

APOLLO

Statistics and probability

HL

Mathematics: applications and interpretation

First assessment 2021

117 questions | questions

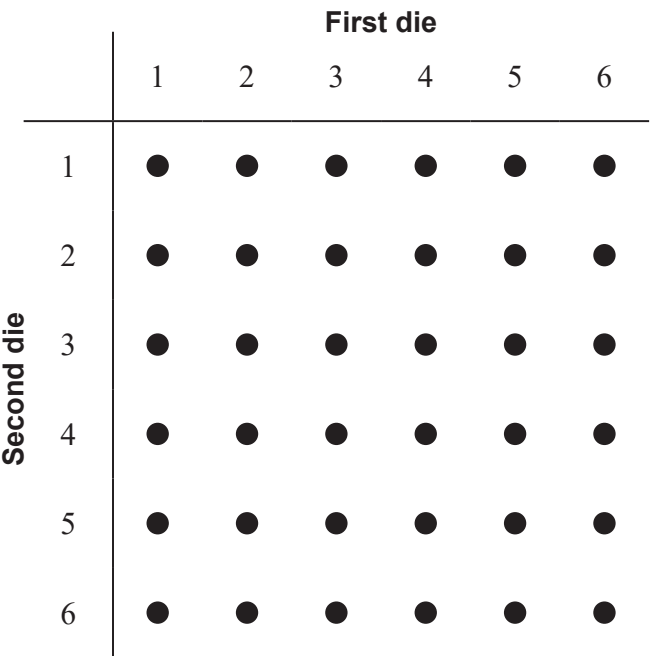
George goes fishing. From experience he knows that the mean number of fish he catches per hour is 1.1. It is assumed that the number of fish he catches can be modelled by a Poisson distribution.

On a day in which George spends 8 hours fishing, find the probability that he will catch more than 9 fish.

[illegible]

8. [Maximum mark: 7]

A game is played where two unbiased dice are rolled and the score in the game is the greater of the two numbers shown. If the two numbers are the same, then the score in the game is the number shown on one of the dice. A diagram showing the possible outcomes is given below.



Let T be the random variable “the score in a game”.

- (a) Complete the table to show the probability distribution of T . [2]

t	1	2	3	4	5	6
$P(T=t)$						

- (b) Find the probability that
- (i) a player scores at least 3 in a game.
 - (ii) a player scores 6, given that they scored at least 3. [3]
- (c) Find the expected score of a game. [2]

(This question continues on the following page).



(Question 8 continued)

[illegible]

Turn over

11. [Maximum mark: 7]

A factory, producing plastic gifts for a fast food restaurant’s Jolly meals, claims that just 1 % of the toys produced are faulty.

A restaurant manager wants to test this claim. A box of 200 toys is delivered to the restaurant. The manager checks all the toys in this box and four toys are found to be faulty.

- (a) Identify the type of sampling used by the restaurant manager. [1]

The restaurant manager performs a one-tailed hypothesis test, at the 10% significance level, to determine whether the factory’s claim is reasonable. It is known that faults in the toys occur independently.

- (b) Write down the null and alternative hypotheses. [2]
- (c) Find the p -value for the test. [2]
- (d) State the conclusion of the test. Give a reason for your answer. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



14. [Maximum mark: 6]

The weights of apples from Tony’s farm follow a normal distribution with mean 158 g and standard deviation 13 g. The apples are sold in bags that contain six apples.

- (a) Find the mean weight of a bag of apples. [2]
- (b) Find the standard deviation of the weights of these bags of apples. [2]
- (c) Find the probability that a bag selected at random weighs more than 1 kg. [2]



Turn over

3. [Maximum mark: 6]

At Springfield University, the weights, in kg, of 10 chinchilla rabbits and 10 sable rabbits were recorded. The aim was to find out whether chinchilla rabbits are generally heavier than sable rabbits. The results obtained are summarized in the following table.

Weight of chinchilla rabbits, kg	4.9	4.2	4.1	4.4	4.3	4.6	4.0	4.7	4.5	4.4
Weight of sable rabbits, kg	4.2	4.1	4.1	4.2	4.5	4.4	4.5	3.9	4.2	4.0

A t -test is to be performed at the 5 % significance level.

- (a) Write down the null and alternative hypotheses. [2]
- (b) Find the p -value for this test. [2]
- (c) Write down the conclusion to the test. Give a reason for your answer. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

9. [Maximum mark: 8]

A newspaper vendor in Singapore is trying to predict how many copies of *The Straits Times* they will sell. The vendor forms a model to predict the number of copies sold each weekday. According to this model, they expect the same number of copies will be sold each day.

To test the model, they record the number of copies sold each weekday during a particular week. This data is shown in the table.

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Number of copies sold	74	97	91	86	112

A goodness of fit test at the 5 % significance level is used on this data to determine whether the vendor’s model is suitable. The critical value for the test is 9.49.

- (a) Find an estimate for how many copies the vendor expects to sell each day. [1]
- (b) (i) State the null and alternative hypotheses for this test.
- (ii) Write down the degrees of freedom for this test.
- (iii) Write down the conclusion to the test. Give a reason for your answer. [7]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



10. [Maximum mark: 6]

A manufacturer of chocolates produces them in individual packets, claiming to have an average of 85 chocolates per packet.

Talha bought 30 of these packets in order to check the manufacturer’s claim.

Given that the number of individual chocolates is x , Talha found that, from his 30 packets, $\sum x = 2506$ and $\sum x^2 = 209\,738$.

- (a) Find an unbiased estimate for the mean number (μ) of chocolates per packet. [1]
- (b) Use the formula $s_{n-1}^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$ to determine an unbiased estimate for the variance of the number of chocolates per packet. [2]
- (c) Find a 95% confidence interval for μ . You may assume that all conditions for a confidence interval have been met. [2]
- (d) Suggest, with justification, a valid conclusion that Talha could make. [1]



Turn over

13. [Maximum mark: 8]

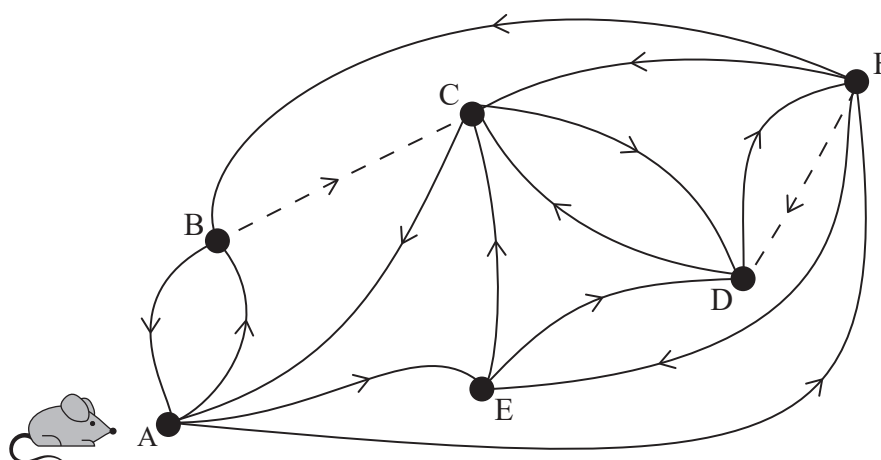
The graph below shows a small maze, in the form of a network of directed routes. The vertices A to F show junctions in the maze and the edges show the possible paths available from one vertex to another.

A mouse is placed at vertex A and left to wander the maze freely. The routes shown by dashed lines indicate paths sprinkled with sugar.

When the mouse reaches any junction, she rests for a constant time before continuing.

At any junction, it may also be assumed that

- the mouse chooses any available normal path with equal probability
- if the junction includes a path sprinkled with sugar, the probability of choosing this path is twice that of a normal path.



- (a) Determine the transition matrix for this graph. [3]
- (b) If the mouse was left to wander indefinitely, use your graphic display calculator to estimate the percentage of time that the mouse would spend at point F. [3]
- (c) Comment on your answer to part (b), referring to at least one limitation of the model. [2]

(This question continues on the following page)



(Question 13 continued)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

15. [Maximum mark: 7]

The number of coffees sold per hour at an independent coffee shop is modelled by a Poisson distribution with a mean of 22 coffees per hour.

Sheila, the shop’s owner wants to increase the number of coffees sold in the shop. She decides to offer a discount to customers who buy more than one coffee.

To test how successful this strategy is, Sheila records the number of coffees sold over a single 5-hour period. Sheila decides to use a 5% level of significance in her test.

- (a) State the null and alternative hypotheses for the test. [1]
- (b) Find the probability that Sheila will make a type I error in her test conclusion. [4]

Sheila finds 126 coffees were sold during the 5-hour period.

- (c) State Sheila’s conclusion to the test. Justify your answer. [2]



Turn over

3. [Maximum mark: 13]

The stopping distances for bicycles travelling at 20 km h^{-1} are assumed to follow a normal distribution with mean 6.76 m and standard deviation 0.12 m .

- (a) Under this assumption, find, correct to four decimal places, the probability that a bicycle chosen at random travelling at 20 km h^{-1} manages to stop

(i) in less than 6.5 m .

(ii) in more than 7 m .

[3]

1000 randomly selected bicycles are tested and their stopping distances when travelling at 20 km h^{-1} are measured.

- (b) Find, correct to four significant figures, the expected number of bicycles tested that stop between

(i) 6.5 m and 6.75 m .

(ii) 6.75 m and 7 m .

[3]

The measured stopping distances of the 1000 bicycles are given in the table.

Measured stopping distance	Number of bicycles
Less than 6.5 m	12
Between 6.5 m and 6.75 m	428
Between 6.75 m and 7 m	527
More than 7 m	33

It is decided to perform a χ^2 goodness of fit test at the 5% level of significance to decide whether the stopping distances of bicycles travelling at 20 km h^{-1} can be modelled by a normal distribution with mean 6.76 m and standard deviation 0.12 m .

- (c) State the null and alternative hypotheses.

[2]

- (d) Find the p -value for the test.

[3]

- (e) State the conclusion of the test. Give a reason for your answer.

[2]

Turn over

5. [Maximum mark: 13]

Long term experience shows that if it is sunny on a particular day in Vokram, then the probability that it will be sunny the following day is 0.8. If it is not sunny, then the probability that it will be sunny the following day is 0.3.

The transition matrix T is used to model this information, where $T = \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix}$.

(a) It is sunny today. Find the probability that it will be sunny in three days' time. [2]

(b) Find the eigenvalues and eigenvectors of T . [5]

The matrix T can be written as a product of three matrices, PDP^{-1} , where D is a diagonal matrix.

(c) (i) Write down the matrix P .

(ii) Write down the matrix D . [2]

(d) Hence find the long-term percentage of sunny days in Vokram. [4]

2. [Maximum mark: 16]

It is known that the weights of male Persian cats are normally distributed with mean 6.1 kg and variance 0.5^2 kg^2 .

(a) Sketch a diagram showing the above information. [2]

(b) Find the proportion of male Persian cats weighing between 5.5 kg and 6.5 kg. [2]

A group of 80 male Persian cats are drawn from this population.

(c) Determine the expected number of cats in this group that have a weight of less than 5.3 kg. [3]

The male cats are now joined by 80 female Persian cats. The female cats are drawn from a population whose weights are normally distributed with mean 4.5 kg and standard deviation 0.45 kg.

(d) Ten female cats are chosen at random.

(i) Find the probability that exactly one of them weighs over 4.62 kg.

(ii) Let N be the number of cats weighing over 4.62 kg.

Find the variance of N . [5]

A cat is selected at random from all 160 cats.

(e) Find the probability that the cat was female, given that its weight was over 4.7 kg. [4]

Turn over

4. [Maximum mark: 18]

In a small village there are two doctors' clinics, one owned by Doctor Black and the other owned by Doctor Green. It was noted after each year that 3.5% of Doctor Black's patients moved to Doctor Green's clinic and 5% of Doctor Green's patients moved to Doctor Black's clinic. All additional losses and gains of patients by the clinics may be ignored.

At the start of a particular year, it was noted that Doctor Black had 2100 patients on their register, compared to Doctor Green's 3500 patients.

- (a) Write down a transition matrix T indicating the annual population movement between clinics. [2]
- (b) Find a prediction for the ratio of the number of patients Doctor Black will have, compared to Doctor Green, after two years. [2]
- (c) Find a matrix P , with integer elements, such that $T = PDP^{-1}$, where D is a diagonal matrix. [6]
- (d) Hence, show that the long-term transition matrix T^∞ is given by $T^\infty = \begin{pmatrix} \frac{10}{17} & \frac{10}{17} \\ \frac{7}{17} & \frac{7}{17} \end{pmatrix}$. [6]
- (e) Hence, or otherwise, determine the expected ratio of the number of patients Doctor Black would have compared to Doctor Green in the long term. [2]

Turn over

5. [Maximum mark: 14]

Hank sets up a bird table in his garden to provide the local birds with some food. Hank notices that a specific bird, a large magpie, visits several times per month and he names him Bill. Hank models the number of times per month that Bill visits his garden as a Poisson distribution with mean 3.1.

- (a) Using Hank's model, find the probability that Bill visits the garden on exactly four occasions during one particular month. [1]
- (b) Over the course of 3 consecutive months, find the probability that Bill visits the garden:
 - (i) on exactly 12 occasions.
 - (ii) during the first and third month only. [5]
- (c) Find the probability that over a 12-month period, there will be exactly 3 months when Bill does not visit the garden. [4]

After the first year, a number of baby magpies start to visit Hank's garden. It may be assumed that each of these baby magpies visits the garden randomly and independently, and that the number of times each baby magpie visits the garden per month is modelled by a Poisson distribution with mean 2.1.

- (d) Determine the least number of magpies required, including Bill, in order that the probability of Hank's garden having at least 30 magpie visits per month is greater than 0.2. [4]

2. [Maximum mark: 28]

A firm wishes to review its recruitment processes. This question considers the validity and reliability of the methods used.

Every year an accountancy firm recruits new employees for a trial period of one year from a large group of applicants.

At the start, all applicants are interviewed and given a rating. Those with a rating of either *Excellent*, *Very good* or *Good* are recruited for the trial period. At the end of this period, some of the new employees will stay with the firm.

It is decided to test how valid the interview rating is as a way of predicting which of the new employees will stay with the firm.

Data is collected and recorded in a contingency table.

	Interview rating		
	Excellent	Very good	Good
Stay	12	20	19
Leave	10	15	24

- (a) Use an appropriate test, at the 5 % significance level, to determine whether a new employee staying with the firm is independent of their interview rating. State the null and alternative hypotheses, the p -value and the conclusion of the test.

[6]

The next year's group of applicants are asked to complete a written assessment which is then analysed. From those recruited as new employees, a random sample of size 18 is selected.

The sample is stratified by department. Of the 91 new employees recruited that year, 55 were placed in the national department and 36 in the international department.

- (b) Show that 11 employees are selected for the sample from the national department.

[2]

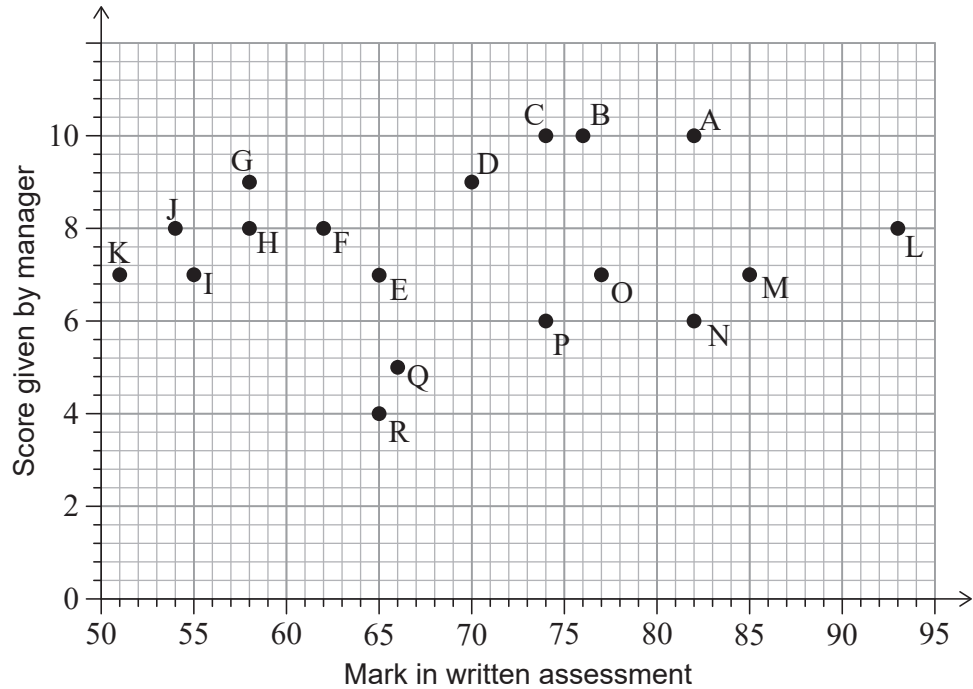
(This question continues on the following page)

(Question 2 continued)

At the end of their first year, the level of performance of each of the 18 employees in the sample is assessed by their department manager. They are awarded a score between 1 (low performance) and 10 (high performance).

The marks in the written assessment and the scores given by the managers are shown in both the table and the scatter diagram.

	National department											International department						
Employee	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Mark in written assessment	82	76	74	70	65	62	58	58	55	54	51	93	85	82	77	74	66	65
Score given by manager	10	10	10	9	7	8	9	8	7	8	7	8	7	6	7	6	5	4



The firm decides to find a Spearman’s rank correlation coefficient, r_s , for this data.

- (c) (i) Without calculation, explain why it might not be appropriate to calculate a correlation coefficient for the whole sample of 18 employees. [2]
- (ii) Find r_s for the seven employees working in the **international** department. [4]
- (iii) Hence comment on the validity of the written assessment as a measure of the level of performance of employees in this department. Justify your answer. [2]

(This question continues on the following page)

Turn over

(Question 2 continued)

The same seven employees are given the written assessment a second time, at the end of the first year, to measure its reliability. Their marks are shown in the table below.

	International department						
	L	M	N	O	P	Q	R
First mark	93	85	82	77	74	66	65
Second mark	90	92	85	73	79	71	65

- (d) (i) State the name of this type of test for reliability. [1]
- (ii) For the data in this table, test the null hypothesis, $H_0 : \rho = 0$, against the alternative hypothesis, $H_1 : \rho > 0$, at the 5% significance level. You may assume that all the requirements for carrying out the test have been met. [4]
- (iii) Hence comment on the reliability of the written assessment. [1]
- (e) The written assessment is in five sections, numbered 1 to 5. At the end of the year, the employees are also given a score for each of five professional attributes: V, W, X, Y and Z.

The firm decides to test the hypothesis that there is a correlation between the mark in a section and the score for an attribute.

They compare marks in **each** of the sections with scores for **each** of the attributes.

- (i) Write down the number of tests they carry out. [1]
- (ii) The tests are performed at the 5% significance level.

Assuming that:

- there is no correlation between the marks in any of the sections and scores in any of the attributes,
- the outcome of each hypothesis test is independent of the outcome of the other hypothesis tests,

find the probability that at least one of the tests will be significant. [4]

- (iii) The firm obtains a significant result when comparing section 2 of the written assessment and attribute X. Interpret this result. [1]

References:

© International Baccalaureate Organization 2021

1. [Maximum mark: 24]

Juliet is a sociologist who wants to investigate if income affects happiness amongst doctors. This question asks you to review Juliet's methods and conclusions.

Juliet obtained a list of email addresses of doctors who work in her city. She contacted them and asked them to fill in an anonymous questionnaire. Participants were asked to state their annual income and to respond to a set of questions. The responses were used to determine a *happiness score* out of 100. Of the 415 doctors on the list, 11 replied.

- (a) (i) Describe **one** way in which Juliet could improve the reliability of her investigation. [1]
- (ii) Describe **one** criticism that can be made about the validity of Juliet's investigation. [1]

Juliet's results are summarized in the following table.

Response	Annual income (\$)	Happiness score
A	65 000	60
B	63 000	52
C	40 000	31
D	125 000	81
E	100 000	48
F	245 000	61
G	48 000	42
H	39 000	40
I	85 000	57
J	92 000	53
K	123 456 789	56

- (b) Juliet classifies response K as an outlier and removes it from the data. Suggest **one** possible justification for her decision to remove it. [1]

(This question continues on the following page)

(Question 1 continued)

- (c) For the remaining ten responses in the table, Juliet calculates the mean happiness score to be 52.5.

- (i) Calculate the mean **annual income** for these remaining responses. [2]
- (ii) Determine the value of r , Pearson's product-moment correlation coefficient, for these remaining responses. [2]

Juliet decides to carry out a hypothesis test on the correlation coefficient to investigate whether increased annual income is associated with greater happiness.

- (d) (i) State why the hypothesis test should be one-tailed. [1]
- (ii) State the null and alternative hypotheses for this test. [2]

The critical value for this test, at the 5% significance level, is 0.549. Juliet assumes that the population is bivariate normal.

- (iii) Determine whether there is significant evidence of a positive correlation between annual income and happiness. Justify your answer. [2]

(This question continues on page 5)

Turn over

— T —

444171210

Blank page

(Question 1 continued)

- (e) Juliet wants to create a model to predict how changing annual income might affect happiness scores. To do this, she assumes that annual income in dollars, X , is the independent variable and the happiness score, Y , is the dependent variable.

She first considers a linear model of the form

$$Y = aX + b.$$

- (i) Use Juliet's data to find the value of a and of b . [1]
- (ii) Interpret, referring to income and happiness, what the value of a represents. [1]

Juliet then considers a quadratic model of the form

$$Y = cX^2 + dX + e.$$

- (iii) Find the value of c , of d and of e . [1]
- (iv) Find the coefficient of determination for each of the two models she considers. [2]
- (v) Hence compare the two models. [1]

Juliet decides to use the coefficient of determination to choose between these two models.

- (vi) Comment on the validity of her decision. [1]

After presenting the results of her investigation, a colleague questions whether Juliet's sample is representative of all doctors in the city.

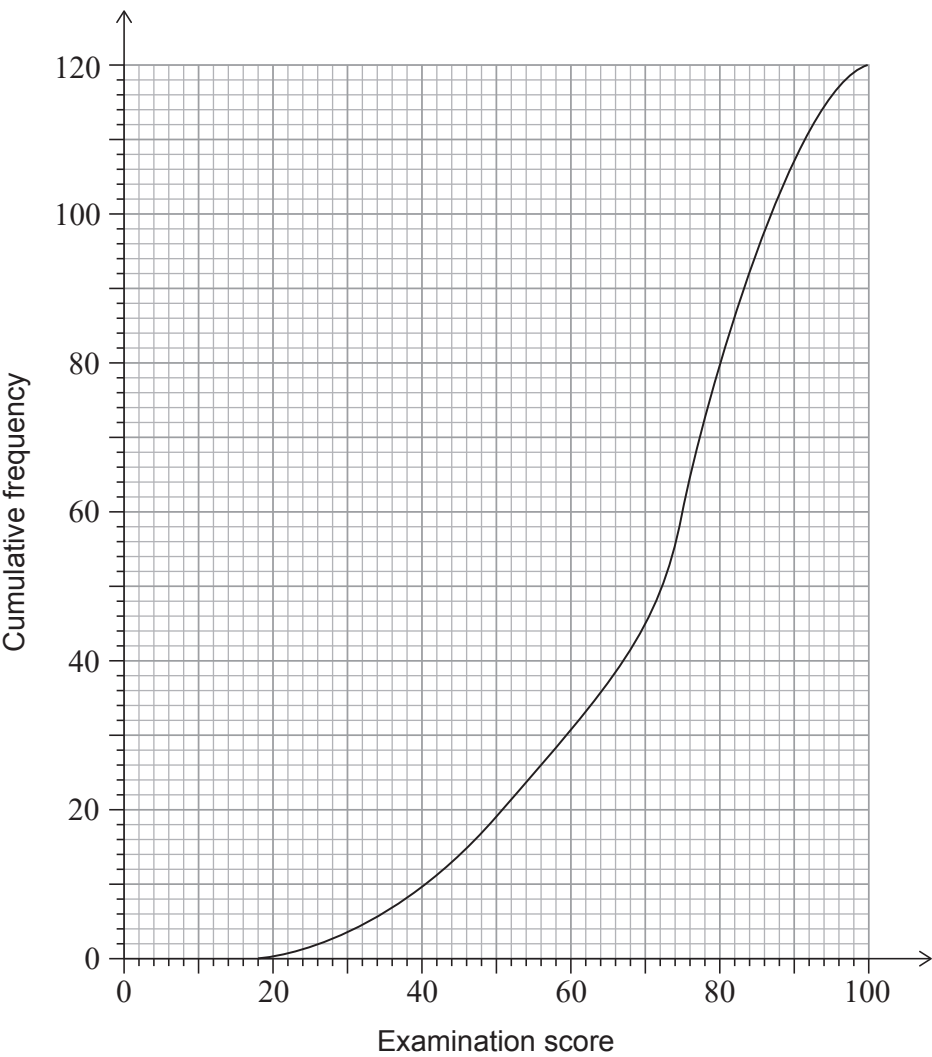
A report states that the mean annual income of doctors in the city is \$80 000. Juliet decides to carry out a test to determine whether her sample could realistically be taken from a population with a mean of \$80 000.

- (f) (i) State the name of the test which Juliet should use. [1]
- (ii) State the null and alternative hypotheses for this test. [1]
- (iii) Perform the test, using a 5% significance level, and state your conclusion in context. [3]

Turn over

7. [Maximum mark: 8]

A group of 120 students sat a history exam. The cumulative frequency graph shows the scores obtained by the students.



- (a) Find the median of the scores obtained. [1]

The students were awarded a grade from 1 to 5, depending on the score obtained in the exam. The number of students receiving each grade is shown in the following table.

Grade	1	2	3	4	5
Number of students	6	13	26	a	b

- (b) Find an expression for a in terms of b . [2]

(This question continues on the following page)



(Question 7 continued)

- (c) The mean grade for these students is 3.65.
- (i) Find the number of students who obtained a grade 5.
 - (ii) Find the minimum score needed to obtain a grade 5.

[5]

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

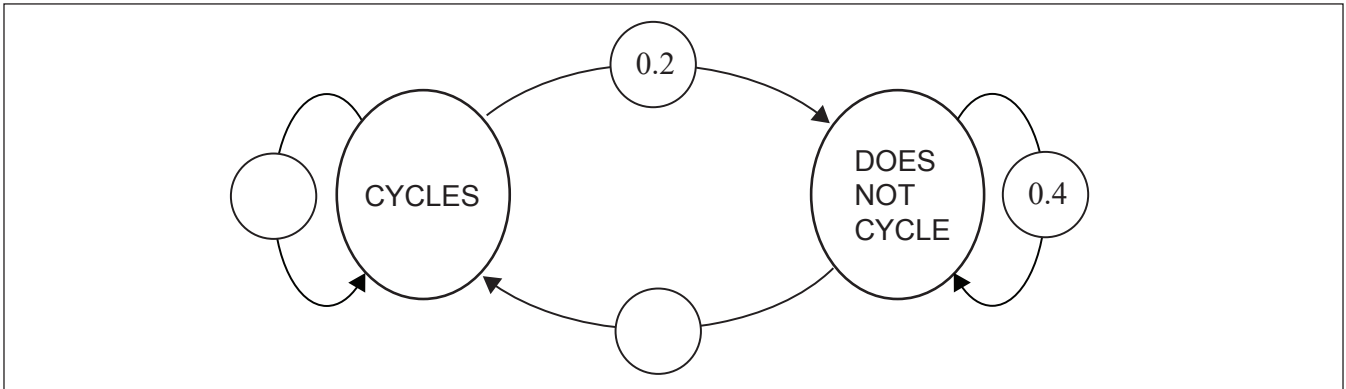
Turn over

9. [Maximum mark: 5]

Katie likes to cycle to work as much as possible. If Katie cycles to work one day then she has a probability of 0.2 of not cycling to work on the next work day. If she does not cycle to work one day then she has a probability of 0.4 of not cycling to work on the next work day.

(a) Complete the following transition diagram to represent this information.

[2]



Katie works for 180 days in a year.

(b) Find the probability that Katie cycles to work on her final working day of the year.

[3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

12. [Maximum mark: 5]

The following table shows the time, in days, from December 1st and the percentage of Christmas trees in stock at a shop on the beginning of that day.

Days since December 1st (d)	1	3	6	9	12	15	18
Percentage of Christmas trees left in stock (x)	100	51	29	21	18	16	14

The following table shows the natural logarithm of both d and x on these days to 2 decimal places.

$\ln(d)$	0	1.10	1.79	2.20	2.48	2.71	2.89
$\ln(x)$	4.61	3.93	3.37	3.04	2.89	2.77	2.64

- Use the data in the second table to find the value of m and the value of b for the regression line, $\ln x = m(\ln d) + b$.
- Assuming that the model found in part (a) remains valid, estimate the percentage of trees in stock when $d = 25$.

[2]

[3]

This image shows a full page of white paper with horizontal dotted lines, typical of primary school writing paper. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

Turn over

14. [Maximum mark: 7]

On Paul’s farm, potatoes are packed in sacks labelled 50 kg. The weights of the sacks of potatoes can be modelled by a normal distribution with mean weight 49.8 kg and standard deviation 0.9 kg.

- (a) Find the probability that a sack is under its labelled weight. [2]
- (b) Find the lower quartile of the weights of the sacks of potatoes. [2]

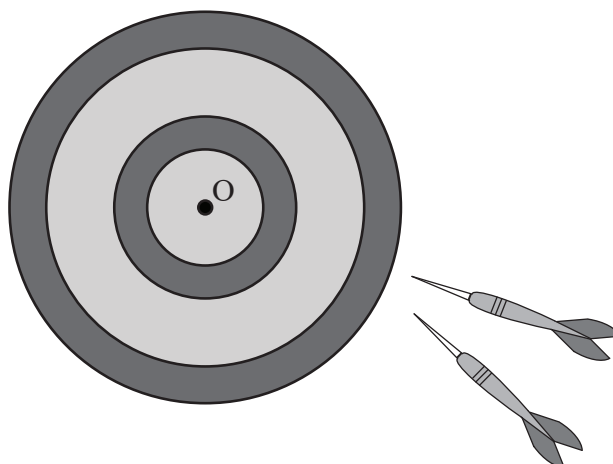
The sacks of potatoes are transported in crates. There are 10 sacks in each crate and the weights of the sacks of potatoes are independent of each other.

- (c) Find the probability that the total weight of the sacks of potatoes in a crate exceeds 500 kg. [3]



3. [Maximum mark: 16]

Arianne plays a game of darts.



The distance that her darts land from the centre, O, of the board can be modelled by a normal distribution with mean 10 cm and standard deviation 3 cm.

(a) Find the probability that

(i) a dart lands less than 13 cm from O.

(ii) a dart lands more than 15 cm from O.

[3]

Each of Arianne's throws is independent of her previous throws.

(b) Find the probability that Arianne throws two consecutive darts that land more than 15 cm from O.

[2]

In a competition a player has three darts to throw on each turn. A point is scored if a player throws **all** three darts to land within a central area around O. When Arianne throws a dart the probability that it lands within this area is 0.8143.

(c) Find the probability that Arianne does **not** score a point on a turn of three darts.

[2]

In the competition Arianne has ten turns, each with three darts.

(d) (i) Find Arianne's expected score in the competition.

(ii) Find the probability that Arianne scores at least 5 points in the competition.

(iii) Find the probability that Arianne scores at least 5 points and less than 8 points.

(iv) Given that Arianne scores at least 5 points, find the probability that Arianne scores less than 8 points.

[9]

7. [Maximum mark: 14]

Loreto is a manager at the Da Vinci health centre. If the mean rate of patients arriving at the health centre exceeds 1.5 per minute then Loreto will employ extra staff. It is assumed that the number of patients arriving in any given time period follows a Poisson distribution.

Loreto performs a hypothesis test to determine whether she should employ extra staff. She finds that 320 patients arrived during a randomly selected 3-hour clinic.

- (a) (i) Write down null and alternative hypotheses for Loreto's test.
- (ii) Using the data from Loreto's sample, perform the hypothesis test at a 5 % significance level to determine if Loreto should employ extra staff.

[7]

Loreto is also concerned about the average waiting time for patients to see a nurse. The health centre aims for at least 95 % of patients to see a nurse in under 20 minutes.

Loreto assumes that the waiting times for patients are independent of each other and decides to perform a hypothesis test at a 10 % significance level to determine whether the health centre is meeting its target.

Loreto surveys 150 patients and finds that 11 of them waited more than 20 minutes.

- (b) (i) Write down null and alternative hypotheses for this test.
- (ii) Perform the test, clearly stating the conclusion in context.

[7]

References:

© International Baccalaureate Organization 2021

1. [Maximum mark: 6]

At the end of a school day, the Headmaster conducted a survey asking students in how many classes they had used the internet.

The data is shown in the following table.

Number of classes in which the students used the internet	0	1	2	3	4	5	6
Number of students	20	24	30	k	10	3	1

(a) State whether the data is discrete or continuous. [1]

The mean number of classes in which a student used the internet is 2.

(b) Find the value of k . [4]

It was not possible to ask every person in the school, so the Headmaster arranged the student names in alphabetical order and then asked every 10th person on the list.

(c) Identify the sampling technique used in the survey. [1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

6. [Maximum mark: 6]

Jae Hee plays a game involving a biased six-sided die.
The faces of the die are labelled -3 , -1 , 0 , 1 , 2 and 5 .
The score for the game, X , is the number which lands face up after the die is rolled.
The following table shows the probability distribution for X .

Score x	-3	-1	0	1	2	5
$P(X=x)$	$\frac{1}{18}$	p	$\frac{3}{18}$	$\frac{1}{18}$	$\frac{2}{18}$	$\frac{7}{18}$

(a) Find the exact value of p . [1]

Jae Hee plays the game once.

(b) Calculate the expected score. [2]

Jae Hee plays the game twice and adds the two scores together.

(c) Find the probability Jae Hee has a **total** score of -3 . [3]



Turn over

A manager wishes to check the mean weight of flour put into bags in his factory. He randomly samples 10 bags and finds the mean weight is 1.478 kg and the standard deviation of the sample is 0.0196 kg.

- Find s_{n-1} for this sample. [2]
- Find a 95 % confidence interval for the population mean, giving your answer to 4 significant figures. [2]
- The bags are labelled as being 1.5 kg weight. Comment on this claim with reference to your answer in part (b). [1]

[illegible]

10. [Maximum mark: 6]

In a coffee shop, the time it takes to serve a customer can be modelled by a normal distribution with a mean of 1.5 minutes and a standard deviation of 0.4 minutes.

Two customers enter the shop together. They are served one at a time.

Find the probability that the total time taken to serve both customers will be less than 4 minutes.

Clearly state any assumptions you have made.

[illegible]

Turn over

12. [Maximum mark: 7]

Product research leads a company to believe that the revenue (R) made by selling its goods at a price (p) can be modelled by the equation.

$$R(p) = cpe^{dp}, \quad c, d \in \mathbb{R}$$

There are two competing models, A and B with different values for the parameters c and d .

Model A has $c = 3$, $d = -0.5$ and model B has $c = 2.5$, $d = -0.6$.

The company experiments by selling the goods at three different prices in three similar areas and the results are shown in the following table.

Area	Price (p)	Revenue (R)
1	1	1.5
2	2	1.8
3	3	1.5

The company will choose the model with the smallest value for the sum of square residuals.

Determine which model the company chose.

A series of horizontal dotted lines for writing.



Turn over

16. [Maximum mark: 6]

The number of fish that can be caught in one hour from a particular lake can be modelled by a Poisson distribution.

The owner of the lake, Emily, states in her advertising that the average number of fish caught in an hour is three.

Tom, a keen fisherman, is not convinced and thinks it is less than three. He decides to set up the following test. Tom will fish for one hour and if he catches fewer than two fish he will reject Emily's claim.

- State a suitable null and alternative hypotheses for Tom's test. [1]
- Find the probability of a Type I error. [2]

The average number of fish caught in an hour is actually 2.5.

- (c) Find the probability of a Type II error. [3]

[illegible]

17. [Maximum mark: 6]

Mr Burke teaches a mathematics class with 15 students. In this class there are 6 female students and 9 male students.

Each day Mr Burke randomly chooses one student to answer a homework question.

In the first month, Mr Burke will teach his class 20 times.

- (a) Find the probability he will choose a female student 8 times. [2]

The Head of Year, Mrs Smith, decides to select a student at random from the year group to read the notices in assembly. There are 80 students in total in the year group. Mrs Smith calculates the probability of picking a male student 8 times in the first 20 assemblies is 0.153357 correct to 6 decimal places.

- (b) Find the number of male students in the year group. [4]



Turn over

2. [Maximum mark: 12]

Slugworth Candy Company sell a variety pack of colourful, shaped sweets.

The sweets are produced such that 80% are star shaped and 20% are shaped like a crescent moon. It is known that 10% of the stars and 30% of the crescent moons are coloured yellow.

(a) A sweet is selected at random.

(i) Find the probability that the sweet is yellow.

(ii) Given that the sweet is yellow, find the probability it is star shaped.

[4]

According to manufacturer specifications, the colours in each variety pack should be distributed as follows.

Colour	Brown	Red	Green	Orange	Yellow	Purple
Percentage (%)	15	25	20	20	10	10

Mr Slugworth opens a pack of 80 sweets and records the frequency of each colour.

Colour	Brown	Red	Green	Orange	Yellow	Purple
Observed Frequency	10	20	16	18	12	4

To investigate if the sample is consistent with manufacturer specifications, Mr Slugworth conducts a χ^2 goodness of fit test. The test is carried out at a 5% significance level.

(b) Write down the null hypothesis for this test.

[1]

(c) **Copy** and complete the following table in your answer booklet.

[2]

Colour	Brown	Red	Green	Orange	Yellow	Purple
Expected Frequency						

(d) Write down the number of degrees of freedom.

[1]

(e) Find the p -value for the test.

[2]

(f) State the conclusion of the test. Give a reason for your answer.

[2]

Turn over

6. [Maximum mark: 14]

A city has two cable companies, X and Y. Each year 20 % of the customers using company X move to company Y and 10 % of the customers using company Y move to company X. All additional losses and gains of customers by the companies may be ignored.

- (a) Write down a transition matrix T representing the movements between the two companies in a particular year. [2]
- (b) Find the eigenvalues and corresponding eigenvectors of T . [4]
- (c) Hence write down matrices P and D such that $T = PDP^{-1}$. [2]

Initially company X and company Y both have 1200 customers.

- (d) Find an expression for the number of customers company X has after n years, where $n \in \mathbb{N}$. [5]
- (e) Hence write down the number of customers that company X can expect to have in the long term. [1]

Turn over

1. [Maximum mark: 27]

Two IB schools, A and B, follow the IB Diploma Programme but have different teaching methods. A research group tested whether the different teaching methods lead to a similar final result.

For the test, a group of eight students were randomly selected from each school. Both samples were given a standardized test at the start of the course and a prediction for total IB points was made based on that test; this was then compared to their points total at the end of the course.

Previous results indicate that both the predictions from the standardized tests and the final IB points can be modelled by a normal distribution.

It can be assumed that:

- the standardized test is a valid method for predicting the final IB points
- that variations from the prediction can be explained through the circumstances of the student or school.

- (a) Identify a test that might have been used to verify the null hypothesis that the predictions from the standardized test can be modelled by a normal distribution. [1]
- (b) State why comparing only the final IB points of the students from the two schools would not be a valid test for the effectiveness of the two different teaching methods. [1]

(This question continues on the following page)

(Question 1 continued)

The data for school A is shown in the following table.

School A

Student number	Gender	Predicted IB points (p)	Final IB points (f)
1	male	43.2	44
2	male	36.5	34
3	female	37.1	38
4	male	30.9	28
5	male	41.1	39
6	female	35.1	39
7	male	36.4	40
8	male	38.2	38
Mean		37.31	37.5

- (c) For each student, the change from the predicted points to the final points ($f - p$) was calculated.
- (i) Find the mean change.
- (ii) Find the standard deviation of the changes. [3]
- (d) Use a paired t -test to determine whether there is significant evidence that the students in school A have improved their IB points since the start of the course. [4]

(This question continues on the following page)

Turn over

(Question 1 continued)

The data for school B is shown in the following table.

School B

Student number	Gender	Final IB points– Predicted IB points ($f - p$)
1	male	8.7
2	female	–1.1
3	female	4.8
4	female	–1.5
5	male	2.5
6	female	3.2
7	female	–1.3
8	female	3.1
Mean		2.3

- (e) (i) Use an appropriate test to determine whether there is evidence, at the 5% significance level, that the students in school B have improved more than those in school A.
- (ii) State why it was important to test that both sets of points were normally distributed. [6]

(This question continues on the following page)

(Question 1 continued)

School A also gives each student a score for effort in each subject. This effort score is based on a scale of 1 to 5 where 5 is regarded as outstanding effort.

Student number	Gender	Predicted IB points	Final IB points	Average effort score
1	male	43.2	44	4.4
2	male	36.5	34	4.2
3	female	37.1	38	4.7
4	male	30.9	28	4.3
5	male	41.1	39	3.9
6	female	35.1	39	4.9
7	male	36.4	40	4.9
8	male	38.2	38	4.3
Mean		37.31	37.5	4.45

It is claimed that the effort put in by a student is an important factor in improving upon their predicted IB points.

- (f) (i) Perform a test on the data from school A to show it is reasonable to assume a linear relationship between effort scores and improvements in IB points. You may assume effort scores follow a normal distribution.
- (ii) Hence, find the expected improvement between predicted and final points for an increase of one unit in effort grades, giving your answer to one decimal place. [4]

A mathematics teacher in school A claims that the comparison between the two schools is not valid because the sample for school B contained mainly girls and that for school A, mainly boys. She believes that girls are likely to show a greater improvement from their predicted points to their final points.

She collects more data from other schools, asking them to class their results into four categories as shown in the following table.

	$(f-p) < -2$	$-2 \leq (f-p) < 0$	$0 \leq (f-p) < 2$	$(f-p) \geq 2$
Male	6	8	10	9
Female	3	8	14	8

- (g) Use an appropriate test to determine whether showing an improvement is independent of gender. [6]
- (h) If you were to repeat the test performed in part (e) intending to compare the quality of the teaching between the two schools, suggest **two** ways in which you might choose your sample to improve the validity of the test. [2]

Turn over

3. [Maximum mark: 7]

Leo is investigating whether a six-sided die is fair. He rolls the die 60 times and records the observed frequencies in the following table:

Number on die	1	2	3	4	5	6
Observed frequency	8	7	6	15	12	12

Leo carries out a χ^2 goodness of fit test at a 5% significance level.

- Write down the null and alternative hypotheses. [1]
- Write down the degrees of freedom. [1]
- Write down the expected frequency of rolling a 1. [1]
- Find the p -value for the test. [2]
- State the conclusion of the test. Give a reason for your answer. [2]

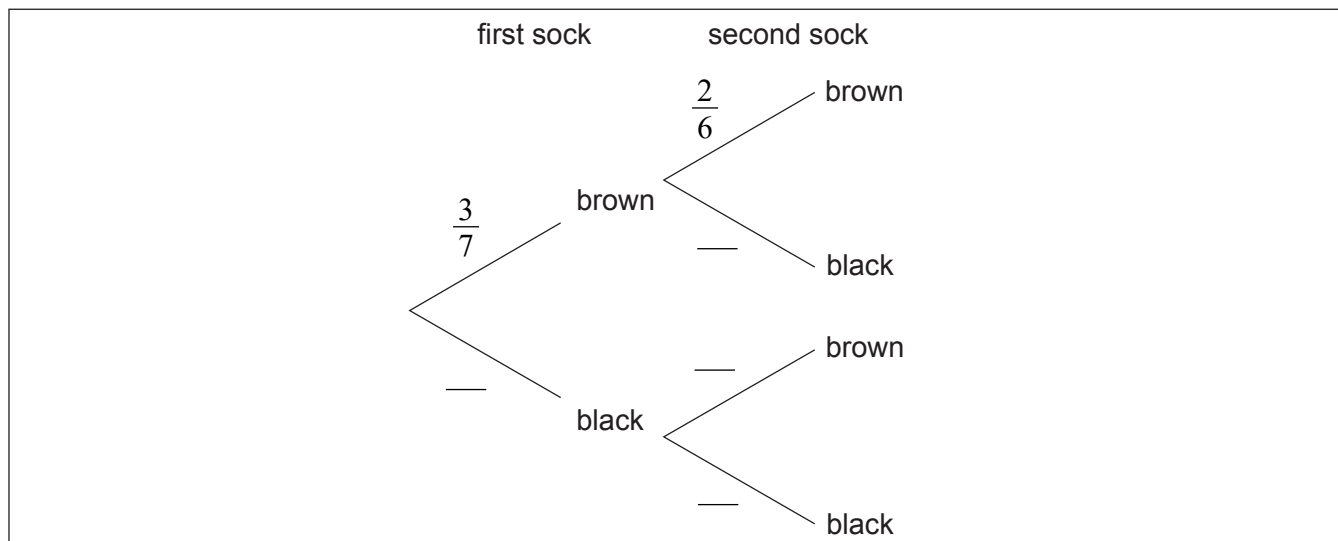
[illegible]

4. [Maximum mark: 7]

Karl has three brown socks and four black socks in his drawer. He takes two socks at random from the drawer.

(a) Complete the tree diagram.

[2]



(b) Find the probability that Karl takes two socks of the same colour.

[2]

(c) Given that Karl has two socks of the same colour find the probability that he has two brown socks.

[3]

[illegible]

Turn over

9. [Maximum mark: 6]

A company produces bags of sugar with a labelled weight of 1 kg. The weights of the bags are normally distributed with a mean of 1 kg and a standard deviation of 100 g. In an inspection, if the weight of a randomly chosen bag is less than 950 g then the company fails the inspection.

- (a) Find the probability that the company fails the inspection. [2]

A statistician in the company suggests it would be fairer if the company passes the inspection when the mean weight of five randomly chosen bags is greater than 950 g.

- (b) Find the probability of passing the inspection if the statistician’s suggestion is followed. [4]



Turn over

15. [Maximum mark: 7]

The number of cars arriving at a junction in a particular town in any given minute between 9:00 am and 10:00 am is historically known to follow a Poisson distribution with a mean of 5.4 cars per minute.

A new road is built near the town. It is claimed that the new road has decreased the number of cars arriving at the junction.

To test the claim, the number of cars, X , arriving at the junction between 9:00 am and 10:00 am on a particular day will be recorded. The test will have the following hypotheses:

- H_0 : the mean number of cars arriving at the junction has not changed,
- H_1 : the mean number of cars arriving at the junction has decreased.

The alternative hypothesis will be accepted if $X \leq 300$.

- (a) Assuming the null hypothesis to be true, state the distribution of X . [1]
- (b) Find the probability of a Type I error. [2]
- (c) Find the probability of a Type II error, if the number of cars now follows a Poisson distribution with a mean of 4.5 cars per minute. [4]



1. [Maximum mark: 6]

A group of 130 applicants applied for admission into either the Arts programme or the Sciences programme at a university. The outcomes of their applications are shown in the following table.

	Accepted	Rejected
Arts programme	17	24
Sciences programme	25	64

- (a) Find the probability that a randomly chosen applicant from this group was accepted by the university. [1]

An applicant is chosen at random from this group. It is found that they were accepted into the programme of their choice.

- (b) Find the probability that the applicant applied for the Arts programme. [2]

Two different applicants are chosen at random from the original group.

- (c) Find the probability that both applicants applied to the Arts programme. [3]

(This question continues on the following page)



(Question 1 continued)

[illegible]

Turn over

3. [Maximum mark: 7]

A polygraph test is used to determine whether people are telling the truth or not, but it is not completely accurate. When a person tells the truth, they have a 20 % chance of failing the test. Each test outcome is independent of any previous test outcome.

10 people take a polygraph test and all 10 tell the truth.

- (a) Calculate the expected number of people who will pass this polygraph test. [2]
- (b) Calculate the probability that exactly 4 people will fail this polygraph test. [2]
- (c) Determine the probability that fewer than 7 people will pass this polygraph test. [3]



Turn over

9. [Maximum mark: 8]

A psychologist records the number of digits (d) of π that a sample of IB Mathematics higher level candidates could recall.

d	2	3	4	5	6	7
Frequency	2	6	24	21	11	3

- (a) Find an unbiased estimate of the population mean of d . [1]
- (b) Find an unbiased estimate of the population variance of d . [2]

The psychologist has read that in the general population people can remember an average of 4.4 digits of π . The psychologist wants to perform a statistical test to see if IB Mathematics higher level candidates can remember more digits than the general population.

- (c) $H_0: \mu = 4.4$ is the null hypothesis for this test.
- (i) State the alternative hypothesis.
- (ii) Given that all assumptions for this test are satisfied, carry out an appropriate hypothesis test. State and justify your conclusion. Use a 5% significance level. [5]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

12. [Maximum mark: 6]

The sex of cuttlefish is difficult to determine visually, so it is often found by weighing the cuttlefish.

The weights of adult male cuttlefish are known to be normally distributed with mean 10 kg and standard deviation 0.5 kg.

The weights of adult female cuttlefish are known to be normally distributed with mean 12 kg and standard deviation 1 kg.

A zoologist uses the null hypothesis that in the absence of information a cuttlefish is male.

If the weight is found to be above 11.5 kg the cuttlefish is classified as female.

- (a) Find the probability of making a Type I error when weighing a male cuttlefish. [2]
- (b) Find the probability of making a Type II error when weighing a female cuttlefish. [2]

90 % of adult cuttlefish are male.

- (c) Find the probability of making an error using the zoologist’s method. [2]



1. [Maximum mark: 16]

The scores of the eight highest scoring countries in the 2019 Eurovision song contest are shown in the following table.

	Eurovision score
Netherlands	498
Italy	472
Russia	370
Switzerland	364
Sweden	334
Norway	331
North Macedonia	305
Azerbaijan	302

(a) For this data, find

(i) the upper quartile.

(ii) the interquartile range.

[4]

(b) Determine if the Netherlands' score is an outlier for this data. Justify your answer.

[3]

(This question continues on the following page)

Turn over

(Question 1 continued)

Chester is investigating the relationship between the highest-scoring countries' Eurovision score and their population size to determine whether population size can reasonably be used to predict a country's score.

The populations of the countries, to the nearest million, are shown in the table.

	Population (x) (millions)	Eurovision score (y)
Netherlands	17	498
Italy	60	472
Russia	145	370
Switzerland	9	364
Sweden	10	334
Norway	5	331
North Macedonia	2	305
Azerbaijan	10	302

Chester finds that, for this data, the Pearson's product moment correlation coefficient is $r = 0.249$.

- (c) State whether it would be appropriate for Chester to use the equation of a regression line for y on x to predict a country's Eurovision score. Justify your answer.

[2]

Chester then decides to find the Spearman's rank correlation coefficient for this data, and creates a table of ranks.

	Population rank (to the nearest million)	Eurovision score rank
Netherlands	3	1
Italy	2	2
Russia	1	3
Switzerland	a	4
Sweden	b	5
Norway	7	6
North Macedonia	8	7
Azerbaijan	c	8

(This question continues on the following page)

(Question 1 continued)

(d) Write down the value of:

(i) a ,

(ii) b ,

(iii) c .

[3]

(e) (i) Find the value of the Spearman's rank correlation coefficient r_s .

(ii) Interpret the value obtained for r_s .

[3]

(f) When calculating the ranks, Chester incorrectly read the Netherlands' score as 478. Explain why the value of the Spearman's rank correlation r_s does not change despite this error.

[1]

Turn over

3. [Maximum mark: 13]

A Principal would like to compare the students in his school with a national standard. He decides to give a test to eight students made up of four boys and four girls. One of the teachers offers to find the volunteers from his class.

- (a) Name the type of sampling that best describes the method used by the Principal. [1]

The marks out of 40, for the students who took the test, are:

25, 29, 38, 37, 12, 18, 27, 31.

- (b) For the eight students find

(i) the mean mark.

(ii) the standard deviation of the marks. [3]

The national standard mark is 25.2 out of 40.

- (c) Perform an appropriate test at the 5% significance level to see if the mean marks achieved by the students in the school are higher than the national standard. It can be assumed that the marks come from a normal population. [5]

- (d) State one reason why the test might not be valid. [1]

Two additional students take the test at a later date and the mean mark for all ten students is 28.1 and the standard deviation is 8.4.

For further analysis, a standardized score out of 100 for the ten students is obtained by multiplying the scores by 2 and adding 20.

- (e) For the ten students, find

(i) their mean standardized score.

(ii) the standard deviation of their standardized score. [3]

Turn over

5. [Maximum mark: 15]

The aircraft for a particular flight has 72 seats. The airline's records show that historically for this flight only 90% of the people who purchase a ticket arrive to board the flight. They assume this trend will continue and decide to sell extra tickets and hope that no more than 72 passengers will arrive.

The number of passengers that arrive to board this flight is assumed to follow a binomial distribution with a probability of 0.9.

- (a) The airline sells 74 tickets for this flight. Find the probability that more than 72 passengers arrive to board the flight. [3]
- (b) (i) Write down the expected number of passengers who will arrive to board the flight if 72 tickets are sold. [2]
- (ii) Find the maximum number of tickets that could be sold if the expected number of passengers who arrive to board the flight must be less than or equal to 72. [2]

Each passenger pays \$150 for a ticket. If too many passengers arrive, then the airline will give \$300 in compensation to each passenger that cannot board.

- (c) Find, to the nearest integer, the expected increase or decrease in the money made by the airline if they decide to sell 74 tickets rather than 72. [8]

Turn over

5. [Maximum mark: 12]

A geneticist uses a Markov chain model to investigate changes in a specific gene in a cell as it divides. Every time the cell divides, the gene may mutate between its normal state and other states.

The model is of the form

$$\begin{pmatrix} X_{n+1} \\ Z_{n+1} \end{pmatrix} = \mathbf{M} \begin{pmatrix} X_n \\ Z_n \end{pmatrix}$$

where X_n is the probability of the gene being in its normal state after dividing for the n th time, and Z_n is the probability of it being in another state after dividing for the n th time, where $n \in \mathbb{N}$.

Matrix $\mathbf{M}$ is found to be $\begin{pmatrix} 0.94 & b \\ 0.06 & 0.98 \end{pmatrix}$.

- (a) (i) Write down the value of b .
(ii) What does b represent in this context? [2]
- (b) Find the eigenvalues of $\mathbf{M}$. [3]
- (c) Find the eigenvectors of $\mathbf{M}$. [3]
- (d) The gene is in its normal state when $n = 0$. Calculate the probability of it being in its normal state
 - (i) when $n = 5$;
 - (ii) in the long term. [4]

1. [Maximum mark: 27]

This question uses statistical tests to investigate whether advertising leads to increased profits for a grocery store.

Aimmika is the manager of a grocery store in Nong Khai. She is carrying out a statistical analysis on the number of bags of rice that are sold in the store each day. She collects the following sample data by recording how many bags of rice the store sells each day over a period of 90 days.

Number of bags of rice sold	0	1	2	3	4	5	6	7	8	9	10
Number of days	1	8	12	11	19	14	13	8	2	0	2

She believes that her data follows a Poisson distribution.

- (a) (i) Find the mean and variance for the sample data given in the table. [2]
- (ii) Hence state why Aimmika believes her data follows a Poisson distribution. [1]
- (b) State one assumption that Aimmika needs to make about the sales of bags of rice to support her belief that it follows a Poisson distribution. [1]

Aimmika knows from her historic sales records that the store sells an average of 4.2 bags of rice each day. The following table shows the expected frequency of bags of rice sold each day during the 90 day period, assuming a Poisson distribution with mean 4.2.

Number of bags of rice sold	≤ 1	2	3	4	5	6	7	≥ 8
Expected frequency	a	11.903	16.665	b	14.698	10.289	6.173	c

- (c) Find the value of a , of b , and of c . Give your answers to 3 decimal places. [5]

(This question continues on the following page)

(Question 1 continued)

Aimmika decides to carry out a χ^2 goodness of fit test at the 5 % significance level to see whether the data follows a Poisson distribution with mean 4.2.

- (d) (i) Write down the number of degrees of freedom for her test. [1]
- (ii) Perform the χ^2 goodness of fit test and state, with reason, a conclusion. [7]

Aimmika claims that advertising in a local newspaper for 300 Thai Baht (THB) per day will increase the number of bags of rice sold. However, Nichakarn, the owner of the store, claims that the advertising will **not** increase the store's overall profit.

Nichakarn agrees to advertise in the newspaper for the next 60 days. During that time, Aimmika records that the store sells 282 bags of rice with a profit of 495 THB on each bag sold.

- (e) Aimmika wants to carry out an appropriate hypothesis test to determine whether the number of bags of rice sold during the 60 days increased when compared with the historic sales records.
- (i) By finding a critical value, perform this test at a 5 % significance level. [6]
- (ii) Hence state the probability of a Type I error for this test. [1]
- (f) By considering the claims of both Aimmika and Nichakarn, explain whether the advertising was beneficial to the store. [3]

Turn over

— T —

4444-1210

Blank page

1. [Maximum mark: 5]

Sergio is interested in whether an adult’s favourite breakfast berry depends on their income level. He obtains the following data for 341 adults and decides to carry out a χ^2 test for independence, at the 10 % significance level.

		Income level		
		Low	Medium	High
Favourite berry	Strawberry	21	39	30
	Blueberry	39	67	42
	Other berry	32	45	26

(a) Write down the null hypothesis. [1]

(b) Find the value of the χ^2 statistic. [2]

The critical value of this χ^2 test is 7.78.

(c) Write down Sergio’s conclusion to the test in context. Justify your answer. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

5. [Maximum mark: 7]

Taizo plays a game where he throws one ball at two bottles that are sitting on a table.
The probability of knocking over bottles, in any given game, is shown in the following table.

Number of bottles knocked over	0	1	2
Probability	0.5	0.4	0.1

- (a) Taizo plays two games that are independent of each other. Find the probability that Taizo knocks over a **total** of two bottles.

[4]

In any given game, Taizo will win k points if he knocks over two bottles, win 4 points if he knocks over one bottle and lose 8 points if no bottles are knocked over.

- (b) Find the value of k such that the game is fair.

[3]



16. [Maximum mark: 8]

The principal of a school is concerned that only 30 % of her students are choosing healthy options from the school canteen. She organizes a campaign to promote healthy eating and decides to test if the campaign has increased the number of students choosing healthy options. She assumes that a student’s choice is independent of other students’ choices.

- (a) Write down suitable hypotheses for this test. [2]

The principal decides to take a random sample of 80 students. She will reject the null hypothesis if at least 31 students choose a healthy option.

- (b) Find the probability that she makes a Type I error. [3]

In fact, the campaign led to 40 % of her students choosing a healthy option.

- (c) Find the probability that she makes a Type II error. [3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

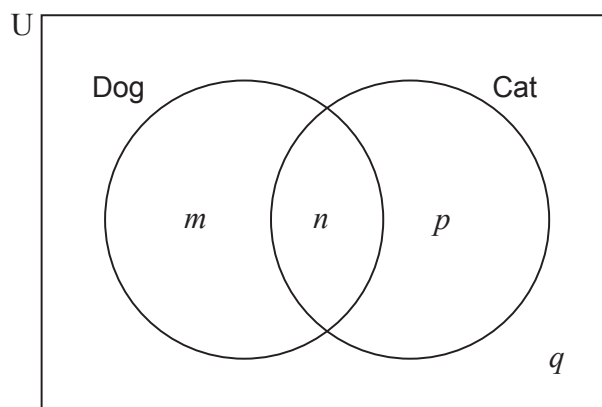
.....



1. [Maximum mark: 15]

At Mirabooka Primary School, a survey found that 68% of students have a dog and 36% of students have a cat. 14% of students have both a dog and a cat.

This information can be represented in the following Venn diagram, where m , n , p and q represent the percentage of students within each region.



(a) Find the value of

(i) m .

(ii) n .

(iii) p .

(iv) q .

[4]

(b) Find the probability that a randomly chosen student

(i) has a dog but does not have a cat.

(ii) has a dog given that they do not have a cat.

[3]

(This question continues on the following page)

(Question 1 continued)

Each year, one student is chosen randomly to be the school captain of Mirabooka Primary School.

Tim is using a binomial distribution to make predictions about how many of the next 10 school captains will own a dog. He assumes that the percentages found in the survey will remain constant for future years and that the events “being a school captain” and “having a dog” are independent.

Use Tim’s model to find the probability that in the next 10 years

- (c) (i) 5 school captains have a dog.
- (ii) more than 3 school captains have a dog.
- (iii) exactly 9 school captains in succession have a dog. [7]

John randomly chooses 10 students from the survey.

- (d) State why John should not use the binomial distribution to find the probability that 5 of these students have a dog. [1]

Turn over

— T —

0022-1202

Blank page

6. [Maximum mark: 18]

A company makes doors for kitchen cupboards from two layers. The inside layer is wood, and its thickness is normally distributed with mean 7 mm and standard deviation 0.3 mm. The outside layer is plastic, and its thickness is normally distributed with mean 3 mm and standard deviation 0.16 mm. The thickness of the plastic is independent of the thickness of the wood.

- (a) Find the probability that a randomly chosen door has a total thickness of less than 9.5 mm. [5]

Eight doors are to be packed into a box to send to a customer. The width of the box is 82 mm. The thickness of each door is independent.

- (b) Find the probability that the total thickness of the eight doors is greater than the width of the box. [4]

The company buys two new machines, A and B, to make the wooden layers. An employee claims that the layers from machine B are thinner than the layers from machine A. In order to test this claim, a random sample is taken from each machine.

The seven layers in the sample from machine A have a thickness, in mm, of

6.23, 7.04, 7.31, 6.79, 6.91, 6.79, 7.47.

- (c) Find the
- (i) mean.
 - (ii) unbiased estimate of the population variance. [3]

The eight layers in the sample from machine B have a mean thickness of 6.89 mm and $s_{n-1} = 0.31$.

- (d) Perform a suitable test, at the 5% significance level, to test the employee's claim. You may assume the thickness of the wooden layers from each machine are normally distributed with equal population variance. [6]

Turn over

1. [Maximum mark: 29]

In this question, you will explore possible approaches to using historical sports results for making predictions about future sports matches.

Two friends, Peter and Helen, are discussing ways of predicting the outcomes of international football matches involving Argentina.

Peter suggests analysing historical data to help make predictions. He lists the results of the most recent 240 matches in which Argentina played, in chronological order, then considers blocks of four matches at a time. He counts how many times Argentina has won in each block. The following table shows his results for the 60 blocks of four matches.

Number of wins for Argentina (per block of four matches)	Frequency
0	0
1	11
2	21
3	21
4	7

- (a) Determine the mean number of wins per block of four matches for Argentina. [2]

Peter thinks that this data can be modelled by a binomial distribution with $n = 4$ and decides to carry out a χ^2 goodness of fit test.

- (b) Use Peter's data to write down an estimate for the probability p for this binomial model. [1]

- (c) (i) Use the binomial model to find the probability that Argentina win zero matches in a block of four matches. [1]

- (ii) Find the expected frequency for zero wins. [2]

(This question continues on the following page)

Turn over

(Question 1 continued)

As some expected frequencies are less than 5, Peter combines rows in his table to produce the following observed frequencies. He then uses his binomial model to find appropriate expected frequencies, correct to one decimal place.

Number of wins for Argentina (per block of four matches)	Observed frequency	Expected frequency
0 or 1	11	10.8
2	21	20.7
3	21	20.7
4	7	7.8

- (d) Peter uses this table to carry out a χ^2 goodness of fit test, to test the hypothesis that the data follows a binomial distribution with $n = 4$, at the 5% significance level.

For this test, state

- (i) the null hypothesis;

[1]
- (ii) the number of degrees of freedom;

[1]
- (iii) the p -value;

[2]
- (iv) the conclusion, justifying your answer.

[2]
- (e) Using Peter’s binomial model, find the probability that Argentina will win at least one of their next four international football matches.

[2]

(This question continues on the following page)

(Question 1 continued)

Helen thinks that a better prediction might be made by considering the transition between matches. To keep the model simple, she decides to use only two states: Argentina won (A) or Argentina did not win (B). Helen looks at Peter’s list of results and counts the number of times that:

- Argentina won, twice in succession (AA),
- Argentina won, then did not win (AB),
- Argentina did not win, then won (BA),
- Argentina did not win, twice in succession (BB).

She recorded the following results.

Transition	Frequency
AA	85
AB	60
BA	62
BB	32

Helen uses the relative frequencies to estimate the probabilities in a transition matrix.

- (f)

(i)

Given that Argentina won the previous match, show that Helen’s estimate for the probability of Argentina winning the next match is $\frac{17}{29}$.

[2]

(ii)

Write down the transition matrix, T , for Helen’s model.

[2]
- (g)

(i)

Show that the characteristic polynomial of T is $1363\lambda^2 - 1263\lambda - 100 = 0$.

[3]

(ii)

Hence or otherwise, find the eigenvalues of T .

[1]

(iii)

Find the corresponding eigenvectors.

[3]

(h)

In her retirement, many years from now, Helen is planning to travel to three consecutive international football matches involving Argentina. Use Helen’s model to find the probability that Argentina will win all three matches.

[4]
- Turn over
- Apollo | Original source crop | 72

2. [Maximum mark: 6]

Carys believes that, on a memory retention test, the mean score of bilingual people (μ_b) will be higher than the mean score of monolingual people (μ_m). Carys gave a memory retention test to a random sample of students in her class. The results are shown in the two tables.

	Scores									
Bilingual	100	94	100	90	100	94	98	98	98	98

	Scores							
Monolingual	97	92	88	98	88	94	100	100

Carys performs a one-tailed t -test at a 5 % level of significance. It is assumed that the scores are normally distributed and the samples have equal variances.

- (a) State the null and alternative hypotheses. [2]
- (b) Calculate the p -value for this test. [2]
- (c) State the conclusion of the test in the context of the question. Justify your answer. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



8. [Maximum mark: 7]

The random variables (X, Y) follow a bivariate normal distribution with product-moment correlation coefficient ρ . The values of six random observations of (X, Y) are shown in the table.

x	6.3	4.1	5.6	9.2	7.8	8.2
y	9.2	4.9	8.9	10.3	8.9	9.8

- (a) State null and alternative hypotheses which could be used to test whether there is a linear correlation between X and Y .

[2]
- (b) Determine the value of

(i) the product-moment correlation coefficient, r , of the sample.

(ii) the corresponding p -value.

[3]
- (c) State whether your result from part (b)(ii) indicates there is sufficient evidence to claim that, at the 5% significance level, X and Y are not linearly correlated.

Give a reason for your answer.

[2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

Please **do not** write on this page.

Answers written on this page
will not be marked.



9. [Maximum mark: 5]

On a specific day, the speed of cars as they pass a speed camera can be modelled by a normal distribution with a mean of 67.3 km h^{-1} .

A speed of 75.7 km h^{-1} is two standard deviations from the mean.

- (a) Find the standard deviation for the speed of the cars. [2]

It is found that 82% of cars on this road travel at speeds between $p \text{ km h}^{-1}$ and $q \text{ km h}^{-1}$, where $p < q$. This interval includes cars travelling at a speed of 74 km h^{-1} .

- (b) Show that the region of the normal distribution between p and q is **not** symmetrical about the mean. [3]

This image shows a full page of white paper with horizontal dotted lines, typical of notebook paper. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

Turn over

11. [Maximum mark: 6]

A shop sells oranges and lemons. The weights of the oranges are assumed to be normally distributed with mean 205 grams and standard deviation 5 grams. The weights of the lemons are assumed to be normally distributed with mean 105 grams and standard deviation 3 grams.

Nelia selects 1 orange and 2 lemons at random and independent of each other. Calculate the probability that the weight of her orange exceeds the combined weight of her lemons.

[illegible]

16. [Maximum mark: 6]

The relationship between the intensity, I , of a light source and the distance, d , from the light source can be modelled by $I = \frac{k}{d^2}$.

Pablo measures the intensity of a light source at different distances. The data collected is shown in the table.

$d(\text{m})$	1	2	5
$I(\text{lm})$	42	11	1.5

Pablo finds the sum of square residuals in the form $1.0641k^2 - 89.62k + c$.

(a) Find the exact value of c . [4]

(b) Hence find the least squares regression curve of the form $I = \frac{k}{d^2}$. [2]



Turn over

2. [Maximum mark: 4]

A company that owns many restaurants wants to determine if there are differences in the quality of the food cooked for three different meals: breakfast, lunch and dinner.

Their quality assurance team randomly selects 500 items of food to inspect. The quality of this food is classified as perfect, satisfactory, or poor. The data is summarized in the following table.

		Quality			Total
		Perfect	Satisfactory	Poor	
Meal	Breakfast	101	124	7	232
	Lunch	68	81	5	154
	Dinner	35	69	10	114
	Total	204	274	22	500

A χ^2 test at the 5 % significance level is carried out to determine if there is significant evidence of a difference in the quality of the food cooked for the three meals.

The critical value for this test is 9.488.

The hypotheses for this test are:

H_0 : The quality of the food and the type of meal are independent.

H_1 : The quality of the food and the type of meal are not independent.

(a) Find the χ^2 statistic. [2]

(b) State, with justification, the conclusion for this test. [2]

(This question continues on the following page)



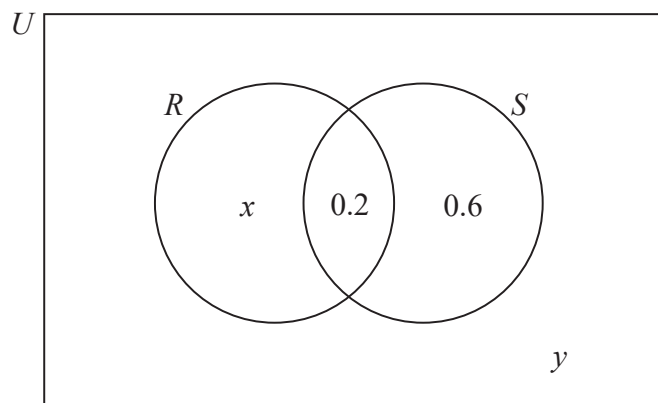
(Question 2 continued)

[illegible]

Turn over

3. [Maximum mark: 7]

The following Venn diagram shows two independent events, R and S . The values in the diagram represent probabilities.



- (a) Find the value of x . [3]
- (b) Find the value of y . [2]
- (c) Find $P(R' | S')$. [2]

[illegible]

5. [Maximum mark: 6]

The lengths of the seeds from a particular mango tree are approximated by a normal distribution with a mean of 4 cm and a standard deviation of 0.25 cm.

A seed from this mango tree is chosen at random.

- (a) Calculate the probability that the length of the seed is less than 3.7 cm. [2]

It is known that 30 % of the seeds have a length greater than k cm.

- (b) Find the value of k . [2]

For a seed of length d cm, chosen at random, $P(4 - m < d < 4 + m) = 0.6$.

- (c) Find the value of m . [2]



7. [Maximum mark: 6]

Akar starts a new job in Australia and needs to travel daily from Wollongong to Sydney and back. He travels to work for 28 consecutive days and therefore makes 56 single journeys. Akar makes all journeys by bus.

The probability that he is successful in getting a seat on the bus for any single journey is 0.86.

- (a) Determine the expected number of these 56 journeys for which Akar gets a seat on the bus. [1]
- (b) Find the probability that Akar gets a seat on at least 50 journeys during these 28 days. [3]

The probability that Akar gets a seat on at most n journeys is at least 0.25.

- (c) Find the smallest possible value of n . [2]



9. [Maximum mark: 9]

At a running club, Sung-Jin conducts a test to determine if there is any association between an athlete’s age and their best time taken to run 100m. Eight athletes are chosen at random, and their details are shown below.

Athlete	A	B	C	D	E	F	G	H
Age (years)	13	17	22	18	19	25	11	36
Time (seconds)	13.4	14.6	13.4	12.9	12.0	11.8	17.0	13.1

Sung-Jin decides to calculate the Spearman’s rank correlation coefficient for his set of data.

- (a) Complete the table of ranks. [2]

Athlete	A	B	C	D	E	F	G	H
Age rank			3					
Time rank							1	

- (b) Calculate the Spearman’s rank correlation coefficient, r_s . [2]
- (c) Interpret this value of r_s in the context of the question. [1]
- (d) Suggest a mathematical reason why Sung-Jin may have decided not to use Pearson’s product-moment correlation coefficient with his data from the original table. [1]
- (e) (i) Find the coefficient of determination for the data from the original table.
- (ii) Interpret this value in the context of the question. [3]

(This question continues on the following page)



(Question 9 continued)

[illegible]

Turn over

10. [Maximum mark: 6]

A chocolate company plans to produce chocolate bars with special flavours. They survey 246 people to determine if there is any particular preference for one of the flavours.

The table below shows the information collected.

Hot chilli	Almond crunch	Spiced Chai	Ginger'n'lime
75	59	46	66

A χ^2 goodness of fit test at the 5 % significance level is carried out on the data.

The critical value for the test is 7.82.

- (a) State the null and alternative hypotheses for this test. [2]
- (b) Perform the test and give your conclusion in context. [4]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



15. [Maximum mark: 6]

A random sample of eight packets of Apollo coffee granules are selected from a supermarket shelf.

The weights of the coffee granules present in each packet are as follows:

222 g 226 g 221 g 228 g 227 g 225 g 222 g 223 g

- (a)
 - (i) Find an unbiased estimate for the mean weight of coffee granules in a packet of Apollo coffee.
 - (ii) Calculate a 95 % confidence interval for the population mean. Give your answer to four significant figures. [3]
- (b) State one assumption you have made in order for your interval to be valid. [1]
- (c) The label of each packet has a description which includes the phrase: “contains 226 g of coffee granules”.

Using your answer to part (a)(ii), briefly comment on the claim on the label. [2]

[illegible]

Turn over

1. [Maximum mark: 13]

The mean annual temperatures for Earth, recorded at fifty-year intervals, are shown in the table.

Year (x)	1708	1758	1808	1858	1908	1958	2008
Temperature °C (y)	8.73	9.22	9.10	9.12	9.13	9.45	9.76

Tami creates a linear model for this data by finding the equation of the straight line passing through the points with coordinates (1708, 8.73) and (1958, 9.45).

- (a) Calculate the gradient of the straight line that passes through these two points. [2]
- (b) (i) Interpret the meaning of the gradient in the context of the question.
(ii) State appropriate units for the gradient. [2]
- (c) Find the equation of this line giving your answer in the form $y = mx + c$. [2]
- (d) Use Tami’s model to estimate the mean annual temperature in the year 2000. [2]

Thandizo uses linear regression to obtain a model for the data.

- (e) (i) Find the equation of the regression line y on x .
(ii) Find the value of r , the Pearson’s product-moment correlation coefficient. [3]
- (f) Use Thandizo’s model to estimate the mean annual temperature in the year 2000. [2]

Turn over

3. [Maximum mark: 17]

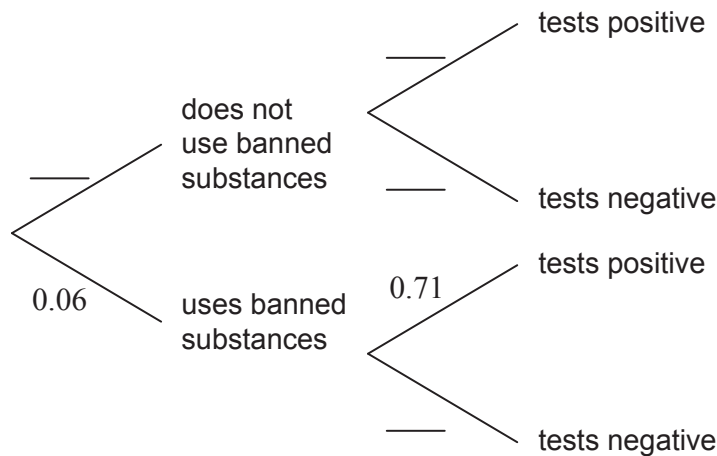
A large international sports tournament tests their athletes for banned substances. They interpret a positive test result as meaning that the athlete uses banned substances. A negative result means that they do not.

The probability that an athlete uses banned substances is estimated to be 0.06.

If an athlete **uses** banned substances, the probability that they will test positive is 0.71.

If an athlete does **not use** banned substances, the probability that they will test negative is 0.98.

- (a) Using the information given, **copy** (into your answer booklet) and complete the following tree diagram. [2]



- (b) (i) Determine the probability that a randomly selected athlete does not use banned substances and tests negative.
- (ii) If two athletes are selected at random, calculate the probability that both athletes do not use banned substances and both test negative. [4]
- (c) (i) Calculate the probability that a randomly selected athlete will receive an **incorrect** test result.
- (ii) A random sample of 1300 athletes at the tournament are selected for testing. Calculate the expected number of athletes in the sample that will receive an incorrect test result. [5]

Team X are competing in the tournament. There are 20 athletes in this team. It is known that none of the athletes in Team X use banned substances.

- (d) Calculate the probability that none of the athletes in Team X will test positive. [4]
- (e) Determine the probability that more than 2 athletes in Team X will test positive. [2]

Turn over

5. [Maximum mark: 15]

Goran is interested in the number of sightings of a particular bird each week in the 50 weeks following the first day of September. He collects some data which is shown in the table.

Number of sightings	0	1	2	3	4	5	More than 5
Number of weeks	8	16	13	8	3	2	0

The sample mean number of sightings per week for this data is 1.76.

- (a) Calculate the unbiased estimate of the population variance of sightings per week. [3]

Goran believes that the data follows a Poisson distribution.

- (b) State why your answer to part (a) supports Goran's belief. [1]

Goran decides to test at the 5% significance level to see if his belief is correct.

His null hypothesis is $X \sim \text{Po}(m)$, where the random variable, X , is defined as the number of sightings per week.

Goran estimates parameter m to be the mean of the sample, 1.76. He calculates the expected frequencies for sightings per week in the 50 weeks after the first day of September. These are shown to two decimal places in the following table.

Number of sightings	0	1	2	3	4	5 or more
Expected frequencies	8.60	15.14	13.32	7.82	j	k

- (c) Find the value of
- j ;
 - k . [5]
- (d) State a reason why Goran should combine groups to conduct his significance test. [1]
- (e) Write down the degrees of freedom for the test. [1]
- (f) Find the p -value for the test. [2]
- (g) State the conclusion of the test. Justify your answer. [2]

5. [Maximum mark: 14]

The three countries of Belgium, Germany and The Netherlands meet at a single point called Vaalserberg.

To support future transport planning, a 10 km circle was drawn around Vaalserberg on a map. A study was carried out over five years to determine what percentage of people living in each of these countries (within the 10 km circular region) either stayed in their own country or moved to another country within the circle.

From this study, the following movements were observed during the five years.

- From Belgium, 5 % moved to Germany, and 0.5 % moved to The Netherlands.
- From Germany, 2 % moved to The Netherlands, and 1.5 % moved to Belgium.
- From The Netherlands, 3 % moved to Germany, and 2 % moved to Belgium.

All additional population changes within the circular region may be ignored.

- (a) Represent the above information in a transition matrix T . [3]

At the end of the study, the population of the Belgian side was 26 000, the population of the German side was 240 000, and the population of The Netherlands side was 50 000.

- (b) By using T , find the expected population of the German side of Vaalserberg 30 years after the end of the study. [4]

For matrix T there exists a steady state vector

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix},$$

where u_1 , u_2 and u_3 are the proportions of the total population on the Belgian side, the German side and The Netherlands side respectively.

The steady state vector $\mathbf{u}$ may be found by solving a system of equations.

- (c) (i) Determine these equations that are to be solved.
- (ii) By solving your system of equations, find $\mathbf{u}$. [3]
- (d) Use your answer to part (c)(ii) to determine the long-term expected population of the German side. [2]
- (e) Suggest two reasons why your answer to part (d) is not likely to be accurate. You may comment on both the model and the situation in context. [2]

6. [Maximum mark: 18]

The gardener in a local park suggested that the number of snails found in the park can be modelled by a Poisson distribution.



- (a) Suggest two observations that the gardener may have made that led him to suggest this model. [2]

Now assume that the model is valid and that the mean number of snails per m^2 is 0.2. The gardener inspects, at random, a 12m^2 area of the park.

- (b) Find the probability that the gardener finds exactly four snails. [3]
- (c) Find the probability that the gardener finds fewer than three snails. [2]
- (d) Find the probability that, in three consecutive inspections, the gardener finds at least one snail per inspection. [3]

Following heavy rain overnight, the gardener wished to determine whether the number of snails found in a random 12m^2 area of the park had increased.

- (e) State the hypotheses for the test. [2]
- (f) Find the critical region for the test at the 1 % significance level. [3]
- (g) Given that the mean number of snails per m^2 has actually risen to 0.75, find the probability that the gardener makes a Type II error. [3]

Turn over

2. [Maximum mark: 29]

This question considers the analysis of several datasets of examination marks using a variety of standard procedures and also an unfamiliar statistical test.

A class of eight students sits two examinations, one in French and one in German.

The marks in these examinations are given in **Table 1**.

Table 1

Student	French mark	German mark
S_1	42	39
S_2	65	66
S_3	82	71
S_4	50	53
S_5	48	32
S_6	73	59
S_7	34	40
S_8	59	56

The maximum mark in both examinations is the same.

You may assume that these data are a random sample from a bivariate normal distribution with mean μ_F for the French examination, mean μ_G for the German examination and Pearson's product-moment correlation coefficient ρ .

Before the examinations were sat, the Head of Languages, Pierre, decided to investigate whether there would be significant evidence of a difference between μ_F and μ_G . He decided to analyse the data using a two-tailed paired t -test with significance level 5%.

(a) Explain briefly

(i) why he chose to use a t -test and not a z -test; [1]

(ii) why he chose to use a two-tailed test and not a one-tailed test. [1]

(b) (i) State suitable hypotheses for the t -test. [1]

(ii) Find the p -value for this test. [2]

(iii) The p -value is a probability. State the event for which it gives the probability. [1]

(iv) State, giving a reason, what conclusion Pierre should reach. [2]

(This question continues on the following page)

(Question 2 continued)

- (c) Pierre believes that students who score well in one language examination tend to score well in the other language examination. He therefore decides to carry out a test at the 5 % significance level to investigate whether there is a positive correlation between the French examination marks and the German examination marks.
- (i) State appropriate hypotheses in terms of ρ . [1]
- (ii) Perform a suitable test and state the p -value. State, in context, the conclusion that Pierre should reach, giving a reason. [4]
- (d) There are actually two more students in this particular class, Paul and Sue. Paul sat the French examination, but he was unable to sit the German examination. Sue sat the German examination, but she was unable to sit the French examination.
- (i) Paul’s mark in the French examination was 58. Use the data in **Table 1** to predict the mark that Paul would have obtained in his German examination. [3]
- (ii) Based on her mark in the German examination, Sue’s mark in the French examination was predicted to be 71. Find the mark she obtained in the German examination. [2]

Six students sit examinations in mathematics and history and their marks are shown in **Table 2**. The Vice Principal, Angela, decides to investigate whether there is any association between the marks obtained in these two subjects.

Table 2

Student	Mathematics mark (x)	History mark (y)
P_1	53	41
P_2	76	70
P_3	50	62
P_4	65	47
P_5	61	66
P_6	84	50

(This question continues on the following page)

(Question 2 continued)

Angela is informed that the maximum mark in each subject is 100.

Angela believes that the data might not be normally distributed, so she investigates what suitable tests are available which do not assume the data is normally distributed. She decides to use an unfamiliar test based on a statistic called Kendall's τ .

Consider n bivariate observations (x_i, y_i) , $i = 1, 2, \dots, n$, such that there are no equal x -values and no equal y -values. Any pair of distinct bivariate observations (x_i, y_i) and (x_j, y_j) is said to be concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$. For n bivariate observations, there are $\frac{n(n-1)}{2}$ distinct pairs.

Kendall's τ is defined as $\frac{2(C-D)}{n(n-1)}$ where C and D denote respectively the number of concordant pairs and discordant pairs.

- (e) (i) Show that the value of Kendall's τ always lies in the interval $[-1, +1]$. [1]
- (ii) For students P_1 and P_2 , show that their pair is concordant. [1]
- (iii) Show that the value of Kendall's τ for the mathematics and history data is 0.2. [4]
- (f) Angela decided to use this statistic in a two-tailed test at the 10% significance level. The critical region for her test is $|\tau| \geq 0.733$.
- (i) State, in words, her null and alternative hypotheses. [1]
- (ii) State the conclusion that Angela should reach. Give a reason for your answer. [2]

Angela now finds that the history marks are actually out of 120. The history teacher advises Angela to scale the history marks so that they are out of 100 and then redo the calculations for the value of τ .

- (g) State, with a reason, whether you agree with the advice given by the history teacher. [2]

References:

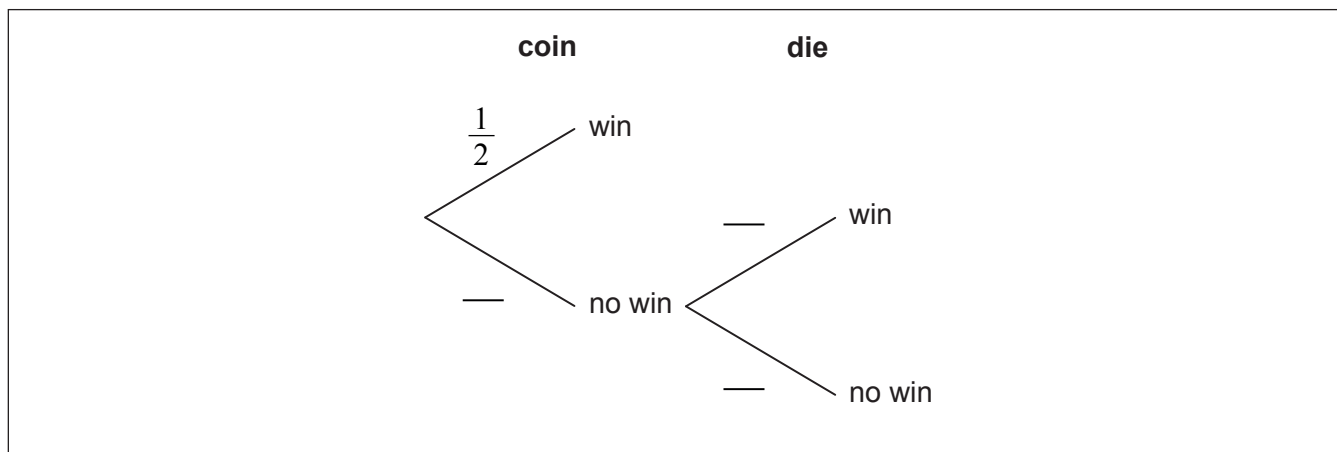
© International Baccalaureate Organization 2023

7. [Maximum mark: 7]

Michèle is playing a game. In the game, she must first flip a fair coin which will result in the coin landing on heads or tails. If the coin lands on heads, then she wins a prize. If it lands on tails, then she has another chance but this time she must roll a fair six-sided dice and get a five or six in order to win a prize.

- (a) Complete the tree diagram by writing in the three missing probabilities.

[2]



- (b) Find the probability that Michèle does **not** win a prize.

[2]

- (c) Given that Michèle won a prize, find the probability that the coin landed on heads.

[3]

[illegible]

10. [Maximum mark: 7]

The decay of a chemical isotope over five years is recorded in **Table 1**. The mass of the chemical M is measured to the nearest gram at the beginning of each year t of the experiment.

Table 1

Time t (years)	1	2	3	4	5
Mass M (grams)	1000	660	517	435	381

It is believed that the decay of the isotope can be modelled by an equation of the form $M = a \times t^b$.

- (a) Use power regression on your graphic display calculator to find the value of a and the value of b .

[2]

The values of t and M can be transformed such that $x = \ln t$ and $y = \ln M$. **Table 2** shows data for x and y to three decimal places.

Table 2

x	0	0.693	1.099	1.386	1.609
y	6.908	6.492	6.248	6.075	5.943

- (b) Find the linear regression equation of y on x , in the form $y = cx + d$. Give the values of c and d to three decimal places.
- (c) Hence, show that this linear regression is equivalent to the power regression found in part (a).

[2]

[3]

(This question continues on the following page)



[illegible]

Apollo | Original source crop | 98

15. [Maximum mark: 7]

The eating habits of students in a school are studied over a number of months. The focus of the study is whether non-vegetarians become vegetarians, and whether vegetarians remain vegetarians.

Each month, students choose between the vegetarian or non-vegetarian lunch options. Once they have chosen for the month, they cannot change the option until the next month.

In any month during the study, it is noticed that the probability of a non-vegetarian becoming vegetarian the following month is 0.1, and that the probability of a vegetarian remaining a vegetarian the following month is 0.8.

This situation can be represented by the transition matrix

$$T = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix}.$$

- (a) Interpret the value 0.9 in T in terms of the changes in the eating habits of the students in the school. [1]
- (b) Find the eigenvalues of matrix T . [3]

One of the eigenvectors of T is $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

- (c) Find another, non-parallel, eigenvector and interpret it in context. [3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

16. [Maximum mark: 6]

When Jef plays basketball, the number of shots he takes during any 6 minutes of play can be modelled by a Poisson distribution with mean 2.5.

- (a) Find the probability that Jef takes less than 7 shots during any 12 minutes of play. [2]

It can be assumed that the outcomes of the shots are independent of each other, and the probability of success of a shot is constant. The probability that Jef is successful with a shot is 0.4.

It can be assumed that the probability of Jef’s success with a shot is independent of the number of shots that he takes.

- (b) Find the probability that during any 6 minutes of play Jef takes fewer than 4 shots and is successful at least once. [4]



Disclaimer:

Content used in IB assessments is taken from authentic, third-party sources. The views expressed within them belong to their individual authors and/or publishers and do not necessarily reflect the views of the IB.

References:

3. Nina Aldin Thune. https://en.wikipedia.org/wiki/Great_Pyramid_of_Giza#/media/File:Kheops-Pyramid.jpg. Licensed under CC BY 2.5 <https://creativecommons.org/licenses/by/2.5/#>. Image adapted.

All other texts, graphics and illustrations © International Baccalaureate Organization 2023



Please **do not** write on this page.

Answers written on this page
will not be marked.



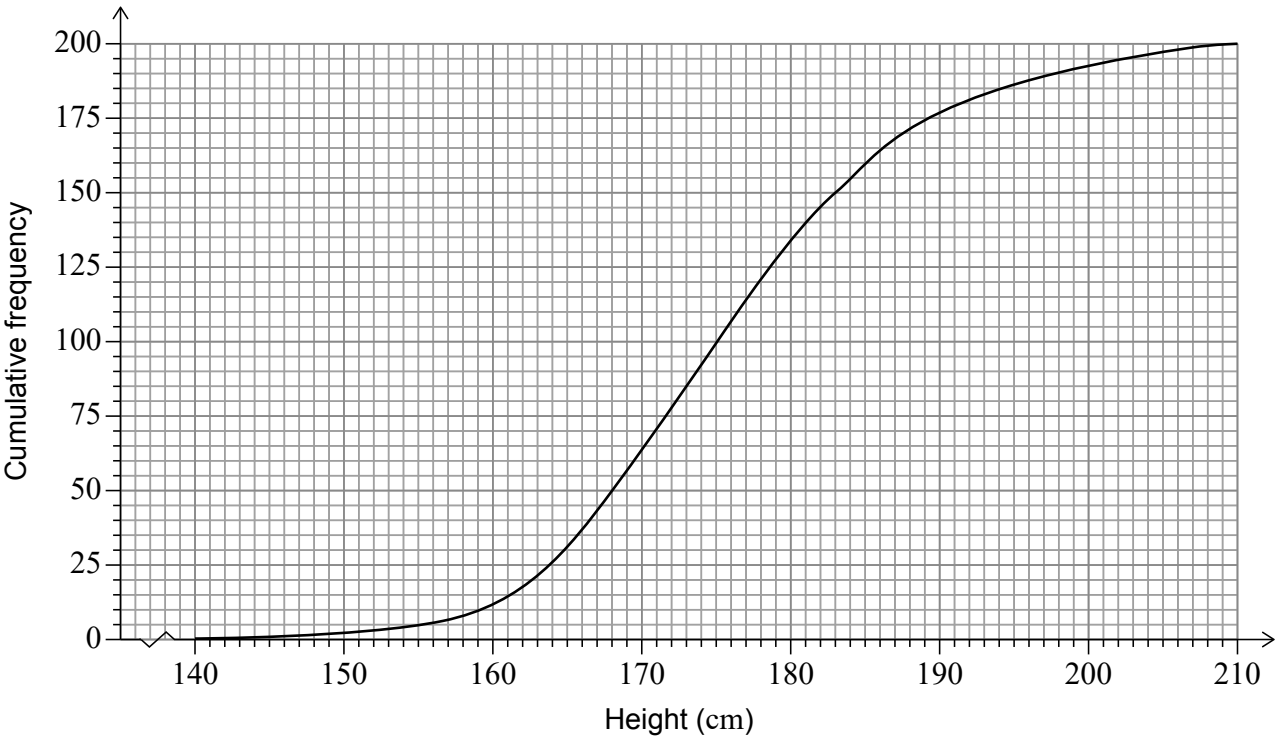
2. [Maximum mark: 18]

The heights, h , of 200 university students are recorded in the following table.

Height (cm)	Frequency
$140 \leq h < 160$	11
$160 \leq h < 170$	51
$170 \leq h < 180$	68
$180 \leq h < 190$	47
$190 \leq h < 210$	23

- (a) (i) Write down the mid-interval value of $140 \leq h < 160$.
- (ii) Calculate an estimate of the mean height of the 200 students. [3]

This table is used to create the following cumulative frequency graph.



- (b) Use the cumulative frequency curve to estimate the interquartile range. [2]

Laszlo is a student in the data set and his height is 204 cm.

- (c) Use your answer to part (b) to estimate whether Laszlo's height is an outlier for this data. Justify your answer. [3]

(This question continues on the following page)

(Question 2 continued)

It is believed that the heights of university students follow a normal distribution with mean 176 cm and standard deviation 13.5 cm.

It is decided to perform a χ^2 goodness of fit test on the data to determine whether this sample of 200 students could have plausibly been drawn from an underlying distribution $N(176, 13.5^2)$.

(d) Write down the null and the alternative hypotheses for the test. [2]

As part of the test, the following table is created.

Height of student (cm)	Observed frequency	Expected frequency
$h < 160$	11	23.6
$160 \leq h < 170$	51	42.1
$170 \leq h < 180$	68	a
$180 \leq h < 190$	47	46.7
$190 \leq h$	23	b

(e) (i) Find the value of a and the value of b .
(ii) Hence, perform the test to a 5 % significance level, clearly stating the conclusion in context. [8]

7. [Maximum mark: 17]

The city of Melba has an adult population of four million people. It is assumed that the weights of adults in Melba can be modelled by a normal distribution with mean 72 kg and standard deviation 10 kg.

- (a) If 10 adults in Melba are chosen independently and at random, find the probability that more than 3 of them have a weight greater than 85 kg.

[4]

Laetitia runs a travel agency in Melba. The elevator to her office is designed to hold a maximum of 8 people.

- (b) Write down a probability distribution that models the total weight of 8 adults chosen independently and at random from Melba.

[3]

The total weight of 8 adults exceeds w on less than 1 % of all occasions that 8 adults enter the elevator.

- (c) Find the value of w .

[2]

A newspaper claims that 42 % of the adults in Melba who go on holiday choose to go abroad. Laetitia believes that this is an overestimation of the true number. During the past month, Laetitia found that 67 of her clients chose a holiday abroad, and 133 chose a holiday that was not abroad.

- (d) Laetitia decides to perform a test using the binomial distribution on her data for the population proportion, p , that go on holiday abroad.
- (i) State **two** assumptions that Laetitia makes in order to conduct the test.
 - (ii) Write down the null and the alternative hypotheses for Laetitia's test, in terms of p .
 - (iii) Using the data from Laetitia's sample, perform the test at a 5 % significance level to determine whether Laetitia's belief is reasonable.

[8]

2. [Maximum mark: 27]

This question is about applying ideas from logarithms, calculus and probability to an unfamiliar mathematical theory called information theory.

Claude Shannon developed a mathematical theory called information theory to measure the information gained when random events occur. He defined the information, I , that is gained when an event with probability p occurs as

$$I = -\ln p$$

where $0 < p \leq 1$. For example, no information is gained ($I = 0$) when an event is certain to occur ($p = 1$).

- (a) (i) Sketch the graph of $I = -\ln p$, for $0 < p \leq 1$, labelling all axes intercepts and asymptotes. [3]
- (ii) Show, using calculus, that I is a decreasing function of p . [3]
- (iii) Interpret what “ I is a decreasing function of p ” means in the given context. [1]
- (b) A computer selects at random an integer x from 1 to 10, inclusive. Each outcome is equally likely.

Alessia is trying to determine the value of x and asks if x is odd.

- (i) Write down the probability that x is odd. [1]
- (ii) Alessia is told that x is odd. Find how much information Alessia gains. [2]

The computer then selects at random an integer y from 1 to 10, inclusive. Each outcome is equally likely.

Daniel is trying to determine the value of y and asks if y is 7. He is told that it is not 7.

- (iii) Find how much information Daniel gains. [2]

(This question continues on the following page)

Turn over

(Question 2 continued)

If a random variable has n possible outcomes with probabilities $p_1, p_2 \dots p_n$, then the expected information gained, $E(I)$, is defined as

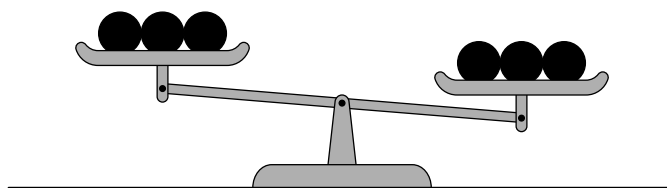
$$E(I) = \sum_{r=1}^n -p_r \ln p_r.$$

- (c) For the integer guessing game described in part (b), when Daniel asks if y is 7, there are two possible outcomes: “ y is 7” or “ y is not 7”.
- (i) Show that the expected information gained by Daniel is 0.325, correct to three significant figures. [2]
- (ii) Alessia asks if x is odd. Show that her expected information gained is greater than Daniel’s expected information gained. [2]

Information theory can be applied to a variety of situations.

- (d) When a coin is flipped, the outcome is either heads or tails. The coin may be biased. Let p be the probability of the outcome being heads.
- (i) Find, in terms of p , the information gained when the outcome is tails. [1]
- (ii) Find, in terms of p , the expected information gained when the coin is flipped once. [1]
- (iii) Hence, find the value of p when the expected information gained is maximized. [2]

A famous puzzle uses 12 balls which appear identical. 11 have the same weight, but one is either lighter or heavier than the others. A pair of scales can be repeatedly used to compare the weights of different combinations of the balls.



The outcome of each weighing can be “balanced”, “left-hand side heavier” or “right-hand side heavier”. The aim of the puzzle is to identify the ball which is the different weight, and whether it is heavier or lighter than the others, in as few weighings as possible.

(This question continues on the following page)

(Question 2 continued)

- (e) Angela wants to decide how many balls should be compared to each other in the **first weighing**. She produces the following table to help plan her strategy.

Number of balls on each side	Probability of balanced	Probability of left-hand side heavier	Probability of right-hand side heavier	Expected information, $E(I)$ (3 decimal places)
1	$\frac{5}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	0.566
2	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	z
3	x	y	y	1.040
4	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1.099
5	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{5}{12}$	1.028
6	0	$\frac{1}{2}$	$\frac{1}{2}$	0.693

- (i) Find the value of x . Justify your answer.

[2]
- (ii) Find the value of y .

[2]
- (iii) Find the value of z .

[2]
- (iv) Use the table to suggest the best choice for Angela’s first weighing. Justify your answer.

[1]
-

2. [Maximum marks: 7]

After taking a mathematics test, Fatima wonders how many more marks she would have achieved if she had spent an extra 1.5 hours studying.

To find out, she randomly selects five students from her class who took the same test and asks them how many hours (t) they spent studying for the test and the marks (m) they achieved. Their responses are shown in the following table.

Hours, t	0	1.2	1.6	2.5	4
Marks, m	45	54	61	72	86

Fatima believes there might be a linear relationship between the time spent studying and the results obtained.

- (a) (i) Find the Pearson’s product moment correlation coefficient, r , for this data.
(ii) Find the least squares regression line of m on t for this data. [4]
- (b) According to her model, find how many more marks Fatima would have achieved if she spent an extra 1.5 hours studying. [2]
- (c) State one reason why the value obtained in part (b) might not be valid. [1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



3. [Maximum mark: 6]

The formula $F = 1.8C + 32$ is used to convert a temperature in degrees Celsius, C , to degrees Fahrenheit, F .

- (a) (i) Find a formula for converting a temperature in degrees Fahrenheit to degrees Celsius.
- (ii) Find the temperature in degrees Celsius that is recorded as 77 degrees Fahrenheit. [3]

Over one year, the mean daily temperature in Mexico City was calculated to be 17 degrees Celsius with a standard deviation of 9 degrees Celsius.

- (b) For the same year, find in degrees Fahrenheit
 - (i) the mean daily temperature in Mexico City.
 - (ii) the standard deviation of the daily temperature in Mexico City. [3]



Turn over



2227-1201

Please **do not** write on this page.

Answers written on this page
will not be marked.



12. [Maximum mark: 8]

Roma is told by the manufacturers of their phone that, when charging from 0%, the percentage P of charge can be modelled by $P = 100 - 100 \times 2^{-t}$, where t is the number of hours since the phone started charging.

- (a) Find the time taken for the amount of charge to go from 0% to 40%, according to this model.

[2]

Roma decides to test the model. They charge the phone from 0% and record the charge of the phone for different values of t . The results are shown in the following table.

t hours	1	2	3	4
P percent	48	74	86	91

- (b) Find the sum of the square residuals (SS_{res}) when using $P = 100 - 100 \times 2^{-t}$ to fit this data. [4]

Roma now uses a quadratic function to model the charge in the phone. They calculate that, when using this model, the sum of the square residuals for these four points is 2.45.

- (c) State one reason why Roma might prefer to use

(i) the quadratic model.

(ii) $P = 100 - 100 \times 2^{-t}$.

[2]



Turn over

17. [Maximum mark: 9]

Phoebe opens a coffee shop, near to a well-established Apollo coffee shop.

After being open for a few months, Phoebe notices that

- 10% of customers who preferred the Apollo coffee shop in one month preferred her coffee shop the following month.
- 25% of customers who preferred her coffee shop in one month preferred the Apollo coffee shop the following month.

She decides to show these changes in the following transition matrix.

$$\begin{pmatrix} 0.9 & 0.25 \\ 0.1 & 0.75 \end{pmatrix}$$

The two eigenvalues for this matrix are 1 and 0.65. An eigenvector corresponding to the eigenvalue of 1 is $\begin{pmatrix} 5 \\ 2 \end{pmatrix}$.

- (a) Find an eigenvector corresponding to the eigenvalue of 0.65. [2]

A diagonal matrix of eigenvalues is $\mathbf{D} = \begin{pmatrix} 0.65 & 0 \\ 0 & 1 \end{pmatrix}$.

- (b) Write down an expression for $\mathbf{D}^n$, giving your answer as a 2×2 matrix in terms of n . [1]

When Phoebe's coffee shop first opened, the Apollo shop had 7000 customers the previous month.

- (c) Assuming all 7000 customers continue to go to one of these coffee shops, find an expression for the number that will favour Phoebe's coffee shop after n months. [6]

(This question continues on the following page)



[illegible]

Disclaimer:

Content used in IB assessments is taken from authentic, third-party sources. The views expressed within them belong to their individual authors and/or publishers and do not necessarily reflect the views of the IB.

References:

7. Saddako, n.d. *Cheetah (Acinonyx jubatus) Running - stock photo*. [image online] Available at: <https://www.gettyimages.co.uk/detail/photo/cheetah-running-royalty-free-image/523244194?phrase=cheetah+speed&adppopup=true> [Accessed 2 May 2023]. Source adapted.
- GlobalP, n.d. *Lion, Panthera leo, 8 years old, standing - stock photo*. [image online] Available at: <https://www.gettyimages.co.uk/detail/photo/lion-panthera-leo-8-years-old-standing-royalty-free-image/134976936?phrase=Lion+standing&adppopup=true> [Accessed 2 May 2023]. Source adapted.

All other texts, graphics and illustrations © International Baccalaureate Organization 2024



Please **do not** write on this page.

Answers written on this page
will not be marked.



Please **do not** write on this page.

Answers written on this page
will not be marked.



2. [Maximum mark: 6]

Jerry makes handcrafted chocolates. On average, 1 in 25 of the chocolates that Jerry makes is flawed. Whether or not a chocolate is flawed is independent of all other chocolates.

- (a) In a batch of 20 chocolates, chosen at random, find the probability that
- (i) two are flawed.
 - (ii) more than two are flawed. [4]

Jerry sells the perfect chocolates for 50 pesos each and the flawed ones for 15 pesos each.

- (b) Calculate the expected number of pesos Jerry makes from selling a batch of 20 randomly selected chocolates. [2]



4. [Maximum mark: 7]

Gustav plays a game in which he first tosses an unbiased coin and then rolls an unbiased six-sided die.

If the coin shows tails, the score on the die is Gustav's final number of points.

If the coin shows heads, one is added to the score on the die for Gustav's final number of points.

- (a) Find the probability that Gustav's final number of points is 7. [2]
- (b) Complete the following table. [3]

Final number of points	1	2	3	4	5	6	7
Probability							

- (c) Calculate the expected value of Gustav's final number of points. [2]

[illegible]

9. [Maximum mark: 6]

The number of traffic accidents at a road junction is modelled by a Poisson distribution with a mean of 0.76 accidents per week.

- (a) Under this model, calculate the probability that
 - (i) there are at least 2 accidents in a particular week.
 - (ii) there will be exactly 3 accidents in a particular 4-week period. [4]

The local traffic authority wishes to determine the probability that, in an 8-week period, fewer than 2 accidents occur in a week on exactly 5 occasions. It assumes that the weekly occurrence of accidents is independent of the week in which these occur.

- (b) State the appropriate model that the traffic authority should use to determine this probability. [2]



Turn over

14. [Maximum mark: 6]

A cup of hot water is placed in a room and is left to cool for half an hour. Its temperature, measured in °C, is recorded every 5 minutes. The results are shown in the table.

Time (mins)	5	10	15	20	25	30
Temperature (°C)	59	52	46	40	35	29

Akira uses the power function $T(t) = at^b + 25$ to model the temperature, T , of the water t minutes after it was placed in the room.

- (a) State what the value of 25 represents in this context. [1]
- (b) Use your graphic display calculator to find the value of a and of b . [3]

Soo Min models the temperature, T , of the water t minutes after it was placed in the room as $T(t) = kc^t + 25$.

- (c) Find the value of k and of c . [1]
- (d) State a reason why Soo Min's model of the temperature is a better fit for the data than Akira's model. [1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



5. [Maximum mark: 16]

The drivers of a delivery company can park their vans overnight either at its headquarters or at home.

Urvashi is a driver for the company. If Urvashi has parked her van overnight at headquarters on a given day, the probability that she parks her van at headquarters on the following day is 0.88. If Urvashi has parked her van overnight at her home on a given day, the probability that she parks her van at home on the following day is 0.92.

- (a) Write down a transition matrix, T , that shows the movement of Urvashi's van between headquarters and home. [2]

On Monday **morning** she collected her van from headquarters where it was parked overnight.

- (b) Find the probability that Urvashi's van will be parked at home at the end of the week on Friday **evening**. [3]
- (c) Write down the characteristic polynomial for the matrix T . Give your answer in the form $\lambda^2 + b\lambda + c$. [2]
- (d) Calculate eigenvectors for the matrix T . [4]
- (e) Write down matrices P and D such that $T = PDP^{-1}$, where D is a diagonal matrix. [2]
- (f) Hence find the long-term probability that Urvashi's van is parked at home. [3]

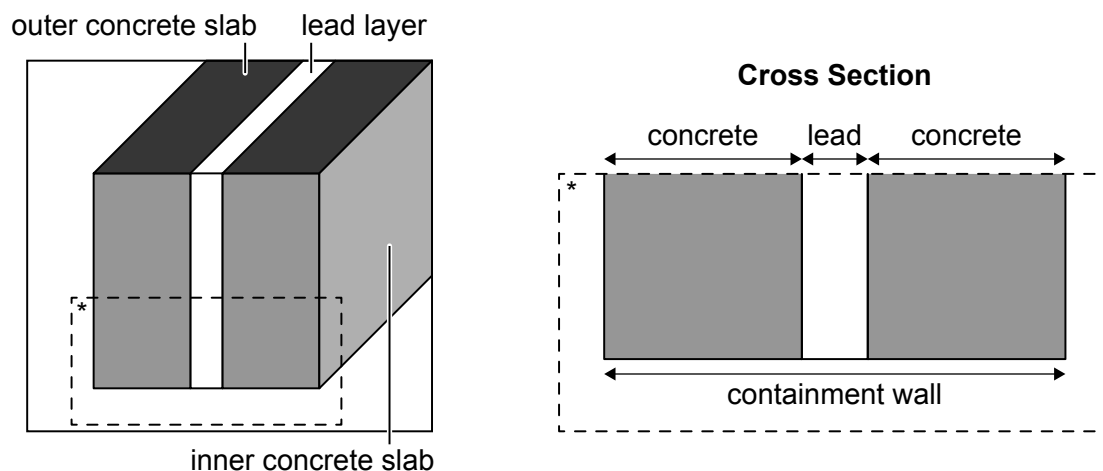
Turn over

Blank page

7. [Maximum mark: 17]

Containment walls to protect against radiation are constructed from two parallel concrete slabs that have a layer of lead between them as shown in the diagram.

diagram not to scale



The width of a concrete slab is modelled by a normal distribution with mean 350 mm and standard deviation 10 mm.

- (a) Find the probability that a randomly chosen concrete slab is less than 340 mm in width. [2]
- (b) Find the endpoints of the interval, symmetric about the mean, such that 95 % of the slabs have a width that lies in this interval. [3]

Stephen assumes the lead layer is also modelled by a normal distribution, but with mean 100 mm and standard deviation 5 mm and is independent of the width of the slabs.

Let W be the random variable that represents the total width of the wall, measured in mm.

- (c) (i) Given that the widths of any two concrete slabs are independent, calculate Stephen's value for the mean and standard deviation of W . [7]
- (ii) Hence find $P(780 < W < 810)$.

(This question continues on the following page)

(Question 7 continued)

There are concerns that the mean and standard deviation for Stephen's model of the lead layer are incorrect. However, his assumption that the model is normal and the width of the lead is independent of the width of the concrete slabs still holds.

On investigation it is found that the total width of the containment wall is normally distributed with mean 810 mm and standard deviation 16 mm. The model for the width of a concrete slab does not change.

(d) Use the results for the **sum** of independent random variables to find a revised value for

(i) the mean of the width of the lead layer.

(ii) the standard deviation of the width of the lead layer.

[4]

Under this revised model, 80% of the lead layers have a width less than k mm.

(e) Calculate the value of k .

[1]

Disclaimer:

Content used in IB assessments is taken from authentic, third-party sources. The views expressed within them belong to their individual authors and/or publishers and do not necessarily reflect the views of the IB.

References:

3. Andyqwe, n.d. *Dumpster truck* [image online] Available at: <https://www.gettyimages.co.uk/detail/photo/dumpster-truck-royalty-free-image/157611454> [Accessed 18 April 2023] Source adapted.

All other texts, graphics and illustrations © International Baccalaureate Organization 2024

1. [Maximum mark: 26]

The purpose of this question is to help a company decide whether or not they should use a new technique to make a component.

A factory produces components for a tractor. They have designed a new technique to produce one of their components that they hope will increase its useful lifespan.

They test 120 components made with the new technique and 240 with the technique they currently use. At the end of 250 hours of use, they check the components, and record whether they have no cracks, minor cracks or major cracks.

The data from the trial are given in the table.

	New technique	Current technique
No cracks	a	88
Minor cracks	54	96
Major cracks	b	56

In total 141 components had no cracks.

- (a) (i) Show that the value of a is 53. [1]
- (ii) Find the value of b . [1]

(This question continues on the following page)

(Question 1 continued)

One of the components from the trial is selected at random.

- (b) Given that this component had minor cracks find the probability that it was produced by the new technique. [2]
- (c) A χ^2 test for independence is performed at the 5% significance level to determine whether a component having no cracks, minor cracks or major cracks is independent of the production technique used.
- (i) State the null and alternative hypotheses. [1]
- (ii) Find the p -value. [2]
- (iii) State the conclusion of the test in context, justifying your answer. [2]
- (d) For the components in the trial that were made with the current technique, show that the proportion which developed cracks is $\frac{19}{30}$. [1]

As an alternative measure, the researchers decide to let p be the probability that a component, made with the new technique, develops cracks. They then test the following hypotheses

$$H_0: p = \frac{19}{30},$$

$$H_1: p < \frac{19}{30}.$$

In a randomly selected sample of 120 components made with the **new technique** let X be the number which developed cracks. The researchers assume that, under the null hypothesis, $X \sim B\left(120, \frac{19}{30}\right)$.

- (e) State one additional assumption that the researchers are making in choosing this distribution. [1]
- (f) Use appropriate data from the trial to perform the test proposed by the researchers, at the 5% significance level. State the conclusion of the test, justifying your answer. [5]
- (g) In comparison with the test in part (c), state one mathematical reason why
- (i) the test in part (f) might be preferred. [1]
- (ii) the test in part (f) might **not** be preferred. [1]

(This question continues on page 5)

Turn over

— T —

2227-1200

Blank page

(Question 1 continued)

For these components, the researchers also consider the mean time taken until cracks develop. It is hoped that using the new technique will increase this value. A second trial is carried out and the times, in hours, taken for cracks to appear is recorded.

The mean time taken for cracks to appear ($\bar{t}$) and the value of s_{n-1} for each technique are given in the following table.

	Number of components tested	$\bar{t}$ (hours)	s_{n-1} (hours)
New technique	100	250.1	3.8
Current technique	200	249.1	3.8

- (h) Perform an appropriate test at the 5 % significance level to determine whether the new technique increases the mean time taken for cracks to appear.
- [7]

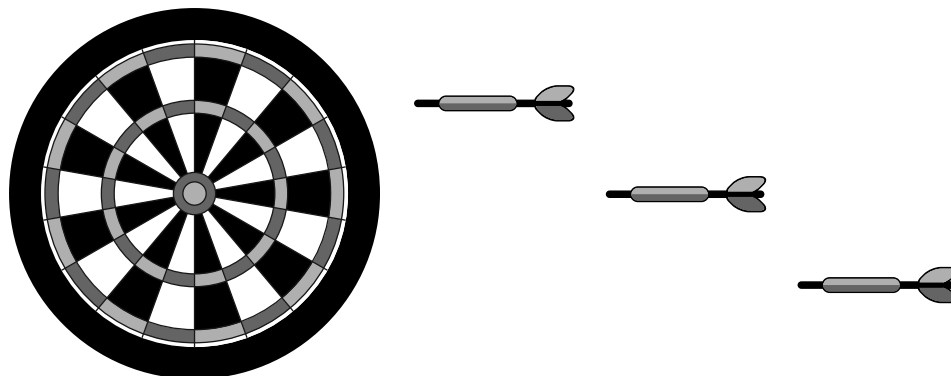
The company decides to go ahead with the new technique and publishes the following statement: “statistical tests show the new technique will significantly increase the time before components crack and need to be replaced”.

- (i) Comment on this statement.
- [1]

1. [Maximum mark: 24]

The following question examines the changes in darts players' scores using two statistical tests.

In the sport of darts, players take turns throwing darts at a board in order to score points.



A player's "three dart average" refers to the mean score achieved when throwing three darts.

Valia aimed to find out whether amateur darts players in her local area improved over a 12-month period. An increase in their "three dart average" would indicate an improvement.

She selected a random sample of eight darts players and recorded their mean "three dart average" from Year 1.

(This question continues on the following page)

(Question 1 continued)

She then recorded their mean “three dart average” from Year 2.

Valia’s results were as follows:

Table 1

Player	Year 1 mean	Year 2 mean
Justin	68.1	70.1
Fallon	72.2	72.0
Rob	65.8	65.0
Michael	73.7	74.9
Mensur	69.8	68.0
Deta	64.5	68.0
Meaghan	69.2	70.6
Peter	64.9	92.9

Valia calculated the median, quartiles and inter-quartile range for each year. The results are shown in **Table 2**.

Table 2

	Year 1 mean	Year 2 mean
Lower quartile	65.35	68
Median	68.65	a
Upper quartile	71	b
Inter-quartile range	5.65	c

- (a) Determine the values of a , b and c .

[3]
- (b) By comparing the results for both years summarized in **Table 2**, state one conclusion, in context, that Valia might be justified in making.

[1]

(This question continues on the following page)

Turn over

(Question 1 continued)

Valia then decided to analyse the data from **Table 1** using a one-tailed paired t -test at the 10% significance level to determine whether the players' averages have increased.

- (c)

State an assumption about the differences in means that is necessary in order for the test to be valid.

[1]
- (d)

(i)

State the null and alternative hypotheses for the test.

[2]

(ii)

Find the p -value.

[2]

(iii)

State the conclusion of the test in context, justifying your answer.

[2]
- (e)

State one way Valia could have reduced the chance of her making

(i)

a Type I error.

[2]

(ii)

a Type II error.

[2]

Valia was not sure the assumption made in part (c) was correct and hence thought the results obtained from her paired t -test may not be valid.

Following further research, Valia decided to use the Wilcoxon signed-rank test, which does not require the assumption she made in part (c).

For this test, the magnitudes of the differences between the Year 2 and Year 1 means are ranked from 1 to 8, with the ranks of the positive differences (P) and the ranks of the negative differences (N) separated in columns.

This is partially shown in the following table, which Valia constructs to perform the test.

Year 1 mean	Year 2 mean	Difference	P	N
68.1	70.1	2.0	A	
72.2	72.0	−0.2		1
65.8	65.0	−0.8		2
73.7	74.9	1.2	3	
69.8	68.0	−1.8		B
64.5	68.0	3.5	C	
69.2	70.6	1.4	D	
64.9	92.9	28	8	
		Total	$\sum P = 28$	$\sum N$

(This question continues on the following page)

(Question 1 continued)

(f) Determine the values of

(i) A , B , C and D . [3]

(ii) $\sum N$. [1]

For this test:

- the Wilcoxon signed-rank test statistic is T = the smaller value from a choice of $\sum P$ or $\sum N$.
- the null hypothesis is that the population's median for "three dart average" is the same in both years.

Valia chooses to carry out the test at the 5 % level of significance. From statistical tables, she determines that the critical region is $T \leq 5$.

(g) (i) State the alternative hypothesis H_1 for this test. [1]

(ii) Write down the value of the test statistic, T . [1]

(iii) Determine the conclusion of the test in context. [2]

(h) Suggest briefly how Valia could assess the reliability of her results for either test. [1]

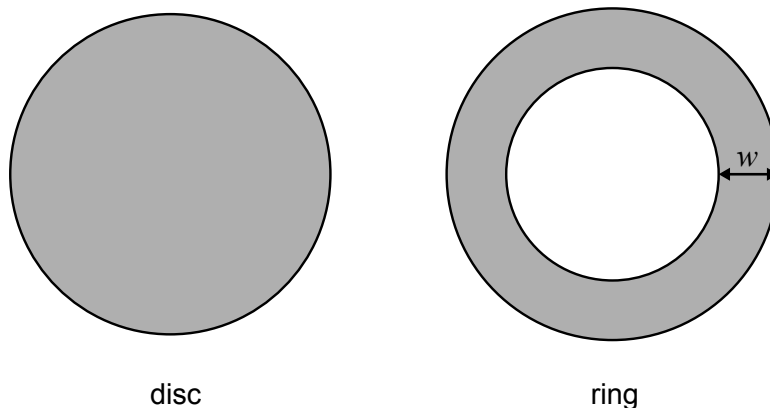
Turn over



15. [Maximum mark: 8]

A metal ring, of width w , is made by first cutting a circular disc and then cutting a circular hole exactly in the centre of this disc. This is shown in the diagram.

diagram not to scale



A machine produces many of these rings each day.

The diameters of the discs are normally distributed with mean 30 cm and standard deviation 0.8 cm.

- (a) Find the distribution of the radii of the discs. [3]

The radii of the **holes** are normally distributed with mean 12 cm and standard deviation 0.25 cm.

- (b) Calculate the probability that the width of a randomly chosen ring is less than 2.5 cm. [5]

(This question continues on the following page)



(Question 15 continued)

[illegible]

Turn over

1. [Maximum mark: 14]

A survey was answered by 20 000 expatriates (people living in a country that is not their own). The data ranked countries in order of the country they felt was best for expatriates. The highest-ranked country was Switzerland.

These results were compared to happiness scores taken from *The World Happiness Report 2022*. The following table shows this data for the top 10 expatriate countries.

Country	Switzerland	New Zealand	Spain	Australia	Cyprus	Portugal	Ireland	United Arab Emirates	France	Netherlands
Expatriate country rank	1	2	3	4	5	6	7	8	9	10
Happiness score	7.5	7.2	6.5	7.2	6.2	6.0	7.0	6.6	6.7	7.4

- (a) For the **happiness score**, find
- (i) the upper quartile

(ii) the interquartile range.
- [4]
- (b) Show that Switzerland’s happiness score is not an outlier for this data.
- [3]

(This question continues on the following page)

(Question 1 continued)

The happiness scores were ranked to calculate Spearman’s rank correlation coefficient, r_s . These ranks are shown in the following table.

Country	Switzerland	New Zealand	Spain	Australia	Cyprus	Portugal	Ireland	United Arab Emirates	France	Netherlands
Happiness score	7.5	7.2	6.5	7.2	6.2	6.0	7.0	6.6	6.7	7.4
Happiness rank	1	a	b	c	9	10	5	7	6	2

- (c)

Write down the value of

(i)

a

(ii)

b

(iii)

c .
- [3]
- (d)

(i)

Find r_s .

(ii)

If France's happiness score is upgraded to 6.9, explain why the value of r_s does not change.
- [3]
- Jose concludes from this data that countries with high happiness scores are likely to be favourite expatriate countries.
- (e)

State, with a reason, whether Jose’s conclusion is appropriate.
- [1]

1. [Maximum mark: 24]

In this question you will use graph theory and transition matrices to solve problems about a manager visiting five factories.

Audrey is the quality control manager for a manufacturing company that has five factories, A, B, C, D and E.

She is planning a route to visit each factory once, starting and finishing from her home, H.

She determines the distance between each location, in kilometres, as shown in the table.

	A	B	C	D	E	H
A		100	150	70	50	40
B	100		80	140	130	85
C	150	80		90	120	160
D	70	140	90		60	100
E	50	130	120	60		70
H	40	85	160	100	70	

Audrey wants to find an upper and lower bound for the shortest total distance travelled on her route.

- (a) Starting at H, use the nearest-neighbour algorithm to find an upper bound. [3]

To find a lower bound, Audrey uses the deleted vertex algorithm and deletes vertex H.

- (b) (i) Use Prim’s algorithm, starting at E, to find the weight of the minimum spanning tree for A, B, C, D and E. You should clearly state the order in which the edges are selected by the algorithm. [3]
- (ii) Hence, find a lower bound. [2]

(This question continues on the following page)

(Question 1 continued)

After her initial visit to all factories, Audrey now decides she will visit one factory each day. She decides which factory to visit according to the following transition matrix, T .

$$\begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{pmatrix} 0 & 0.3 & 0.1 & 0 & 0.1 \\ 0 & 0 & 0.2 & 0.2 & 0.4 \\ 0 & 0.2 & 0.2 & p & 0.2 \\ 0.7 & 0.25 & q & r & 0 \\ 0.3 & 0.25 & 0.2 & 0.2 & 0.3 \end{pmatrix} \end{matrix}$$

After visiting factory D, there is a probability of 0.4 that Audrey will visit factory C next.

- (c) Write down the value of
- (i) p [1]
 - (ii) q [1]
 - (iii) r . [1]
- (d) Audrey first visits factory A.
- (i) Write down the initial state matrix, S_0 . [1]
 - (ii) Find the probability that the fifth factory that Audrey visits is C. [2]
 - (iii) Find the probability that the fifth factory that Audrey visits is the same as the second factory she visits. [5]
- (e) Over a long period of time,
- (i) find the proportion of Audrey's visits that are to factory A [2]
 - (ii) find the expected distance Audrey travels in a day, given that she always travels directly from home to a factory and then back home. [3]

Turn over

2. [Maximum mark: 31]

In this question, you will first consider a statistical model for the number of fish caught in a lake and then consider a differential equation to model the growth of fish in the lake.

Althea enjoys fishing in a lake near her home. She usually fishes for 4 hours each day, and she records the length of each fish before releasing it back in the lake.

Althea decides to perform a χ^2 goodness of fit test, at the 5% significance level, to determine whether the number of fish caught can be modelled by a Poisson distribution.

She uses her records from the last 50 days to produce this table.

Number of fish caught per day	0	1	2	3	4	5	>5
Frequency	2	7	12	11	10	8	0

- (a) Calculate the mean number of fish caught per day. [2]
- (b) Hence, calculate the expected number of days, in a 50-day period, that Althea would catch no fish according to the Poisson model. [3]

Althea calculates the expected frequencies and records them in this table.

Number of fish caught per day	≤1	2	3	4	≥5
Expected frequency	10.9	11.6	11.2	8.05	8.25

- (c) Explain why Althea has combined the columns for 0 and 1. [1]
- (d) For this χ^2 goodness of fit test,

(i) write down a suitable null hypothesis [1]

(ii) write down the number of degrees of freedom [1]

(iii) find the p -value [3]

(iv) write down the conclusion, justifying your answer. [2]

(This question continues on the following page)

(Question 2 continued)

A factory is built on the edge of the lake, and Althea is concerned about the impact of pollution on the length of the fish. Based on her records, before the factory was built, the length of the fish could be modelled by a normal distribution with a mean of 50 cm.

After the factory is built, Althea collects a random sample of 15 fish and calculates that $\bar{x} = 48.6$ cm and $s_{n-1} = 2.56$ cm.

Althea decides to test if there is sufficient evidence of a decrease in the population's mean length. The hypotheses for her test are $H_0: \mu = 50$ and $H_1: \mu < 50$.

- (e) Perform the test, at the 5% significance level, to show that there is evidence of a decrease in the population's mean length. [3]

Althea now decides to consider the length of an individual fish in the lake as it grows. She researches the von Bertalanffy growth model, which states that

$$\frac{dL}{dt} = k(L_\infty - L),$$

where:

- L is the length of the fish (in centimetres) at time t (in months since birth)
- $L_\infty = \lim_{t \rightarrow \infty} L$, the length the fish approaches as it grows
- k is a constant.

- (f) Justifying your answer,
- (i) state whether k is positive, negative, or could be either [2]
- (ii) state the value of $\frac{dL}{dt}$ as $t \rightarrow \infty$ [2]
- (iii) state a parameter in the model that could have been affected by the factory pollution in the lake. [2]

At $t = 0$, a fish has length L_0 cm.

- (g) Sketch a graph of $\frac{dL}{dt}$ against t . You should label the coordinates of any axes intercepts and the equation of any asymptotes on your sketch. [3]
- (h) Solve the differential equation $\frac{dL}{dt} = k(L_\infty - L)$, using the initial condition. Write your answer in the form $L = f(t)$. [6]

9. [Maximum mark: 7]

In a year, Louis attends school for a total of 180 school days. Louis leaves his house at exactly 7:45 am every day and takes public transportation from his house to school.

His travel times to school are normally distributed with a mean of 35.5 minutes and a standard deviation of 5.2 minutes. The travel times are independent of each other.

He is late if he arrives at school after 8:30 am.

(a) Find the probability that Louis is late on at least 10 days in the year. [4]

Mrs. Beale is responsible for recording students who arrive late. However, Mrs. Beale does not always accurately mark lateness.

On days when Louis is late, there is a 35% chance that Mrs. Beale will still mark him as being on time. Mrs. Beale’s accuracy for one school day is independent of her accuracy for any other school day.

Louis will receive a detention if he is marked late to school at least 4 times during the year.

(b) Find the probability of Louis receiving a detention in the year. [3]



Turn over

3. [Maximum mark: 6]

Two judges, Brett and Clarence, rank the skill levels of eight sheepdogs in a competition. The sheepdogs are labelled A to H and the judges rank the dogs as shown in the table.

Rank	1	2	3	4	5	6	7	8
Brett	A	C	D	B	E	F	G	H
Clarence	A	B	D	C	E	G	F	H

- Write down the rank that Brett awards sheepdog B. [1]
- Calculate Spearman's rank correlation coefficient for these data. [4]
- Comment on your answer to part (b) in terms of the ranks awarded by Brett and Clarence. [1]

[illegible]

Turn over

6. [Maximum mark: 5]

The temperatures at 5 different locations on a coral reef were measured. The mean of this sample was 20.1°C and the standard deviation of this sample, s_n , was 3.2°C .

- (a) Find an unbiased estimate of the population variance.

[2]
- (b) Assuming that the temperatures are normally distributed, find a 95 % confidence interval for the mean temperature on the coral reef.

[2]
- (c) Using your answer to part (b), determine if it is plausible that the mean temperature on the coral reef could be 17°C .

[1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

13. [Maximum mark: 9]

A biologist uses a wire frame to count the number of worms in a 1 m^2 section.

She models the number of worms found in each 1 m^2 section as following a Poisson distribution with mean 1.2 .

- (a) Find the probability of observing exactly one worm in one 1 m^2 section. [1]
- (b) Find the probability of observing at least one worm in one 1 m^2 section. [1]

The biologist looks at 5 independent 1 m^2 sections.

- (c) Find the probability of observing a total of five worms in 5 sections. [2]
- (d) Find the probability of observing exactly one worm in all 5 sections. [2]
- (e) Find the probability of observing at least one worm in exactly 3 of the 5 sections. [3]



14. [Maximum mark: 6]

A vet wants to find a relationship between the age in days of a breed of puppy (d) and its weight in kg(w).

To do this he collects a large quantity of data and plots two graphs.

He finds the regression line for each graph. His results are summarized in the table.

	Horizontal axis	Vertical axis	Gradient	Intercept on vertical axis	R^2
Graph 1	d	$\ln w$	0.00571	1.54	0.72
Graph 2	$\ln d$	$\ln w$	0.302	0.693	0.95

Based on these results, find the best of the two possible relationships between w and d .

Express your relationship in the form $w = f(d)$ where f is a simplified expression.

Justify your choice of expression.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

3. [Maximum mark: 7]

Prakash is the leader of a customer service team and is interested in determining whether there is a relationship between a customer's satisfaction level and the type of service interaction they have experienced.

He collects data from a random sample of 250 customers and tracks their satisfaction level after three types of service interactions: in-person, online chat bots and website contact forms.

He categorizes the satisfaction levels as satisfied, neutral and dissatisfied.

He records the data in the following table.

		Satisfaction level		
		Satisfied	Neutral	Dissatisfied
Type of service interaction	In-person	35	30	23
	Online chat bots	31	39	23
	Website contact forms	19	28	22

Prakash performs a χ^2 test for independence at the 5% significance level.

The critical value is 9.488.

The null hypothesis, H_0 , is the satisfaction level and the type of service interaction are independent.

(a) State the alternative hypothesis for this test. [1]

(b) Find the degrees of freedom for this test. [1]

(c) Find χ^2_{calc} , the chi-squared test statistic. [2]

Prakash concludes that there is sufficient evidence to reject the null hypothesis.

(d) (i) State whether Prakash is correct. Justify your answer.

(ii) Write down the conclusion for this test in context. [3]

(This question continues on the following page)



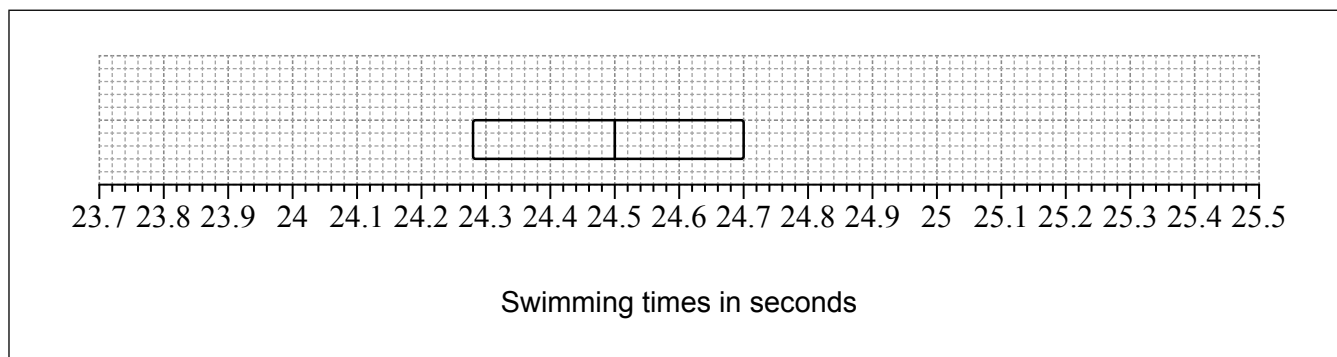
[illegible]

Turn over

4. [Maximum mark: 7]

The times, in seconds, for the fastest 16 women in a 50 m freestyle swimming championship event were recorded. All swimmers recorded different times.

Part of a box and whisker diagram for these times is shown in the following diagram.



- (a) Write down the number of swimmers who took more than 24.70 seconds to complete the race. [1]

- (b) Find the interquartile range (IQR) for the data. [2]

An outlier is defined as a value that satisfies one of the following:

- more than $1.5 \times \text{IQR}$ below the lower quartile
- more than $1.5 \times \text{IQR}$ above the upper quartile.

Of the 16 women, the two fastest swimmers took 23.96 and 24.12 seconds and the two slowest women took 25.12 and 25.40 seconds to complete the race.

- (c) (i) Show that only one of these times is an outlier.

- (ii) Complete the box and whisker diagram above. [4]

(This question continues on the following page)



This image shows a full page of white paper with horizontal dotted lines. The lines are evenly spaced and run across the width of the page, providing a guide for handwriting practice. There are no margins, text, or other markings on the page.



Apollo | Original source crop | 151

Please **do not** write on this page.

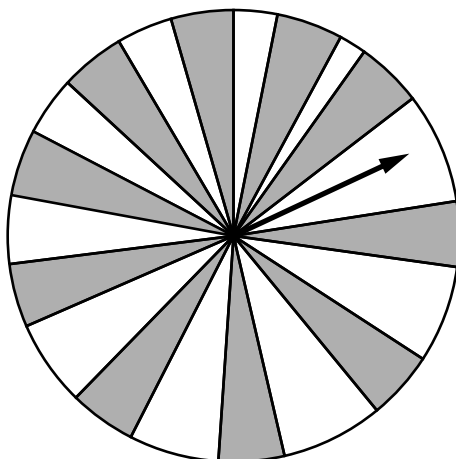
Answers written on this page
will not be marked.



6. [Maximum mark: 6]

A circular disk has 10 white sectors and 10 shaded sectors. The size of the angles, made at the centre of the circle by the white sectors, form an arithmetic sequence. The common difference of this sequence is 3° and the first term is 6° .

An arrow is attached to the centre of the circle as shown in the following diagram.



The arrow can be spun and has an equal chance of stopping at any point.

- (a) Find the probability that the arrow will stop in a white sector. [3]

In a game, the arrow is spun 4 times and the player wins if the arrow stops in a white sector at least 3 times.

- (b) Find the probability that the player wins the game. [3]

(This question continues on the following page)



7. [Maximum mark: 7]

A team of researchers is using a model to predict the relative happiness of different countries. To do this, a value x is calculated based on easily measured parameters, for example, life expectancy, or available social support. It is assumed that higher values of x indicate greater happiness.

To test the model a survey is conducted in six countries, A, B, C, D, E and F. In these countries the level of happiness is assessed directly using questionnaires and given a score y , out of 10, with higher scores indicating greater happiness.

To select the countries for the survey, all countries are divided into three equal groups based on wealth and two countries are chosen randomly from each group.

(a) Write down the name of this type of sampling. [1]

The results of the survey, along with the value obtained from the model, are given in the following table.

Country	A	B	C	D	E	F
Value from the model (x)	12.3	15.2	14.1	18.5	20.1	19.2
Happiness score (y)	5.2	7.3	6.2	6.9	8.0	7.2

The researchers will accept the model is a valid predictor of happiness score if the Pearson's product-moment correlation coefficient, r , is greater than 0.8.

(b) (i) Find the value of r .
(ii) Hence state whether the model can be regarded as a valid predictor of happiness score. [3]

(c) Find the equation of the regression line y on x . [1]

For a particular country $x = 17.2$.

(d) Use the regression line to predict the happiness score for this country. [2]

(This question continues on the following page)



[illegible]

Apollo | Original source crop | 156

12. [Maximum mark: 5]

Tajinder believes that the number of cars arriving at an intersection every minute follows a Poisson distribution. Under this assumption he decides to test the null hypothesis that the mean is 12.4 cars per minute against the alternative hypothesis that the mean is greater than 12.4.

The test consists of observing the intersection for 10 minutes and then rejecting the null hypothesis if more than 130 cars arrive.

- (a)

(i)

State the meaning of a Type I error.
- (ii)

Find the probability of a Type I error for Tajinder's test.

[4]
- (b)

Write down one reason why a Poisson distribution might not be a suitable model for the number of cars arriving at the intersection.

[1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



15. [Maximum mark: 6]

A Markov chain has two states, A and B. The transition matrix for the Markov chain is

A

B

A

B

$$\begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix}.$$

One of the eigenvalues for this matrix is equal to 1.

- (a)

Find the eigenvector corresponding to the eigenvalue of 1.

[3]
- (b)

Hence, find an expression, in terms of p and q , for the proportion of time spent in state A.

[2]
- (c)

Hence or otherwise, find the proportion of time spent in state A when $p = 0.3$ and $q = 0.6$.

[1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

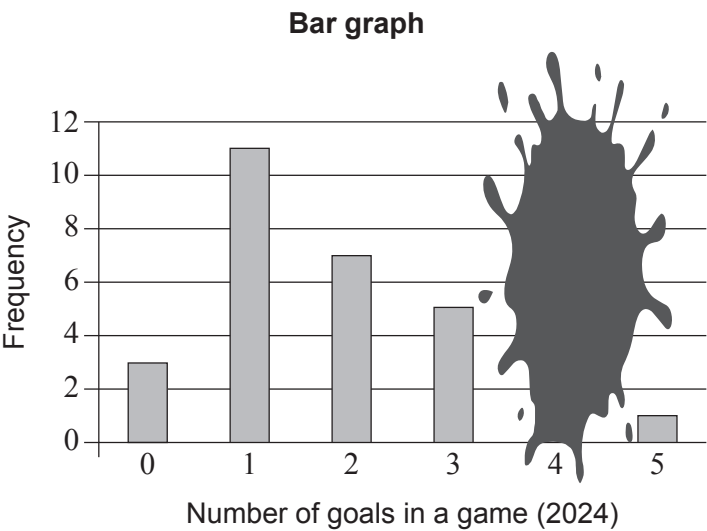


Turn over

1. [Maximum mark: 15]

Paul has a bar graph for the total number of goals scored in each game of a soccer tournament in 2024. The bar graph is shown below, however the frequency of 4 goals in a game is unreadable.

Paul uses this bar graph to create a frequency table.



Frequency table

Number of goals in a game (2024)	Frequency
0	3
1	11
2	7
3	k
4	p
5	1

(a) Write down the value of k . [1]

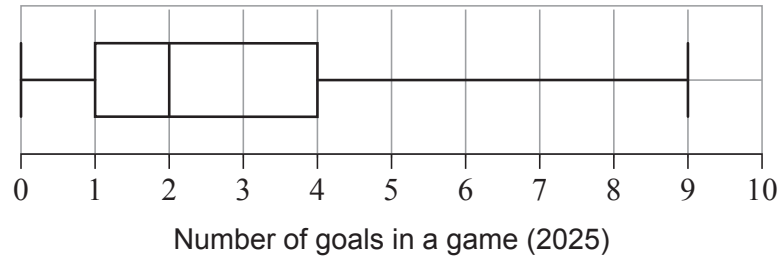
Paul knows that the mean number of goals per game scored during the tournament was 2.2.

(b) (i) Write down an equation for the mean in terms of p .
(ii) Determine the value of p . [3]

(This question continues on the following page)

(Question 1 continued)

Data for the number of goals per game in the **2025 soccer tournament** are shown in the following box and whisker diagram.



After comparing the box and whisker diagram from the 2025 tournament with the frequency table from the 2024 tournament, Paul concludes that the distribution of goals is consistent between the two tournaments.

- (c) State two observations that support Paul’s conclusion using values from the data to compare any **two** of:

range, symmetry, median, and interquartile range. [3]

Paul plans to watch all the games from the **2024** tournament in a random order. He will watch each game once.

For the first game he watches, he defines event F as:

“scoring either 0 goals or exactly 1 goal”.

- (d) Write down the event(s) from the table that are equivalent to F' . There may be more than one correct event. [2]

Event	Description
A	Scoring exactly 2 goals in this game
B	Scoring more than 1 goal in this game
C	Scoring at least 2 goals in this game
D	Scoring either 0 goals or exactly 1 goal in all games except this game
E	Not scoring 0 goals or exactly 1 goal in any game

- (e) If exactly 1 goal was scored in the first game Paul watches, write down the probability that exactly 1 goal was scored in the second game he watches. Give your answer as a fraction. [2]

- (f) Calculate the probability that 5 goals were scored in the first game that Paul watches **and** 0 goals were scored in the second game he watches. [4]

Turn over

5. [Maximum mark: 15]

A zoologist collects a sample of cane beetles. He measures their length and categorizes them as “small” meaning from 10 to 12 mm long, “medium” meaning from 12 to 16 mm long and “large” meaning from 16 to 18 mm long. He also notes their sex and records the frequencies in the following table.

		Length, x mm		
		Small $10 < x \leq 12$	Medium $12 < x \leq 16$	Large $16 < x \leq 18$
Sex	Female	42	25	19
	Male	61	27	12

- (a) Find how many cane beetles are in the zoologist's sample. [1]
- (b) Based on this data set, estimate the mean length of a cane beetle. [2]

Two female beetles are chosen at random with replacement from the sample.

- (c) Find the probability that they are both categorized as small. [3]
- (d) Test, at the 5% significance level, the null hypothesis that length category and sex are independent. State the p -value of your test and write your conclusion in context. Justify your answer. [4]

Let ϕ be the population proportion of cane beetles that are male.

- (e) Test, at the 5% significance level, the hypothesis that more than 45% of cane beetles are male. Write the null and alternative hypotheses. State the p -value of your test and write your conclusion in context. [5]

7. [Maximum mark: 17]

On any given day, the probability that Emlyn charges his phone depends only on whether he charged it the previous day.

If he charged his phone the previous day, the probability he charges it today is 0.4.

If he did not charge his phone the previous day, the probability he charges it today is p .

On day n this can be represented using the vector $\mathbf{v}_n$ where

$$\mathbf{v}_n = \begin{pmatrix} \text{probability that Emlyn charges his phone on day } n \\ \text{probability that Emlyn does not charge his phone on day } n \end{pmatrix}$$

A Markov chain model is formed where

$$\mathbf{v}_{n+1} = \mathbf{M}\mathbf{v}_n.$$

Matrix $\mathbf{M}$ is of the form $\begin{pmatrix} a & p \\ b & 1-p \end{pmatrix}$.

(a) Write down the value of

(i) a .

(ii) b .

[2]

(b) On day zero Emlyn charges his phone. Find the probability

(i) that Emlyn charges his phone on all days from $n = 1$ to $n = 4$.

(ii) that Emlyn charges his phone on day 4, when $p = 0.7$.

[5]

(c) Demonstrate that, for all values of p , one eigenvector of $\mathbf{M}$ is $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and hence state the associated eigenvalue.

[4]

(d) Find, in terms of p , the steady state probability that Emlyn charges his phone on a given day.

[4]

In the long term, Emlyn wants to charge his phone on at least 60% of days.

(e) Find the minimum value of p required for this to occur.

[2]

Disclaimer:

Content used in IB assessments is taken from authentic, third-party sources. The views expressed within them belong to their individual authors and/or publishers and do not necessarily reflect the views of the IB.

References:

1. naisupakit, 2016. *Triangle Pillow tradition native Thai style pillow*. [image online] Available at: <https://www.gettyimages.co.uk/detail/photo/triangle-pillow-tradition-native-thai-style-pillow-royalty-free-image/623127206> [Accessed 9 April 2024]. Source adapted.

All other texts, graphics and illustrations © International Baccalaureate Organization 2025

6. [Maximum mark: 15]

Daphne runs a coffee shop and needs to decide whether she should hire more staff.

Over a long period of time, the staff collect data on the number of customers, C , who are queuing when a new customer enters the shop. The probability distribution of C is shown in the following table.

Number of customers in queue, c	0	1	2	3	≥ 4
$P(C = c)$	0.21	0.34	0.28	0.17	0

- (a) (i) Find the probability that there are at least two customers in the queue when a new customer arrives.

- (ii) Find $E(C)$.

[3]

The time in seconds, T , taken to serve a single customer can be modelled by the normal distribution $T \sim N(115, 28^2)$.

Daphne is told that an estimate for the expected time a customer will need to wait before being served can be found by calculating $E(C) \times E(T)$.

- (b) Find the value of $E(C) \times E(T)$.

[2]

Daphne now decides that she will employ a new member of staff if the probability that a customer has to wait more than three minutes before being served is greater than 0.4.

- (c) Using the distribution of T given above, find the probability that it takes more than three minutes to serve a randomly selected customer.

[2]

- (d) Find the probability it takes more than three minutes in total to serve two randomly selected customers. You may assume all serving times are independent of all other serving times.

[4]

Daphne assumes that when a new customer arrives the person at the front of the queue has only just reached the service counter. She also assumes that the probability that it takes more than three minutes to serve three customers is equal to 1.

- (e) Using Daphne's assumptions and the probabilities for C given in the table above

- (i) find the probability a customer just arriving at the coffee shop will wait more than three minutes before being served.

- (ii) Hence state whether Daphne will decide to employ more staff.

[4]

Turn over

1. [Maximum mark: 29]

The following question explores how sequences, series and Markov chains may be used in modelling the number of customers in a commercial setting.

In a town, there are three stores: Aroma, Bodega and Clover.

Ashley is the manager of Aroma. She gathers data to determine whether there is significant movement of customers between the three stores over the course of one year.

She found that:

- 91 % of Aroma customers stayed with Aroma, 5 % moved to Bodega, and 4 % moved to Clover.
- 95 % of Bodega customers stayed with Bodega, 4 % moved to Aroma, and 1 % moved to Clover.
- 92 % of Clover customers stayed with Clover, 6 % moved to Aroma, and 2 % moved to Bodega.

This information is used to form a transition matrix, T .

- (a) Write down the transition matrix T . [3]

It is assumed that the movement of customers between stores remains constant from year to year.

- (b) Determine the percentage of Clover customers expected to move to Aroma over a 5-year period. [4]
- (c) Find the steady state vector for T . [2]
- (d) Hence, state the percentage of Clover customers expected to move to Aroma in the long term. [1]

(This question continues on the following page)

(Question 1 continued)

Ashley's initial findings suggested that 6% of Clover's customers moved to Aroma over the course of one year. Ashley is instructed to increase this figure so that at least 40% of Clover's customers move to Aroma **in the long term**.

It may be assumed that all other annual percentage changes remain the same, other than the percentage that stay with Clover.

- (e) Determine the minimum integer percentage to which the 6% figure will need to be raised to achieve this objective, justifying your answer. [3]
- (f) State, with an explanation, whether Bodega's manager would benefit from Ashley attracting more of Clover's customers. [2]
- (g) Suggest a contextual reason why the annual percentage changes are unlikely to be constant from year to year. [1]

Ashley moves to a new store, Dusk, in a different town.

She notes that in her first week, Dusk had 600 customers and that every week, the number of customers increases by 30. Therefore, Ashley models the number of customers as an arithmetic sequence (**Model 1**).

- (h) Use **Model 1** to find the number of customers Ashley can expect Dusk to have in her 10th week. [2]

Ashley wants to increase the number of customers at Dusk, so she introduces a loyalty scheme after her 5th week. She anticipates that the number of **new customers** she can expect each week should be 12% greater than the number of **new customers** the previous week (**Model 2**).

- (i) Show that the number of customers in Ashley's 6th week, according to **Model 2**, is 753.6. [2]
- (j) Use **Model 2** to find how many customers Ashley should expect in her 10th week. [3]
- (k) By comparing the two models, determine the first week in which there would be an expected difference of at least 500 customers. [6]

Turn over

2. [Maximum mark: 26]

The following question uses statistical tests to compare the weights of eggs in different situations.

Farmer Giles owns a chicken farm and sells eggs at the local market. He assumes that the weights of his eggs follow a normal distribution, with mean 52.0 g and standard deviation 3.7 g.

Giles selects 25 eggs at random.

- (a) Find the probability that at least 5 of these eggs weigh at least 55.0 g. [4]
- (b) Calculate the probability that the mean weight of these 25 eggs lies within 1 gram of the population mean. [3]

In an effort to increase the mean weight of his eggs, Farmer Giles gives the chickens a new food, *Chick Crackle*, for one month. He then randomly selects 10 eggs and weighs them. His results, in grams, are shown in **Table 1**.

Table 1

54.5	52.4	50.4	54.3	54.2	53.1	50.2	51.8	52.2	53.5
------	------	------	------	------	------	------	------	------	------

Farmer Giles assumes the weights of all his eggs are still normally distributed but believes the mean and standard deviation of the weights may have changed.

- (c) Perform a suitable test at the 5 % significance level to investigate whether the mean weight of his eggs has increased from 52.0 g. [6]
- (d) Find a 99 % confidence interval for the mean weight of Giles's eggs, after feeding his chickens *Chick Crackle*. [2]

To analyse further the effect of *Chick Crackle*, Giles decides to choose 8 specific chickens. He weighs a randomly selected egg from each chicken before feeding them *Chick Crackle*. He then weighs another randomly selected egg from each chicken after feeding them *Chick Crackle* for one month, to determine whether the new feed has the required effect.

His results, in grams, are shown in **Table 2**.

Table 2

Chicken	A	B	C	D	E	F	G	H
Egg weight before	53.8	51.5	49.3	52.3	46.4	55.0	51.2	52.4
Egg weight after	55.1	51.9	50.0	54.8	47.3	53.5	52.2	54.4

- (e) (i) State a suitable test that Giles should use. [1]
- (ii) State one assumption that Giles needs to make before carrying out the test. [1]

(This question continues on the following page)

(Question 2 continued)

- (f) Perform the test at the 5 % significance level, stating your conclusion in context. [6]

From the next village, Farmer Ray offers Farmer Giles a random sample of 10 eggs from his own chickens and requests that Giles carries out a t -test to see whose farm produces the heavier eggs.

Farmer Giles recalls from his statistics degree that a t -test is only valid if both population variances are the same.

- (g) Given that Giles uses his data from **Table 1**, calculate the value of S_G^2 , the unbiased estimator for the population variance of his eggs. [1]

He sets out to determine the maximum value of S_R^2 , the unbiased estimator for the population variance of Ray's eggs, for which the t -test is considered valid. To do this, he uses a test called the F -test.

The test statistic for the F -test in this case is $F = \frac{S_R^2}{S_G^2}$.

The critical value at the 5 % level for comparing two samples of 10 eggs is $F_{crit} = 3.1789$.

- (h) Determine the maximum value of S_R^2 for the t -test to be valid. [2]

1. [Maximum mark: 28]

This question is about comparing the academic performance of two schools.

At age 18, all students in school A and school B take the same final exam. Augustin is studying the results in these schools.

Augustin chooses to take a representative sample of size six from each school. For each student in the sample, he will conduct an interview.

- (a) (i) State one advantage of increasing the sample size. [1]
- (ii) State one disadvantage of increasing the sample size. [1]

The data in **Table 1** shows the results of the final exam as a percentage for each of the six students, sampled from school A.

Table 1

Candidate number	Final exam
10001	63.5
10003	52.5
10012	50.7
10045	42.8
10073	44.7
10090	56.1

The mean result for the sample from school A is 51.7 to three significant figures.

- (b) Find the value of s_{n-1} for the sample from school A. [2]

(This question continues on the following page)

(Question 1 continued)

The value of s_{n-1} for the sample from school B is 7.66 to three significant figures.

Augustin makes the following claim:

“The spread of results in school A must be less than the spread of results in school B.”

- (c) Make one criticism of Augustin’s claim. [1]

The examination board claims that the final exam results in each school are approximately normally distributed. You may assume that this claim is correct.

Augustin decides to use a pooled t -test to compare the mean results of school A and of school B.

- (d) (i) State the condition regarding population variances required to use a pooled t -test. [1]
 (ii) State whether Augustin should use a pooled t -test in this case. Justify your answer. [2]

Prior to collecting data, Augustin believed that the mean result of school B was higher than that of school A. From his data he finds that the mean final exam result for the sample from school B is exactly 60.

- (e) (i) State appropriate null and alternative hypotheses for the pooled t -test. [2]
 (ii) Find the p -value. [2]
 (iii) Given that the test is carried out at the 5% significance level, state the appropriate conclusion in context. Justify your answer. [2]

All students in the two schools took the same entry exam at age 11.

Augustin wants to determine if there is evidence of any correlation between the entry exam result and final exam result. For the 12 students in the sample, Augustin collects their entry exam results.

The Pearson’s product moment correlation coefficient between the results for the entry exam and the final exam is $r = 0.876$ to three significant figures. The critical value of r at the 5% significance level is 0.576.

- (f) (i) Assuming all requirements are met, perform a test at the 5% significance level.
 State the hypotheses and justify your conclusion. [4]
 (ii) If the requirements are not met, state an alternative test. [1]

(This question continues on page 5)

Turn over

— T —

222071310

Blank page

(Question 1 continued)

The examination board uses a model to make a prediction of a student’s final exam result ($\hat{y}$) based on their entry exam result (x).

The model used is:

$$\hat{y} = 0.37x + 37.6$$

- (g) Give, in context, an interpretation of the gradient 0.37 in the model. [1]

For each student, Augustin uses the value $y - \hat{y}$ rounded to one decimal place to measure the extent to which a school helped to improve a student’s results. He calls this “school value added”. This is shown in **Table 2**.

Table 2

Candidate number	School	Entry exam (x)	Final exam (y)	School value added ($y - \hat{y}$)
10001	A	53.2	63.5	6.2
10003	A	38.7	52.5	q
10012	A	32.0	50.7	1.3
10045	A	20.0	42.8	−2.2
10073	A	23.2	44.7	−1.5
10090	A	42.4	56.1	2.8
20004	B	72.0	60.8	−3.4
20008	B	69.5	60.3	−3.0
20023	B	52.3	47.1	−9.9
20044	B	84.1	69.9	1.2
20061	B	74.7	64.6	−0.6
20083	B	66.9	57.3	−5.1

- (h) (i) Show that the value of q in **Table 2** is 0.6. [2]
- (ii) Assuming that the appropriate requirements are met, use a pooled t -test at a 5% significance level to determine if the mean “school value added” is higher in school A than in school B. Write down your null and alternative hypotheses and justify your conclusion. [4]
- (i) Using Augustin’s results, explain how each school could claim they are performing better than the other school. [2]

Turn over

1. [Maximum mark: 7]

Under controlled driving conditions, Jacob investigated the fuel efficiency of his car when using premium fuel compared to standard fuel.

Jacob recorded the distance travelled per litre (km L^{-1}) using standard fuel for six days and then using premium fuel for seven days. This information is shown in the following tables.

	Standard fuel						
Distance travelled per litre (km L^{-1})	8.1	8.2	7.9	8.2	7.9	8.0	

	Premium fuel						
Distance travelled per litre (km L^{-1})	8.3	8.4	8.1	8.3	8.2	8.0	8.2

At the 5% significance level, Jacob performs a t -test to determine whether there is sufficient evidence that his car travels further using premium fuel compared to standard fuel.

- (a) State one mathematical assumption made for this test to be valid. [1]
- (b) Write down the

(i) null hypothesis.

(ii) alternative hypothesis. [2]
- (c) (i) Find the p -value.
- (ii) State your conclusion to the test in context. Justify your answer. [4]

(This question continues on the following page)



(Question 1 continued)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....



Turn over

— T —

002071301

Please **do not** write on this page.
Answers written on this page
will not be marked.



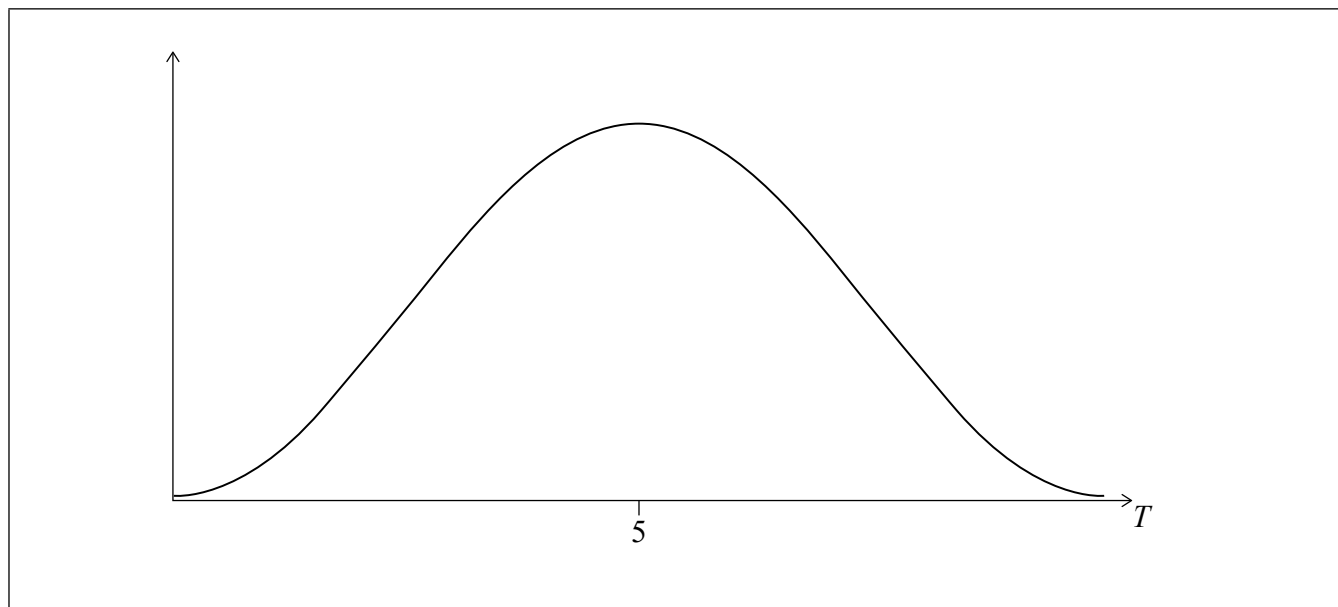
3. [Maximum mark: 5]

In a reaction test, the time, T seconds, taken to complete the test can be modelled by a normal distribution with a mean of 5 seconds and a standard deviation of 1 second.

- (a) Find the probability that a randomly selected person completes the reaction test in less than 4 seconds.

[2]

The graph of the distribution of T is shown in the following diagram.



The slowest 15% of participants will be given training to reduce their time to complete the test.

- (b) (i) Shade the region on the given diagram that corresponds to these participants.
- (ii) Find the least time taken to complete the test by a person who will be given training.

[3]

(This question continues on the following page)



(Question 3 continued)

[illegible]

Turn over

9. [Maximum mark: 6]

A help desk receives an average of 34 calls each hour. The number of calls received can be modelled by a Poisson distribution.

The help desk staff can answer a maximum of 22 calls in any 30-minute period.

- (a) Find the probability that some calls are not answered in a 30-minute period. [3]

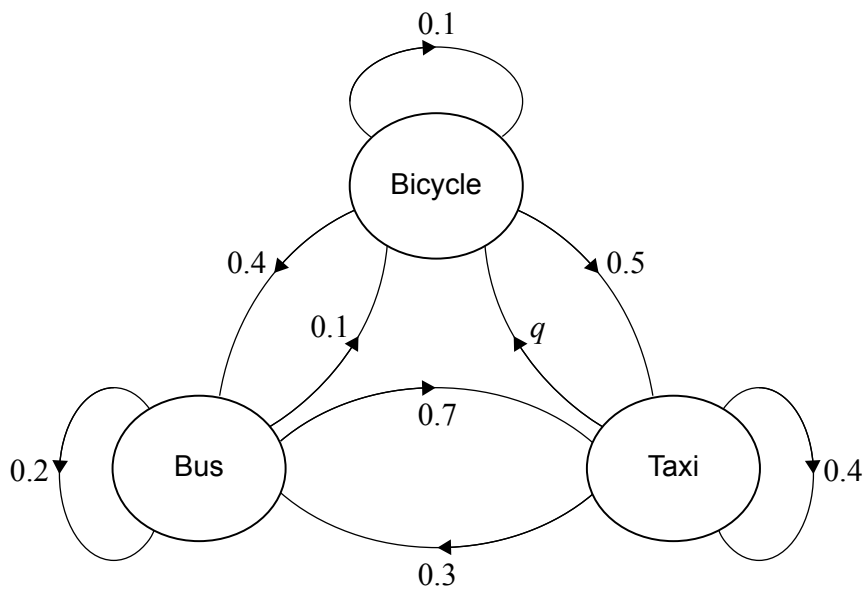
A supervisor is deciding whether to employ extra staff for the help desk. They monitor the phone calls for a random 30-minute period each day for 10 days.

- (b) Find the probability that all the monitored calls are answered on at least 9 of these 10 days. [3]



16. [Maximum mark: 8]

Each day Anirudh goes to work by bus (B) or bicycle (C) or taxi (T). The probability of him using a mode of transport depends on which mode of transport he used the previous day, and is shown in the following transition diagram.



- (a) Write down the value of q . [1]

On day 1, Anirudh goes to work by bus.

- (b) Write down the probability that he goes to work by taxi on day 2. [1]
- (c) Find the probability that the mode of transport Anirudh uses on day 3 is the same as the one he uses on day 2. [3]
- (d) Complete the following transition matrix. [1]

$\begin{pmatrix} 0.2 & 0.4 & 0.3 \\ 0.1 & 0.1 & \square \\ 0.7 & \square & \square \end{pmatrix}$

- (e) Hence, find the long-term proportion of days that he goes to work by bicycle. [2]

(This question continues on the following page)



(Question 16 continued)

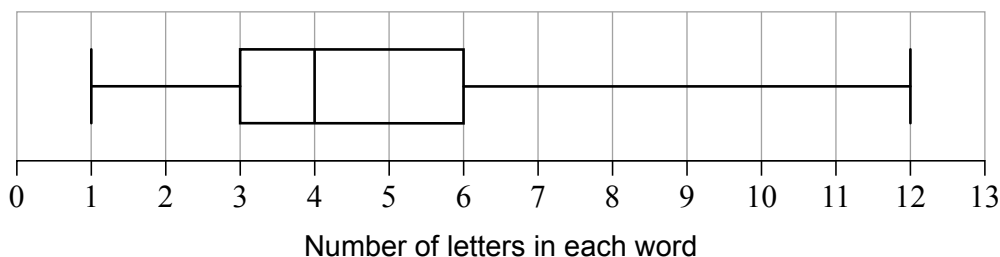
[illegible]

Turn over

1. [Maximum mark: 15]

Dr Petrillo wrote a short scientific essay. He analysed the readability of his essay by counting the number of letters in each word.

Dr Petrillo constructs a box and whisker diagram for his data.



(a) Write down

- (i) the median;
- (ii) the upper quartile, Q_3 ;
- (iii) the interquartile range, IQR.

[3]

Dr Petrillo now wants to modify his diagram to show any outliers. He considers the longer words in his data and uses the following formula:

$$\text{outliers} > (1.5 \times \text{IQR}) + Q_3.$$

Words with at least k letters are considered outliers.

(b) Find the value of k .

[2]

(This question continues on the following page)

(Question 1 continued)

Dr Petrillo further considers the outliers and sees no reason to exclude them from his analysis.

The length of each word in the essay, n , and its associated frequency are given in the following table.

n	$1 \leq n \leq 2$	$3 \leq n \leq 4$	$5 \leq n \leq 6$	$7 \leq n \leq 8$	$9 \leq n \leq 10$	$11 \leq n \leq 12$	$n \geq 13$
Frequency	289	480	220	136	90	45	0

- (c) Use the mid-interval values to calculate an estimate of the mean number of letters in each word. [3]

Dr Petrillo conducts a χ^2 goodness of fit test at the 1 % significance level, to test the following null hypothesis:

H_0 : The frequency of the number of letters in each word in his essay is consistent with the English language.

- (d) Write down the alternative hypothesis for this test. [1]

The observed and expected frequencies of the number of letters in each word in his essay are listed in the following table.

n	$1 \leq n \leq 2$	$3 \leq n \leq 4$	$5 \leq n \leq 6$	$7 \leq n \leq 8$	$9 \leq n \leq 10$	$11 \leq n \leq 12$	$n \geq 13$
Observed frequency	289	480	220	136	90	45	0
Expected frequency	252.0	491.4	239.4	151.2	88.2	31.8	6.0

- (e) (i) Write down the number of degrees of freedom.
 (ii) Find the χ^2 statistic for this test.
 (iii) Find the p -value for this test. [4]

The critical value for this test, at the 1 % significance level, is 16.812.

- (f) State whether Dr Petrillo should reject the null hypothesis. Justify your answer. [2]

Turn over

— T —

002071302

Blank page

5. [Maximum mark: 17]

A company sells spicy snacks. Over a long period of time, the company finds that 60% of people who taste the snacks will buy them. The company wants to test a new recipe to see if a greater percentage of people who taste these new snacks will buy them.

The company surveys the first 80 customers at a local supermarket who agree to taste the new snacks.

- (a) State the name of the sampling method used. [1]

Out of the 80 people who taste the new snacks, 54 buy them. The company performs a hypothesis test to determine whether the new recipe is an improvement.

It assumes the responses of the 80 people are independent of each other. The null hypothesis is

$$H_0: p = 0.6,$$

where p is the proportion of people who buy the new snacks after tasting them.

- (b) (i) Write down H_1 , the alternative hypothesis.
- (ii) Perform a suitable test, at the 5% significance level, stating the conclusion in context. [6]
- (c) (i) Find the critical region for this test.
- (ii) Given that $p = 0.65$, find the probability that the company makes a Type II error. [4]

During the survey, the company also recorded the age of each person. They want to find out if a person's age is independent of whether they will buy these snacks. The data for the 80 people is shown in the following table.

Age, x	Will buy	Will not buy
$x < 20$	22	8
$20 \leq x < 30$	18	5
$30 \leq x < 40$	6	8
$40 \leq x$	8	5

- (d) Perform a suitable test, at the 5% significance level, to determine if age is independent of whether someone will buy these snacks. [6]

Turn over

Blank page