Write your name here

Candidate surname	Other names

Pearson Edexcel Level 3 GCE

Centre Number	Candidate Number

Mathematics

Advanced

Statistics

Practice Questions

Paper Reference **9MA0/31**

You must have:
Mathematical Formulae and Statistical Tables, calculator

Total Marks

This is a collection of stand alone Statistics practice questions written as an additional resource for the GCE 2017 Mathematics specification.

- There are **11 questions** in this document.
- The marks for each question are shown in brackets.
- The questions are ramped in order of difficulty.
- Mark schemes can be found in the accompanying document on the Pearson website and Emporium.

This **is not** an exam paper so there is no time allocation or a set number of total marks. Teachers can use the questions as they wish to support teaching and learning. If any of the questions are being set as a test, students should be advised to follow the standard guidance for A Level Statistics exams:

Instructions

- Use black ink or ball-point pen.
- If a pencil is used for diagrams/sketches/graphs it must be dark (HB or B).
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.

Information

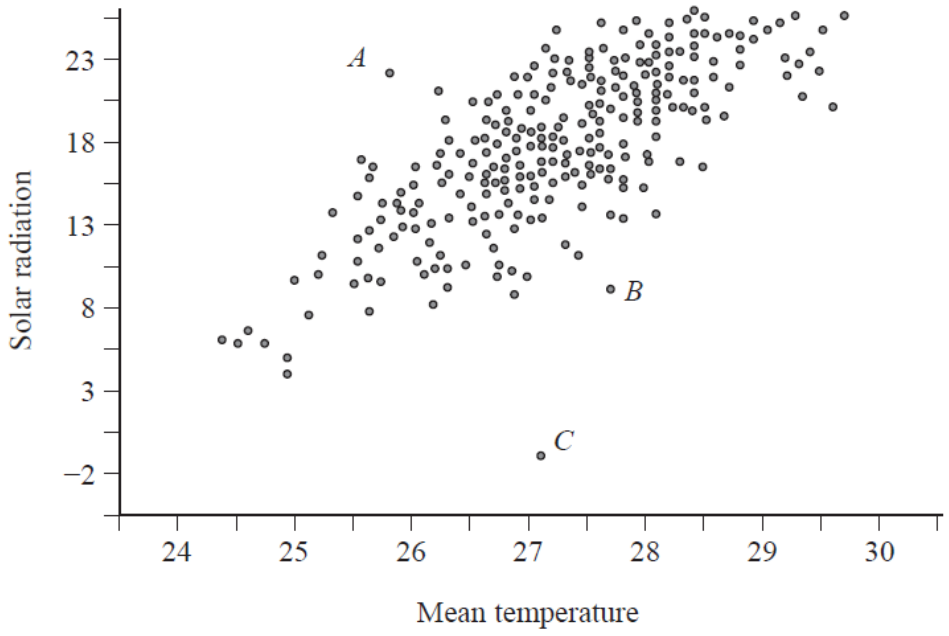
- The marks for each question are shown in brackets - use this as a guide as to how much time to spend on each question.

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.

1. Some researchers collected data on the levels of solar radiation and surface temperature for a specific site in Malaysia.

The data was analysed and the scatter diagram below was created.



The 3 data points, labelled *A*, *B* and *C* in the scatter diagram were commented on by the researchers.

(a) Suggest why the researchers commented on these 3 points.

(1)

One of the researchers calculates the product moment correlation coefficient for these data as 0.7473

(b) Comment on whether or not this value is consistent with the data shown in the scatter diagram.
Give a reason for your answer.

(1)

(Total for Question 1 is 2 marks)

(Image source: Scatter plot of solar radiation versus surface temperature modified from 'Relationship between the solar radiation and surface temperature in Perlis', Daut, I. Yusof, MI. Ibrahim, S. Irwanto, M. and Nsurface, G (2012) https://www.researchgate.net/publication/269340779_Relationship_between_the_Solar_Radiation_and_Surface_Temperature_in_Perlis
[first accessed August 2020])

2. The admissions director of a large university is studying data for the previous year's applicants to the university.

(a) Explain how she could select a systematic sample of size 100 from the 16 000 applicants to the university for the previous year.

(2)

(b) Given that the university made offers to 4000 of these applicants,

(i) use a suitable model to estimate the probability that her sample contains more than 25 applicants who received offers from the university.

(ii) State an assumption you have made in using your model.

(4)

(Total for Question 2 is 6 marks)

3. An organisation trains its drivers to one of 3 levels – Basic, Standard and Advanced.

They have a record of the drivers that they trained in the last year and those trained more than 1 year ago.

The table gives information about the types of drivers and the time since training.

		Driver level		
		Basic	**Standard**	**Advanced**
Time since training	**Trained in last year**	142	207	86
	Trained more than 1 year ago	180	161	144

The organisation plans to take a sample of 80 drivers stratified by driver level and time since training.

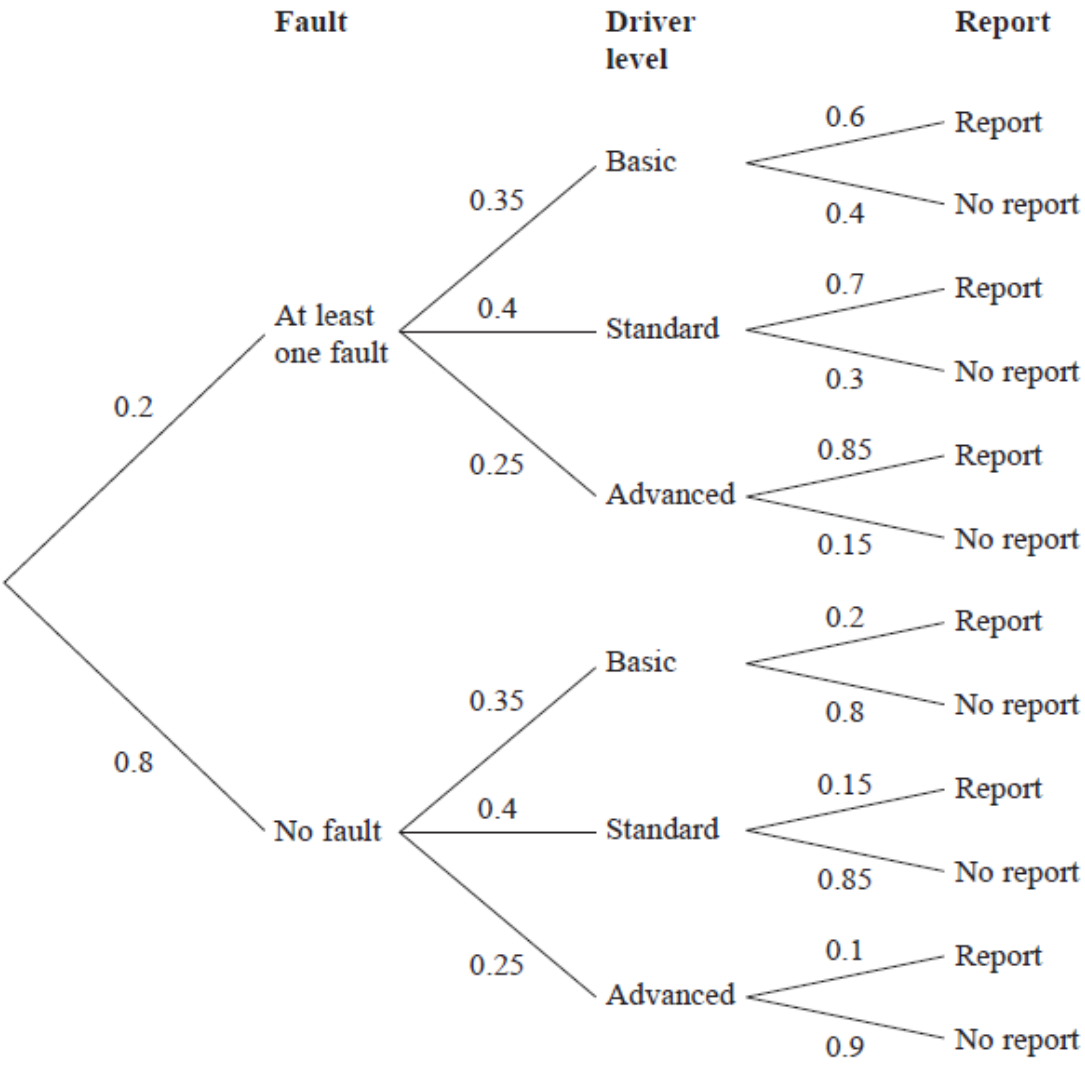
(a) Work out how many Advanced drivers, trained more than 1 year ago, should be sampled.

(2)

Each driver is allocated a random vehicle at the start of their shift and required to check it and report any faults.

The tree diagram below gives information about the probability of:

- the vehicle having a fault
- the level of the driver checking the vehicle
- a fault being reported


Fault
Driver level
Report
0.2
At least one fault
0.35
Basic
0.6
Report
0.4
No report
0.4
Standard
0.7
Report
0.3
No report
0.25
Advanced
0.85
Report
0.15
No report
0.8
No fault
0.35
Basic
0.2
Report
0.8
No report
0.4
Standard
0.15
Report
0.85
No report
0.25
Advanced
0.1
Report
0.9
No report

On Monday, all vehicles are allocated and have been checked by their drivers.
A vehicle is chosen at random.

(b) Find the probability that the vehicle has at least one fault which was not reported.

(3)

A fault report is chosen at random and was from a Basic level driver.

(c) Find the probability that the vehicle has at least one fault.

(3)

(Total for Question 3 is 8 marks)

---

4. Jon is using the large data set to carry out investigations into the weather.
He randomly selects 10 days from the large data set for Leuchars in 2015.
The daily total sunshine, in hours, for each of these days is shown below

5.2	4.8	6.2	0.2	0.2	10.5	0.8	2.9	4.6	8.3

(a) Find the value of the median of this sample.

(2)

Given that the lower quartile is 0.8, the upper quartile is 6.2 and that outliers may be defined as values outside the interval

$$(Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR})$$

(b) show that there are no outliers in this sample.

(3)

Jon uses the data from his random sample of 10 days from Leuchars.
He calculates the product moment correlation coefficient between daily total sunshine and daily total rainfall for this sample.

His value is $r = -0.7414$

(c) Using Jon's data test whether or not there is evidence of a negative correlation between daily total sunshine and daily total rainfall at Leuchars in 2015.
You should

- state your hypotheses clearly
- use a 5% level of significance
- state the critical value used in the test

(3)

(d) Using your knowledge of the large data set,

(i) state the units used for the daily total rainfall.

Jon randomly selects a day from Leuchars in 2015 from the large data set.
It has 15.9 hours of daily total sunshine.

(ii) Explain why this day cannot have been in October.

(2)

Jon wishes to use the daily total sunshine to try to predict the daily total rainfall on a particular day.

(e) (i) State which would be the explanatory variable in this case.

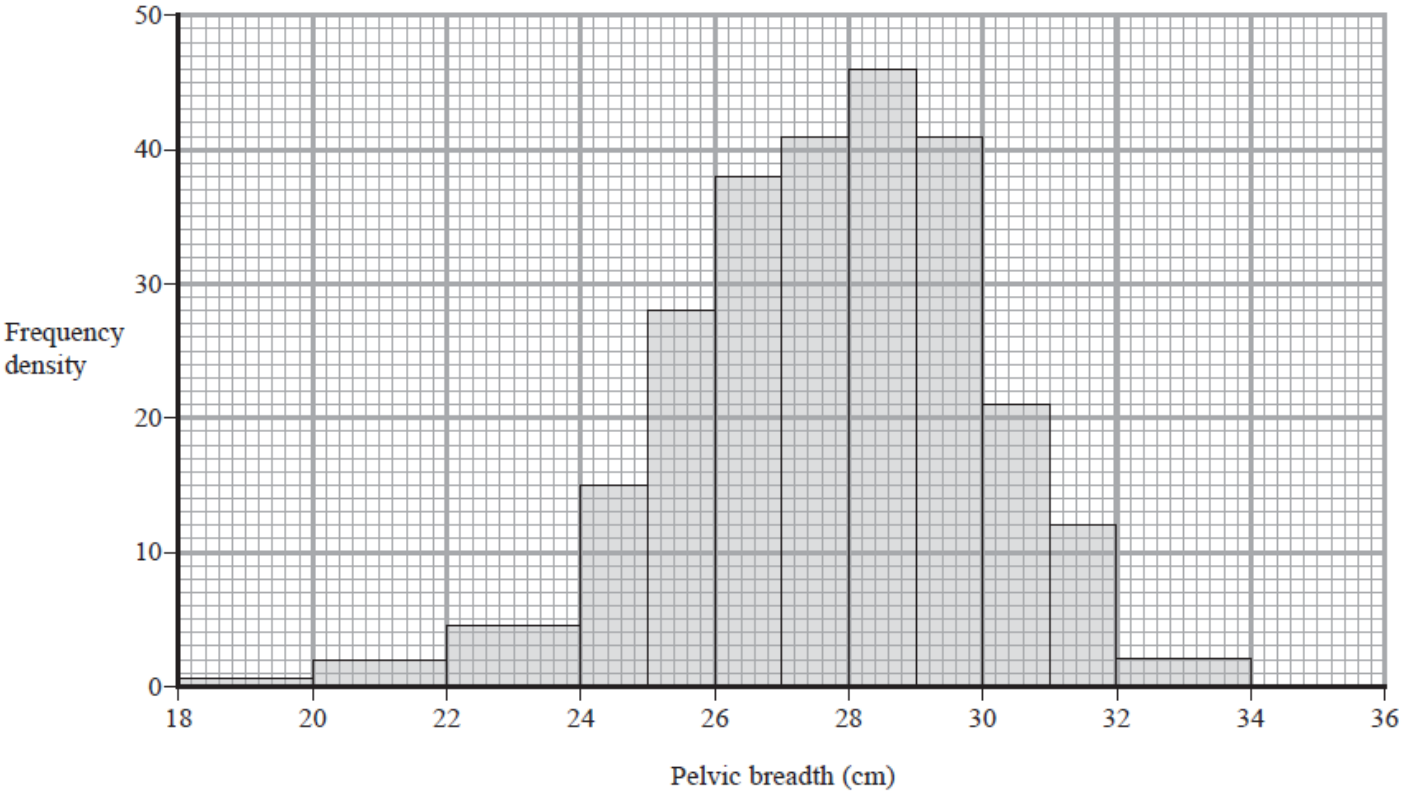
(ii) Explain why Jon's product moment correlation coefficient does not necessarily justify predicting the daily total rainfall from the daily total sunshine.

(2)

(Total for Question 4 is 12 marks)

5. A scientist is collecting data about body measurements in adults.

The results for pelvic breadth, in centimetres, for the female adults are summarised by the histogram and summary statistics below.



(a) Calculate the frequency of female adults with pelvic breadths above 31cm.

(2)

Statistical software was used to calculate the following summary statistics for these data.

$n = 260$ $\sum x = 7171.2$ $S_{xx} = 1379.0$

(b) Calculate the mean and the standard deviation for these data.

(3)

An outlier is defined as a value
more than 3 × standard deviation above the mean
more than 3 × standard deviation below the mean

(c) Based on the information given, show that there is at least one outlier in the pelvic breadths for female adults.

(3)

Summary statistics were also produced for the pelvic breadths, in centimetres, for a similar sized sample of male adults.

	Mean	Standard deviation
Males	28.1	1.95

The scientist believes that the males have a larger and more variable pelvic breadth.

(d) State, giving reasons, whether or not the statistics support the scientist's beliefs.

(2)

The data for male pelvic breadth contained no outliers.

(e) Without doing any further calculations

(i) explain the effect of omitting the female outliers on your calculations in part (b)

(2)

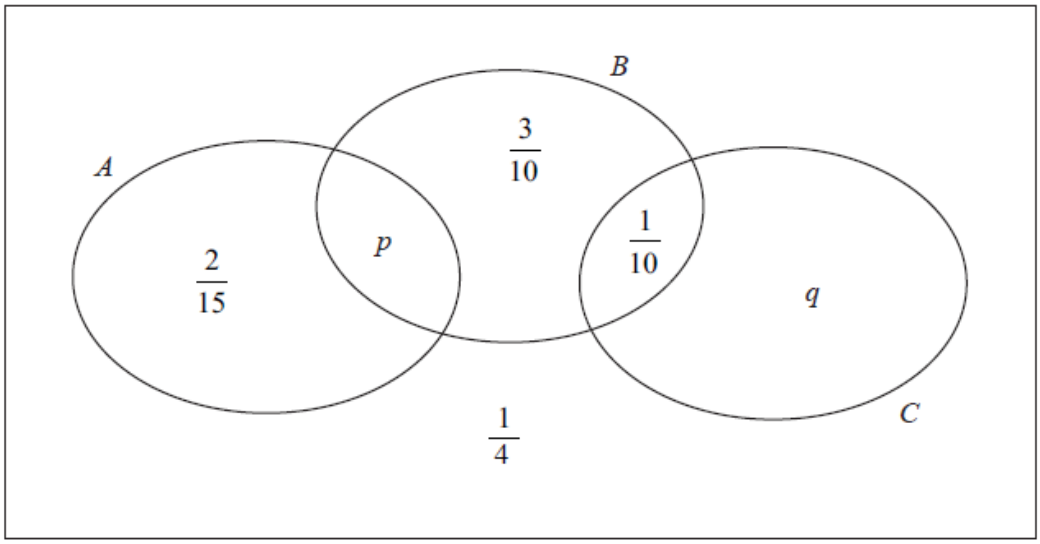
(ii) comment on the effect this may have had on your answer to part (d)

(1)

(Total for Question 5 is 13 marks)

(Data source: Pelvic breadth data manipulated from data sets found in 'http://mathsci.solano.edu/mac/Minitab%20Data/Utts%20Datasets/readme/readme.html#handheight') [first accessed August 2020]

6. The events A, B and C occur with the probabilities shown in the Venn diagram below.



(a) Explain why events A and C cannot be independent.

(1)

(b) Find P([$A \cup C$]')

(1)

Given that events A and B are independent,

(c) find the value of p.

(6)

(d) Find the exact value of P(A|[$B \cup C$]')

(2)

(Total for Question 6 is 10 marks)

7. An internet service provider (ISP) claims to have a mean download speed of at least 195 Mbps in *Seapron*.
A random sample of 36 properties in *Seapron* which use this ISP is taken.
The download speed, x Mbps, is recorded for each of these properties.
The data is summarised below

$$\sum x = 6786 \qquad \sum x^2 = 1\,307\,268$$

(a) Calculate

(i) the mean of these download speeds,

(ii) the variance of these download speeds.

(3)

The standard deviation of the download speed from this ISP in *Seapron* is known to be 30 Mbps.

Assuming that the download speed may be modelled by a normal distribution,

(b) test whether or not the sample provides evidence that the mean download speed in *Seapron* is less than the ISP claims.
State your hypotheses clearly and use a 5% level of significance.

(5)

Sam models the download speed from this ISP in *Seapron*, X, by

$$X \sim \mathrm{N}(195, 30^2)$$

(c) Calculate $\mathrm{P}(X < 125)$

(1)

Sam is examining the raw data from the sample summarised above.
Sam finds that there are 3 sample values of less than 125 Mbps in the sample of 36 values.

(d) (i) Use the answer to part (c) to calculate the probability of there being at least 3 sample values of less than 125 Mbps.

(ii) Explain why the hypothesis test in part (b) might not be valid.

(3)

(Total for Question 7 is 12 marks)

8. Kathleen is exploring the large data set.

She wants to know if there is a correlation between temperatures in different locations.

Here are the daily mean temperatures for the first 10 days of May 2015 for Beijing and for Perth.

Beijing	17.5	20.0	19.2	18.5	21.1	17.1	18.8	18.0	13.0	9.7
Perth	15.8	16.4	16.1	9.7	12.0	13.8	14.0	15.2	15.0	16.5

(a) Calculate the product moment correlation coefficient for the daily mean temperatures for Beijing and for Perth for the first 10 days of May 2015.

(1)

Kathleen selects a random sample of 30 days from 2015.

She calculates the product moment correlation coefficient for daily mean temperatures for these 30 days for Heathrow and Hurn.

The product moment correlation coefficient for these data is 0.8198

(b) Test whether or not the correlation between daily mean temperatures for Heathrow and for Hurn is positive.
You should
- state your hypotheses clearly
- use a 5% level of significance
- state the critical value used in the test

(3)

For these 30 days she also calculates the product moment correlation coefficient for daily mean temperatures for Heathrow and Leuchars.

The product moment correlation coefficient for these data is 0.5612

(c) Compare the product moment correlation coefficients for daily mean temperatures for Heathrow and Hurn and for Heathrow and Leuchars.
You should
- comment on the strength of any relationships
- give an interpretation of these correlation coefficients
- suggest a reason for the difference in the values of the correlation coefficients

(3)

(Total for Question 8 is 7 marks)

9. An ornithologist has data on the number of breeding pairs of bald eagles for selected years from 1963 to 2000.

The ornithologist believes that the number of breeding pairs of bald eagles can be modelled by

$$x = a(b^T) \qquad \text{(I)}$$

where T is the number of years since 1950, x is the number of breeding pairs of bald eagles and a and b are constants.

(a) Show that the equation can be written in the form $\log_{10} x = mT + c$

(3)

The ornithologist codes the data using the coding $Y = \log_{10} x$ and $X = T$ and obtains the model

$Y = 0.0348X + 2.0827$

(b) Determine the values of a and b.

(3)

The ornithologist wants to use the model (I) to estimate the number of breeding pairs of bald eagles in 2020.

(c) Explain why this would not be appropriate.

(1)

There is no record of the number of breeding pairs of bald eagles for 1970.

(d) Use the model (I) to estimate how many breeding pairs of bald eagles there would have been in 1970.

(2)

(Total for Question 9 is 9 marks)

(Data source: Data on Bald Eagle pairs from 'https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/exponential-relationships-1-of-6/'
[first accessed August 2020]

10. A machine puts crisps into packets.
The weight of crisps in each packet, M grams, follows a normal distribution with mean 40 g

Given that 20% of packets contain more than 42 g of crisps,

(a) find, to 2 decimal places, the value of k such that $P(k < A < 40) = 0.40$

(5)

Eighteen packets of crisps are selected at random.

(b) Find the probability that fewer than 3 of these packets contain more than 42 g of crisps.

(2)

A second machine makes larger packets of crisps of weight Y grams, where Y is normally distributed with standard deviation 7.5 g

A manager believes that the mean weight of crisps put into each packet is greater than 175 g

A random sample of 10 packets from this second machine was found to have a mean weight of crisps of 178.5 g

(c) Test whether or not the mean weight of crisps in larger packets, filled by the second machine, is greater than 175 g
State your hypotheses clearly and use a 5% level of significance.

(5)

(Total for Question 10 is 12 marks)

11. A village fete has a stall with a game called 'hook-a-duck'.
Players use a hook on the end of a piece of string to catch a plastic duck floating in a pool.
Each duck looks the same but has a number written on the bottom; 1, 2, 3 or 4
There are 20 ducks in the pool, 5 with each number.

Each time a duck is hooked it is put back in the pool ready for the next player.

(a) State the distribution which may be used to model the number on a duck hooked by a randomly selected player.

(1)

Calculate the probability that of the first 3 ducks hooked,

(b) all 3 have the same number,

(2)

(c) exactly 2 of them have the same number.

(3)

During the day 200 ducks are hooked.
The random variable X is defined as the number of ducks out of 200 that have an even number.

(d) State an appropriate distribution which may be used to model the random variable X.

(1)

(e) State a necessary assumption about hooking the ducks which must be true for this distribution to be a good model.

(1)

(f) Use this model to calculate the probability that, out of the 200 ducks hooked that day, at least 110 ducks have even numbers.

(2)

Jamie suggests using a normal approximation to estimate P(X □110)

(g) State, giving a reason, whether or not this suggestion is reasonable.

(1)

(h) Calculate the error in using Jamie's suggestion.
You should show your working clearly and give your answer to 2 significant figures.

(4)

(Total for Question 11 is 15 marks)