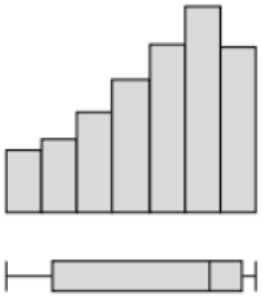
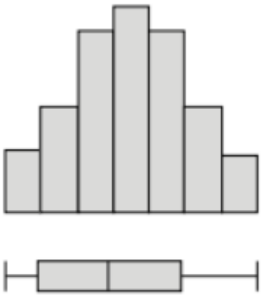
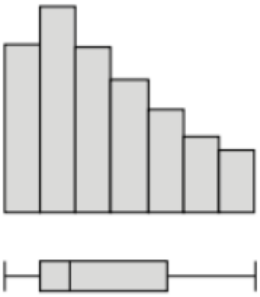
ANALYSIS OF DATA

1 Statistical displays

Statistical displays are used to analyse data, and to determine the shape of the data distribution. A data set can be symmetrical (evenly distributed), or skewed in the negative (left) or positive (right) direction. A skewed data display can also be described as stretched.

You can use histograms and box plots to illustrate the terms.

Negative skew (stretched towards the negative, or left, side of the diagram)	Symmetrical (relatively evenly distributed)	Positive skew (stretched towards the positive, or right, side of the diagram)
		

When describing the shape of a data distribution, the median and IQR are included, rather than the mean and standard deviation. This is because the mean and standard deviation can be affected by outliers.

Example 25

Describe the following data set, including a comment about its shape. The data set represents the marks obtained on a test, out of 50, by a class of 20 students.

14, 26, 49, 45, 46, 23, 24, 25, 25, 48, 49, 28, 33, 35, 37, 38, 38, 39, 41, 43

Solution

Enter the data into your technology to find the five-number summary as well as the mean and population standard deviation. Write your answers correct to 2 decimal places, when necessary:

Minimum = 14, $Q_1 = 25.5$, Median = 37.5, $Q_3 = 44$, Maximum = 49, Mean = 35.3, Standard deviation = 9.89

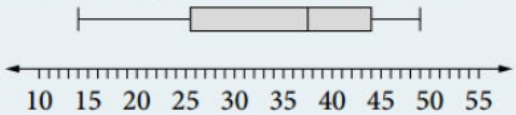
$$\text{IQR} = 44 - 25.5 = 18.5$$

$$1.5 \times \text{IQR} = 1.5 \times 18.5 = 27.75$$

$$Q_1 - 1.5 \times \text{IQR} = 25.5 - 27.75 = -2.250$$

$$Q_3 + 1.5 \times \text{IQR} = 44 + 27.75 = 71.75$$

All marks are in the interval from −2.25 to 71.75, so there are no outliers.

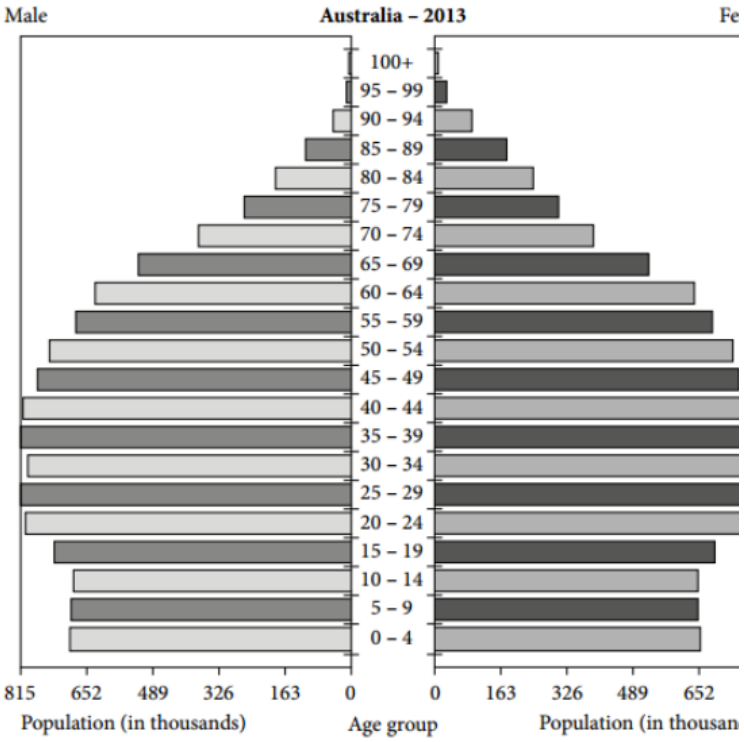


The data set contains no outliers, but is slightly negatively skewed. This is shown in the box plot where the median is closer to the upper quartile. At least 50% of the students achieved 37.5 or more, with no more than 25% obtaining less than 25.5.

ANALYSIS OF DATA

2 Composite bar graphs

A **composite bar graph** allows two data sets to be visually compared. The example below shows a comparison of a populatio based on gender and age. This particular sort of graph is sometimes referred to as a population pyramid.



3 Back-to-back stem-and-leaf plots

Another way of comparing two data sets is by using a **back-to-back stem-and-leaf plot**. These have a central stem with the leaves moving away from the stem in either direction. The example below shows the height data, measured to the nearest centimetre, for two different hockey teams.

Hockey players' heights (cm)

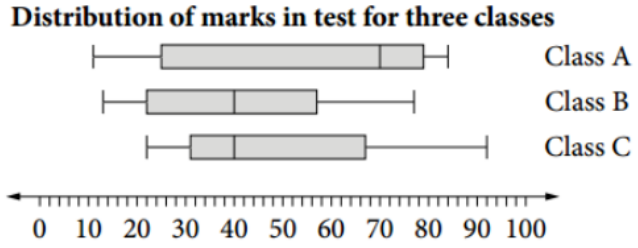
Team A		Team B
9 8 5	15	1 4 6 7
6 5 4 4 1	16	0 8
9 8 8 6 5	17	2 4 4 6
	18	3 6
	19	0

Key: 17|9 = 179

Note: Both sides have the lowest leaf values next to the stem column. So, the left-hand data set is written right to left, but the right-hand data set is written left to right.

4 Parallel box plots

A **parallel box plot** draws several box plots on the same scale. The example below compares the marks for three different classes on the same test.



Example 26

After first obtaining the five-number summary, draw parallel box plots for these two data sets.

Class A: 22, 25, 34, 12, 49, 50, 44, 48, 27, 32, 42, 44, 28, 50

Class B: 33, 37, 30, 40, 48, 49, 50, 42, 46, 21, 17, 18, 22, 25

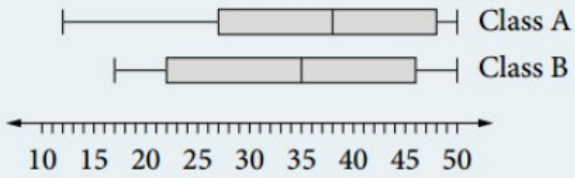
Solution

Class A: 12, 22, 25, 27, 28, 32, 34, 42, 44, 44, 48, 49, 50, 50.

Minimum = 12, $Q_1 = 27$, Median = 38, $Q_3 = 48$, Maximum = 50

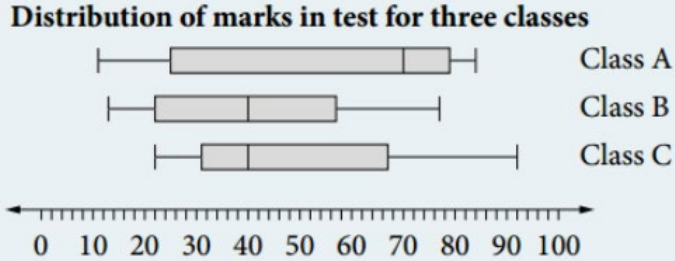
Class B: 17, 18, 21, 22, 25, 30, 33, 37, 40, 42, 46, 48, 49, 50.

Minimum = 17, $Q_1 = 22$, Median = 35, $Q_3 = 46$, Maximum = 50



Example 27

The following parallel box plots show the distribution of marks in a test for three classes.



Compare the box plots.

Solution

The box plots indicate that Class A is negatively skewed. Class B is close to symmetrical and Class C is positively skewed.

The highest mark was in Class C and the lowest was in Class A.

At least 50% of Class A obtained 70 or more, less than 25% of Class B obtained 70 or more and about 25% of Class C obtained 70 or more.

Classes B and C have the same median mark which is much less than the median mark for Class A.