Please check the examination details below before entering your candidate information

Candidate surname	Other names

Centre Number | Candidate Number

HA061457363

Pearson Edexcel Level 3 GCE

Wednesday 5 June 2024

Morning (Time: 2 hours)

Paper reference **9ST0/01**

Statistics

Advanced

PAPER 1: Data and Probability

You must have:
Statistical formulae and tables booklet, Calculator

Total Marks

Candidates may use any calculator allowed by Pearson regulations.
Calculators must not have retrievable mathematical formulae stored in them.

Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B).
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided – *there may be more space than you need.*
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- Unless otherwise stated, inexact answers should be given to three significant figures.
- Unless otherwise stated, statistical tests should be carried out at the 5% significance level.

Information

- A booklet 'Statistical formulae and tables' is provided.
- There are 8 questions in this question paper. The total mark for this paper is 80.
- The marks for **each** question are shown in brackets – *use this as a guide as to how much time to spend on each question.*

Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.
- If you change your mind about an answer, cross it out and put your new answer and any working underneath.

Turn over

Pearson

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

1 The Olympic Games are held **every 4 years**.

The least squares regression line for the winning times in the men's 100 metre races for each Olympic Games held after 1900 was calculated.

The calculated equation, where t is the winning time in seconds, and n is the number of years after 1900 that the time was achieved, is

$$t = 10.878 - 0.0106n$$

[Source: https:www.liveabout.com/100-meter-mens-olympic-medalists-3259179]

(a) Interpret the value of 0.0106 in context.

(2)

The expected decrease in winning time for the men's 100 m race is 0.0106 seconds for every year which passes

This means the expected decrease in winning time is 0.0424 seconds per Olympic games

No Olympic Games were held in 1940 because of the Second World War.

(b) Estimate what the winning time for the men's 100 metre race in 1940 would have been, had it been run.

Give your answer correct to **two** decimal places.

(1)

$t = 10.878 - 0.0106 \times 40$ $(n = 40)$

$= 10.45$ seconds

(c) Explain why the least squares regression line may be unsuitable for predicting future winning times in the Olympic Games, men's 100 metre races.

(2)

Predicting future winning times is extrapolation and therefore may be unreliable - the winning times may not decrease in a linear way in the future

Also, the data used includes data from the early 20th century and timing technology may have improved.

(Total for Question 1 is 5 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

BLANK PAGE

2 Leona is analysing the **total** number of goals scored in football matches. She estimated the probability of a certain number of total goals being scored by the two teams in a match and displayed them in a table.

Leona's table is shown in **Figure 1**

Total goals	0	1	2	3	4	5	6
Probability	0.1	0.15	0.3	0.25	0.1	0.05	0.05

Figure 1

A match where **both** teams score the same number of goals is called a draw.

(a) Using Leona's table,

(i) explain why the probability of a draw is at least 0.1

(1)

For example - if both teams score no goals, this is a draw.
The probability of no goals scored is 0.1
But this is only 1 option, there are other possible options so the probability of a draw is at least 0.1

(ii) find the highest possible probability of a match ending in a draw.

(2)

If the match ends in a draw, both teams scored the same number of goals, so the total number of goals would be even.

so P(even number of goals) $= 0.1 + 0.3 + 0.1 + 0.05$
$= \underline{0.55}$

(b) State **one** limitation of using Leona's table, given in **Figure 1**

(1)

e.g.
- An even number of goals does not mean the match will end in a draw
- Only total number of goals up to 6 is considered
- Leona estimated the probabilities

etc.

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA

Question 2 continued

Leona models the number of matches with 3 or more goals scored in a tournament of 64 matches as a binomial distribution.

(c) State the parameters for Leona's model.

(2)

$n = 64$

$p = 0.25 + 0.1 + 0.05 + 0.05 = 0.45$

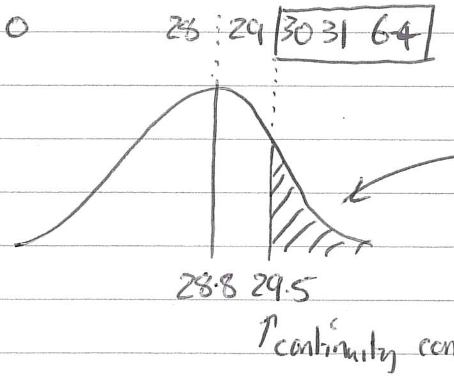
(d) Use **distributional approximation** to estimate the probability that, in a football tournament with 64 matches, at least 30 matches have 3 or more goals scored.

(4)

Let X be the number of matches where 3 or more goals were scored.

$X \sim B(64, 0.45) \approx N(28.8, 15.84)$

$\left(np = 64 \times 0.45 = 28.8 \quad np(1-p) = 64 \times 0.45 \times 0.55 = 15.84 \right)$



using calculator: $\underline{0.430}$

(e) Comment on the reliability of your approximation used in (d).

(2)

Since $n \geq 30$ $(64 \geq 30)$ and p is close enough to 0.5 $\left(\frac{1}{3} \leq p \leq \frac{2}{3}\right)$

(or $64 \times 0.45 = 28.8 \geq 10$ and $64 - 28.8 = 35.2 \geq 10$)

then the normal approximation may be reliable to use.

(f) State **two** reasons why Leona's binomial distribution model may **not** be appropriate.

(2)

- Matches with at least 3 goals scored may not be independent - if a team plays a match and improves as a result, they may score more goals in a later match.
- The probability of scoring 3 or more goals may not be the same for each match - some games involve equally matched teams where goal chance is low.

(Total for Question 2 is 14 marks)

3 Ned is a Physics education researcher at a university.

Ned wants to investigate whether there is any evidence of association between the number of sports activities that a student takes part in each week and their Physics grade.

(a) Design an experiment for Ned to carry out. Your answer should include details regarding

- what data he should gather, and how he can gather it
- how Ned should choose his sample
- which hypothesis test could be carried out with his data
- the hypotheses that should be used for this test.

(6)

- Ned should take a large sample of students ($n \geq 30$)
- Using the university enrolment register, number the students and use a random number generator to generate e.g. 100 numbers without repeats and select the students assigned those numbers
- Ask each student how many sports activities they do per week and their Physics grade

* - Group the students into categories (e.g. 0 activities, 1-2 activities etc) for sports and also into grades (e.g. High, Medium, low)
- Use a χ^2 test for association with the hypotheses

H_0: There is no association between number of sports activities and physics grade

H_1: There is an association between number of sports activities and physics grade

OR † - Record the number of sporting activities and the Physics grade score

- Use a PMCC / Spearman's Rank test with the hypotheses

(PMCC)
$H_0: \rho = 0$
$H_1: \rho \neq 0$

(Spearman's Rank)
H_0: There is no association between number of sports activities and Physics grade
H_1: There is an association between number of sports activities and Physics grade.

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 3 continued

At the end of the first year, Physics students sit a multiple-choice test with 70 questions. Each question has 4 possible answers. One student, Zamira, has not attended any lectures so guesses every answer.

(b) Making any necessary assumptions, calculate the probability that Zamira scores more than 30% on the test.

(3)

Let X be the number of correct answers Zamira gets.

$X \sim B(70, 0.25)$ $\left(\frac{1}{4} = 0.25\right)$

30% of 70 = 30% × 70 = 21

$P(X > 21)$: 0 20 21 $\boxed{22 \quad 70}$

0.8644 1-0.8644

using calculator. = 0.136

$P(X>21) = 0.136$

(c) For **each** assumption you made in (b), give a reason why that assumption might **not** be valid.

(2)

- There may be some questions Zamira may know subconsciously so the probability of guessing correct answers may not be the same each time
- There may be questions which are related so reading one question may provide a hint to a future question, so her answers may not be independent.

(Total for Question 3 is 11 marks)

4 Dahlia is running an experiment to investigate whether the consumption of an edible grain, quinoa, reduces blood glucose levels. Dahlia wants to be able to easily replicate her experiment.

From a large group of student volunteers, Dahlia selected those aged 18 to 35 years with no health issues for her experiment.

The table in **Figure 2** shows the numbers of her selected volunteers, in six groups, by age and sex at birth.

		Sex at birth	
		Female	**Male**
Age	**18-24**	51	37
	25-29	34	26
	30-35	21	18

Figure 2

Dahlia randomly chose 10 volunteers from each of the six groups in the table.

For each group, for lunch every day for four weeks, 5 were asked to eat 75 g of quinoa and 5 were asked to eat 75 g of couscous. Couscous is another edible grain, known to have little effect on blood glucose level.

Volunteers had their blood glucose levels measured by a monitor. Measurements from the monitor were obtained at the start of the four weeks, then weekly throughout the four weeks. Each measurement was taken at 10:00 on a Monday.

(a) State **both** the blocking factors Dahlia has used.

(1)

Age and Sex at birth

(b) Explain the purpose of taking the reading at the same time each week.

(1)

To reduce the experimental error arising from the variation in blood glucose levels at different times in the week.

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA

Question 4 continued

(c) State **two** other measures that Dahlia has taken to ensure the experiment can be easily replicated.

(2)

- She chose 10 volunteers from each category combination and split them equally for quinoa and couscous
- Everybody ate 75g of their grain
- Everybody ate their grain at the same time each day

etc.

Dahlia finds that the mean blood glucose level for the volunteers eating the quinoa each day for four weeks, reduces from 7.8 mmol/L to 7.6 mmol/L. She concludes that consumption of quinoa reduces blood glucose levels.

(d) Comment on the validity of this conclusion, giving reasons for your answer.

(2)

This may not be valid because

e.g.
- The blood glucose levels of those eating couscous may also have decreased so it may be due to another factor other than quinoa
- Although there is a decrease, it may not be significant since she doesn't appear to have accounted for the variance
- Her statement is too definite - there may be evidence to suggest her conclusion but it may not be true.

(e) State **two** improvements that Dahlia could make to her experiment.

(2)

- Use a larger sample of volunteers from each group
- Use a larger range of ages
- Take more measurements over the 4 weeks
- Extend the time period to 6 weeks

etc.

(Total for Question 4 is 8 marks)

5 Ayesha parks her bicycle at the train station on days when she catches the train to work.

She is considering moving home to a new location, and decides to find data on bicycle thefts at various train stations.

The railway stations with the most bicycle thefts in England are shown in **Figure 3**

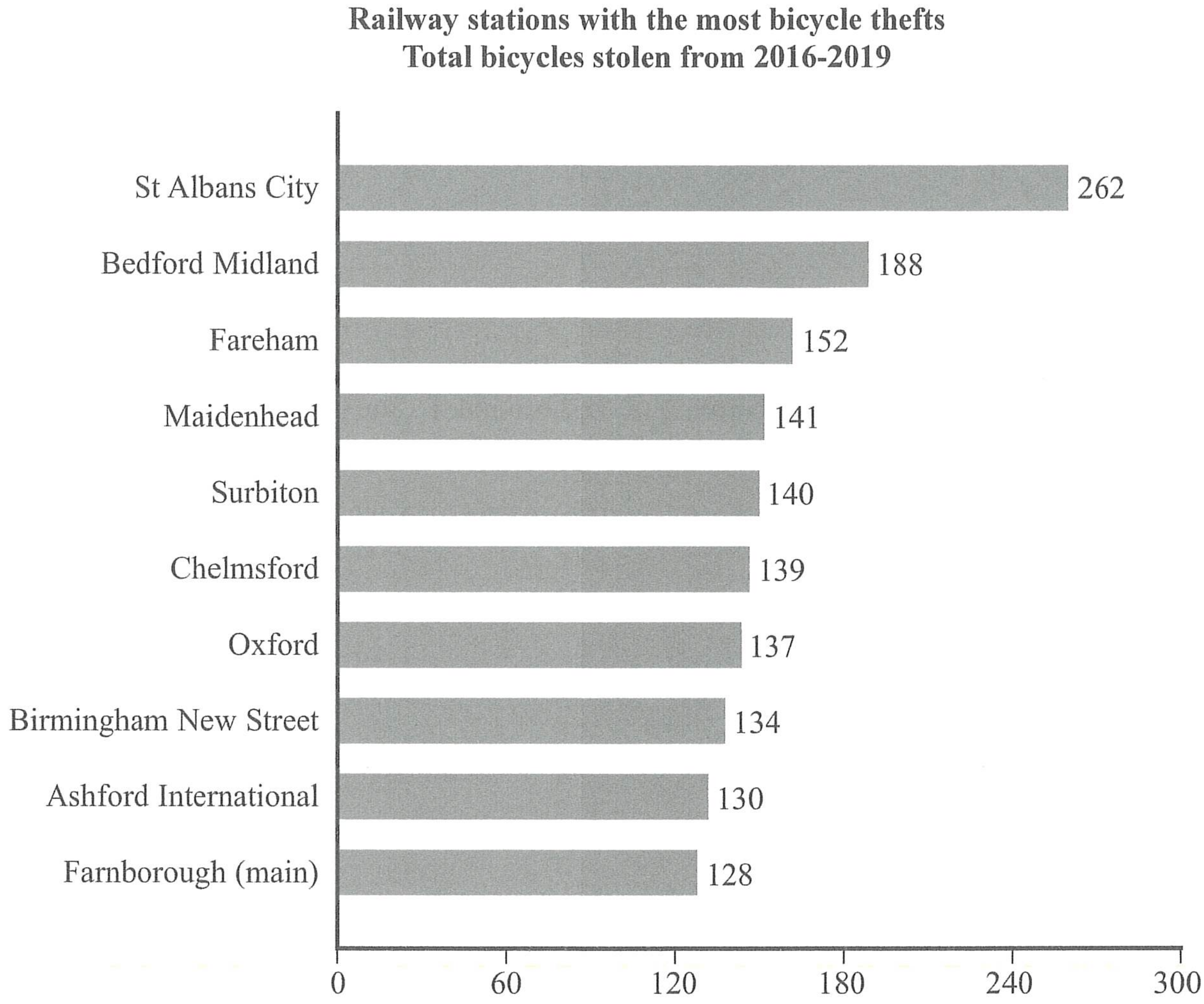


Figure 3

[Source: https://www.bbc.co.uk/news/uk-england/49154673, Data source: British transport police]

DO NOT WRITE IN THIS AREA

Question 5 continued

(a) Show that the number of bicycles stolen from St Albans City station is approximately 86% higher than the number stolen from Maidenhead station.

(2)

Number of bikes stolen from St. Albans = 262
Number of bikes stolen from Maidenhead = 141

$\frac{262}{141} = 1.858 \approx 1.86$

So the number of bikes stolen from St. Albans is approximately an 86% increase from that for Maidenhead.

Ayesha claims that a bicycle parked at St Albans City station is 86% more likely to be stolen than a bicycle parked at Maidenhead Station.

(b) Do you agree with Ayesha's claim?

Explain your answer.

(2)

This is not true. There is an 86% increase in the total number of bikes stolen in St. Albans compared to Maidenhead, but it doesn't represent the probability of a single bike being stolen.

Question 5 continued

As well as data on the railway stations with the most bicycle thefts in England, the British Transport Police also has data on the number of secure bicycle lockers available at each station.

Ayesha calculates Pearson's product moment correlation coefficient, r, between the number of secure bicycle lockers at a station and the total number of bicycles stolen at that station between 2016–19 for all stations in England.

(c) Suggest a practical reason

(i) why the value for r might be negative,

(1)

As the number of secure bike lockers increases, there will be more secure places to put bikes so stealing them may be harder, so the number of thefts may decrease.

(ii) why the value for r might be positive.

(1)

As the number of bike lockers increases, there may be more bikes at the station making it a target for thieves, so the number of thefts may increase

(d) Suggest **three other** pieces of data that should be considered in order to analyse how likely a bicycle is to be stolen from a given station.

(3)

- The general crime rate in the local area
- The level of security at the station
- The time of day the bikes were stolen
- How expensive the stolen bikes were worth etc.

(Total for Question 5 is 9 marks)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

BLANK PAGE

6 (a) Explain the difference between a discrete distribution and a continuous distribution.

(2)

A discrete distribution is where the outcomes can be counted and can only take certain values.

A continuous distribut is where the outcomes can be any number in a range of values, and must be measured.

Erin owns several vans that she uses for her parcel delivery company. She allocates 8 hours for the delivery of all the parcels in a van.

When a parcel is delivered, the van driver scans the bar code on the parcel.

To check that the drivers are delivering parcels on time, one parcel in each van is chosen at random. When the chosen parcel is delivered, Erin receives a notification.

Erin models the amount of time taken for a driver to deliver the randomly chosen parcel as a continuous uniform distribution from 0 to 8, as shown in **Figure 4**

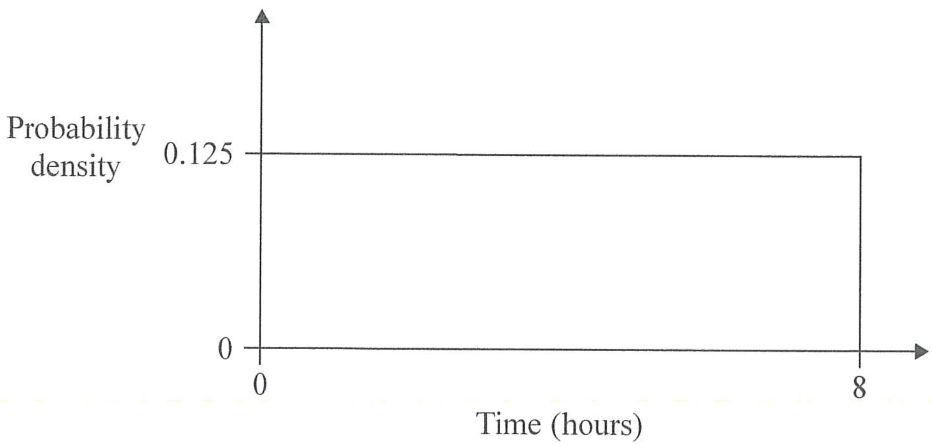
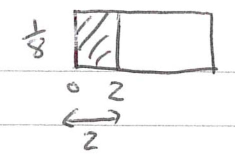


Figure 4

(b) Using Erin's model, find the probability that

(i) a driver scans the selected parcel within 2 hours of starting their deliveries,

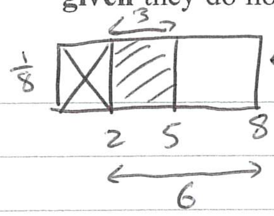
(1)



$= 2 \times \frac{1}{8} = \frac{1}{4} = 0.25$

Question 6 continued

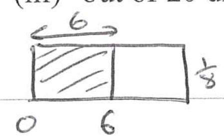
(ii) a driver scans the selected parcel within 5 hours of starting their deliveries, **given** they do not scan it during the first 2 hours, **(2)**



now out of $6 \times \frac{1}{8} = \frac{3}{4} = 0.75$

$$\frac{3 \times \frac{1}{8}}{0.75} = \underline{\underline{0.5}}$$

(iii) out of 20 drivers, at least 10 of them have delivered their parcel within 6 hours. **(3)**



$p = 6 \times \frac{1}{8} = 0.75$

Let X be the number of drives who deliver their parcel within 6 hours.

$X \sim B(20, 0.75)$

$P(X \geq 10)$

0 9 $\boxed{10 \quad 11 \quad 20}$

using calculator → 0.00394 $1 - 0.00394 = 0.996$

$= \underline{\underline{0.996}}$

(c) Give a comment

(i) in support of Erin's model, **(1)**

The parcel selected is at random so it could be as likely as any other to be scanned within 8 hours

(ii) against Erin's model. **(1)**

Some vans may deliver all parcels in a short space of time and not need the full 8 hours.

(Total for Question 6 is 10 marks)

7 The Poisson distribution is considered a suitable model for the occurrence of major earthquakes.

Harriet models the number of worldwide major earthquakes, occurring in a year, as a random variable X following a Poisson distribution with mean 16

[Source: https://policyadvice.net/insurance/insights/earthquake-statistics/]

(a) Using Harriet's model, find the probability there are

(i) exactly 10 major earthquakes occurring in one year,

(1)

$X \sim Po(16)$ per year

$P(X=10) = 0.0341$ (calculator)   0   9 $\boxed{10}$ 11 ...

(ii) 12 or more major earthquakes occurring in one year,

(2)

$P(X \geq 12)$

0 11 $\boxed{12 \quad 13 \quad \ldots}$

0.1270 $1-0.1270 = 0.873$

using calculator

$= \underline{0.873}$

(iii) more than 2, but fewer than 6, major earthquakes occurring in a 3-month period.

(3)

16 per 12 months $\Rightarrow$ $16 \div 4 = 4$ per 3 months

$X \sim Po(4)$ per 3 months

0 1 2 $\boxed{3 \quad 5}$ 6 7 ...

0.2381   0.7851

$P(2 < X < 6) = 0.7851 - 0.2381 = \underline{0.547}$

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 7 continued

Harriet wants to calculate the probability that there are exactly 30 major earthquakes occurring in a 2-year period, given that 13 major earthquakes occurred in the first year.

She calculates

$$\frac{P(Y = 30)}{P(X = 13)}$$

where Y is a random variable following the Poisson distribution with a mean of 32

(b) Explain why Harriet's calculation is incorrect, and how this probability should have been calculated.

(2)

$P(Y=30)$ is the probability of 30 major earthquakes over 2-years but does not account for 13 being specifically in the first year.

The correct calculation should have been $P(X=17)$. If 13 have already occurred in Year 1, then 17 will have to occur in year 2.

(c) Calculate the minimum whole number of days that are necessary for the probability of 1 or more major earthquakes occurring within that number of days to be at least 95%

Trial and improvement may be used for this question.

(4)

$X \sim Po\left(\frac{16}{365}\right)$ per day

$P(X \geq 1)$ $\quad$ 0 $\boxed{1 \quad 2 \quad \ldots}$

$\leq 5\%$ $\quad$ $\geq 95\%$

So $P(X=0) \leq 5\%$. $\quad$ (calculator)

$X \sim Po\left(\frac{16}{365} \times 68\right)$ per 68 days $\quad$ $P(X=0) = 0.0508 > 0.05$

$X \sim Po\left(\frac{16}{365} \times 69\right)$ per 69 days $\quad$ $P(X=0) = 0.0486 < 0.05$

So 69 days

(Total for Question 7 is 12 marks)

8 Ingrid is taking part in a chess tournament.

Each game she plays in the tournament is against a randomly selected player that she has not played before.

In this tournament, there are 3 players that Ingrid classifies as "stronger" than herself, 7 players she classifies as "similar" to herself and 5 players she classifies as "weaker" than herself.

Ingrid estimates her probability in this tournament that a game she plays ends in a win, draw or loss with the following probabilities, shown in **Figure 5**

		Ingrid's result		
		Win	**Draw**	**Loss**
Ingrid's classification	**Stronger**	15%	40%	45%
	Similar	25%	50%	25%
	Weaker	45%	40%	15%

Figure 5

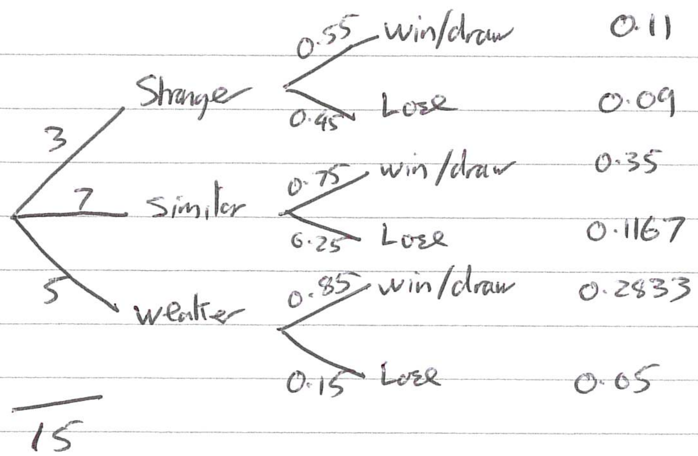
(a) Find the probability that the first person Ingrid plays is a weaker player than her.

(1)

Number of weaker playes = 5 so $\frac{5}{15} = \frac{1}{3} = 0.333$
total number of playes = 15

(b) Find the probability that Ingrid loses her first game.

(3)



$P(\text{Lose}) = 0.09 + 0.1167 + 0.05$

$= \underline{\underline{0.257}}$

Question 8 continued

(c) Find the probability that her first 3 games are against one stronger, one similar and one weaker player and that she wins all 3 games. Candidates play each other once only.

(4)

P(Strong/similar/Weak) (in that order) $= \dfrac{3 \times 7 \times 5}{15 \times 14 \times 13} = \dfrac{1}{26}$

Number of combinations:
St S, W
St W S,
S, St W
S, W St
W St S,
W S, St
$\}$ 6 ways.

win against each one: $0.15 \times 0.25 \times 0.45 = 0.016875$

so P(Strong, similar, weak in any order and win 3 times) $= \dfrac{1}{26} \times 6 \times 0.016875$

$= \underline{0.00389}$

Some chess tournaments do not randomly pair players, but follow a Swiss model where, after the first round, players play other players with similar results to themselves.

(d) If the tournament Ingrid is playing in follows the Swiss model, how would this change the probabilities used in (c)?

(3)

After the first match, if Ingrid wins then she is more likely to meet a stronger opponent in subsequent rounds

This means the probability of meeting weaker opponents will decrease and the probability of winning will also decrease.

So the probability in (c) will likely decrease too.

(Total for Question 8 is 11 marks)

TOTAL FOR PAPER IS 80 MARKS

BLANK PAGE

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA