1. A geyser is a hot spring which erupts from time to time. For two geysers, the duration of each eruption, x minutes, and the waiting time until the next eruption, y minutes, are recorded.

(a) For a random sample of 50 eruptions of the first geyser, the correlation coefficient between x and y is 0.758. The critical value for a 2-tailed hypothesis test for correlation at the 5% level is 0.279. Explain whether or not there is evidence of correlation in the population of eruptions. [2]

The scatter diagram in Fig. 9 shows the data from a random sample of 50 eruptions of the second geyser.

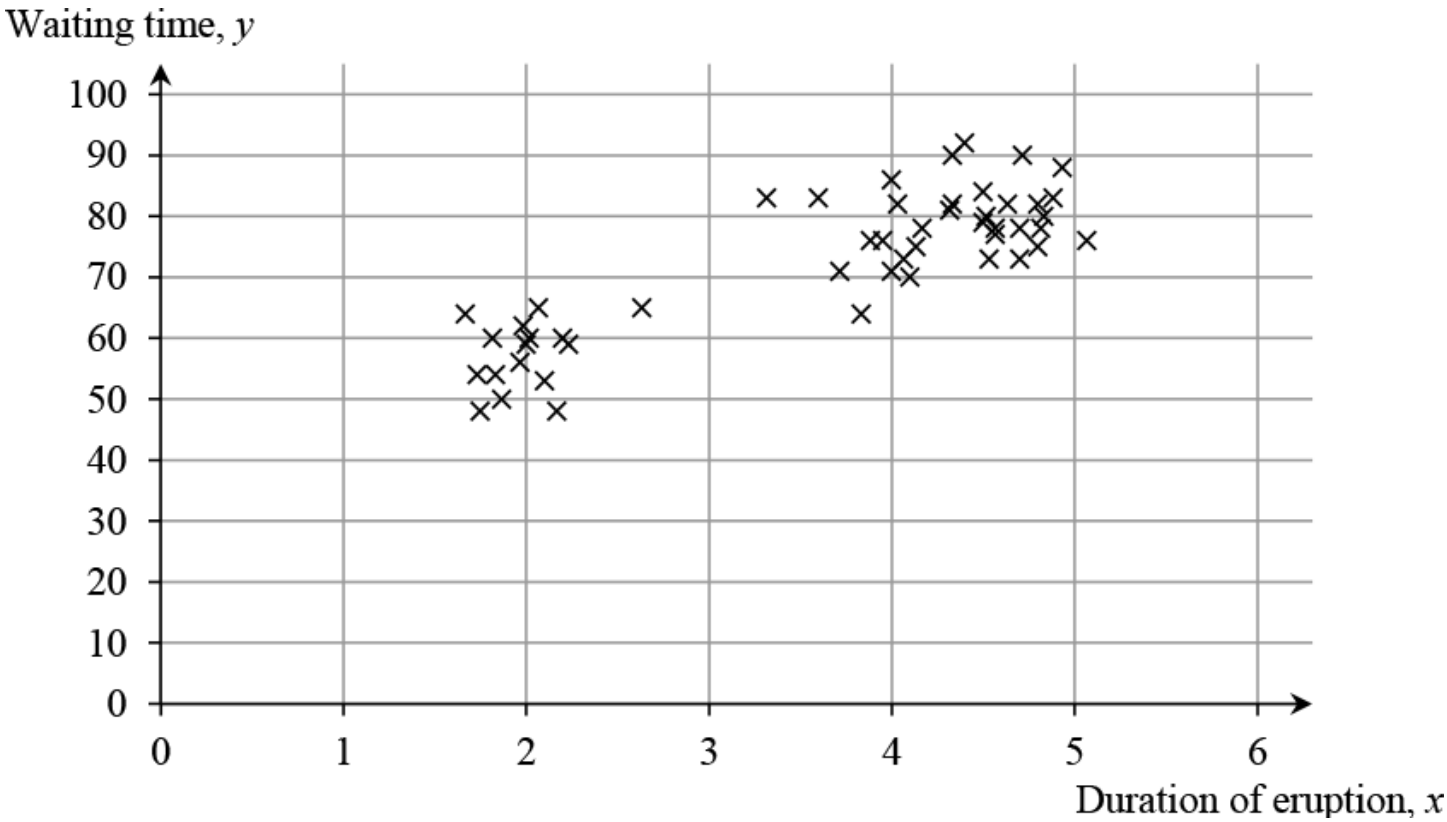

Fig.9

(b) Stella claims the scatter diagram shows evidence of correlation between duration of eruption and waiting time. Make two comments about Stella's claim. [2]

2. Nigel investigated migration in and out of different areas in the United Kingdom. He considered all the data available for long term international inflow and outflow rates per thousand resident population in 2012. Fig. 9.1 shows a scatter diagram for Nigel's sample after the data had been cleaned.

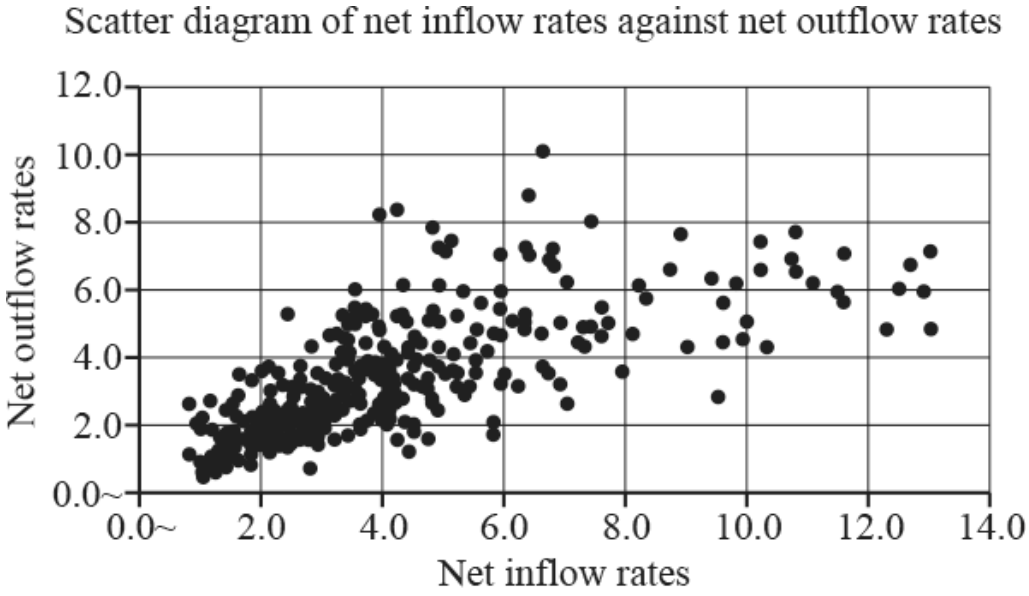

Fig. 9.1

Nigel used software to generate summary statistics. Fig. 9.2 shows summary statistics for inflow rates.

Statistics	
n	391
mean	6.8
s	8.579
$\sum x$	2644.9
$\sum x^2$	46598.17
min	0.8
Q_1	2.5
median	4.0
Q_3	6.8
max	91.8

Fig. 9.2

Nigel decided to clean the data by discarding all outliers, and so the pairs of values with long term inflow rates greater than 13.25 were discarded.

(a) Use information from Fig. 9.2 to show how Nigel calculated the value 13.25. **[1]**

Nigel then adopted a similar procedure by considering summary statistics for the long term net outflow rates. As a result he obtained a sample of size 334 drawn from a population of 391.

(b) State **one** strength and **two** weaknesses of Nigel's method for selecting his sample. **[3]**

Nigel found that the product moment correlation coefficient for these data is 0.6916. He used an online statistics calculator to find the associated p-value. A screenshot of the result is shown in Fig. 9.3.

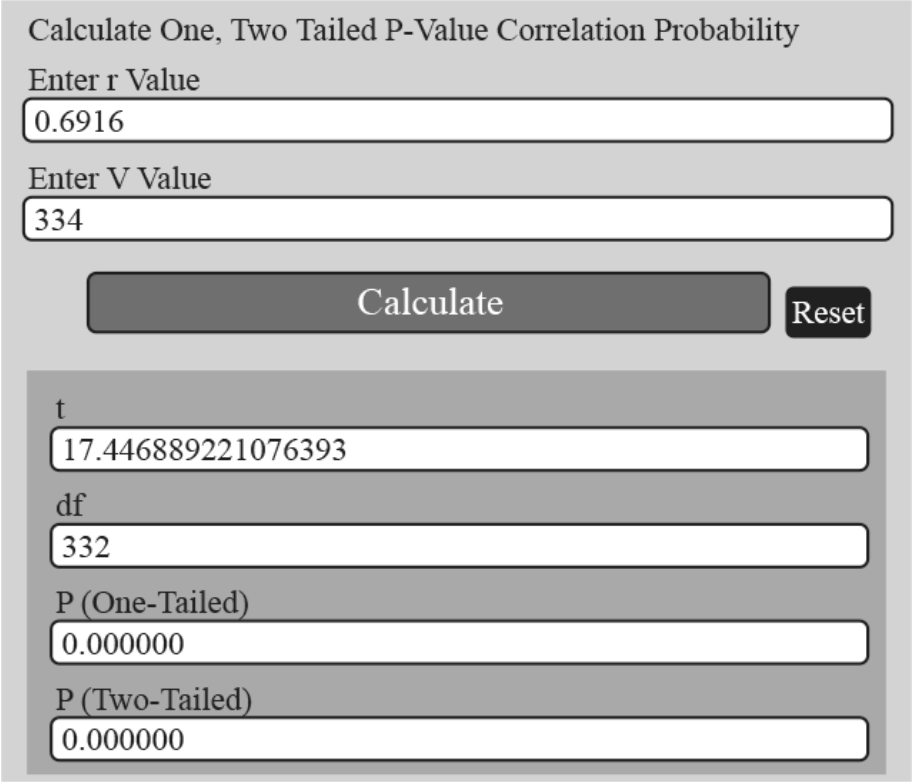

Fig. 9.3

Nigel made the following statement.

"My analysis proves that there is positive correlation between long term net inflow and long term net outflow rates per thousand resident population. Therefore high long term inflow rates cause high long term outflow rates."

(c) Give **three** reasons why Nigel's statement is flawed. **[3]**

3. The pre-release material includes data on unemployment rates in different countries. A sample from this material has been taken. All the countries in the sample are in Europe. The data have been grouped and are shown in Fig 14.1.

Unemployment rate	0–	5–	10–	15–	20–	35–50
Frequency	15	21	5	5	2	2

Fig. 14.1

A cumulative frequency curve has been generated for the sample data using a spreadsheet. This is shown in Fig. 14.2.

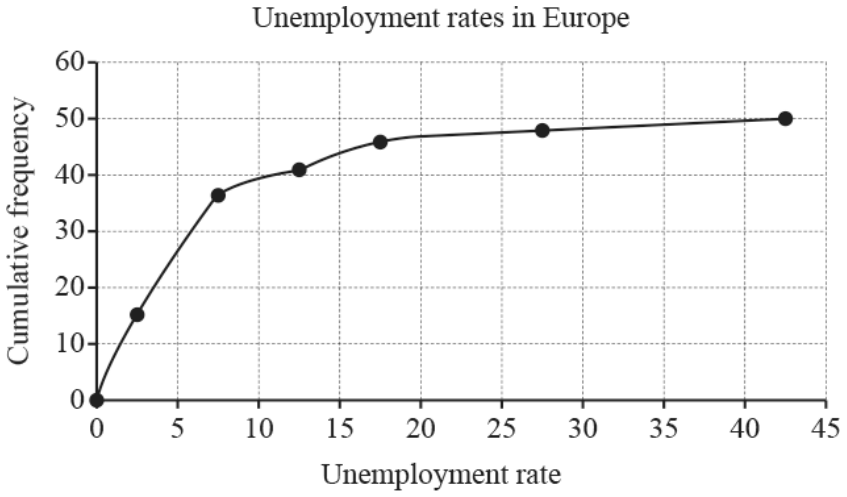

Fig. 14.2

Hodge used Fig. 14.2 to estimate the median unemployment rate in Europe. He obtained the answer 5.0. The correct value for this sample is 6.9.

(a) (i) There is a systematic error in the diagram.
- Identify this error.
- State how this error affects Hodge's estimate. [2]

(ii) There is another factor which has affected Hodge's estimate.
- Identify this factor.
- State how this factor affects Hodge's estimate. [2]

(b) Use your knowledge of the pre-release material to give another reason why any estimation of the median unemployment rate in Europe may be unreliable. [1]

(c) Use your knowledge of the pre-release material to explain why it is very unlikely that the sample has been randomly selected from the pre-release material. [1]

The scatter diagram shown in Fig. 14.3 shows the unemployment rate and life expectancy at birth for the 47 countries in the sample for which this information is available.

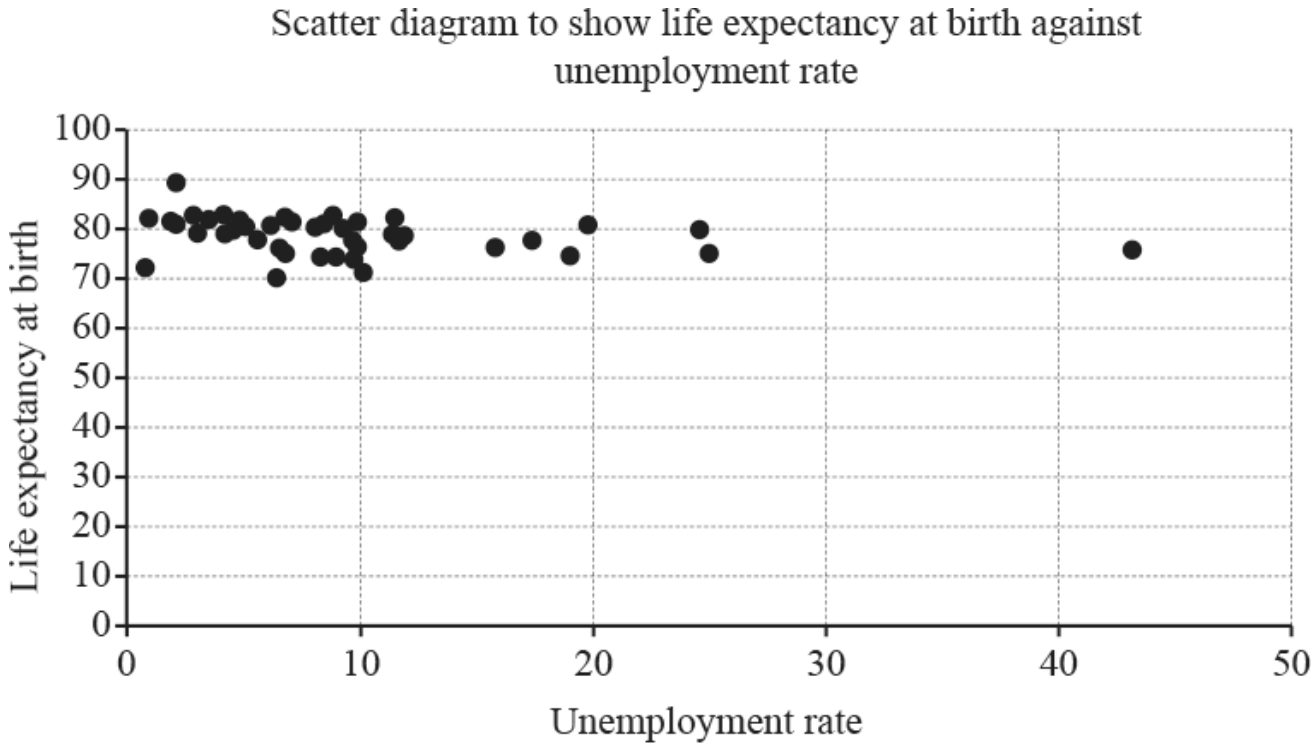

Fig. 14.3

The product moment correlation coefficient for the 47 items in the sample is – 0.2607.
The p-value associated with $r = -0.2607$ and $n = 47$ is 0.0383.

(d) Does this information suggest that there is an association between unemployment rate and life expectancy at birth in countries in Europe? [2]

Hodge uses the spreadsheet tools to obtain the equation of a line of best fit for this data.

(e) The unemployment rate in Kosovo is 35.3, but there is no data available on life expectancy. Is it reasonable to use Hodge's line of best fit to estimate life expectancy at birth in Kosovo? [1]

END OF QUESTION paper

Mark scheme

Question			Answer/Indicative content	Marks	Guidance
1		a	0.758 > 0.279 So there is sufficient evidence of correlation (in the population)	M1(AO1.1) A1(AO2.2b) [2]	Oe but not evidence of positive correlation.
		b	E.g. diagram shows positive correlation overall, but the data consists of two distinct clusters. E.g. neither of the two clusters show evidence of correlation	B1(AO2.3) B1(AO2.2b) [2]	Accept other suitable correct comments
			Total	4	
2		a	= 6.8 + 1.5 × (6.8 – 2.5) [= 13.25]	B1(AO3.1b) [1]	
		b	Strength – large sample size Weakness – sample is not random – removal of outliers may not be justifiable	B1(AO2.2a) B1(AO2.3) B1(AO2.4) [3]	
		c	eg pmcc may not be appropriate as scatter does not appear to be linear eg sample not random so calculating a correlation coefficient may not be valid eg correlation and causation are not the same thing eg the results of a hypothesis test are only suggestive, not conclusive (although highly suggestive in this case)	E1(AO2.4) E1(AO2.4) E1(AO2.4) [3]	Allow 1 mark for each distinct (sensible) reason up to a maximum of 3
			Total	1	

3		a	(i) the cumulative frequencies have been plotted against the mid-points of the class intervals, mis-plotting [at centre of each class] reduces estimate (by 2.5) oe	B1 (AO 2.4) B1 (AO 2.4) [2]	 Examiner's Comments Candidates who did well in this question recognised that the cumulative frequencies had been plotted at the mid-point of the intervals instead of at the upper limit. or eg Hodge used the graph (instead of the raw data)
			(ii) grouped data has been used grouping has slightly reduced the error introduced by misplotting (because the error is less than 2.5)	B1 (AO 2.4) B1 (AO 2.4) [2]	Examiner's Comments Candidates understood that grouping the data affects the accuracy of the result and commented accordingly. Candidates who did less well made comments about whether the points had been joined by straight lines or a curve.
		b	percentage unemployment is often estimated oe	E1 (AO 2.4) [1]	allow data (on percentage unemployment) is not available for all countries in Europe oe Examiner's Comments Candidates who did less well based comments on general geographical or economic ideas, rather than specifically related to issues related to the estimation of median values.
		c	there are many other countries in the pre-release material; it is very unlikely that a random sample would only include European countries.	E1 (AO 2.4) [1]	 Examiner's Comments Candidates who did well on part (b) and part (c) were familiar with the pre-release material made appropriate comments.

		d	negative correlation / association (may be embedded) comparison of *p*-value with 0.05 or 0.01 or other appropriate significance level and supporting comment	B1 (AO 2.2b) B1 (AO 2.2b) [2]	if B0B0 allow SC2 for eg comment on no significant association justified by comparison of *p*-value with appropriate significance level (eg 0.025) Examiner's Comments Candidates who did well commented on the nature of the correlation. They compared the p-value with a significance level and then made an appropriate deduction. Candidates who did less well compared the correlation coefficient with the p-value or did not comment on the association at all.
		e	(even though this is interpolation), the scatter / weak correlation / presence of an outlier would suggest that the use of of a line of best fit is inappropriate	E1 (AO 2.2b) [2]	allow explanation based on the value for Kosovo being an outlier or on it lying in the (large) gap in the scatter Examiner's Comments Candidates who did well commented on the nature of the scatter or the position of 35.3 relative to the given values to justify their comment.
			Total	9	