

In this chapter you will learn:

- what a random variable is and how to list its possible values
- how to predict the average value of a random variable
- how to calculate the probability of getting a certain number of successes over a fixed number of trials – the binomial distribution
- about a distribution that models many naturally occurring random variables – the normal distribution.

18 Probability distributions

Introductory problem

A casino offers a game in which a coin is tossed repeatedly. If the first head occurs on the first throw you get £2, if the first head occurs on the second throw you get £4, if the first head is on the third throw you get £8, and so on, with the prize doubling each time. How much should the casino charge for this game if they want to make a profit?

In statistics we find the mean, standard deviation and other measures of centre or spread from data we have already collected. In real life, however, it is often useful to be able to predict these quantities in advance. Even though, in a random situation, it is impossible to predict the outcome of a single event, such as one roll of a die, it turns out that if you look at enough events, the average can be predicted quite precisely.

The idea of being able to predict averages but not individual events is central to many areas of knowledge. For example, economists cannot predict what an individual will do when interest rates increase, but they can predict the average effect on the economy. A physicist knows that in a waterfall, any particular water molecule may actually be moving upwards, but on average the flow is definitely going to be downwards.



18A Random variables

A **random variable** is a quantity whose value depends on chance – for example, the outcome when a die is rolled. If the probabilities associated with each possible value are known, useful mathematical calculations can be made. A random variable is conventionally represented by a capital letter, and the values that the random variable can take are represented by the corresponding lower-case letter. For instance, if we let the random variable X be the outcome when a die is rolled, in one particular experiment you may find that $x = 2$.

Recall that the sample space of a random event is the list of all possible outcomes of the event. The sample space of a random variable, together with the probabilities associated with all the values in the list, is called the **probability distribution** of the variable; this information is best displayed in a table.

Worked example 18.1

Make a table to show the probability distribution of the outcome of rolling a fair six-sided die.

Make a list of all the values that the random variable can take. Then write down the probability of each value occurring.

Let X be the outcome of rolling the die.

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The probabilities in a probability distribution cannot be just any numbers.

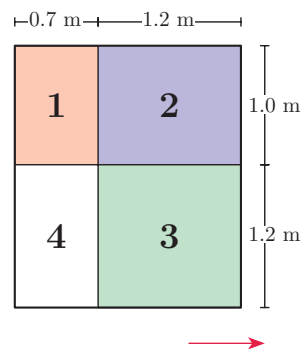
KEY POINT 18.1

The total of all the probabilities in a probability distribution must always equal 1.

This fact can be useful if we do not have complete information about the probabilities.

Worked example 18.2

In a game at a fair, a ball is thrown at a rectangular target. The dimensions of the target (in metres) are as shown. The probability of hitting each region is proportional to its area. The prize for hitting a region is a number of chocolates equal to the number shown in that region. Find the probability distribution of the number of chocolates won.



continued . . .

Show the possible values in a table.

The probability of each value is proportional to the area of the region; this can be expressed as probability = $k \times \text{area}$.

Use the fact that the probabilities add up to 1.

We can now calculate all the probabilities.

Since we are not asked for exact values, we can round them to 3SF.

Let X = number of chocolates won

x	1	2	3	4
$P(X = x)$	$0.7k$	$1.2k$	$1.44k$	$0.84k$

$$0.7k + 1.2k + 1.44k + 0.84k = 1$$



$$\therefore k = 0.239$$

x	1	2	3	4
$P(X = x)$	0.167	0.287	0.344	0.201

An obvious question to ask about a random variable is what value is it *most likely* to have. This value, whose associated probability is the highest, is called the **mode**. In the above example, the random variable X has mode 3, which means that the most likely number of chocolates you will win is three. A random variable may not have a mode – for example, the outcomes of a fair die are all equally likely – or it may have more than one mode. In particular, if the largest probability corresponds to two of the outcomes, the random variable is said to be **bimodal**.

Another question we might ask is: if we were to play the above game many times, *on average* how many chocolates would we expect to win? The answer is not necessarily the same as the most likely outcome. We will see how to answer this question in the next section.

Exercise 18A

 In this exercise you will need to use tools from chapter 17, in particular tree diagrams. For question 2(c) you may  want to look back at chapter 6 on geometric sequences.

- For each of the following, make a table to represent the probability distribution of the random variable described.

- A fair coin is thrown four times. The random variable W is the number of tails obtained.
- Two fair dice are thrown. The random variable D is the difference between the larger and the smaller score, or zero if they are the same.
- A fair die is thrown once. The random variable X is calculated as half the result if the die shows an even number, or one higher than the result if the die shows an odd number.
- A bag contains six red and three green counters. Two counters are drawn at random from the bag without replacement. The random variable G is the number of green counters remaining in the bag.
- Karl picks a card at random from a standard pack of 52 cards. If he draws a diamond, he stops; otherwise, he replaces the card and continues to draw cards at random, with replacement, until he has either drawn a diamond or drawn a total of four cards. The random variable C is the total number of cards drawn.
- Two fair four-sided spinners, each labelled 1, 2, 3 and 4, are spun. The random variable X is the product of the two values shown.

2. Find the missing value k for each probability distribution.

(a) (i)

x	3	7	9	11
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	k

(ii)

x	5	6	7	10
$P(X = x)$	0.2	0.3	k	0.5

(b) (i) $P(Y = y) = ky$ for $x = 1, 2, 3, 4$

(ii) $P(X = x) = \frac{k}{x}$ for $x = 1, 2, 3, 4$

(c) (i) $P(X = x) = k(0.1)^x$ for $x \in \mathbb{N}$

(ii) $P(R = r) = k(0.9)^r$ for $r \in \mathbb{N}$



\mathbb{N} is the set of natural numbers, $\{0, 1, 2, 3, \dots\}$.



See Prior Learning section G on the CD-ROM for a review of number sets.



3. In a game, a player rolls a biased four-sided die. The probability of each possible score is shown below.

Score	1	2	3	4
Probability	$\frac{1}{3}$	$\frac{1}{4}$	k	$\frac{1}{5}$

Find the probability that the total score is 4 after two rolls.

[5 marks]

18B Expectation of a discrete random variable

EXAM HINT

The expectation gives the *theoretical* expected mean. In any particular series of trials, the actual mean value may be different. Also, do not confuse the expected value with the most likely value (the mode).

If you roll a fair die many times, you would expect the average score to be around 3.5, the mean of the numbers 1 to 6. However, if the sides of the die were labelled 1, 2, 3, 4, 6 and 6, you would expect the average to be higher, because the probability of getting a 6 is higher than for each of the other possible outcomes. The **expectation** of a random variable is the average value you would get if you were to repeatedly measure the variable an infinite number of times.

KEY POINT 18.2

The **expectation** (or mean value) of a random variable X is written $E(X)$ and calculated as

$$E(X) = \mu = \sum_x x P(X = x)$$

Worked example 18.3

The random variable X has probability distribution as shown in the table. Calculate $E(X)$.

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{10}$

The formula tells us to calculate the product $xP(X = x)$ for all possible values of x and add up the results.

$$\begin{aligned}
 E(X) &= 1 \times \frac{1}{10} + 2 \times \frac{1}{4} + 3 \times \frac{1}{10} + 4 \times \frac{1}{4} \\
 &\quad + 5 \times \frac{1}{5} + 6 \times \frac{1}{10} \\
 &= \frac{7}{2}
 \end{aligned}$$

Just as the mean of a set of integers need not be an integer itself, so the expectation of a random variable need not be one of the values that the variable can take.

Exercise 18B

1. Calculate the expectation of each of the following random variables.

(a) (i)

x	1	2	3	4
$P(X = x)$	0.4	0.3	0.2	0.1

(ii)

w	8	9	10	11
$P(W = w)$	0.4	0.3	0.2	0.1

(b) (i) $P(X = x) = \frac{x^2}{14}$ for $x = 1, 2, 3$

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

2. A random variable X has probability distribution $P(X = x) = k(x + 1)$ for $x = 2, 3, 4, 5, 6$.


(a) Show that $k = 0.04$.

(b) Find $E(X)$. [5 marks]

3. The random variable V has probability distribution as shown in the table and $E(V) = 6.3$. Find the value of k .

v	1	2	5	8	k
$P(V = v)$	0.2	0.3	0.1	0.1	0.3

[4 marks]

-  4. A random variable X has its probability distribution given by $P(X = x) = k(x + 3)$, where x is 0, 1, 2 or 3.

(a) Show that $k = \frac{1}{18}$.

(b) Find the exact value of $E(X)$. [6 marks]

5. The probability distribution of a random variable X is given by $P(X = x) = kx(4 - x)$ for $x = 1, 2, 3$.

(a) Find the value of k .

(b) Find $E(X)$. [6 marks]

6. A fair six-sided die with sides numbered 1, 1, 2, 2, 2, 5 is thrown. Find the expected mean of the score. [6 marks]

7. The table below shows the probability distribution of a random variable X .

x	0	1	2	3
$P(X = x)$	0.1	p	q	0.2

Given that $E(X) = 1.5$, find the values of p and q . [6 marks]

8. A biased die with four faces is used in a game. A player pays 5 counters to roll the die. The table below shows the possible scores on the die, the probability of each score, and the number of counters the player receives in return for each score.

Score	1	2	3	4
Probability	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{20}$
Number of counters player receives	4	5	15	n

Find the value of n so that the player gets an expected profit of 3.25 counters per roll. [5 marks]

9. In a game, a player pays an entrance fee of $\$n$. He then selects one number from 1, 2, 3 or 4 and rolls three 4-sided dice.

If his chosen number appears on all three dice, he wins four times his entrance fee.

If his number appears on exactly two of the dice, he wins three times the entrance fee.

If his number appears on exactly one die, he wins $\$1$.

If his number does not appear on any of the dice, he wins nothing.

- (a) Copy and complete the following probability table.

Player's profit (\$)	$-n$		$2n$	$3n$
Probability		$\frac{27}{64}$		

- (b) The game organiser wants to make a profit over many plays of the game. Given that he must charge a whole number of cents, what is the minimum entrance fee the organiser should charge? [10 marks]

18C The binomial distribution

Some probability distributions come up so often that they have been given names and formal notation. One of the most important of these is the **binomial distribution**.

A binomial distribution arises when an experiment (or 'trial') with two possible outcomes is repeated a set number of times. The word 'binomial' refers to the two possible outcomes; conventionally, one of them is called a 'success' and the other a 'failure'. If the probability of 'success' remains constant and the trials are conducted independently of each other, then the total number of successes can be modelled using the binomial distribution.

KEY POINT 18.3

The binomial distribution models the number of successful outcomes in a fixed number of trials, provided the following conditions are satisfied:

- Each trial has two possible outcomes.
- The trials are independent of each other.
- The probability of success is the same in every trial.

If n is the number of trials, p is the probability of a success, and the random variable X is the total number of successes, then X follows a binomial distribution with n trials and probability of success p , written $X \sim B(n, p)$.

EXAM HINT

You need to know the conditions under which the binomial distribution can be used and be able to interpret them in context.

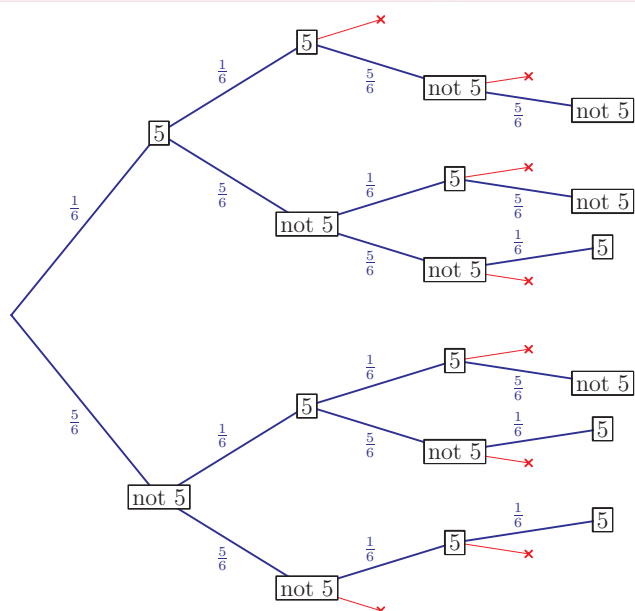
So what is this distribution? Let us consider a specific example: suppose a die is rolled four times; what is the probability of getting exactly two fives?

There are four trials, so $n = 4$. In this context, it makes sense to label getting a 5 as 'success', so we have $p = \frac{1}{6}$. The probability of a 'failure' (getting any number other than 5) is therefore $\frac{5}{6}$. One

way of getting two fives is if on the first two rolls we get a five and on the next two rolls we get something else. The probability

of this happening is $\frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$. But this is not

the only way in which two fives can occur. The two fives may be on the first and third or second and fourth rolls. In fact, we have to consider all the ways in which we can pick two trials out of the four for the 5 to turn up in. This can be done by drawing a tree diagram. We only need to include the branches along which exactly two fives occur.



Each of the paths giving 'two fives' has the same probability,

$\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$. The total number of paths giving the outcome

'two fives' is 6. So if X is the random variable 'number of fives thrown when four dice are rolled', then we can say that

$$P(X=2) = 6 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2.$$

Generalising this reasoning, consider n trials in which the probability of a success is p , and suppose we are interested in the probability of obtaining r successes. If we imagine representing the n trials on a tree diagram, each relevant path (along which exactly r successes occur) will have probability $p^r (1-p)^{n-r}$, because r of the outcomes are successes (each occurring with probability p) and the remaining $n-r$ outcomes are failures (each occurring with probability $1-p$). It turns out that the number of paths that give r successes is given by the binomial coefficient $\binom{n}{r}$. This leads to the following general formula for the probabilities of the binomial distribution.

KEY POINT 18.4

If $X \sim B(n, p)$, then

$$P(X=r) = \binom{n}{r} p^r (1-p)^{n-r} \text{ for } r=0,1,2,\dots,n$$

The useful thing about identifying a binomial distribution is that you can then apply standard results, such as the formula

We met binomial coefficients in chapter 7 when studying the binomial expansion. You can find binomial coefficients with your calculator or by using the formula given in the Formula booklet.

above, without having to go through the earlier argument with tree diagrams every time.

Worked example 18.4

Rohir has a 30% chance of correctly answering a multiple-choice question. He takes a test in which there are ten multiple-choice questions.

- What is the probability that Rohir gets exactly four of the questions correct? Give your answer to five significant figures.
- Suggest which requirements for a binomial distribution might not be satisfied in this situation.

Define the random variable (if not already given in the question).

Identify the probability distribution.

Write down the formula for the probability required, and calculate the answer.

Consider the conditions for the distribution to apply.

Are there any requirements that are not met in this context? There are two outcomes, and trials are independent (answering one question does not make it easier or harder to answer another).

(a) Let X be the number of correct answers out of the ten.

$$X \sim B(10, 0.3)$$

$$n = 10, p = 0.3, r = 4$$


$$P(X = 4) = \binom{10}{4} (0.3)^4 (0.7)^6 \\ = 0.20012 \text{ (5SF)}$$

(b) Binomial distribution requires:
two outcomes at each trial
trials independent of each other
constant probability of success in each trial.

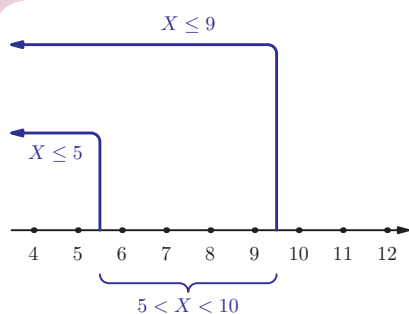
The questions may not all be of the same difficulty, so there may not be a constant probability of success.

Most calculators can find binomial probabilities if you specify the values of n , p and r . Sometimes you may want to find the probability of a *range* of numbers of successes (r values); you could in principle work out the probabilities for the different r values and then add them all up, but this can be very time-consuming. Fortunately, calculators usually have a function that gives the probability of getting up to (and including) a certain number of successes – this is called a **cumulative probability**.

EXAM HINT

 See Calculator skills sheet 12 on the CD-ROM for how to find binomial probabilities and cumulative probabilities on your calculator.





For example, if you are asked to find the probability that the number of successes is greater than 10, you can use your calculator to find the cumulative probability $P(X \leq 10)$; then the probability you want, $P(X > 10)$, is just $1 - P(X \leq 10)$. If you need the probability of getting more than 5 but fewer than 10 successes, this is $P(5 < X < 10) = P(X \leq 9) - P(X \leq 5)$, as shown on the number line.

Worked example 18.5

Anna shoots at a target 15 times. The probability that she hits the target on any shot is 0.6. Find the probability that she hits the target more than 5 but at most 10 times.

Define the random variable.

Let X be the number of times Anna hits the target.

State the probability distribution.

$X \sim B(15, 0.6)$

Write down the probability to be found, and express it in terms of probabilities that can be found on the GDC.

From GDC:
 $P(5 < X \leq 10) = P(X \leq 10) - P(X \leq 5)$
 $= 0.7827 - 0.0338$
 $= 0.749 \quad (3 \text{ SF})$

EXAM HINT

Even when you are using a calculator to find probabilities, you should still write your solution in correct mathematical notation (not calculator notation). You must state what distribution you used and which probabilities you found using the GDC. Remember to give the answer to 3 significant figures.

Now that we can calculate probabilities for different numbers of successes in a binomial distribution, we can ask what is the expected mean number of successes. As we noted in section 18B, this is not the same as the most likely number of successes.

Suppose that all students in a large school take a multiple-choice test with 12 questions, each with 5 possible answers (only one of which is correct). If the students all decide to guess answers randomly, what is the expected average number of correct answers? The scores for each individual student will vary, but it seems plausible that the average will be around

$$12 \times \frac{1}{5} = 2.4.$$

Using the formula for the expectation, it can be proved that this is indeed the case. It is also possible to show that the variance of the scores is $12 \times \frac{1}{5} \times \frac{4}{5} = 1.92$.



KEY POINT 18.5

If $X \sim B(n, p)$, then its expectation (mean) and variance are given by


$$E(X) = np$$

$$\text{Var}(X) = np(1 - p)$$

In answering questions, we can use these formulas without explanation, but we must make it clear what distribution is being used.

 The variance was introduced in section 16B; it is the square of the standard deviation, a measure of the spread of values of a random variable. 

Worked example 18.6

 A bag contains a large number of balls, two-thirds of which are green. Ten balls are selected one by one, with each ball being replaced before the next one is selected. Find the mean and standard deviation of the number of green balls selected.

Define the random variable.

Let G be the number of green balls selected.

State the distribution. There are 10 trials, and in each trial the probability of selecting a green ball is always $\frac{2}{3}$, so G follows a binomial distribution.

$$G \sim B\left(10, \frac{2}{3}\right)$$

Use the formula for expectation.

$$E(X) = 10 \times \frac{2}{3} = \frac{20}{3}$$

Use the formula for variance. Then take the square root to get the standard deviation.

$$\text{Var}(X) = 10 \times \frac{2}{3} \times \frac{1}{3} = \frac{20}{9}$$

$$\therefore \sigma = \sqrt{\frac{20}{9}} = \frac{2\sqrt{5}}{3}$$

Although most of the time you will be using your calculator to find binomial probabilities, in some situations you may need to use the formula in Key point 18.4, as in the next example.

Worked example 18.7

A random variable X has distribution $B(15, p)$. Given that $P(X = 9) = 0.105$, find the possible values of p .

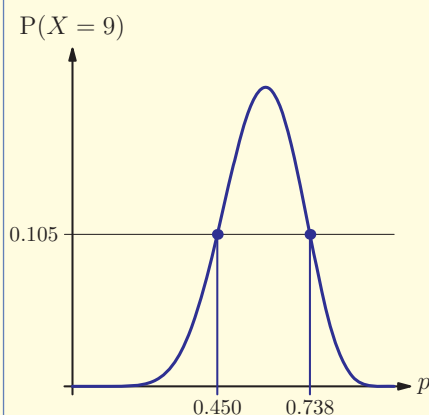
Use the formula in Key point 18.4 to write $P(X = 9)$ in terms of p . Set this equal to 0.105 to get an equation for p .

$$P(X = 9) = \binom{15}{9} p^9 (1-p)^6 = 0.105$$



continued ...

Use the GDC to solve this equation.
Remember that only values between
0 and 1 make sense for p .



From GDC, $p = 0.450$ or 0.738 (3 SF)

Exercise 18C



1. The random variable X has a binomial distribution with $n = 8$ and $p = 0.2$. Calculate the following probabilities.

- | | |
|---------------------------|------------------------|
| (a) (i) $P(X=3)$ | (ii) $P(X=4)$ |
| (b) (i) $P(X \leq 3)$ | (ii) $P(X \leq 2)$ |
| (c) (i) $P(X > 3)$ | (ii) $P(X > 4)$ |
| (d) (i) $P(X < 5)$ | (ii) $P(X < 3)$ |
| (e) (i) $P(X \geq 3)$ | (ii) $P(X \geq 1)$ |
| (f) (i) $P(3 < X \leq 6)$ | (ii) $P(1 \leq X < 4)$ |



2. Given that $Y \sim B\left(5, \frac{1}{2}\right)$, find the exact value of

- | | |
|-----------------------|--------------------|
| (a) (i) $P(Y=1)$ | (ii) $P(Y=0)$ |
| (b) (i) $P(Y \geq 1)$ | (ii) $P(Y \leq 1)$ |
| (c) (i) $P(Y > 4)$ | (ii) $P(Y \leq 3)$ |

3. Find the mean and variance of the following random variables.

- | | |
|--|---|
| (a) (i) $Y \sim B\left(100, \frac{1}{10}\right)$ | (ii) $X \sim B\left(16, \frac{1}{2}\right)$ |
| (b) (i) $X \sim B(15, 0.3)$ | (ii) $Y \sim B(20, 0.35)$ |
| (c) (i) $Z \sim B\left(n-1, \frac{1}{n}\right)$ | (ii) $X \sim B\left(n, \frac{2}{n}\right)$ |

4. (a) Jake beats Marco at chess in 70% of their games. Assuming that this probability is constant and that the results of games are independent of each other, what is the probability that Jake will beat Marco in at least 16 of their next 20 games?
- (b) On a television channel, the news is shown at the same time each day. The probability that Salia watches the news on a given day is 0.35. Calculate the probability that on 5 consecutive days she watches the news on exactly 3 days.
- (c) Sandy is playing a computer game and needs to accomplish a difficult task at least three times in five attempts in order to pass the level. There is a 1 in 2 chance that he accomplishes the task each time he tries, unaffected by how he has done before. What is the probability that he will pass to the next level?

5. 15% of students at a large school travel by bus. A random sample of 20 students is taken.

- (a) Explain why the number of students in the sample who travel by bus follows only approximately a binomial distribution.
- (b) Use the binomial distribution to estimate the probability that exactly five of the students in the sample travel by bus.

[3 marks]

6. A coin is biased so that when it is tossed the probability of obtaining heads is $\frac{2}{3}$. The coin is tossed 4050 times. Let X be the number of heads obtained. Find the expected value of X .

[3 marks]

7. A biology test consists of eight multiple-choice questions. Each question has four answers, only one of which is correct. At least five correct answers are required to pass the test. Sheila has not studied for the test, so answers each question at random.

- (a) What is the probability that Sheila answers exactly five questions correctly?
- (b) What is the expected number of correct answers Sheila will give?
- (c) Find the variance of the number of correct answers Sheila gives.
- (d) What is the probability that Sheila manages to pass the test?

[7 marks]

8. Suppose that 0.8% of people in a country have a particular cold virus at any time. On a single day, a doctor sees 80 patients.

- (a) What is the probability that exactly two of them have the virus?
- (b) What is the probability that three or more of them have the virus?
- (c) State an assumption you have made in these calculations. [5 marks]

9. Given that $Y \sim B(12, p)$ and that the mean of Y is 4.8, find the value of p . [3 marks]

10. With a fair die, which is more likely: rolling 3 sixes in 4 throws or rolling a five or a six in 5 out of 6 throws? [6 marks]

11. A drawer contains 5 red socks and 5 blue socks. Two socks are removed without replacement.

- (a) Show that the probability of taking a red sock second depends on the outcome of the first sock taken, i.e. the events are dependent.
- (b) Show that the probability of taking a red sock second equals the probability of taking a red sock first, i.e. the probability is constant.



12. Over a one-month period, Ava and Sven play a total of n games of tennis. The probability that Ava wins any game is 0.4. The result of each game played is independent of any other game played. Let X denote the number of games won by Ava over the one-month period.

- (a) Find an expression for $P(X = 2)$ in terms of n .
- (b) If the probability that Ava wins two games is 0.121 correct to three decimal places, find the value of n . [5 marks]



13. A die is biased so that the probability of rolling a six is p . If the probability of rolling 2 sixes in 12 throws is 0.283 (to three significant figures), find the possible values of p correct to two decimal places. [5 marks]



14. X is a binomial random variable where the number of trials is 5 and the probability of success in each trial is p . Find the possible values of p if $P(X = 3) = 0.3087$. [5 marks]



Question 10 is the problem posed to Pierre de Fermat in 1654 by a professional gambler who could not understand why he was losing. It inspired Fermat (with the assistance of Pascal) to set up probability as a rigorous mathematical discipline.

18D The normal distribution

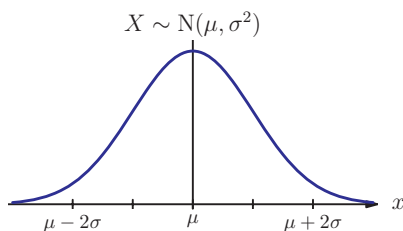
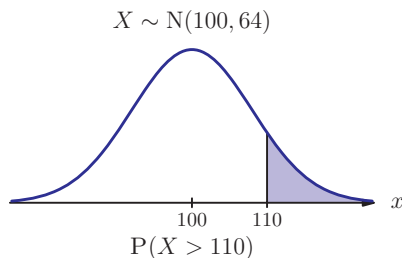
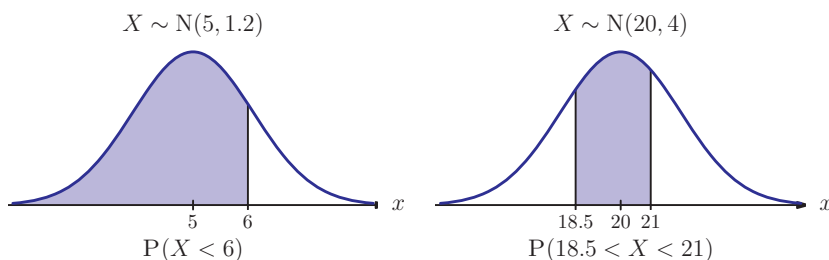
There are many situations where a random variable is most likely to be close to its average value, and values further away from the average become increasingly unlikely. The **normal distribution** is a model of such situations.

The normal distribution is a *continuous* distribution, used to model continuous data or a continuous random variable, where each value can be any real number within a certain interval – examples include measurements of height and time. For a continuous distribution, unlike for the *discrete* distributions we met in sections 18A–C, we cannot list all possible values of the variable in a table; we can only calculate the probability of the variable taking values in a specified range.

A continuous distribution is commonly represented by a curve, where the probability of the variable being between two specified values is equal to the *area* under the curve between those two values. The diagrams below show several examples of normal distributions. The values of the random variable X are plotted on the x -axis, and the shaded area represents the probability given beneath each graph.

EXAM HINT

Be careful with the notation: σ^2 is the variance, so $X \sim N(10, 9)$ has standard deviation $\sigma = 3$.



To describe a particular normal distribution, all that is needed is its mean and variance. If a random variable X follows a normal distribution with mean μ and variance σ^2 , we write $X \sim N(\mu, \sigma^2)$. The peak of the normal distribution curve is at the mean. The standard deviation σ determines the width (i.e. spread) of the curve; values of x that lie further than two standard deviations from the mean are very unlikely.



Historically, cumulative probabilities for the normal distribution, i.e. probabilities of the form $P(X \leq x)$, were recorded in tables. Many people still use such tables if they don't have access to a calculator. Since it is not feasible to make tables for every possible combination of μ and σ values, values of the random variable need to be converted into 'Z-scores', described below, before they can be looked up in a table.

You can find probabilities of normal distributions using your calculator. You need to enter the mean and standard deviation, and specify values x_1 and x_2 to obtain the probability $P(x_1 \leq X \leq x_2)$. You can also calculate probabilities of the form $P(X \geq x)$ or $P(X \leq x)$.

It is often helpful to sketch a graph to get a visual representation of the probability you are trying to find. Graphs can also provide a useful check of your answer, as they show you whether you should expect the probability to be smaller or greater than 0.5.

EXAM HINT

See Calculator skills sheet 13 on the CD-ROM for details of how to calculate normal probabilities.



Worked example 18.8

The average height of people in a town is 170 cm, and the standard deviation of the heights is 10 cm. What is the probability that a randomly selected resident is

- (a) shorter than 165 cm
- (b) between 180 cm and 190 cm tall
- (c) taller than 176 cm?

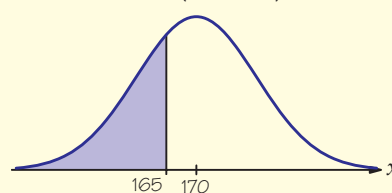
State the distribution.

Identify the probability to be found and use a calculator to find it.

Let X be the height of a town resident. Then
 $X \sim N(170, 100)$

(a) $P(X < 165)$

$X \sim N(170, 100)$

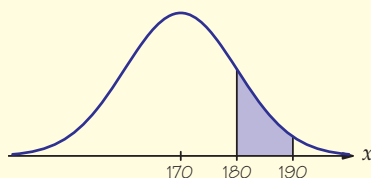


$P(X < 165) = 0.309$ (3 SF) from GDC

continued ...

(b) $P(180 < X < 190)$

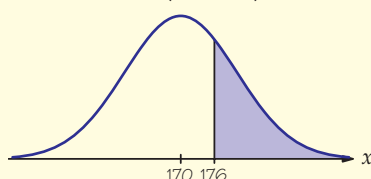
$X \sim N(170, 100)$



$P(180 < X < 190) = 0.136$ (3 SF) from GDC

(c) $P(X > 176)$

$X \sim N(170, 100)$



$P(X > 176) = 0.274$ (3 SF) from GDC

If a normally distributed random variable has mean 120, should a value of 150 be considered unusually large? The answer depends on how spread out the variable is, which is measured by its standard deviation. If the standard deviation of this variable is 30, then a value around 150 would be quite common; however, if the standard deviation were 5, then 150 would be very unusual.

The probability of a random variable being less than a given value is called **cumulative probability**. For a normally distributed variable, it turns out that this probability depends only on the number of standard deviations the value is from the mean. This distance in terms of number of standard deviations is called the **Z-score**.

KEY POINT 18.6

For $X \sim N(\mu, \sigma^2)$, the Z-score of the value x measures the number of standard deviations that x is away from the mean:

$$z = \frac{x - \mu}{\sigma}$$



Worked example 18.9

Suppose that $X \sim N(15, 6.25)$.

- How many standard deviations is $x = 16.1$ away from the mean?
- Find the value of X which is 1.2 standard deviations below the mean.

The number of standard deviations away from the mean is measured by the Z-score.

6.25 is the variance! We need to take the square root to find the standard deviation.

Values below the mean have a negative Z-score.

$$(a) \quad z = \frac{x - \mu}{\sigma}$$

$$\sigma = \sqrt{6.25} = 2.5$$

$$\therefore z = \frac{16.1 - 15}{2.5} = 0.44$$

So 16.1 is 0.44 standard deviations away from the mean.

$$(b) \quad z = -1.2$$

means that

$$-1.2 = \frac{x - 15}{2.5}$$

$$\Rightarrow x - 15 = -3$$

$$\Rightarrow x = 12$$

If we have a random variable $X \sim N(\mu, \sigma^2)$, we can create a new random variable Z which takes values equal to the Z-scores of the values of X . In other words, for each x there is a corresponding $z = \frac{x - \mu}{\sigma}$; this is called the *standardised value*.

It turns out that, whatever the original mean and standard deviation of X , this new random variable Z always has a normal distribution with mean 0 and variance 1, called the **standard normal distribution**: $Z \sim N(0, 1)$. This is an extremely important property of normal distributions, and it is especially useful in situations when the mean and standard deviation of X are not known (see section 18E).

Think about the transformations of graphs that we studied in chapter 5: the normal distribution curves for X and Z are related via a horizontal translation by μ units and a horizontal stretch with scale factor σ .

Before graphical calculators became available (which is not so long ago!), people used tables showing cumulative probabilities of the standard normal distribution. Because of their importance, these probabilities were given special notation: $\Phi(z) = P(Z \leq z)$.

Although you don't have to use this notation, you should understand what it means as you may still encounter it in some books.



KEY POINT 18.7

The (cumulative) probabilities of X and Z are related by

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

Worked example 18.10

Let $X \sim N(6, 0.5^2)$. Write the following in terms of probabilities of Z .

- (a) $P(X \leq 6.1)$ (b) $P(5 < X < 7)$ (c) $P(X > 6.5)$

Here $x = 6.1$, so we calculate the corresponding $z = \frac{x - \mu}{\sigma}$.

The relationship between probabilities of X and Z in Key point 18.7 is stated for cumulative probabilities, so convert to that form first.

(a) $\mu = 6, \sigma = 0.5$

$$\begin{aligned} P(x \leq 6.1) &= P\left(Z \leq \frac{6.1 - 6}{0.5}\right) \\ &= P(Z \leq 0.2) \end{aligned}$$

(b) $P(5 < X < 7) = P(X < 7) - P(X < 5)$

$$\begin{aligned} &= P\left(Z < \frac{7 - 6}{0.5}\right) - P\left(Z < \frac{5 - 6}{0.5}\right) \\ &= P(Z < 2) - P(Z < -2) \\ &= P(-2 < Z < 2) \end{aligned}$$

(c) $P(X > 6.5) = 1 - P(X \leq 6.5)$

$$\begin{aligned} &= 1 - P\left(Z \leq \frac{6.5 - 6}{0.5}\right) \\ &= 1 - P(Z \leq 1) \\ &= P(Z > 1) \end{aligned}$$

As you can see from parts (b) and (c) of the above example, you don't actually have to convert probabilities into the form $P(X \leq k)$ every time; you can simply replace the x values by the corresponding z values (Z-scores).

Exercise 18D

1. Find the following probabilities.

(a) $X \sim N(20, 100)$

(i) $P(X \leq 32)$ (ii) $P(X < 12)$

(b) $Y \sim N(4.8, 1.44)$

(i) $P(Y > 5.1)$ (ii) $P(Y \geq 3.4)$

(c) $R \sim N(17, 2)$

(i) $P(16 < R < 20)$ (ii) $P(17.4 < R < 18.2)$

(d) Q has a normal distribution with mean 12 and standard deviation 3.

(i) $P(Q > 9.4)$ (ii) $P(Q < 14)$

(e) F has a normal distribution with mean 100 and standard deviation 25.

(i) $P(|F - 100| < 15)$ (ii) $P(|F - 100| > 10)$

2. Find the Z-score corresponding to the given value of X .

(a) (i) $X \sim N(12, 2^2)$, $x = 13$

(ii) $X \sim N(38, 7^2)$, $x = 45$

(b) (i) $X \sim N(20, 9)$, $x = 15$

(ii) $X \sim N(162, 25)$, $x = 160$

3. Given that $X \sim N(16, 2.5^2)$, write the following in terms of probabilities of the standard normal variable.

(a) (i) $P(X < 20)$ (ii) $P(X < 19.2)$

(b) (i) $P(X \geq 14.3)$ (ii) $P(X \geq 8.6)$

(c) (i) $P(12.5 < X < 16.5)$ (ii) $P(10.1 \leq X \leq 15.5)$

4. The battery life of a certain brand of laptop batteries follows a normal distribution with mean 16 hours and standard deviation 5 hours. A particular battery has a life of 10.2 hours.

(a) How many standard deviations below the mean is this battery life?

(b) What is the probability that a randomly chosen battery has a life shorter than this? [6 marks]

5. Weights of a certain breed of cat follow a normal distribution with mean 16 kg and variance 16 kg^2 . In a sample of 2000 such cats, estimate the number that will have a weight above 13 kg.

[6 marks]

6. When Ali participates in long-jump competitions, the lengths of his jumps are normally distributed with mean 5.2 m and standard deviation 0.7 m.
- What is the probability that Ali will record a jump between 5 m and 5.5 m?
 - Ali needs to jump 6 m to qualify for the school team.
 - What is the probability that he will qualify with a single jump?
 - If he is allowed three jumps, what is the probability that he will qualify for the school team? [7 marks]
7. If $D \sim N(250, 400)$, find
- $P(D > 265 \cap D < 280)$
 - $P(D > 265 | D < 280)$
 - $P(D < 242 \cup D > 256)$ [6 marks]
8. If $Q \sim N(4, 160)$, find
- $P(5 < |Q|)$
 - $P(Q > 5 | 5 < |Q|)$ [6 marks]
9. The weights of apples in a certain shipment are normally distributed with mean weight 150 g and standard deviation 25 g. Supermarkets classify apples as 'medium' if their weights are between 120 g and 170 g.
- What proportion of apples in this shipment are 'medium'?
 - In a bag of 10 apples, what is the probability that there are at least 8 medium apples? [6 marks]
10. The wingspan of a species of pigeon is normally distributed with mean 60 cm and standard deviation 6 cm. A pigeon of this species is chosen at random.
- Find the probability that its wingspan is greater than 50 cm.
 - Given that this pigeon's wingspan is greater than 50 cm, find the probability that it is greater than 55 cm. [6 marks]

- 11.** Grains of sand are believed to have normally distributed widths with mean 2 mm and variance 0.25 mm^2 .
- Find the probability that a randomly chosen grain of sand is wider than 1.5 mm.
 - The sand is passed through a filter which blocks grains wider than 2.5 mm. The sand that has passed through the filter is examined. What is the probability that a randomly chosen grain of filtered sand is wider than 1.5 mm? [6 marks]

- 12.** The amount of paracetamol per tablet of a pain-relieving medicine is normally distributed with mean 500 mg and standard deviation 160 mg. A dose containing less than 300 mg is ineffective in alleviating toothache. In a trial of this medicine on 20 people suffering from toothache, what is the probability that two or more people get less than the effective dose? [6 marks]

18E The inverse normal distribution

In section 18D we saw how to find probabilities when we were given information about a normally distributed random variable, such as values that it should lie between. In real life there are many situations where we need to work backwards from given probabilities to estimate corresponding values of the variable. Doing so requires the **inverse normal distribution**.

KEY POINT 18.8

For a given value of probability p , the inverse normal distribution gives the value of x such that $P(X \leq x) = p$.

EXAM HINT

Remember that p must be a cumulative probability.



Many books also use the $\Phi(z)$ notation mentioned in section 18D to write inverse normal distributions: If $P(X \leq x) = p$, then $\Phi^{-1}(p) = z = \frac{x - \mu}{\sigma}$.

You can use your calculator to find values of an inverse normal distribution: you need to enter the values of p , μ and σ , and then the calculator will return the value of x . If the probability you are given is $P(X > x)$, you may have to calculate $P(X \leq x) = 1 - P(X > x)$ first.

EXAM HINT



Calculator skills sheet 13 on the CD-ROM explains how to work with inverse normal distributions. Some calculators are able to find the value of x such that $P(X > x) = p$ as well.



Worked example 18.11

The size of men's feet is thought to be normally distributed with mean 22 cm and variance 25 cm^2 . A shoe manufacturer wants only 5% of men to be unable to find shoes large enough for them. How big should their largest shoe be?

Convert the question into mathematical terms.

Use the inverse normal distribution. Since the information we are given is $P(X > x)$, we first convert it into a cumulative probability of the form $P(X \leq x)$.

Let X = length of a man's foot

Then $X \sim N(22, 25)$

We need the value of x such that

$$P(X > x) = 0.05$$

$$P(X \leq x) = 1 - P(X > x) = 0.95$$

$$\Rightarrow x = 30.2 \text{ (from GDC)}$$

So their largest shoe must fit a foot 30.2 cm long.

EXAM HINT

Estimating parameters will involve solving equations – sometimes simultaneous equations. As the numbers are usually not 'nice', you may want to use your calculator.

One of the main applications of statistics is to estimate parameters of the distribution from information found in the data. For example, suppose we know that a certain random variable can be described as normally distributed, but we don't know the mean or standard deviation. How can we estimate the unknown parameter(s) using other information available from the data? This is where the standard normal distribution comes in useful: we can replace all the X values by their Z -scores, which follow a distribution $N(0,1)$ where all parameters are known.

Worked example 18.12

The masses of gerbils are thought to be normally distributed with standard deviation 8.3 g. It is found that 30% of gerbils have a mass of more than 65 g. Estimate the mean mass of a gerbil.

Convert the question into mathematical terms.

Convert the probability into the form $P(X \leq k)$.

Use the inverse normal distribution for Z and $z = \frac{x - \mu}{\sigma}$.

Now we can solve for μ .

Let X = mass of a gerbil

Then $X \sim N(\mu, 8.3^2)$

and we know $P(X > 65) = 0.3$

$$\therefore P(X \leq 65) = 0.7$$

$$P(Z \leq z) = 0.7 \Rightarrow z = 0.524 \text{ (from GDC)}$$

$$\text{where } z = \frac{x - \mu}{\sigma} = \frac{65 - \mu}{8.3}$$

$$\therefore \frac{65 - \mu}{8.3} = 0.524$$

$$\Rightarrow \mu = 60.6 \text{ g}$$

Exercise 18E

- If $X \sim N(14, 49)$, find the value of x for which
 - $P(X < x) = 0.8$
 - $P(X < x) = 0.46$
 - If $X \sim N(36.5, 10)$, find the value of x for which
 - $P(X > x) = 0.9$
 - $P(X > x) = 0.4$
 - If $X \sim N(0, 12)$, find the value of x for which
 - $P(|X| < x) = 0.5$
 - $P(|X| < x) = 0.8$
- If $X \sim N(\mu, 4)$, find μ given that
 - $P(X > 4) = 0.8$
 - $P(X > 9) = 0.2$
 - If $X \sim N(8, \sigma^2)$, find σ given that
 - $P(X \leq 19) = 0.6$
 - $P(X \leq 0) = 0.3$
- If $X \sim N(\mu, \sigma^2)$, find μ and σ given that
 - $P(X > 7) = 0.8$ and $P(X < 6) = 0.1$
 - $P(X > 150) = 0.3$ and $P(X < 120) = 0.4$
 - $P(X > 0.1) = 0.4$ and $P(X \geq 0.6) = 0.25$
 - $P(X > 700) = 0.8$ and $P(X \geq 400) = 0.99$

4. IQ tests are designed to have a mean of 100 and a standard deviation of 20. What IQ score is needed to be in the top 2% of all scores? [5 marks]
5. Rabbits' masses are normally distributed with an average of 2.6 kg and a variance of 1.44 kg^2 . A vet decides that the heaviest 20% of rabbits are 'obese'. What is the minimum mass of an obese rabbit? [5 marks]
6. A manufacturer knows that his machines produce bolts whose diameters follow a normal distribution with standard deviation 0.02 cm. He takes a random sample of bolts and finds that 6% of them have diameter greater than 2 cm. Find the mean diameter of the bolts. [6 marks]
7. The times taken for students to complete a test are normally distributed with a mean of 32 minutes and standard deviation of 6 minutes.
- Find the probability that a randomly chosen student completes the test in less than 35 minutes.
 - 90% of students complete the test in less than t minutes. Find the value of t .
 - A random sample of 8 students had the time they spent on the test recorded. Find the probability that exactly 2 of these students completed the test in less than 30 minutes. [7 marks]
8. An old textbook says that the range of data can be estimated as 6 times the standard deviation. If the data is normally distributed, what percentage of the data is within this range? [6 marks]
9. (a) 30% of sand from Playa Gauss falls through a sieve with gaps of width 1 mm, but 90% passes through a sieve with 2 mm gaps. Assuming that the diameters of the grains are normally distributed, estimate the mean and standard deviation of the sand grain diameter.
- (b) 80% of sand from Playa Fermat falls through a sieve with gaps of width 2 mm, and 40% of this filtered sand passes through a sieve with 1 mm gaps. Assuming that the diameters of the grains are normally distributed, estimate the mean and standard deviation of the sand grain diameter. [7 marks]

Summary

- A **random variable** is a quantity whose value depends on chance. The **probability distribution** of the random variable is a list of all the possible outcomes together with their associated probabilities.
- Even though the outcome of any one observation of a random variable is impossible to determine, the **expectation** – i.e. the expected mean value – of observations can be predicted quite accurately by the formula

$$E(X) = \sum_x x P(X = x)$$

- Among a fixed number n of independent trials (each with two possible outcomes) with a constant probability p of success in each trial, the total number of successes X follows a **binomial distribution**: $X \sim B(n, p)$.
 - The probability of getting x successes is $P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$, $r = 0, 1, \dots, n$, which can be found using your calculator.
 - The mean ($E(X)$) of the binomial distribution is np .
 - The variance ($\text{Var}(X)$) is $np(1-p)$.
- The **normal distribution** is a continuous distribution which can be used to model many physical situations. A normal distribution is completely defined by its mean μ and variance σ^2 . Given that $X \sim N(\mu, \sigma^2)$, you can use a calculator to find probabilities of the form $P(x_1 \leq X \leq x_2)$, $P(X \leq x)$ or $P(X \geq x)$ if you enter the values of μ , σ and x_1 , x_2 or x .
- If we know probabilities relating to a variable that follows a normal distribution, we can deduce information about the values of the variable by using the **inverse normal distribution**: $P(X \leq x) = p$, where p is probability.
- For $X \sim N(\mu, \sigma^2)$, the **Z-score** of the value x measures the number of standard deviations that x is from the mean:

$$z = \frac{x - \mu}{\sigma}$$

The random variable Z whose values are these Z-scores follows a normal distribution with mean 0 and standard deviation 1, called the **standard normal distribution** ($Z \sim N(0, 1)$). This can be useful when trying to calculate the μ and σ^2 of a given normal distribution.

If we put in a **cumulative probability** p , a calculator can tell us the value of z such that $P(Z \leq z) = p$, from which we can then deduce information about the value of x , μ or σ using the Z-score relation above.

Introductory problem revisited

A casino offers a game in which a coin is tossed repeatedly. If the first head occurs on the first throw you get £2, if the first head occurs on the second throw you get £4, if the first head is on the third throw you get £8, and so on, with the prize doubling each time. How much should the casino charge for this game if they want to make a profit?

The probability of getting heads on the first throw is $\frac{1}{2}$, so $P(\text{win } £2) = \frac{1}{2}$. The probability of the first head being on the second throw is $P(\text{tails}) \times P(\text{heads}) = \frac{1}{4}$, so $P(\text{win } £4) = \frac{1}{4}$. The probability of the first head being on the third throw is $P(\text{tails}) \times P(\text{tails}) \times P(\text{heads}) = \frac{1}{8}$, so $P(\text{win } £8) = \frac{1}{8}$.

If the random variable X is the number of pounds won, then the probability distribution is as follows:

X	2	4	8	...	2^n	...
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$...	$\frac{1}{2^n}$...

Therefore the expected amount of winnings is

$$E(X) = 2 \times \frac{1}{2} + 4 \times \frac{1}{4} + 8 \times \frac{1}{8} + \dots = 1 + 1 + 1 + \dots$$

This sum continues for ever, so $E(X) = \infty$; that is, the expected payout over a long period of time is infinite – the casino cannot possibly charge enough to cover the expected payout.

Even though the expected payout is infinite, if you were offered the opportunity to play this game you should think twice. The calculation of $E(X)$ assumes that you can play the game infinitely many times, and in reality this is of course not possible. This is an example of a famous fallacy known as 'Gambler's Ruin'.



Mixed examination practice 18

Short questions

1. A factory that makes bottles knows that, on average, 1.5% of its bottles are defective. Find the probability that in a randomly selected sample of 20 bottles, at least one bottle is defective. [4 marks]

2. A biased die with four faces is used in a game. A player pays 10 counters to roll the die and receives a number of counters equal to the value shown on the die. The table below shows the different values on the die and the probability of each occurring.

Value	1	5	10	N
Probability	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{10}$

Find the value represented by N , given that the player has an expected loss of one counter each time he plays the game. [5 marks]

3. The test scores of a group of students are normally distributed with mean 62 and variance 144.
- (a) Find the percentage of students with scores above 80.
- (b) What is the lowest score achieved by the top 5% of the students? [6 marks]

4. When a boy bats at baseball, the probability that he hits the ball is 0.4. In a practice session he gets pitched 12 balls; let X denote the total number of balls he hits. Assuming that his attempts are independent of each other, find

- (a) $E(X)$
- (b) $P(X \leq \text{Var}(X))$ [5 marks]

5. The adult of a certain breed of dog has average height 0.7 m with variance 0.05 m^2 . If the heights follow a normal distribution, find the probability that of six independently selected dogs of this breed, exactly four are over 0.75 m tall. [5 marks]

6. When Robyn shoots an arrow at a target, the probability that she hits the target is 0.6. In a competition she has eight attempts to hit the target. If she gets at least seven hits on target, she will qualify for the next round.

- (a) Find the probability that she hits the target exactly 4 times.
- (b) Find the probability that she fails to qualify for the next round. [6 marks]

7. A company producing light bulbs knows that the probability of a new light bulb being defective is 0.5%.
- Find the probability that a pack of six light bulbs contains at least one defective bulb.
 - Mario buys 20 packs of six light bulbs. Find the probability that more than four of the packs contain at least one defective light bulb. [6 marks]
8. 200 people are asked to estimate the size of an angle: 16 gave an estimate which was less than 25° , and 42 gave an estimate which was greater than 35° . Assuming that the data follows a normal distribution, estimate its mean and standard deviation. [6 marks]
9. When a fair die is rolled n times, the probability of getting at most two sixes is 0.532 correct to three significant figures.
- Find the value of n .
 - Find the probability of getting exactly two sixes. [7 marks]

Long questions

- A bag contains a very large number of ribbons. One-quarter of the ribbons are yellow and the rest are blue. Ten ribbons are selected at random from the bag.
 - Find the expected number of yellow ribbons selected.
 - Find the probability that exactly six of the selected ribbons are yellow.
 - Find the probability that at least two of the selected ribbons are yellow.
 - Find the most likely number of yellow ribbons selected.
 - What assumption have you made about the probability of selecting a yellow ribbon? [11 marks]
- The probability that each student forgets to do homework is 5%, independently of whether other students do homework or not. If at least one student forgets to do homework, the whole class has to do a test.
 - If there are 12 students in a class, find the probability that the class will have to do a test.
 - For a class with n students, write down an expression for the probability that the class will have to do a test.
 - Hence find the smallest number of students in the class such that the probability of the class having to do a test is at least 80%. [12 marks]

3. Two children, Alan and Belle, each throw two fair cubical dice simultaneously. The score for each child is the sum of the two numbers shown on their respective dice.

- (a) (i) Calculate the probability that Alan obtains a score of 9.
(ii) Calculate the probability that Alan and Belle both obtain a score of 9.
- (b) (i) Calculate the probability that Alan and Belle obtain the same score.
(ii) Deduce the probability that Alan's score exceeds Belle's score.
- (c) Let X denote the largest number shown on the four dice.

- (i) Show that $P(X \leq x) = \left(\frac{x}{6}\right)^4$ for $x = 1, 2, \dots, 6$.
- (ii) Copy and complete the following probability distribution table.

X	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{1296}$	$\frac{15}{1296}$				$\frac{671}{1296}$

- (iii) Calculate $E(X)$.

[13 marks]

(© IB Organization 2002)