Subject – Math AI(Standard Level)
Topic - Statistics and Probability
Year - May 2021 – Nov 2024
Paper -1
Questions

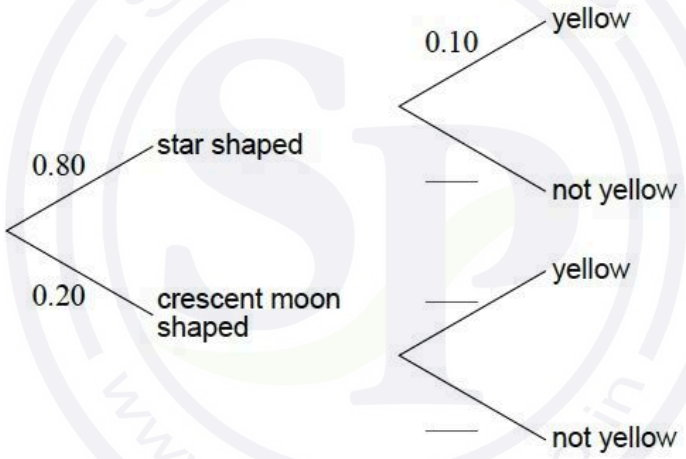
Question 1

[Maximum mark: 14]

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) Using the given information, **copy** and complete the following tree diagram. [2]



(b) 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]

Question 2

[Maximum mark: 13]

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

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

(i) in less than $6.5\,\text{m}$.

(ii) in more than $7\,\text{m}$. [3]

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

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

(i) $6.5\,\text{m}$ and $6.75\,\text{m}$.

(ii) $6.75\,\text{m}$ and $7\,\text{m}$. [3]

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

Measured stopping distance	Number of bicycles
Less than $6.5\,\text{m}$	12
Between $6.5\,\text{m}$ and $6.75\,\text{m}$	428
Between $6.75\,\text{m}$ and $7\,\text{m}$	527
More than $7\,\text{m}$	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 $20\,\text{km h}^{-1}$ can be modelled by a normal distribution with mean $6.76\,\text{m}$ and standard deviation $0.12\,\text{m}$.

(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]

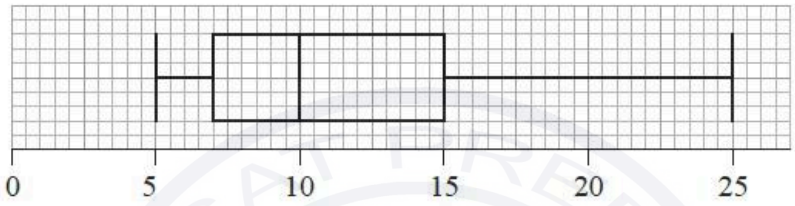
Question 3

[Maximum mark: 18]

As part of his mathematics exploration about classic books, Jason investigated the time taken by students in his school to read the book *The Old Man and the Sea*. He collected his data by stopping and asking students in the school corridor, until he reached his target of 10 students from **each** of the literature classes in his school.

(a) State which of the two sampling methods, systematic or quota, Jason has used. [1]

Jason constructed the following box and whisker diagram to show the number of hours students in the sample took to read this book.



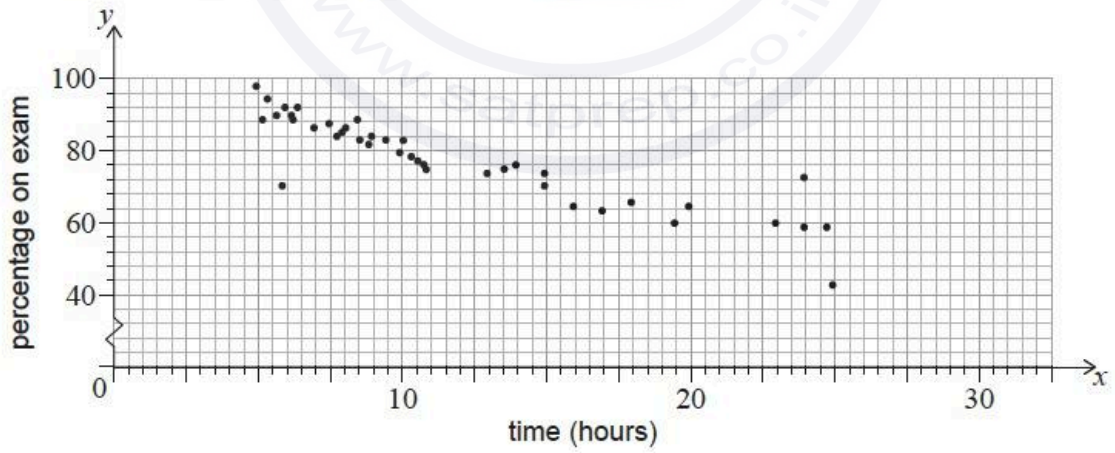
(b) Write down the median time to read the book. [1]

(c) Calculate the interquartile range. [2]

Mackenzie, a member of the sample, took 25 hours to read the novel. Jason believes Mackenzie's time is not an outlier.

(d) Determine whether Jason is correct. Support your reasoning. [4]

For each student interviewed, Jason recorded the time taken to read *The Old Man and the Sea* (x), measured in hours, and paired this with their percentage score on the final exam (y). These data are represented on the scatter diagram.



(e) Describe the correlation. [1]

Jason correctly calculates the equation of the regression line y on x for these students to be

$$y = -1.54x + 98.8.$$

He uses the equation to estimate the percentage score on the final exam for a student who read the book in 1.5 hours.

(f) Find the percentage score calculated by Jason. [2]

(g) State whether it is valid to use the regression line y on x for Jason's estimate. Give a reason for your answer. [2]

Jason found a website that rated the 'top 50' classic books. He randomly chose eight of these classic books and recorded the number of pages. For example, Book H is rated 44th and has 281 pages. These data are shown in the table.

Book	A	B	C	D	E	F	G	H
Number of pages (n)	4215	863	585	1225	366	209	624	281
Top 50 rating (t)	1	2	5	7	13	22	40	44

Jason intends to analyse the data using Spearman's rank correlation coefficient, r_s.

(h) Copy and complete the information in the following table. [2]

Book	A	B	C	D	E	F	G	H
Rank – Number of pages	1							
Rank – Top 50 Rating	1							

(i) (i) Calculate the value of r_s.

(ii) Interpret your result. [3]

Question 4

[Maximum mark: 14]

It is known that the weights of male Persian cats are normally distributed with mean $6.1\,\text{kg}$ and variance $0.5^2\,\text{kg}^2$.

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

(b) Find the proportion of male Persian cats weighing between $5.5\,\text{kg}$ and $6.5\,\text{kg}$. [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 $5.3\,\text{kg}$. [3]

(d) It is found that 12 of the cats weigh more than $x\,\text{kg}$. Estimate the value of x. [3]

(e) Ten of the cats are chosen at random. Find the probability that exactly one of them weighs over $6.25\,\text{kg}$. [4]

Question 5

[Maximum mark: 19]

A medical centre is testing patients for a certain disease. This disease occurs in 5% of the population.

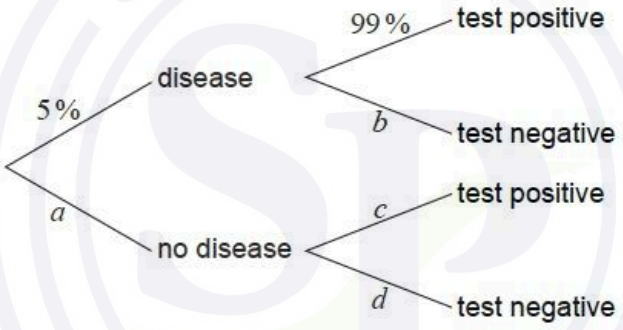
They test every patient who comes to the centre on a particular day.

(a) State the sampling method being used. [1]

It is intended that if a patient has the disease, they test "positive", and if a patient does not have the disease, they test "negative".

However, the tests are not perfect, and only 99% of people who have the disease test positive. Also, 2% of people who **do not** have the disease test positive.

The tree diagram shows some of this information.



(b) Write down the value of

(i) a.

(ii) b.

(iii) c.

(iv) d. [4]

(c) Use the tree diagram to find the probability that a patient selected at random

(i) will not have the disease and will test positive.

(ii) will test negative.

(iii) has the disease given that they tested negative. [8]

(d) The medical centre finds the actual number of positive results in their sample is different than predicted by the tree diagram. Explain why this might be the case. [1]

The staff at the medical centre looked at the care received by all visiting patients on a randomly chosen day. All the patients received at least one of these services: they had medical tests (M), were seen by a nurse (N), or were seen by a doctor (D). It was found that:

- 78 had medical tests,
- 45 were seen by a nurse;
- 30 were seen by a doctor;
- 9 had medical tests and were seen by a doctor and a nurse;
- 18 had medical tests and were seen by a doctor but were not seen by a nurse;
- 11 patients were seen by a nurse and had medical tests but were not seen by a doctor;
- 2 patients were seen by a doctor without being seen by nurse and without having medical tests.

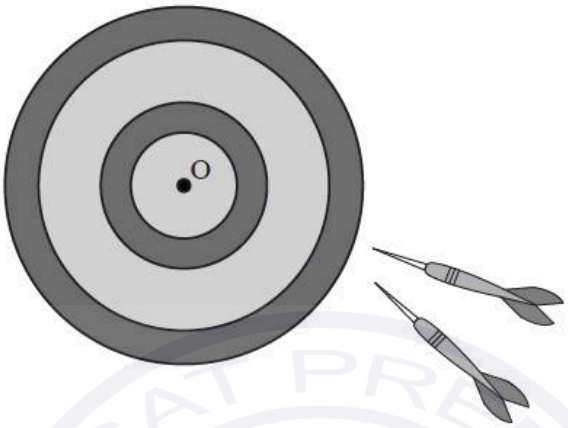
(e) Draw a Venn diagram to illustrate this information, placing all relevant information on the diagram. [3]

(f) Find the total number of patients who visited the centre during this day. [2]

Question 6

[Maximum mark: 14]

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 $10\,\text{cm}$ and standard deviation $3\,\text{cm}$.

(a) Find the probability that

(i) a dart lands less than $13\,\text{cm}$ from O.

(ii) a dart lands more than $15\,\text{cm}$ 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 $15\,\text{cm}$ 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 the probability that Arianne scores at least 5 points in the competition.

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

(iii) Given that Arianne scores at least 5 points, find the probability that Arianne scores less than 8 points. [7]

Question 7

[Maximum mark: 16]

A group of 1280 students were asked which electronic device they preferred. The results per age group are given in the following table.

	Age			
Preferred device	**11–13**	**14–16**	**17–18**	**Total**
Laptop	143	160	153	456
Tablet	205	224	131	560
Mobile phone	72	128	64	264
Total	420	512	348	1280

(a) A student from the group is chosen at random. Calculate the probability that the student

(i) prefers a tablet.

(ii) is 11–13 years old and prefers a mobile phone.

(iii) prefers a laptop **given that** they are 17–18 years old.

(iv) prefers a tablet or is 14–16 years old. [9]

A χ^2 test for independence was performed on the collected data at the 1 % significance level. The critical value for the test is 13.277.

(b) State the null and alternative hypotheses. [1]

(c) Write down the number of degrees of freedom. [1]

(d) (i) Write down the χ^2 test statistic.

(ii) Write down the p-value.

(iii) State the conclusion for the test in context. Give a reason for your answer. [5]

Question 8

[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]

Question 9

[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]

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

(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.

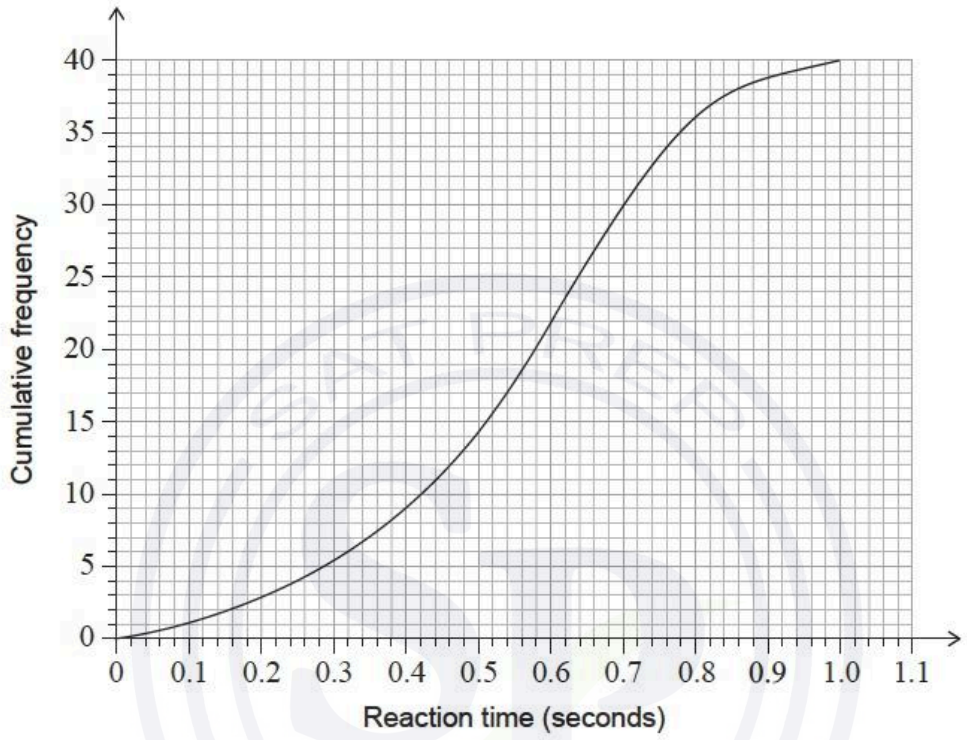
(ii) Interpret the value obtained for r_s. [3]

(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]

Question 10

[Maximum mark: 17]

Mackenzie conducted an experiment on the reaction times of teenagers. The results of the experiment are displayed in the following cumulative frequency graph.



(a) Use the graph to estimate the

(i) median reaction time;

(ii) interquartile range of the reaction times. [4]

(b) Find the estimated number of teenagers who have a reaction time greater than 0.4 seconds. [2]

(c) Determine the 90th percentile of the reaction times from the cumulative frequency graph. [2]

Mackenzie created the cumulative frequency graph using the following grouped frequency table.

Reaction time, t (s)	Frequency
$0 < t \leq 0.2$	3
$0.2 < t \leq 0.4$	a
$0.4 < t \leq 0.6$	13
$0.6 < t \leq 0.8$	14
$0.8 < t \leq 1.0$	b

(d) Write down the value of

(i) a;

(ii) b. [2]

(e) Write down the modal class from the table. [1]

(f) Use your graphic display calculator to find an estimate of the mean reaction time. [2]

Upon completion of the experiment, Mackenzie realized that some values were grouped incorrectly in the frequency table. Some reaction times recorded in the interval $0 < t \leq 0.2$ should have been recorded in the interval $0.2 < t \leq 0.4$.

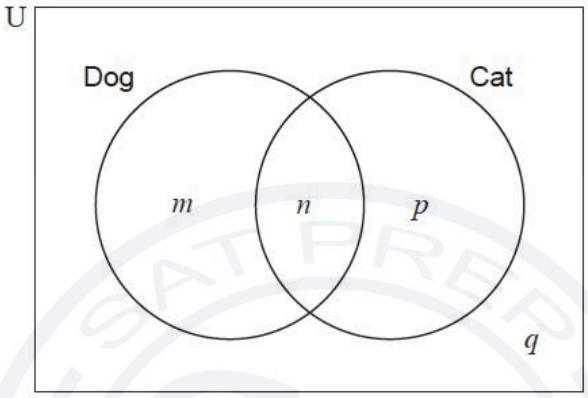
(g) Suggest how, if at all, the estimated mean and estimated median reaction times will change if the errors are corrected. Justify your response. [4]

Question 11

[Maximum mark: 16]

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 percentage of students who have a dog or a cat or both. [1]

(c) 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]

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

(d) (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.

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

Question 12

[Maximum mark: 17]

Elsie, a librarian, wants to investigate the length of time, T minutes, that people spent in her library on a particular day.

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

Elsie's data for 160 people who visited the library on that particular day is shown in the following table.

T (minutes)	$0 \leq T < 20$	$20 \leq T < 40$	$40 \leq T < 60$	$60 \leq T < 80$	$80 \leq T < 100$
Frequency	50	62	k	14	8

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

(c) (i) Write down the modal class.

(ii) Write down the mid-interval value for this class. [2]

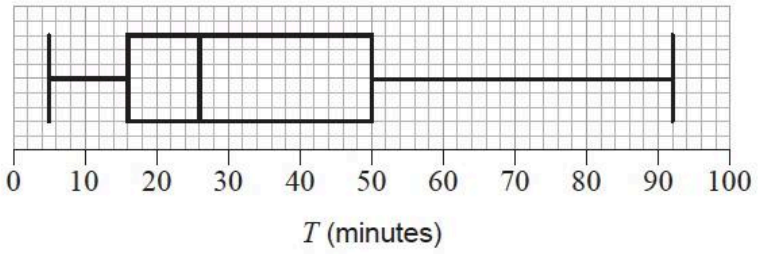
(d) Use Elsie's data to calculate an estimate of the mean time that people spent in the library. [2]

(e) Using the table, write down the maximum possible number of people who spent 35 minutes or less in the library on that day. [1]

Elsie assumes her data to be representative of future visitors to the library.

(f) Find the probability a visitor spends at least 60 minutes in the library. [2]

The following box and whisker diagram shows the times, in minutes, that the 160 visitors spent in the library.



(g) Write down the median time spent in the library. [1]

(h) Find the interquartile range. [2]

(i) Hence show that the longest time that a person spent in the library is not an outlier. [3]

Elsie believes the box and whisker diagram indicates that the times spent in the library are not normally distributed.

(j) Identify one feature of the box and whisker diagram which might support Elsie's belief. [1]

Question 13

[Maximum mark: 17]

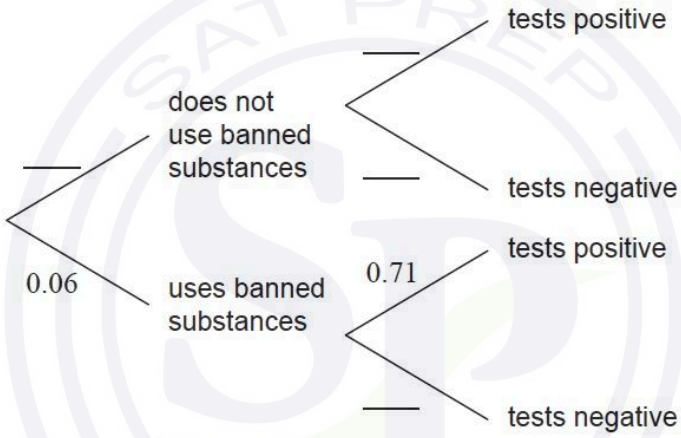
A large international sports tournament tests their athletes for banned substances.
They interpret a positive test result as meaning that the athlete uses banned substances.
A negative result means that they do not.

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]

Question 14

[Maximum mark: 15]

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]

Thandizo uses his regression line to predict the year when the mean annual temperature will first exceed $15\,°\text{C}$.

(g) State two reasons why Thandizo's prediction may not be valid. [2]

Question 15

[Maximum mark: 17]

It is claimed that a new remedy cures 82% of the patients with a particular medical problem.

This remedy is to be used by 115 patients, and it is assumed that the 82% claim is true.

(a) Find the probability that exactly 90 of these patients will be cured. [3]

(b) Find the probability that at least 95 of these patients will be cured. [2]

(c) Find the variance in the possible number of patients that will be cured. [2]

The probability that at least n patients will be cured is less than 30%.

(d) Find the least value of n. [3]

A clinic is interested to see if the mean recovery time of their patients who tried the new remedy is less than that of their patients who continued with an older remedy. The clinic randomly selects some of their patients and records their recovery time in days. The results are shown in the table below.

	Recovery time (days)								
Group N (new remedy)	12	13	9	13	14	15	17		

	Recovery time (days)								
Group O (old remedy)	17	11	10	18	20	22	14	15	18

The data is assumed to follow a normal distribution and the population variance is the same for the two groups. A t-test is used to compare the means of the two groups at the 10% significance level.

(e) State the appropriate null and alternative hypotheses for this t-test. [2]

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

(g) State the conclusion for this test. Give a reason for your answer. [2]

(h) Explain what the p-value represents. [1]

Question 16

[Maximum mark: 18]

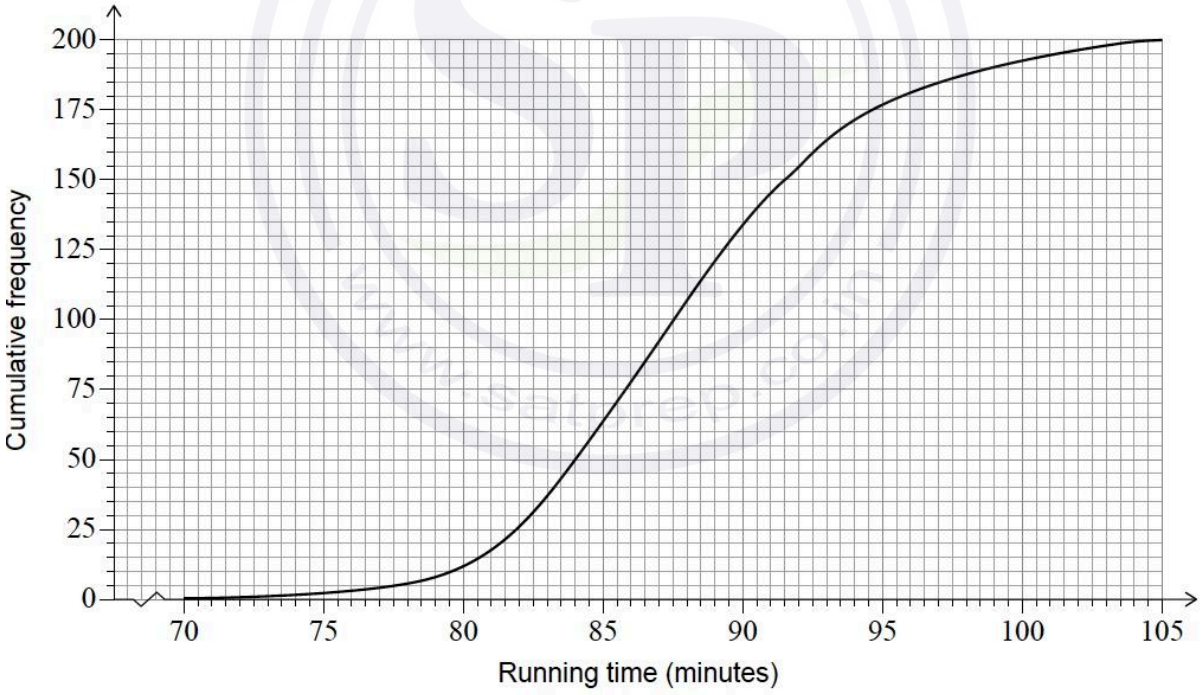
The running time, t (minutes), of 200 family movies are recorded in the following table.

Running time (minutes)	Frequency
$70 \le t < 80$	11
$80 \le t < 85$	51
$85 \le t < 90$	68
$90 \le t < 95$	47
$95 \le t < 105$	23

(a) (i) Write down the mid-interval value of $70 \le t < 80$.

(ii) Calculate an estimate of the mean running time of the 200 movies. [3]

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



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

"Star Feud" is a movie in the data set and its running time is 100 minutes.

(c) Use your answer to part (b) to estimate whether "Star Feud's" running time is an outlier for this data. Justify your answer. [3]

It is believed that the running times of family movies follow a normal distribution with mean 88 minutes and standard deviation 6.75 minutes.

It is decided to perform a χ^2 goodness of fit test on the data to determine whether this sample of 200 movies could have plausibly been drawn from an underlying distribution $\mathrm{N}(88,\ 6.75^2)$.

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

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

Movie running time (minutes)	Observed frequency	Expected frequency
$t < 80$	11	23.6
$80 \le t < 85$	51	42.1
$85 \le t < 90$	68	a
$90 \le t < 95$	47	46.7
$95 \le t$	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]

Question 17

[Maximum mark: 18]

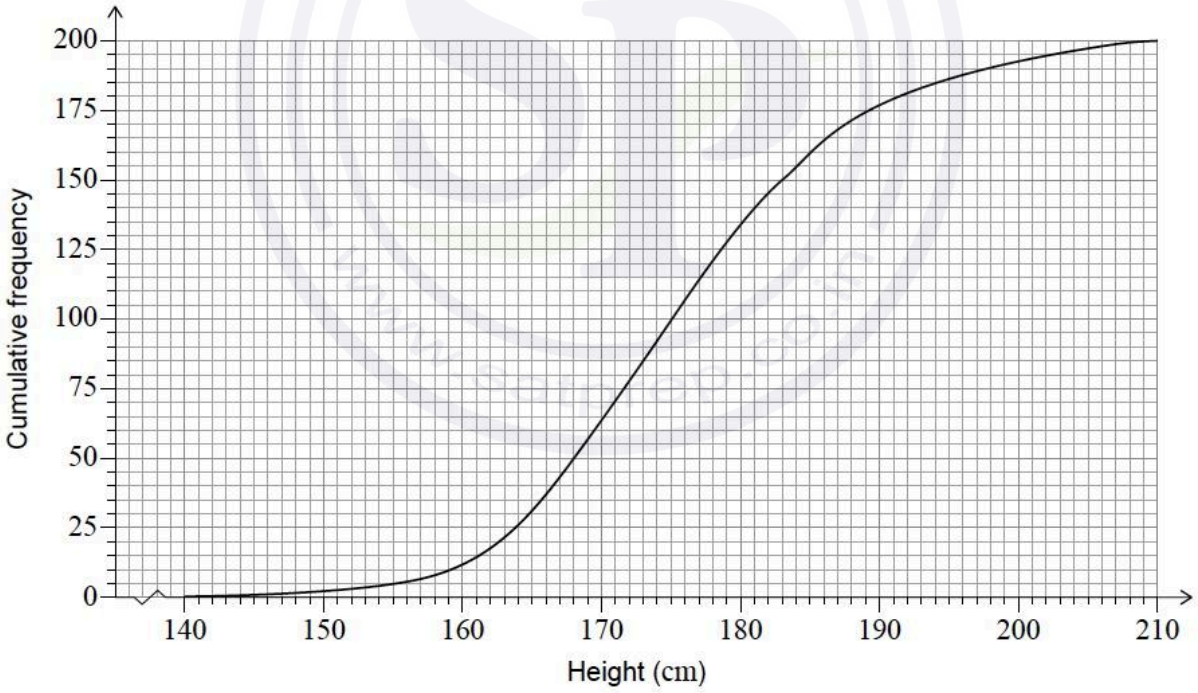
The heights, h, of 200 university students are recorded in the following table.

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]

It is believed that the heights of university students follow a normal distribution with mean $176\,\text{cm}$ and standard deviation $13.5\,\text{cm}$.

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 $\text{N}(176,\ 13.5^2)$.

(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]

Question 18

[Maximum mark: 19]

A recent study found that the heights of Dutch women can be modelled by a normal distribution with mean $170.7\,\text{cm}$ and standard deviation $6.3\,\text{cm}$.

A Dutch woman is chosen at random.

(a) Calculate the probability that her height is

(i) less than $160\,\text{cm}$.

(ii) between $160\,\text{cm}$ and $170\,\text{cm}$. [4]

27% of Dutch women have a height of more than h metres.

(b) Calculate the value of h. [2]

Janneke selects a random sample of 200 Dutch women from Amsterdam and measures their heights. She wants to determine whether this sample could have been chosen from a normally distributed population with mean of $170.7\,\text{cm}$ and standard deviation of $6.3\,\text{cm}$.

She performs a χ^2 goodness of fit test at the 5% significance level. She begins by creating the following frequency table.

Height, h	Observed frequency	Expected frequency
$h \leq 160$	6	8.943
$160 < h \leq 170$	85	a
$170 < h \leq 180$	92	b
$h > 180$	17	13.99

(c) Calculate, correct to four significant figures, the value of

(i) a.

(ii) b. [2]

The hypotheses for Janneke's test are

H_0: the heights are drawn from a normally distributed population with mean $170.7\,\text{cm}$ and standard deviation $6.3\,\text{cm}$

H_1: the heights are not drawn from a normally distributed population with mean $170.7\,\text{cm}$ and standard deviation $6.3\,\text{cm}$

(d) Write down the degrees of freedom for this test. [1]

The critical value for this test is 7.815.

(e) Perform the χ^2 goodness of fit test and state your conclusion, justifying your reasoning. [4]

Gundega claims that, on average, Latvian women are taller than Dutch women.

Random samples of 10 Latvian women and 10 Dutch women are chosen, and their heights are measured.

Heights of Latvian women (cm)									
171	163	180	159	169	182	166	168	171	170

Heights of Dutch women (cm)									
173	182	181	166	175	161	169	165	172	169

Gundega performs a t-test at the $5\,\%$ significance level. It is assumed that the populations are normally distributed and have equal variances.

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

(g) Perform the t-test and state the conclusion, justifying your reasoning. [4]

Question 19

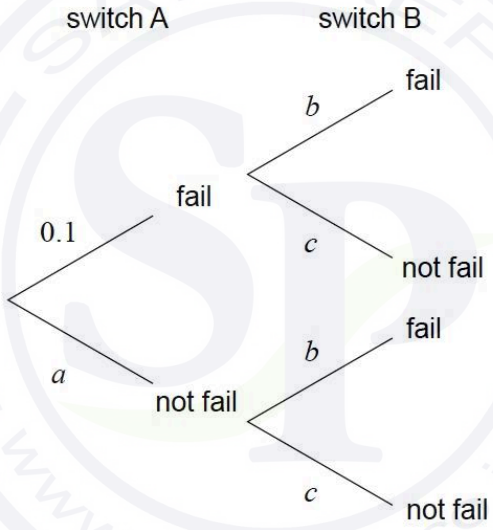
[Maximum mark: 12]

A type of generator will only function if a particular switch is working. The generator has a main switch, A, and a 'back up' switch, B.

The manufacturer claims the probability of switch A failing within one month of being fitted is 0.1 and the probability of the cheaper switch B failing within one month is 0.3. Whether or not a switch fails is independent of the state of the other switch.

If both switches fail, the generator needs to shut down to replace the switches. Both switches are replaced after a month of use (whether they have failed or not) or whenever the generator needs to be shut down.

The following tree diagram shows the probabilities of a switch failing within one month of them both being replaced, assuming the manufacturer's claim is correct.



(a) Write down the values of

(i) a.

(ii) b.

(iii) c. [2]

(b) Hence find the probability that the generator needs to shut down within one month of the switches being replaced. [1]

The owner of the generator is suspicious of the switch manufacturer's claims, so they look back through the past 200 occasions when the switches were replaced. The records show whether no switches, one switch or two switches had failed.

The data the owner collected are shown in the following table.

No switch fails	One switch fails	Two switches fail
118	72	10

(c) Show that the expected value of no switches failing in the generator, during the last 200 occasions when the switches were replaced, is 126. [2]

(d) Perform a χ^2 goodness of fit test at the 5% significance level to test whether the manufacturer's claims are correct using the following hypotheses.

H_0: The manufacturer's claims are correct.
H_1: The manufacturer's claims are not both correct. [7]

Question 20

[Maximum mark: 15]

The company Fred Express delivers packages. From past experience, the time taken, T, to deliver a package follows a normal distribution with mean 64 hours and standard deviation 12 hours.

(a) State $\mathrm{P}(T < 64)$. [1]

(b) Find $\mathrm{P}(44 < T < 64)$. [2]

30% of packages are delivered in less than k hours.

(c) (i) Sketch a diagram of this normal distribution, shading the region that represents $\mathrm{P}(T < k)$.

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

For quality control, the manager randomly selects five outgoing packages. These selections are independent.

(d) Find the probability that exactly two of these packages are delivered in less than k hours. [3]

Fred Express charges a fixed amount of $\$4.50$ for any package weighing $1\,\mathrm{kg}$ or less. Heavier packages are charged an additional fee of $\$2.00$ per kg. This fee is applied for any weight **in excess of** $1\,\mathrm{kg}$. For example, a $1.5\,\mathrm{kg}$ package is charged an additional $\$1.00$.

(e) Write down an expression for the amount charged to deliver a package of weight $x\,\mathrm{kg}$, where $x > 1$. [2]

(f) Find the amount Fred Express charges for a $5.3\,\mathrm{kg}$ package. [1]

Meiling is charged $\$7.20$ for the delivery of a package.

(g) Find the weight of Meiling's package. [2]

Question 21

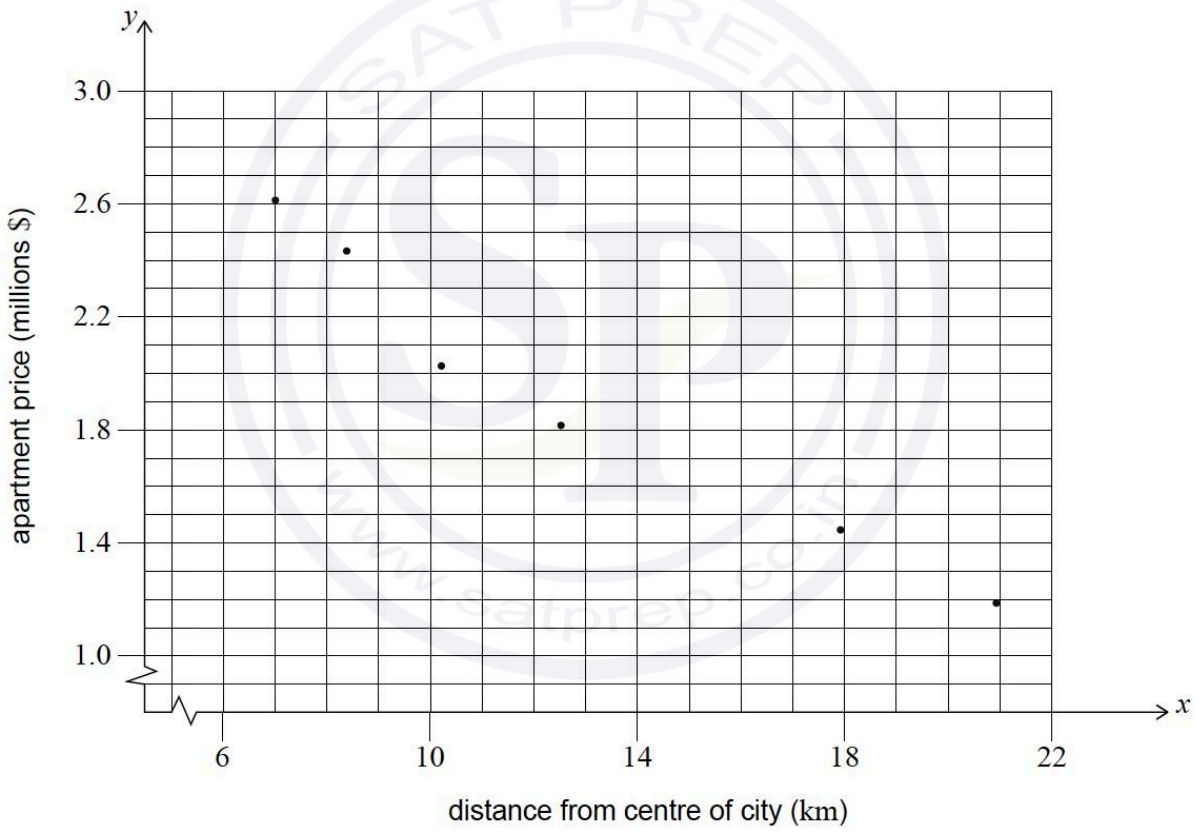
[Maximum mark: 19]

Xavie conducted a study to see if there is a relationship between the price of an apartment, y, and its distance, x, from the city centre of Melbourne.

They took a random sample of six typical apartments along a train line in the city. Xavie obtained the data shown in the following table.

x (kilometres)	7.0	8.4	10.3	12.5	17.8	20.9
y (millions of dollars)	2.61	2.44	2.03	1.81	1.45	1.18

A plot of these data is seen in the following graph.



(a) Write down the value of the Spearman's rank correlation coefficient, r_s. [1]

(b) (i) Find the Pearson's product-moment correlation coefficient, r.

(ii) Use your value of r to state which **two** of the following would best describe the correlation between the variables.

Positive	Negative	Strong	Weak	No correlation

[4]

The relationship between the variables can be modelled by the regression equation $y = ax + b$.

(c) (i) Write down the value of a.

(ii) Write down the value of b.

(iii) According to this model, state in context what the value of b represents. [3]

(d) Xavie uses the regression equation to estimate the price of a typical apartment located 19.6km from the city centre.

(i) Find this estimated price.

(ii) State two reasons that Xavie might use to justify the validity of this estimate. [5]

To verify whether this relationship applies in a different direction from the city centre, Xavie considers two locations, A and B, both an equal distance from the city centre. They take a random sample of seven apartments from each location and record the prices (in millions of dollars) in the following tables.

Apartment price in location A						
1.21	1.25	1.31	1.32	1.58	1.95	2.13

Apartment price in location B						
1.51	1.58	1.69	2.61	2.72	2.81	2.95

Xavie conducts a t-test, at the 5 % level of significance, to see if the mean apartment price in location A is different to the mean apartment price in location B. They assume the population variances are the same.

For this test, Xavie takes the null hypothesis to be $\mu_A = \mu_B$.

(e) Write down the alternative hypothesis. [1]

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

(g) State the conclusion of the test. Justify your answer. [2]

(h) State one additional assumption Xavie has made about the distributions to conduct this test. [1]

Question 22

[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]

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]

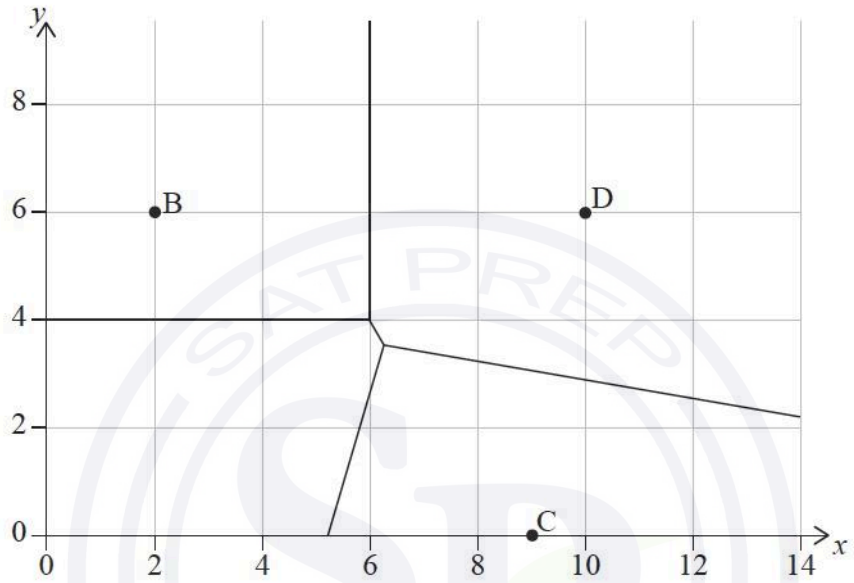
Jose concludes from this data that countries with high happiness scores are likely to be favourite expatriate countries.

(e) State, with a reason, whether Jose's conclusion is appropriate. [1]

Question 23

[Maximum mark: 13]

The following grid shows a restaurant's floorplan. There are four food stations centred at points A, B, C and D. The Voronoi diagram for these four points is shown. Point A is not shown.



One unit represents 1 metre.

Point B is located at $(2, 6)$.

The equation of the perpendicular bisector of $[\mathrm{AB}]$ is $y = 4$.

(a) Write down the coordinates of A. [1]

Points C and D are located at $(9, 0)$, and $(10, 6)$, respectively.

(b) Find

(i) the coordinates of the midpoint of $[\mathrm{CD}]$

(ii) the equation of the perpendicular bisector of $[\mathrm{CD}]$. Give your answer in the form $y = mx + c$. [6]

Customers in the restaurant take their food from the nearest food station.

The following table shows the average waiting time, in minutes, for each food station.

Food station	A	B	C	D
Average waiting time (minutes)	9	5	10	7

(c) Using nearest-neighbour interpolation, find the average waiting time for a customer at point $(4, 6)$. [1]

The restaurant owner wishes to determine whether customers spend more money during the weekend. She decides to use a two-sample t-test at a 5% level of significance.

For this test, the null hypothesis is:

$$H_0 : \mu_1 = \mu_2$$

where μ_1 is the mean amount of money spent by all customers on weekdays, and μ_2 is the mean amount of money spent by all customers on weekends.

(d) State the alternative hypothesis. [1]

The owner considers collecting her sample data from 2 weekdays and 1 weekend by recording the amount spent on the first 12 orders on each of these days.

(e) (i) State which one of the following methods best describes the owner's planned sampling:

convenience, systematic, or stratified.

(ii) State one disadvantage of this sampling technique. [2]

The restaurant owner instead collects the data using simple random sampling.

The p-value is 0.0897 for the sampled data.

(f) State the conclusion for the test. Justify your answer. [2]

Question 24

[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]

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]

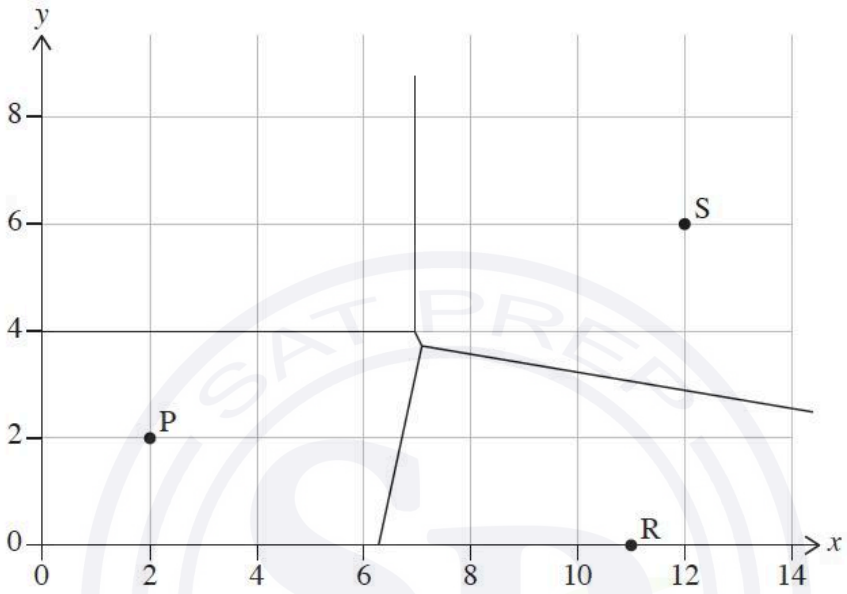
Jose concludes from this data that countries with high happiness scores are likely to be favourite expatriate countries.

(e) State, with a reason, whether Jose's conclusion is appropriate. [1]

Question 25

[Maximum mark: 13]

The following grid shows a restaurant's floorplan. There are four food stations centred at points P, Q, R and S. The Voronoi diagram for these four points is shown. Point Q is not shown.



One unit represents 1 metre.

Point P is located at $(2, 2)$.

The equation of the perpendicular bisector of $[\mathrm{PQ}]$ is $y = 4$.

(a) Write down the coordinates of Q. [1]

Points R and S are located at $(11, 0)$, and $(12, 6)$, respectively.

(b) Find

(i) the coordinates of the midpoint of $[\mathrm{RS}]$

(ii) the equation of the perpendicular bisector of $[\mathrm{RS}]$. Give your answer in the form $y = mx + c$. [6]

Customers in the restaurant take their food from the nearest food station.

The following table shows the average waiting time, in minutes, for each food station.

Food station	P	Q	R	S
Average waiting time (minutes)	8	4	9	6

(c) Using nearest-neighbour interpolation, find the average waiting time for a customer at point $(6, 8)$. [1]

The restaurant owner wishes to determine whether customers spend more money during the weekend. She decides to use a two-sample t-test at a 5% level of significance.

For this test, the null hypothesis is:

$$H_0 : \mu_1 = \mu_2$$

where μ_1 is the mean amount of money spent by all customers on weekdays, and μ_2 is the mean amount of money spent by all customers on weekends.

(d) State the alternative hypothesis. [1]

The owner considers collecting her sample data from 2 weekdays and 1 weekend by recording the amount spent on the first 12 orders on each of these days.

(e) (i) State which one of the following methods best describes the owner's planned sampling:

convenience, systematic, or stratified.

(ii) State one disadvantage of this sampling technique. [2]

The restaurant owner instead collects the data using simple random sampling.

The p-value is 0.0897 for the sampled data.

(f) State the conclusion for the test. Justify your answer. [2]