Practice 1

1. Measures of central tendency

Summary of key points

1 A variable that can take any value in a given range is a **continuous** variable.

2 A variable that can take only specific values in a given range is a **discrete** variable.

3 A **grouped frequency distribution** consists of **classes** and their related **class frequencies**

Classes	30–31	**32–33**	34–35

For the class 32–33

Lower class boundary	is	31.5
Upper class boundary	is	33.5
Class width	is	$33.5 - 31.5 = 2$
Class mid-point	is	$\frac{1}{2}(31.5 + 33.5) = 32.5$

4 The **mode** or **modal class** is the value or class that occurs most often.

5 The **median** is the middle value when the data is put in order.

6 For **discrete** data divide n by 2. When $\frac{n}{2}$ is a whole number find the mid-point of the corresponding term and the term above. When $\frac{n}{2}$ is not a whole number, round the number up and pick the corresponding term.

7 For **continuous** grouped data divide n by 2 and use interpolation to find the value of the corresponding term.

8 The **mean** is given by

$\bar{x} = \frac{\sum x}{n}$ or $\frac{\sum fx}{\sum f}$ (for grouped data x is the mid-point of the class).

9 If set A, of size n_1, has mean $\bar{x}_1$ and set B, of size n_2, has a mean $\bar{x}_2$ then the **mean of the combined** set of A and B is $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

10 When the data values are large, you can use coding to make the numbers easier to work with. To find the mean of the original data, find the mean of the coded data, equate this to the coding used and solve.

1 The following figures give the number of children injured on English roads each month for a certain period of seven months.

55 72 50 66 50 47 38

a Write down the modal number of injuries. **b** Find the median number of injuries.

c Calculate the mean number of injuries.

2 The mean Science mark for one group of eight students is 65. The mean mark for a second group of 12 students is 72. Calculate the mean mark for the combined group of 20 students.

3 A computer operator transfers an hourly wage list from a paper copy to her computer. The data transferred is given below:

£5.50 £6.10 £7.80 £6.10 £9.20 £91.00 £11.30

a Find the mean, mode and median of these data.

The office manager looks at the figures and decides that something must be wrong.

b Write down, with a reason, the mistake that has probably been made.

4 A piece of data was collected from each of nine people. The data was found to have a mean value of 35.5. One of the nine people had given a value of 42 instead of a value of 32.

a Write down the effect this will have had on the mean value.

The correct data value of 32 is substituted for the incorrect value of 42.

b Calculate the new mean value.

5 On a particular day in the year 2007, the prices (x) of six shares were as follows:

807 967 727 167 207 767

Use the coding $y = \frac{x - 7}{80}$ to work out the mean value of the shares.

6 The coded mean of employee's annual earnings (£x) for a store is 18. The code used was $y = \frac{x - 720}{1000}$. Work out the uncoded mean earnings.

7 Different teachers using different methods taught two groups of students. Both groups of students sat the same examination at the end of the course. The students' marks are shown in the grouped frequency table.

a Work out an estimate of the mean mark for Group *A* and an estimate of the mean mark for Group *B*.

b Write down whether or not the answer to **a** suggests that one method of teaching is better than the other. Give a reason for your answer.

Exam Mark	Frequency Group *A*	Frequency Group *B*
20–29	1	1
30–39	3	2
40–49	6	4
50–59	6	13
60–69	11	15
70–79	10	6
80–89	8	3

8 The table summarises the distances travelled by 150 students to college each day.

a Use interpolation to calculate the median distance for these data.

The mid-point is x and the corresponding frequency is f. Calculations give the following values

$\sum fx = 1056$

b Calculate an estimate of the mean distance for these data.

Distance (nearest km)	Number of students
0–2	14
3–5	24
6–8	70
9–11	32
12–14	8
15–17	2

9 Chloe enjoys playing computer games.
Her parents think that she spends too much time playing games. Chloe decides to keep a record of the amount of time she plays computer games each day for 50 days. She draws up a grouped frequency table to show these data.

The mid-point of each class is represented by x and its corresponding frequency is f.
(You may assume $\sum fx = 1275$)

a Calculate an estimate of the mean time Chloe spends on playing computer games each day.

Time to the nearest minute	Frequency
10–14	1
15–19	5
20–24	18
25–29	15
30–34	7
35–39	3
40–44	1

Chloe's parents thought that the mean was too high and suggested that she try to reduce the time spent on computer games. Chloe monitored the amount of time she spent on games for another 40 days and found that at the end of 90 days her overall mean time was 26 minutes.

b Find an estimate for the mean amount of time Chloe spent on games for the 40 days.

c Comment on the two mean values.

10 The lifetimes of 80 batteries, to the nearest hour, is shown in the table opposite.

a Write down the modal class for the lifetime of the batteries.

b Use interpolation to find the median lifetime of the batteries.

The mid-point of each class is represented by x and its corresponding frequency by f, giving $\sum fx = 1645$.

c Calculate an estimate of the mean lifetime of the batteries.

Lifetime (hours)	Number of batteries
6–10	2
11–15	10
16–20	18
21–25	45
26–30	5

Another batch of 12 batteries is found to have an estimated mean lifetime of 22.3 hours.

d Find the mean lifetime for all 92 batteries.

Summary of key points

1 The **range** of a set of data is the difference between the highest and lowest value in the set.

2 The **quartiles**, Q_1, Q_2, Q_3, Q_4, split the data into four parts. To calculate the **lower quartile, Q_1,** divide n by 4.

3 For discrete data for the lower quartile, Q_1, divide n by 4. To calculate the upper quartile, Q_3, divide n by 4 and multiply by 3. When the result is a whole number find the mid-point of the corresponding term and the term above. When the result is not a whole number round the number up and pick the corresponding term.

4 For continuous grouped data for Q_1, divide n by 4 and for Q_3 divide n by 4 and multiply by 3. Use interpolation to find the value of the corresponding term.

5 The **interquartile range** is $Q_3 - Q_1$.

6 **Percentiles** split the data into 100 parts.

- **To calculate the** xth percentile, P_x, find the value of the $\frac{xn}{100}$th term.
- The n% to m% interpercentile range $= P_m - P_n$.

7 The **deviation** of an observation x from the mean is given by $x - \bar{x}$.

- **Variance** $= \frac{\sum(x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$
- **Standard deviation** $= \sqrt{\textbf{Variance}}$

8 To work out the variance and standard deviation for a frequency table and a grouped frequency distribution where x is the mid-point of the class, let f stand for the frequency, then $n = \sum f$

- Variance $= \frac{\sum f(x - \bar{x})^2}{\sum f}$ or $\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$

9 When the data values are large you can use coding to make the numbers easier to work with. To find the standard deviation of the original data, find the standard deviation of the coded data and either multiply this by what you divided by, or divide this by what you multiplied by.

1 For the set of numbers:

2 5 3 8 9 10 3 2 15 10 5 7 6 7 5

a work out the median,

b work out the lower quartile,

c work out the upper quartile,

d work out the interquartile range.

2 A frequency distribution is shown below

Class interval	1–20	21–40	41–60	61–80	81–100
Frequency (f)	5	10	15	12	8

Work out the interquartile range.

3 A frequency distribution is shown below.

Class interval	1–10	11–20	21–30	31–40	41–50
Frequency (f)	10	20	30	24	16

a Work out the value of the 30th percentile.

b Work out the value of the 70th percentile.

c Calculate the 30% to 70% interpercentile range.

4 The heights(h) of 10 mountains were recorded to the nearest 10 m and are shown in the table below.

h	1010	1030	1050	1000	1020	1030	1030	1040	1020	1000

Use the coding $y = \frac{h - 10}{100}$ to find the standard deviation of the heights.

5 The times it took a random sample of runners to complete a race are summarised in the table.

Time taken (t minutes)	**Frequency**
20–29	5
30–39	10
40–49	36
50–59	20
60–69	9

a Work out the lower quartile.

b Work out the upper quartile.

c Work out the interquartile range.

The mid-point of each class was represented by x and its corresponding frequency by f giving:

$\sum fx = 3740 \qquad \sum fx^2 = 183\,040$

d Estimate the variance and standard deviation for these data.

6 20 birds were caught for ringing. Their wing spans (x) were measured to the nearest centimetre.

The following summary data was worked out:

$\sum x = 316$ $\qquad\qquad \sum x^2 = 5078$

a Work out the mean and the standard deviation of the wing spans of the 20 birds.

One more bird was caught. It had a wing span of 13 centimetres

b Without doing any further working say how you think this extra wing span will affect the mean wing span.

7 The heights of 50 clover flowers are summarised in the table.

Heights in mm (h)	**Frequency**
90–95	5
95–100	10
100–105	26
105–110	8
110–115	1

a Find Q_1.

b Find Q_2.

c Find the interquartile range.

d Use $\sum fx = 5075$ and $\sum fx^2 = 516\,112.5$ to find the standard deviation.

Summary of key points

1 A stem and leaf diagram is used to order and present data.

2 A stem and leaf diagram reveals the shape of the data and enables quartiles to be found.

3 Two sets of data can be compared using back-to-back stem and leaf diagrams.

4 An outlier is an extreme value. You will be told what rule to apply to identify outliers.

5 A box plot represents important features of the data. It shows quartiles, maximum and minimum values and any outliers.

6 Box plots can be used to compare two sets of data.

7 If data are continuous and summarised in a group frequency distribution, the data can be represented by a histogram.

8 Diagrams, measures of location and measures of spread can be used to describe the shape (skewness) of a data set.

9 You can describe whether a distribution is skewed using

- quartiles
- shape from box plots
- measures of location
- formula $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ (larger value means greater skew).

1 Jason and Perdita decided to go for a touring holiday on the continent for the whole of July. They recorded the number of kilometres they travelled each day. The data are summarised in the stem and leaf diagram below.

stem	leaf									Key: 15|5 means 155 kilometres
15	5									
16	4	8	9							
17	3	5	7	8	8	8	9	9	9	
18	4	4	5	5	8					
19	2	3	4	5	5	6				
20	4	7	8	9						
21	1	2								
22	6									

a Find Q_1, Q_2, and Q_3

Outliers are values that lie outside $Q_1 - 1.5(Q_3 - Q_1)$ and $Q_3 + 1.5(Q_3 - Q_1)$.

b Find any outliers.

c Draw a box plot of these data.

d Comment on the skewness of the distribution.

2 Sophie and Jack do a survey every day for three weeks. Sophie counts the number of pedal cycles using Market Street. Jack counts the number of pedal cycles using Strand Road. The data they collected are summarised in the back-to-back stem and leaf diagram.

Sophie	Stem	Jack
9 9 7 5	0	6 6
7 6 5 3 3 2 2 2 1 1	1	1 1 5
5 3 3 2 2	2	1 2 2 2 3 7 7 8 9
2 1	3	2 3 4 7 7 8
	4	2

Key: 5|0|6 means Sophie counts 5 cycles and Jack counts 6 cycles

a Write down the modal number of pedal cycles using Strand Road.

The quartiles for these data are summarised in the table below.

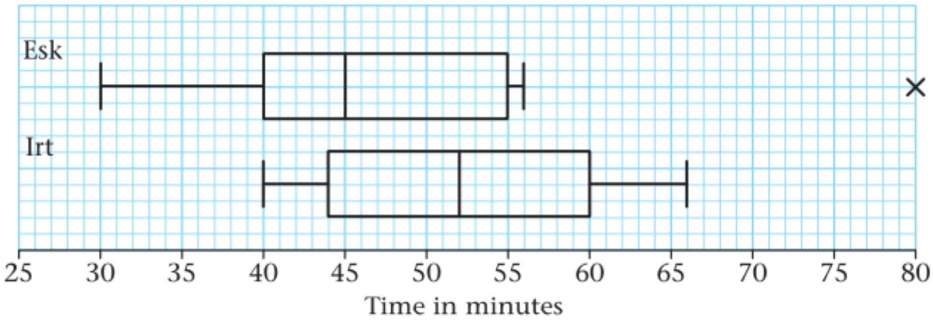
	Sophie	**Jack**
Lower quartile	X	21
Median	13	Y
Upper quartile	Z	33

b Find the values for X, Y and Z.

c Write down the road you think has the most pedal cycles travelling along it overall. Give a reason for your answer.

4 Fell runners from the Esk Club and the Irt Club were keen to see which club had the fastest runners overall. They decided that all the members from both clubs would take part in a fell run. The time each runner took to complete the run was recorded.

The results are summarised in the box plot.



a Write down the time by which 50% of the Esk Club runners had completed the run.

b Write down the time by which 75% of the Irt Club runners had completed the run.

c Explain what is meant by the cross (×) on the Esk Club box plot.

d Compare and contrast these two box plots.

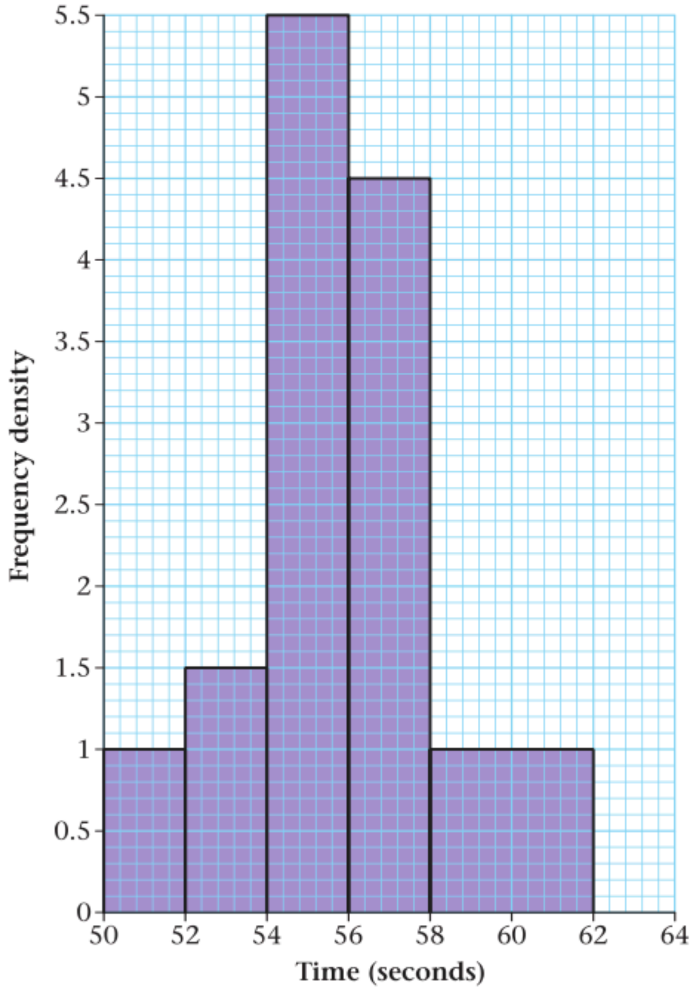
e Comment on the skewness of the two box plots.

f What conclusions can you draw from this information about which club has the fastest runners?

5 The histogram shows the time taken by a group of 58 girls to run a measured distance.

a Work out the number of girls who took longer than 56 seconds.

b Estimate the number of girls who took between 52 and 55 seconds.



6 The table gives the distances travelled to school, in km, of the population of children in a particular region of the United Kingdom.

Distance, km	0–1	1–2	2–3	3–5	5–10	10 and over
Number	2565	1784	1170	756	630	135

A histogram of this data was drawn with distance along the horizontal axis. A bar of horizontal width 1.5 cm and height 5.7 cm represented the 0–1 km group.
Find the widths and heights, in cm to one decimal place, of the bars representing the following groups:

a 2–3, **b** 5–10.

7 The labelling on bags of garden compost indicates that the bags weigh 20 kg.
The weights of a random sample of 50 bags are summarised in the table opposite.

Weight in kg	**Frequency**
14.6–14.8	1
14.8–18.0	0
18.0–18.5	5
18.5–20.0	6
20.0–20.2	22
20.2–20.4	15
20.4–21.0	1

a On graph paper, draw a histogram of these data.

b Estimate the mean and standard deviation of the weight of a bag of compost.

[You may use $\sum fy = 988.85$, $\sum fy^2 = 19\,602.84$]

c Using linear interpolation, estimate the median.

One coefficient of skewness is given by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}.$$

d Evaluate this coefficient for the above data.

e Comment on the skewness of the distribution of the weights of bags of compost.