PART B

Chapter 5

Descriptive statistics

Contents:

2 a For the categorical variables in question 1, write down two or three possible categories. (In all cases but one, there will be more than three categories possible.) Discuss your answers.

b For each of the quantitative variables (discrete and continuous) identified in question 1, discuss as a class the range of possible values you would expect.

INVESTIGATION 1 STATISTICS FROM THE INTERNET

In this investigation you will be exploring the web sites of a number of organisations to find out the topics and the types of data that they collect and analyse.

Note that the web addresses given here were operative at the time of writing but there is a chance that they will have changed in the meantime. If the address does not work, try using a search engine to find the site of the organisation.

What to do:

Visit the site of a world organisation such as the United Nations (www.un.org) or the World Health Organisation (www.who.int) and see the available types of data and statistics.

B PRESENTING AND INTERPRETING DATA

ORGANISING CATEGORICAL DATA

A **tally and frequency table** can be used to organise categorical data.

For example, a survey was conducted on the type of fuel used by 50 randomly selected vehicles.

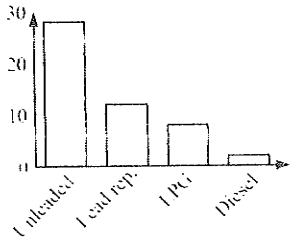
The variable 'type of fuel' is a categorical variable because the information collected for each vehicle can only be one of the four categories: Unleaded, Lead Replacement, LPG or Diesel. The data has been tallied and organised in the given frequency table:

Fuel type	*Tally*	*Freq.*
Unleaded	𝍸 𝍸 𝍸 𝍸 𝍸 |	28
Lead Rep	𝍸 𝍸 ||	12
LPG	𝍸 ||	8
Diesel	||	2
	Total	50

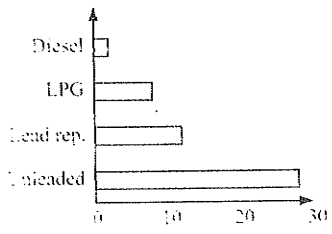
DISPLAYING CATEGORICAL DATA

Acceptable graphs to display the 'type of fuel' categorical data are:

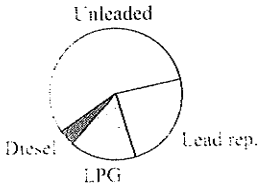
Vertical column graph



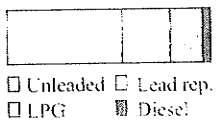
Horizontal bar chart



Pie chart



Segment bar chart



For categorical data, the **mode** is the category which occurs most frequently.

ORGANISING DISCRETE NUMERICAL DATA

Discrete numerical data can be organised:

- in a **tally and frequency table**
- using a **dot plot**
- using a **stem-and-leaf plot** (also called a **stemplot**).

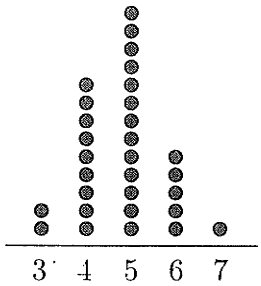
Stemplots are used when there are many possible data values. The stemplot is a form of grouping of the data which displays frequencies but retains the actual data values.

Examples:

- **frequency table**

Number	*Tally*	*Freq.*
3	||	2
4	𝍸 ||||	9
5	𝍸 𝍸 |||	13
6	𝍸	5
7	|	1

- **dot plot**



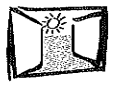
- **stemplot**

Example:

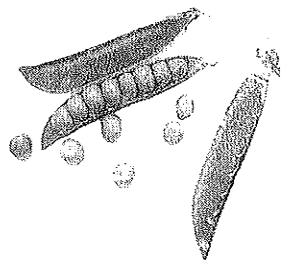
Stem	*Leaf*
0	9
1	7 1
2	8 3 6 7 6 4
3	9 3 5 5 6 8 2 1
4	7 9 3 4 2
5	1

As data is collected it can be entered directly into a carefully set up tally table, dot plot or stemplot blank sheet.

THE PEA PROBLEM



A farmer wishes to investigate the effect of a new organic fertiliser on his crops of peas. He is hoping to improve the crop yield by using the fertiliser. He set up a small garden which was subdivided into two equal plots and planted many peas. Both plots were treated the same except for the use of the fertiliser on one, but not the other. All other factors such as watering were as normal.



A random sample of 150 pods was harvested from each plot at the same time and the number of peas in each pod counted. The results were:

Without fertiliser

4 6 5 6 5 6 4 6 4 9 5 3 6 8 5 4 6 8 6 5 6 7 4 6 5 2 8 6 5 6 5 5 5 4 4 4 6 7 5 6
7 5 5 6 4 8 5 3 7 5 3 6 4 7 5 6 5 7 5 7 6 7 5 4 7 5 5 5 6 6 5 6 7 5 8 6 8 6 7 6
6 3 7 6 8 3 3 4 4 7 6 5 6 4 5 7 3 7 7 6 7 7 4 6 6 5 6 7 6 3 4 6 6 3 7 6 7 6 8 6
6 6 6 4 7 6 6 5 3 8 6 7 6 8 6 7 6 6 6 8 4 4 8 6 6 2 6 5 7 3

With fertiliser

6 7 7 4 9 5 5 5 8 9 8 9 7 7 5 8 7 6 6 7 9 7 7 7 8 9 3 7 4 8 5 10 8 6 7 6 7 5 6 8
7 9 4 4 9 6 8 5 8 7 7 4 7 8 10 6 10 7 7 7 9 7 7 8 6 8 6 8 7 4 8 6 8 7 3 8 7 6 9 7
6 9 7 6 8 3 9 5 7 6 8 7 9 7 8 4 8 7 7 7 6 6 8 6 3 8 5 8 7 6 7 4 9 6 6 6 8 4 7 8
9 7 7 4 7 5 7 4 7 6 4 6 7 7 6 7 8 7 6 6 7 8 6 7 10 5 13 4 7 7

For you to consider:

- Can you state clearly the problem that the farmer wants to solve?
- How has the farmer tried to make a fair comparison?
- How could the farmer make sure that his selection is at random?
- What is the best way of organising this data?
- What are suitable methods of display?
- Are there any abnormally high or low results and how should they be treated?
- How can we best indicate the most typical pod size?
- How can we best indicate the spread of possible pod sizes?
- What is the best way to show 'typical pod size' and the spread?
- Can a satisfactory conclusion be made?

ORGANISATION AND DISPLAY OF DISCRETE DATA

In **The Pea Problem**, the **discrete quantitative variable is**: *The number of peas in a pod.*

To organise the data a tally/frequency table could be used. We count the data systematically and use a '|' to indicate each data value.

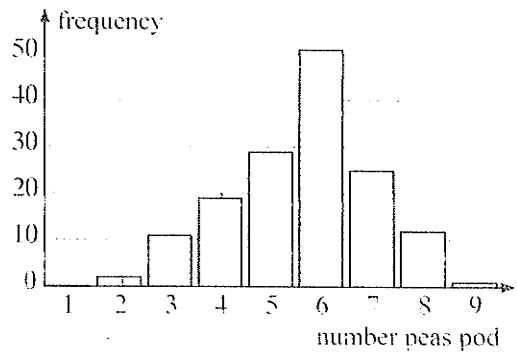
Remember that 𝍸 represents 5.

Below is the table for *Without fertiliser*:

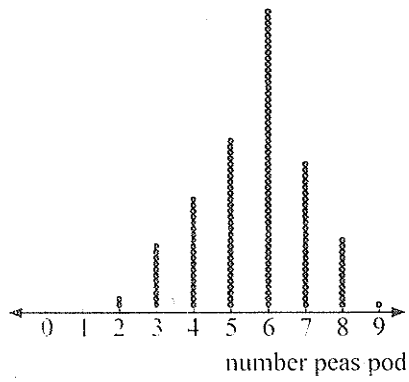
Number of peas/pod	*Tally*	*Frequency*
1		0
2	||	2
3	𝍸 𝍸 ||	11
4	𝍸 𝍸 𝍸 |||	19
5	𝍸 𝍸 𝍸 𝍸 𝍸 |||	29
6	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	51
7	𝍸 𝍸 𝍸 𝍸 𝍸	25
8	𝍸 𝍸 |||	12
9	|	1

A **dot plot** could be used to organise and display the results. or a **column graph** could be used to display the results.

Column graph of *Without fertiliser*



Dot plot of *Without fertiliser*



DISCUSSION

Are there any advantages/disadvantages in using a dot plot rather than a column graph?

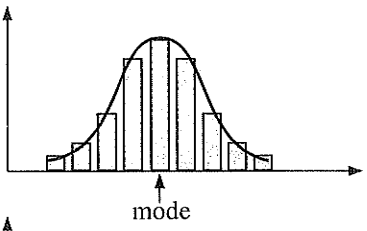
From both graphs we can make observations and calculations such as:

- 6 peas per pod is the mode of the *Without fertiliser* data.
- 8.7% of the pods had fewer than 4 peas in them, etc.

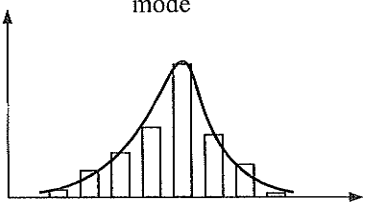
DESCRIBING THE DISTRIBUTION OF THE DATA SET

Many data sets show **symmetry** or **partial symmetry** about the mode.

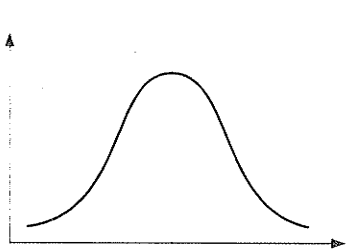
If we place a curve over the column graph (or dot plot) we see that this curve shows symmetry and we say that we have a **symmetrical distribution** of the data.

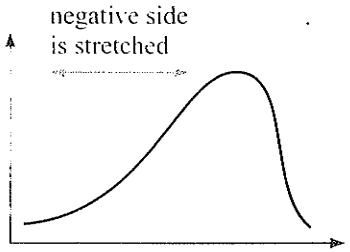


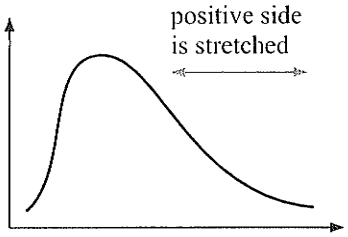
For the *Without fertiliser* data we have: This distribution is said to be **negatively skewed** as if we compare it with the symmetrical distribution it has been 'stretched' on the left (or negative) side of the mode.



So we have:



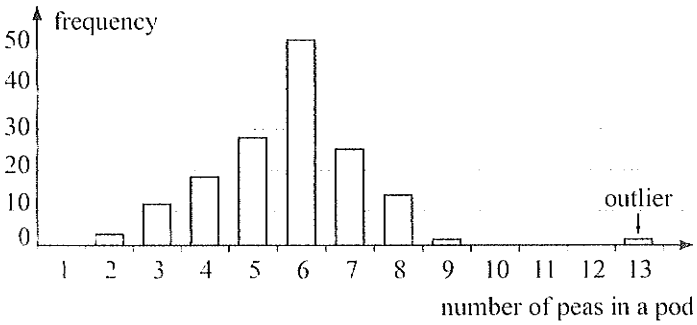




OUTLIERS

Outliers are data values that are either much larger or much smaller than the general body of data. Outliers appear separated from the body of data on a frequency graph.

For example, if the farmer in **The Pea Problem** (page **113**) found one pod in the *Without fertiliser* sample contained 13 peas, then the data value 13 would be considered an outlier. It is much larger than the other data in the sample. On the column graph it would appear separated.



EXERCISE 5B

1 State whether the following quantitative (or numerical) variables are discrete or continuous:

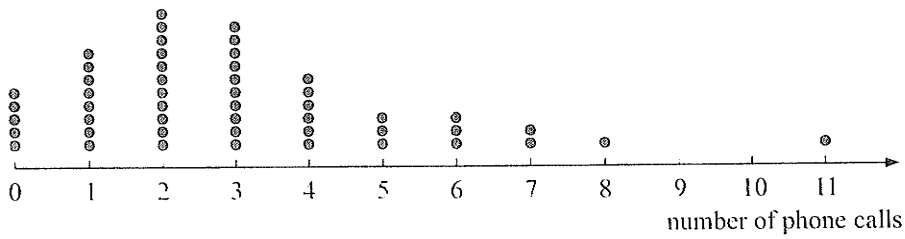
a the time taken to run 100 metres
b the maximum temperature reached on a January day
c the number of matches in a box
d the weight of luggage taken on an aircraft
e the time taken for a battery to run down
f the number of bricks needed to build a house
g the number of passengers on a bus
h the number of minutes spent on the internet per day

2 A class of 20 students was asked "How many pets do you have in your household?" and the following data was collected:

0 1 2 2 1 3 4 3 1 2 0 0 1 0 2 1 0 1 0 1

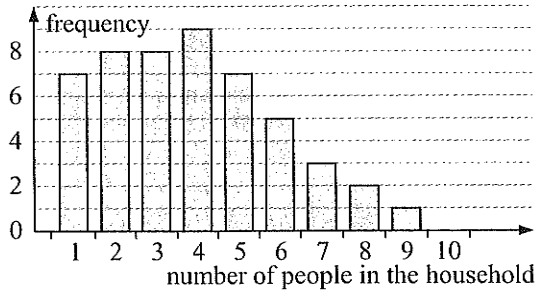
a What is the variable in this investigation?
b Is the data discrete or continuous? Why?
c Construct a dotplot to display the data. Use a heading for the graph, and scale and label the axes.
d How would you describe the distribution of the data? (Is it symmetrical, positively skewed or negatively skewed? Are there any outliers?)
e What percentage of the households had no pets?
f What percentage of the households had three or more pets?

3 For an investigation into the number of phone calls made by teenagers, a sample of 50 fifteen-year-olds were asked the question "How many phone calls did you make yesterday?" The following dotplot was constructed from the data:



a What is the variable in this investigation?
b Explain why the data is discrete numerical data.
c What percentage of the fifteen-year-olds did not make any phone calls?
d What percentage of the fifteen-year-olds made 5 or more phone calls?
e Copy and complete:
"The most frequent number of phone calls made was ....."
f Describe the distribution of the data.
g How would you describe the data value "11"?

4 A randomly selected sample of households has been asked, 'How many people live in your household?' A column graph has been constructed for the results.



a How many households gave data in the survey?

b How many of the households had only one or two occupants?

c What percentage of the households had five or more occupants?

d Describe the distribution of the data.

5 The number of matches in a box is stated as 50 but the actual number of matches has been found to vary. To investigate this, the number of matches in a box has been counted for a sample of 60 boxes:

51 50 50 51 52 49 50 48 51 50 47 50 52 48 50 49 51 50 50 52
52 51 50 50 52 50 53 48 50 51 50 50 49 48 51 49 52 50 49 50
50 52 50 51 49 52 52 50 49 50 49 51 50 50 51 50 53 48 49 49

a What is the variable in this investigation?

b Is the data continuous or discrete numerical data?

c Construct a frequency table for this data.

d Display the data using a bar chart.

e What percentage of the boxes contained exactly 50 matches?

6 Revisiting **The Pea Problem**. For the *With fertiliser* data:

a Organise the data in a tally-frequency table.

b Draw a column graph of the data.

c Are there any outliers?

d What evidence is there that the fertiliser 'increases the number of peas in a pod'?

e Can it be said that the fertiliser will increase the farmer's pea crop and his profits?

C GROUPED DISCRETE DATA

A local kindergarten is concerned about the number of vehicles passing by between 8.45 am and 9.00 am. Over 30 consecutive week days they recorded data.

The results were: 27, 30, 17, 13, 46, 23, 40, 28, 38, 24, 23, 22, 18, 29, 16,
35, 24, 18, 24, 44, 32, 52, 31, 39, 32, 9, 41, 38, 24, 32

In situations like this we group the data into **class intervals**.

It seems sensible to use class intervals of length 10 in this case.

The tally frequency table is:

Number of cars	*Tally*	*Frequency*
0 to 9	|	1
10 to 19	𝍸	5
20 to 29	𝍸 𝍸	10
30 to 39	𝍸 |	9
40 to 49	𝍸	4
50 to 59	|	1
	Total	30