Pearson Edexcel Level 3 GCE

Tuesday 20 June 2023

Afternoon   Paper reference **9MA0/31**

Mathematics
Advanced
PAPER 31: Statistics

You must have:
Mathematical Formulae and Statistical Tables (Green), calculator

Total Marks

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.
- Values from statistical tables should be quoted in full. If a calculator is used instead of tables the value should be given to an equivalent degree of accuracy.
- Inexact answers should be given to three significant figures unless otherwise stated.

Information

- A booklet 'Mathematical Formulae and Statistical Tables' is provided.
- The total mark for this part of the examination is 50. There are 6 questions.
- 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.

Turn over ▶

1. The Venn diagram, where p and q are probabilities, shows the three events A, B and C and their associated probabilities.

(a) Find P(A)

(1)
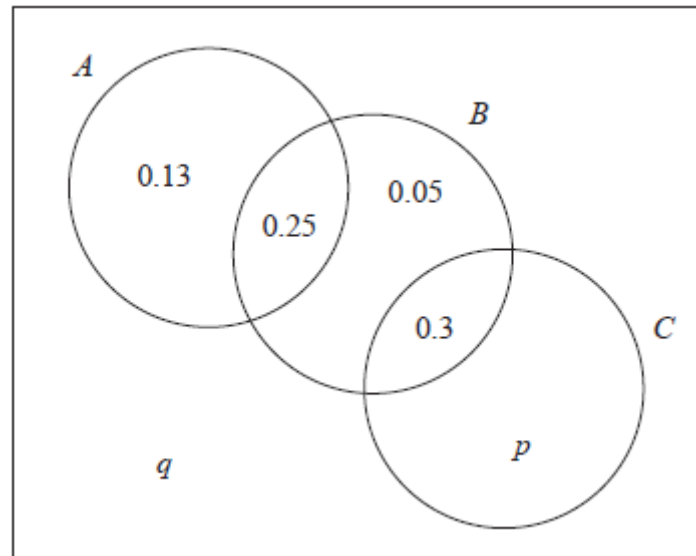
The events B and C are independent.

(b) Find the value of p and the value of q

(3)

(c) Find P($A|B'$)

(2)

(Total for Question 1 is 6 marks)

1. The Venn diagram, where p and q are probabilities, shows the three events A, B and C and their associated probabilities.
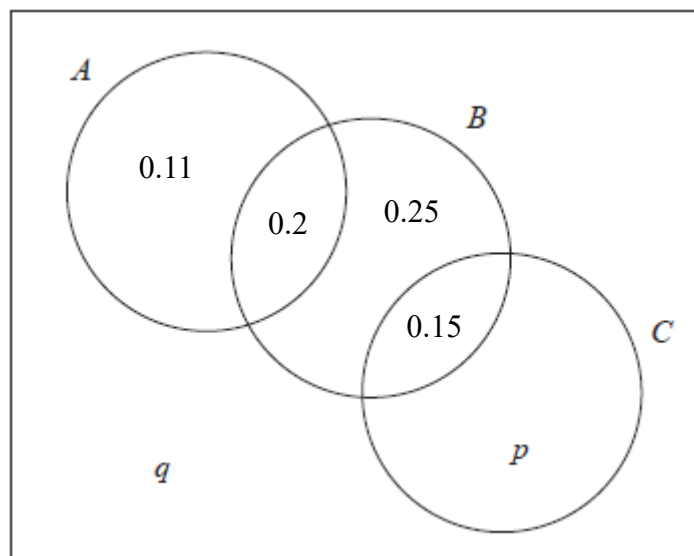

(a) Find P(A)

(1)

The events B and C are independent.

(b) Find the value of p and the value of q

(3)

(c) Find P($B|A'$)

(2)

(Total for Question 1 is 6 marks)

2. A machine fills packets with sweets and $\frac{1}{7}$ of the packets also contain a prize.

The packets of sweets are placed in boxes before being delivered to shops.
There are 40 packets of sweets in each box.

The random variable T represents the number of packets of sweets that contain a prize in each box.

(*a*) State a condition needed for T to be modelled by $B(40, \frac{1}{7})$

(1)

A box is selected at random.

(*b*) Using $T \sim B(40, \frac{1}{7})$ find

    (i)   the probability that the box has exactly 6 packets containing a prize,

    (ii) the probability that the box has fewer than 3 packets containing a prize.

(2)

Kamil's sweet shop buys 5 boxes of these sweets.

(*c*) Find the probability that exactly 2 of these 5 boxes have fewer than 3 packets containing a prize.

(2)

Kamil claims that the proportion of packets containing a prize is less than $\frac{1}{7}$

A random sample of 110 packets is taken and 9 packets contain a prize.

(*d*) Use a suitable test to assess Kamil's claim.
You should

- state your hypotheses clearly
- use a 5% level of significance

(4)

(Total for Question 2 is 9 marks)

2. A machine fills packets with sweets and $\frac{1}{9}$ of the packets also contain a prize.

The packets of sweets are placed in boxes before being delivered to shops.
There are 60 packets of sweets in each box.

The random variable T represents the number of packets of sweets that contain a prize in each box.

(a) State a condition needed for T to be modelled by $\mathrm{B}\left(60,\frac{1}{9}\right)$

(1)

A box is selected at random.

(b) Using $T \sim \mathrm{B}\left(60,\frac{1}{9}\right)$ find

    (i) the probability that the box has exactly 4 packets containing a prize,

    (ii) the probability that the box has fewer than 6 packets containing a prize.

(2)

Vineet's sweet shop buys 7 boxes of these sweets.

(c) Find the probability that exactly 3 of these 7 boxes have fewer than 6 packets containing a prize.

(2)

Vineet claims that the proportion of packets containing a prize is less than $\frac{1}{9}$

A random sample of 200 packets is taken and 17 packets contain a prize.

(d) Use a suitable test to assess Vineet's claim.
    You should

    •   state your hypotheses clearly

    •   use a 10% level of significance

(4)

(Total for Question 2 is 9 marks)

3. Ben is studying the Daily Total Rainfall, x mm, in Leeming for 1987

He used all the data from the large data set and summarised the information in the following table.

x	0	0.1–0.5	0.6–1.0	1.1–1.9	2.0–4.0	4.1–6.9	7.0–12.0	12.1–20.9	21.0–32.0	tr
Frequency	55	18	18	21	17	9	9	6	2	29

(*a*) Explain how the data will need to be cleaned before Ben can start to calculate statistics such as the mean and standard deviation.

(2)

Using all 184 of these values, Ben estimates $\sum x = 390$ and $\sum x^2 = 4336$

(*b*) Calculate estimates for

    (i)   the mean Daily Total Rainfall,

    (ii) the standard deviation of the Daily Total Rainfall.

(3)

Ben suggests using the statistic calculated in part (b)(i) to estimate the annual mean Daily Total Rainfall in Leeming for 1987

(*c*) Using your knowledge of the large data set,

    (i)   give a reason why these data would not be suitable,

    (ii) state, giving a reason, how you would expect the estimate in part (b)(i) to differ from the actual annual mean Daily Total Rainfall in Leeming for 1987

(2)

(Total for Question 3 is 7 marks)

3. Mike is studying the Daily Total Rainfall, x mm, in Leuchars for 1987

He used all the data from the large data set and summarised the information in the following table.

x	0	0.1–0.5	0.6–1.0	1.1–1.9	2.0–4.0	4.1–6.9	7.0–12.0	12.1–20.9	21.0–32.0	tr
Frequency	42	35	12	9	21	19	9	5	2	30

(a) Explain how the data will need to be cleaned before Mike can start to calculate statistics such as the mean and standard deviation.

(2)

Using all 184 of these values, Mike estimates $\sum x = 423$ and $\sum x^2 = 4373$

(b) Calculate estimates for

   (i) the mean Daily Total Rainfall,

   (ii) the standard deviation of the Daily Total Rainfall.

(3)

Mike suggests using the statistic calculated in part (b)(i) to estimate the annual mean Daily Total Rainfall in Leuchars for 1987

(c) Using your knowledge of the large data set,

   (i) give a reason why these data would not be suitable,

   (ii) state, giving a reason, how you would expect the estimate in part (b)(i) to differ from the actual annual mean Daily Total Rainfall in Leuchars for 1987

(2)

(Total for Question 3 is 7 marks)

4. A study was made of adult men from region *A* of a country.
It was found that their heights were normally distributed with a mean of 175.4 cm and standard deviation 6.8 cm.

(*a*) Find the proportion of these men that are taller than 180 cm.

(1)

A student claimed that the mean height of adult men from region *B* of this country was different from the mean height of adult men from region *A*.

A random sample of 52 adult men from region *B* had a mean height of 177.2 cm

The student assumed that the standard deviation of heights of adult men was 6.8 cm both for region *A* and region *B*.

(*b*) Use a suitable test to assess the student's claim.
   You should

   • state your hypotheses clearly

   • use a 5% level of significance

(4)

(*c*) Find the *p*-value for the test in part (*b*)

(1)

(Total for Question 4 is 6 marks)

4. A study was made of adult women from region *A* of a country.
It was found that their heights were normally distributed with a mean of 155.2 cm and standard deviation 4.2 cm.

(a) Find the proportion of these men that are taller than 158 cm.

(1)

A student claimed that the mean height of adult women from region *B* of this country was different from the mean height of adult women from region *A*.

A random sample of 60 adult women from region *B* had a mean height of 156.3 cm

The student assumed that the standard deviation of heights of adult women was 4.2 cm both for region *A* and region *B*.

(b) Use a suitable test to assess the student's claim.
  You should

  - state your hypotheses clearly

  - use a 5% level of significance

(4)

(c) Find the *p*-value for the test in part (b)

(1)

(Total for Question 4 is 6 marks)

5. Tisam is playing a game.
She uses a ball, a cup and a spinner.

The random variable X represents the number the spinner lands on when it is spun.
The probability distribution of X is given in the following table

x	20	50	80	100
P($X = x$)	a	b	c	d

where a, b, c and d are probabilities.

To play the game

- the spinner is spun to obtain a value of x

- Tisam then stands x cm from the cup and tries to throw the ball into the cup

The event S represents the event that Tisam successfully throws the ball into the cup.

To model this game Tisam assumes that

- $P(S|\{X=x\}) = \dfrac{k}{x}$ where k is a constant

- $P(S \cap \{X = x\})$ should be the same whatever value of x is obtained from the spinner

Using Tisam's model,

(a) show that $c = \dfrac{8}{5}b$

(2)

(b) find the probability distribution of X

(5)

Nav tries, a large number of times, to throw the ball into the cup from a distance
of 100 cm.
He successfully gets the ball in the cup 30% of the time.

(c) State, giving a reason, why Tisam's model of this game is not suitable to describe
Nav playing the game for all values of X

(1)

(Total for Question 5 is 8 marks)

5. Elina is playing a game.
She uses a ball, a cup and a spinner.

The random variable X represents the number the spinner lands on when it is spun.
The probability distribution of X is given in the following table

x	10	40	70	100
P($X = x$)	a	b	c	d

where a, b, c and d are probabilities.

To play the game

- the spinner is spun to obtain a value of x

- Elina then stands x cm from the cup and tries to throw the ball into the cup

The event S represents the event that Elina successfully throws the ball into the cup.

To model this game Elina assumes that

- $P(S|\{X=x\}) = \dfrac{k}{x}$ where k is a constant

- $P(S \cap \{X=x\})$ should be the same whatever value of x is obtained from the spinner

Using Elina's model,

(a) show that $d = \dfrac{10}{7}c$

(2)

(b) find the probability distribution of X

(5)

Matt tries, a large number of times, to throw the ball into the cup from a distance of 70 cm.
He successfully gets the ball in the cup 25% of the time.

(c) State, giving a reason, why Elina's model of this game is not suitable to describe Matt playing the game for all values of X

(1)

(Total for Question 5 is 8 marks)

6. A medical researcher is studying the number of hours, T, a patient stays in hospital following a particular operation.

The histogram on the next page summarises the results for a random sample of 90 patients.

(a) Use the histogram to estimate $P(10 < T < 30)$

(2)

For these 90 patients the time spent in hospital following the operation had

- a mean of 14.9 hours

- a standard deviation of 9.3 hours

Tomas suggests that T can be modelled by $N(14.9, 9.3^2)$

(b) With reference to the histogram, state, giving a reason, whether or not Tomas' model could be suitable.

(1)

Xiang suggests that the frequency polygon based on this histogram could be modelled by a curve with equation

$$y = kxe^{-x} \quad 0 \le x \le 4$$

where

- x is measured in **tens of hours**

- k is a constant

(c) Use algebraic integration to show that

$$\int_0^n xe^{-x}\,dx = 1 - (n + 1)e^{-n}$$

(4)

(d) Show that, for Xiang's model, $k = 99$ to the nearest integer.

(3)

(e) Estimate $P(10 < T < 30)$ using

   (i) Tomas' model of $T \sim N(14.9, 9.3^2)$

(1)

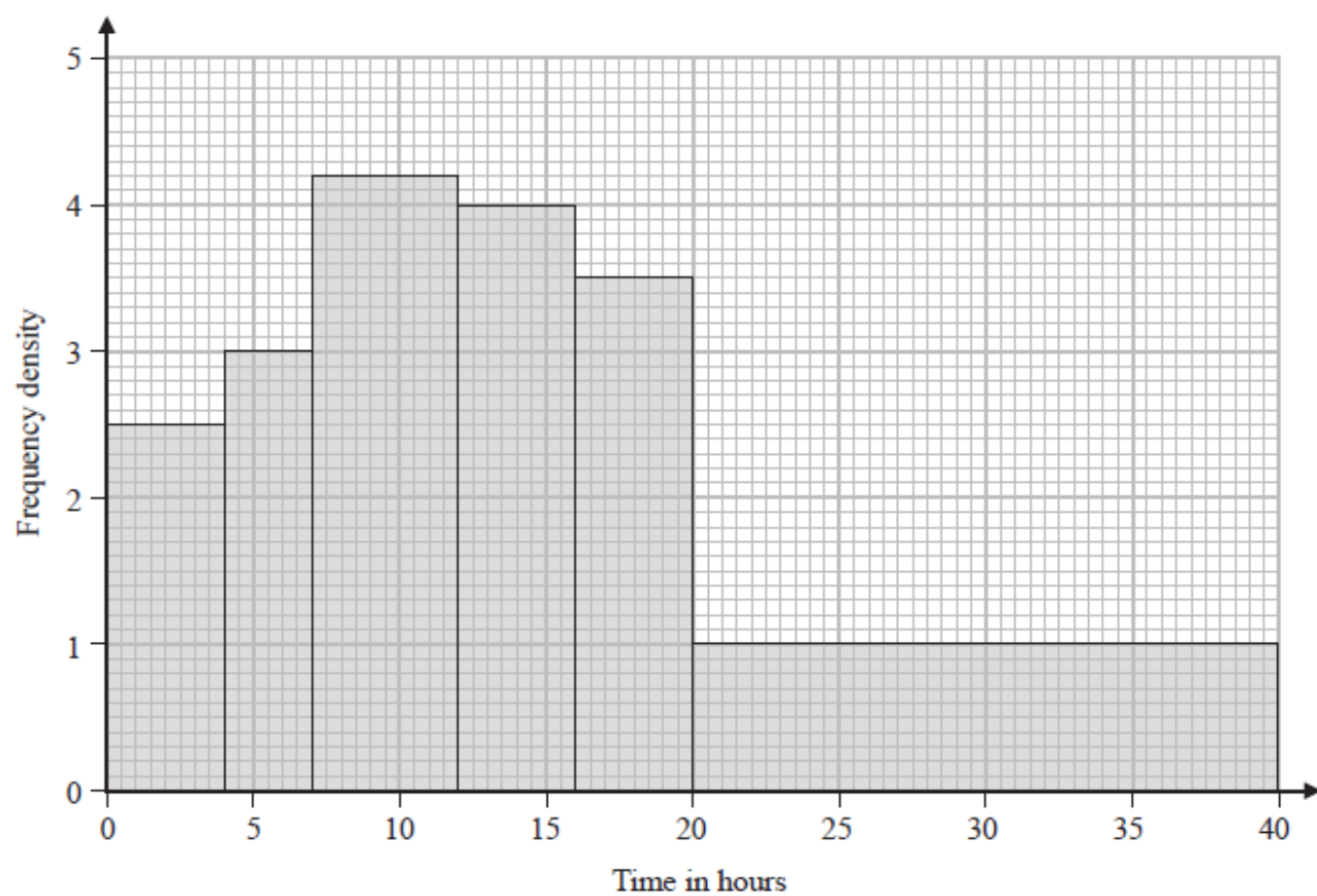   (ii) Xiang's curve with equation $y = 99xe^{-x}$ and the answer to part (c)

(2)

The researcher decides to use Xiang's curve to model $P(a < T < b)$

(f) State one limitation of Xiang's model.

(1)

(Total for Question 6 is 14 marks)

6. A medical researcher is studying the number of hours, T, a patient stays in hospital following a particular operation.

The histogram on the next page summarises the results for a random sample of 100 patients.

(a) Use the histogram to estimate $P(20 < T < 40)$

(2)

For these 100 patients the time spent in hospital following the operation had

- a mean of 32.9 hours

- a standard deviation of 19.8 hours

Tim suggests that T can be modelled by $N(32.9, 19.8^2)$

(b) With reference to the histogram, state, giving a reason, whether or not Tims' model could be suitable.

(1)

Xi suggests that the frequency polygon based on this histogram could be modelled by a curve with equation

$$y = k(x+1)e^{-x+1} \qquad 0 \le x \le 8$$

where

- x is measured in **tens of hours**

- k is a constant

(c) Use algebraic integration to show that

$$\int_0^n (x+1)e^{-x+1}dx = 2e - (n+2)e^{-n+1}$$

(4)

(d) Show that, for Xi's model, $k = 18$ to the nearest integer.

(3)

(e) Estimate $P(20 < T < 40)$ using

(i) Tims' model of $T \sim N(32.9, 19.8^2)$

(1)

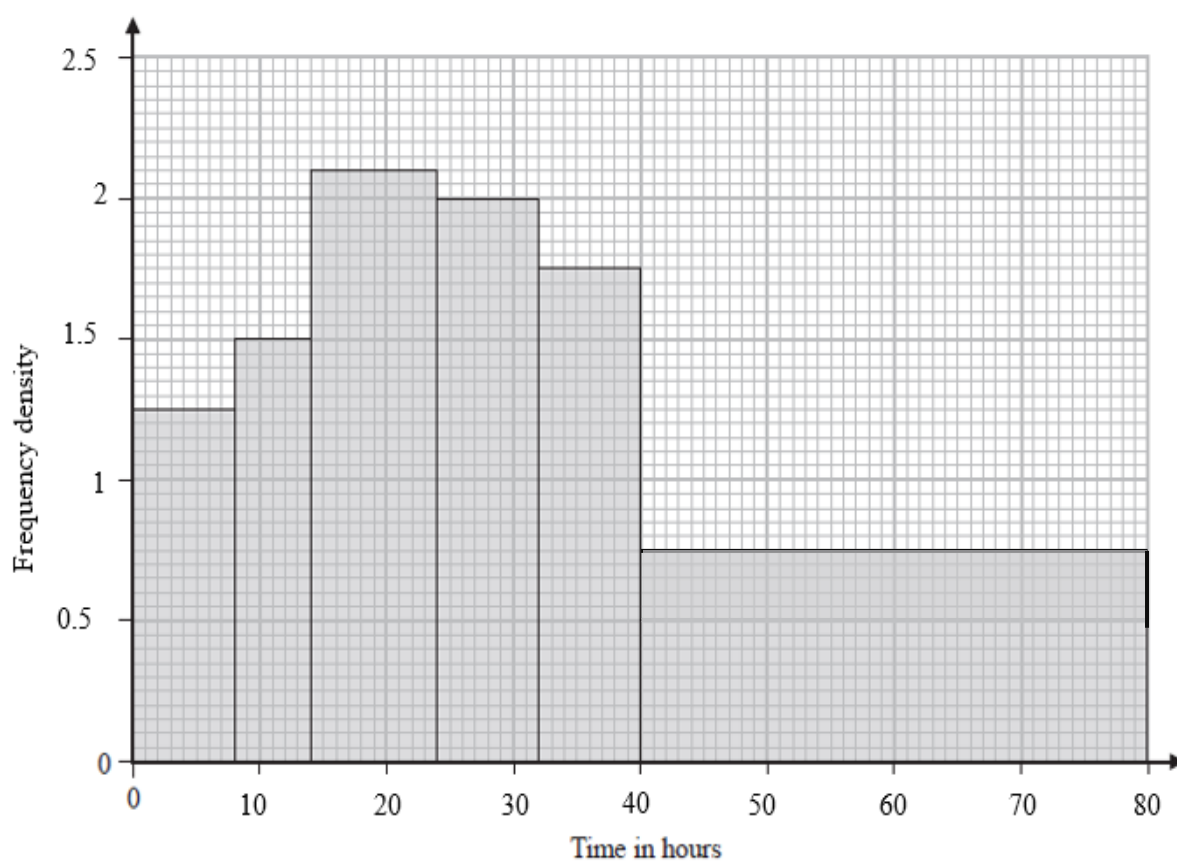(ii) Xi's curve with equation $y = 18(x+1)e^{-x+1}$ and the answer to part (c)

(2)

The researcher decides to use Xi's curve to model $P(a < T < b)$

(f) State one limitation of Xi's model.

(1)

(Total for Question 6 is 14 marks)