Please check the examination details below before entering your candidate information

Candidate surname	Other names

Pearson Edexcel
Level 3 GCE

Centre Number □□□□□ Candidate Number □□□□

Tuesday 18 June 2019

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). Coloured pencils and highlighter pens must not be used.
- **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, statistical tests should be carried out at the 5% significance level.
- When a calculator is used, the answer should be given to three significant figures unless otherwise stated.

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 Eva has been asked by the manager of a gym to select a sample of members. Each one of these people will be sent a questionnaire to complete by post.

She has been told that half of the sample should be aged 60 years or over.

(a) Explain how Eva could select a stratified sample of 50 gym members to post the questionnaires to.

(4)

Eva has been told that she needs 50 completed questionnaires for her results analysis.

(b) Explain why your sampling method described in (a) is unlikely to be sufficient for Eva's analysis.

(1)

(c) Describe how you would change your sampling method described in (a) to allow for this.

(2)

(Total for Question 1 is 7 marks)

2 Dirk is working on a piece of statistics coursework.

He has been given a database containing various statistics for each of the countries of the world, and he has been asked to analyse the relationship between two variables of his choice.

He has chosen

- Number of domesticated chickens in 2014
- Number of human births in 2014

He decides to randomly select a sample of 10 countries to analyse, in order to save time and reduce the chance of errors when entering the data into his calculator.

(a) Explain how Dirk could have analysed the **whole data set** accurately and efficiently.

(1)

The data for Dirk's sample is shown in **Figure 1**.

Country	Human births (H)	Domesticated chickens (D)
Costa Rica	79 556	23 400
Mauritania	121 168	4 600
Croatia	39 423	9 803
Senegal	551 827	54 513
Pakistan	4 885 785	430 000
Colombia	830 539	149 078
Oman	113 533	4 600
Singapore	45 460	3 500
Greece	94 760	32 062
United Kingdom	802 219	159 000

(Data sources: CIA World Factbook, FAOSTAT)

Figure 1: Data for Dirk's sample

(b) For the data in Dirk's sample, find the value of Pearson's product-moment correlation coefficient between H (human births) and D (domesticated chickens).

Interpret this value in context.

(2)

(c) For the data in Dirk's sample, find the equation of the least squares regression line in the form $D = a + bH$.

Interpret your values of a and b in context.

(4)

(d) Explain why Dirk should not use the equation from part (c) to estimate the number of domesticated chickens in Liechtenstein (in 2014), which had 399 human births (in 2014).

(1)

Dirk humorously suggests that the extremely high value of Pearson's product-moment correlation coefficient is evidence that it is in fact chickens responsible for delivering babies (rather than storks).

(e) Explain why this reasoning is flawed.

Give an alternative explanation for this high value.

(2)

(Total for Question 2 is 10 marks)

3 Rhodri is a web analyst working for Wikipedia. He has been told that articles on Wikipedia that do not trend have an approximately constant rate of page views, with individual page views occurring randomly and independently.

During 2016, the Wikipedia article entitled '*Poisson distribution*' had a mean of 2.80 page views per minute (correct to 3 significant figures).

(a) Assuming that this article did not trend in 2016, find

(i) the probability that, in a randomly chosen one-minute period, the article has precisely 2 page views,

(2)

(ii) the probability that, in a randomly chosen five-minute period, the article has more than 20 page views.

(2)

Rhodri is analysing a daily log of page views for the article.

He notices an occurrence of an unexpectedly long time period of 2 minutes and 25 seconds between page views. He suspects that the website may have been down during that particular time period and therefore people were unable to access the article.

(b) Find the probability that the time between two randomly-selected consecutive page views was at least 2 minutes and 25 seconds.

(4)

(c) Does your answer to (b) support Rhodri's suspicion that the website was down? Explain your answer, using numerical evidence where appropriate.

(3)

Rhodri produces a bar chart showing page views per day for the '*Poisson distribution*' Wikipedia article in 2016. An extract from his bar chart is shown in **Figure 2**.

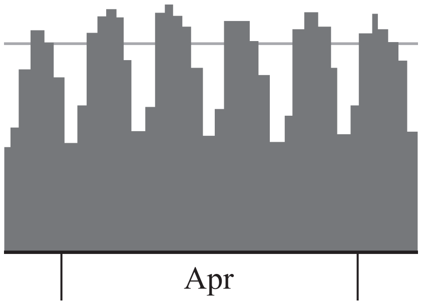


Figure 2

(d) Describe the variation shown in **Figure 2**.

Give a possible reason for this variation in context.

(2)

A second extract from Rhodri's bar chart is shown in **Figure 3**.

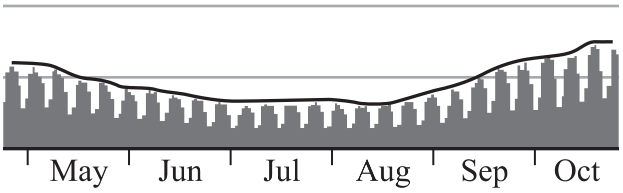


Figure 3

(e) Describe the variation in the height of the highest bars for part of the year.

Give a possible reason for this variation in context.

(2)

A third extract from Rhodri's bar chart is shown in **Figure 4**.

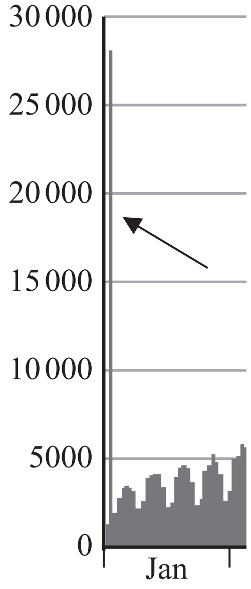
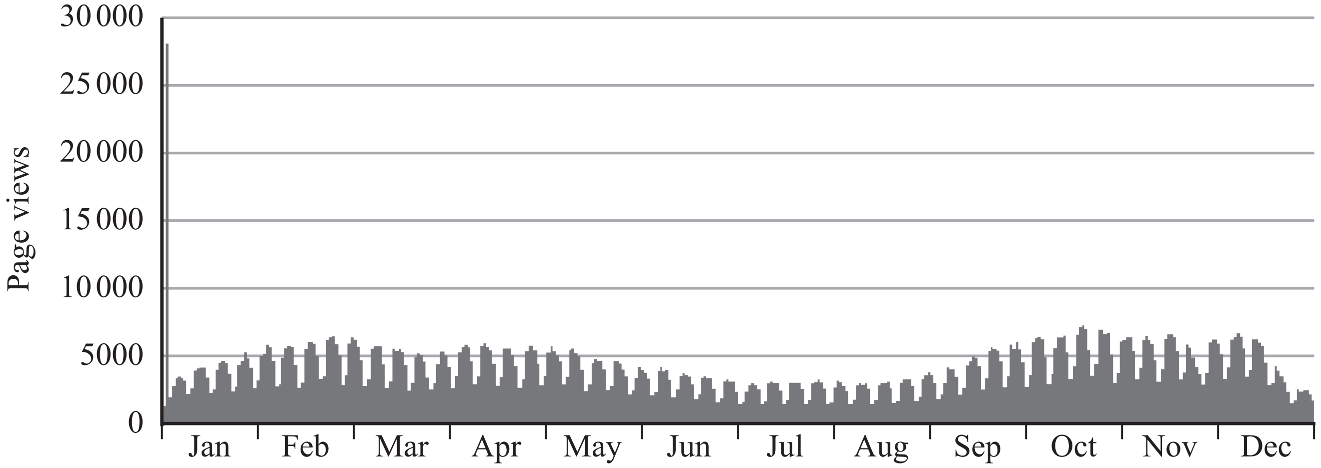


Figure 4

(f) Give **two** possible explanations, in context, for the one very tall bar in **Figure 4**.

(2)

Rhodri’s complete bar chart is shown in **Figure 5**.



(Data source: Wikipedia Pageviews Analysis)

Figure 5: Page views per day for the ‘*Poisson distribution*’ Wikipedia article in 2016

(g) Give **two** reasons why a Poisson distribution with a mean of 2.8 page views per minute might be unsuitable for modelling the data presented in **Figure 5**.

(2)

(Total for Question 3 is 19 marks)

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

4 Kayoko wants to investigate whether caffeine intake affects mental arithmetic ability in students at her university.

Describe how you would design an experiment for Kayoko to investigate this relationship.

You should try to minimise bias.

(5)

(Total for Question 4 is 5 marks)

5 Brutus works for a large vehicle-hire company with over 900 vehicles and multiple branches across Wales.

He is part of the planning team for a new branch to be opened next year.

He is currently investigating costs associated with car maintenance.

The costs for maintenance are split into servicing costs and repair costs. The data for 2018 is located in two separate tables in a database.

The **first five** rows of each table are presented in **Figure 6** and **Figure 7**.

Car_ID	Services_Count	Repairs_Needed	Latest_Mileage	Total_Servicing_Costs
01532	2	FALSE	34 201	£602.02
01561	1	FALSE	28 563	£213.46
01563	2	TRUE	24 033	£519.09
01566	2	FALSE	53 472	£711.87
01567	1	TRUE	42 118	£403.96

Figure 6: Table for 'Car servicing 2018'

Car_ID	Repairs	Labour_Hours	Total_Repair_Costs
01563	Offside dashboard heater element replaced	1	£53.20
01567	Bodywork - rust	2	£245.82
01580	Oil light - sensor	1	£75.22
01604	Nearside front rim and tyre replaced, front axle tested	3	£373.77
01621	Offside mirror - shattered, casing cracked	1	£189.10

Figure 7: Table for 'Car repairs 2018'

(a) Explain how you would use database software to get both maintenance costs (servicing and repairs) for each car in a single table.

(3)

Brutus uses this table to find the total maintenance costs for each car in 2018. He then uses this data to produce the histogram in **Figure 8**.

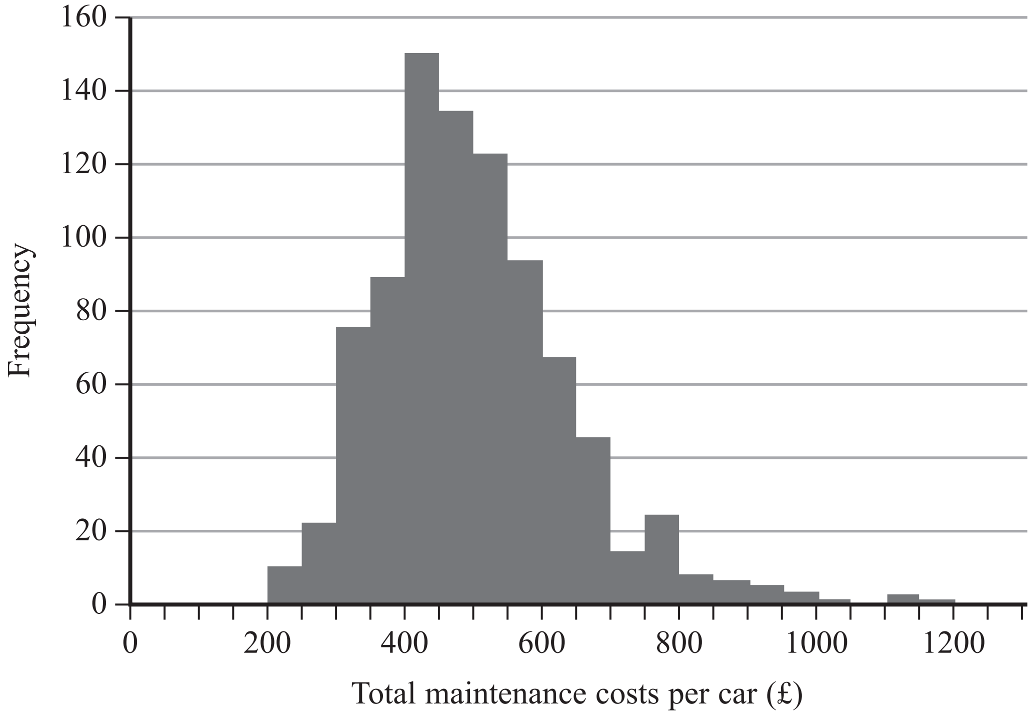


Figure 8: Total maintenance costs for cars in 2018

Brutus chooses to use a normal distribution to model this data.

(b) Explain **one** feature of **Figure 8** that

(i) may not support Brutus's choice of distribution to model this data,

(1)

(ii) supports Brutus's choice of distribution to model this data.

(1)

Brutus's next step is to investigate corresponding maintenance costs for vans.

He also chooses to model these costs as a normal distribution.

The data produces the following summary statistics.

	Mean	**SD**
Cars	£511.36	£168.65
Vans	£885.12	£232.78

Brutus has allocated £10 000 for **total** maintenance costs in the first year.

The new branch plans to buy **ten** new cars and **four** new vans.

(c) Making any necessary assumptions, use Brutus's models to estimate the probability that the new branch will have to pay more than £10 000 in **total** maintenance costs during the first year.

(6)

(d) Give **two** reasons **in context** why the estimate given in (c) may not be reliable.

Do not make further comment on the shape of the distribution.

(2)

(Total for Question 5 is 13 marks)

6 A bag contains 9 fair coins and 1 ‘double-header’ coin that has a head on **both** sides.

One coin is selected at random from this bag and tossed three times.

This coin shows heads on each toss.

Given this information, find the probability that the selected coin is the double-header.

You may find a tree diagram useful in answering this question.

(5)

Question 6 continued

(Total for Question 6 is 5 marks)

7 Zener cards were once used in experiments designed to detect mind-reading abilities.

Each Zener card is one of five designs, as presented in **Figure 9**.

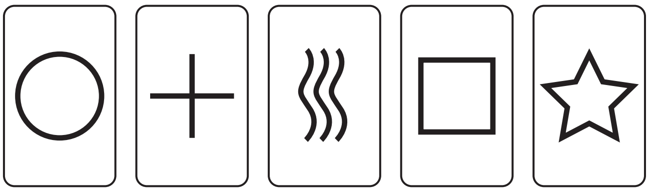


Figure 9: Five designs of Zener cards

The experiments involved one person looking at a randomly-chosen Zener card and another person, who could not see the card, writing down which card they thought it was. This was then repeated for a large number of Zener cards.

In a selection of experiments in 1933–1934, Hubert Pearce correctly identified 558 cards out of 1850.

(a) Using a suitable **approximate** distribution, estimate the probability of correctly identifying **at least** 558 cards out of 1850 by chance.

You should take all precautions to avoid rounding errors.

Give your answer in standard form to 3 significant figures.

(5)

The actual probability (not using an approximation) of correctly identifying at least 558 Zener cards out of 1850 is 2.19×10^{-25} (correct to 3 significant figures).

(b) Describe how you could explain this level of probability to a member of the general public, in terms of an event with a comparable probability.

You should use **calculations** involving **one** of the following:

- the probability of winning The National Lottery's Lotto jackpot with a single selection is approximately 1 in 14 million,
- any probabilities involving dice rolls.

(3)

During the experiments, Hubert Pearce was sitting in a university library, and the psychologist running the experiments was sitting in a classroom.

The psychologist turned over one Zener card per minute from a large shuffled pack of cards and wrote down the results. Hubert Pearce would try to guess the card and write down his guess.

Neither man was supervised during the experiments.

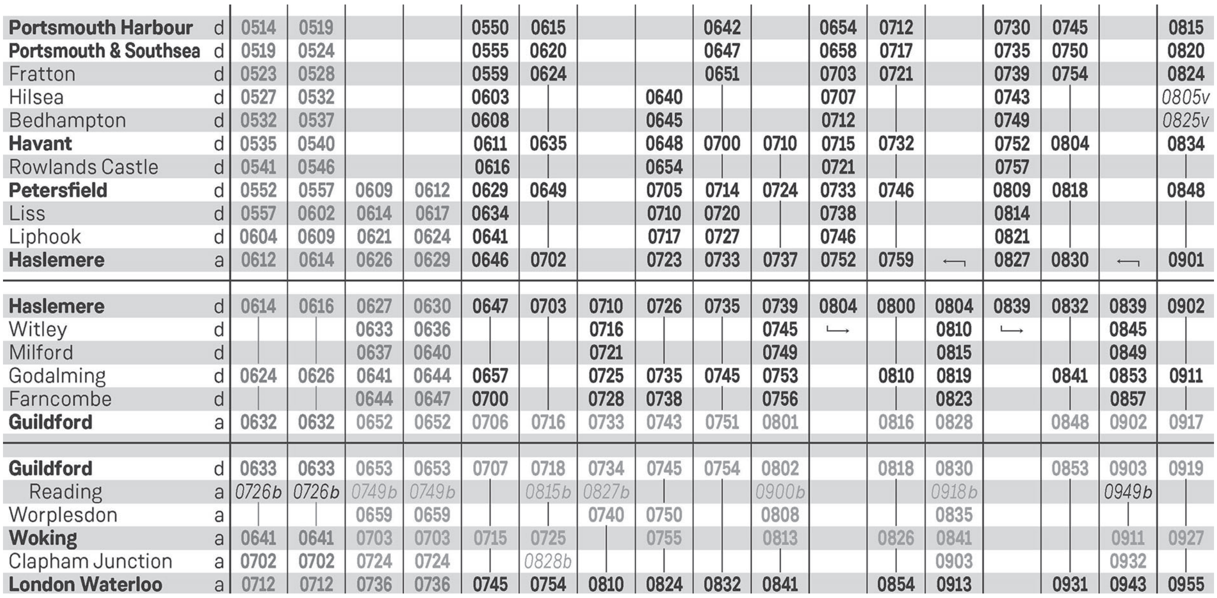
Based on these experiments, the psychologist declared Hubert Pearce a mind-reader.

(c) Do you agree with the psychologist's conclusion? Explain your answer.

(2)

(Total for Question 7 is 10 marks)

8 South Western Railway operates the train service that runs from Portsmouth to London Waterloo. A page from the timetable is shown in **Figure 10**.



Portsmouth Harbour	d	0514	0519			0550	0615			0642		0654	0712		0730	0745		0815
Portsmouth & Southsea	d	0519	0524			0555	0620			0647		0658	0717		0735	0750		0820
Fratton	d	0523	0528			0559	0624			0651		0703	0721		0739	0754		0824
Hilsea	d	0527	0532			0603	│		0640	│		0707	│		0743	│		*0805v*
Bedhampton	d	0532	0537			0608	│		0645	│		0712	│		0749	│		*0825v*
Havant	d	0535	0540			0611	0635		0648	0700	0710	0715	0732		0752	0804		0834
Rowlands Castle	d	0541	0546			0616	│		0654	│	│	0721	│		0757	│		│
Petersfield	d	0552	0557	0609	0612	0629	0649		0705	0714	0724	0733	0746		0809	0818		0848
Liss	d	0557	0602	0614	0617	0634	│		0710	0720	│	0738	│		0814	│		│
Liphook	d	0604	0609	0621	0624	0641	│		0717	0727	│	0746	│		0821	│		│
Haslemere	a	0612	0614	0626	0629	0646	0702		0723	0733	0737	0752	0759	⟵	0827	0830	⟵	0901
Haslemere	d	0614	0616	0627	0630	0647	0703	0710	0726	0735	0739	0804	0800	0804	0839	0832	0839	0902
Witley	d	│	│	0633	0636	│	│	0716	│	│	0745	⟶	│	0810	⟶	│	0845	│
Milford	d	│	│	0637	0640	│	│	0721	│	│	0749		│	0815		│	0849	│
Godalming	d	0624	0626	0641	0644	0657	│	0725	0735	0745	0753		0810	0819		0841	0853	0911
Farncombe	d	│	│	0644	0647	0700	│	0728	0738	│	0756		│	0823		│	0857	│
Guildford	a	0632	0632	0652	0652	0706	0716	0733	0743	0751	0801		0816	0828		0848	0902	0917
Guildford	d	0633	0633	0653	0653	0707	0718	0734	0745	0754	0802		0818	0830		0853	0903	0919
Reading	a	*0726b*	*0726b*	*0749b*	*0749b*	│	*0815b*	*0827b*	│	│	*0900b*		│	*0918b*		│	*0949b*	│
Worplesdon	a	│	│	0659	0659	│	│	0740	0750	│	0808		│	0835		│	│	│
Woking	a	0641	0641	0703	0703	0715	0725	│	0755	│	0813		0826	0841		│	0911	0927
Clapham Junction	a	0702	0702	0724	0724	│	*0828b*	│	│	│	│		│	0903		│	0932	│
London Waterloo	a	0712	0712	0736	0736	0745	0754	0810	0824	0832	0841		0854	0913		0931	0943	0955

(Source: South Western Railway)

Figure 10: Extract from Portsmouth to London Waterloo timetable

Mark is a consultant who is currently modelling the waiting times of passengers travelling from Godalming to London Waterloo on a Monday morning.

In a simple model, Mark assumes that a passenger arrives at the platform at Godalming Station randomly at some time between 7:00 and 8:00, with all times equally likely.

You should assume that a passenger can board a train up to the instant it departs the station.

(a) Using this model, and making any necessary assumptions:

(i) show that the probability that the passenger will have to wait less than 5 minutes for a train is $\frac{1}{3}$

You may find a diagram useful.

(2)

(ii) find the probability that the passenger will have to wait less than 10 minutes for a train.

(2)

Mark now models the waiting times of such a passenger who uses the train every weekday.

(b) Using this model, find the probability that the passenger will have to wait less than 5 minutes for at least two days in a week (Monday to Friday).

(2)

Portsmouth Harbour	d	0514	0519			0550	0615			0642		0654	0712		0730	0745		0815
Portsmouth & Southsea	d	0519	0524			0555	0620			0647		0658	0717		0735	0750		0820
Fratton	d	0523	0528			0559	0624			0651		0703	0721		0739	0754		0824
Hilsea	d	0527	0532			0603	│		0640	│		0707	│		0743	│		0805v
Bedhampton	d	0532	0537			0608	│		0645	│		0712	│		0749	│		0825v
Havant	d	0535	0540			0611	0635		0648	0700	0710	0715	0732		0752	0804		0834
Rowlands Castle	d	0541	0546			0616	│		0654	│	│	0721	│		0757	│		│
Petersfield	d	0552	0557	0609	0612	0629	0649		0705	0714	0724	0733	0746		0809	0818		0848
Liss	d	0557	0602	0614	0617	0634	│		0710	0720	│	0738	│		0814	│		│
Liphook	d	0604	0609	0621	0624	0641	│		0717	0727	│	0746	│		0821	│		│
Haslemere	a	0612	0614	0626	0629	0646	0702		0723	0733	0737	0752	0759	←	0827	0830	←	0901
Haslemere	d	0614	0616	0627	0630	0647	0703	0710	0726	0735	0739	0804	0800	0804	0839	0832	0839	0902
Witley	d	│	│	0633	0636	│	│	0716	│	│	0745	→	│	0810	→	│	0845	│
Milford	d	│	│	0637	0640	│	│	0721	│	│	0749		│	0815		│	0849	│
Godalming	d	0624	0626	0641	0644	0657	│	0725	0735	0745	0753		0810	0819		0841	0853	0911
Farncombe	d	│	│	0644	0647	0700	│	0728	0738	│	0756		│	0823		│	0857	│
Guildford	a	0632	0632	0652	0652	0706	0716	0733	0743	0751	0801		0816	0828		0848	0902	0917
Guildford	d	0633	0633	0653	0653	0707	0718	0734	0745	0754	0802		0818	0830		0853	0903	0919
Reading	a	0726b	0726b	0749b	0749b	│	0815b	0827b	│	│	0900b		│	0918b		│	0949b	│
Worplesdon	a	│	│	0659	0659	│	│	0740	0750	│	0808		│	0835		│	│	│
Woking	a	0641	0641	0703	0703	0715	0725	│	0755	│	0813		0826	0841		│	0911	0927
Clapham Junction	a	0702	0702	0724	0724	│	0828b	│	│	│	│		│	0903		│	0932	│
London Waterloo	a	0712	0712	0736	0736	0745	0754	0810	0824	0832	0841		0854	0913		0931	0943	0955

(Source: South Western Railway)

Figure 10: Extract from Portsmouth to London Waterloo timetable

Mark then decides to model a passenger's arrival time at Godalming Station using a normal distribution.

He measures the time, T, in **minutes after 7:00**.

In his model, T has mean 30 and standard deviation 10

(c) Using this new model, find the approximate probability that the passenger will have to wait less than 10 minutes for a train on a Monday morning.

Give your answer to 3 decimal places.

You may find a diagram useful.

(3)

Question 8 continued

(d) State an assumption you have made about the trains for the probabilities found in parts (a)–(c) to be reliable.

Comment on the validity of this assumption.

(2)

(Total for Question 8 is 11 marks)

TOTAL FOR PAPER IS 80 MARKS

BLANK PAGE