Please check the examination details below before entering your candidate information

Candidate surname	Other names

Pearson Edexcel Level 3 GCE

Centre Number | Candidate Number

Tuesday 16 June 2020

Afternoon (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 9 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.

Turn over ▶

P62709A

1/1/1/1/

P 6 2 7 0 9 A 0 1 2 8

Pearson

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

1 Shenandoah works at a medical practice. She wants to find out about patients' opinions on a planned extension to the practice and the disruption this may cause.

She plans to ask several questions to a sample of patients.

For **each** of the following methods, state **one** advantage, and **two** potential sources of bias.

(a) Method A: Give a questionnaire, to be filled out whilst waiting, to each patient in the waiting room at the practice on a certain morning.

(i) one advantage

(1)

(ii) two potential sources of bias

(2)

Question 1 continued

(b) Method B: Number the list of all patients registered at the practice and use a simple random sample (without replacement) to select a sample of patients.

Email a link, to an online survey, to each of these patients.

(You may assume that all patients have given their permission to be contacted by email.)

(i) one advantage

(1)

(ii) two potential sources of bias

(2)

(Total for Question 1 is 6 marks)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

2 For **each** of the **random variables identified in part (a) and part (b)** decide which **one** of the following best applies.

- The Poisson distribution is a **suitable** model.
  In this case, you should write down the value of the parameter (if known).
- The suitability of the Poisson distribution as a model is **dependent** on some additional factor(s).
  In this case, you should state what factor(s) the suitability depends on.
- The Poisson distribution is **unsuitable** as a model.
  In this case, you should explain in context why it is unsuitable.

(a) X = the number of buses that pass a fixed point on a busy road during a one-hour interval between 08:00 and 09:00

Experimental data produced the summary statistic $\bar{x} = 7.6$

(3)

Question 2 continued

(b) Y = the number of meteorites of mass greater than 10 grams that hit the Earth's surface in a 10-year interval.

A study estimated that the average number of such meteorites that hit the Earth's surface is 7.73×10^{-5} per square kilometre per year.

The Earth has a surface area of 5.10×10^{8} square kilometres.

(3)

(Total for Question 2 is 6 marks)

3 Sylvia runs a course at her local university. She decides to assess her students using both a coursework assignment and an examination.

Sylvia believes that the skills shown in the coursework are more important than those shown in the examination, so she decides to mark the coursework as an integer score out of 100, and the examination as an integer score out of 50

The scores for the first year of Sylvia's course assessments are given in **Figure 1** and **Figure 2**.

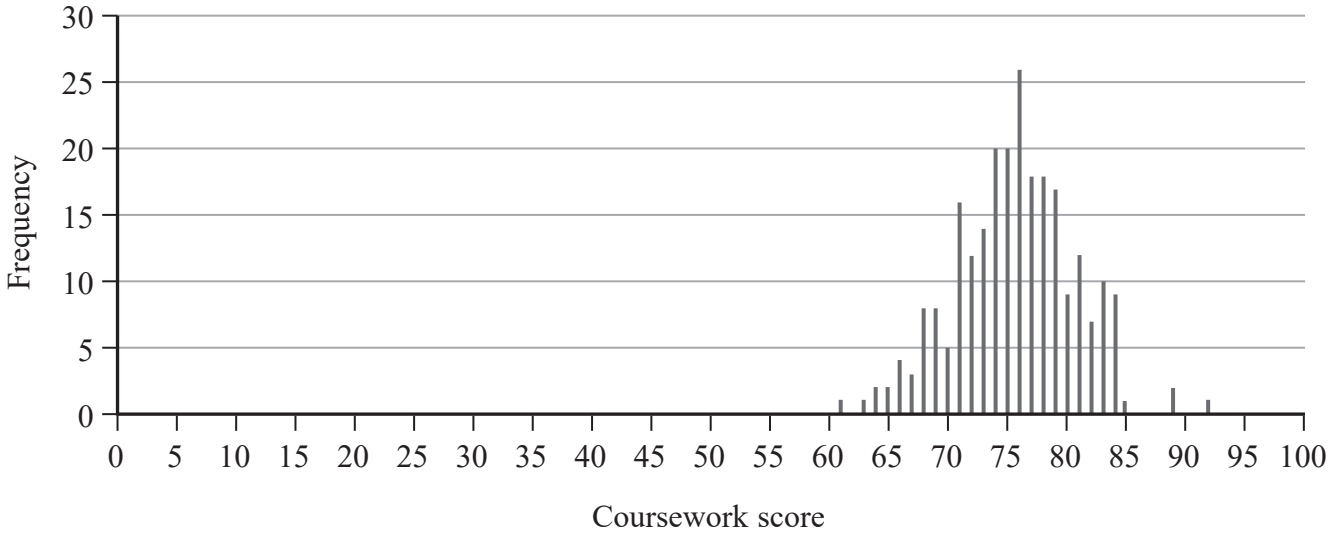

Figure 1: Coursework scores in first year of assessments

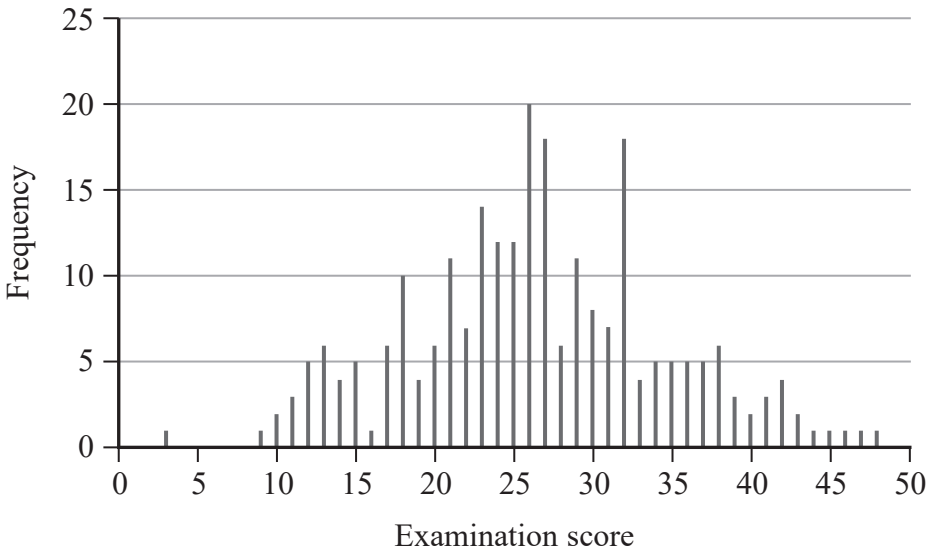

Figure 2: Examination scores in first year of assessments

Question 3 continued

Sylvia looks at **Figure 1** and **Figure 2** and decides that the normal distribution might be an appropriate model for each of the populations from which the data were taken. Tests for normality confirm her decision.

Hence, Sylvia uses normal distributions to model the scores.

(a) Describe:

(i) one advantage of Sylvia's model,

(1)

(ii) one disadvantage of Sylvia's model.

(1)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 3 continued

Summary statistics from the data are given below.

	Coursework	Examination
Mean	76.75	27.39
Standard deviation	4.76	8.19

(b) Assuming that this year's students are a representative sample of all future students, use Sylvia's model to estimate the **coursework** score of a student at the 10th percentile of the coursework scores (the score below which 10% of the coursework scores lie).

You should give your answer to the nearest integer.

(3)

(c) Assuming that this year's students are a representative sample of all future students, use Sylvia's model to estimate the **examination** score of a student at the 90th percentile of the examination scores (the score below which 90% of the examination scores lie).

You should give your answer to the nearest integer.

(1)

Question 3 continued

Each student's final score is the total of their coursework score and their examination score.

Student A scored at the 10th percentile in the coursework and at the 90th percentile in the examination.

Student B scored at the 90th percentile in the coursework with a score of 83, and at the 10th percentile in the examination with a score of 17. This gave Student B a total score of 100.

(d) Do you think that Sylvia has successfully weighted the marks so that the total score reflects her goal that coursework scores are more important than the examination scores in the final score achieved?

Explain your answer.

(4)

(e) What recommendations would you make for Sylvia when designing next year's assessments?

(2)

(Total for Question 3 is 12 marks)

4 London Fire Brigade keeps detailed records of every incident the fire service responds to. The full dataset contains 31 fields of data for each individual incident dating from 1st January 2017 until the present day.

Some selected fields are shown in **Figure 3**.

Field name	Data type	Description
IncidentNumber	Number	ID number of incident
DateOfCall	Date/Time	Date of call
TimeOfCall	Date/Time	Time of call
IncidentGroup	Short Text	Type of incident ('Fire', 'Special Service', or 'False Alarm')
PropertyCategory	Short Text	Type of property/location (e.g. 'Dwelling', 'Non-residential', 'Outdoor')
Postcode_full	Short Text	Full postcode of location
IncGeo_BoroughName	Short Text	Name of London Borough
Easting_m	Number	UTM geographical coordinate of location (East)
Northing_m	Number	UTM geographical coordinate of location (North)
IncidentStationGround	Short Text	Name of fire station attending incident
FirstPumpArriving_AttendanceTime	Number	Time (in seconds) for first fire engine to arrive at the location
NumPumpsAttending	Number	Number of fire engines attending the incident.

(Source: UK Open Government Licence v2)

Figure 3: Table structure of 'London Fire Brigade Incident data from January 2017'

Explain how you would use database software to find the top 10 fastest arrival times for 'Fire' incidents during **2018**.

(4)

Question 4 continued

(Total for Question 4 is 4 marks)

5 A game is played with a pack of 24 cards, each of which has a picture of a man or a woman on it, as well as his/her name.

If a card is picked at random, the following events are defined:

H: The person in the picture is wearing a hat.
W: The person in the picture is a woman.
R: The person in the picture has red hair.
B: The person in the picture has a beard.

When the whole pack is considered, some of the frequencies of each event are presented in the Venn diagram in **Figure 4**.

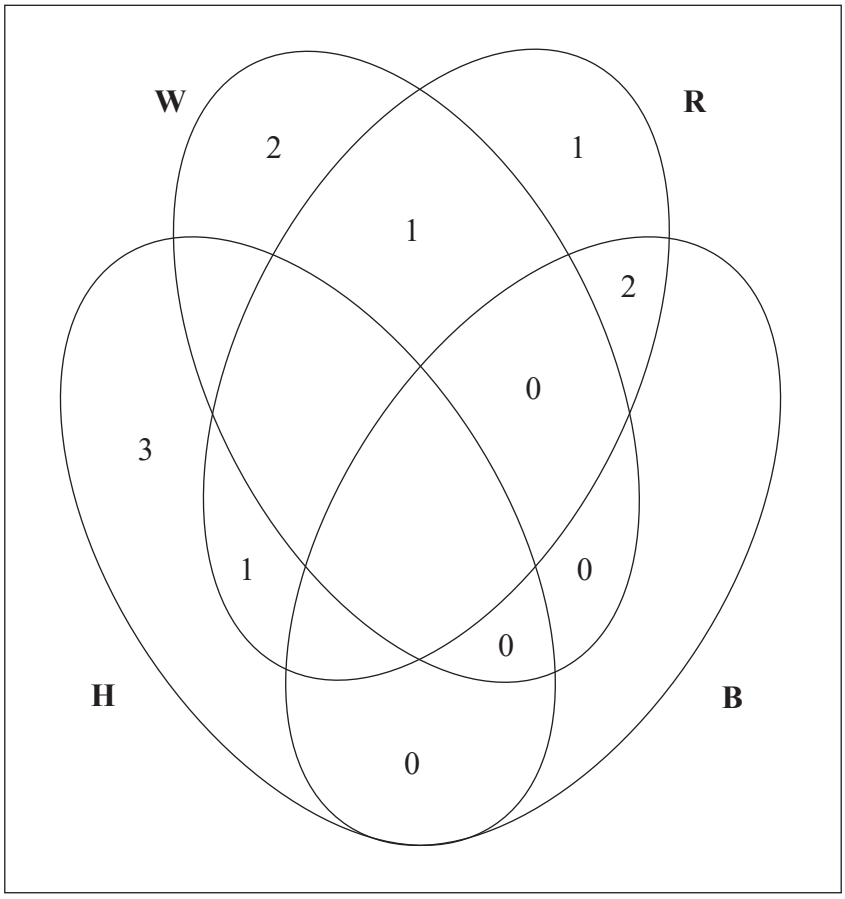

Figure 4

In the pack of cards:

- there is one red-haired **man** with a hat and a beard,
- there are no red-haired women wearing a hat,
- there are nine people with a beard,
- exactly **half** of the women are wearing hats.

(a) Fully complete the Venn diagram in **Figure 4**.

(5)

Question 5 continued

A card is randomly drawn from the pack.

(b) Find the following probabilities:

(i) $P(H \cap W' \cap R \cap B)$

(1)

(ii) $P(H' \cap W' \cap R' \cap B')$

(1)

(iii) $P(W \cap B)$

(1)

(iv) $P(R|H)$

(2)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 5 continued

(c) Describe a person who satisfies $H' \cap W' \cap R' \cap B'$

(1)

The game is played as follows:

- The cards are placed face-up on the table.
- Player Y selects a card at random without telling Player X what it is, and writes down the name of the person on the card on a piece of paper.
- Player X deduces which card Player Y selected by asking questions that require only yes/no answers.
- The object of the game is for Player X to identify the correct card with the fewest number of questions.

Statistically, the most efficient tactic for Player X is to ask a question that gets closest to dividing the cards in half with yes and no answers.

(d) Suggest a question based on **one** of the events H, W, R, or B, for Player X to ask to most effectively divide the cards.

(1)

Question 5 continued

(e) Suggest a new question that takes into account more than one of the events, which would be even more effective at dividing the cards than the question suggested in part (d).

(2)

(Total for Question 5 is 14 marks)

6 A certain model of concrete mixer lorry has a full capacity of approximately $6.12\,m^3$

In practice, however, a certain amount of air is contained inside the lorry.

A large building company has calculated that the volume, $X m^3$, of concrete delivered by a single lorry is approximately normally distributed with the following summary statistics:

Mean	5.67
Standard deviation	0.12

You may assume independence between such volumes of concrete throughout this question.

(a) Find the probability that:

(i) the volume of concrete delivered by a single lorry is more than $5.75\,m^3$

(1)

(ii) the total volume of concrete delivered by **two** lorries is less than $11\,m^3$

(3)

Question 6 continued

T = total volume of concrete delivered (in m^3) by k lorries.

$$T = X_1 + X_2 + \ldots + X_k$$

(b) Specify fully the distribution of T.

(2)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 6 continued

The construction of the Three Gorges Dam in China required 27.2 million m^3 of concrete.

Li is in charge of planning for the construction of a similar-sized dam. She needs to estimate the total number, k, of lorry loads of concrete she must have delivered during construction in order to be **99% sure** she has enough to build the dam.

She makes the rough calculations shown in **Figure 5** on a pad of paper.

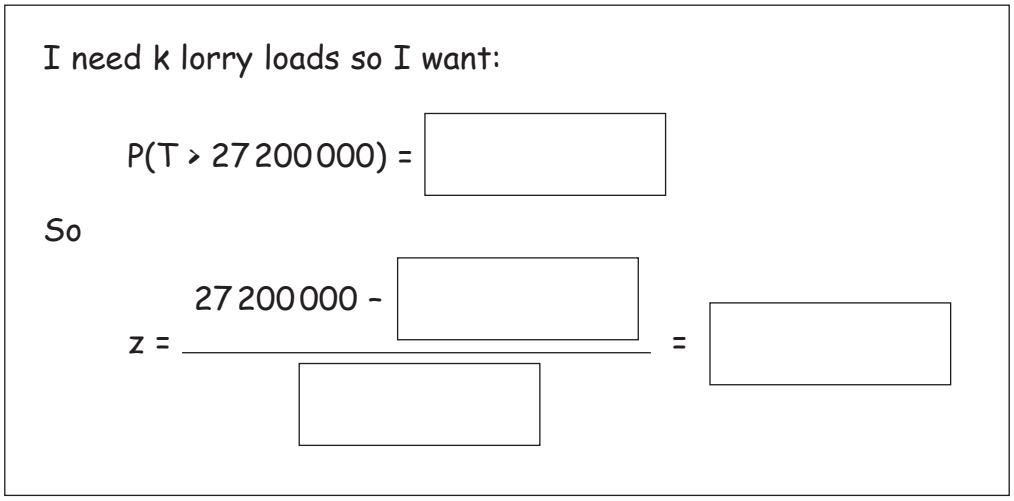

Figure 5: Li's calculations and working (part 1)

(c) Using your answer to (b), fill in the blanks in Li's working in **Figure 5**.

(3)

Question 6 continued

Li's working continues in **Figure 6**.

Using equation solver on calculator,

$\sqrt{k} = 2190.270752$

Figure 6: Li's calculations and working (part 2)

(d) How many concrete mixer lorry loads should Li plan to purchase so that she is **99% sure** she will have enough to build the dam?

(2)

(Total for Question 6 is 11 marks)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

7 Wilf is a town planner investigating the feasibility of building a new village in Oxfordshire. The village will not have its own shops, so he wants to model the potential shopping patterns of new residents.

He researches the surrounding towns and cities, and decides that the towns and cities shown in **Figure 7** are the most likely shopping destinations.

Name of town/city	Description	Size (population, 2011 census)	Distance from new village site (km)
Faringdon	Small town	7 121	5.1
Wantage	Small town	11 327	16.3
Witney	Medium town	27 522	17.5
Swindon	Large town	182 441	24.6
Oxford	City	150 200	25.7

Figure 7

He uses a version of the 'gravity model of human interaction' to estimate the proportion of the population's shopping trips to a specific location (town/city).

His model uses the following formula:

$$P(T) = \frac{1}{1000} \times \frac{S}{D^2}$$

Where:
T = Town/city
S = Size of town/city (population)
D = Distance to the town/city (km)

Wilf's calculations have been started in **Figure 8**.

Town/city (T)	Size (S)	Distance (D)	Estimated proportion of shopping trips to the town/city (P(T)) [to 4 decimal places]
Faringdon	7 121	5.1	0.2738
Wantage	11 327	16.3	0.0426
Witney	27 522	17.5	0.0899
Swindon	182 441	24.6	
Oxford	150 200	25.7	

Figure 8

(a) Complete the table in **Figure 8**.

(2)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 7 continued

(b) Estimate the proportion of the population's shopping trips that will be spent in a small or medium town.

(1)

By design, the estimated proportions in **Figure 8** add up to a total less than 1

(c) Give a possible explanation for this.

(1)

The proposed village has 300 houses planned.

The UK average number of people per household is 2.4

The UK average number of total household shopping trips per week is 1.9

(d) Making any necessary assumptions, estimate how many more shoppers Faringdon should expect per week if the village is built.

You should clearly explain or show how you obtained your estimate.

(3)

Question 7 continued

At a later stage, Wilf attempts to refine his model by investigating towns and cities further afield. He stops when he discovers the following:

Name of town/city	Description	Size (population, 2011 census)	Distance from new village site (km)
London	City	9 787 426	119.5

(e) Give **two** reasons why Wilf's model may be unsuitable if London is included.

You should provide numerical evidence to support your reasons where necessary.

(3)

(Total for Question 7 is 10 marks)

8 In the 1980s, a man was put on trial as a DNA match had been found between his hair and a hair found at the crime scene. There was no other evidence associating him with the crime.

Throughout this question, you may assume that the following information is valid.

- The probability of the man's guilt is estimated at $10^{-6} = 0.000\,001$ (when DNA evidence has not been considered)
- The probability of a DNA match between the hair at the crime scene and a randomly chosen **innocent** male adult is approximately $\frac{1}{7000}$
- A guilty man is guaranteed to have a DNA match.

An 'expert witness' stated in the court that the probability of the man's guilt, given that the DNA was a match, was $\frac{6999}{7000}$

Consider the following events:

G = The man was guilty
D = There was a DNA match

You should use $P(G) = 10^{-6} = 0.000\,001$

(a) Use Bayes' theorem or an appropriate diagram to show that the expert witness's conclusion is incorrect.

(6)

Question 8 continued

Question 8 continued

(b) In light of this finding, what recommendations would you give to the justice system with regard to the use of DNA matching evidence?

(2)

(Total for Question 8 is 8 marks)

9 If a foul is committed in basketball, the player who had the foul committed against them is usually awarded **two** 'free throws'. A point is scored from a 'free throw' if the player successfully throws the ball into the basket.

Aoife is a keen basketball player. Last year she achieved the statistics given in **Figure 9** when awarded two 'free throws'.

		Second free throw	
		Scored	**Did not score**
First free throw	**Scored**	78	39
	Did not score	22	12

Figure 9: Aoife's free throw statistics last year

- Professional basketball players should score from at least 70% of 'free throws'.
- For professional basketball players, the outcome of their second 'free throw' is approximately independent of the outcome of their first 'free throw'.

Determine whether the 'free throw' statistics in **Figure 9** suggest that Aoife is taking 'free throws' at a professional level.

You should support your argument with appropriate numerical evidence.

You should use your knowledge of probability theory in support of your answer.

You should **not** use a hypothesis test.

(9)

Question 9 continued

(Total for Question 9 is 9 marks)

TOTAL FOR PAPER IS 80 MARKS

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

BLANK PAGE