BOOK 1: TABLE OF CONTENTS

BOOK 2: TABLE OF CONTENTS

(continued overleaf)

CORE: DATA ANALYSIS

Extended response test 2

Writing time: 20 minutes

Structure of Test

Number of questions	*Number of questions to be answered*	*Number of marks*
4	4	15

Question 1

The ordered stemplot below shows the annual membership charges, in dollars, of 19 tennis clubs.

Stem	Leaf
7	8
8	3 9
9	1 2 4 7 7 9
10	1 3 5 6 7 8
11	0 1 3 5

Key: 7|8 = 78

a. Calculate the percentage of the tennis clubs that charge over \$100 for membership, correct to one decimal place.

1 mark

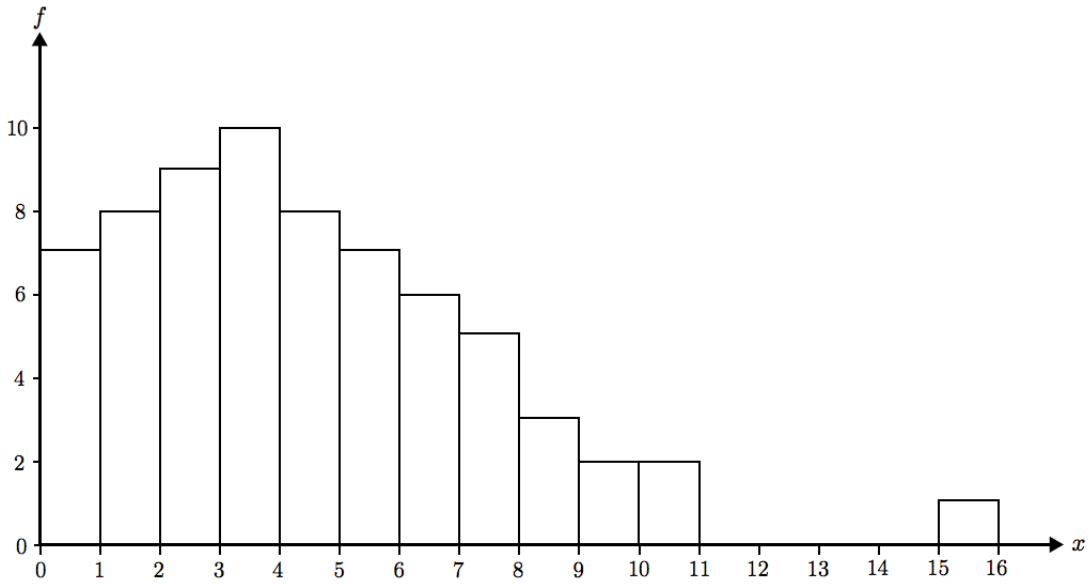
b. Find the median and mean values for the given data correct to the nearest dollar.

2 marks

c. Explain which of the values calculated in **part b.** is a more appropriate measure of centre for this data.

1 mark

Question 2



a. Name the type of data that is displayed in the above histogram.

1 mark

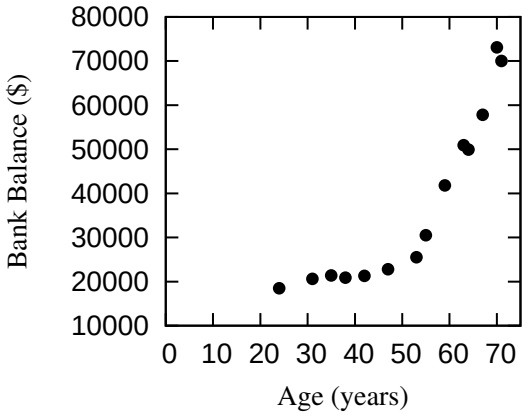
b. Describe the histogram in terms of shape, outliers, centre, and spread.

2 marks

Question 3

The ages and bank balances of 14 people were recorded and the data summarised as shown below.

Bank Balance ($)	Age (years)
18,500	24
20,900	38
30,500	55
50,900	63
21,300	42
41,800	59
20,600	31
22,800	47
49,900	64
70,000	71
21,400	35
25,500	53
57,800	67
73,100	70



a. Which is the dependent variable?

1 mark

The relationship between *age* and *bank balance* is non-linear.

b. Apply a transformation to the data that allows *bank balance* to be predicted from Age^2. Write the least squares regression equation in terms of these variables, with coefficients correct to one decimal place.

2 marks

c. Use the equation found in **part b.** to predict the bank balance of Ashley, who is 45 years old, to the nearest dollar.

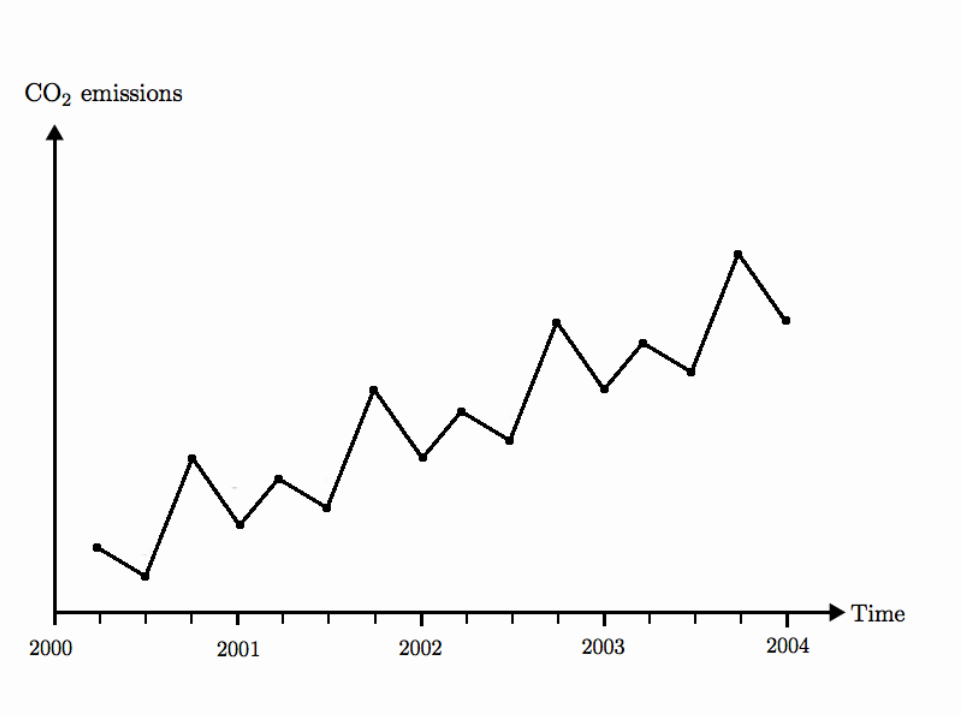
1 mark

d. Ashley actually had a bank balance of \$27,300. Determine the residual value for Ashley's bank balance to the nearest dollar.

1 mark

Question 4

The time series below displays recorded data of carbon dioxide emissions over time.



a. Describe the time series plot in terms of trend and variation.

1 mark

b. Use 3-median smoothing to smooth the time series. Mark the smoothed data points with a cross (×).

2 marks

END OF TEST

CORE: DATA ANALYSIS

Extended response test 2

Model solutions and marking scheme

Question 1a.

$$\frac{10}{19} \times 100\% \quad \approx \quad 52.6\%$$

- 1 answer mark for correct percentage.

Question 1b.

$$\begin{aligned} \text{mean} &= 100 \\ \text{median} &= 101 \end{aligned}$$

- 1 answer mark for correct median value.
- 1 answer mark for correct mean value.

Question 1c.

The median is a more appropriate measure of centre for the data as the distribution is skewed.

- 1 answer mark for correctly choosing the median, with reference to the skewed distribution.

Question 2a.

Numerical, continuous

- 1 answer mark for identifying that the variable is numerical and continuous.

Question 2b.

The data is positively skewed, with one apparent outlier between 15 and 16, a median of ~ 4, and an IQR of ~ 4 and/or a range of ~ 16.

- 1 answer mark for identifying skew and outlier.
- 1 answer mark for a reasonable estimate for the median and one measurement of spread (at least one of the range or IQR).

Question 3a.

Bank balance

- 1 answer mark for identifying bank balance as the dependent variable.

Question 3b.

$$\begin{aligned} Bank\ balance &= 2929.2 + 12.1 \times Age^2 \end{aligned}$$

- 1 answer mark for the correct equation.

Question 3c.

$$\begin{aligned} Predicted\ bank\ balance &= 2929.2 + 12.1 \times 45^2 \\ &= \$27,432 \end{aligned}$$

- 1 answer mark for the correct predicted bank balance.

Question 3d.

$$\begin{aligned} resid. &= 27300 - 27432 \\ &= -\$132 \end{aligned}$$

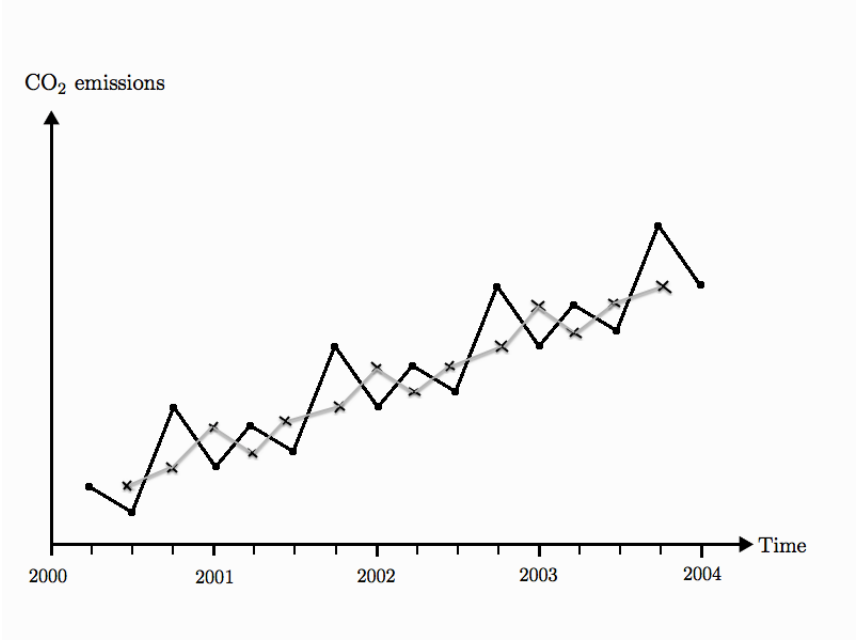
- 1 answer mark for the correct residual value, including the negative sign.

Question 4a.

The time series shows seasonal variation and an upward trend.

- 1 answer mark for correctly identifying the seasonal variation and the upward trend.

Question 4b.



- 1 answer mark for the correct allocation of points.
- 1 answer mark for connecting the points with straight lines.

CORE: DATA ANALYSIS

Extended response test 2

Detailed solutions

Question 1a.

In this question, there are membership charges for 19 tennis clubs, 10 of which charge membership over \$100 (the number of figures over 100 is shown in the stems for 10 and 11.

The percentage of clubs with membership charges over \$100 is therefore $\frac{10}{19} \times 100\% = 52.6\%$.

Question 1b.

The median is the $\frac{n+1}{2}$th data point. $\frac{19+1}{2} = 10$, so we want the 10th data point, which is \$101.

The mean is calculated as the sum of all the data values, divided by the number of all data values. The mean value is \$100, to the nearest dollar.

Note that both of these figures can be calculated using *Data and Statistics* and *one-variable statistics*.

Question 1c.

The mean and the median are both measures of centre. The difference between them is that the mean, because it is an average of the data figures, is affected by the skew or outliers, whilst the median is called a "resistant statistic" because it is the middle figure, and is not affected by skew or outliers.

The median is generally considered an appropriate measure of centre for all types of data spreads. The mean is generally considered an appropriate measure of centre when the relationship is roughly symmetrical - when there are no outliers and when there is no obvious skew.

In the data distribution of the stemplot, it appears there is a negative skew. The median is therefore appropriate for the data; however, the mean is not.

Question 2a.

The variable that is represented in the histogram is a numerical variable (as can be seen because the values on the horizontal axis are numerical). Because the columns are from 0 to 1 and 1 to 2 and so on, the variable is continuous, as the column represents any number between the two labels.

If the column were to be drawn with the number label in the middle of the column, this would represent a discrete variable, but this is not the case in this example.

Question 2b.

Describing histograms is usually done by reference to shape, outliers, centre, and spread. In this example, the histogram is positively skewed because the tail is to the right hand side, and there is an apparent outlier (an extreme figure that is inconsistent with the majority of the data) between 15 and 16.

The total number of data points in the histogram is 68, so the median is the $\frac{68+1}{2} = 34.5$th data point, which occurs between the bars for 3 to 4 and 4 to 5. The median is therefore likely to be close to 4. Q1 would therefore be the $\frac{34+1}{2} = 17.5$th point (between 2 and 3) and Q3 would be the $\frac{35+68}{2} = 51.5$th point (between 6 and 7), so the interquartile range could be at most $7-2=5$ or at least $6-3=3$, so any answer from 3 to 5 is reasonable.

The range is the highest data value minus the lowest, which in this example is $16-0=16$.

Shape relates to the skew, and if significant, modal value of the histogram. In this case, the histogram is positively skewed.

Outliers are extreme values that are far away from the remainder of the data. In this case, there is an apparent outlier in the 15-16 interval.

Centre can be measured either using median or mean. Referring to question **1c.**, the mean is inappropriate for the data given there is an outlier. The median is approximately 4.

Spread refers either to the range, the difference between the highest and lowest values of the data (in this case, around 14-16), or the interquartile range, the difference between Q3 and Q1 (the spread of the middle 50% of the data), in this case, around 3-5.

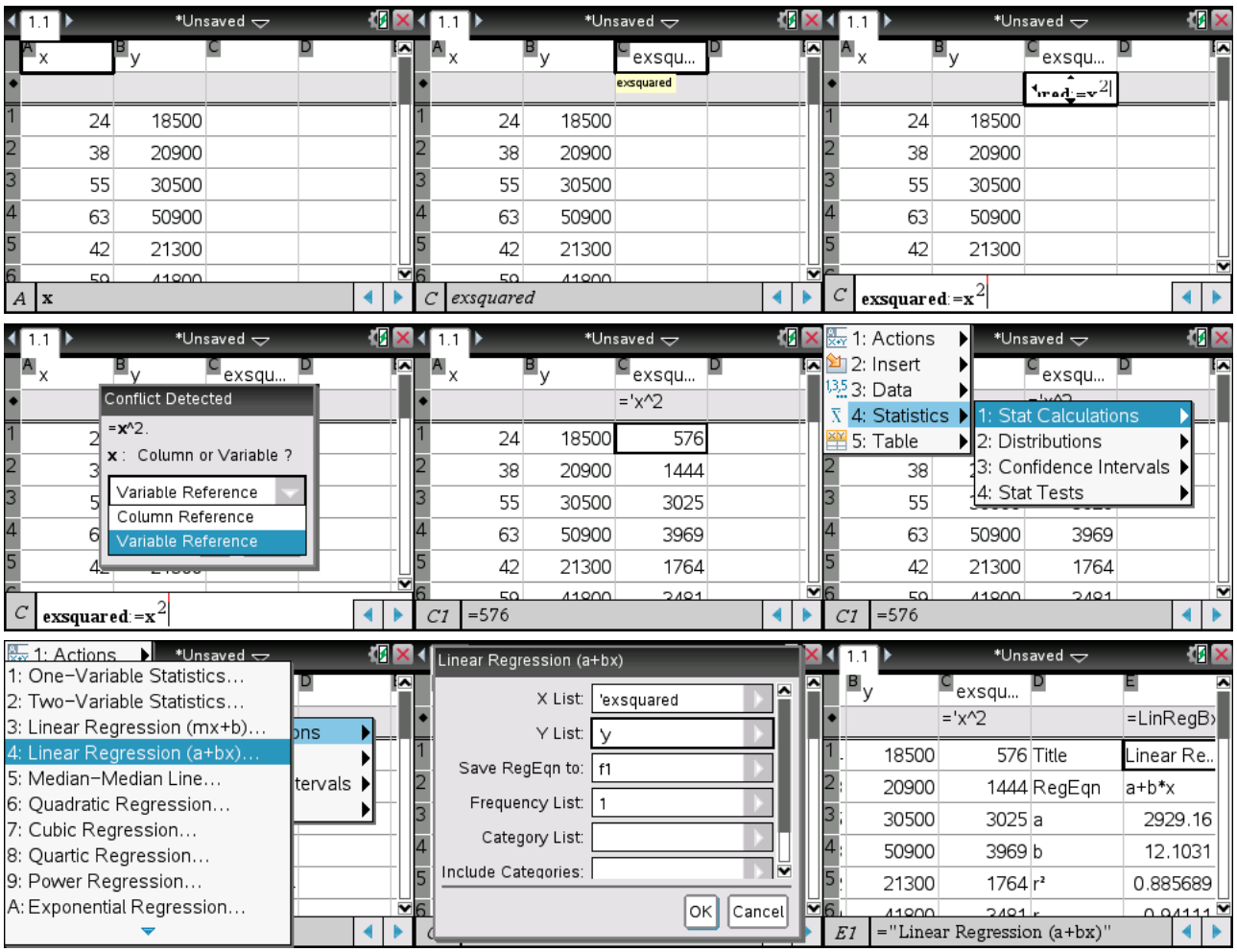
Question 3a.

Since age does not generally depend on other variables, it is the independent variable. We would consider it likely that bank balance depends on age. Bank balance is therefore the dependent variable.

Question 3b.

In order to perform the transformation, we must use the values of Age^2 instead of Age as the "x" variable. To do this, we need to find the x^2 data values, by introducing a new column, in this example called "exsquared" (note that the column title cannot begin with x because it is already in use to represent Age). Then, in the grey row (the 2nd row), enter the line: $=x^2$ so that the column exsquared shows the values of the "x" column, squared. Select variable reference and then press enter.

After this, we can do our regular linear regression, using the "exsquared" column as the x-value and the original y column as the y-value. This will give us y against x^2 as is required.



The answer, to one decimal place, expressed in terms of the variables, is therefore

$$Bank\ balance = 2929.2 + 12.1 \times Age^2.$$

Question 3c.

Substituting an age of 45 into the equation from part **b.**:

$$\begin{aligned} Predicted\ bank\ balance &= 2929.2 + 12.1 \times 45^2 \\ &= \$27,432 \end{aligned}$$

Question 3d.

Residual values are calculated as the actual value minus the predicted value.

$$\begin{aligned} resid. &= 27300 + 27432 \\ &= -132 \end{aligned}$$

Note that the residual value is negative; you should make sure to include the sign in your answer, as it indicates that the actual value is less than the predicted value (rather than greater, which a positive answer would suggest).

Question 4a.

Trend refers to whether the plot seems, overall, to be going upwards, downwards, or have no trend (remaining even). In this time series plot, there is an upward trend, as the data values increase with time.

Variation refers to whether the changes (the increases and decreases in the data) occur within 1 year (seasonal variation), or over periods of over 1 year (cyclical variation), or whether there is no pattern in the increases and decreases (random variation). In this time series plot, there is seasonal variation.

The time series plot therefore shows an upward trend and seasonal variation.

Question 4b.

In order to do 3-median smoothing, the new data point should occur at whichever is the middle value of the point one before the point in question, the actual point in question, and the point after the point in question. For example, the 3-median point for the time value shown as 2002 must be shown vertically as the median height of the actual 2002 point, and the points before and after. In this case, the median vertical height is the point after, so the median point at 2002 should be drawn at the same height as the actual point immediately after the 2002 point.

The smoothed plot should appear as follows: