430$$

This would appear to show that you would expect to make a profit (or a positive gain) playing this lottery! However you do have to buy the ticket first, which costs UK£2.00 so you should expect a **loss** (a negative gain) of UK£1.57. The attraction of the game is based on the desire to win a large prize or contribute to good causes. But it should not be a surprise that you would **expect** to make a loss on any one game.

**TOK**

Do you rely on intuition to help you make decisions?

If $X$ is a discrete random variable that represents the gain of a player, then if $E(X) = 0$, the game is **fair.**

## Example 21

Some students have a meeting to design a dice game to raise funds for charity as part of a CAS project. Some of the decisions made in the meeting are lost.

This incomplete probability distribution table remains:

| $x$ (prize in US$) | 1 | 2 | 4 | 6 | 7 |
|---|---|---|---|---|---|
| $P(X = x)$ | $\frac{11}{40}$ | $\frac{1}{4}$ | | | $\frac{1}{8}$ |

The students also recall that $E(X) = \frac{67}{20}$ and that the probability distribution function generalizes to a linear model.

**a** Determine the missing entries in the table and hence find the probability distribution function.

**b** Find the smallest entry fee the students could set for playing the game in order to predict a profit. Comment on the advantages and disadvantages of a number of possible entry fees.

**a** Let the missing probabilities be represented by $a$ and $b$.

Then $a + b = 1 - \frac{11}{40} - \frac{1}{4} - \frac{1}{8} = \frac{7}{20}$, also

$1 \times \frac{11}{40} + 2 \times \frac{1}{4} + 4a + 6b + 7 \times \frac{1}{8} = \frac{67}{20}$

Hence $a + b = \frac{7}{20}$ and $4a + 6b = \frac{17}{10}$.

So $a = \frac{1}{5}$ and $b = \frac{3}{20}$.

The completed table is :

| $x$ (prize in US$) | 1 | 2 | 4 | 6 | 7 |
|---|---|---|---|---|---|
| $P(X = x)$ | $\frac{11}{40}$ | $\frac{10}{40}$ | $\frac{8}{40}$ | $\frac{6}{40}$ | $\frac{5}{40}$ |

Hence $f(x) = P(X = x) = \frac{1}{40}(12 - x)$

Representing the unknown quantities with a variable and writing down true statements involving them is a problem-solving strategy.

The probabilities must add to 1. You can apply formula for the expected value.

Solve the system of simultaneous equations.

Look for a pattern in the probability distribution table or alternatively apply the general equation for a linear function $y = mx + c$.

**b** $E(X) = \frac{77}{20} = \text{US\$3.85}$ so charging a player US$3.85 would be a fair game. Therefore charging US$3.86 would predict a small profit, but this could easily give a loss. Perhaps charging US$4.00 would be more practical and it would predict a larger profit.

Apply the formula for the expected value and reflect critically.

## Exercise 7J

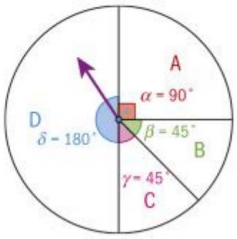
**1** The discrete random variable $B$ has probability distribution function given by $P(B = b) = k(4 - b)$ for $b = 0, 1, 2, 3$. Find $k$ and $E(B)$.

**2** Apply your probability distribution table from Exercise 7I question 4 to find the expected number of male births in a set of triplets. Interpret your result.

**3** A handbag contains seven coins and three keys. Two items are taken out of the handbag one after the other and not replaced. Find the expected number of keys taken out of the handbag.

**4** A handbag contains five coins, four keys and eight mints. Two items are taken out of the handbag one after the other and not replaced. Find the expected number of mints taken out of the handbag.

**5** Alexandre is designing a game. A spinning arrow rotates and stops on one of the regions A, B, C or D as shown in the diagram.




Alexandre proposes the prizes shown in the table and that the game should cost US$5 to play.

| Letter | A | B | C | D |
|---|---|---|---|---|
| Prize | US$3 | US$7 | US$5 | US$2 |

Determine whether Alexandre's game is fair and justify your answer.

**6** Ten thousand US$l0 lottery tickets are sold. One ticket wins a prize of US$5000, five tickets each win US$1000, and ten tickets each win US$200. Find:

**a** the probability of winning each prize in the lottery

**b** the expected gain from one ticket

**c** the price of a ticket to make the lottery a fair game.

**7** Two fair tetrahedral dice with faces numbered 1, 2, 3 and 4 are thrown. The discrete random variable $D$ is defined as the **product** of the two numbers thrown.

**a** Find the probability distribution table of $D$.

**b** Find $P(D \text{ is a square number} \mid D < 8)$.

**c** In a CAS fundraising game, a prize of US$12 is won if $D$ is odd, and a prize of US$6 is won if $D$ is even. Find the price of a ticket to ensure this is a fair game.

**8** Xsquared Potato Crisps runs a promotion for a week. In 0.01% of the hundreds of thousands of bags produced there are gold tickets for a round-the-world trip. Let $B$ represent the number of bags of crisps opened until a gold ticket is found.

**a** Find $P(B = 1)$, $P(B = 2)$ and $P(B = 3)$.

**b** Hence show that the probability distribution function of $B$ is

$f(b) = P(B = b) = 0.0001(0.9999)^{b-1}$

**c** State the domain of $f(b)$.

**d** Determined to win a ticket, Yimo buys 10 bags of crisps. Find the probability that she finds a gold ticket after opening no more than 10 bags.

## Developing inquiry skills

Read again the second question from the opening scenario.

Can this question be modelled by a discrete probability distribution?

How could you find its probability distribution? What assumptions would have to be made?

# 7.6 Modelling the number of successes in a fixed number of trials

You learned in Section 7.5 that a probability distribution function can be found as a generalization of a random process.

An example of the process you will learn about in this section is found in the work of cognitive psychologists Daniel Kahneman and Amos Tversky (1972).