Please check the examination details below before entering your candidate information

Candidate surname	Other names

Centre Number | Candidate Number

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

Candidates may use any calculator allowed by Pearson regulations. Calculators must not have the facility for symbolic algebra manipulation, differentiation and integration, or 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.
- 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

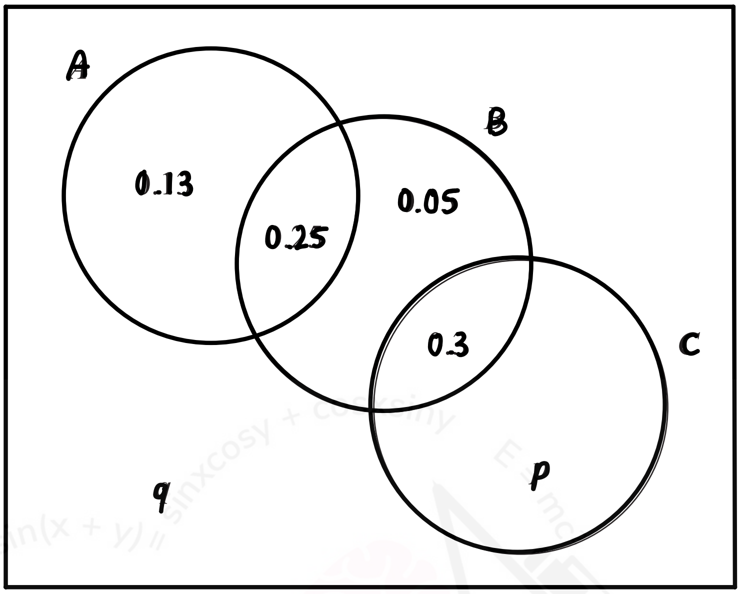
P72819A

N:1/1/1/

P 7 2 8 1 9 A 0 1 2 0

Pearson

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.

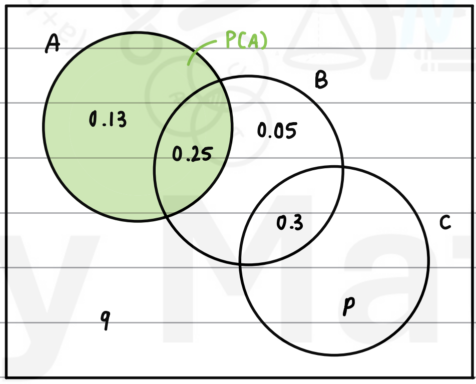
(b) Find the value of p and the value of q

(3)

(c) Find $\mathrm{P}(A \mid B')$

(2)

(a)



P(A): "everything included in set A"

P(A) = 0.13 + 0.25 = 0.38

∴ P(A) = 0.38 B1

(b) "B and C are independent"

Formula for independent events:

$P(A) \times P(B) = P(A \cap B)$

$P(B) = 0.6$

$P(C) = 0.3 + p$

$P(B \cap C) = 0.3$

Substitute: $P(B) \times P(C) = 0.6(0.3 + p)$

$= 0.18 + 0.6p$ M1

$P(B \cap C) = 0.3$

$\therefore 0.18 + 0.6p = 0.3$

$0.6p = 0.12$

$p = 0.2$ value of p A1

Question 1 continued

$\sum$ probabilities $=1$

$0.13 + 0.25 + 0.05 + 0.3 + p + q = 1$

$0.73 + 0.2 + q = 1$

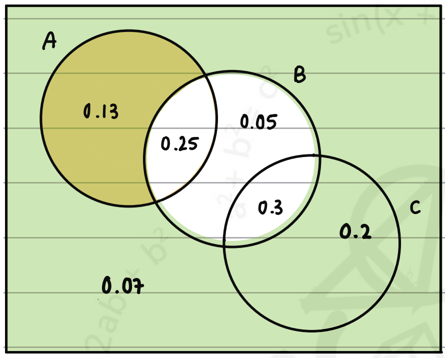
$q = 0.07$ value of q B1

(c) $P(A|B')$ "given that"

Formula for "given that":

$P(A|B) = \frac{P(A\cap B)}{P(B)}$

$\therefore$ For our case: $P(A|B') = \frac{P(A\cap B')}{P(B')}$ M1



$P(B')$ is everything but B → shaded in green

$P(A\cap B')$ everything that's both A and not B

↳ shaded in orange

$P(B') = 0.13 + 0.2 + 0.07 = 0.4$

$P(A\cap B') = 0.13$

Substitute → $P(A|B') = \frac{0.13}{0.4} = \frac{13}{40}$ A1

(Total for Question 1 is 6 marks)

constant probability

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. set number of trials

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)

(a) → Prizes must be placed in packets independantly of eachother
→ The probability that a packet contains a prize must be constant B1

(b) $T \sim B(40, \frac{1}{7})$
i. "exactly" → equal to
$P(X=6) = 0.17273 \longrightarrow$ 0.173 to 3sf B1
ii. "fewer than" → $<$
$P(X<3) = P(X\leq 2) = 0.061587 \longrightarrow$ 0.0616 to 3sf B1

DO NOT WRITE IN THIS AREA

Question 2 continued

(c) $K \rightarrow$ number of boxes with less than 3 packets with a prize   define variable

$K \sim B(5, P(X<3)) \longrightarrow K \sim B(5, 0.0616)$ M1

he bought 5 boxes   probability that less than 3 packets have a prize (from part (b)ii.)

$P(K=2) = 0.031344 \longrightarrow 0.0313$ to 3sf A1

"exactly 2"

(d) Hypotheses $X \sim B(110, \frac{1}{7})$ M1 A1

$H_0: p = \frac{1}{7}$ $P(X \leq 9) = 0.038292 < 0.05 \therefore$ falls in critical region so there is sufficient

$H_1: p < \frac{1}{7}$ "9 samples contain a prize"   evidence to reject H_0. There is evidence

B1   to support Kamil's claim A1

Question 2 continued

Question 2 continued

(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 $\sum x^2 = 4336$

n

(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)

(a) We need to replace "tr" (trace) with a numerical value, e.g. 0.025 (very small amount) M1A1

(b) i. Formula for mean:

$\bar{x} = \frac{\sum x}{n}$ $\bar{x} = \frac{390}{184} = 2.119 \rightarrow$ 2.12 to 3sf. B1

ii. Formula for SD:

$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$ $\sigma = \sqrt{\frac{4336}{184} - \left(\frac{390}{184}\right)^2} = 4.367 \rightarrow$ 4.37 to 3sf A1

M1

(c) i. The large data set only covers May-October B1

ii. The estimate would be lower than the actual since the winter months (which would have more rain) are missing. B1

Question 3 continued

(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.

μ σ

(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.

two-tailed!

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)

(a) $A \rightarrow$ height from region A   define variable

$A \sim N(175.4, 6.8^2)$

$P(A > 180) = 0.24937 \longrightarrow \boxed{0.249 \text{ to } 3sf}$

"taller than 180cm"   B1

(b) $B \rightarrow$ height from region B   define variable

$B \sim N(175.4, 6.8^2)$

This part talks about "mean" ∴ we will use $\bar{B}$ (sample mean) variable

Formula for Sample mean:

$X \sim N(\mu, \sigma^2) \longrightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

enter $\frac{6.8}{\sqrt{52}} = \sigma$ into your calculator!

Apply the formula.

$B \sim N(175.4, 6.8^2) \longrightarrow \bar{B} \sim N(175.4, \frac{6.8^2}{52})$ M1

found value   significance level / 2 → two-tailed test! two tails, 0.025 each → 0.05 total!

Hypotheses

$H_0: \mu = 175.4$ B1

$H_1: \mu \neq 175.4$

two tailed test.

$P(\bar{B} > 177.2) = 0.02814 > 0.025$ ∴ does not fall in the critical region and there is insufficient evidence to reject H_0.

When you are given a value to test within a two-tailed test, you compare it to H_0 to figure out which tail to test. If it's $> H_0$, test upper tail, if $< H_0$ test lower tail!   A1

The claim is $\boxed{\text{not supported}}$ A1

Question 4 continued

(c) "p-value" is the actual significance level.

We need to multiply what we got for the upper tail in (b) by 2:

$2 \times 0.02814 = 0.05628\ldots$

Convert that into a percentage! $p = 5.6\%$ B1

(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
$\mathrm{P}(X = x)$	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

- $\mathrm{P}(S \mid \{X = x\}) = \dfrac{k}{x}$ where k is a constant
- $\mathrm{P}(S \cap \{X = x\})$ should be the same whatever value of x is obtained from the spinner

this constant for all x's.

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. $p = 0.3$

(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)

(a) We can use the given "$P(S \cap \{X = x\})$ is constant".

This means: $P(S \cap \{X=50\}) = P(S \cap \{X=80\}) = Q$ – this is just a constant so you can set up simultaneous equations!

Use given "$P(S \mid \{X = x\}) = \frac{k}{x}$" to get $P(S)$ in both cases:

"and" and

$P(S \cap \{X = 50\}) = \frac{k}{50} \times b = Q_1$

$P(S \mid \{X=50\})$ $P(X=50)$

$P(S \cap \{X = 80\}) = \frac{k}{80} \times c = Q_2$ M1

Equate Q_1 and Q_2: $\frac{k}{80} \times c = \frac{k}{50} \times b$ cancel k's

$c = \frac{80}{50} b \longrightarrow \therefore c = \frac{8}{5} b$ hence shown A1

P 7 2 8 1 9 A 0 1 2 2 0

Question 5 continued

(b) to get the complete probability distribution of X we need to get a, b, c and d.

Use the same method as in (a) to get equations that all in terms of b.

for a: $P(S \cap \{X=20\}) = P(S \cap \{X=80\})$

$\frac{\cancel{k}}{20} \times a = \frac{\cancel{k}}{50} \times b$ $a = \frac{2}{5}b$ M1A1

for c: $c = \frac{8}{5}b$ (from (a))

for d: $P(S \cap \{X=100\}) = P(S \cap \{X=80\})$

$\frac{\cancel{k}}{100} \times d = \frac{\cancel{k}}{50} \times b$ $d = 2b$

We know that $\sum$ probabilities $= 1 \therefore a+b+c+d = 1$

Substitute the equations in terms of b we got for a, c and d:

$\frac{2}{5}b + b + \frac{8}{5}b + 2b = 1 \longrightarrow 5b = 1 \therefore b = \frac{1}{5}$ M1A1

Now substitute b back in:

$a = \frac{2}{5}\left(\frac{1}{5}\right) = \frac{2}{25}$

$c = \frac{8}{5}\left(\frac{1}{5}\right) = \frac{8}{25}$

$d = 2\left(\frac{1}{5}\right) = \frac{2}{5}$ A1

$\therefore$ Probability distribution X:

x	20	50	80	100
$P(X=x)$	$\frac{2}{25}$	$\frac{1}{5}$	$\frac{8}{25}$	$\frac{2}{5}$

(c) From what we're told, we understand that

$P(S \mid \{X=100\}) = 0.3$ for Nav's experiment

so as $P(S \mid \{X=x\}) = \frac{k}{x}$, $0.3 = \frac{k}{100} \longrightarrow k = 30$

With $k=30 \longrightarrow P(S \mid \{X=20\}) = \frac{30}{20} \longrightarrow p > 1$ which is not possible as probabilities are always < 1.

So the model won't work. B1

Question 5 continued

Question 5 continued

(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 page opposite summarises the results for a random sample of 90 patients.

(a) Use the histogram to estimate $\mathrm{P}(10 < T < 30)$

(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

normal

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

(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

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

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 $\mathrm{P}(10 < T < 30)$ using

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

(1)

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

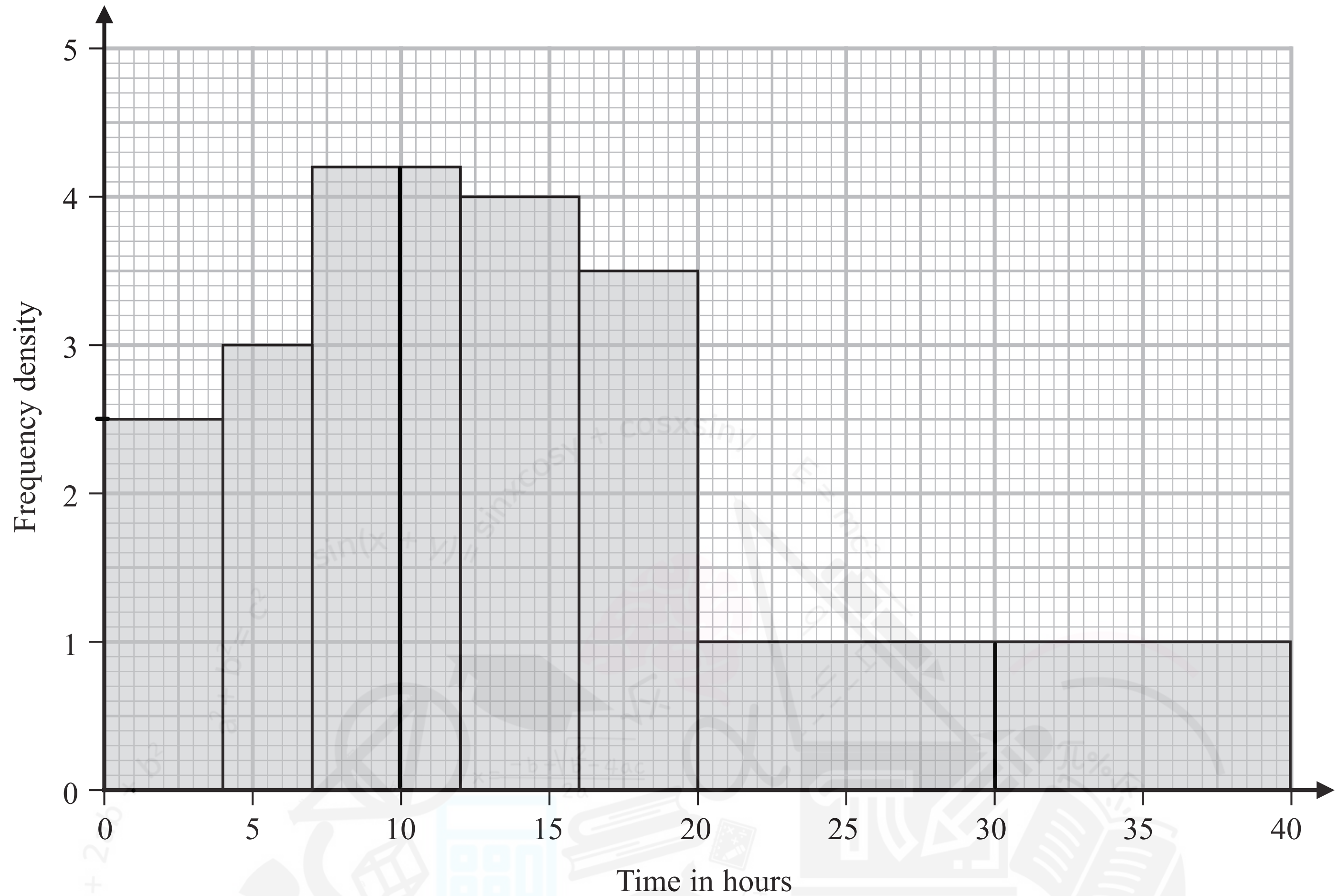
(2)

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

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

(1)

Question 6 continued



(a) We need to get the total area of the bars.

$0 \to 4: \quad 4 \times 2.5 = 10$

$4 \to 7: \quad 3 \times 3 = 9$

$7 \to 11: \quad 5 \times 4.2 = 21$

$11 \to 16: \quad 4 \times 4 = 16$

$16 \to 20: \quad 4 \times 3.5 = 14$

$20 \to 40: \quad 20 \times 1 = 20$

$\therefore$ total area:

$10 + 9 + 21 + 16 + 14 + 20 = 90$

"random sample of 90 patients" $\therefore \frac{90}{90} = 1$

Now let's find the area $P(10 < T < 30)$:

$10 \to 11: \quad 4.2 \times 2 = 8.4$

from above $11 \to 20: \quad 30$

$20 \to 30: \quad 10 \times 1 = 10$

$\therefore$ area: 48.1 units. M1

hence $P(10 < T < 30) = \frac{48.1}{90} = \frac{121}{225} = 0.53777\ldots \longrightarrow$ 0.538 to 3sf A1

area $10<T<30$

total

(b) It's not suitable as the data are <u>not</u> symmetric. (Remember the normal distribution is a symmetric bell-shaped curve) B1

P 7 2 8 1 9 A 0 1 7 2 0

Question 6 continued

(c)

$$\int_0^n xe^{-x}\,dx$$

we see this is a multiplication, $\therefore$ we have to use Integration by Parts.

★ Integration by Parts: $\int vu'dx = vu - \int uv'dx$

OR

SDI - table

Method 1 Formula

$u = -e^{-x} \leftarrow \int u'dx - u' = e^{-x}$

$v = x - \frac{dv}{dx} \rightarrow v' = 1$

$$\int_0^n xe^{-x}\,dx$$

when you see a linear term like x, set it as your v!

$= x(-e^{-x}) - \int (1)(-e^{-x})dx$ M1A1

$= -xe^{-x} + \int e^{-x}dx$

$= \left[-xe^{-x} - e^{-x}\right]_0^n$

Substitute limits: $= (-ne^{-n} - e^{-n}) - (-\cancel{0e^{-0}} - e^{-(0)})$ dM1

$= 1 - ne^{-n} - e^{-n}$ factorize e^{-n}

$= 1 - e^{-n}(n+1)$ hence shown A1

Method 2 SDI table

Sign	Differentiate	Integrate
+	x	e^{-x}
−	1	$-e^{-x}$
+	0	e^{-x}
−		$-e^{-x}$

M1A1

$\therefore \int_0^n xe^{-x}\,dx = \left[-xe^{-x} - e^{-x}\right]_0^n$

Substitute limits: $= (-ne^{-n} - e^{-n}) - (-\cancel{0e^{-0}} - e^{-(0)})$ dM1

$= 1 - ne^{-n} - e^{-n}$ factorize e^{-n}

$= 1 - e^{-n}(n+1)$ hence shown A1

(d)

40 is the maximum as x is tens of hours, $n = \frac{40}{10} = 4$

$k\int_0^4 xe^{-x}\,dx = 90$ since the area we calculated above is 90.

M1

$k(1 - e^{-4}(4+1)) = 90$ solve for k M1

$k(1 - 5e^{-4}) = 90$

$k = \frac{90}{(1 - 5e^{-4})} = 99.0729... \longrightarrow \therefore k = 99$ shown A1

Question 6 continued

(e) i. $P(10 < T < 30) = 0.64863 \longrightarrow$ 0.649 to 3sf B1

ii. # of patients: $99\int_1^3 xe^{-x}\,dx$

$= 99\left[(1-4e^{-3})-(1-4e^{-1})\right]$ M1

$= 53.1\ldots$

Probability $= \dfrac{53.1\ldots}{90} = 0.59027 \longrightarrow$ 0.590 to 3sf A1

(f) The patients may stay longer than 40h. B1

Question 6 continued

(Total for Question 6 is 14 marks)

TOTAL FOR STATISTICS IS 50 MARKS