Centre No.						Paper Reference	Surname	Initial(s)
Candidate No.						6 6 8 3 / 0 1	Signature	

Paper Reference(s)

6683/01

Edexcel GCE

Statistics S1

Solutions

Advanced/Advanced Subsidiary

Thursday 9 June 2005 – Morning

Time: 1 hour 30 minutes

Materials required for examination	Items included with question papers
Mathematical Formulae (Lilac or Green)	Nil

Candidates may use any calculator EXCEPT those with the facility for symbolic algebra, differentiation and/or integration. Thus candidates may NOT use calculators such as the Texas Instruments TI 89, TI 92, Casio CFX 9970G, Hewlett Packard HP 48G.

Examiner's use only

Team Leader's use only

Question Number	Leave Blank
1	
2	
3	
4	
5	
6	
7	
Total	

Instructions to Candidates

In the boxes above, write your centre number, candidate number, your surname, initial(s) and signature.
Check that you have the correct question paper.
You must write your answer for each question in the space following the question.
Values from the statistical tables should be quoted in full. When a calculator is used, the answer should be given to an appropriate degree of accuracy.

Information for Candidates

A booklet 'Mathematical Formulae and Statistical Tables' is provided.
Full marks may be obtained for answers to ALL questions.
The marks for individual questions and the parts of questions are shown in round brackets: e.g. **(2)**.
There are 7 questions in this question paper.
The total for this question paper is 75.
There are 20 pages in this question paper. Any blank pages are indicated.

Advice to Candidates

You must ensure that your answers to parts of questions are clearly labelled.
You must show sufficient working to make your methods clear to the examiner. Answers without working may gain no credit.

Printer's Log. No.
N20910A
W850/R6683/57570 4/7/3/3/2/6/3/

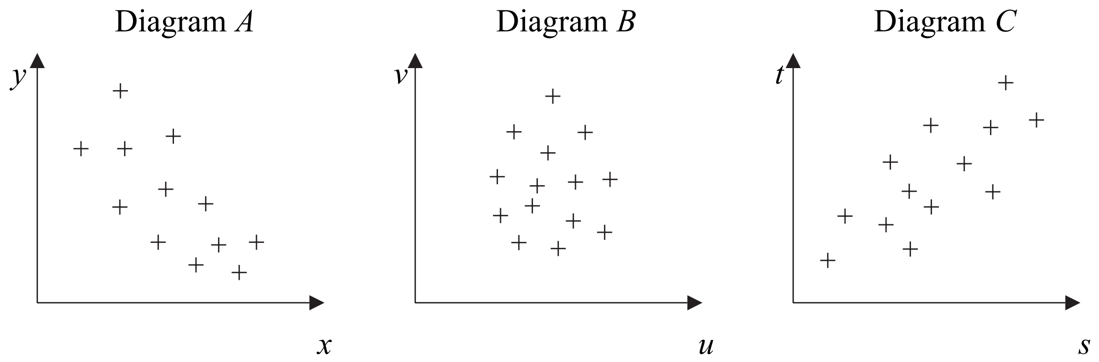
N 2 0 9 1 0 A 0 1 2 0

Turn over

edexcel

Leave blank

1. The scatter diagrams below were drawn by a student.



The student calculated the value of the product moment correlation coefficient for each of the sets of data.

The values were

$$0.68 \qquad -0.79 \qquad 0.08$$

Write down, with a reason, which value corresponds to which scatter diagram.

(6)

A −0.79 Negative correlation apparent

B 0.08 No correlation apparent

C 0.68 Positive correlation apparent

Leave blank

2. The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners	cf	FD
41–45	4	4	4 ÷ 5 = 0.8
46–50	19	23	19 ÷ 5 = 3.8
51–60	53	76	53 ÷ 10 = 5.3
61–70	37	113	37 ÷ 10 = 3.7
71–90	15	128	15 ÷ 20 = 0.75
91–150	6	134	6 ÷ 60 = 0.1

(a) Give a reason to justify the use of a histogram to represent these data. **(1)**

Grouped frequency table for continuous data

(b) Calculate the frequency densities needed to draw a histogram for these data.
(DO NOT DRAW THE HISTOGRAM)
(2)

(c) Use interpolation to estimate the median Q_2, the lower quartile Q_1, and the upper quartile Q_3 of these data.
(4)

The mid-point of each class is represented by x and the corresponding frequency by f. Calculations then give the following values

$$\Sigma fx = 8379.5 \quad \text{and} \quad \Sigma fx^2 = 557489.75$$

(d) Calculate an estimate of the mean and an estimate of the standard deviation for these data.
(4)

One coefficient of skewness is given by

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}.$$

(e) Evaluate this coefficient and comment on the skewness of these data.
(4)

(f) Give another justification of your comment in part (e).
(1)

c) 134 ÷ 4 = 33.5 Q_1 at 33.5 item

134 ÷ 2 = 67 Q_2 at 67 item

134 ÷ 4 × 3 = Q_3 at 100.5 item

$Q_1 = 50.5 + \frac{10.5}{53} \times 10 = 52.48$

N 2 0 9 1 0 A 0 4 2 0

Leave blank

Question 2 continued

$Q_2 = 50.5 + \frac{44}{53} \times 10 = 58.80$

$Q_3 = 60.5 + \frac{24.5}{37} \times 10 = 67.12$

d) $\Sigma fx = 8379.5$ $\Sigma fx^2 = 557489.75$

Estimate $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{8379.5}{134} = 62.53$

Estimate s.d. $= \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \bar{x}^2}$

$= \sqrt{\frac{557489.75}{134} - 62.53358^2}$

$= 15.81$

e) $\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$. $= \frac{67.12 - 2 \times 58.80 + 52.48}{67.12 - 52.48}$

$= 0.1366$

Distribution has slight positive skew

f) mean 62.53 > median 58.80

indicates positive skew

Leave blank

3. A long distance lorry driver recorded the distance travelled, m miles, and the amount of fuel used, f litres, each day. Summarised below are data from the driver's records for a random sample of 8 days.

The data are coded such that $x = m - 250$ and $y = f - 100$.

$$\Sigma x = 130 \qquad \Sigma y = 48 \qquad \Sigma xy = 8880 \qquad S_{xx} = 20\,487.5$$

(a) Find the equation of the regression line of y on x in the form $y = a + bx$. **(6)**

(b) Hence find the equation of the regression line of f on m. **(3)**

(c) Predict the amount of fuel used on a journey of 235 miles. **(1)**

No longer on syllabus

However, still worth doing using these formulae from the old syllabus formulae book

Correlation and regression

For a set of n pairs of values (x_i, y_i)

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$$

The regression coefficient of y on x is $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$

Least squares regression line of y on x is $y = a + bx$ where $a = \bar{y} - b\bar{x}$

a) $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{S_{xx}}$

N 2 0 9 1 0 A 0 8 2 0

$$b = \frac{8880 - \frac{130 \times 48}{8}}{20{,}487.5} = 0.395363$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{130}{8} = 16.25$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{48}{8} = 6$$

$$a = \bar{y} - b\bar{x} = 6 - 0.3954 \times 16.25 = -0.4253$$

Regression Line $y = -0.4253 + 0.3954x$

$$y = -0.425 + 0.395x$$

b) The data are coded such that $x = m - 250$ and $y = f - 100$.

$$f - 100 = -0.425 + 0.395(m - 250)$$

$$f = -0.425 + 0.395m - 0.395363 \times 250 + 100$$

$$f = 0.734 + 0.395m$$

c) $m = 235$

$$f = 0.734 + 0.395 \times 235$$

$$f = 93.55 \text{ litres}$$

$$f = 93.6 \text{ litres}$$

Leave blank

4. Aeroplanes fly from City A to City B. Over a long period of time the number of minutes delay in take-off from City A was recorded. The minimum delay was 5 minutes and the maximum delay was 63 minutes. A quarter of all delays were at most 12 minutes, half were at most 17 minutes and 75% were at most 28 minutes. Only one of the delays was longer than 45 minutes.

An outlier is an observation that falls either 1.5× (interquartile range) above the upper quartile or 1.5× (interquartile range) below the lower quartile.

(a) On the graph paper opposite draw a box plot to represent these data. **(7)**

(b) Comment on the distribution of delays. Justify your answer. **(2)**

(c) Suggest how the distribution might be interpreted by a passenger who frequently flies from City A to City B. **(1)**

a) IQR $= 28 - 12 = 16$ min

Top end outliers above $28 + \frac{3}{2} \times 16 = 52$

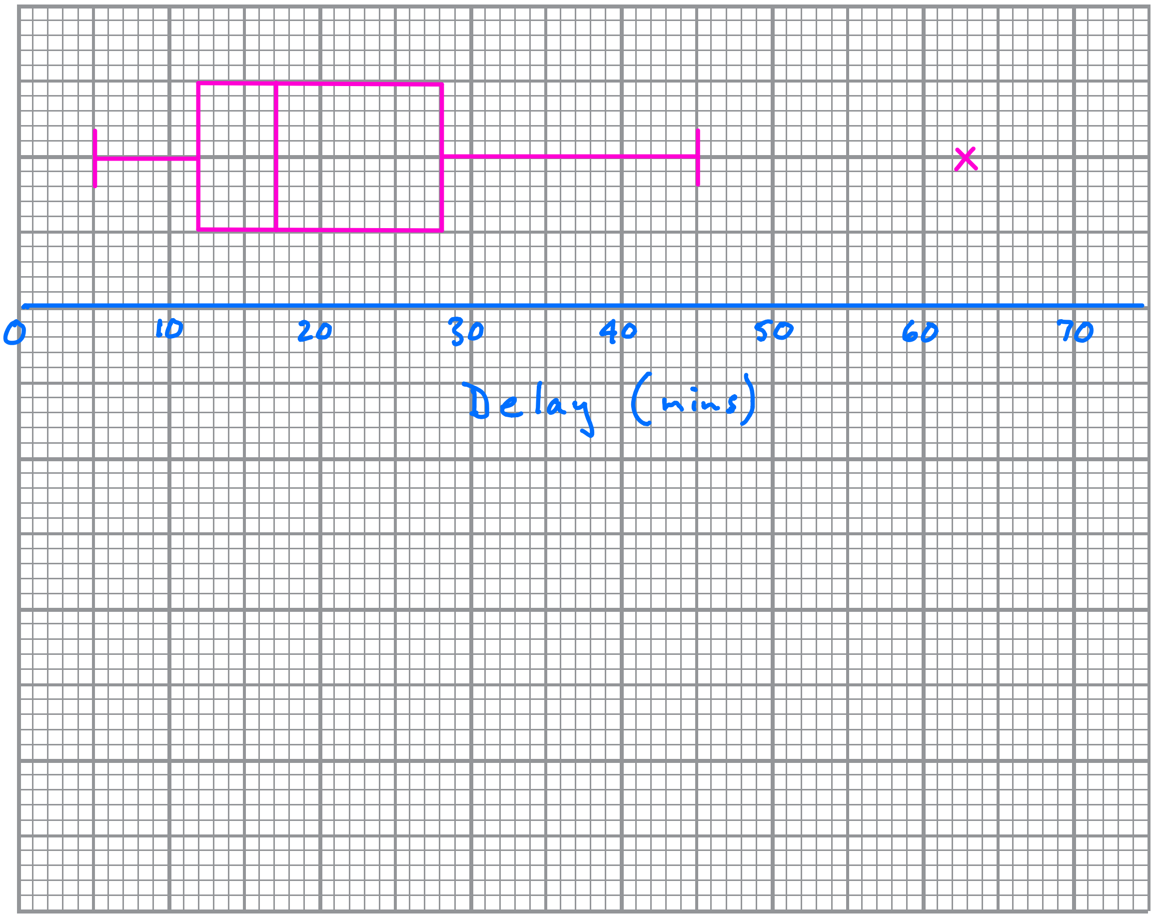
Bottom end outliers below $12 - \frac{3}{2} \times 16 = -12$

So 1 outlier at top end
0 outliers at bottom end

N20910A01020

Leave blank

Question 4 continued



b) Distribution shows positive skew

$Q_3 - Q_2 > Q_2 - Q_1$

$11 > 5$

c) Most delays likely to be a reasonably small proportion of total journey time.

Q4

(Total 10 marks)

Leave blank

5. The random variable X has probability function

$$P(X = x) = \begin{cases} kx, & x = 1, 2, 3, \\ k(x+1), & x = 4, 5, \end{cases}$$

where k is a constant. No longer on syllabus

(a) Find the value of k.
(2)

(b) Find the exact value of $E(X)$.
(2)

(c) Show that, to 3 significant figures, $\text{Var}(X) = 1.47$.
(4)

(d) Find, to 1 decimal place, $\text{Var}(4 - 3X)$.
(2)

Leave blank

Question 5 continued

Leave blank

6. A scientist found that the time taken, M minutes, to carry out an experiment can be modelled by a normal random variable with mean 155 minutes and standard deviation 3.5 minutes.

Find

(a) $P(M > 160)$. **(3)**

(b) $P(150 \leqslant M \leqslant 157)$. **(4)**

(c) the value of m, to 1 decimal place, such that $P(M \leqslant m) = 0.30$. **(4)**

a) By calc $P(M > 160) = 0.0766$

b) By calc $P(150 \leq M \leq 157) = 0.6396$

c) By calc Area beneath $m = 0.30$

$m = 153.16$

$m \approx 153.2$ minutes to 1 d.p.

Leave blank

7. In a school there are 148 students in Years 12 and 13 studying Science, Humanities or Arts subjects. Of these students, 89 wear glasses and the others do not. There are 30 Science students of whom 18 wear glasses. The corresponding figures for the Humanities students are 68 and 44 respectively.

A student is chosen at random.

Find the probability that this student

(a) is studying Arts subjects, **(4)**

(b) does not wear glasses, given that the student is studying Arts subjects. **(2)**

Amongst the Science students, 80% are right-handed. Corresponding percentages for Humanities and Arts students are 75% and 70% respectively.

A student is again chosen at random.

(c) Find the probability that this student is right-handed. **(3)**

(d) Given that this student is right-handed, find the probability that the student is studying Science subjects. **(3)**

	Sci	Hum	Arts	Total
Glasses	18	44	27	89
No Glasses	12	24	23	59
Total	30	68	50	148

a) $P(\text{Arts student}) = \frac{50}{148} = \frac{25}{74}$

b) $P(\text{No Glasses} \mid \text{Arts}) = \frac{23}{50}$

c) $P(\text{Right-handed})$

$= \frac{30}{148} \times 0.8 + \frac{68}{148} \times 0.75 + \frac{50 \times 0.7}{148} = \frac{55}{74}$

N 2 0 9 1 0 A 0 1 8 2 0

Leave blank

Question 7 continued

d) $P(\text{Sci} \mid \text{Right-handed}) = \dfrac{P(\text{Sci} \cap \text{Right-handed})}{P(\text{Right-handed})}$

$$= \frac{\frac{30}{148} \times 0.8}{\frac{55}{74}} = \frac{12}{55}$$

Q7

(Total 12 marks)

TOTAL FOR PAPER: 75 MARKS

END

N20910A02020