13 Representing multiple outcomes: random variables and probability distributions

Probability enables us to quantify the likelihood of events occurring and evaluate risk. The probabilities in the diverse situations described below can be modelled to help you make predictions and well informed decisions.

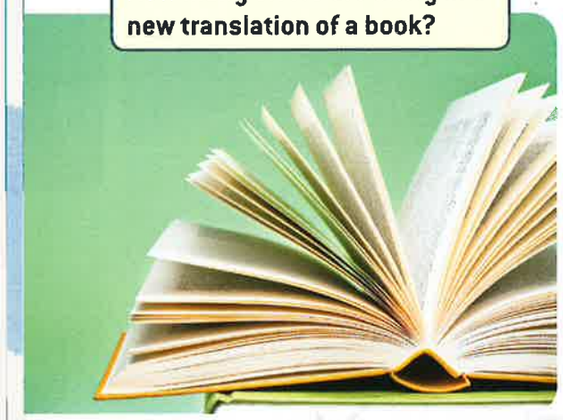
Concepts

- Representation
- Validity

Microconcepts

- Discrete random variables
- Binomial distribution
- Normal distribution
- Probability distribution function
- Continuous probability distribution function
- Discrete probability distribution function
- Probability distribution
- Cumulative distribution function
- Discrete and continuous data
- Expected value
- Variance
- Poisson distribution
- Parameters
- Probability density functions

How many errors are likely in a new translation of a book?



How can an airline manage their booking system so that overbooking the flight maintains profitability while treating their customers fairly?

How can an engineer quantify the risk that an unacceptable number of micro-scopic flaws are present in an aircraft wing?

How likely is it that a sample of fish is representative of the entire population?

I woke with a premonition: tomorrow my café would be inspected by the hygiene authority. I knew 3 months ago to expect a visit within a year, but not *when*. I'd expected longer than 3 months to prepare. I set out to arrive at the café early, but had to wait for 19 minutes at the bus stop. I usually have to wait 5 minutes for a bus on average. Then 3 came along at once!

On arrival I checked the machines. Two of the four percolators on the espresso machine were not functioning! Was this bad luck, I wondered? I knew that each percolator had a one in a thousand chance of not functioning on any given day … I started to think about this problem but decided to focus on fixing the percolators.

I had to open the espresso machine combination lock to access the percolators. I was given the 3-digit code years ago but had lost it! I could remember that the first two digits were the same. I decided to guess. After 34 trials I found the right code and the door swung open. I fixed the percolators: but did I have enough coffee? I knew the average weight of a bag was 3 kg and that it could range from 2.8 to 3.2 kg … how sure could I be that I would not run out? Matt the barista rushed in. "Four inspectors came!" he exclaimed. "Why the past tense?" I enquired. "They were too heavy for the maximum elevator load of 300 kg and they didn't have time to use the stairs" came the answer … time to relax with an espresso I figured.

- How many situations are there in this tale that involve probability?
- What kind of variables are involved?
- How could you represent and quantify the variables?
- Write down some examples where chance has played a role in *your* life.

Developing inquiry skills

How could you estimate the probability of each event? Do the results surprise you? How good are people at evaluating risk?

Think about the questions in this opening problem and answer any you can. As you work through the chapter, you will gain mathematical knowledge and skills that will help you to answer them all.

Before you start

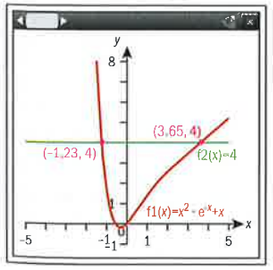
You should know how to:

1 Find the mean of a data set.
eg Find the mean of:

x_i	4	7	8	10
f_i	2	1	5	7

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{125}{15} \approx 8.33$$

2 Use technology to solve equations.
eg Use technology to solve $x^2 e^{-x} + x = 4$



$x = -1.23$ or 3.65

Skills check

Click here for help with this skills check

1 Find the mean value of x.

x_i	1	2	3	6	11
f_i	9	7	3	2	1

2 Solve the equation $x^3 e^{-x} + \ln(x) = 1$

13.1 Modelling random behaviour

You have learned in Chapter 5 how to quantify the probability of an event and the probability of combined events. In this chapter you will apply these concepts to model random behaviour in various processes that occur in real life.

For example, in Chancer's cafe in the opening scenario the number of failing percolators on a given day is a quantity that changes randomly. The number of buses that arrive at a bus stop in a 5-minute interval of time is another quantity that changes randomly. These quantities are determined by different processes.

We use the following terminology and notation:

Let X represent the number of failing percolators 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 {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 1, and they must each satisfy $0 \leq P(X=x) \leq 1$.

You will determine the probabilities in the second row in section 13.2. This row 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.

In this example you have explored data sets for patterns in order to determine the probability distribution. You can also explore processes.

Investigation 1

Daniel is playing a dice game in which he throws five fair cubical dice until he throws five equal numbers. He tires of this game after throwing the dice 617 times without success, and plans to carry out a simulation with a spreadsheet instead.

He writes down the definition of a discrete random variable T: "T is the number of trials taken until I throw five equal numbers with five fair cubical dice."

1 What is the sample space for T?

2 What is $P(T=1)$?

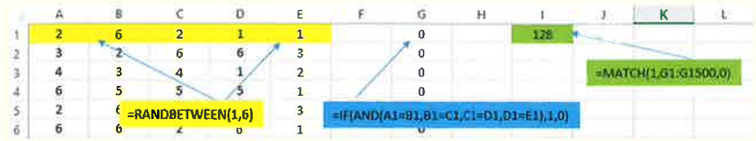
3 How many trials would Daniel expect to carry out before he can expect to have thrown five equal numbers once?

TOK

"Those who have knowledge, don't predict. Those who predict, don't have knowledge." – Lao Tzu.

Why do you think that people want to believe that an outside influence such as an octopus or a groundhog can predict the future?

You can test your answer with a spreadsheet as shown below:



a Use a random number generator such as "= RANDBETWEEN(1, 6)" in each one of the first five columns of the first row to simulate the five dice.

b Typing "=IF(AND(A1=B1,B1=C1,C1=D1,D1=E1),1,0)" in the sixth column makes the spreadsheet give the answer 1 if five equal numbers are thrown, 0 otherwise.

c Copy and paste the first row down to the 1500th row to simulate 1500 throws of the five dice.

d Typing "=MATCH(1,G1:G1500,0)" will find the number of the first trial of the 1500, if any, in which five identical numbers were thrown. This is the value of the random variable T. Pressing F9 will generate another 1500 trials.

Find $P(T=2), P(T=3), P(T=4), \ldots$ Generalize to find a function $f(t) = P(T=t)$. This is called a **probability distribution function**.

4 What is the general statement for $f(t) = P(T=t)$?

5 What should the values of $f(t)$ add up to on its domain and why?

6 Find $\sum f(t)$ over all possible values of T.

7 **Conceptual** How can a discrete probability distribution function be found?

Definition

A **discrete probability distribution function** $f(t)$ assigns to each value of the random variable a corresponding probability. $f(t) = P(T=t)$.

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

The **discrete cumulative distribution function** $F(t)$ assigns to each value of the random variable its corresponding cumulative probability.

$F(t) = P(T \leq t) = \sum_{n=a}^{t} f(n)$ where a is the minimum value of the domain of $f(t)$.

$F(t)$ is commonly referred to by the abbreviation "cdf".

Example 1

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

a Represent the probability distribution of S as:

i a table of values **ii** a bar chart.

b Hence find the probabilities:

i $P(S > 7)$ **ii** $P(S$ is at most $5)$ **iii** $P(S \leq 6 | S > 2)$.

Continued on next page

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

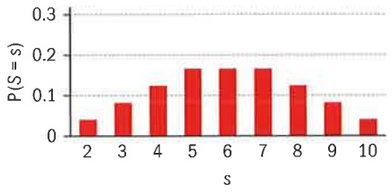
Draw a sample space diagram.

i

s	2	3	4	5	6	7	8	9	10
$\mathbf{P}(S=s)$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$

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

ii



Don't forget to label both axes.

b i $\mathrm{P}(S>7)=\frac{1}{8}+\frac{1}{12}+\frac{1}{24}=\frac{1}{4}$

ii $\mathrm{P}(S\le 5)=\frac{1}{24}+\frac{1}{12}+\frac{1}{8}+\frac{1}{6}=\frac{5}{12}$

Take care to interpret "at most 5" correctly.

iii $\mathrm{P}(S\le 6\,|\,S>2)=\frac{\mathrm{P}(S\le 6\cap S>2)}{\mathrm{P}(S>2)}$

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

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

Example 2

The probability distribution of a discrete random variable U is defined by $\mathrm{P}(U=u)=k(u-3)(8-u)$, $u\in\{4, 5, 6, 7\}$.

a Find the value of k and hence represent the probability distribution of U.

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

c Hence predict the mean value of U after a large number of trials. 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

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

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

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 $A=n\mathrm{P}(A)$.

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

u	4	5	6	7
Frequency	20	30	30	20

Each number of occurrences is a frequency.

Hence the mean value of U with these frequencies is

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

Use the formula for the population mean:

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

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 example more briefly as

$$\frac{20\times4+30\times5+30\times6+20\times7}{100}=4\times\frac{20}{100}+5\times\frac{30}{100}+6\times\frac{30}{100}+7\times\frac{20}{100}=4\times\frac{1}{5}+5\times\frac{3}{10}+6\times\frac{3}{10}+7\times\frac{1}{5}$$

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

The **expected value** of a discrete random variable X is the mean score that would be expected if X was repeated many times. It is calculated using the formula: $\mathrm{E}(X)=\mu=\sum_x x\mathrm{P}(X=x)$

Investigation 2

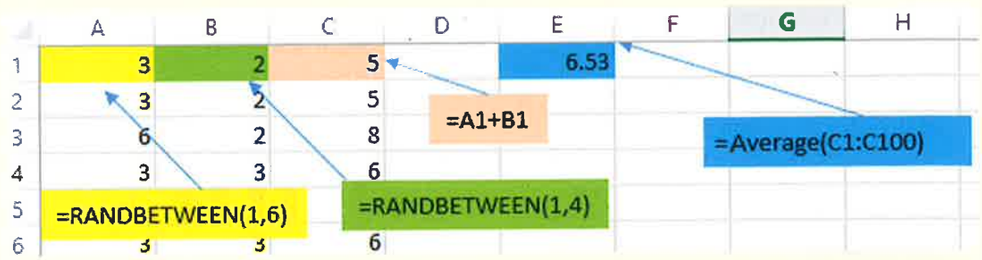
Consider the distribution in worked example 1: "A fair cubical dice and a fair tetrahedral dice are thrown. The discrete random variable S is defined as the sum of the numbers on the two dice." Consider an experiment in which 100 values of S are calculated in 100 trials.

Predict the average of these 100 values of S values 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 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?

Application: Many countries organize national lotteries in which adults buy a ticket giving 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 £37 billion 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 (£)
1st (Jackpot)	$\frac{1}{45\,057\,474}$	5 421 027
2nd	$\frac{1}{7\,509\,579}$	44 503
3rd	$\frac{1}{144\,415}$	1 018
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, the cost of a ticket is £2.00 so you expect a *loss* (a negative gain) of £1.57. The attraction of the game is based on the desire to win a large prize and/or contribute to good causes. But it should not be a surprise that you would *expect* to make a loss on any one game.

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

Example 3

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 $)	1	2	4	6	7
$P(X=x)$	$\frac{11}{40}$	$\frac{1}{4}$			$\frac{1}{8}$

The students also recall that $E(X) = \frac{67}{20}$.

a Determine the rest of the table. Given the probabilities follow a linear model, write down an expression for $P(X = x)$.

b What is the smallest cost the students could set for playing the game in order to predict a profit? What would you recommend as the cost of the game and why?

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 $b = \frac{3}{20}$

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 and the formula for the expected value can be applied.

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 $y = mx + c$. This is called generalizing to a linear model.

Continued on next page

The completed table is :

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

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

b $E(X) = \frac{67}{20} = \$3.35$ so charging a player \$3.35 would be a fair game. Therefore charging \$3.36 would predict a small profit, but this could easily give a loss. Perhaps charging \$4.00 would be more practical and it would predict a larger profit.

Apply the formula for the expected value and reflect critically.

Designing a CAS project

Students are working to raise money for charity in a CAS project during a school fair. Adults are invited to pay to play a simple game that gives the chance to win cash prizes. The rules are as follows:

A fair cubical dice is thrown. To play the game once costs \$5. For each outcome the prizes are:

Outcome	1	2	3	4	5	6
Prize (\$)	3	3	3	4	5	6

The expected gain from this game is

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

a Is this a fair game? Would you expect it to make a profit for the charity?

b You can test the profitability of the game using a spreadsheet to generate 100 trials.

Enter the prizes in cells A1, B1, C1, D1, E1, F1 and "=RANDBETWEEN(1,6)" in cell H1. Then type in cell J1: ="IF(H1=1,A1,IF(H1=2,B1,IF(H1=3,C1, IF(H1=4,D1,IF(H1=5,E1,IF(H1=6,F1))))))"

Copy and drag all cells down to the 100th row. Use the spreadsheet to find the total winnings and the total gain in the 100 trials.

Work in a small group on these design tasks for the CAS project:

c Suggest a change to the game to make it fair by changing the cost.

d Suggest changes to the game to make it fair by changing the distribution of the prizes.

e Suggest changes to the game that make it profitable but more attractive to players.

Exercise 13A

1 Consider these three tables. State which table(s) could not represent a discrete probability distribution and why.

a

b	1	2	3	4
$P(B=b)$	0.1	0.2	0.4	0.4

b

b	4	3	1	0
$P(B=b)$	−0.2	0.2	0.6	0.4

c

b	1	2	3.104	4
$P(B=b)$	0.1	0.2	0.3	0.4

2 The probability distribution of a discrete random variable A is defined by this table:

a	5	8	9	10	11	12
$P(A=a)$	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 $P(A > 8 | A \leq 11)$ **f** $E(A)$

3 X-squared 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 $P(B = 1)$, $P(B = 2)$, $P(B = 3)$.

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 ten bags of crisps. Find the probability that she finds a golden ticket after opening no more than ten bags.

4 The cumulative probability distribution function of a discrete random variable C is defined by this table:

c	1	2	3	4	5
$P(C \leq c)$	0.07	0.09	0.26	0.72	1

a Calculate the probability distribution of C.

b Find $E(C)$.

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

a Represent the probability distribution of D in a table.

b Find $P(D$ is a square number $|D < 8)$.

6 Marin throws three coins. He wins \$15 if three heads occur, \$5 if exactly two heads occur, \$$y$ if only one head occurs and \$2 if no heads occur. M is the discrete random variable representing Marin's winnings.

a Find $E(M)$ in terms of y.

b Find the value of y that makes the game fair if Marin pays \$7 to play one game.

7 Two batteries are required to fit into a handheld whisk in a restaurant kitchen. The batteries are selected at random from a box holding three Fastcell and four Econbatt batteries. F is the number of Fastcell batteries selected to fit into the whisk.

a Find the probability distribution table.

b Find $E(F)$.

8 **a** Ten identically shaped discs are in a bag: two of them are black and the rest white. Discs are drawn at random from the bag and not replaced. Let G be the number of discs drawn until the first black one is drawn.

i Find the probability distribution function $f(g) = P(G = g)$.

ii Find $E(G)$.

b If instead each disc is replaced before the next is drawn, repeat part **a**.

c Show that for each case, the sum of all probabilities on the domain is 1.

9 The marketing team of Xsquared crisps wishes to explore how their marketing campaign can be generalized. If p represents the probability that a golden ticket is found in a randomly chosen bag of crisps then the probability distribution function of B, the number of bags of crisps opened until a gold ticket is found, is given by $f(b) = \mathrm{P}(B = b) = p(1-p)^{b-1}$ where $b \in \mathbb{Z}^+$ and the cumulative distribution function is $F(b) = \mathrm{P}(B \leq b) = 1 - (1-p)^b$.

a Show that for this discrete probability distribution, if $y \geq 0$ then $\mathrm{P}(B \geq x + y \mid B \geq x) = \mathrm{P}(B \geq y + 1)$.

b Interpret $\mathrm{P}(B \geq 15 \mid B \geq 10) = \mathrm{P}(B \geq 6)$.

Developing inquiry skills

Return to the opening problem.

Which of the situations in the café involve discrete probability distributions?

13.2 Modelling the number of successes in a fixed number of trials

You learned in section 13.1 that a probability distribution function can be found as a generalization of a random process. One 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). One question they posed in a survey was:

"All families of six children in a city were surveyed. In 72 families, the exact order of births of boys and girls was GBGBBG. What is your estimate of the number of families surveyed in which the exact order of births was BGBBBB?"

The median estimate was 30, suggesting the participants in the survey judged that GBGBBG was more than twice as likely an outcome as BGBBBB. Psychologists have studied this "representativeness fallacy" in further research of subjective judgments and biases. In this section, you will learn the mathematics needed to model situations like this.

You can model the family of six children as a sequence of six independent trials in which the probability of a male birth is constant over the six trials. A convenient way to experience this process is by flipping a coin six times.

TOK

Is it possible to reduce all human behaviour to a set of statistical data?

Investigation 3

Preliminary experience:
Carry out ten trials of flipping a coin six times, recording each trial as a sequence of H (heads) and T (tails). Compare your results with others in your class.

- How many of your trials resulted in outcomes like HTHTTH which *appear* more random than THTTTT?
- Discuss this claim: "The probability of a total of three heads in six trials is more than the probability of a total of one head in six trials". Justify your answer.

International-mindedness

The physicist Frank Oppenheimer wrote: "Prediction is dependent only on the assumption that observed patterns will be repeated."

This is the danger of extrapolation. There are many examples of its failure in the past, for example share prices, the spread of disease, climate change.

Investigation 4

In this investigation, represent all probabilities as fractions. Do not simplify the fractions.

A fair coin is tossed twice. Let X be the discrete random variable equal to the number of heads tossed in two trials of a fair coin. Use a tree diagram to represent all the possibilities in the sample space and hence find the probabilities to complete the probability distribution table.

x	0	1	2
$\mathrm{P}(X = x)$			

Repeat these steps for three trials and then four trials.

1 What is the connection between your results and the pattern found here in Pascal's triangle?

2 What do these numbers represent?

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
... ...

3 Predict the probability distribution table for five trials.

4 Would the probability distribution tables for 0 trials and 1 trial be consistent with the pattern in Pascal's triangle?

The numbers in the Pascal's triangle are **binomial coefficients**. You can calculate them with technology. Make a prediction of the sum of a row and the next two rows and check your answer with technology.

Definition: The binomial coefficients in row $(n+1)$ of the triangle are represented by the following notation:

$${}^nC_0, {}^nC_1, {}^nC_2, \ldots, {}^nC_r, {}^nC_{n-1}, {}^nC_n.$$

EXAM HINT

In examinations, binomial probabilities will be found using technology.

Continued on next page

Use your probability distributions from parts **3** and **4** and binomial coefficients to make a general statement for the probability distribution function:

5 **Factual** Complete: "Let X be the discrete random variable equal to the number of heads tossed in n trials of a fair coin.

Then $P(X=x) =$ _____ for $x \in \{0, 1, \ldots, \_\_\}$"

6 **Factual** The experiment is changed so that the coin is not fair and it is thrown five times.
$P(H) = p, P(T) = 1 - p$

Complete the general statement $P(X=x) =$ _____ for $x \in \{0, 1, \ldots, \_\_\}$

This investigation leads to the formal definition of the **binomial distribution**:

In a sequence of n independent trials of an experiment in which there are exactly two outcomes "success" and "failure" with constant probabilities P(success) $= p$, P(failure) $= 1 - p$, if X denotes the discrete random variable equal to the number of successes in n trials, then the probability distribution function of X is

$$P(X = x) = {}^nC_x\, p^x (1-p)^{n-x},\ x \in \{0, 1, 2, \ldots, n\}$$

These facts are summarized in words as "X is distributed binomially with parameters n and p" and in symbols as $X \sim B(n, p)$.

HINT
These probabilities can be worked out using your GDC.

The cumulative probability distribution function is

$$P(X \le x) = \sum_{i=0}^{x} {}^nC_i\, p^i (1-p)^{n-i}, x \in \{0, 1, 2, \ldots, n\}$$

You can use the binomial distribution to reflect on the questions at the start of this section.

7 Use the binomial distribution to find the probability of having exactly three boys in a family of six.

8 What is the probability of the outcome GBGBBG?

9 **Factual** Which part of the formula for the binomial distribution counts possibilities?

10 **Conceptual** What situations are described by binomial distributions?

Example 4

For each situation, state if the random variable is distributed binomially. If so, find the probability asked for.

a A coin is biased so that the probability of a head is 0.74. The coin is tossed seven times. A is the number of tails. Find $P(A = 5)$.

b A bag contains 12 white chocolates and 7 dark chocolates. A chocolate is selected at random and its type noted and then eaten. This is repeated five times. B is the number of dark chocolates eaten. Find $P(B = 7)$.

c A bag contains 10 red, 1 blue and 7 yellow dice. A dice is selected at random and its colour noted and replaced. This is repeated 12 times. C is the number of yellow dice recorded. Find $P(C \le 6)$.

d In a multiple-choice test of 20 questions, students must select the correct answer from five different options. Valentina guesses each of the 20 answers. D is the number of correct answers Valentina guesses. Find $P(D \ge 10)$.

e Ciaran plays a lottery in which the probability of buying a winning ticket is 0.001. E is the number of tickets Ciaran buys until he wins a prize. Find $P(E < 7)$.

a Each toss of the coin is independent of the others. There are exactly two outcomes and a fixed number of trials. Therefore $A \sim B(7, 0.26)$. $P(A = 5) = 0.0137$	Write the distribution. This clarifies your thoughts as well as awarding you method marks in the examination since you have demonstrated your knowledge and understanding. Use technology to find the binomial probability using the probability distribution function (pdf). Write the answer to three significant figures.
b Since the probability of selecting a dark chocolate is dependent on what was selected in previous trials, the trials are not independent so the binomial distribution is not an appropriate model for A.	
c Since the dice are replaced at each trial, the probability of success is constant and equal to $\frac{7}{18}$. Therefore $C \sim B\left(12, \frac{7}{18}\right)$. $P(C \le 6) = 0.861$	Use technology to find the binomial probability using the cumulative probability distribution function (cdf). Write the answer to three significant figures.
d Assuming Valentina pays no attention at all to the questions asked, $D \sim B\left(20, \frac{1}{5}\right)$. $P(D \ge 10) = 0.00260$	Use technology to find the binomial probability using the cumulative probability distribution function (cdf). For some GDCs you need to use the fact that $P(D \ge 10) = 1 - P(D \le 9)$. Write the answer to three significant figures.
e There is not a fixed number of trials, so the binomial distribution is not an appropriate model for E.	

From your findings in the investigation and from this example, how can you understand from the context of a problem that the binomial distribution is an appropriate model to apply?

Example 5

Solve the problems, stating any assumptions and interpretations you make.

a In a family of six children, find

i the probability that there are exactly three girls

ii the probability that exactly three consecutive girls are born.

b A study shows that 0.9% of a population of over 4 000 000 carries a virus. Find the smallest size of sample from the population required in order that the probability of the sample having no carriers is less than 0.4.

a i $G \sim B(6, 0.5)$. $P(G = 3) = 0.3125$	Assuming that "boy" and "girl" are the only two outcomes, the probability of each is 0.5, the gender of each child is independent of the others and that G represents the number of girls, then $G \sim B(6, 0.5)$.
ii Three consecutive girls are born: GGGBBB BGGGBB BBGGGB BBBGGG	This is considerably smaller than 0.3125 since there are many more combinations of three girls that are not consecutive than are consecutive.
Each of these four outcomes has probability $\left(\frac{1}{2}\right)^6$, so P(three consecutive girls are born) $= 4 \times \left(\frac{1}{2}\right)^6 = \frac{1}{16} = 0.0625$	Examine your result critically and check that it is feasible.
b In a population this size, the binomial distribution is an appropriate model since sampling without replacement does not alter the probability of choosing a carrier significantly. If C is the number of carriers chosen in a sample of size n then $C \sim B(n, 0.009)$.	State the assumptions.
Find the smallest value of n so that $P(C = 0) < 0.4$ $P(C = 0) = (1 - 0.009)^n < 0.4$, so $0.991^n < 0.4$	Write down the distribution.
$\Rightarrow \ln(0.991^n) < \ln(0.4)$ $\Rightarrow n\ln(0.991) < \ln(0.4)$ $\Rightarrow n > \frac{\ln(0.4)}{\ln(0.991)} = 101.351$	Translate the problem into an inequality. $\ln(x)$ is an increasing one-to-one function so the inequality sign does not change. $\ln(0.991) < 0$ so the inequality sign does change.
Hence $n = 102$ is the minimum value required.	Interpret the decimal. Alternatively, the problem can be solved using technology.

In section 13.1 you learned that the expected value of a discrete random variable X is $E(X) = \mu = \sum_x xP(X = x)$. You can use this to deduce that the expected value of the binomial distribution $X \sim B(n, p)$ is $E(X) = np$.

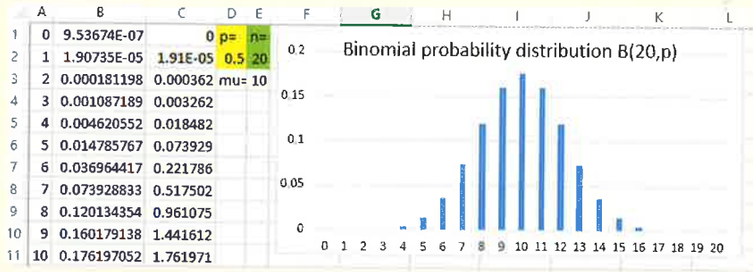
In statistics, you learned about the measures of central tendency (mean, median and mode), and measures of dispersion (range, interquartile range and standard deviation).

Probability distributions also have equivalents for the standard deviation and the variance.

The variance of a probability distribution X is written as $\text{Var}(X)$ and is the variance you would expect to find in the results if X was repeated many times. The standard deviation of X is written as $\text{Std}(X)$ and is the square root of the variance.

Investigation 5

A An experimental perspective



	A	B	C	D	E
1	0	9.53674E-07	0	p=	n=
2	1	1.90735E-05	1.91E-05	0.5	20
3	2	0.000181198	0.000362	mu=	10
4	3	0.001087189	0.003262		
5	4	0.004620552	0.018482		
6	5	0.014785767	0.073929		
7	6	0.036964417	0.221786		
8	7	0.073928833	0.517502		
9	8	0.120134354	0.961075		
10	9	0.160179138	1.441612		
11	10	0.176197052	1.761971		

a Use a spreadsheet to represent $X \sim B(20, p)$ as a probability distribution table and a bar chart by following these steps.

i In column A type the numbers 0, 1, 2, ..., 20 down to cell A21.

ii Fill in cells D1, D2, E1 and E2 as shown above.

iii Type "=BINOM.DIST(A1,E2,D2,FALSE)" in cell B1. These are the probabilities of the events $P(X = 0), P(X = 1), ..., P(X = 20)$, where $X \sim B(20, p)$.

iv Type "=A1*B1" in cell C1.

v Copy and drag cells B1 and C1 down to row 21.

vi Type "=sum(C1:C21)" in cell E3.

vii Add a chart to display the values of X on the x-axis and the corresponding probabilities on the y-axis.

b Which cell displays $E(X)$?

B With your spreadsheet, alter the values of p and of n to explore the effect these parameters have on the spread of the distribution.

- Compare and contrast the spread of $X \sim B(20, 0.15)$ with that of $X \sim B(20, 0.5)$.
- Compare and contrast the spread of $X \sim B(20, 0.15)$ with that of $X \sim B(20, 0.85)$.
- Compare and contrast the spread of $X \sim B(5, 0.85)$ with that of $X \sim B(20, 0.85)$.

Continued on next page

1 How can you make the spread greatest for a fixed number of trials by changing the probability of success?

2 How can you make the spread greater for a fixed probability of success by changing the number of trials?

3 Which parameters of $X \sim \mathrm{B}(n, p)$ affect $\mathrm{Var}(X)$?

If $X \sim \mathrm{B}(n, p)$ then $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$

TOK

What does it mean to say that mathematics can be regarded as a formal game lacking in essential meaning?

Exercise 13B

1 Given $X \sim \mathrm{B}(6, 0.29)$ find the probabilities:

a $\mathrm{P}(X = 4)$ **b** $\mathrm{P}(X \leq 4)$

c $\mathrm{P}(1 \leq X < 4)$ **d** $\mathrm{P}(X \geq 2)$

e $\mathrm{P}(X \leq 4 | X \geq 2)$

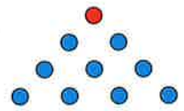
f Use your answers to determine if $X \leq 4$ and $X \geq 2$ are independent events.

2 A fair octahedral dice numbered 1, 2, …, 8 is thrown seven times. Find the probability that at least three prime numbers are thrown.

3 a Given $Y \sim \mathrm{B}(n, 0.4)$ and $\mathrm{E}(Y) = 5.2$, find n and $\mathrm{Var}(Y)$.

b Given $Z \sim \mathrm{B}(9, p)$ and $\mathrm{Var}(Z) = 1.44$, show that there are two possible values of p.

4 David plays a game at a fair. He throws a ball towards a pattern of ten holes in this formation:



The aim of the game is to have the ball fall into the red hole to win a point. One game consists of throwing ten balls. Assume David has no skill whatsoever at aiming, the ball is equally likely to fall in each of the holes and that a ball thrown must fall through one of the holes.

a Find the probability that David scores at least 5 points in a game.

b David plays six games. Find the probability that he scores no points in at least two games.

5 In a mathematics competition, students try to find the correct answer from five options in a multiple-choice exam of 25 questions. Alex decides his best strategy is to guess all the answers.

a State an appropriate model for the random variable A, the number of questions Alex gets correct.

b Find the probability that Alex gets at most five questions correct.

c Find the probability that Alex gets at least seven questions correct.

d Find the probability that Alex gets no more than three questions correct.

e Write down $\mathrm{E}(A)$ and interpret this value.

f Find the probability that Alex scores more than expected.

g In the test, a correct answer is awarded with 4 points. An incorrect answer incurs a penalty of 1 point. If Alex guesses all questions, find the expected value of his total points for the examination.

h Four students in total decide to guess all their answers. What is the probability that at least two of the four students will get seven or more questions correct?

6 Calcair buys a new passenger plane with 538 seats. For the first flight of the new plane all 538 tickets are sold. Assume that the probability that an individual passenger turns up to the airport in time to take their seat on the plane is 0.91.

a Model the distribution of the random variable T as the number of passengers that arrive on time to take their seat, stating any assumptions you make.

b Find $\mathrm{P}(T = 538)$ and interpret your answer.

c Find $\mathrm{P}(T \geq 510)$ and interpret your answer.

d Calcair knows that it is highly likely that there will be some empty seats on any flight unless it sells more tickets than seats. Find the smallest possible number of tickets sold so that $\mathrm{P}(T \geq 510)$ is at least 0.1.

e How many tickets should Calcair sell so that the expected number of passengers turning up on time is as close to 538 as possible?

f For this number of tickets sold, find $\mathrm{P}(T = 538)$ and $\mathrm{P}(T > 538)$. Interpret your answers.

7 You are given $X \sim \mathrm{B}(n, p)$. Apply calculus to the formula for $\mathrm{Var}(X)$ to find the value of p that gives the most dispersion (spread) of the probability distribution.

8 Johanna is designing a dice game. She buys three five-sided dice but is not convinced that they are fair dice. In her game, the three dice are thrown and the number of "1"s thrown is counted. In 200 trials, the following data is collected:

Number of 1s thrown	Frequency
0	79
1	83
2	9
3	29

a Assuming the dice are fair, model the data and calculate the expected frequencies of each outcome.

b Is there evidence that the dice are not fair?

13.3 Modelling the number of successes in a fixed interval

In this section, you will learn another discrete probability distribution that models a different process to the binomial distribution: the Poisson distribution.

Investigation 6

Nicolas is often late for school and he is reflecting on how bad his punctuality is, and the reasons for it. Nicolas collects data and identifies two discrete random variables of interest to him: L and U. L is the number of times in a school week of five days that Nicolas is late. Nicolas estimates that the probability that he is late on any given day is fixed at 0.05 and that lateness on any given day is independent of his punctuality or lack of it on any other day.

Nicolas relies on a bus service to pick him up near his home and drop him off near school.

Continued on next page

U is the number of buses arriving in a five-minute interval at Nicolas's bus stop. The bus company tells him that there is on average one bus per five minutes arriving at Nicolas's stop, but he knows from experience that sometimes a full ten minutes go by with no buses arriving, whereas on other occasions three buses will arrive in the same minute. Nicolas notices that traffic congestion, roadworks, and competing bus companies serve to randomize when buses arrive at his stop as well as other factors like poor weather and breakdowns.

1 **Factual** What are all the similarities and differences between the random variables L and U?

U satisfies the assumptions for the Poisson distribution, which are:

U counts the number of occurrences of an event (a bus arrives) in a given interval. The interval may have dimensions of time or of space.

An average rate of occurrences in a given time interval is given and the rate is uniform over the whole time-interval being considered.

Occurrences in a time interval are independent and occurrences cannot occur at the same time or position.

2 **Factual** For the following random variables, which ones can be modelled by the Poisson distribution and which by the binomial distribution?

A is the number of grammatical errors in five pages of a book translated from English to Albanian. The average number of errors per ten pages is four.	B is the number of faulty switches in a random sample of 10 switches chosen from a production line. The probability of a switch being faulty is 0.07.	C is the number of customers arriving between 0810 and 0820 at Chancer's café (from the opening problem) for morning coffee. On average, two customers arrive each minute between 0800 and 0900.
D is the number of goals recorded in a ten-minute interval during 60 football matches, assuming that on average 2.7 goals are scored per game.	E is the number of five-minute intervals in one hour in which there are at least 10 customers arriving at Chancer's café given that the probability of at least 10 customers arriving in five minutes is fixed at 0.21.	F is the number of accidents on a motorway in seven weekends given that the average number of accidents per weekend is 2.1.

3 **Conceptual** What situations does the Poisson process model?

4 **Conceptual** How do the parameters and sample space of the Poisson distribution and the binomial distribution compare and contrast?

If X satisfies these requirements:

- X counts the number of occurrences of an event in a given interval. The interval may have dimensions of time or space.
- An average rate of occurrences in a given time interval is given (α in this case), and is uniform across all the time intervals being considered.
- Occurrences in a time interval are independent and occurrences cannot occur at the same time or position.

The probability distribution function of X is $\mathrm{P}(X = x) = \dfrac{\mathrm{e}^{-\alpha}\alpha^{x}}{x!}$,

$x \in \{0, 1, 2, \ldots\}$. This formula is not required for the exam.

These facts are summarized in words as "X follows a Poisson distribution with parameter α" and in symbols as $X \sim \mathrm{Po}(\alpha)$.

EXAM HINT
In examinations, Poisson probabilities will be found using technology. Make sure you know how to do this.

Example 6

a Assume that the number of goals scored in a football match can be modelled by the Poisson distribution with parameter 2.9. Let G be the number of goals in a particular match. Find:

i $\mathrm{P}(G = 4)$ **ii** $\mathrm{P}(G \leq 3)$ **iii** $\mathrm{P}(G \geq 4)$.

b Let L be the number of goals scored in five matches. Write down the distribution of L and use it to find $\mathrm{P}(L \leq 10 | L \geq 2)$.

a $G \sim \mathrm{Po}(2.9)$	State the distribution.
i 0.162 **ii** 0.670 **iii** 0.330	Use technology to find the probabilities. Write the answers correct to three significant figures.
b $5 \times 2.9 = 14.5$	L must have parameter 14.5 by direct proportion since it is the average number of goals in five matches. State the distribution.
$L \sim \mathrm{Po}(14.5)$ $\mathrm{P}(L \leq 10 \mid L \geq 2) = \dfrac{\mathrm{P}(L \leq 10 \cap L \geq 2)}{\mathrm{P}(L \geq 2)} = \dfrac{\mathrm{P}(2 \leq L \leq 10)}{\mathrm{P}(L \geq 2)}$	Use the formula for conditional probability.
$\mathrm{P}(L \leq 10 \mid L \geq 2) = 0.145$	Use technology to find the probabilities. Write the answer correct to three significant figures.

Example 7

Trains at a busy railway station are occasionally cancelled due to staff shortages, breakdowns, a lack of available trains and many other causes. Assume that there are on average 2.31 cancelled trains per day and that the number of cancelled trains C can be modelled by $C \sim \mathrm{Po}(2.31)$.

a Find the probability that there will be four or more cancellations on a given day.

b Find the probability that there will be at least 81 cancellations in the month of March.

c Find, in two different ways, the probability that there are no cancellations in a working week of five days.

d In a working week of five days, find the probability that there will be four or more cancellations on exactly three of these days.

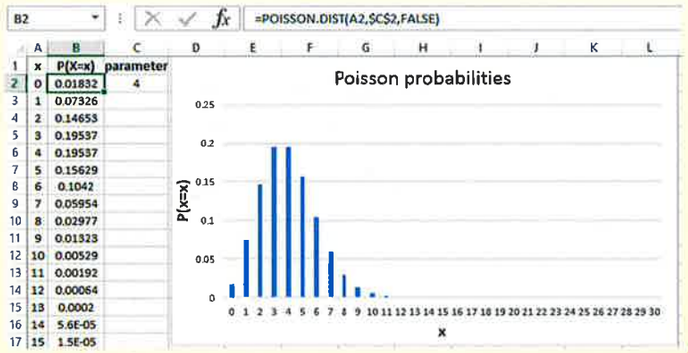
Continued on next page

a	$P(C \geq 4) = 1 - P(C \leq 3)$ $P(C \geq 4) = 0.203$	Interpret the problem, write down the event and apply complementary events to solve a smaller, related problem. Use technology to find the probability and write your answer to three significant figures.
b	Let M be the number of cancellations in March. Then $M \sim \text{Po}(71.61)$. $P(M \geq 81) = 0.147$	Write down the distribution with mean 2.31×31 since March has 31 days. Use technology to find the probability and write your answer to three significant figures.
c	One way is to find $(P(C = 0))^5$ where $C \sim \text{Po}(2.31)$, and the other is to find $P(W = 0)$ where $W \sim \text{Po}(11.55)$. Both methods give the probability 0.00000964 to three significant figures.	Apply the laws of independent events. Model the number of cancelled trains in one working week.
d	Let D be the number of days in a working week of five days on which there are at least four cancellations. Then $D \sim \text{B}(5, 0.203)$. $P(D = 3)$ is required. $P(D = 3) = 0.0531$	Define a random variable completely and clearly. Apply the binomial distribution to model the five days as independent trials with a probability of success 0.203 and write down the distribution and the event.

You have learned how to find the expected value and the variance of the binomial distribution. You can use technology to explore the mean and variance of the Poisson distribution.

Investigation 7

You can use technology to visualize the Poisson distribution with a spreadsheet. Enter the formula =POISSON.DIST(A2,C2,FALSE) in cell B2. Copy and drag down to row 31 to have 30 values of the probability distribution.



1 Use your spreadsheet and the formula for expected value, $E(X) = \sum_x xP(X = x)$, to find the expected value of the Poisson distribution for your chosen parameter.

Are 30 enough values to get an accurate result? Justify your answer.

Alter the value of the parameter in order to see the shape of the probability distribution function.

2 When is the shape of the distribution skewed? When is it more symmetrical?

3 **Factual** How do the mean and variance of the distribution change when the parameter is changed?

4 **Conceptual** How can you infer directly from the definition of the Poisson model that the parameter is the mean?

5 How can you infer from the shape of the distribution that the variance is related to the mean?

6 **Conceptual** What does the parameter of the Poisson distribution model?

If $X \sim \text{Po}(\alpha)$, then $E(X) = \text{Var}(X) = \alpha$
The proof of this result is beyond the scope of this book.

Example 8

The random variable T is modelled by a Poisson distribution. Given that $P(T > 2) = 0.53$, find the variance of T.

Let the Poisson parameter be λ $P(X > 3) = 1 - P(X \leq 3) = 0.53$ $P(X \leq 3) = 0.47$ $\lambda = 3.82$ Variance $= 3.82$	This equation can be solved using the cumulative Poisson probability function on the GDC with the Poisson parameter as the unknown variable. The variance is equal to the parameter for the Poisson distribution.

Exercise 13C

1 Use technology to find the following probabilities given that $A \sim \text{Po}(6.2)$.

a $P(A = 2)$ **b** $P(A < 6)$

c $P(A \geq 7 | A > 5)$

d $P(A$ is no more than $4)$

e $P(A$ is more than 8 given that A is at least $3)$

2 Given $Z \sim \text{Po}(\beta)$ and $P(Z = 0) = 0.301$, find the variance of Z.

3 Show that if $Y \sim \text{Po}(\alpha)$ and $\text{P}(Y = 1) = 0.15$ then there are two possible solutions for α. Explain how the two solutions both apply correctly to the event $\text{P}(Y = 1) = 0.15$.

4 A quantity of 278 pumpkin seeds are put in a dough mixture used to make 10 loaves of bread and mixed thoroughly. Let S be the number of pumpkin seeds in a loaf. State the probability distribution of S and any assumptions made. Find the most likely number of pumpkin seeds found in a loaf.

5 In Example 7, the train regulators decide to investigate how to punish poor punctuality with a penalty payment if there are four or more cancellations in a day and reward good punctuality for no cancellations with a bonus reward according to this table:

Number of cancellations in a day	4 or more	1, 2 or 3	0
Consequence	Penalty of £10 000	No penalty or reward	Bonus of ??

Given that the distribution of C, the number of cancellations per day, is Po(2.31) calculate the bonus payment that would make this fair.

6 The number of telephone calls per ten minutes to an IT support helpline is recorded in this table:

Number of calls	0	1	2	3	4	5 or more
Number of hours	15	30	28	14	7	8

Investigate whether this data appears to be modelled by a Poisson distribution.

7 Assume that the number of bacteria B in a petri dish are modelled by a Poisson distribution with a mean of two bacteria per square cm. A microbiologist selects two distinct areas of a petri dish at random and counts the number of bacteria. The first area measures 4.2 cm^2 and the second 1.7 cm^2.

Find the probability there are no bacteria in the petri dish by:

a considering each of the areas separately

b considering the two areas as a single area.

Verify that you get the same answer in each case.

8 The number of cars breaking down on a motorway is modelled by a Poisson distribution. The average number of breakdowns per kilometre of motorway is 0.597 per hour. Find the probability that:

a there are no breakdowns on a 5 km length of motorway in one hour

b there are three or more breakdowns on a 10 km length of motorway in a day

c there are fewer than 11 breakdowns per day on a 1 km section of motorway on at least five days of a week.

9 A car hire company has a fleet of ten cars that it hires out by the day. The number of requests for a day hire R is modelled by a Poisson distribution with mean 5.1 per day.

a Find the number of days in a year on which the owners of the company expect to have no custom.

b Find the number of days in 100 days on which the owners of the company expect to have to turn away customers because all the cars are rented out.

Developing inquiry skills

Look back at the opening scenario. Do any of the situations in the tale of Chancer's café involve the Poisson distribution?

TOK

A model might not be a perfect fit for a real-life situation, and the results of any calculations will not necessarily give a completely accurate depiction. Does this make it any less useful?

13.4 Modelling measurements that are distributed randomly

So far you have studied two discrete probability distributions: the binomial and Poisson distributions, and how these distributions can model real-world data. While discrete random variables are quantities that can be **counted** using integers, there are many data sets in fields of nature, society and science that consist of **measurements** using real numbers. For example, the height Y metres of an adult human chosen at random is an example of a **continuous** random variable. The reaction times of a learner driver and the speeds of cars on a highway are other examples.

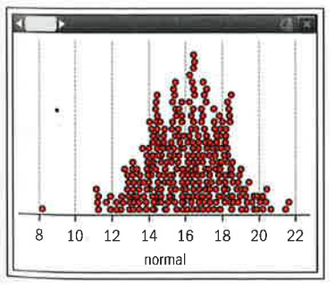
Terminology	**Explanation**
Y is a **continuous random variable.**	Continuous: Y can be found by measuring and is therefore a real number.
	Random: Y is the result of a random process.
	Variable: Y can take any value in a domain which is a subset of $\mathbb{R}$.
	If Y is the height of an adult human, Y could take any values such that $0.67 \leq Y \leq 2.72$ according to the *Guinness Book of World Records*.

Y can be measured to various degrees of accuracy and can take any value in its domain, hence the sample space of Y cannot be represented as a list of numbers that can be counted. You will learn how to model one example of a continuous random variable in this section: the normal distribution.

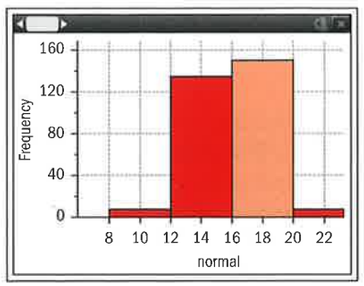
Consider the lifetimes L of 300 batteries measured to the nearest second in a quality control investigation. The data can be presented in many ways.

International-mindedness

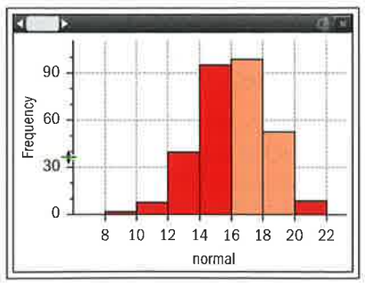
The Galton board, also known as a quincunx or bean machine, is a device for statistical experiments named after English scientist Sir Francis Galton. It consists of an upright board with evenly spaced nails or pegs driven into its upper half, where the nails are arranged in staggered order, and a lower half divided into a number of evenly spaced rectangular slots. In the middle of the upper edge, there is a funnel into which balls can be poured. Each time a ball hits one of the nails, it can bounce right or left with the same probability. This process gives rise to a binomial distribution of in the heights of heaps of balls in the lower slots and the shape of a normal or bell curve.



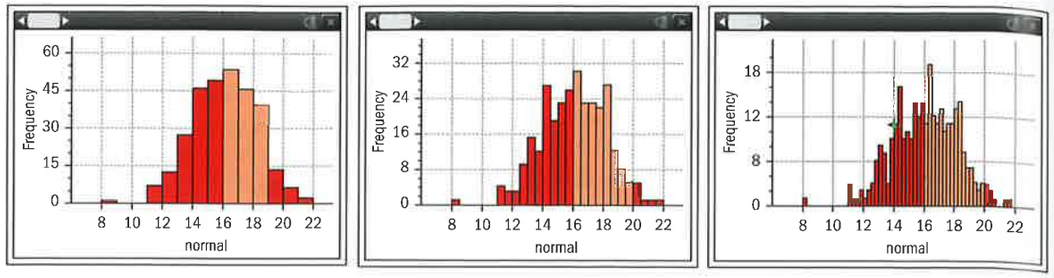
A dot plot of the 300 data points. Note the symmetry and the fact that batteries lasting very long or very short periods of time are very rare.



A frequency histogram with the interval $16 \leq L < 20$ showing a frequency of 150. The data is broadly symmetric in this representation.



The interval $16 \leq L < 18$ has a frequency of 98 and $18 \leq L < 20$ has a frequency of 52, confirming the total frequency of 150 for $16 \leq L < 20$.



These three histograms have increasingly small widths to their bars. All three histograms have a symmetric shape and total frequency of 150 for $16 \leq L < 20$. The symmetric pattern fluctuates as the class intervals change in width.

The distribution of the battery lifetimes, shown above, display the characteristics of the normal distribution. When a data follows a normal distribution most of its values lie close to an average value and as you move away from the average there will be fewer and fewer values.

Probability density functions

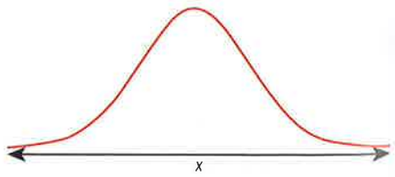
All continuous random variables have an associated **probability density function**. This has the property such that if X is a continuous random variable with probability density function f then $\mathrm{P}(a < X < b) = \int_a^b f(x)\mathrm{d}x$.

Because $f(x) \geq 0$ for all values of x (as you cannot have negative probabilities) the $\mathrm{P}(a < X < b)$ can also be thought of as the area under the curve $y = f(x)$ between a and b.

The probability density function for a normal distribution has the equation

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

Some properties of the curve depend on the parameters μ and σ but all curves with this probability density function have the same basic shape, often referred to a **bell curve**.



If you look back at the graphs for the distribution of battery life times you can see this shape is probably best illustrated by the histogram with a bar width of 0.5. If battery life did follow a normal distribution you would expect as the sample size increased the bar chart of frequencies would fit this theoretical model increasingly well.

Investigation 8

Using the dataset you are going to analyse W, the length from carpal joint to wing tip, of 4000 blackbirds.

1 Is W a discrete random variable?

	E	F	G	H	I	J	K	L
	Wing	Weight	Day	Month	Year	Time		
	133	95	21	12	2006	14		114
	134	106	25	11	2012	9		144
	135	125	29	1	1994	9		112
	135	113	5	2	1994	10		114
	135	111	12	2	1994	8		116
	134	105	15	2	1994	8		118
	136	111	11	2	2004	8		120
	127	103	23	2	2004	13		122
	127	102	26	2	2004	9		124
	126	104	25	2	2004	11		126
	125	97	26	2	2004	9		128
	135	102	27	2	2004	10		130
	135	126	27	2	2004	15		132
	135	121	2	3	2004	12		134
	125	88	27	5	2004	14		136
	123	90	8	6	2004	17		138
	135	101	30	6	2004	13		140
	129	100	7	9	2004	8		142
	129	98	4	3	2005	16		144
	129	97	6	3	2005	14		146

In cell L2 (green) type "=min(E2:E4001)".

In cell L3 (blue) type "=max(E2:4001)".

These give the maximum and minimum values of W in this sample of 4000.

Type 112, 114, 116 as shown and drag down to 146 in cell L21.

These will give your class intervals for a histogram. These are referred to as bins by the software.

Use the output to create a histogram of the 4000 values of W.

2 Does the histogram for W show a symmetric bell-shaped curve?

3 Can it be modelled by a normal distribution?

Repeat the experiment for E, the life expectancy at birth data for 224 countries, found on the CIA website: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html

4 Factual Is E a continuous random variable?

5 Factual Does the histogram for E show a symmetric bell-shaped curve?

6 Factual Can it be modelled by a normal distribution?

Repeat the experiment using other data sets that interest you.

7 Which of these measurements do you feel could be modelled by a normal distribution?

- **a** Baby birth weight
- **b** Age of mother of new baby
- **c** Number of hairs on head of a new baby
- **d** Number of toes on a new baby
- **e** Annual salary of adults aged between 20 and 25 years
- **f** Life expectancy in Sweden
- **g** The number of births on each day of January 1978 in the USA
- **h** Age of humans in Haiti
- **i** Journey time of a delivery van
- **j** Height of sunflowers
- **k** Annual salary of professional footballers
- **l** IQ scores of 2501 undergraduate students

8 Conceptual In which contexts do you expect the normal distribution to be an appropriate model?

Investigation 9

Use your GDC to explore the parameters of the normal distribution by varying the values of the mean and the standard deviation in the function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ and seeing how this affects the shape of the curve.

1 **Conceptual** What does the shape of $f(x)$ tell you about where the probability is distributed most/least densely in the normal distribution?

2 What are the coordinates of the maximum point of the function in terms of μ and σ?

3 What is the equation of the asymptote of the function?

4 **Factual** Which parameter affects the position of the axis of symmetry of the function?

5 **Factual** Which parameter affects the gradient of the function?

6 **Factual** In a data set, how do you quantify the central tendency of the data? How do you quantify the spread of the data?

7 **Factual** What letters do we use to represent mean and variance?

8 **Factual** What is the normal distribution function?

9 **Conceptual** What do the parameters of $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ model?

A normal distribution is defined by the parameters μ and σ, where μ is the mean of the distribution and σ is its standard deviation.

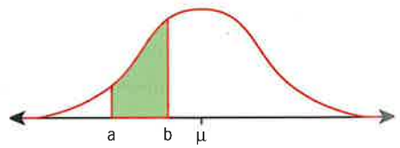
If X is distributed with a mean of μ and a standard deviation of σ we would write $X \sim N(\mu, \sigma^2)$.

Notice that this notation gives the variance rather than the standard deviation.

Finding probabilities

$P(a < X < b)$ is the area under the probability density function of X between the values of a and b.

This area cannot be found by integrating the probability density function, but all GDCs will have a function that allows the area, and hence the probability, to be calculated if the mean and standard deviation are known.



Example 9

If $X \sim N(10, 4)$ find

a $P(9 < X < 12)$ **b** $P(X < 13)$ **c** $P(X > 7)$.

a $P(9 < X < 12) = 0.533$ **b** $P(X < 13) = 0.933$ **c** $P(X > 7) = 0.933$	Most GDCs will require you to input the standard deviation rather than the area, so enter $\sigma = 2$. When sketching a normal curve the mean will always be on the line of symmetry, hence a sketch will demonstrate why the answers to parts **b** and **c** are equal.

Example 10

The lengths of trout in a fish farm are normally distributed with a mean of 39 cm and a standard deviation of 6.1 cm.

a Find the probability that a trout caught in the fish farm is less than 35 cm long.

b Cliff catches five trout in an afternoon. Find the probability that at least two of the trout are more than 35 cm long. State any assumptions you make.

c Find the probability that a trout caught is longer than 42 cm given that it is longer than 40 cm.

d Determine if the events $L > 42$ and $L > 40$ are independent.

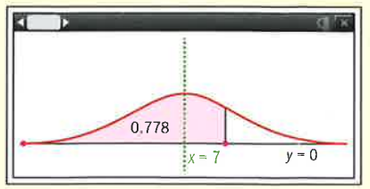
a Let L represent the length of a randomly selected trout. Then $L \sim N(39, 6.1^2)$. $P(L < 35) = 0.256$	Write down the random variable, the distribution and the event to clarify your thoughts and to demonstrate knowledge and understanding. $P(L < 35)$ can be found on a GDC.
b Let C represent the event that five of the fish are more than 35 cm long. Then $C \sim B(5, 0.744)$, assuming that the length of each fish caught is independent of the others. $P(C \geq 2) = 0.983$	Five fish caught can be represented as five trials. Write down the distribution and the event to clarify your thoughts and to demonstrate knowledge and understanding of why this particular distribution was chosen. Demonstrate understanding that $1 - 0.256 = 0.744$ is the probability of success. $P(C \geq 2)$ can then be found on a GDC.
c $P(L > 42 \mid L > 40) = \frac{P(L > 42)}{P(L > 40)}$ $P(L > 42 \mid L > 40) = 0.716$	Apply the formula for conditional probability. $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
d Since $P(L > 42 \mid L > 40) = 0.716$ and $P(L > 42) = 0.311$, the events are not independent.	Apply the definition of independent events: If $P(A \mid B) = P(A)$ then A and B are independent.

Investigation 10

Given $X \sim N(7, 1.5^2)$, you can visualize a value $F(8)$ of the cumulative distribution function $F(8) = P(X \leq 8)$ on a sketch.

The shaded area shows the probability $P(X \leq 8) = 0.748$ quantified by $F(8)$.

1 In this diagram, why does the fact that $8 > 7$ guarantee that $P(X \leq 8) > 0.5$?



Use sketch diagrams to answer the following.

2 Find $P(X \geq 8)$ in terms of $F(8)$.

3 Find $P(X \leq 6)$ in terms of $F(8)$.

Continued on next page

4 Find $P(X \geq 6)$ in terms of $F(8)$.

5 Find $P(6 \leq X \leq 8)$ in terms of $F(8)$.

Check your answers using your GDC.

6 **Factual** What is the total area under the function?

7 How do $P(X \geq 6)$ and $P(X > 6)$ compare? Why?

8 Let $X \sim N(10,\ 1.7^2)$. Find the following probabilities.

$P(10-1.7 \leq X \leq 10+1.7), P(10-2\times1.7 \leq X \leq 10+2\times1.7), P(10-3\times1.7 \leq X \leq 10+3\times1.7)$

9 **Factual** Repeat question **8** with your own values of μ and σ. What do you notice?

Write your findings in a table:

If $X \sim N(\mu, \sigma^2)$, then:	
$P(\mu-\sigma \leq X \leq \mu+\sigma) =$	
$P(\mu-2\sigma \leq X \leq \mu+2\sigma) =$	
$P(\mu-3\sigma \leq X \leq \mu+3\sigma) =$	

10 **Conceptual** Given an interval of values of the random variable $X \sim N(\mu, \sigma^2)$, what does the cumulative probability function F quantify?

Up to this point we have been given values of the variable X and been asked to find probabilities. It is also possible to find a value of X given a probability.

If $F(x) = P(X < x) = p$ for some probability p, then we can write

$F^{-1}(p) = x$

This function is often referred to as the inverse cumulative normal function and is on all GDCs.

Example 11

For $X \sim N(21,9)$,

1 find x given that:

a $P(X < x) = 0.8$ b $P(X > x) = 0.4$

2 a find a and b given that $P(a < X < b) = 0.68$ and a and b are an equal distance either side of the mean.

b Verify that this supports the statement that approximately 68% of all data for a normally distributed population is likely to lie within one standard deviation of the mean.

1 a $x = 23.5$ **b** $P(X > x) = 0.4 \Rightarrow P(X < x) = 0.6$ $x = 21.8$	This is obtained from a GDC using the inverse cumulative normal function. Some GDCs can work out the value of x without converting to $F(x)$, but for many finding $P(X < x)$ is the first stage.

2 a $P(X < b) = 0.5 + 0.34 = 0.84$ $b = 24.0$ $a = 18.0$ **b** The two values are 21 ± 3	From a sketch it can be seen the $P(X < b)$ is half the whole area under the curve, plus half of 0.68.

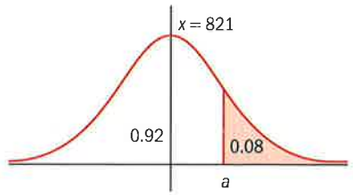
Example 12

The weights of cauliflowers purchased by a supermarket from their suppliers are distributed normally with mean 821 g and standard deviation 40 g.

Cauliflowers weighing less than 750 g are classified as small.

a Predict the number of cauliflowers classified as small in a sample of 400 cauliflowers.

b The heaviest 8% of cauliflowers are classified as oversized and re-packaged. Find the range of weights of cauliflowers classified as oversized.

a Let W represent the weight of a randomly selected cauliflower. Then $W \sim N(821, 40^2)$. The expected number of cauliflowers classified as small is: $400 \times P(W \leq 750) = 15.1796$ 15 cauliflowers are predicted to be classified as small.	Write down the random variable and the distribution to clarify your thoughts and to demonstrate knowledge and understanding. Apply the formula for the expected number of occurrences and use technology to find the probability.
b  The value of a is 877.203. Cauliflowers weighing at least 877 g will be classified as oversized.	A sketch helps you orientate the answer in the correct place. You can see already that the lower limit for classification as oversized must be greater than 821. You have the probability, so you will need to use the inverse cumulative normal distribution on your GDC. Take care to use the correct cumulative probability of 0.92. Interpret the result to state the range.

International-mindedness

The normal curve is also known as the Gaussian curve, and is named after the German mathematician, Carl Friedrich Gauss (1777–1855), who used it to analyse astronomical data. This is seen on the old 10 Deutsche Mark notes.

Exercise 13D

1 $T \sim N(17.1, 3.1^2)$. Estimate these probabilities without technology:

a $P(T < 17.1)$ **b** $P(T < 14)$

c $P(T > 20.2)$ **d** $P(14 \leq T < 23.3)$

e $P(T < 7.8)$ **f** $P(T < 23.3 | T > 20.2)$

2 $Q \sim N(4.03, 0.7^2)$. Find these probabilities with technology:

a $P(Q < 4)$ **b** $P(Q < 3.4)$

c $P(Q > 5)$ **d** $P(3.5 \leq Q < 4.5)$

e $P(T < 4.9 | T > 2.9)$

3 $R \sim N(22\,129, 300^2)$. Find the interquartile range of R, ie the values between which lie the middle 50% of the distribution.

4 $S \sim N(0, 1.35^2)$. Find the value of k if $P(|S| > k) = 0.57$

5 A random variable X is distributed normally with mean 372 and standard deviation 13.

a Find $P(X \leq 381)$. Show your answer on a sketch.

b Given that $P(X > t) = 0.17$, find t. Show your answer on a sketch.

6 An electronics company produces batteries with a lifespan that is normally distributed with mean 182 days and a standard deviation of 10 days.

a Find the probability that a randomly selected battery lasts longer than 190 days.

b In a sample of seven batteries chosen for a quality control inspection, find the probability that no more than three of them last longer than 190 days.

c If a battery is guaranteed to last up to 165 days, find the probability that the battery will cease to function before the guarantee runs out.

d Hence predict the number of batteries in a batch of 10 000 that would not last the duration of the guarantee.

7 The distance travelled to and from work each day by employees in a central business district is modelled by a normal distribution with mean 16 km and standard deviation 5 km.

a Find the probability that a randomly chosen employee travels between 13 km and 15.3 km each day.

b 13% of employees travel more than x km each day to and from work. Find the value of x.

c Records show that when snow falls, 91% of employees who live further than 14 km from the central business district will fail to get to work. Assuming that all employes who live closer than 14 km do get to work, predict how many of the 23 109 employees will fail to get to work on a snow y day.

8 A courier service in a city centre analyzes their performance and finds that the delivery times of their drivers are normally distributed with a mean of 23 minutes and a standard deviation of 5 minutes. The management of the courier service want to promote "Delivery to you within m minutes or your money back!" in an advertising campaign. What value of m should they choose in order to be at least 99% sure not to pay the customer their money back?

9 A nurse has a daily schedule of home visits to make. He has two possible routes suggested to him by an app on his phone for the journey to his first patient, Nur. Assume that the journey times are normally distributed in each case.

Route A has a mean of 42 minutes and a standard deviation of 8 minutes.

Route B has a mean of 50 minutes and a standard deviation of 3 minutes.

a Distinguish between the advantages and disadvantages of each route.

b The nurse starts his journey at 8.15am and must be at Nur's house by 9.00am. Which route should he take?

c If on five consecutive days, the nurse leaves home at 8.15am and takes route A, find the probability that he arrives at Nur's house:

i by 9.00am on all five days

ii by 9.00am on at least three of the five days

iii by 9.00am on exactly three consecutive days.

Developing inquiry skills

Return to the opening problem.

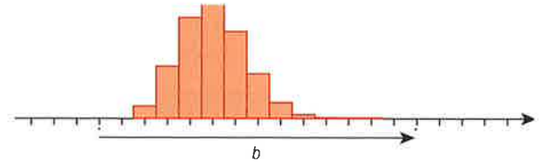
Do any of the situations in the café involve the normal distribution?

13.5 Mean and variance of transformed or combined random variables

You have learned how to transform coordinates and graphs of functions.

You can transform random variables too: a random variable X may be transformed to a related random variable $Y = aX + b$ where $a, b \in \mathbb{R}$. This is a **linear transformation** of the random variable X.

Consider a random variable X, with the distribution shown below. If a fixed value of b is added to each of the possible values of X we would obtain the distribution $X + b$ shown on the right of the diagram.



It can be seen that the expected value of the new distribution will be greater by the value b compared with X, hence $E(X + b) = E(X) + b$.

As the distribution is no more spread out than it was previously and as $Var(X)$ is a measure of spread it follows that $Var(X + b) = Var(X)$.

Now consider multiplying all the possible values of X by a. As all the values of X are multiplied by a then the expected value will also be multiplied by a. On this occasion the spread of the values will also increase by a factor of a. For example, if the lowest value was 4 and the highest 10 then the range would be 6. After the transformation the lowest value would be $4a$ and the highest $10a$ so the range would be $6a$. Hence the standard deviation would also be multiplied by a and so the variance by a^2.

This can be summarized as $E(aX + b) = aE(X) + b$ and $Var(aX) = a^2 Var(X)$.

Combining the ideas above we obtain

$E(aX+b) = aE(X)+b$ and $\text{Var}(aX+b) = a^2\text{Var}(X)$

Example 13

a T is a discrete random variable with $E(T) = 4.01$ and $\text{Var}(T) = 1.1$. Find:

i $E(2.5T+1)$ **ii** $\text{Var}(5T-2)$ **iii** $E(-0.5T-3)$ **iv** $\text{Var}(0.2T+4)$

b $C \sim B(7, 0.2)$. Find: **i** $E(3C+2)$ **ii** $\text{Var}(0.9T+1.2)$

c $D \sim \text{Po}(4.2)$. Find: **i** $E(-1.7D+5.1)$ **ii** $\text{Var}(0.9D+1.2)$

a i $E(2.5T+1) = 2.5\times4.01+1 = 11.025$ **ii** $\text{Var}(5T-2) = 5^2\times1.1 = 27.5$ **iii** $E(-0.5T-3) = -0.5\times4.01-3 = -5.005$ **iv** $\text{Var}(0.2T+4) = 0.044$	Apply the formulae for the expected value and variance of a linear transformation of a random variable.
b $E(C) = 7\times0.2 = 1.4$ $\text{Var}(C) = 7\times0.2\times0.8 = 1.12$, hence **i** $E(3C+2) = 3\times1.4+2 = 6.2$ **ii** $\text{Var}(0.9T+1.2) = 0.9^2\times1.12 = 0.9072$	Apply the formulae for the expected value and variance of the binomial distribution.
c $E(D) = 4.2 = \text{Var}(D)$ **i** $E(-1.7D+5.1) = -1.7\times4.2+5.1 = -2.04$ **ii** $\text{Var}(0.9D+1.2) = 0.9^2\times4.2 = 3.402$	Apply the formulae for the expected value and variance of the Poisson distribution.

You can combine two or more random variables to make a new random variable, just as you can combine two functions to make a new function. For example two independent random variables X and Y may be combined to make a new random variable $Z = aX + bY$ where $a, b \in \mathbb{R}$.

Z is a **linear combination** of the random variables X and Y.

The results of the previous section can be applied to a linear combination of random variables. If X and Y are two random variables:

$E(aX \pm bY) = aE(X) \pm bE(Y)$

$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$

In the case of the variance we need X and Y to be independent for this result to hold. Whether you add or subtract the two variables the variances are always added as demonstrated below:

$\text{Var}(X-Y) = \text{Var}(X+(-Y)) = \text{Var}(X)+\text{Var}(-Y)$

$= \text{Var}(X)+(-1)^2\,\text{Var}(Y) = \text{Var}(X)+\text{Var}(Y)$

International-mindedness

French mathematicians Abraham De Moivre and Pierre Laplace were involved in the early work on the applications of the normal curve.

De Moivre developed the normal curve as an approximation of the binomial theorem in 1733 and Laplace used the normal curve to describe the distribution of errors in 1783, and in 1810 to prove the Central Limit Theorem.

The results above can be generalized to these results for linear combinations of n independent random variables X_i, $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$.

$E(a_1X_1+a_2X_2+\ldots+a_nX_n) = a_1E(X_1)+a_2E(X_2)+\ldots+a_nE(X_n)$ and

$\text{Var}(a_1X_1+a_2X_2+\ldots+a_nX_n) = a_1^2\text{Var}(X_1)+a_2^2\text{Var}(X_2)+\ldots+a_n^2\text{Var}(X_n)$.

The second result (with variances) is only valid when the random variables are independent.

Example 14

$A \sim \text{Po}(2.3)$, $M \sim B(8, 0.2)$ and $U \sim N(11.7, 1.7^2)$ are three independent random variables. Calculate:

a $E(3M-2U)$ **b** $\text{Var}(A-3M+0.7U)$.

a $E(M) = 8\times0.2 = 1.6$, $E(U) = 11.7$ $E(3M-2U) = 3E(M)-2E(U) = 3\times1.6-2\times11.7 = -18.6$	Apply your knowledge of the means of the binomial and normal distributions. Apply the formula for the expected value of a linear combination of two independent random variables.
b $\text{Var}(A) = 2.3$, $\text{Var}(U) = 1.7^2$ $\text{Var}(M) = 8\times0.2\times0.8 = 1.28$ $\text{Var}(A-3M+0.7U)$ $= \text{Var}(A)+9\text{Var}(M)+0.49\text{Var}(U)$ $= 2.3+9\times1.28+0.49\times1.7^2$ $= 15.2361 \approx 15.2$	Apply your knowledge of the variance of the Poisson, binomial and normal distributions. Take care to show the steps in your working because arithmetical errors are easy to make. Give the answer to three significant figures.

Exercise 13E

1 The random variable F is such that $E(3F+1) = 6$ and $\text{Var}(3.1-2F) = 7$

Calculate **a** $E(F)$ **b** $\text{Var}(F)$

2 C and D are two independent Poisson random variables such that $E(C) = 3$ and $\text{Var}(D) = 6.1$. Calculate:

a $E(9C-4D)$ **b** $\text{Var}(D+0.2C)$

For questions **3**, **4**, **5**, and **6** use the following information:

Random variable	Expected value	Variance
U	4.01	1.2
V	2.7	0.4
W	12.9	3
X	7.81	2.11

3 Calculate:

a $E(2U+9)$

b $E(4X-0.1)$

c $\text{Var}(0.9W+10)$

d $\text{Var}(7-3V)$

e $E(U+4X-2W)$

f $\text{Var}(2V+0.8U-0.9X+W)$

4 Given that $a > 0$, $E(aU+b) = 5.1$ and $\text{Var}(aV+b) = 2$, calculate a and b.

5 Given that $E(aW+bX) = 6$ and $E(aU+bV) = 2$, calculate a and b.

6 Given that $a > 0$, $E(aW+bX) = 15.1$ and $\text{Var}(aU+bV) = 2$, calculate a and b.

7 Denise has a fair three-sided spinner numbered 1, 2 and 2. David has a fair four-sided spinner numbered 3, 4, 5 and 5. Let T represent the number scored on Denise's spinner and H the number scored on David's.

a Construct probability distribution tables for T and for H.

b Hence find:

i $\mathrm{E}(T)$

ii $\mathrm{E}(H)$

Denise and David play a game in which they spin their spinners and add the numbers obtained. Let S represent the sum of the two numbers obtained.

c Find $\mathrm{E}(S)$.

d Construct the probability distribution table for $S = T + H$ and use it to confirm your answer to part **c**.

8 Zeinab carries out an experiment to investigate two games she is designing. She uses two fair four-sided dice numbered 3, 4, 5 and 5.

In the first game, she takes her two dice, throws them and records the total as K.

In the second game, she throws one of her dice, doubles the number obtained and records the answer as L.

Karim argues that Zeinab is wasting her time thinking of these as two different games because the distributions of K and of L are identical, but Zeinab says otherwise. Determine who is correct and for what reason.

Developing inquiry skills

Return to the opening problem.

Do any of the situations in the café involve a linear combination of random variables?

13.6 Distributions of combined random variables

In section 13.5 you found the mean and variance of linear transformations of a random variable and of linear combinations of two or more random variables. In this section you will find out more about what distributions may be followed by random variables that are the results of linear transformations or combinations of other random variables.

For example, you can investigate the distribution of a sum of two independently distributed Poisson random variables using an intuitive argument and an exploration of data as follows.

TOK

How well do models, such as the Poisson distribution, fit real-life situations?

Investigation 11

Dulcinea counts the number of wild flowers in two fields as part of her biology fieldwork. She finds that the number of wild flowers in field A follows a Poisson distribution with mean 3 flowers per square metre whereas the number of wild flowers in a separate field B follows a Poisson distribution with mean 1.5 flowers per square metre. Dulcinea conjectures that if $A \sim \mathrm{Po}(3)$ and $B \sim \mathrm{Po}(1.5)$ then if $T = A + B$, T must follow a Poisson distribution. She searches for evidence with which to justify her conjecture.

1 **Factual** What are $\mathrm{E}(A+B)$ and $\mathrm{Var}(A+B)$?

Is this information consistent with Dulcinea's conjecture? Discuss.

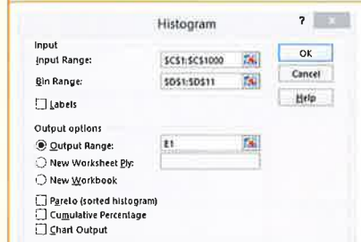
2 Since A and B both satisfy the conditions for the Poisson distribution and the two fields are separate and hence independent of each other, does it follow that $T = A + B$ must also satisfy the Poisson conditions? Discuss.

Dulcinea says she can't believe her conjecture until she sees some statistical evidence with her own eyes.

Follow these steps to create the same data sets as Dulcinea.

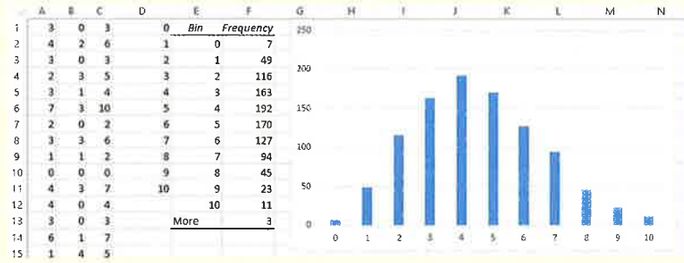
- Choose Data from the main menu and Data Analysis from the far right of the screen.
- Choose Random Number Generator and fill in the dialogue box as below to simulate 1000 Poisson experiments with mean (λ) 3 in the cells A1: A1000.

Repeat these steps to place 1000 random Poisson experiments with mean 1.5 in cells B1:B1000.
In cell C1 type "=A1+B1" and drag down to cell C1000.



Type 0, 1, 2, 3, …, 10 in cells D1:D11.
These are the classes for the grouped frequency data that will be put in a table with the top left-hand corner in cell E1.

- Choose Data from the main menu and Data Analysis from the far right of the screen.
- Choose Histogram.



Bin	Frequency
0	7
1	49
2	116
3	163
4	192
5	170
6	127
7	94
8	45
9	23
10	11
More	3

Dulcinea wants to check if the histogram that follows from the grouped frequency table really is from a Poisson distribution. Use the same steps to create 1000 Poisson experiments with mean 4.5 in column G and add this data to your histogram and compare.

3 **Conceptual** Given two independently distributed Poisson random variables $X \sim \mathrm{Po}(\lambda)$ and $Y \sim \mathrm{Po}(\alpha)$, what can you predict about the distribution of $Z = X + Y$?

Care must be taken when combining the average rates of two Poisson distributions, as shown in the next example.

Example 15

Richard is waiting for a bus. There are two bus companies, K Bus and L Express, who serve his bus stop. The numbers of arrivals follow Poisson distributions. On average, there is one bus every 5 minutes arriving from K Bus and one bus every 12 minutes from L Express.

Richard can use either company to get to his destination.

Assuming that the buses arrive independently of each other, find the probability that at least two buses arrive in a 10-minute interval.

On average two K Buses arrive every 10 minutes. On average $\frac{5}{6}$ L Express buses arrive every 10 minutes. Then $K \sim \text{Po}(2)$ and $L \sim \text{Po}\left(\frac{5}{6}\right)$.	Use proportionality to find the parameters required for the distribution of the number of buses every 10 minutes. Write down the distributions to clarify your thoughts and show knowledge and understanding.
Let $T = K + L$. Then $T \sim \text{Po}\left(2\frac{5}{6}\right)$ We require $\text{P}(T \geq 2)$. $\text{P}(T \geq 2) = 0.775$	Write down the event. Find the answer with technology.

You can investigate the distribution of a sum of two independently distributed normal random variables in a similar way.

Investigation 12

Milagros explores the time taken for her daily commute. She found that the time taken for her bus to arrive T is distributed normally with mean 5 minutes and standard deviation 1.7 minutes. The time Y for all the passengers to embark and pay for their tickets before the bus can depart is distributed normally with mean 0.5 minutes and standard deviation 0.2 minutes. Milagros conjectures that the total time taken for her bus to arrive and then begin its journey $S = T + Y$ must be distributed normally as well but she wants to see evidence.

1 Find $\text{E}(S)$ and $\text{Var}(S)$.

Milagros uses technology to create a sample of 1000 random numbers T such that $T \sim \text{N}(5, 1.7^2)$ and saves this in a list called t. She then creates a sample of 1000 random numbers Y such that $Y \sim \text{N}(0.5, 0.2^2)$ and saves this in a list called y.

She adds the lists to find a new list $s = t + y$, and then finds the mean and the standard deviation of s.

2 Do her findings confirm your answer for **1**?

Milagros constructs a histogram of the data in s.

Use technology to investigate if a normal model fits the histogram.

3 Does the mean and standard deviation of the normal model for s fit your prediction?

4 **Factual** Is the normal distribution an appropriate model for $S = T + Y$?

Milagros wants to explore further.

She wants to know if she can predict the distribution of a linear combination $S = aT + bY$ of two independent normally distributed random variables.

Investigate by using a variety of examples such as $S = 3T - 2Y$ and repeating the above process.

5 **Conceptual** Given two independently distributed normal random variables $X \sim \text{N}(\mu_1, \sigma_1^2)$ and $Y \sim \text{N}(\mu_2, \sigma_2^2)$, what can you predict about the distribution of $T = aX + bY$?

6 **Conceptual** How can you determine the probability distribution of a linear combination of n independent normal random variables?

Note that the proof of your results is outside the scope of this course.
The results of this investigation are very useful in applications and problem solving.

Example 16

Packing boxes used by a removal company come in three sizes: small, regular and large. After the boxes are filled weights S, R and L are all modelled by normal distributions with these parameters.

	Mean (kg)	Standard deviation (kg)
S	5	2
R	12.1	5
L	30.2	7

a One full box of each size is chosen at random. Find the probability that the large box weighs less than the regular box and three times the small box combined.

b One full large box and four small boxes are chosen at random. Find the probability that the total weight is more than 60 kg.

a $\text{P}(L < R + 3S)$ is required.

Write down the event.

$\text{P}(L < R + 3S) = \text{P}(L - R - 3S < 0)$

Rearrange the event into a linear combination of random variables.

$\text{E}(L - R - 3S) = \text{E}(L) - \text{E}(R) - 3\text{E}(S)$
$= 30.2 - 12.1 - 3 \times 5 = 3.1$

Apply the formula for the expected value of the linear combination.

$\text{Var}(L - R - 3S) =$
$\text{Var}(L) + \text{Var}(R) + 9\text{Var}(S) =$
$7^2 + 5^2 + 9 \times 2^2 = 110$

Apply the formula for the variance of the linear combination.

$L - R - 3S \sim \text{N}(3.1, 110)$

$\text{P}(L - R - 3S < 0) = 0.384$

State the distribution of the linear combination and find the probability of the event using technology, writing the answer to three significant figures.

Continued on next page

b Let the weights of the four small boxes be represented by S_1, S_2, S_3, S_4.

$P(L + S_1 + S_2 + S_3 + S_4 > 60)$ is required.

Write the event, showing that there are five separate independent random variables.

$E(L + S_1 + S_2 + S_3 + S_4) =$

$= E(L) + E(S_1) + E(S_2) + E(S_3) + E(S_4)$

$= 50.2$

Apply the formula for the expected value of the linear combination.

$Var(L + S_1 + S_2 + S_3 + S_4)$

$= Var(L) + Var(S_1) + Var(S_2) + Var(S_3) + Var(S_4)$

$= 7^2 + 2^2 + 2^2 + 2^2 + 2^2 = 65$

Apply the formula for the variance of the linear combination.

$L + S_1 + S_2 + S_3 + S_4 \sim N(50.2, 65)$

$P(L + S_1 + S_2 + S_3 + S_4 > 60) = 0.112$

State the distribution of the linear combination and find the probability of the event using technology, writing the answer to three significant figures.

Investigation 13

An important application of the distribution of a linear combination of independent normally distributed random variables is determining the distribution of the **sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ in which each X_i, $i = 1, 2, \ldots, n$ in a sample of size n are selected from a large population without replacement that follows a normal distribution whose parameters are known: $X \sim N(\mu, \sigma^2)$.

You can assume X_i, $i = 1, 2, \ldots, n$ are independent even though the sampling is without replacement in a large population.

Apply the results from earlier in the section to answer the following.

1 What is the distribution of $X_1 + X_2$?

2 What is the distribution of $X_1 + X_2 + X_3$?

3 What is the distribution of $X_1 + X_2 + \ldots + X_n$?

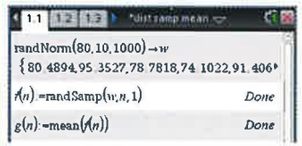
4 Hence find the **distribution of the sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Application: Adult passengers who book tickets for a flight can be thought of as a sample from a population. Assume that the adults booking tickets for a flight are from a population that follows the normal distribution with mean 80 kg and standard deviation 10 kg. You can visualise the distribution of the mean of the passengers' weights for different sizes of samples for each of these cases: commuter (19 passengers), regional (70 passengers), single-aisle (200 passengers) and double-aisle (400 passengers). Aircraft have a random selection of passengers on each flight and each aircraft has a total weight limit beyond which a flight is not safe and hence not permitted.

International-mindedness

Belgian scientist Lambert Quetelet applied the normal distribution to human characteristics (l'homme moyen) in the 19th century. He noted that characteristics such as height, weight and strength were normally distributed.

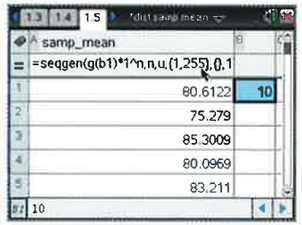
In this simulation you will create a population of 1000 passengers, select many samples of each size, find the mean of each sample and explore the distribution of your sample means.



A list of 1000 random numbers drawn from the normal population following $N(80, 10^2)$ is stored in the list w.

The function $f(n)$ is defined with the symbol ":=" as a list of n numbers selected without replacement from w.

The function $g(n)$ is defined as the mean of the numbers in the list $f(n)$

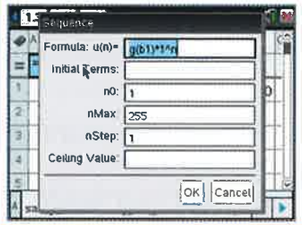


In a new spreadsheet, type 10 in cell B1.

This will be the size of your first samples.

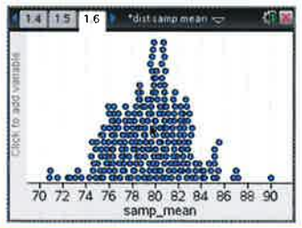
Define the list of data in column 1 as samp_mean.

You can type $seqgen(g(b1)*1\hat{}(n), n, u, \{1,255\}, \{\}, 1)$ or navigate to 'generate sequence' through the menu, as below



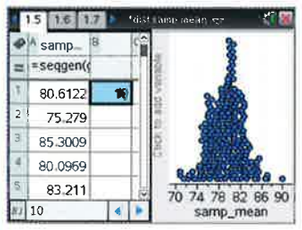
The $1\hat{}n$ instructs the GDC to re-calculate the sequence and hence take a new sample.

In total, 255 samples of size 10 are selected from the population of 1000 and their means calculated and placed in column A.



Add a new Data and statistics page.
Select samp_mean on the x-axis.

This dot plot shows the distribution of the 255 means of your 255 samples of size 10.



Group the page layout, so you can see the spreadsheet and the dotplot on the same screen.

Now change the value in cell b1 to 19 to simulate the distribution of sample means for a commuter aircraft.

Repeat for all the other sizes of aircraft.

Use the GDC to find the mean and standard deviation of your 255 sample means for each size of aircraft.

Continued on next page

5 As the sample size is increased, what happens to the mean of the distribution of sample means?

6 As the sample size is increased, what happens to the standard deviation of the distribution of sample means?

7 How does your answer for question 4 help explain your answer for questions 5 and 6?

8 **Factual** As the size of the aircraft increases, what is the effect on the confidence the pilot has about the average weight of the passengers in the aircraft?

9 **Factual** As the size of a sample increases, what happens to the probability that the sample mean will differ from the population mean by a given amount?

10 **Conceptual** What does the **distribution of sample means** of n independent normally distributed random variables help you to predict?

Save your work for use in the next investigation.

Example 17

A food standards authority survey finds that the number of calories in takeaway meals from a national restaurant chain *Speedfood* is distributed normally with a mean of 1900 calories and a standard deviation of 80 calories.

Assume that the recommended daily intake of calories is 2000 calories.

Find the probability of the following.

a A randomly chosen takeaway from *Speedfood* contains more than the recommended daily intake.

b The average calorie intake of a family of five resulting from eating a meal from *Speedfood* is more than the recommended daily intake of calories.

c Reflect on your results.

a Let C represent the calorific content in one randomly selected takeaway.	State the random variable, its distribution and the event.
Then $C \sim N(1900, 80^2)$.	
$P(C > 2000)$ is required.	
$P(C > 2000) = 0.106$	Give the answer to three significant figures.
b Let $\overline{C}$ represent the mean calorific content of the five randomly selected takeaway meals.	
Then $\overline{C} \sim N\left(1900, \frac{80^2}{5}\right)$	Apply the formula for the distribution of sample means.
$P(\overline{C} > 2000)$ is required.	
$P(\overline{C} > 2000) = 0.00260$	
c The mean calorific content of five meals is far less likely to be above the daily limit than the calorific content of a single meal. This is due to a smaller standard deviation of the sample mean.	

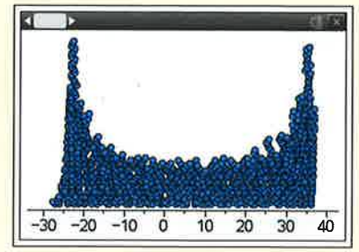
You have investigated the distribution of the mean of a sample of n from a large population without replacement that follows a **normal distribution** whose parameters are known. But many populations follow other distributions. You can find, in a similar way, the distribution of the sample mean when the population follows **any** distribution.

TOK

Discuss the statement "Without the central limit theorem, there could be no statistics of any value within the human sciences."

Investigation 14

Using technology, change the population away from a normal population to **any** type of data set of 1000.



As an example, first try using the function $30\sin(n) + \ln(n)$ to create the list then create your own. For each example, note the mean and the standard deviation of your population.

Plot the values created in the suggested example and see in this case that the distribution followed is not normal. For the purposes of this investigation, we will call this a "concave up" distribution.

Using a random integer list is another good example to try – it may be referred to as a "uniform" distribution.

Then carry on the same analysis for the distribution of sample means as in Investigation 13.

Include distributions that are skewed or ones that show no obvious shape at all.

Note that you will re-calibrate the axes on your GDC each time.
Use sample sizes of at least 30.

In each case, read off the parameters of the normal model the GDC fits to the histogram.

Use your own examples and data to fill in a table:

Distribution	Population mean	Population standard deviation	Parameters of distribution of sample mean with sample size n, μ_n and σ_n
Concave up			n / μ_n / σ_n: 30; 50; 100
Uniform			n / μ_n / σ_n: 30; 50; 100
...	...		n / μ_n / σ_n: 30; 50; 100
...			

1 **Factual** In every distribution you create, is the distribution of sample means modelled by a normal distribution when the sample size n is at least 30?

Continued on next page

2 **Factual** Do your findings support this statement of the **central limit theorem:** "The mean $\bar{X}$ of a sample of size n taken from **any** population with mean μ and standard deviation σ can be modelled by a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ provided that n is at least 30"?

3 How do the assumptions and conclusions of the central limit theorem compare and contrast with those of the distribution of the sample means of n independently distributed random variables?

4 **Conceptual** What does the **central limit theorem** help you to predict?

Example 18

The number of emergency calls per hour to a hospital between 9.30am and 10.30am each day follows a Poisson distribution with parameter 6.3. Use the central limit theorem to find the probability that in 40 days the mean of the number of calls between 9.30am and 10.30am is less than 5.4.

The population sampled from has mean 6.3 and variance 6.3.	Apply your knowledge of the Poisson distribution.
Let $\bar{X}$ represent the mean of the number of calls in the 40 periods of time, 9.30am to 10.30am.	Define the random variable.
Then $\bar{X} \sim \mathrm{N}\left(6.3, \frac{6.3}{40}\right)$. $\mathrm{P}(\bar{X} < 5.4)$ is required.	Apply the central limit theorem and state the event.
$\mathrm{P}(\bar{X} < 5.4) = 0.0117$	Find the probability required with technology and give the answer to three significant figures.

Example 19

Scientists are comparing the growth of two types of wheat. Type A has an mean height of 23.0 cm with a standard deviation of 3.1 cm. Type B has a mean height of 22.5 cm with a standard deviation of 4.1 cm.

a Find the probability that a sample of 50 blades of wheat taken from type A has a mean greater than 24 cm.

b Find the probability that a sample of 50 blades taken from sample A has a greater mean than a sample of 50 blades taken from sample B.

a Let X be the distribution of the height of wheat from type A	Even though we do not know the distribution of X we can use the central limit theorem to find the distribution of $\bar{X}$
$\bar{X} \sim \mathrm{N}\left(23.0, \frac{3.1^2}{50}\right)$	
$\mathrm{P}(\bar{X} > 24) = 0.0113$	
b Let Y be the distribution of height of wheat from type B	
$\bar{Y} \sim \mathrm{N}\left(22.5, \frac{4.1^2}{50}\right)$	

$\mathrm{P}(\bar{X} > \bar{Y}) = \mathrm{P}(\bar{X} - \bar{Y} > 0)$	Rearranging to form a single distribution.
The distribution of $\bar{X} - \bar{Y}$ is	
$\mathrm{N}\left(23.0 - 22.5, \frac{3.1^2}{50} + \frac{4.1^2}{50}\right)$	
$= \mathrm{N}(0.5, 0.5284)$	
$\mathrm{P}(\bar{X} - \bar{Y} > 0) = 0.754$	

Developing inquiry skills

Look back at the opening scenario. Add a paragraph to the tale of the café that involves the central limit theorem.

Exercise 13F

1 A cookery book is translated from Russian into English. Each page has two separate sections, Ingredients and Method, which are translated by different translators working independently. Assuming the number of translation errors in the Ingredients section and the Method section follow Poisson distributions with parameter 0.7 and 0.6 respectively, find the probability that on a given page there is at least one error.

2 The maximum load a lift can carry is 375 kg. The weights of adult males and females are distributed normally with means 85 kg and 60 kg and standard deviations 10 kg and 7 kg respectively. Find the probability that a load of three women and two men would be too heavy for the lift to function, assuming that the weights of the five adults are independent of each other.

3 An emergency food parcel consists of three bottles of water, two bananas and two bars of chocolate. The weights of all the items comprising the parcel are distributed normally as follows:

Item	Mean (g)	Standard deviation (g)
Water	300	2
Banana	180	7
Chocolate	100	2
Box	16	0.5

What is the probability that the entire food parcel weighs more than 1.5 kg?

4 Saida collects data to investigate her email inbox. Find the probability that in a 24-hour period she receives more than 40 emails assuming that the emails arrive independently of each other at these constant rates. Saida receives junk email at an average rate of 1.5 per hour and personal emails at a rate of 2 per day.

5 Lengths of bamboo pole are cut in a hardware store into three sizes called short, regular and long. The lengths of each size follow normal distributions with parameters given in the following table:

Size	Mean (cm)	Standard deviation (cm)
Short	40	2.1
Regular	80	3.7
Long	120	4

Find the probability that:

a One short and one regular laid end to end are longer than one long.

b Three short lengths laid end to end are shorter than one long.

6 A supermarket purchases cut flowers from two suppliers, A and B. The heights of the flowers have parameters as shown in this table:

Supplier	Mean height (cm)	Standard deviation (cm)
A	110	25.0
B	123	8.1

a A random sample of 40 cut flowers is taken from each supplier. Find the probability that the mean of the sample from A is greater than that from B.

b Find the probability that the sample mean from A is between 108 and 112 cm.

c In 150 samples each of 50 flowers from B, in how many would you expect the sample mean to be greater than 125 cm?

7 A cylindrical part of an engine is manufactured with a radius that is distributed normally with a mean of 5.1 mm and a standard deviation of 0.1 mm. The cylinder must fit in a circular hole which has a diameter that is distributed normally with mean 10 mm and standard deviation 0.1 mm. What is the probability that the cylinder will fit in the hole?

8 Assume that the heights of adult males in Argentina and Egypt are distributed normally with means of 172 cm and 170 cm respectively and that the standard deviation of both populations is 7 cm. A sample of 25 adult males from each country is chosen at random. For each country find the probability that the sample mean is greater than 175 cm.

9 Assuming the heights of adult females in Croatia C are found to be distributed normally with mean 170 cm and standard deviation 6 cm, find the minimum sample size n such that $P(168 < \bar{C} < 172)$ is at least 0.9.

Chapter summary

- A **discrete random variable** T takes natural number values and varies randomly. A discrete random variable can be represented as a frequency histogram, a table of values or a **discrete probability distribution function.**
- The discrete probability distribution function (often abbreviated to pdf) $f(t) = P(T = t)$ represents the probability of the random variable taking a particular value in the domain.
- The values of the probability distribution function must be **at least zero** and their sum must be **exactly equal to 1**.
- The probability distribution function can be found by generalizing a process or by modelling a data set.
- The **cumulative distribution function** $F(t) = P(T \leq t) = \sum_{n=a}^{t} f(n)$, where a is the minimum value of the domain of $f(t)$, quantifies the probability of the random variable taking a value less than or equal to a particular value of the domain.
- The **expected value** of a discrete random variable X is $E(X) = \mu = \sum_{x} xP(X = x)$
- The **binomial distribution** models the process of a sequence of n independent trials of an experiment in which there are exactly two outcomes, "success" and "failure" with constant probabilities $P(\text{success}) = p$, $P(\text{failure}) = 1 - p$. You can use your GDC to calculate the probability distribution function of X.
- These facts are summarized in words as "X is distributed binomially with parameters n and p" and in symbols as $X \sim B(n, p)$.
- n and p are the **parameters** of the binomial distribution of X.
- The expected value of $X \sim B(n, p)$ is $E(X) = np$, and the variance is $\text{Var}(X) = np(1 - p)$.
- The **Poisson distribution** models the number of occurrences of an event in a given interval. The interval may have dimensions of time or of space. The average rate of occurrences α in a time interval is given and is uniform throughout all the times being considered. Occurrences in a time interval are independent and occurrences cannot occur at the same time or position. You can use your GDC to calculate the probability distribution function of X.

- These facts are summarized in words as "X follows a Poisson distribution with parameter α" and in symbols as $X \sim \text{Po}(\alpha)$.
- The expected value of $X \sim \text{Po}(\alpha)$ is $E(X) = \alpha$, and the variance is $\text{Var}(X) = \alpha$.
- Whereas discrete probability distributions involve **counting** outcomes, **continuous** probability distributions involve **measuring** a random variable.
- A continuous random variable can be represented as a **continuous probability density function** or its graph.
- The probability that a continuous random variable takes a particular value in the domain is zero. The area under the curve is 1.
- The parameters of the distribution are μ and σ^2. You write "X follows a normal distribution with a mean μ and variance σ^2" or $X \sim N(\mu, \sigma^2)$.
- The graph of the normal probability density function is a symmetric bell shape with these properties:
  - The axis of symmetry is $x = \mu$
  - $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$
  - $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$
  - $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$
- A random variable X can be transformed to the random variable $Y = aX + b$. The expected value and the variance of $Y = aX + b$ can be found by use of the formulae $E(Y) = aE(X) + b$ and $\text{Var}(Y) = a^2\text{Var}(X)$.
- Two independent random variables X and Y may be combined to make a new random variable $Z = aX + bY$ where $a, b \in \mathbb{R}$. The expected value and the variance of $Z = aX + bY$ can be found by use of the formulae $E(aX \pm bY) = aE(X) \pm bE(Y)$ and $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$.
- This can be generalized for linear combinations of n independent random variables X_i, $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$:
- $E(a_1X_1 \pm a_2X_2 \pm \ldots \pm a_nX_n) = a_1E(X_1) \pm a_2E(X_2) \pm \ldots \pm a_nE(X_n)$ and $\text{Var}(a_1X_1 \pm a_2X_2 \pm \ldots \pm a_nX_n) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + \ldots + a_n^2\text{Var}(X_n)$
- Given two independently distributed Poisson random variables $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\alpha)$, $T = X + Y$ follows a Poisson distribution with parameter $\lambda + \alpha$.
- Given two independently distributed normal random variables $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, $T = aX + bY$ follows a normal distribution $T \sim N(a\mu_1 + b\mu_2, a^2\sigma^2{}_1 + b^2\sigma^2{}_2)$.
- An important application of this is the **distribution of sample means.**
- If $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for which each X_i, $i = 1, 2, \ldots, n$ are selected from a large normally distributed population without replacement so that $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- Further, if $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and each X_i, $i = 1, 2, \ldots, n$ come from **any** distribution, then the **central limit theorem** states that the distribution of $\bar{X}$ can be modelled by a normal distribution $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ provided that n is at least 30.

Developing inquiry skills

In Chancer's café in the opening scenario, how many examples of probability distributions can you find?

How many of them are discrete? How many are continuous?

How many of these distributions have you learned about in this chapter?

How many of the situations involve combinations of random variables?

Do you need any more information to help you be sure of your choice of model?

How can you name and define precisely the distributions in the Chancer's café scenario that you have not yet learned?

Chapter review

Click here for a mixed review exercise

1 The discrete random variable D is distributed as shown in this table:

d	0	1	2	3	4
$\mathbf{P}(D = d)$	0.3	$p+q$	0.15	$p-q$	$p+2q$

a Given that $E(D) = 1.4$, find p and q.

b Find $P(D = 3 \mid D > 1)$.

2 The sides of a fair cubical dice are numbered 2, 3, 3, 4, 4 and 5. The dice is thrown twice and the outcomes are added to give a total T.

a Represent the probability distribution of T in a table.

b A game is designed so that if T is prime, a prize of $\$T$ is won. If T is a square number, the player must pay $\$x$. If T is any other number, no prize or payment is made. Find the value of $\$x$ so that the game is fair.

3 **a** Show that there are 16 factors of 120.

b A 120-sided dice is thrown six times. Find the probability that:

i exactly three factors of 120 are thrown

ii at least three factors of 120 are thrown

iii at most three factors of 120 are thrown

iv a factor of 120 is thrown on exactly three consecutive throws.

4 The number of accidents on a stretch of motorway follows a Poisson distribution with mean 1.21 accidents per day.

a Find the probability that on a randomly chosen day there are three or more accidents.

b Find the probability that on two consecutive days there is a total of exactly three accidents.

c Find the probability that in one week there are more than four days on which there are no accidents.

5 Assume that the weights of 18-year-old males in a population follow a normal distribution with mean 65 kg and standard deviation 11 kg.

a Find the probability that a randomly chosen male weighs more than 70 kg.

b Find the interquartile range of the weights in the population.

c Find the weight exceeded by 7.3% of the population.

d Eight boys are chosen at random from the population. Find the probability that at most three of them weigh at least 70 kg.

e In a sample of 1000 taken from the population, estimate how many would weigh below 60 kg.

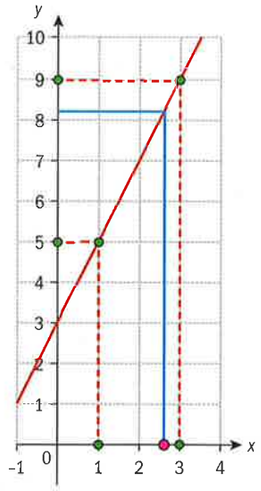
6 A soft drinks manufacturer produces an energy drink in two sizes, R (regular) and F (family). A survey shows that the number of calories in each size follows a normal distribution with parameters as shown in this table:

Size	Mean (calories)	Standard deviation (calories)
R	160	5
F	430	9

a One bottle of each size is selected at random. Find the probability that the family bottle contains less than three times the calories of the regular bottle.

b One family bottle and three regular bottles are selected at random. Find the probability that the family bottle contains more that the total number of calories in the three regular bottles.

7 A computer game design is based on the straight line l with equation $y = 2x + 3$. The red point shown on the x-axis moves at varying speeds along the x-axis. The gamer must choose a time to "kick" a "ball" vertically at l. As the ball reaches l, it changes direction to move horizontally towards the y-axis.



The score is determined by where the ball touches the y-axis as follows:

Interval	Points scored
$8 < y \leq 9$	10
$6 < y \leq 8$	5
$5 < y \leq 6$	1

Let F represent the x value chosen by the gamer as the position she kicks the ball. If $F \sim N(2.3, 0.3)$, find the expected number of points she scores in 127 games.

8 **a** A and B are two independently distributed Poisson random variables such that $E(A) = 3.1$ and $Var(B) = 4.7$. For each of the following random variables, state if the random variable follows the Poisson distribution, giving a justification for those which do not.

$C = A - B$, $D = 3A + 9$, $E = A + B$, $F = 2.9B$

b An online retailer sells a special type of battery at a rate of seven per week. The number of batteries sold follows a Poisson distribution. Find the smallest number of batteries that the retailer should have in stock at the start of the week in order to be at least 99% certain that they can meet the demand for these batteries.

c The random variable U is modelled by a Poisson distribution. Given that $P(U > 2) = 0.401$ find the variance of U.

9 **a** X is a discrete random variable with expectation 3.07 and standard deviation 0.8. Y and Z are independent of each other and independent of X. Given that $Y \sim B(5, 0.27)$, $Z \sim Po(3.81)$ and $T = X + Y + 2Z$, find $E(T)$ and $Var(T)$.

b A sample of size 35 is from a population modelled by the distribution of T. Find $P(11 \leq \overline{T} < 13)$.

c Find the minimum sample size n such that $P(11 \leq \overline{T} < 13)$ is at least 0.95.

Exam-style questions

10 **P1:** In a game, the discrete random variable X represents the number of counters a player wins. The probability distribution for X is given in the table below.

$X=x$	−3	−2	−1	0	1	2	3
$P(X=x)$	$\frac{1}{20}$	$\frac{2}{20}$	a	b	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

The mean of X is zero.

Find the values of **i** a **ii** b. (4 marks)

11 **P2:** Sue has an electrical fault in her house. The random variable X represents the number of times that the power goes off in a day. It can be assumed that X satisfies the Poisson distribution with mean of 5.

a Find the probability that Sue experiences exactly five power cuts in a day. (2 marks)

b Find the probability that Sue experiences less than five power cuts in a day. (2 marks)

c Find the probability that Sue experiences more than 33 power cuts in a week. (3 marks)

Mavis also has an electrical fault in her house. Let Y represents the number of times in a day that the power goes off in her house. Y satisfies a Poisson distribution with mean of 4.

It can be assumed that X and Y are independent, since Sue and Mavis live a large distance apart.

d Find the probability that the combined number of power cuts that both Sue and Mavis experience in a day is 8 or less. (3 marks)

A new random variable is defined by $Z = 3X + 2Y$.

e For the random variable Z, find **i** the mean **ii** the variance. (3 marks)

f State (with a reason) whether Z could satisfy a Poisson distribution. (2 marks)

12 **P1:** On a particular piece of road, the probability that a car is speeding is $\frac{1}{4}$.

The speeds of the cars are independent of each other.

A police officer sets up a speed-trap on this road.

a If the police officer records the speeds of 10 cars, find the probability that exactly three cars are speeding. (3 marks)

b If he records the speeds of 100 cars, find the probability that no more than 27 cars are speeding. (2 marks)

If he catches at least 50 cars for speeding, the policeman gets promoted.

c Assuming he catches every car that is speeding, find the smallest number of cars whose speed he has to record, for him to be at least 75% certain of being promoted. (3 marks)

13 **P1:** The mass, M, of Olympic marathon runners is normally distributed with mean of 60 kg and standard deviation of 3 kg. Let T be the total mass of the three runners who gain the gold, silver and bronze medals. It can be assumed that their masses are independent.

a Find the probability that $T > 175$. (4 marks)

The mass, H, of Olympic shot putters is normally distributed with mean of 160 kg and standard deviation of 5 kg.

b Find the probability that the mass of the gold medallist in the shot put is smaller than 2.5 times the mass of the gold medallist in the marathon. (5 marks)

14 **P1:** The mass of hens' eggs is normally distributed with mean of 50 grams and standard deviation of 4 grams. An egg is classified as large if it weighs more than 55 grams.

a Find the percentage of eggs that will be classified as large. (2 marks)

Eggs are put in boxes of six eggs.

b Find the probability that a random box of eggs has at least one large egg. (3 marks)

15 **P1:** The random variable, X, satisfies the Poisson distribution with mean of μ and $P(X=9) = \frac{1}{2}P(X=7)$.

a Find the value of μ. (3 marks)

16 **P2:** When Phil answers a multiple-choice question with four possible answers, he guesses the correct answer at random. Hence, the probability that he obtains the correct answer is $\frac{1}{4}$.

For a single multiple-choice question, let the discrete random variable X equal 1 if he has the question correct, and 0 if he has it wrong. Hence X satisfies the $B\left(1, \frac{1}{4}\right)$ distribution.

a For the random variable X, write down **i** the mean **ii** the variance. (2 marks)

A random sample of 100 values of X are taken (ie Phil takes an exam consisting of 100 multiple-choice questions). Let the sample mean $\frac{\sum_1^{100} X_i}{n}$ be denoted by $\bar{X}$. The Central Limit Theorem states that the distribution of $\bar{X}$ can be approximated by a normal distribution.

b Write down the **i** mean **ii** variance of $\bar{X}$, when approximated by a normal distribution. (3 marks)

c Use this normal approximation to find and estimate for $P(\bar{X} > 0.305)$. (2 marks)

d Hence, write down $P\left(\sum_1^{100} X_i > 30.5\right)$ when using this normal approximation. (2 marks)

e **i** State the true distribution that $T = \sum_1^{100} X_i$ satisfies and **ii** find the exact value of $P(T > 30.5)$, which can be construed as the probability of Phil passing the exam, if the pass mark was 31. (4 marks)

17 **P2:** A discrete random variable, X, has a probability distribution function given by

x	1	2	3
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

a Find $P(X \leq 2)$. (1 mark)

b Giving your answers as fractions, calculate:

i $E(X)$ (3 marks)

The variance of X is $\frac{5}{9}$. Let the random variable $Y = 4X$.

c Find **i** $E(Y)$ **ii** $\text{Var}(Y)$. (3 marks)

Let the random variable $T = X_1 + X_2 + X_3$, be the total of three independent values of X.

d Find **i** $E(T)$ **ii** $\text{Var}(T)$. (3 marks)

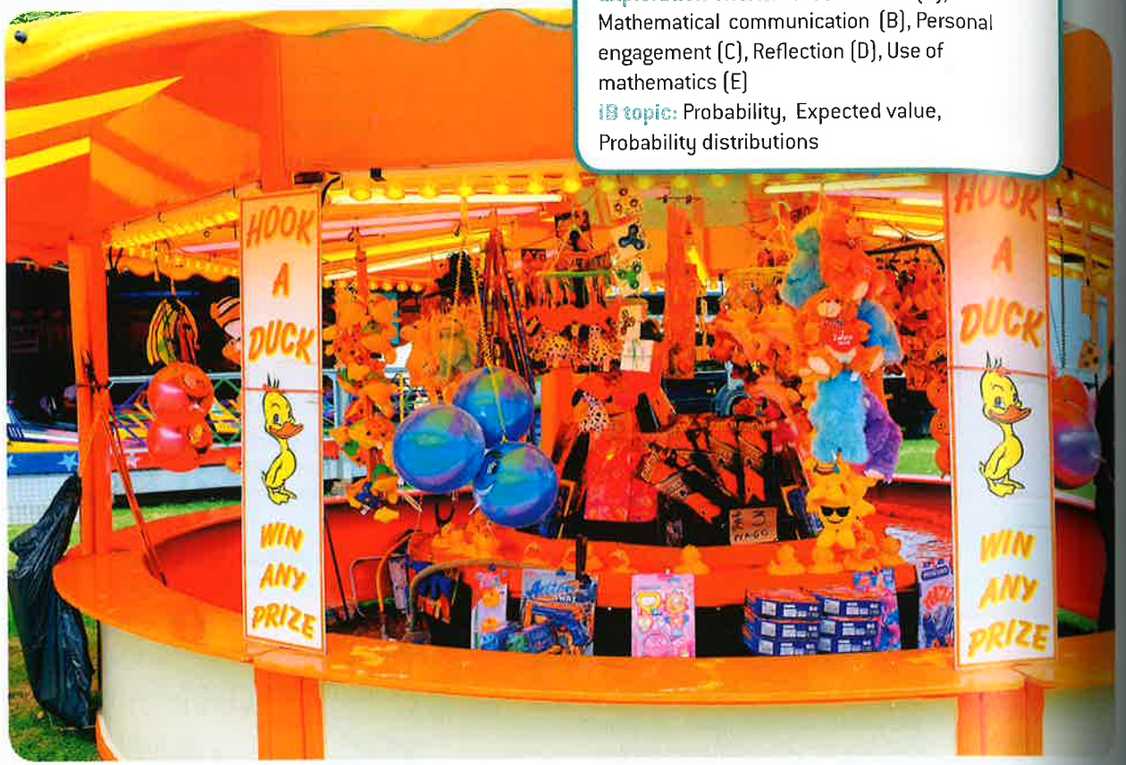
e Let the random variable $R = \frac{1}{X}$.

i Calculate $E(R)$ giving the answer as a fraction.

ii Hence, determine whether or not the statement $E\left(\frac{1}{X}\right) = \frac{1}{E(X)}$ is true. You must justify your answer. (3 marks)

Fair game!

Approaches to learning: Collaboration, Communication, Self-management
Exploration criteria: Presentation (A), Mathematical communication (B), Personal engagement (C), Reflection (D), Use of mathematics (E)
IB topic: Probability, Expected value, Probability distributions



In this chapter you have been looking at probability distributions and understanding **expected value** and the meaning of a **fair game**.

The task

In your pairs or groups of three, your task is to design your own fully functioning game for other students to play.

In the task you will need to think about:

- What equipment can you use to create probabilities?
- How can you make your game exciting/appealing?
- How can you make sure you make a profit from your game?

Your game must be unique, and **not** copied from an existing game!

Part 1: Design and understand the probabilities involved in your game

In your groups decide on the game you will produce.

Brainstorm some ideas first.

You may need to trial the game in your group to check it works before you finalize everything.

Provide a brief overview of your game and make sure that you are able to explain the probabilities involved for your teacher to check through.

Think about:

- What equipment do players need to play the game?
- How much will the game cost to play?
- What will the prizes be?

Create a set of clear, step-by-step instructions to explain your game to players.

A player should be able to read the instructions and then be able to start playing your game immediately.

Part 2: Set up the game in a "class fair"

You will need to provide any required equipment for your game.

Your game needs to be attractive and eye-catching.

Play the games in class.

Part 3: Reflect on the success or otherwise of your game

Either in a written report, interview or in a video reflect on the success or otherwise of your game.

Think about:

- Explain how your game worked. What went well?
- What did you do as the game manager and what did your participants do as game players?
- Provide accurate analysis of the mathematics behind your game.
- Was your game popular? Why? Why not?
- What was your expected profit per game? Is this what you actually received? Why? Why not? Did the experimental probability match the theoretical probability?
- If the game was not fair, how could you change the game to make it fair?
- If anything didn't go as planned, what went wrong?
- What did you learn from other games that would improve your game? What would you improve or change if you were to do this task again?

Extension

As you reflect on this task and on this chapter as a whole, consider one, some or all of these questions regarding mathematics, probability and gambling.

You can approach each of these questions from a combination of mathematics and TOK concepts.

- What does "the house always wins" mean? Is it **fair** that casinos should make a profit? Is there such a thing as "ethical gambling"?
- Could and/or should mathematics and mathematicians help increase incomes in gambling? What is the ethical responsibility of a mathematician?
- What is luck? How would a mathematician explain luck?