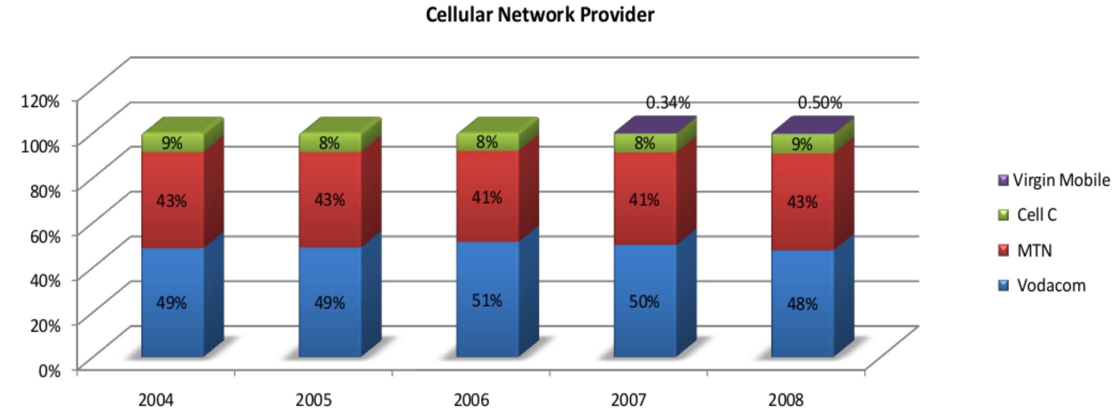
Table of Contents

Data and Data Collection....................3
Data....................3
Primary data: 3
Quantitative data: 3
Qualitative data: 4
Secondary data: 4
Data collection 4
Worked example 1....................5
Worked example 2....................5
Worked example 3....................6
Exercise : Data and data collection....................6

Summarising data....................7
Averages or measures of central tendency....................7
Mean 8
Median....................8
Worked Example : Median....................9
Worked Example : Median....................10
Mode....................10
Worked Example : Mode....................11
Worked example : Measures of central tendency....................12
Measures of Dispersion....................12
Range....................12
Worked Example : Range....................13
Quartiles....................13
Worked example : Quartiles....................14
Inter-quartile Range....................15
The semi-interquartile range 16
Percentiles....................16
Five number summary....................16
Box and whisker plots....................16
Worked Example :....................17
Box and whisker plot....................17
Worked Example : All measures of central tendency and dispersions....................19
Exercises : All measures of central tendency and dispersions....................19
Grouping Data....................20
Worked example 1 : Grouping data....................20
Exercises : Grouping data....................21
Measures of central tendency and dispersion for grouped data....................23
Worked example 1 : Measures of central tendency for grouped data....................23
Worked example : Measures of central tendency for grouped data....................24
Exercises : Measures of central tendency for grouped data....................26
Applications....................27
Exercises - Application of Statistics....................27
End of chapter summary....................28
End of Chapter Exercises....................29

Statistics 9

Information in the form of numbers, graphs and tables is all around us; on television, on the radio or in the newspaper. We are exposed to crime rates, sports results, rainfall, government spending, rate of HIV/AIDS infection, population growth and economic growth. The figure below is an example of statistical information about the market share of local cellular service provider in South Africa.



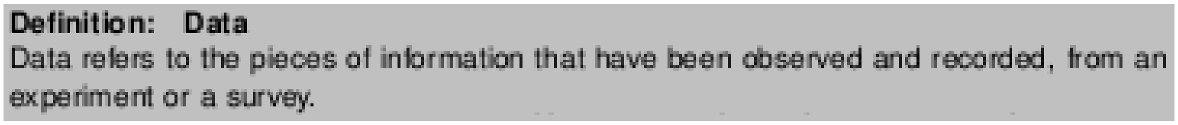
A visual representation of the market share of local cellular service providers

This chapter demonstrates how mathematics can be used to manipulate data, to represent or misrepresent trends and patterns and to provide solutions that are directly applicable to the world around us. Skills relating to the collection, organisation, display, analysis and interpretation of information that were introduced in earlier grades are developed further.

Data and Data Collection

Data

The collection of data has been introduced in earlier grades as a method of obtaining answers to questions about the world around us. This work will be briefly reviewed.



Definition: Data
Data refers to the pieces of information that have been observed and recorded, from an experiment or a survey.

The word "data" is the plural of the word "datum", and therefore one should say, "the data are" and not "the data is".

Data can be classified as primary or secondary, and primary or secondary data can be classified as qualitative or quantitative.

Primary data:

describes the original data that have been collected. This type of data is also known as raw data. Often the primary data set is very large and is therefore summarised or processed to extract meaningful information.

Quantitative data:

Quantitative data deals with numbers and are also called numerical data

- It can be measured or counted and written as numbers
- Length, time, weight, ages, cost, height, etc.
- Numerical data can be discrete or continuous
- **Discrete data** are counted and can only take certain values. Examples are :
  - The number of SMS messages sent per day.
  - The number of students in a class. (You cannot have or count half a student.)

- **Continuous data** are measured and take any value. Examples are:
  - The amount of money spent on airtime by a learner in a year.
  - The length of time of a cellphone conversation.
  - Learner's heights or their times for running the 100m sprint.

These measurements are not restricted to fixed values but can even be measured in fractions of a second, millimeters or cents and take on any value.

- Calculations can be done with the values to find the totals, averages and other meaningful

Qualitative data:

Qualitative data deals with descriptions and are also called descriptive or categorical data

- It can be observed but can not be measured or written as numbers
- Choices, colours, feelings, tastes, gender, etc
- Examples are:
    - Which cellular provider a learner prefers
    - How satisfied a customer is with their cellular service provider

- It is not possible to do the same calculations to find the average

Secondary data:

Secondary data is primary data that has been summarised or processed. For example, the percentage of market share held by each cellular operator would be secondary data because it is processed from a number of individual responses

Transforming primary data into secondary data through analysis, grouping or organisation into secondary data is the process of generating information.

Data collection

Data is collected to provide answers that help with understanding a particular situation. Common ways of gathering data include questionnaires, surveys and interviews , experiments or
The most important aspect of each method of data collecting is to clearly formulate the question that is to be answered. The details of the data collection should therefore be structured to take your question into account.

Before the data collecting starts, it is important to decide how much data is needed to make sure that the results give an accurate reflection to the required answers. Ideally, the study should be designed to maximise the amount of information collected while minimising the effort. The concepts of populations and samples is vital to minimising effort. The following terms should be familiar:

Population: describes the entire group under consideration in a study. For example, if you wanted to know how much learners in your schools spend on airtime, then the population would be all the learners in your school.

Sample: describes a group chosen to represent the population under consideration in a study. For example, for the survey on airtime spending, you might select a sample of learners, maybe one from each class.

Random sample: describes a sample chosen from a population in such a way that each member of the population has an equal chance of being chosen.

The most accurate results are obtained if the entire population is sampled for the survey, but this is expensive and time-consuming. The next best method is to randomly select a sample of subjects for the interviews.

Worked example 1

Question:
Andrew is interested in becoming an airtime reseller to his classmates. He would like to know how much business he can expect from them. He asked each of his 20 classmates how many SMS messages they sent during the previous day. The results were:

20	0	30	11	13	9	16	13	17	9
3	14	9	13	15	13	12	7	14	13

1. Is this data set qualitative or quantitative? If the data set is qualitative, identify whether it is discrete or continuous.
2. Identify the population and sample.
3. Do you think this survey will provide Andrew with a good estimate of the business he can expect from his classmates?

Solution:

1. The number of SMS messages is **quantitative** or numerical data. Because it can be counted it can be classified as **discreet.**
2. Andrew is interested in selling to his class, so the population would be all the learners in his class. His sample is also all the learners in his class and therefore covers the entire population.

Worked example 2

Question:
Andrew would like to know who the most popular cellular provider is among learners in his school. This time Andrew randomly selects 20 learners from the entire school and asks them which cellular provider they currently use. The 20 results were:

Cell C	MTN	Vodacom	Vodacom
Vodacom	MTN	MTN	Vodacom
Vodacom	Virgin Mobile	Vodacom	Vodacom
MTN	Cell C	Vodacom	Virgin Mobile
Vodacom	8ta	MTN	MTN

1. Is this data set qualitative or quantitative?
2. Can the average choice be calculated from this data?
3. If Andrew uses this data to buy airtime from the different suppliers, do you think he will have the right stock to supply to his class?

Solution:

1. The name of preferred operator is **qualitative** or categorical data. The choice cannot be expressed in numbers but can be assigned to one of five categories.
2. It is possible to count the number learners which fall into each category but we cannot perform

any numerical calculations on their actual choice.

3. Would Andrew be able to tell from this information which cellular provider sells the most airtime in South Africa and why?

Worked example 3

Question:

There are regulations in South Africa related to bread production to protect consumers. By law if a loaf of bread is not labelled, it must weigh 800g, with the leeway of 5 percent under or 10 percent over.

Vishnu is interested in how a well known national retailer measures up to this standard. He visited his local branch and recorded the masses of 10 different loaves of bread for a week.

Sample	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Sample 1	802.39	787.78	815.74	807.41	801.48	786.59	799.01
Sample 2	796.76	798.93	809.68	798.72	818.26	789.08	805.99
Sample 3	802.5	793.63	785.37	809.3	787.65	801.45	799.35
Sample 4	819.59	812.62	809.05	791.13	805.28	817.76	801.01
Sample 5	801.21	795.86	795.21	820.39	806.64	819.54	796.67
Sample 6	789	796.33	787.87	799.84	789.45	802.05	802.2
Sample 7	788.99	797.72	776.71	790.69	803.16	801.24	807.32
Sample 8	808.8	780.38	812.61	801.82	784.68	792.19	809.8
Sample 9	802.37	790.83	792.43	789.24	815.63	799.35	791.23
Sample 10	796.2	817.57	799.05	825.96	807.89	806.65	780.23

1. Is this data set qualitative or quantitative? If it is quantitative, specify whether it is discrete or continuous. Explain your answer.
2. Do you think Andrew would be able to tell from this survey whether customers around South Africa are getting value for their money from this supplier?

Solution:

Exercise : Data and data collection

Question:

In response to a question from Adrew, Phumza records the call duration in seconds of the last 40 phone calls she made from her cellphone.

136	109	3	242	131	5	14	14	265	2
1023	97	63	207	1	86	4	11	221	64
103	199	3	234	3	299	115	4	172	327
6	2	125	147	3	300	6	22	182	14

1. Is this data set qualitative or quantitative? If quantitative, specify whether it is discreet or continuous?

Summarising data

Once the data has been collected, it must be organised in a manner that allows for the information to be extracted most efficiently. For this reason it is useful to be able to summarise the data set by calculating a few quantities that give information about the central values and how the data values are spread about in the data set.

Averages or measures of central tendency

The three measures of central tendency whiich is used most commonly are the mean, the median and the mode.

Mean

Definition: Mean
The mean of a data set, x, denoted by $\overline{x}$, is the average of the data values, and is calculated as:

$$\overline{x} = \frac{\text{sum of all values}}{\text{number of all values}} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \qquad (17.1)$$

Method: Calculating the mean

1. Find the total of the data values in the data set.
2. Count how many data values there are in the data set.
3. Divide the total by the number of data values.

Exercise 17.2: Mean What is the mean of $x = \{10, 20, 30, 40, 50\}$?

Solution to Exercise

Step 1.

$$10 + 20 + 30 + 40 + 50 = 150 \qquad (17.2)$$

Step 2. There are 5 values in the data set.

Step 3.

$$150 \div 5 = 30 \qquad (17.3)$$

Step 4. $\therefore$ the mean of the data set $x = \{10, 20, 30, 40, 50\}$ is 30.

Median

Median

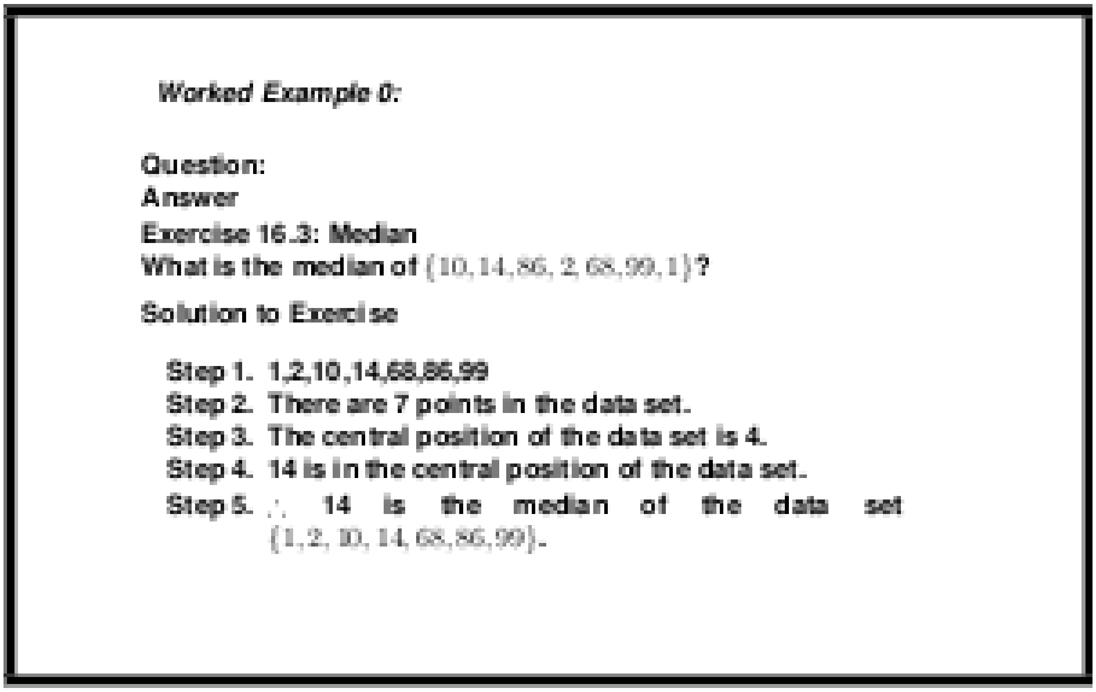
Definition: Median
The median of a set of data is the data value in the central position, when the data set has been arranged from highest to lowest or from lowest to highest. There are an equal number of data values on either side of the median value.

The median is calculated from the raw, ungrouped data, as follows.

Method: Calculating the median

- Order the data from smallest to largest or from largest to smallest.
- Count how many data values there are in the data set.
- Find the data value in the central pos

Worked Example : Median



Worked Example 0:

Question:

Answer

Exercise 16.3: Median

What is the median of $\{10, 14, 86, 2, 68, 99, 1\}$?

Solution to Exercise

Step 1. 1,2,10,14,68,86,99

Step 2. There are 7 points in the data set.

Step 3. The central position of the data set is 4.

Step 4. 14 is in the central position of the data set.

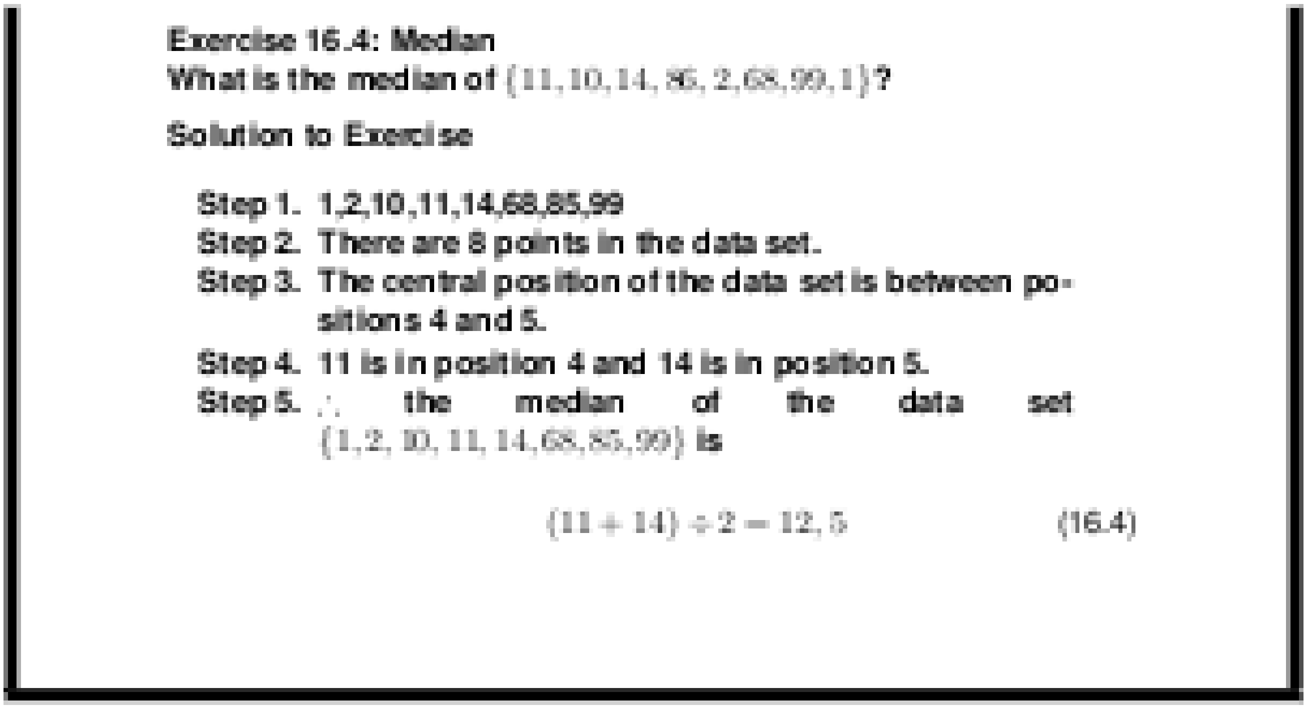
Step 5. $\therefore$ 14 is the median of the data set $\{1, 2, 10, 14, 68, 86, 99\}$.

This example has highlighted a potential problem with determining the median. It is very easy to determine the median of a data set with an odd number of data values, but what happens when there is an even number of data values in the data set?

When there is an even number of data values, the median is the mean of the two middle points.

TIP : An easy way to determine the central position or positions for any ordered data set is to take the total number of data values, add 1, and then divide by 2. If the number you get is a whole number, then that is the central position. If the number you get is a fraction, take the two whole numbers on either side of the fraction, as the positions of the data values that must be averaged to obtain the median.

Worked Example : Median



Exercise 16.4: Median

What is the median of $\{11, 10, 14, 86, 2, 68, 99, 1\}$?

Solution to Exercise

Step 1. 1,2,10,11,14,68,85,99

Step 2. There are 8 points in the data set.

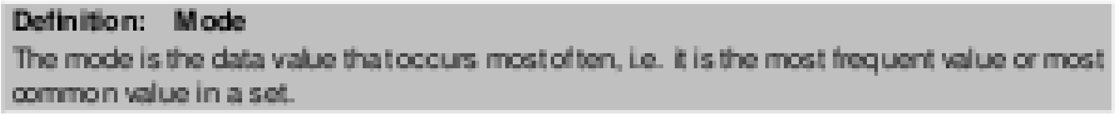
Step 3. The central position of the data set is between positions 4 and 5.

Step 4. 11 is in position 4 and 14 is in position 5.

Step 5. $\therefore$ the median of the data set $\{1, 2, 10, 11, 14, 68, 85, 99\}$ is

$$(11 + 14) \div 2 = 12,5 \qquad (16.4)$$

Mode



Definition: Mode

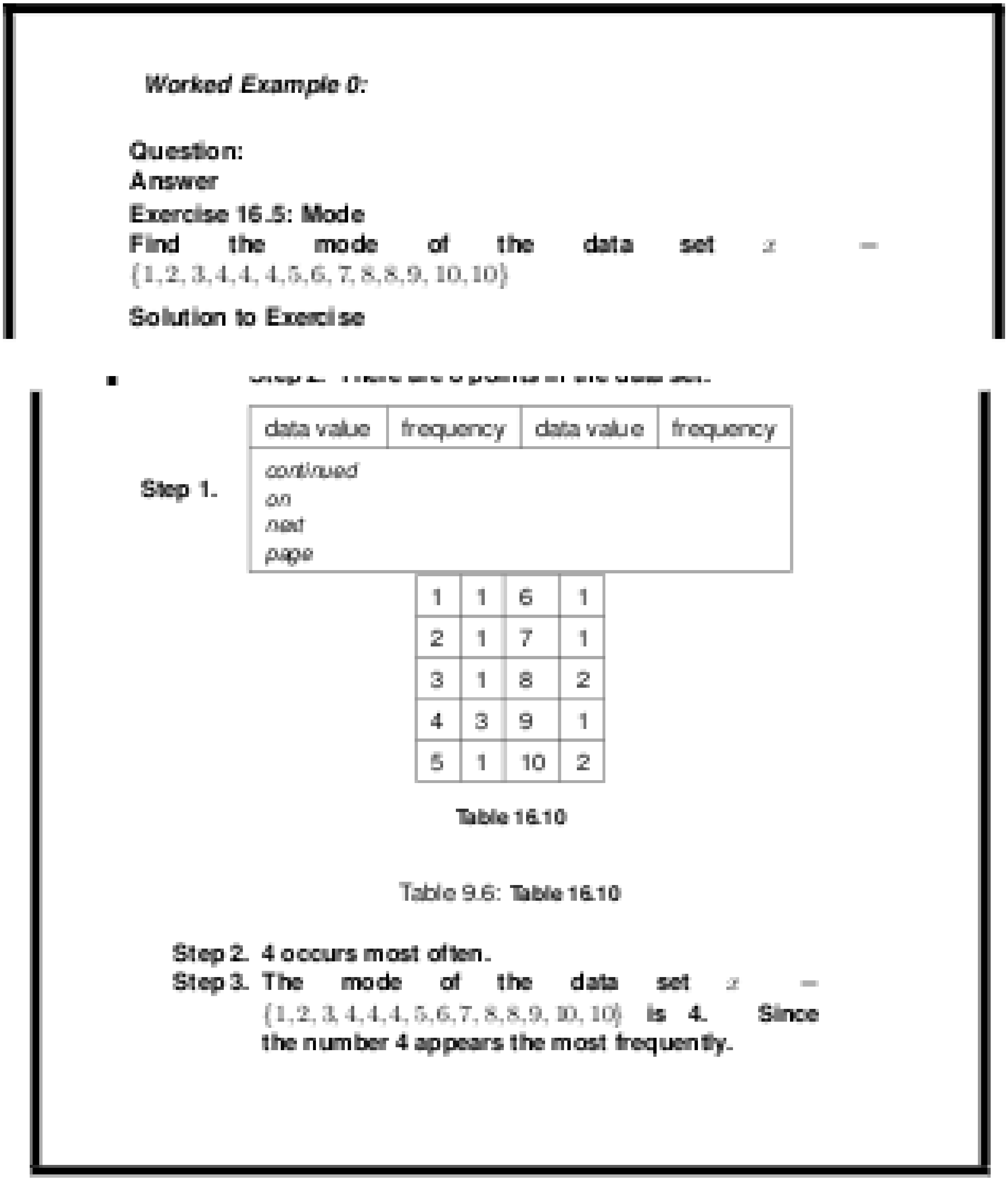
The mode is the data value that occurs most often, i.e. it is the most frequent value or most common value in a set.

Method: Calculating the mode

- Count how many times each data value occurs.

The mode is the data value that occurs the most.
The mode is calculated from grouped data, or single data items.

Worked Example : Mode



Worked Example 0:

Question:

Answer

Exercise 16.5: Mode

Find the mode of the data set $x = \{1,2,3,4,4,4,5,6,7,8,8,9,10,10\}$

Solution to Exercise

Step 2. There are [illegible] points in the data set.

Step 1.

data value	frequency	data value	frequency
continued *on* *next* *page*			
1	1	6	1
2	1	7	1
3	1	8	2
4	3	9	1
5	1	10	2

Table 16.10

Table 9.6: **Table 16.10**

Step 2. 4 occurs most often.

Step 3. The mode of the data set $x = \{1,2,3,4,4,4,5,6,7,8,8,9,10,10\}$ is 4. Since the number 4 appears the most frequently.

A data set can have more than one mode. For example, both 2 and 3 are modes in the set 1, 2, 2, 3, 3. If all points in a data set occur with equal frequency, it is equally accurate to describe the data set as having many modes or no mode.

Worked example : Measures of central tendency

Question:

Andrew is interested in becoming an airtime reseller to his classmates. He would like to know how much business he can expect from them. He asked each of his 20 classmates how many SMS messages they sent during the previous day. The results were:

20	0	30	11	13	9	16	13	17	9
3	14	9	13	15	13	12	7	14	13

1. Calculate the mean, mode and median number of messages

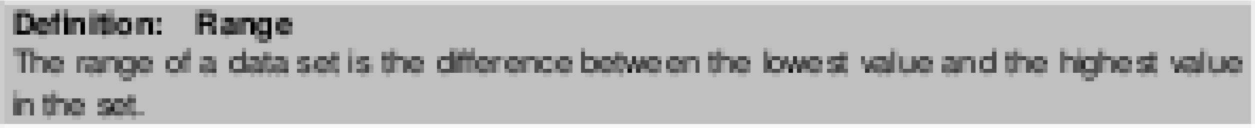
Answer:

1.

Measures of Dispersion

The mean, median and mode are measures of central tendency, i.e. they provide information on the central data values in a set. When describing data it is sometimes useful (and in some cases necessary) to determine the spread of a distribution. Measures of dispersion provide information on how the data values in a set are distributed around the mean value. Some measures of dispersion are **range, quartiles** and **percentiles**.

Range



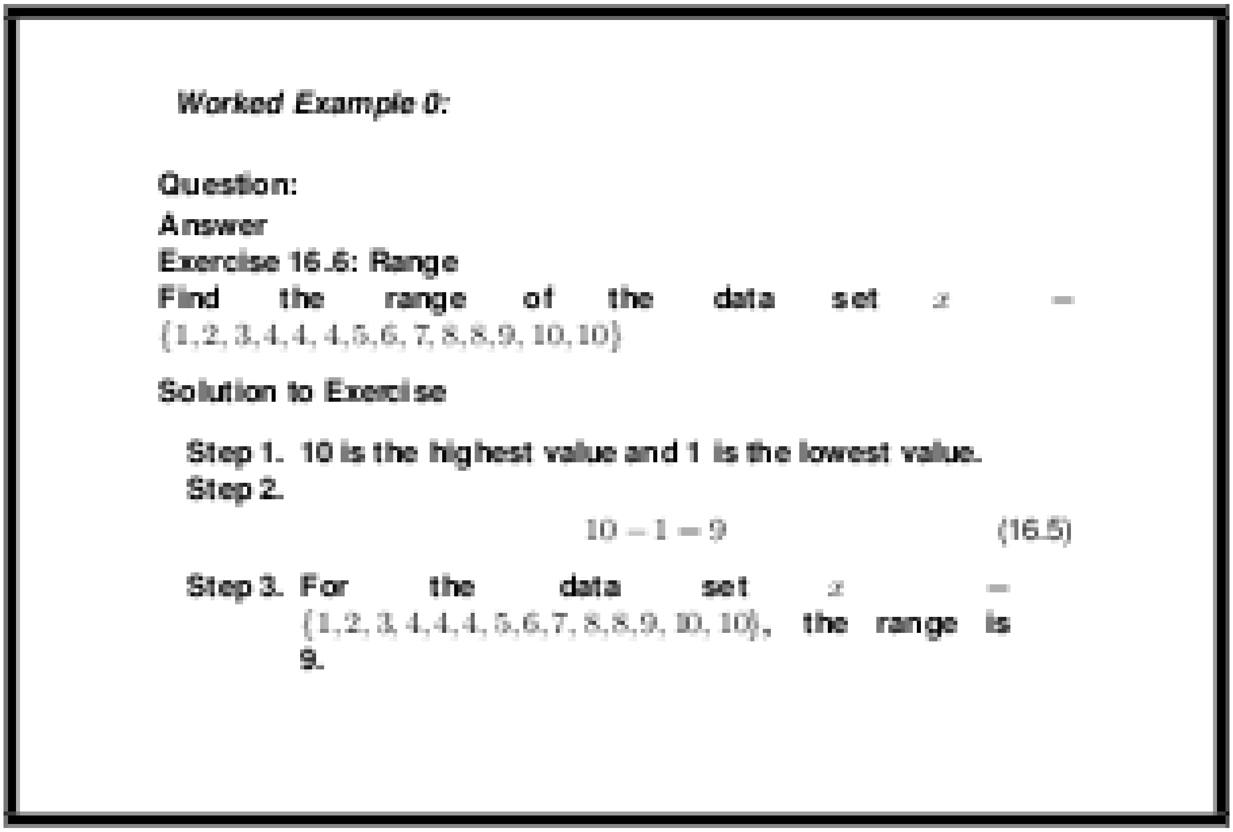
Definition: Range

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

Method: Calculating the range

- Find the highest value in the data set.
- Find the lowest value in the data set.
- Subtract the lowest value from the highest value. The difference is the range.

Worked Example : Range



Worked Example 0:

Question:

Answer

Exercise 16.6: Range

Find the range of the data set $x = \{1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 8, 9, 10, 10\}$

Solution to Exercise

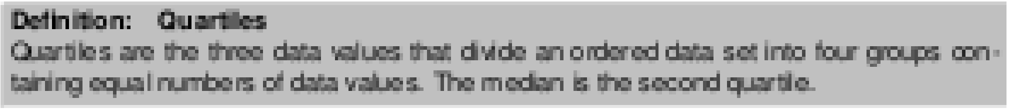
Step 1. 10 is the highest value and 1 is the lowest value.

Step 2.

$$10 - 1 = 9 \qquad (16.5)$$

Step 3. For the data set $x = \{1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 8, 9, 10, 10\}$, the range is 9.

Quartiles

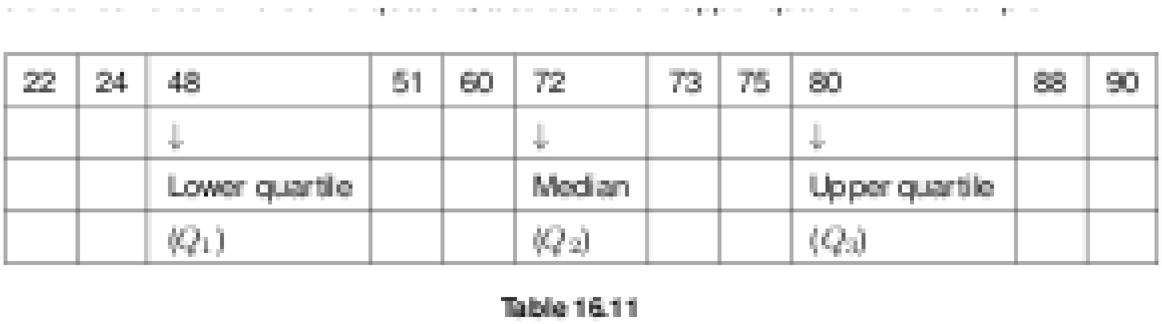


Definition: Quartiles

Quartiles are the three data values that divide an ordered data set into four groups containing equal numbers of data values. The median is the second quartile.

The quartiles of a data set are formed by the two boundaries on either side of the median, which divide the set into four equal sections. The lowest 25% of the data being found below the first quartile value, also called the lower quartile. The median, or second quartile divides the set into two equal sections. The lowest 75% of the data set should be found below the third quartile, also called the upper quartile.

For example:



22	24	48	51	60	72	73	75	80	88	90
		↓			↓			↓		
		Lower quartile			Median			Upper quartile		
		(Q_1)			(Q_2)			(Q_3)		

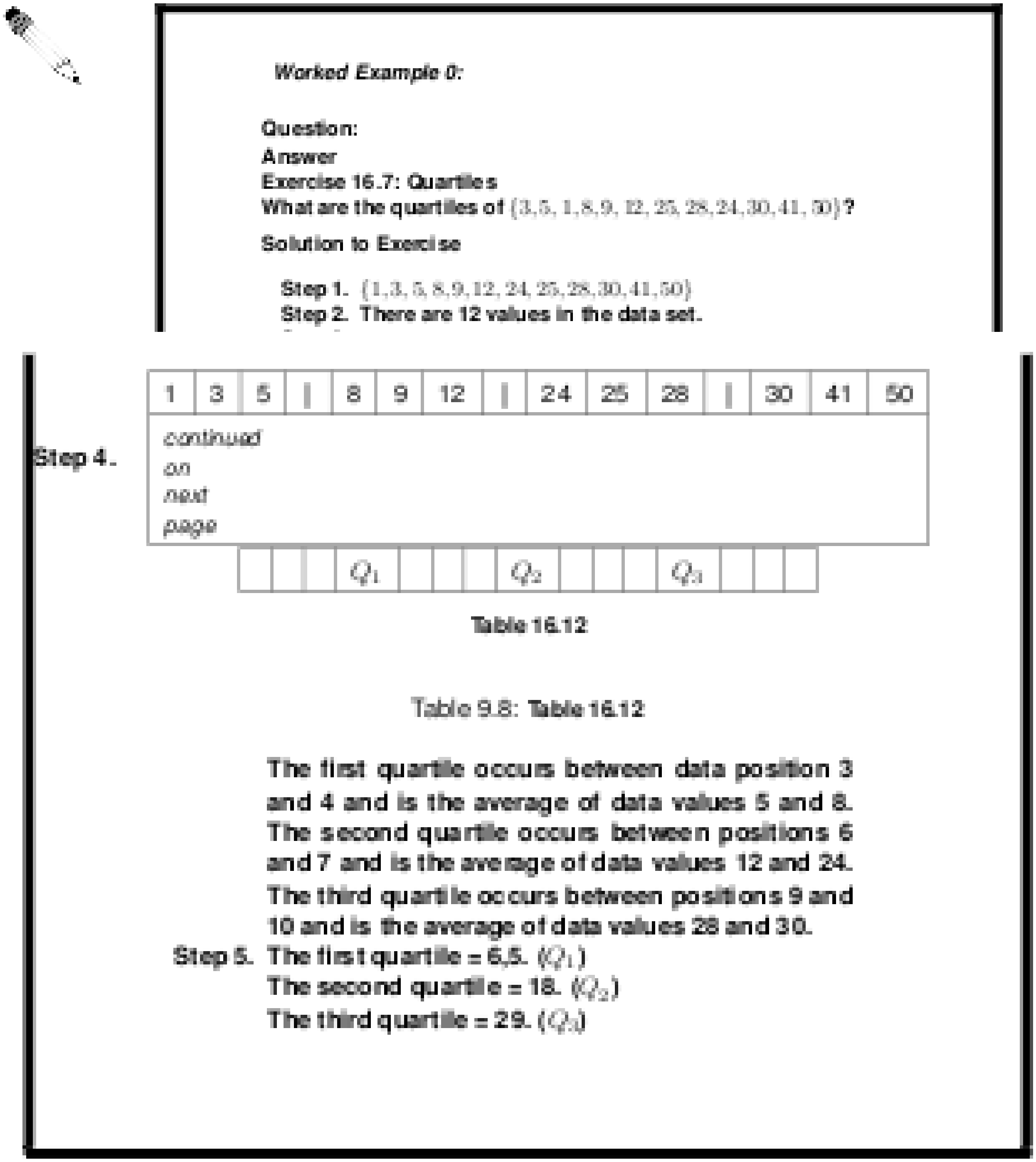
Table 16.11

Table 9.7: Table 16.11

Method: Calculating the quartiles

- Order the data from smallest to largest or from largest to smallest.
- Count how many data values there are in the data set.
- Divide the number of data values by 4. The result is the number of data values per group.
- Determine the data values corresponding to the first, second and third quartiles using the number of data values per quartile.

Worked example : Quartiles



Worked Example 0:

Question:

Answer

Exercise 16.7: Quartiles

What are the quartiles of $\{3, 5, 1, 8, 9, 12, 25, 28, 24, 30, 41, 50\}$**?**

Solution to Exercise

Step 1. $\{1, 3, 5, 8, 9, 12, 24, 25, 28, 30, 41, 50\}$

Step 2. **There are 12 values in the data set.**

Step 4.

1	3	5	|	8	9	12	|	24	25	28	|	30	41	50
continued on next page														
			Q_1				Q_2				Q_3			

Table 16.12

Table 9.8: **Table 16.12**

The first quartile occurs between data position 3 and 4 and is the average of data values 5 and 8. The second quartile occurs between positions 6 and 7 and is the average of data values 12 and 24. The third quartile occurs between positions 9 and 10 and is the average of data values 28 and 30.

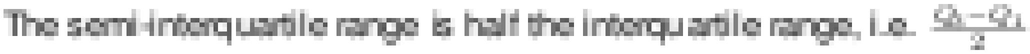
Step 5. **The first quartile = 6,5.** (Q_1)
The second quartile = 18. (Q_2)
The third quartile = 29. (Q_3)

Inter-quartile Range

Definition: Inter-quartile Range

The inter quartile range is a measure which provides information about the spread of a data set, and is calculated by subtracting the first quartile from the third quartile, giving the range of the middle half of the data set, trimming off the lowest and highest quarters, i.e.

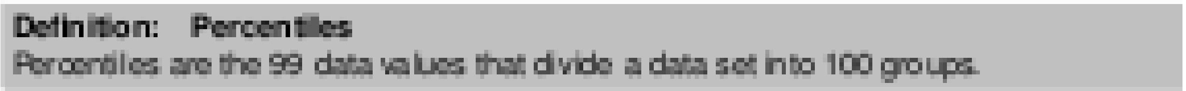
The semi-interquartile range



The semi-interquartile range is half the interquartile range, i.e. $\frac{Q_3-Q_1}{2}$.

Percentiles

Definition: Percentiles



Definition: Percentiles
Percentiles are the 99 data values that divide a data set into 100 groups.

The calculation of percentiles is identical to the calculation of quartiles, except the aim is to divide the data values into 100 groups instead of the 4 groups required by quartiles.

Method: Calculating the percentiles

- Order the data from smallest to largest or from largest to smallest.
- Count how many data values there are in the data set.
- Divide the number of data values by 100. The result is the number of data values per group.
- Determine the data values corresponding to the first, second and third quartiles using the number of data values per quartile.

Five number summary

We can summarise a data set by using the five number summary. The five number summary gives the lowest data value, the highest data value, the median, the first (lower) quartile and the third (higher) quartile. Consider the following set of data: 5, 3, 4, 6, 2, 8, 5, 4, 6, 7, 3, 6, 9, 4, 5. We first order the data as follows: 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 9. The lowest data value is 2 and the highest data value is 9. The median is 5. The first quartile is 4 and the third quartile is 6. So the five number summary is: 2, 4, 5, 6, 9.

Box and whisker plots

The five number summary can be shown graphically in a box and whisker plot. The main features of the box and whisker diagram are shown in Figure 16.4. The box can lie horizontally (as shown) or vertically. For a horizontal diagram, the left edge of the box is placed at the first quartile and the right edge of the box is placed at the third quartile. The height of the box is arbitrary, as there is no y-axis. Inside the box there is some representation of central tendency, with the median shown with a vertical line dividing the box into two. Additionally, a star or asterix is placed at the mean value, centered in the box in the vertical direction. The whiskers which extend to the sides reach the minimum and maximum values. This is shown for the data set: 5, 3, 4, 6, 2, 8, 5, 4, 6, 7, 3, 6, 9, 4, 5.

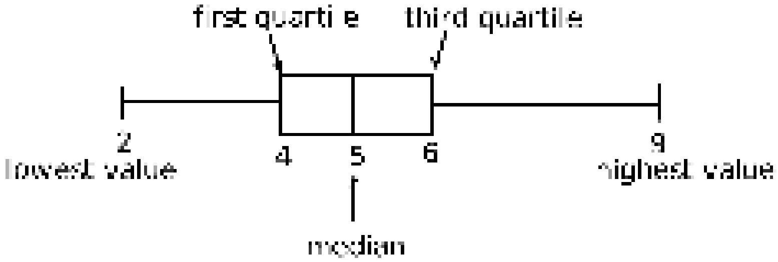


Figure 16.4: Main features of a box and whisker plot

Figure 16.4: Main features of a box and whisker plot

Worked Example :

Box and whisker plot

Worked Example 0:

Question:

Answer

Exercise 16.9

Draw a box and whisker diagram for the data set: $x = \{1,25; 1,5; 2,5; 2,5; 3,1; 3,2; 4,1; 4,25; 4,75; 4,8; 4,95; 5,1\}$.

Solution to Exercise

Step 1. $\text{Minimum} = 1,25$

$\text{Maximum} = 5,10$

The position of first quartile is between 3 and 4.

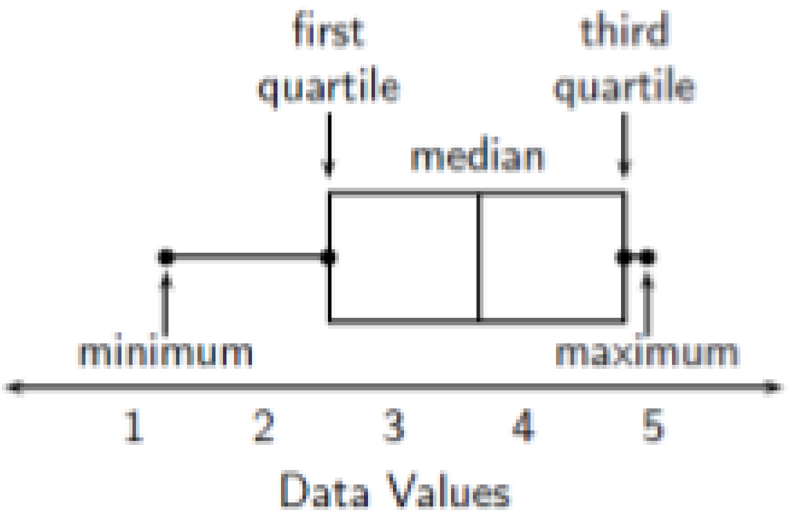
The position of second quartile is between 6 and 7.

The position of third quartile is between 9 and 10.

The data value between 3 and 4 is: $\frac{1}{2}(2,5+2,5) = 2,5$

The data value between 6 and 7 is: $\frac{1}{2}(3,2+4,1) = 3,65$

The data value between 9 and 10 is: $\frac{1}{2}(4,75+4,8) = 4,775$



Step 2.

Worked Example : All measures of central tendency and dispersions

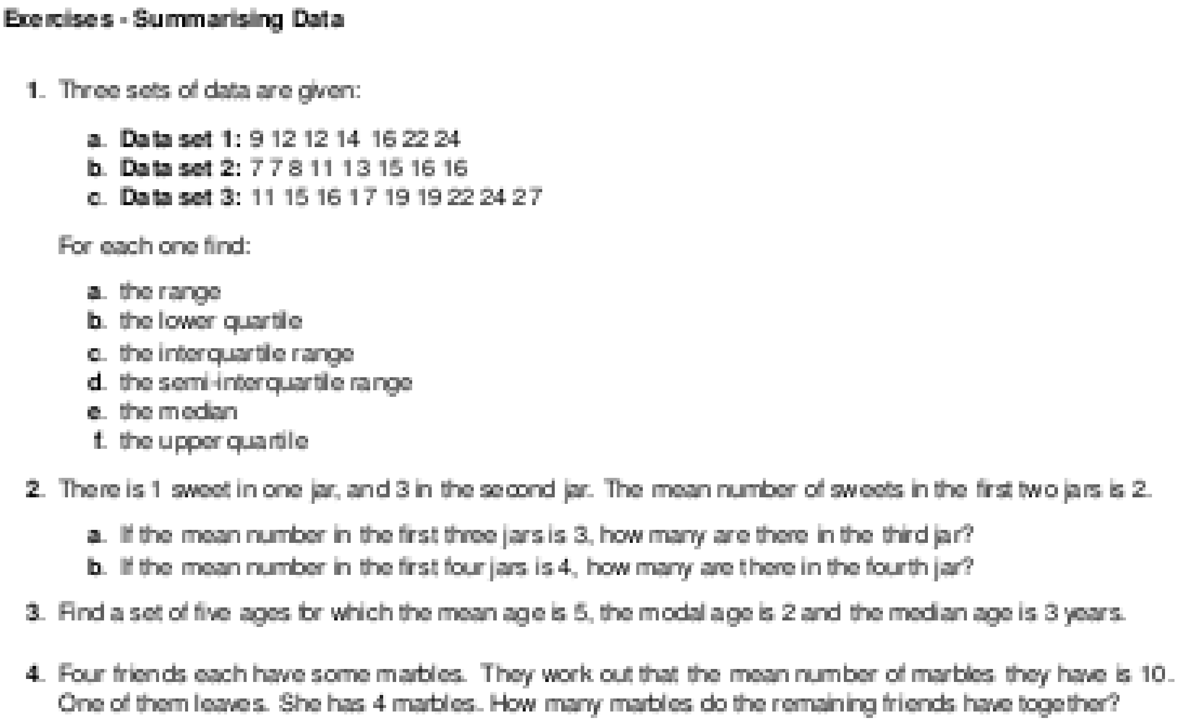
Question:

In response to a question from Adrew, Phumza records the call duration in seconds of the last 40 phone calls she made from her cellphone.

126	231	3	272	228	308	14	14	285	2
1023	327	288	30	1	133	4	11	11	310
101	137	3	245	3	109	6	4	341	95
6	2	137	9	3	154	256	22	224	14

1. What is the range?
2. Identify the quartiles....

Exercises : All measures of central tendency and dispersions



Exercises - Summarising Data

1. Three sets of data are given:

    a. **Data set 1:** 9 12 12 14 16 22 24
    b. **Data set 2:** 7 7 8 11 13 15 16 16
    c. **Data set 3:** 11 15 16 17 19 19 22 24 27

    For each one find:

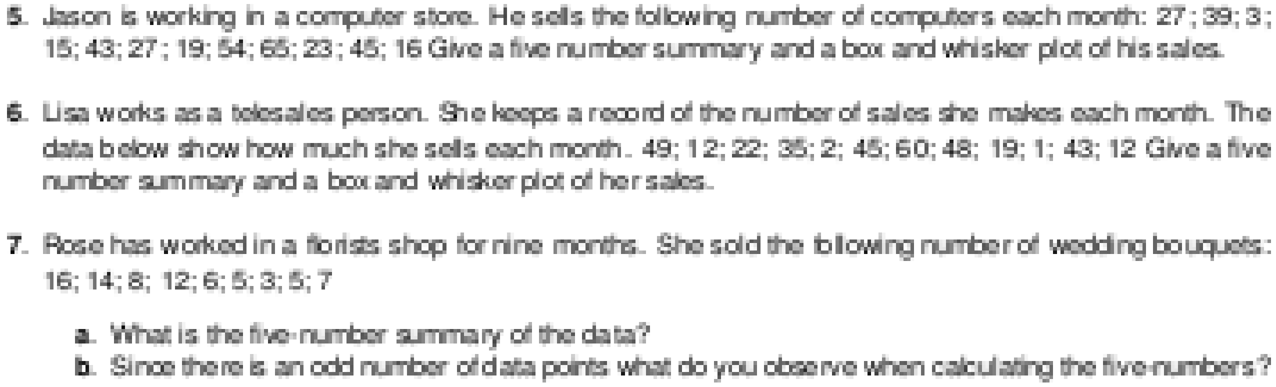
    a. the range
    b. the lower quartile
    c. the interquartile range
    d. the semi-interquartile range
    e. the median
    f. the upper quartile

2. There is 1 sweet in one jar, and 3 in the second jar. The mean number of sweets in the first two jars is 2.

    a. If the mean number in the first three jars is 3, how many are there in the third jar?
    b. If the mean number in the first four jars is 4, how many are there in the fourth jar?

3. Find a set of five ages for which the mean age is 5, the modal age is 2 and the median age is 3 years.

4. Four friends each have some marbles. They work out that the mean number of marbles they have is 10. One of them leaves. She has 4 marbles. How many marbles do the remaining friends have together?



5. Jason is working in a computer store. He sells the following number of computers each month: 27 ; 39; 3 ; 15; 43; 27 ; 19; 54; 65; 23 ; 45; 16 Give a five number summary and a box and whisker plot of his sales.

6. Lisa works as a telesales person. She keeps a record of the number of sales she makes each month. The data below show how much she sells each month . 49; 12; 22; 35; 2; 45; 60; 48; 19; 1; 43; 12 Give a five number summary and a box and whisker plot of her sales.

7. Rose has worked in a florists shop for nine months. She sold the following number of wedding bouquets: 16; 14; 8; 12; 6; 5; 3; 5; 7

   a. What is the five-number summary of the data?
   b. Since there is an odd number of data points what do you observe when calculating the five-numbers?

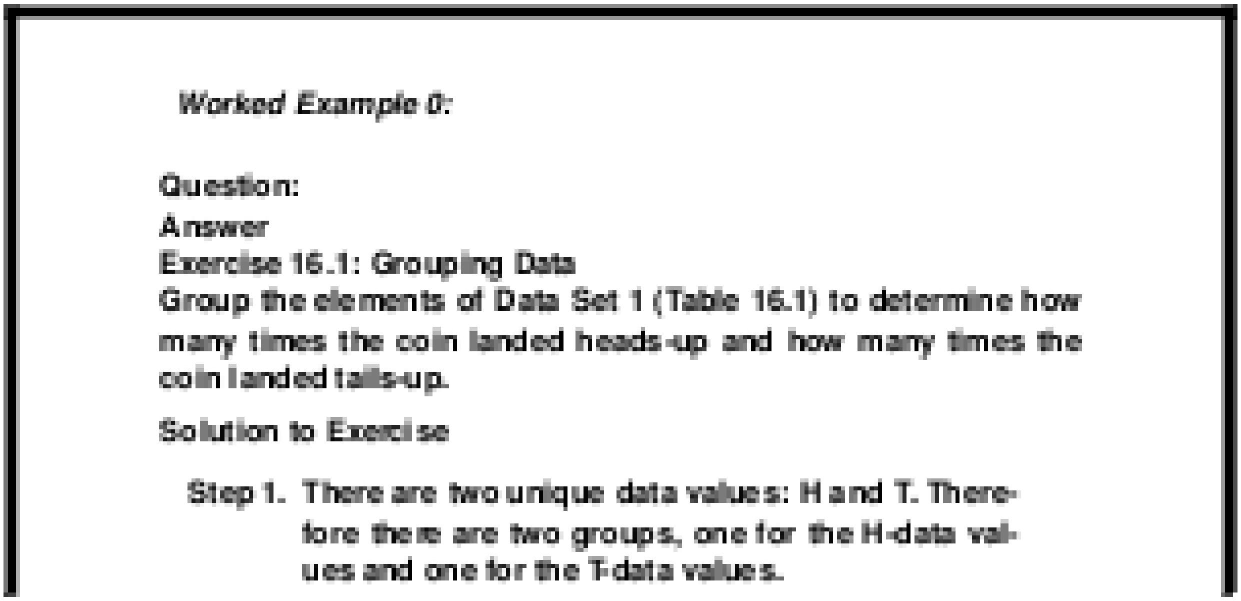
Grouping Data

One of the first steps to processing a large set of raw data is to arrange the data values together into a smaller number of groups, and then count how many of each data value there are in each group. The groups are usually based on some sort of interval of data values, so data values that fall into a specific interval, would be grouped together. The grouped data is often presented graphically or in a frequency table. (Frequency means "how many times")

Worked example 1 : Grouping data

A fair coin was tossed 100 times and the values on the top face were recorded.

H	T	T	H	H	T	H	H	H	H
H	H	H	H	T	H	H	T	T	T
T	T	H	T	T	H	T	H	T	H
H	H	T	T	H	T	T	H	T	T
T	H	H	H	T	T	H	T	T	H
H	T	T	T	T	H	T	T	H	H
T	T	H	T	T	H	T	T	H	T
H	T	T	H	T	T	T	T	H	T
T	H	T	T	H	H	H	T	H	T
T	T	T	H	H	T	T	T	H	T



Worked Example 0:

Question:

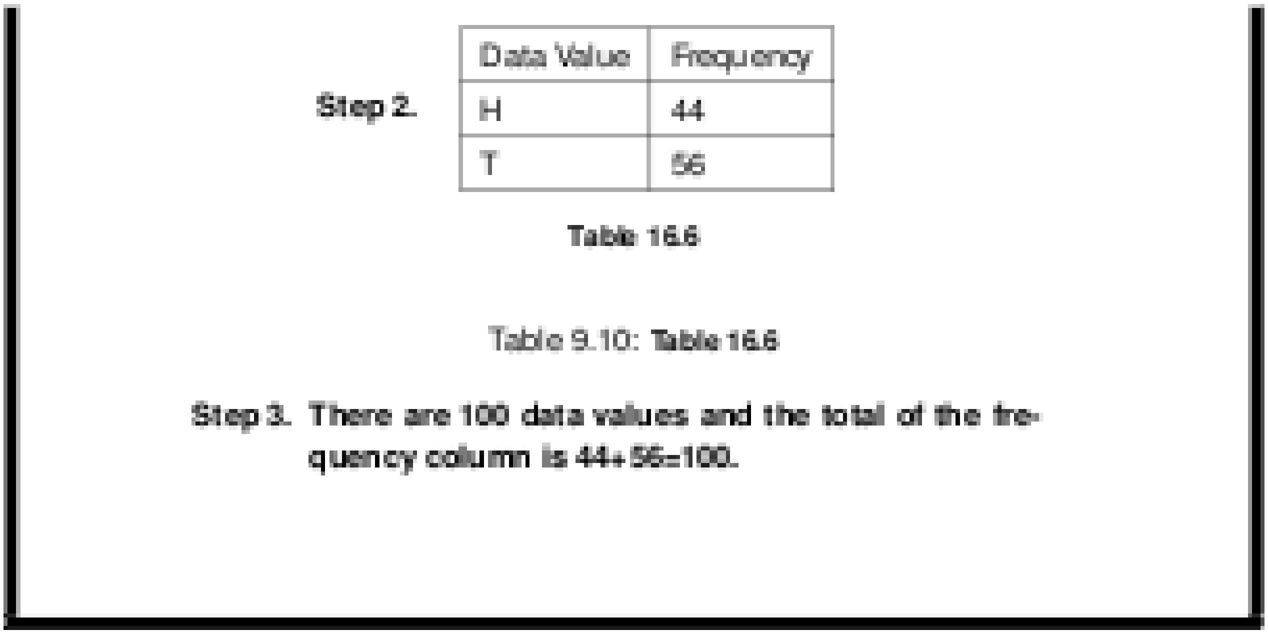
Answer

Exercise 16.1: Grouping Data

Group the elements of Data Set 1 (Table 16.1) to determine how many times the coin landed heads-up and how many times the coin landed tails-up.

Solution to Exercise

Step 1. There are two unique data values: H and T. Therefore there are two groups, one for the H-data values and one for the T-data values.



Step 2.

Data Value	Frequency
H	44
T	56

Table 16.6

Table 9.10: Table 16.6

Step 3. There are 100 data values and the total of the frequency column is 44+56=100.

Exercises : Grouping data

Exercises - Grouping Data

1. The height of 30 learners are given below. Fill in the grouped data below. (Tally is a convenient way to count in 5's. We use IIII to indicate 5.)

142	163	169	132	139	140	152	168	139	150
161	132	162	172	146	152	150	132	157	133
141	170	156	155	169	138	142	160	164	168

Table 16.7

Table 9.11: **Table 16.7**

Group	Tally	Frequency
$130 \leq h < 140$		
$140 \leq h < 150$		
$150 \leq h < 160$		
$160 \leq h < 170$		
$170 \leq h < 180$		

continued on next page

Table 16.8

Table 9.12: **Table 16.8**

2. An experiment was conducted in class and 50 learners were asked to guess the number of sweets in a jar. The following guesses were recorded.

56	49	40	11	33	33	37	29	30	59
21	16	38	44	38	52	22	24	30	34
42	15	48	33	51	44	33	17	19	44
47	23	27	47	13	25	53	57	28	23
36	35	40	23	45	39	32	58	22	40

Table 16.9

Table 9.13: **Table 16.9**

Draw up a grouped frequency table using intervals 11-20, 21-30, 31-40, etc.

Measures of central tendency and dispersion for grouped data

We can apply the concepts of mean, median and mode to data that has been grouped. Grouped data does not have individual data points, but rather has the data organized into groups or bins.

- To calculate the **mean** we need to add up all the frequencies and divide by the total. We do not know what the actual data values are, so we approximate by choosing the midpoint of each group. We then multiply those midpoint numbers by the frequency. Then we add these numbers together to find the approximate total of the masses. The modal group is the group with the highest frequency.

- The **median group** is the group that contains the middle terms.

- Measures of dispersion can also be found for grouped data.
  - The **range** is found by subtracting the smallest number in the lowest bin from the largest number in the highest bin. The quartiles are found in a similar way to the median.

Worked example 1 : Measures of central tendency for grouped data

Question:
There are regulations in South Africa related to bread production to protect consumers. By law if a loaf of bread is not labelled, it must weigh 800g, with the leeway of 5 percent under or 10 percent over.

Vishnu is interested in how a well known national retailer measures up to this standard. He visited his local branch and recorded the masses of 10 different loaves of bread for a week.

Sample	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Sample 1	802.39	787.78	815.74	807.41	801.48	786.59	799.01
Sample 2	796.76	798.93	809.68	798.72	818.26	789.08	805.99
Sample 3	802.5	793.63	785.37	809.3	787.65	801.45	799.35
Sample 4	819.59	812.62	809.05	791.13	805.28	817.76	801.01
Sample 5	801.21	795.86	795.21	820.39	806.64	819.54	796.67
Sample 6	789	796.33	787.87	799.84	789.45	802.05	802.2
Sample 7	788.99	797.72	776.71	790.69	803.16	801.24	807.32
Sample 8	808.8	780.38	812.61	801.82	784.68	792.19	809.8
Sample 9	802.37	790.83	792.43	789.24	815.63	799.35	791.23
Sample 10	796.2	817.57	799.05	825.96	807.89	806.65	780.23

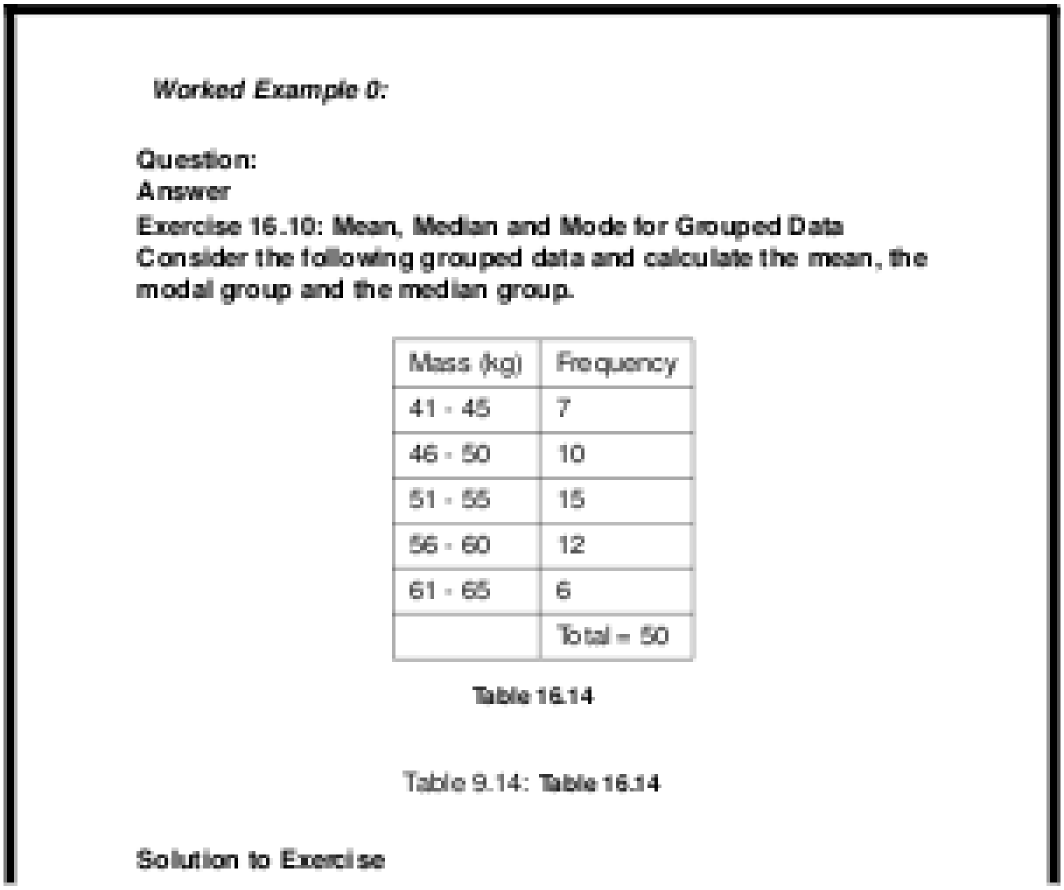
We can group this set of data by arranging

Weight	Frequency
775 < x < 780	1
780 < x < 785	3
785 < x < 790	10
790 < x < 795	7
795 < x < 800	14
800 < x < 805	12
805 < x < 810	12
810 < x < 815	2
815 < x < 820	7
820 < x < 825	1
825 < x < 830	1

1. Find the modal interval
2. An estimate of the mean
3. An estimate of the median weight of a load of bread.

Solution:

Worked example : Measures of central tendency for grouped data



Worked Example 0:

Question:

Answer

Exercise 16.10: Mean, Median and Mode for Grouped Data

Consider the following grouped data and calculate the mean, the modal group and the median group.

Mass (kg)	Frequency
41 - 45	7
46 - 50	10
51 - 55	15
56 - 60	12
61 - 65	6
	Total = 50

Table 16.14

Table 9.14: **Table 16.14**

Solution to Exercise

Step 1. To calculate the mean we need to add up all the masses and divide by 50. We do not know actual masses, so we approximate by choosing the midpoint of each group. We then multiply those midpoint numbers by the frequency. Then we add these numbers together to find the approximate total of the masses. This is show in the table below.

Mass (kg)	Midpoint	Frequency	Midpt × Freq
41 - 45	(41+45)/2 = 43	7	43 × 7 = 301
46 - 50	48	10	480
51 - 55	53	15	795
56 - 60	58	12	696
61 - 65	63	6	378
		Total = 50	Total = 2650

Table 16.15

Table 9.15: Table 16.15

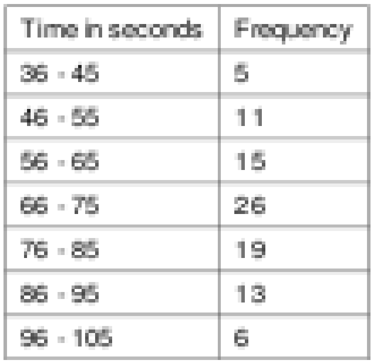
Step 2. The mean = $\frac{2650}{50} = 53$.
The modal group is the group 51 - 53 because it has the highest frequency.
The median group is the group 51 - 53, since the 25th and 26th terms are contained within this group.

Exercises : Measures of central tendency for grouped data

More mean, modal and median group exercises.

In each data set given, find the mean, the modal group and the median group.

1. Times recorded when learners played a game.

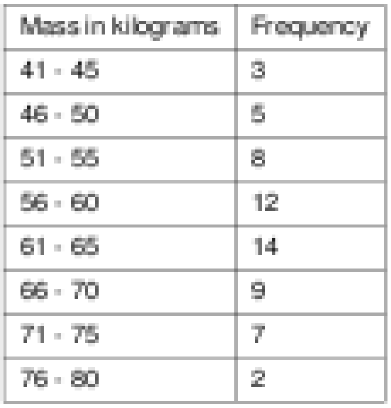


Time in seconds	Frequency
36 - 45	5
46 - 55	11
56 - 65	15
66 - 75	26
76 - 85	19
86 - 95	13
96 - 105	6

Table 16.16

Table 9.16: **Table 16.16**

2. The following data were collected from a group of learners.



Mass in kilograms	Frequency
41 - 45	3
46 - 50	5
51 - 55	8
56 - 60	12
61 - 65	14
66 - 70	9
71 - 75	7
76 - 80	2

Table 16.17

Table 9.17: **Table 16.17**

Applications

Many people take statistics and just blindly apply it to life or quote it. This, however, is not wise since the data that led to the statistics also needs to be considered. A well known example of several sets of data that lead to the same statistical analysis (the process of examining data and determining values such as central tendency, etc.) but are in fact very different is Anscombe's quartet. This is shown in . In Grade 11 you will learn about the methods used to represent data graphically. For now, however, you should simply appreciate the fact that we can plot data values on the Cartesian plane in a similar way to plotting graphs. If each of the datasets in Anscombe's quartet are analysed statistically, then one finds that the mean, variance, correlation and linear regression (these terms will be explained in later grades) are identical. If, instead of analysing the data statistically, we simply plot the data points we can see that the data sets are very different. This example shows us that it is very important to consider the underlying data set as well as the statistics that we obtain from the data. We cannot simply assume that just because we know the statistics of a data set, we know what the data set is telling us. For general interest, some of the ways that statistics and data can be misinterpreted are given in the following extension section.

Figure 16.6: Anscombe's quartet

In many cases groups can gain an advantage by misleading people with the misuse of statistics. Companies misuse statistics to attempt to show that they are performing better than a competitor, advertisers abuse statistics to try to convince you to buy their product, researchers misuse statistics to attempt to show that their data is of better quality than it really is, etc.

Common techniques used include:

- Three dimensional graphs.
- Axes that do not start at zero.
- Axes without scales.
- Graphic images that convey a negative or positive mood.
- Assumption that a correlation shows a necessary causality.
- Using statistics that are not truly representative of the entire population.
- Using misconceptions of mathematical concepts

For example, the following pairs of graphs show identical information but look very different. Explain why.

Exercises - Application of Statistics

1. A company has tried to give a visual representation of the increase in their earnings from one year to the next. Does the graph below convince you? Critically analyse the graph.

2. In a study conducted on a busy highway, data was collected about drivers breaking the speed limit

and the colour of the car they were driving. The data were collected during a 20 minute time interval during the middle of the day, and are presented in a table and pie chart below.

Conclusions made by a novice based on the data are summarised as follows:

- "People driving white cars are more likely to break the speed limit."
- "Drivers in blue and red cars are more likely to stick to the speed limit."
- Do you agree with these conclusions? Explain.

3. A record label produces a graphic, showing their advantage in sales over their competitors. Identify at least three devices they have used to influence and mislead the readers impression.

4. In an effort to discredit their competition, a tour bus company prints the graph shown below. Their claim is that the competitor is losing business. Can you think of a better explanation?

5. To test a theory, 8 different offices were monitored for noise levels and productivity of the employees in the office. The results are graphed below.

The following statement was then made: "If an office environment is noisy, this leads to poor productivity."

Explain the flaws in this thinking.

End of chapter summary

- Data types can be divided into primary and secondary data. Primary data may be further divided into qualitative and quantitative data.
- We use the following as measures of central tendency
  - mean
  - mode
  - median
- The median is the centre data value in a data set that has been ordered from lowest to highest
- The mode is the data value that occurs most often in a data set.

- The following are measures of dispersion:
  - The range of a data set is the difference between the lowest value and the highest value in the set.
  - Quartiles are the three data values that divide an ordered data set into four groups containing equal numbers of data values. The median is the second quartile.
  - Percentiles are the 99 data values that divide a data set into 100 groups.
  - The inter quartile range is a measure which provides information about the spread of a data

set, and is calculated by subtracting the first quartile from the third quartile, giving the range of the middle half of the data set, trimming off the lowest and highest quarters, i.e. Q3 − Q1 .

    - Half of this value is the semi-interquartile range.

- The five number summary is a way to summarise data. A box and whisker plot is a graphical representation of the five number summary.
- Random errors are found in all sets of data and arise from estimating data values. Bias or systematic error occurs when you consistently under or over estimate data values.
- You must always consider the data and the statistics that summarise the data

End of Chapter Exercises