

Statistics 1 for AQA contents

- 1 Review: collecting and processing data 6**
 - A Experiment or survey 6
 - B Sampling methods 7
 - random, systematic, stratified and quota sampling; questionnaire design
 - C Recording and presenting data 10
 - stem-and-leaf diagram, grouped frequency table, histogram, proportional pie chart
 - D Averages 15
 - mode, median, cumulative frequency (graph and linear interpolation), mean from frequency table (ungrouped and grouped)
 - E Percentiles 19
 - interquartile range, box-and-whisker diagram
 - F Secondary data 22
 - Mixed questions 23
- 2 Variance and standard deviation 24**
 - A Measures of spread 24
 - variance, standard deviation
 - B Σ notation 26
 - for variance and standard deviation
 - C Calculating the standard deviation 27
 - using a calculator
 - D Scaling 29
 - E Working with frequency distributions 31
 - mean and standard deviation from grouped data
 - F Choosing measures 33
 - appropriateness of mode, median, mean, range, interquartile range, standard deviation, variance
- 3 Probability 38**
 - A Outcomes and events 38
 - set notation, Venn diagram, probability of complement, addition law, mutually exclusive events
 - B Conditional probability 42
 - C Independent events 44
 - relation to conditional probability, multiplication law
 - D Tree diagrams 46
 - conditional probability 'in reverse'
 - Mixed questions 50
- 4 Discrete random variables 52**
 - A Probability distributions 52
 - B Mean, variance and standard deviation 55
- 5 Binomial distribution 58**
 - A Pascal's triangle 58
 - probabilities for binomial distribution with $p = \frac{1}{2}$, combinations, $\binom{n}{r}$ notation
 - B Unequal probabilities 62
 - C Using the binomial distribution 63
 - conditions for use, calculations using formula
 - D Using tables of the binomial distribution 66
 - E Mean, variance, standard deviation 68
 - Mixed questions 70

6 Normal distribution 73

- A Proportions 73**
between a given number of standard deviations above or below the mean, shape, symmetry
- B The normal probability distribution 77**
continuous random variable, area for probability, standard normal distribution, use of table
- C Solving problems 80**
transformation to standard normal variable
- D Further problems 83**
percentage points table
- E Other continuous distributions 86**
Mixed questions 87

7 Estimation 90

- A The sampling distribution of the mean 90**
sample statistic as estimator of a population parameter, unbiased estimator of population mean
- B Reliability of estimates 93**
standard error of the sample mean
- C Further problems 95**
sampling distribution as a normal distribution
- D The central limit theorem 98**
estimating with non-normal populations
- E Estimating the variance 101**
unbiased estimator
- F Using estimated variances 104**
Mixed questions 107

8 Confidence intervals 110

- A Estimating with confidence 110**
point estimate, standard error of the estimator
- B Confidence intervals 112**
for population with normal distribution
- C Other population distributions 117**
use of the central limit theorem
- D Using an estimated variance 119**
Mixed questions 122

9 Linear regression 124

- A The least squares regression line 124**
- B Explanatory variables 128**
identifying response and explanatory variable, interpretation of intercept and gradient of regression line
- C Scaling 130**
- D Residuals 131**
identification of outliers, check of plausibility of model
- Mixed questions 133

10 Correlation 136

- A Measuring correlation 136**
positive and negative correlation, covariance, product moment correlation coefficient
- B Scaling 142**
- C Interpreting correlation 142**
spurious correlation, effect of outliers, non-linear relationships

Excel functions 147

Tables 148

- Cumulative binomial distribution function 148
- Normal distribution function 154
- Percentage points of the normal distribution 155

Answers 156

Index 186

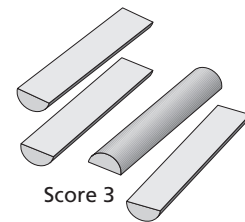
4 Discrete random variables

In this chapter you will learn

- what is meant by 'discrete random variable' and 'probability distribution'
- how to find the mean, variance and standard deviation of a discrete random variable

A Probability distributions (answers p 164)

'Senet' is a board game known to have been played in ancient Egypt. Instead of dice, four 'sticks' are thrown. Each stick is rounded on one side and flat on the other. The score depends on the number of sticks that land flat side up.



Number of flat sides up	0	1	2	3	4
Score	5	1	2	3	4

It is impossible to tell, just from the shape, what is the probability of a stick landing flat side up, although 'flat side up' seems more likely than 'curved side up'.

The only way to get an estimate of the probability is by experiment: throw the stick many times and find the **relative frequency** of flat side up.

In such an experiment a stick was thrown 100 times and landed flat side up 70 times. This gives 0.7 as an estimate of the probability of landing flat side up. The probability of landing curved side up is thus estimated as 0.3.

These estimates can be used to find the probability of each of the five scores. Assume that each stick falls independently of the others, either F (flat side up) or C (curved side up). We can list all the different possible ways the four sticks could land:

- FFFF (only outcome with 4 flat sides up)
 FFFC FFCF FCFF CFFF (outcomes with 3 flat sides up)
 FFCC FCCF CCFF FCFC CFCF CFFC (with 2 flat sides up)
 and so on

A1 Complete this list of all the possible outcomes for the four sticks.

The sixteen outcomes are not equally likely. For example, the probability of getting FFFF is $0.7 \times 0.7 \times 0.7 \times 0.7$ (because the four sticks fall independently of each other), or 0.2401. This is also the probability of getting a score of 4.

There are four different ways to get a score of 3: FFFC, FFCF, FCFF, CFFF.

The probability of each of these ways is $0.7^3 \times 0.3 = 0.1029$.

So the probability of getting a score of 3 is $4 \times 0.1029 = 0.4116$.

A2 Find the probability of getting a score of (a) 2 (b) 1 (c) 5

(Reminder: 5 is scored when no stick lands flat side up.)

The score is an example of a **discrete random variable**. This means a variable that can take individual values (usually integers), each with a given probability.

Let S stand for the score. (Capital letters are used for random variables.)

$P(S = 3)$ means ‘the probability that $S = 3$ ’.

We have already found that $P(S = 3) = 0.4116$.

The complete set of probabilities (the **probability distribution** of S) is shown in this table.

s	1	2	3	4	5
$P(S = s)$	0.0756	0.2646	0.4116	0.2401	0.0081

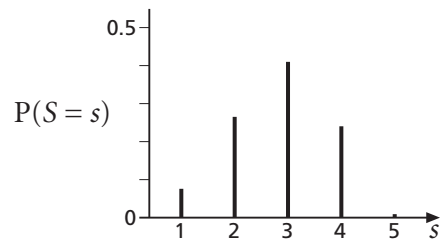
Notice that the small letter s is used for individual values of the random variable S .

$P(S = s)$ is also called the **probability function** of S .

A3 What is the sum of all the probabilities in the table?

The probability distribution can also be shown in a ‘stick graph’. The total of the heights of the sticks is 1.

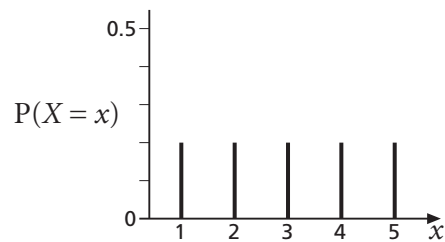
The probability distribution would be useful to someone wanting to design a computer version of Senet. The scoring device would need to have this distribution if it were to behave like the sticks.



The simplest type of probability distribution is one where all the probabilities are equal (as, for example, with the score on a fair dice).

Such a distribution is called **uniform**.

The random variable X shown here has a uniform distribution.



Exercise A (answers p 164)

1 A company that makes games has the idea of using a dice in the shape of a cuboctahedron. This has two kinds of face, squares and triangles.



The company makes a prototype dice and rolls it 500 times. It lands with square uppermost 300 times and triangle uppermost 200 times.

(a) Write down the estimates of the probabilities of ‘square’ and ‘triangle’.

In the planned game, two of the dice are rolled. If both land triangle uppermost the player scores 3. If both land square uppermost the player scores 2. Otherwise the player scores 1.

(b) Let S be the score. Using a tree diagram or otherwise, find $P(S = 1)$, $P(S = 2)$ and $P(S = 3)$.

Make a table and a stick graph showing the probability distribution of S .

- 2** A pack contains four cards: Ace, King, Queen, Jack. In a game played with this pack, the scoring system is as follows.
- Pick a card at random: if it is the Ace, score 3; if it isn't, pick again without replacing the first card. If the card picked this time is the Ace, score 2; otherwise pick again without replacing. If you get the Ace this time, score 1. Otherwise score 0.

- (a) With the help of a tree diagram, make a table showing the probability distribution of the score S . Sketch a stick graph of the distribution.
- (b) What type of distribution is it?

- 3** Two children play a game where they roll two ordinary dice. The score D is the difference between the numbers on the dice. (The difference is always positive.)

		1st dice					
		1	2	3	4	5	6
2nd dice	1	0	1	2	3		
	2	1	0	1			
	3	2	1	0			
	4						
	5						
	6						

- (a) Copy and complete the table on the right which shows the value of D for every possible outcome.

- (b) Copy and complete this table to show the probability distribution of D .

d	0	1	2	3	4	5
$P(D = d)$	$\frac{6}{36}$					

- (c) Sketch a stick graph of the distribution.
- (d) One child suggests a simple dice game: 'I win if the difference is less than 3; you win if the difference is 3 or more.' Is this a fair game? If not, suggest a different, but fair, rule for winning, still based on differences.

- 4** A game is played with a single ordinary dice. The scoring system is as follows.
- Roll a six first time: score 3; otherwise roll again.
- Roll a six second time: score 2; otherwise roll again.
- Roll a six third time: score 1, otherwise score 0.
- Make a table showing the probability distribution of the score, S .

- 5** The probability function of a discrete random variable X is defined by $P(X = x) = \frac{1}{10}x$, $x = 1, 2, 3, 4$.
- For example, $P(X = 4) = \frac{1}{10} \times 4 = 0.4$
- Find $P(X = 1)$, $P(X = 2)$, $P(X = 3)$ and $P(X = 4)$ and show that they add up to 1.

- *6** X is a discrete random variable. The probability function $P(X = x)$ is defined by $P(X = x) = kx^2$, $x = 1, 2, 3, 4$.
- (a) Write down, in terms of k , the values of $P(X = 1)$, $P(X = 2)$, $P(X = 3)$ and $P(X = 4)$.
- (b) Explain why the value of k must be $\frac{1}{30}$.

B Mean, variance and standard deviation (answers p 164)

Here again is the probability distribution of the score S when throwing the Senet sticks.

s	1	2	3	4	5
$P(S = s)$	0.0756	0.2646	0.4116	0.2401	0.0081

Imagine throwing the sticks 10 000 times.

The number of times you would expect to get a score of 1 is $10\,000 \times 0.0756 = 756$.

Similarly, a score of 2 would be expected 2646 times, and so on.

These frequencies can be used to calculate a mean score:

$$\text{Mean score} = \frac{(1 \times 756) + (2 \times 2646) + (3 \times 4116) + (4 \times 2401) + (5 \times 81)}{10000} = 2.84 \quad (\text{to 2 d.p.})$$

It was unnecessary to multiply all the probabilities by 10 000 and then divide by 10 000 at the end. The mean score can be calculated using the probabilities themselves:

$$\begin{aligned} \text{Mean score} &= (1 \times 0.0756) + (2 \times 0.2646) + (3 \times 0.4116) + (4 \times 0.2401) + (5 \times 0.0081) \\ &= 2.84 \quad (\text{to 2 d.p.}) \end{aligned}$$

2.84 is the mean of the random variable S .

To calculate it, each possible value of the random variable is multiplied by its probability, and the products are added together.

$$\text{Mean of } S = \sum s \times P(S = s)$$

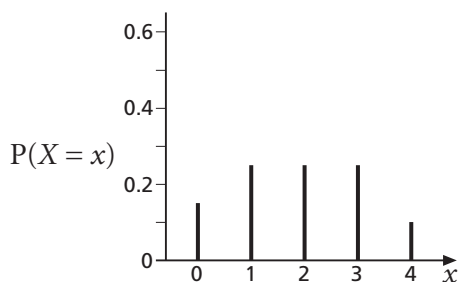
- B1** Find the mean of the random variable X whose probability distribution is shown in this table.

x	0	1	2	3
$P(X = x)$	0.35	0.3	0.2	0.15

- B2** The probability distributions for the scores X and Y in two different games are given below.

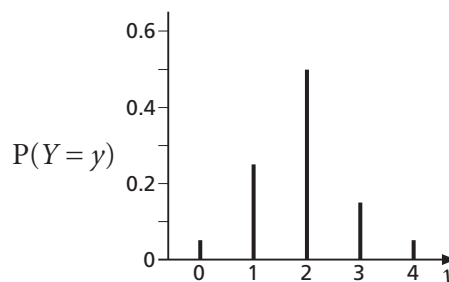
Game A

x	0	1	2	3	4
$P(X = x)$	0.15	0.25	0.25	0.25	0.10



Game B

y	0	1	2	3	4
$P(Y = y)$	0.05	0.25	0.50	0.15	0.05



- (a) Calculate the mean score for each game.
 (b) The distributions are alike as far as the mean score is concerned. What is different about them?

The **variance** of a random variable is defined in a similar way to the variance of a frequency distribution. The score X in game A in question B2 will be used here as an example.

The mean of a random variable is usually denoted by the Greek letter μ ('mu').

In this case, $\mu = 1.9$.

The deviation of each possible value from the mean is found.

This is $x - \mu$.

Each deviation is squared: $(x - \mu)^2$.

Each squared deviation is multiplied by the corresponding probability $P(X = x)$.

The total, 1.49, is the variance.

x	$x - \mu$	$(x - \mu)^2$	$P(X = x)$	$(x - \mu)^2 \times P(X = x)$
0	-1.9	3.61	0.15	0.5415
1	-0.9	0.81	0.25	0.2025
2	0.1	0.01	0.25	0.0025
3	1.1	1.21	0.25	0.3025
4	2.1	4.41	0.10	0.4410

Variance = 1.49

K The **standard deviation** of a random variable is denoted by σ (the small Greek letter 'sigma'). This is the square root of the variance, so the variance is denoted by σ^2 .

In the example above, $\sigma^2 = 1.49$, so $\sigma = \sqrt{1.49} = 1.22$ (to 3 s.f.).

K The definitions of the mean and variance are:

$$\mu = \sum x \times P(X = x)$$

$$\sigma^2 = \sum (x - \mu)^2 \times P(X = x)$$

As in the case of a frequency distribution, an equivalent expression for the variance is easier to use in practice:

$$\sigma^2 = \sum x^2 \times P(X = x) - \mu^2$$

- B3** Calculate the variance and standard deviation of the score in game B in question B2. Which of the two games has the wider variation in scores? How could you also tell this from the graphs?

Exercise B (answers p 164)

- 1** The probability distribution of the score S on an ordinary dice is shown here. Calculate

s	1	2	3	4	5	6
$P(S = s)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- (a) the mean of S (b) the variance of S (c) the standard deviation of S

- 2** Two ordinary dice are rolled. The score D is the difference between the two numbers (as in exercise A, question 3).

Here is the probability distribution of D .

d	0	1	2	3	4	5
$P(D = d)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Calculate

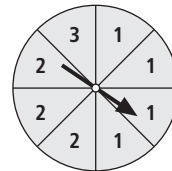
- (a) the mean of D (b) the variance of D (c) the standard deviation of D

Key points

- A discrete random variable takes individual values (usually integers), each with a given probability. The probability distribution can be shown in a table or a graph. (p 53)
- The mean, μ , of a discrete random variable is defined by $\mu = \sum x \times P(X = x)$. (p 55)
- The variance, σ^2 , is defined by $\sigma^2 = \sum (x - \mu)^2 \times P(X = x)$.
An equivalent form of this equation is $\sigma^2 = \sum x^2 \times P(X = x) - \mu^2$. (p 56)
- Standard deviation = $\sqrt{\text{variance}}$. (p 56)

Test yourself (answers p 165)

- 1 The diagram shows a fair spinner.
Let X be the score when the spinner is spun.
The table shows part of the probability distribution of X .



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

- (a) Copy and complete the table.
(b) Calculate the mean of X .
(c) Calculate the variance of X .
- 2 (a) The probability distribution of the litter size S in a species of animal is given in this table.

s	1	2	3	4
$P(S = s)$	0.3	0.2	0.3	0.2

- (i) Find $P(S \geq 3)$.
(ii) Calculate the mean of S .
(iii) Calculate the variance of S .
- (b) The probability distribution of the litter size T in a second species is as follows.

t	1	2	3	4
$P(T = t)$	0.05	0.4	0.5	0.05

- (i) Calculate the mean of T .
(ii) Calculate the variance of T .
- (c) (i) In which species is litter size larger, on average?
(ii) In which species is there wider variation in litter size?