Please check the examination details below before entering your candidate information

Candidate surname	Other names

Pearson Edexcel Level 3 GCE

Centre Number	Candidate Number

Time 2 hours

Paper reference **9ST0/01**

Statistics

Advanced
PAPER 1: Data and Probability

You must have:
Statistical formulae and tables booklet
Calculator

Total Marks

Candidates may use any calculator allowed by Pearson regulations.
Calculators must not have retrievable mathematical formulae stored in them.

Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B).
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided – *there may be more space than you need*.
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- Unless otherwise stated, inexact answers should be given to three significant figures.
- Unless otherwise stated, statistical tests should be carried out at the 5% significance level.

Information

- A booklet 'Statistical formulae and tables' is provided.
- There are 8 questions in this question paper. The total mark for this paper is 80.
- The marks for **each** question are shown in brackets – *use this as a guide as to how much time to spend on each question*.

Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.
- If you change your mind about an answer, cross it out and put your new answer and any working underneath.
- Good luck with your examination.

Turn over ▶

P66655A

1/1/1/1/1/

P 6 6 6 5 5 A 0 1 2 0

Pearson

Answer ALL questions. Write your answers in the spaces provided.

1 Josh works as a proofreader at a company. He is currently correcting the errors in spelling, punctuation, and grammar in the text of a new novel.

He has to make a record of every change he makes to the text.

The author's last novel had 136 pages and needed a lot of corrections. The bar chart in **Figure 1** shows the distribution of text errors per page in this last novel.

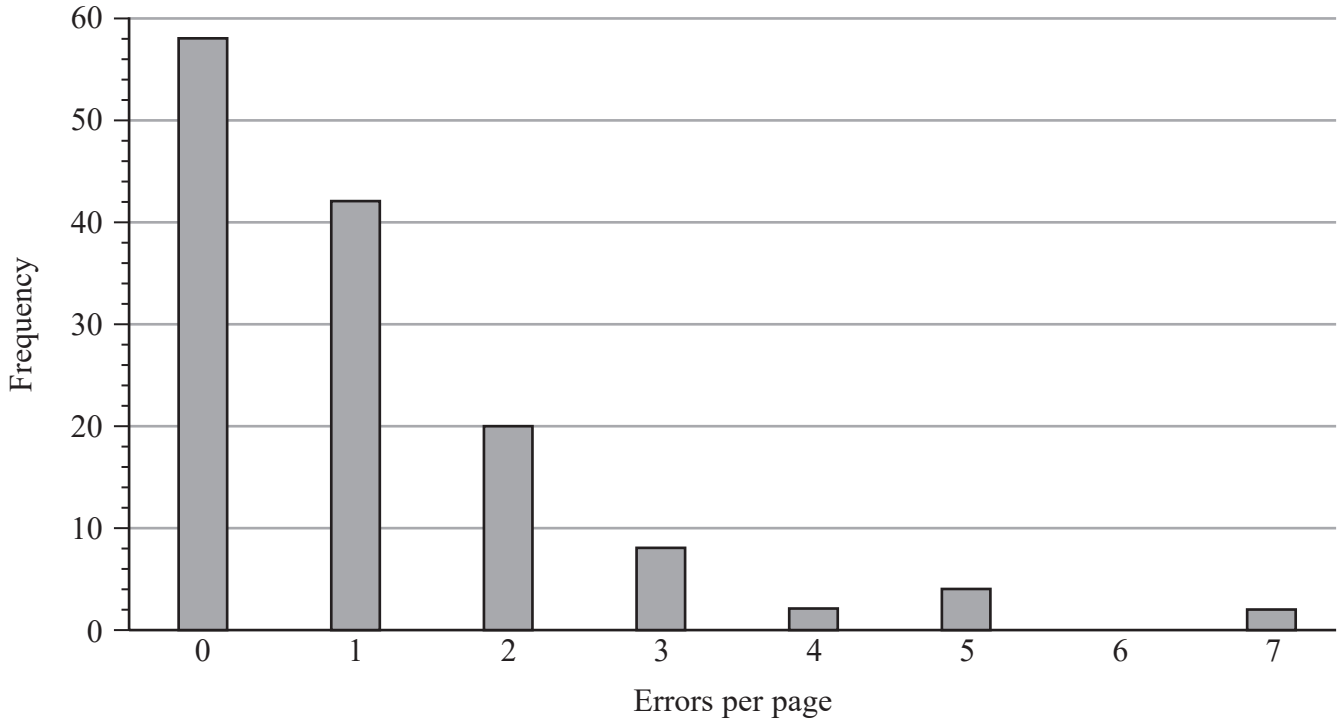


Figure 1

(a) Assuming that the new novel has a similar distribution of errors to the last novel, estimate,

(i) the probability that a randomly chosen page has fewer than 2 errors,

(2)

(ii) the expected number of errors per page.

(3)

Question 1 continued

The new novel has 223 pages.

(b) Assuming that the new novel has a similar distribution of errors to the last novel, estimate the expected number of errors in the new novel.

(1)

(c) Give one reason why the assumption used in (a) and (b) may be valid.

(1)

(d) Give one reason why the assumption used in (a) and (b) may **not** be valid.

(1)

(Total for Question 1 is 8 marks)

2 When a signal is converted from analogue to digital, a small amount of information is lost. This loss is called the quantisation error. The probability density function of the quantisation error, measured in LSBs (least significant bits) is shown in **Figure 2**.

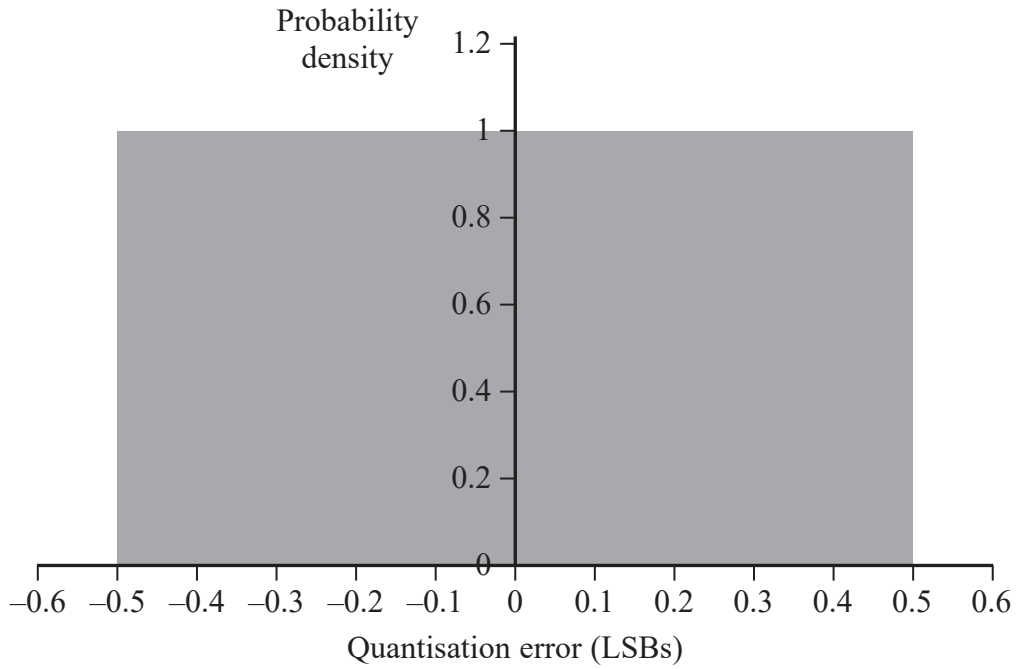


Figure 2

(a) At a randomly chosen time, find the probability that

(i) the quantisation error is 0

(1)

(ii) the quantisation error is greater than 0.28

(2)

Question 2 continued

Jo-Anne is investigating the error levels of radio signals detected from space. She classes an error that is no more than 0.1 LSBs away from 0 as a 'minor error'.

(b) At a randomly chosen time, find the probability that the quantisation error is a 'minor error'.

(2)

Jo-Anne takes quantisation error readings once every second during a 30-second interval.

(c) Find the probability that at least 10 of the readings show a 'minor error'.

(3)

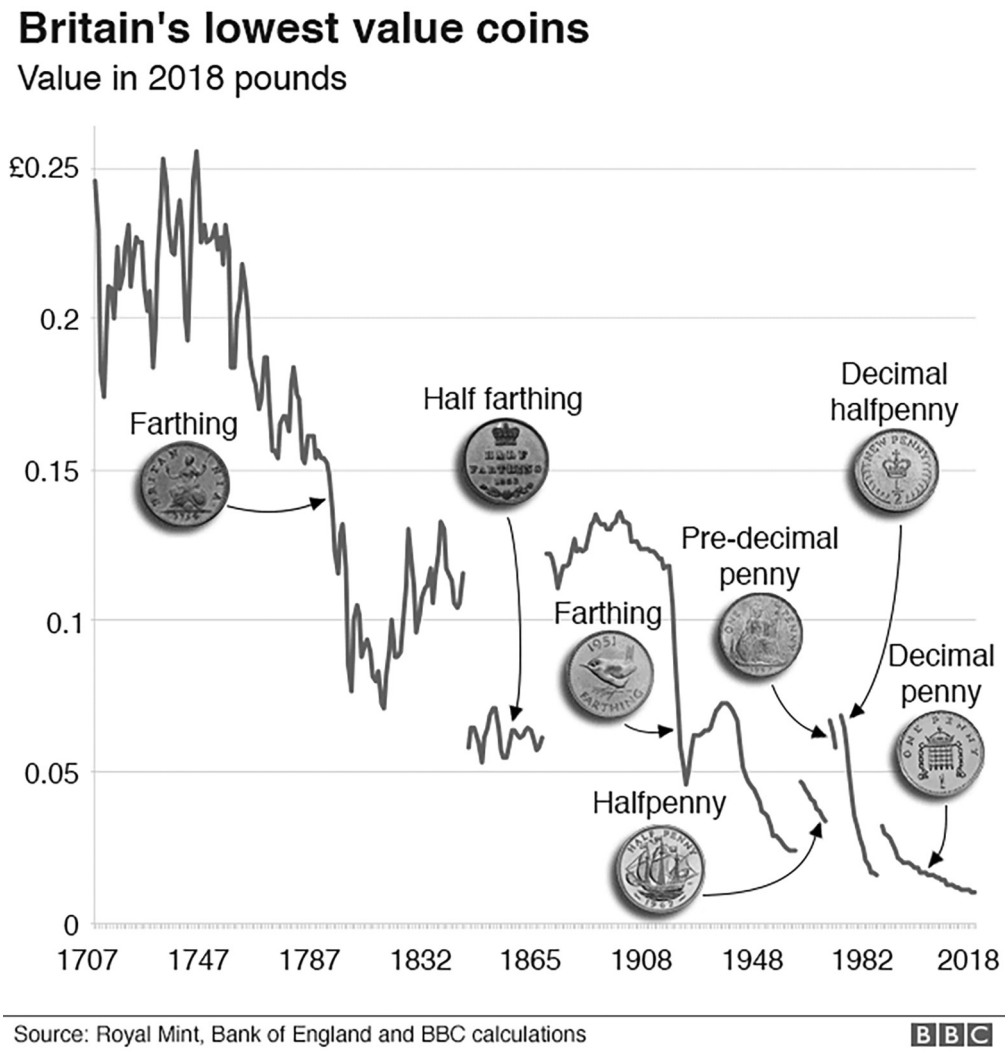
(d) State an assumption that needs to be made to ensure the answer in (c) is valid.

(1)

(Total for Question 2 is 9 marks)

3 In 2019, the UK Treasury completed a consultation on the use of 1p and 2p coins, which led to a decision to keep these coins in circulation for the foreseeable future.

The BBC News team ran an investigation into the lowest value British coins since 1707 (when the Kingdom of Great Britain was formed).



[Source: *https://www.bbc.co.uk/news/business-48153442*]

Figure 3

Question 3 continued

Figure 3 was published on a website to be read by the general public.

(a) Make **two** criticisms of the graph in **Figure 3**.

(2)

(b) Using the graph in **Figure 3**, determine how many farthings had the same value as a halfpenny (**not** the decimal halfpenny).

(1)

(Total for Question 3 is 3 marks)

4 Kit is researching a type of mould that grows on oak and ash trees.

He decides to take a sample of 20 trees from a local arboretum to study.

The arboretum has a detailed database listing all of the trees growing there.

Approximately 5% of the trees in the arboretum are oak trees, and approximately 3% are ash trees.

(a) State the name of each of the following three sampling methods.

A Kit assigns a number to each of the oak trees and ash trees in the arboretum, and uses a random number generator to pick 10 oak trees and 10 ash trees.

(2)

B Kit chooses 10 oak trees near the southern entrance to the arboretum, and 10 ash trees near the western entrance.

(1)

C Kit asks the manager of the arboretum to select a sample of 10 oak trees and 10 ash trees that she feels is representative of the trees in the arboretum.

(1)

There are 917 oak trees in the arboretum.

(b) Explain how you would use technology to get a systematic sample of 10 oak trees in a list format to be printed out.

(4)

Question 4 continued

Kit decides that he definitely wants a sample of 10 oak trees and 10 ash trees.

He decides to use sampling method C.

(c) Give, in context, one advantage and one disadvantage of Kit's selected sampling method.

(2)

(Total for Question 4 is 10 marks)

5 Test A is a simple saliva test designed to detect whether an adult is a user of a particular psychoactive drug.

This test is used in a large secure facility where random testing for this drug routinely takes place.

Previous research shows that 1% of adults living in the facility are users of this drug.

If an adult is a user of this drug, the probability of Test A returning a positive result is 97%

If an adult is **not** a user of this drug, the probability of Test A returning a (false) positive result is 4%

(a) Show that the probability that a randomly chosen adult living in this facility, who had a positive Test A result, actually is a user of this drug is 0.1968, correct to 4 decimal places.

(6)

Question 5 continued

(b) Explain why the probability given in (a) may be of concern.

(1)

Test A is inexpensive and easy to administer.

Another, Test B, also detects the presence of the psychoactive drug.

Test B is considerably more accurate than Test A. However, Test B is more expensive, and involves the need for trained medical staff and a more invasive test procedure.

(c) State a situation where the facility manager in charge of monitoring this drug usage might decide to use **both** Test A and Test B.

You should also **explain** why the use of both tests would be beneficial both to the manager and the adults living in the facility.

(4)

(Total for Question 5 is 11 marks)

6 Yasmine is investigating the effectiveness of some home cures for hiccups, such as drinking water, being startled, or biting on a lemon. She plans to run 500 trials using a large sample of volunteers. Her experiment runs as follows.

Yasmine induces hiccups on a volunteer and starts a timer. After a minute, she tries to 'cure' the hiccups using one of the methods, then observes whether the hiccups have stopped.

(a) State the conditions under which the Poisson distribution would be a suitable model for the number of hiccups in the first minute.

(3)

Yasmine decides that the Poisson distribution is a suitable model, and she calculates that the mean number of hiccups in the first minute for the sample of volunteers is 8.5

After drinking water, Volunteer A hiccups only 2 times in the next minute.

(b) Assuming that the cure has **not** worked, and that hiccups are still occurring at this constant rate, find the probability that a volunteer would hiccup 2 times or fewer.

(2)

After being startled, Volunteer B does not hiccup for 30 seconds (half a minute).

(c) Assuming that the cure has **not** worked, and that hiccups are still occurring at this constant rate, find the probability that a volunteer would not hiccup for the first 30 seconds.

(4)

Question 6 continued

Yasmine is planning to observe the volunteers for 30 seconds after the 'cure' before moving on to the next volunteer.

(d) Do you think 30 seconds is a sufficient amount of time for Yasmine to wait?

You should explain your answer using your answer to (c) and any other relevant information found in this question.

(3)

(Total for Question 6 is 12 marks)

7 Snooker is a sport played between two players. The players participate in a number of games, called Frames, each of which is won by one of the players.

In the final match of the Snooker World Championship, the winner is the first player to win 18 Frames.

The Frames are played in four sessions, as follows:

Session 1	**(Sunday afternoon)**	8 Frames
Session 2	**(Sunday evening)**	9 Frames
Session 3	**(Monday afternoon)**	(up to) 8 Frames
Session 4	**(Monday evening)**	(up to) 10 Frames

[Note that the maximum possible number of Frames is 35, with a score of 18–17]

You may assume that the probability of a player winning each Frame remains constant, with results between Frames being independent of one another.

This year's final is to be played between Gordon and Ping.

Gordon is considered to be a better player than Ping, with the probability of Gordon winning a Frame equal to 0.55

(a) After the first two sessions (17 Frames), find the expected number of Frames won by Gordon.

(2)

(b) After the first two sessions, find the probability that Gordon will have won more Frames than Ping.

(3)

Question 7 continued

(c) Calculate the probability that the match will finish within the first three sessions (i.e. in the first 25 Frames).

(3)

Each of sessions 1–4 has a different group of nearly 1000 ticket-buying spectators, and each is broadcast live on national television.

(d) Explain why the probability found in (c) may be of concern to the championship organisers.

(1)

(e) Suggest **two** possible solutions for the organisers.

(2)

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA

Question 7 continued

(f) Discuss the validity of the assumptions made throughout this question.

(3)

(Total for Question 7 is 14 marks)

BLANK PAGE

8 Therese is studying some data collected at the Large Hadron Collider (LHC), a particle collider located in France and Switzerland.

The data contains light-intensity measurements from two different sensors (left and right), measured immediately after a collision of protons. When enough light strikes the sensor, a measurement can be made of its intensity.

Therese wants to investigate the relationship between the two measurements.

She takes a random sample of 15 collisions and produces the following summary statistics using a spreadsheet program.

Pearson: $r = 0.8000$

Regression: $y = 1.8392 + 1.0771x$

The data for this sample is given in **Figure 4**.

Collision	Left sensor measurement (x)	Right sensor measurement (y)	Residual (to 2 d.p.)
A	3.59	3.58	
B	0	0.88	–0.96
C	10.28	12.32	–0.59
D	1.97	2.68	–1.28
E	3.04	5.81	0.70
F	2.45	5.11	0.63
G	0	9.81	7.97
H	7.38	4.03	–5.76
I	4.83	5.65	–1.39
J	1.25	4.47	1.28
K	14.19	21.81	4.69
L	1.42	3.53	0.16
M	0	4.51	2.67
N	1.29	0	–3.23
O	0.86	0	–2.77

[Data source: *DOI:10.7483/OPENDATA.OPERA.XF0L.CYE0*]

Figure 4

(a) For each of the sensor measurements, write down the modal value.

(1)

Question 8 continued

(b) Considering practical aspects of the context, explain why the values found in part (a) may be expected to be the modal values.

(2)

(c) Find the missing residual for collision A.

(2)

Therese identifies collision G as a significant outlier.

(d) Describe, in context, a situation in which it would be appropriate to discard the data for this collision in any analysis of the data.

(1)

Therese decides that it is appropriate to **discard** the data for collision G.

(e) For the new dataset of 14 collisions, find

(i) Pearson's product moment correlation coefficient, r

(1)

(ii) the least squares regression line in the form $y = a + bx$

(2)

Question 8 continued

Therese's assistant also calculated values for r, a and b on this new dataset of 14 values. However, he entered the data in the reverse order in his spreadsheet. He entered the left sensor data in the column Therese labelled (y) and the right sensor data in the column Therese labelled (x).

(f) Which of the calculated values in (e) would be expected to remain the same?

Give a reason for your answer.

(2)

(g) Comment on Therese's decision to quote the values of r, a and b correct to 4 decimal places.

(2)

(Total for Question 8 is 13 marks)

TOTAL FOR PAPER IS 80 MARKS