Chapter Four.

Boxplots, histograms and more about describing distributions.

Box and whisker diagrams (boxplots).

A simple diagram that shows the way a set of scores is distributed is a **box and whisker diagram** or **box plot**. This type of diagram does <u>not</u> show all the individual scores (unlike a dot frequency diagram which does show all of the scores) but instead it concentrates our attention on specific features of the data.

Just as the median divides the distribution into two halves then so the **quartiles** divide the distribution into four quarters. Box and whisker diagrams show the locations of:

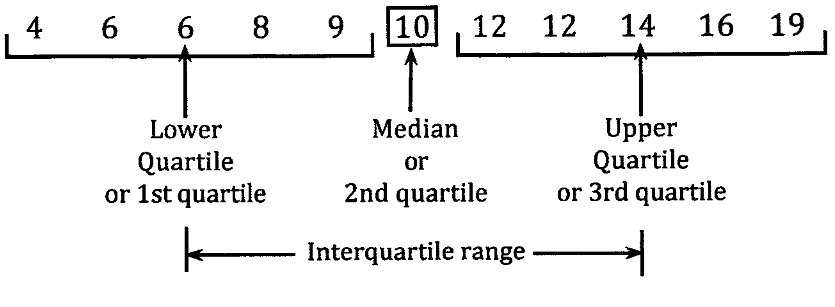
the **lowest score**,
the **highest score**,
the **median**,
the **lower quartile**
and the **upper quartile**.

Using this **five-number summary** boxplots give a visual impression of the location of the data and how widely spread it is. The range (highest score minus lowest score) and the **interquartile range** (upper quartile minus lower quartile) can be determined.

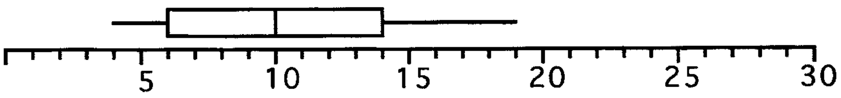
For example consider the set of scores

12, 6, 10, 19, 9, 12, 4, 14, 8, 16, 6.

Listing the scores in order allows the quartiles and interquartile range to be determined:



A box plot can then be drawn with the "box" extending from the lower quartile to the upper quartile with a line in the box indicating the median. "Whiskers" then extend from the lower quartile to the lowest score and from the upper quartile to the highest score.



Example 1

Draw box and whisker plots for each of the following sets of scores.

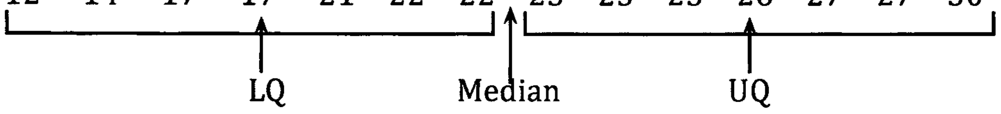
(a) 12, 22, 22, 23, 27, 14, 27, 23, 21, 30, 26, 17, 23, 17.

(b) 7, 11, 11, 11, 8, 17, 10, 12, 10, 14, 9, 15, 9.

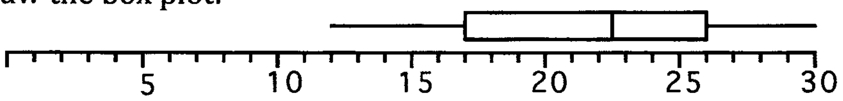
(c) 21, 18, 28, 30, 23, 17, 30, 27, 28, 19, 29, 20.

(a) Order the scores and find the median, lower quartile and upper quartile:

12 14 17 17 21 22 22 23 23 23 26 27 27 30

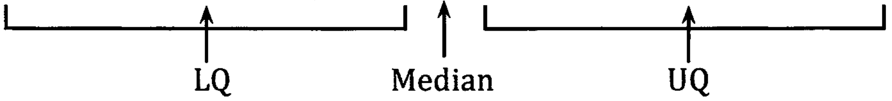


Hence draw the box plot:

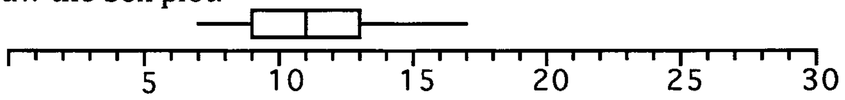


(b) Order the scores and find the median, lower quartile and upper quartile:

7 8 9 9 10 10 11 11 11 12 14 15 17

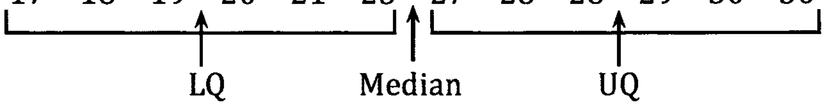


Hence draw the box plot:

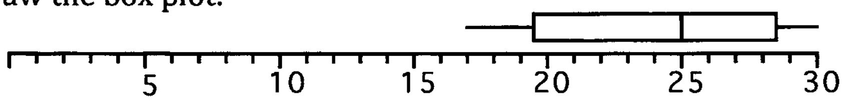


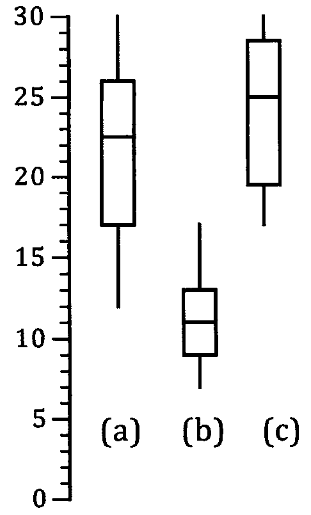
(c) Order the scores and find the median, lower quartile and upper quartile:

17 18 19 20 21 23 27 28 28 29 30 30



Hence draw the box plot:

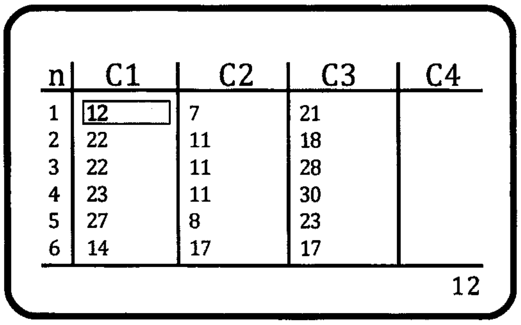




Note

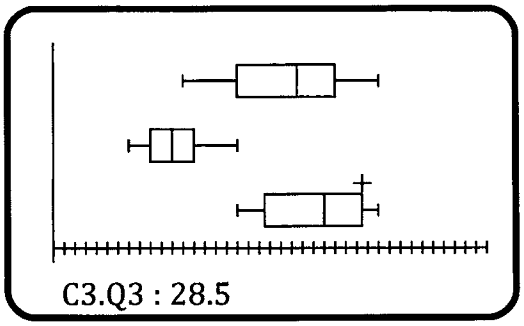
- The box and whisker diagrams shown in the previous example have all been drawn horizontally but they may also be drawn vertically, as shown on the right.

- Some graphic calculators can display data as box plots. The display below left shows the data of example 1 (remaining data could be viewed by scrolling down) and the display below right shows the box plots.



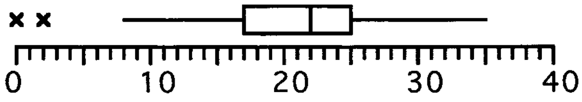
n	C1	C2	C3	C4
1	12	7	21	
2	22	11	18	
3	22	11	28	
4	23	11	30	
5	27	8	23	
6	14	17	17	

12



C3.Q3 : 28.5

- A possible refinement of this type of diagram is to indicate scores that may be considered as unusually high or unusually low compared to the others, as separate points on the diagram (i.e. show *outliers* as separate points.) The whiskers are drawn to include all scores that are within 1·5 times the interquartile range of the nearest quartile. Any scores outside that are considered outliers and are marked separately. The diagram below shows an example of this. The interquartile range is 8. The whiskers only extend to include marks that are no more than 12 marks (= 1·5 × 8) from the nearest quartile. Any scores beyond this are marked with a ✖.



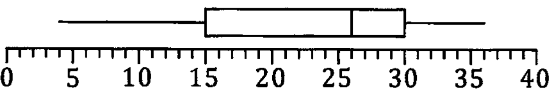
Most of the questions in this book will involve the simpler boxplots where the whiskers are drawn to the lowest and highest scores. However the idea of using "more than 1·5 × interquartile range beyond the upper and lower quartiles" as the criteria for identifying possible outliers should be remembered.

Exercise 4A.

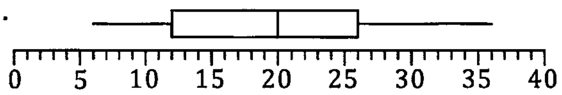
For each of the box and whisker diagrams shown in numbers 1 to 4 state:

(a) the median, (b) the lower quartile, (c) the upper quartile,
(d) the lowest score, (e) the highest score, (f) the interquartile range.

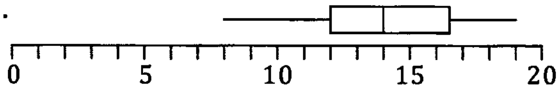
1.



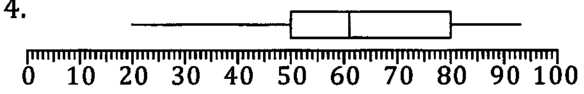
2.



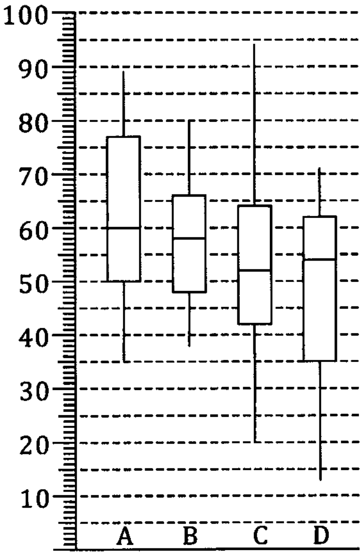
3.



4.



5. Four year 10 maths classes, A, B, C and D, take the same test, marked out of 100. The diagram on the right shows box plots for the results.
   (a) In which class is the student who scored the highest mark?
   (b) In which class is the student who scored the lowest mark?
   (c) Which class had the highest median?
   (d) Which class had the lowest median?
   (e) Which class had the smallest interquartile range?
   (f) Which class had the greatest range of marks?
   (g) Which class had the smallest range of marks?



Draw box plots for each of the following data sets.

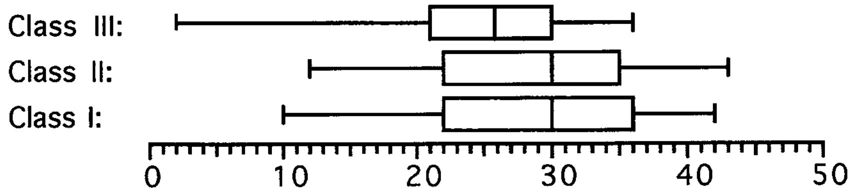
6. 5, 6, 11, 12, 12, 15, 16, 18, 18, 19, 20, 22, 25, 29, 31.

7. 11, 14, 7, 16, 16, 5, 14, 14, 24, 7, 12, 15, 14, 9.

8. 7, 10, 17, 23, 9, 12, 20, 2, 15, 5, 10, 12, 1.

9. 1, 14, 11, 25, 16, 14, 1, 1, 7, 18, 20, 5.

10. The box plots on the right are for scores achieved by three classes, in the same test.



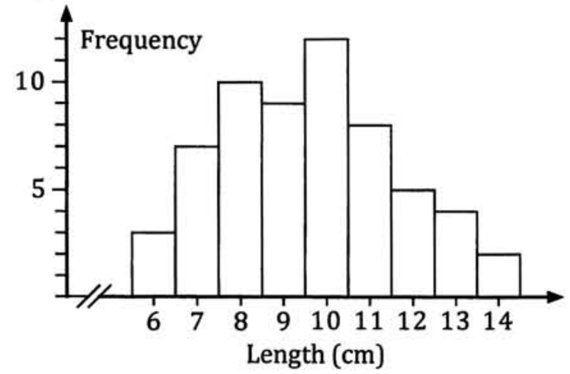
Comment on each of the following statements.
   (a) Class III had more scores below the median than above it.
   (b) The class I marks were more spread out than the class II marks.
   (c) The class III marks were more spread out than the class I marks.
   (d) The class I marks and the class II marks were similarly distributed.
   (e) Based on this test the top student in class III would be the twenty fifth student if they moved to class I.
   (f) Class III had lots of students who scored a lower mark than the lowest mark from the other two classes.

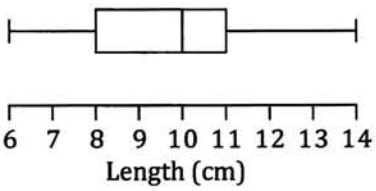
Boxplot or histogram?

The lengths of the beaks of sixty male birds of a particular species were measured and the lengths, recorded to the nearest centimetre, were as follows:

8	7	10	13	11	9	9	7	10	12	8	11	7	11	8
11	13	9	14	10	7	6	10	8	11	10	8	9	10	7
9	12	11	8	6	8	10	12	9	7	10	8	12	6	9
7	10	14	9	8	10	13	8	10	12	11	11	13	9	10

Below left shows the data displayed as a histogram and below right it is displayed as a boxplot.





Question: Which is the better form of display?

Answer: Well they are each useful in their own way and each allows us to visualise how the data is distributed.

Box plots can be drawn quickly, allow a five number summary consisting of lowest score, lower quartile, median, upper quartile and highest score to be readily obtained and the range and the interquartile range to be determined. Their compact nature and ease of production allows several boxplots to be drawn in close proximity thus allowing distributions to be compared easily.

Histograms convey the overall "shape" of a distribution allowing aspects such as symmetry, grouping, gaps, modes etc to be noticed. They allow the mean and the standard deviation of the data to be determined, or at least estimated if grouped data is involved. However certain features can be hidden if we choose too few or two many class intervals.

Hence which is "better" depends upon how much information we are wanting to show, whether we want a quickly produced visual summary of the data or a more detailed picture.

Thus box plots and histograms are both useful methods of data display, each enabling us, in their own way, to build up a picture of how a set of scores are distributed. They complement each other. Sometimes both forms of display may be given for the same set of data, as was the case above. Each form of display provides information about three key aspects of a data set:

- its **location**, (Where is it?)
- its **dispersion**, (How spread out is it?)
- its **shape**. (What does it look like?)

The three aspects of **location**, **spread** (dispersion) and **shape**, together with any other information we may notice about a distribution, were the aspects of a distribution that we were encouraged to consider when describing a distribution in chapter two. The work of chapter three now allows us to include mention of standard deviation when considering spread and the following section covers **skewness**, an aspect we can consider when describing the shape of a distribution.

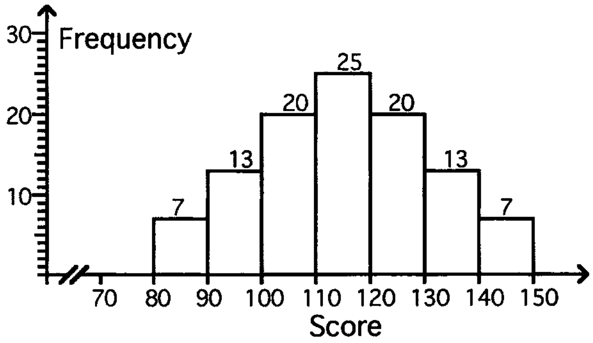
More about the shape of a distribution – skewness.

Consider the histogram shown on the right. Using the centre of each class interval to determine the mean and standard deviation of the distribution gives:

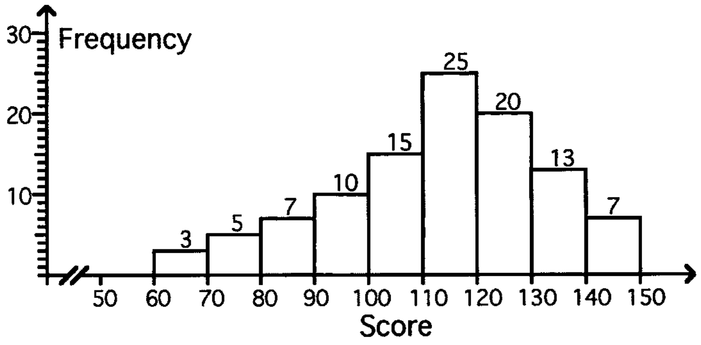
Mean = 115

(As we would expect considering the symmetry of the histogram.)

St. dev^n. (σ_n) = 16·04 (2 d.p.)



Suppose we now take the scores to the left of the central column, 40 scores in this case, and spread them further out on this left side, as shown on the second histogram. The median score would still lie in the central column, and would therefore be unchanged, but consider what will have happened to the mean and the standard deviation?



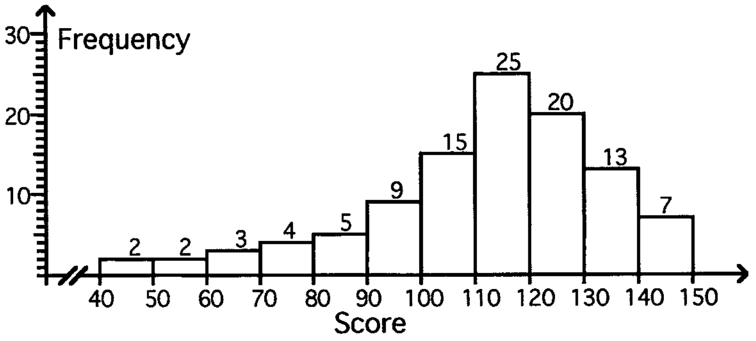
We now have more extreme scores to the left of the centre score and these scores will drag the mean left, and increase the standard deviation.

Mean = 112·71 Standard deviation (σ_n) = 19·58 (2 d.p.)

Spreading the left half of the distribution further left drags the mean further left and increases the standard deviation yet more.

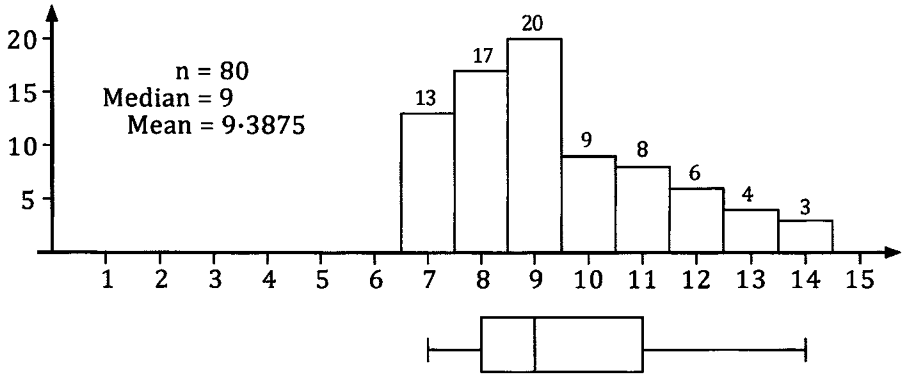
Mean = 111·38

St. dev^n. (σ_n) = 22·39 (2 d.p.)



Whilst the first histogram was symmetrical we say that the second and third histograms are **skewed to the left**, also referred to as being **negatively skewed**.

Skewed to the right (positively skewed).

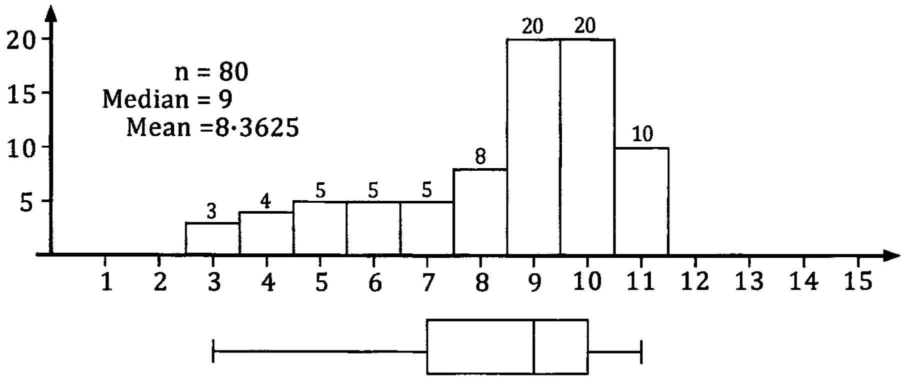


If a distribution is skewed to the right, i.e. positively skewed, the longer "tail" will be in the positive direction. The mean will usually be to the right of the median, i.e. for most positively skewed distributions we would expect

mean > median

because the "tail" of high scores to the right will tend to drag the mean right.
The box plot will tend to be longer to the right of the median than it is to the left.

Skewed to the left (negatively skewed).



If a distribution is skewed to the left, i.e. negatively skewed, the longer "tail" will be in the negative direction. The mean will usually be to the left of the median, i.e. for most negatively skewed distributions we would expect

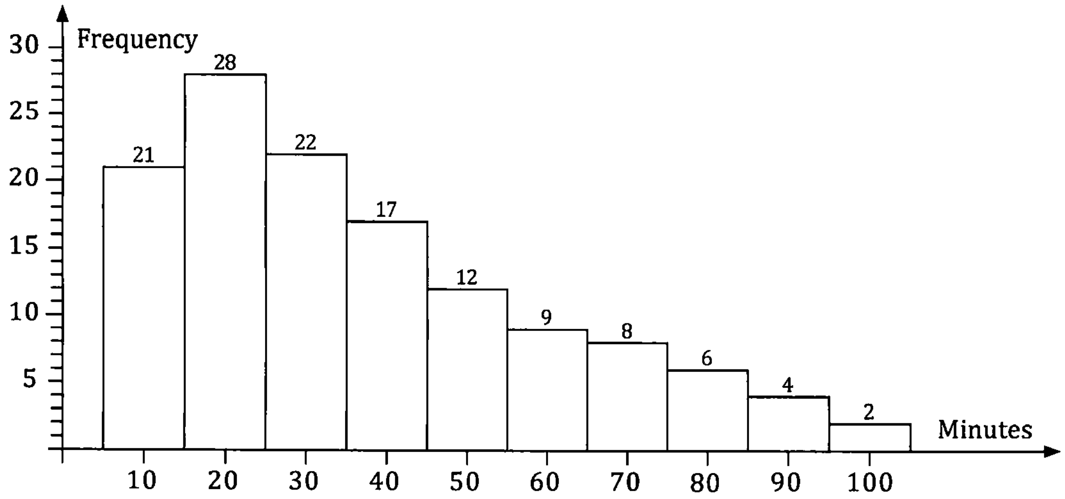
mean < median

because the "tail" of low scores will tend to drag the mean to the left.
The box plot will tend to be longer to the left of the median than it is to the right.

Note: The explanation of skewness given here is somewhat simplistic and does not, for example, consider what skewness might mean for a multimodal distribution, nor does it attempt to "quantify" skewness. However the explanation given here is sufficient for a basic understanding of the idea.

Example 2

Following a traffic warning that due to a number of accidents and roadworks long delays were likely to occur for people making their way home, the workers of one company decided that they would each record how long it took them to get home that evening. The histogram below shows the distribution of recorded times.



Describe the distribution of times.

129 times were recorded.

← Any relevant information perhaps not covered in location, spread and shape?

An estimate for the mean time is 37·6 minutes.
The median time lies in the 25 to 35 minute class.

← Comment about location. Mean, median (whichever can be determined).

The times were spread out from about 5 minutes to 105 minutes, i.e. the range was about 100 minutes.
An estimate for the standard deviation (σ_n) is 23·4 minutes.

← Comment about spread. Range, interquartile range, standard deviation (whichever can be determined).

15 - 25 minutes is the modal class (which is just the second of the ten classes).
The long tail to the right indicates that the distribution is positively skewed.

← Comment on shape. Symmetry, modality, skewness, as appropriate.

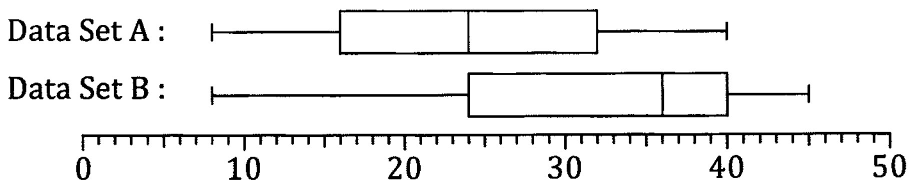
Whilst the times ranged from approx 5 minutes to 105 minutes over half (55%) were between 5 minutes and 35 minutes.

← Anything else you notice of relevance.

Descriptions could also include mention of: gaps,
clusters,
more dense/less dense regions,
extreme values or outliers.

Example 3

Compare the distributions shown in the box plots below.



The median for data set B, 36, is much higher than that of data set A, 24.
Both data sets have a lowest score of 8 but set B has the greater highest score, 45, compared to 40 for set A.

← Compare location.

Data set B has a range of 37 compared to 32 for set A.
Both data sets have an interquartile range of 16.

← Compare spread.

The box plot for set A is symmetrical and each quarter of the scores span 8 marks. On the other hand the longer left whisker of set B and the greater part of the box being to the left of the median suggest the set B marks are skewed to the left.

← Compare shape.

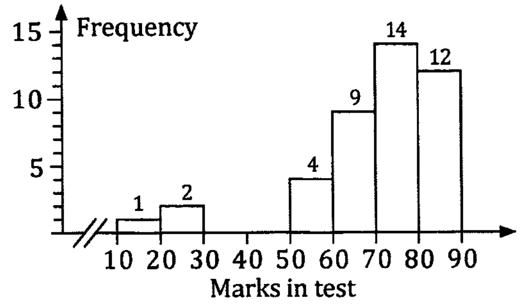
The median score in set A is the same as the lower quartile score in set B.
The top 25% of the marks from set B exceeded the top mark in set A.

← Anything else of relevance.

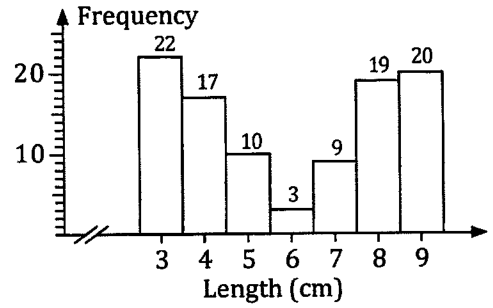
Exercise 4B.

Describe each of the distributions shown in questions 1 to 6.

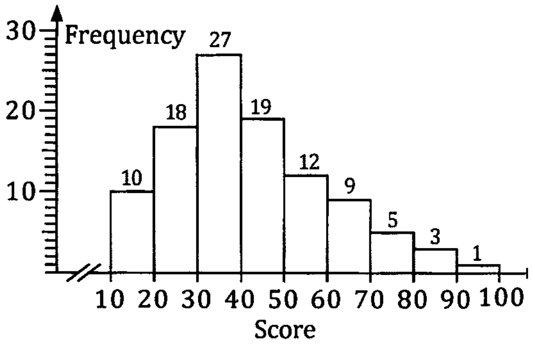
1.



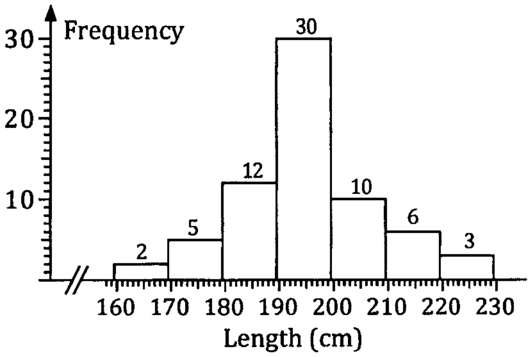
2.



3.



4.



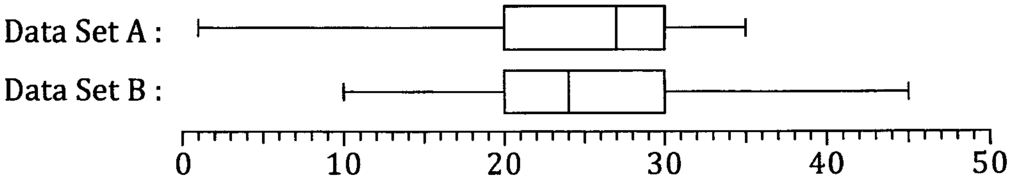
5.

Score	31-35	36-40	41-45	46-50	51-55	56-60	61-65
Frequency	15	12	16	15	13	14	15

6.

Score (x)	$0\leq x<10$	$10\leq x<20$	$20\leq x<30$	$30\leq x<40$	$40\leq x<50$	$50\leq x<60$	$60\leq x<70$
Frequency	46	29	13	6	3	2	1

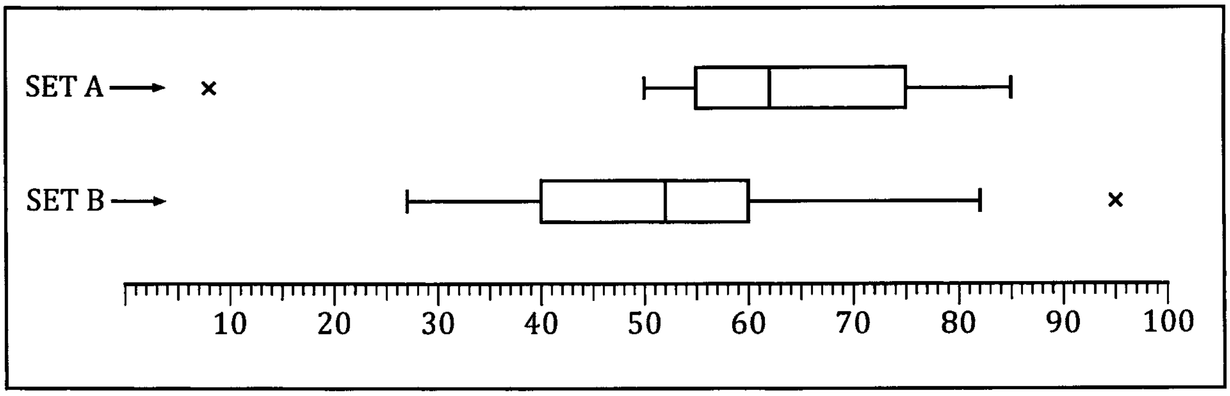
7. Compare the distributions shown in the box plots below.



8. Compare the distributions shown in the box plots below which were formed using the results of two maths classes, set A and set B, taking the same test, with set A being the top set and expected to do better than set B which was the second set. Each boxplot shows any outliers that are more than

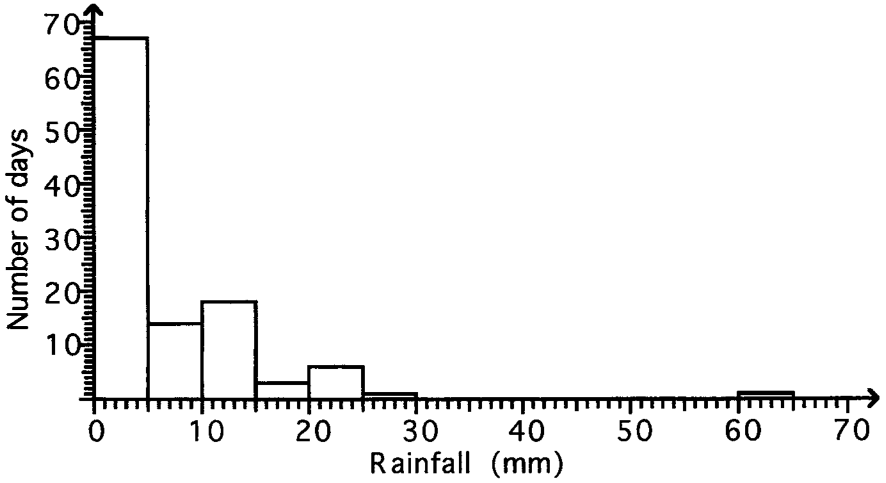
1·5 × the interquartile range from the nearest quartile

as separate crosses



9. The rainfall figures recorded at a Regional Meteorology Station for each day that some rain fell at the location, from 1st January to 31st December of a particular year, are shown in the table and graph below:

Rainfall (x mm)	$0< x <5$	$5\le x <10$	$10\le x <15$	$15\le x <20$	$20\le x <25$	$25\le x <30$	$60\le x <65$
Number of days	67	14	18	3	6	1	1



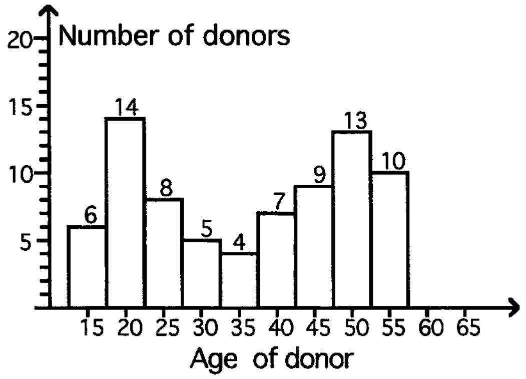
Write a report describing the rainfall for this region in the year for which the data applies.

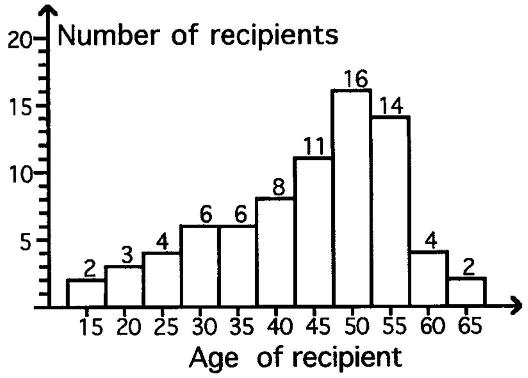
10. The table below shows the scores obtained by the 196 students sitting a particular examination.

Score	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80	81 to 90	91 to 100	101 to 110	111 to 120
Number of students	2	3	11	12	17	32	37	41	24	17

Write a summary describing the performance of the students in this examination.

11. A survey of the age of the donors and the recipients of a particular organ transplant procedure led to the following histograms:





Write a report describing and comparing the data.

Miscellaneous Exercise Four.

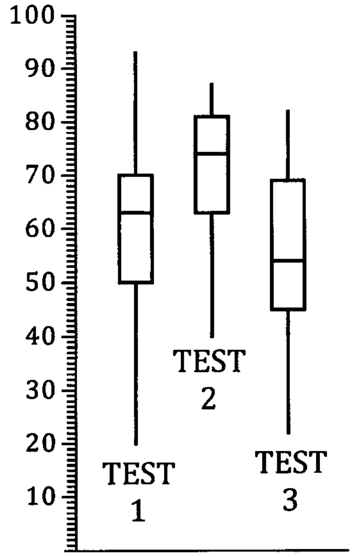
This miscellaneous exercise may include questions involving the work of this chapter, the work of any previous chapters, and the ideas mentioned in the preliminary work section at the beginning of the book.

1. Find the mean and the standard deviation (correct to two decimal places) of the following set of scores both with and without the outlier included.

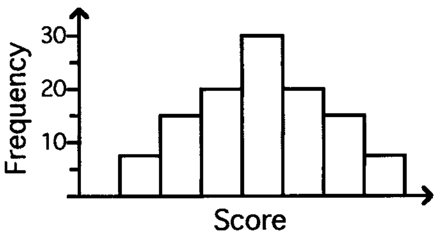
   4, 5, 6, 6, 8, 9, 9, 9, 34.

2. The diagram on the right shows boxplots for percentage scores in three tests taken by a maths class.

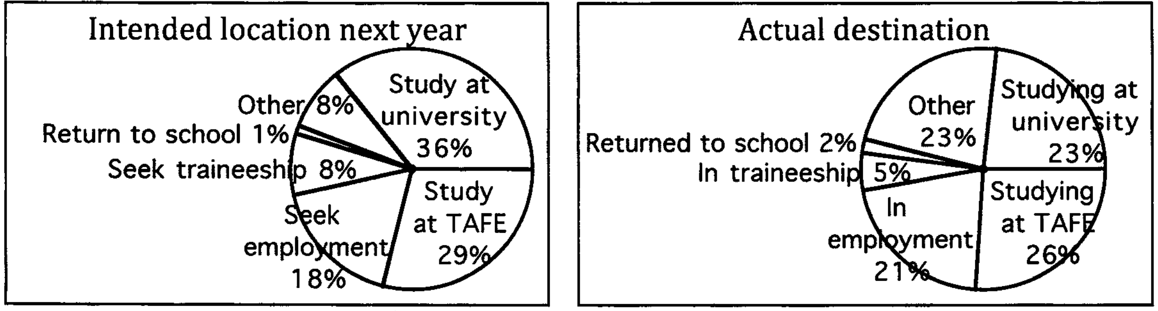
   (a) In which test was the highest mark scored?
   (b) In which test was the lowest mark scored?
   (c) Which test had the highest median?
   (d) Which test had the greatest interquartile range?
   (e) Which test had the greatest range of marks?
   (f) Which test had the smallest range of marks?
   (g) What percentage of the students scored 50% or above in test one?
   (h) In which test did at least three quarters of the students achieve a mark of more than 60%?



3. If five of the scores from the central column of the histogram shown on the right were removed from the data, would the standard deviation increase or decrease?
   Justify your answer with appropriate reasoning.



4. The pie chart below left shows (for a particular year) the intended destinations for the following year for the year 12 students in Western Australia. The pie chart below right shows where these same students actually were the following year.

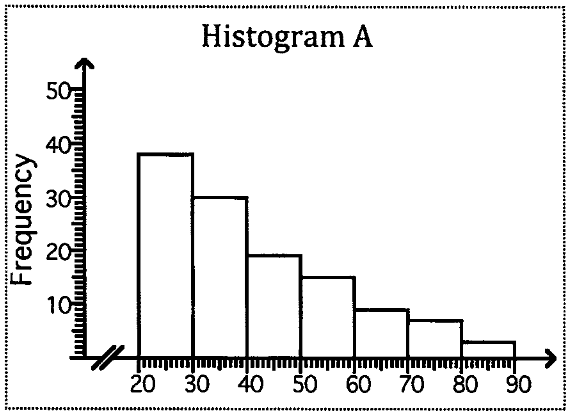


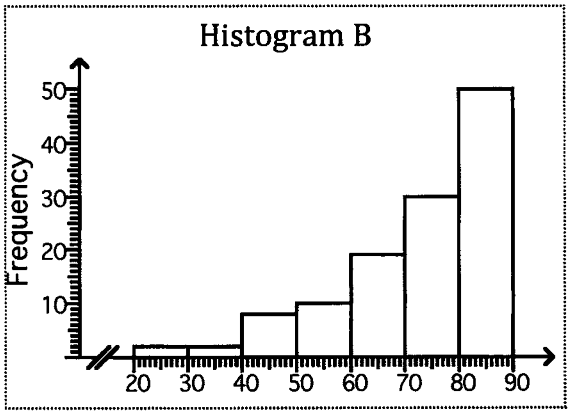
[Source of data: Western Australian Department of Education and Training.]

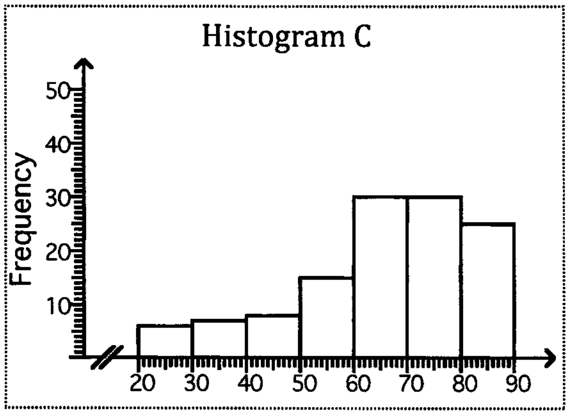
Imagine you are a newspaper reporter asked to write a short article using some of the information contained in these pie charts. Concentrate on just one or two sectors, eg university study or TAFE study, and write the article with an appropriate headline included.

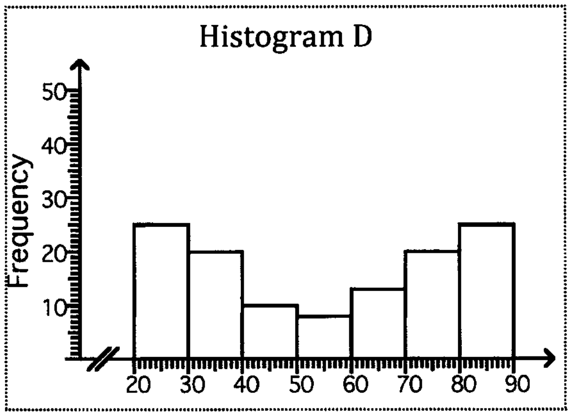
(There were approximately 18 000 year 12 students in WA that year.)

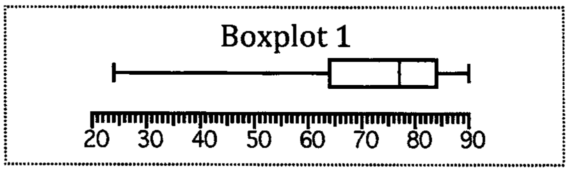
5. Given that the four sets of data that were used to create the histograms shown below were also used to create the four boxplots shown match each histogram with its corresponding boxplot.

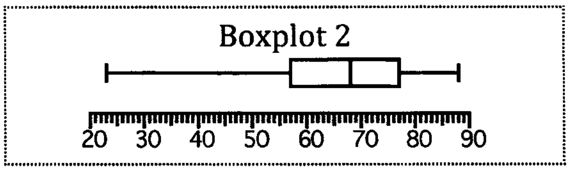

Histogram A

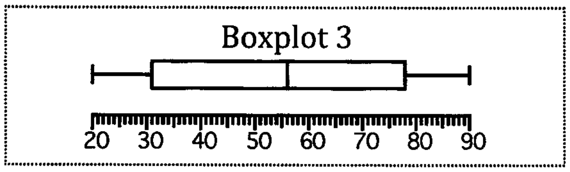

Histogram B

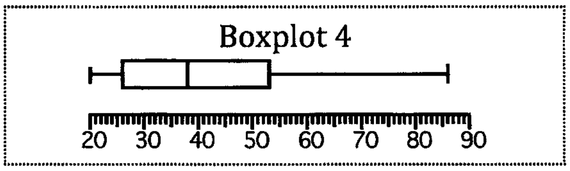

Histogram C


Histogram D


Boxplot 1


Boxplot 2


Boxplot 3


Boxplot 4

6. Classifying an outlier as:

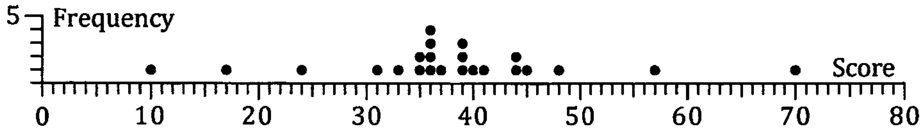
any score that is more than 1·5 × the interquartile range from either quartile 1 or quartile 3, whichever is the nearer

which scores in the following distributions would be classified as outliers?

(a) A distribution with a lower quartile of 28 and upper quartile of 40.

(b) A distribution with a median of 35, which is 5 above the lower quartile and 13 below the upper quartile.

(c)



(d)

73	63	79	91	87	75	89	77	79	57
87	82	71	85	81	98	62	85	83	87
65	93	80	89	88	42	91	68	80	75
88	83	68	82	74	80	79	78	50	88
80	49	87	77	83	78	86	62	76	80

7. A test involved ten questions and was sat by 120 students. Copy and complete the following table showing the marks obtained. (Each question was either correct, 1 mark, or incorrect, 0 marks.)

☞ Whilst you may not be familiar with the term *cumulative frequency* used in the third row of the table, with thought you should be able to determine what it means.

Mark	0	1	2	3	4	5	6	7	8	9	10
Frequency	3	0	5				15	21	18		
Cumulative frequency	3	3	8	15	28	42				113	120

(a) How many students scored a mark of 9?

(b) How many students scored a mark less than 10?

(c) What percentage of students scored a mark greater than 7 ?

(d) What fraction of students scored 3 or less?

(e) Display the data as a boxplot.

(f) Display the data as a frequency histogram.

(g) Describe the distribution.

8. Find, correct to two decimal places, the standard deviation of the five numbers:

$(a-4)$, $(a-2)$, $(a+1)$, $(a+3)$, $(a+7)$.