Statistics - Grade 10 [CAPS]*

Free High School Science Texts Project

Based on *Statistics - Grade 10*† by

Rory Adams

Free High School Science Texts Project

Mark Horner

Heather Williams

This work is produced by The Connexions Project and licensed under the
Creative Commons Attribution License ‡

1 Introduction

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.

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.

2 Recap of Earlier Work

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.

2.1 Data and Data Collection

2.1.1 Data

Definition 1: Data
Data refers to the pieces of information that have been observed and recorded, from an experiment or a survey. There are two types of data: primary and secondary. The word "data" is the plural of the word "datum", and therefore one should say, "the data are" and not "the data is".

*Version 1.3: Jun 13, 2011 9:25 pm -0500
†http://cnx.org/content/m32627/1.4/
‡http://creativecommons.org/licenses/by/3.0/

Data can be classified as *primary* or *secondary*, and primary or secondary data can be classified as *qualitative* or *quantitative*. Figure 1 summarises the classifications of data.

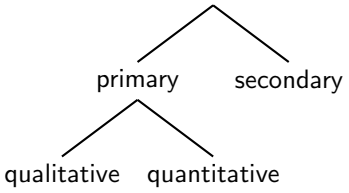


Figure 1: Classes of data.

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.

Qualitative data: is information that cannot be written as numbers, for example, if you were collecting data from people on how they feel or what their favourite colour is.

Quantitative data: is information that can be written as numbers, for example, if you were collecting data from people on their height or weight.

Secondary data: is primary data that has been summarised or processed, for example, the set of colours that people gave as favourite colours would be secondary data because it is a summary of responses.

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

2.1.2 Purpose of Collecting Primary Data

Data is collected to provide answers that help with understanding a particular situation. Here are examples to illustrate some real world data collections scenarios in the categories of qualitative and quantitative data.

2.1.3 Qualitative Data

- The local government might want to know how many residents have electricity and might ask the question: "Does your home have a safe, independent supply of electricity?"
- A supermarket manager might ask the question: “What flavours of soft drink should be stocked in my supermarket?" The question asked of customers might be “What is your favourite soft drink?” Based on the customers’ responses (i.e. which flavours are chosen), the manager can make an informed decision as to what soft drinks to stock.
- A company manufacturing medicines might ask “How effective is our pill at relieving a headache?” The question asked of people using the pill for a headache might be: “Does taking the pill relieve your headache?” Based on responses, the company learns how effective their product is.
- A motor car company might want to improve their customer service, and might ask their customers: “How can we improve our customer service?”

2.1.4 Quantitative Data

- A cell phone manufacturing company might collect data about how often people buy new cell phones and what factors affect their choice, so that the cell phone company can focus on those features that would make their product more attractive to buyers.
- A town councillor might want to know how many accidents have occurred at a particular intersection, to decide whether a robot should be installed. The councillor would visit the local police station to research their records to collect the appropriate data.
- A supermarket manager might ask the question: "What flavours of soft drink should be stocked in my supermarket?" The question asked of customers might be "What is your favourite soft drink?" Based on the customers' responses (i.e. the number of customers who liked soft drink A), the manager can make an informed decision as to what soft drinks to stock.

However, it is important to note that different questions reveal different features of a situation, and that this affects the ability to understand the situation. For example, if the first question in the list was re-phrased to be: "Does your home have electricity?" then if you answered yes, but you were getting your electricity from a neighbour, then this would give the wrong impression that you did not need an independent supply of electricity.

2.2 Methods of Data Collection

The method of collecting the data must be appropriate to the question being asked. Some examples of data collecting methods are:

1. Questionnaires, surveys and interviews
2. Experiments
3. Other sources (friends, family, newspapers, books, magazines and the Internet)

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.

For example, questionnaires, interviews or surveys would be most appropriate for the list of questions in "Purpose of Collecting Primary Data" (Section 2.1.2: Purpose of Collecting Primary Data).

2.3 Samples and Populations

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 many learners in your school got the flu each winter, then your 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 winter flu, 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.

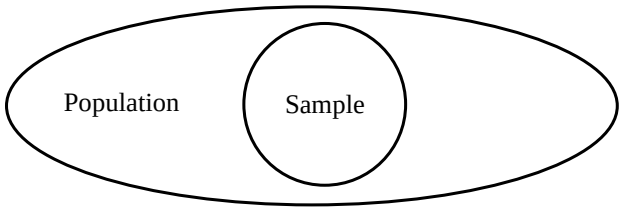


Figure 2

Choosing a representative sample is crucial to obtaining results that are unbiased. For example, if we wanted to determine whether peer pressure affects the decision to start smoking, then the results would be different if only boys were interviewed, compared to if only girls were interviewed, compared to both boys and girls being interviewed.

Therefore questions like: "How many interviews are needed?" and "How do I select the candidates for the interviews?" must be asked during the design stage of the sampling process.

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. This means that whatever the method used to select subjects for the interviews, each subject has an equal chance of being selected. There are various methods of doing this for example, names can be picked out of a hat or can be selected by using a random number generator. Most modern scientific calculators have a random number generator or you can find one on a spreadsheet program on a computer.

So, if you had a total population of 1 000 learners in your school and you randomly selected 100, then that would be the sample that is used to conduct your survey.

3 Example Data Sets

The remainder of this chapter deals with the mathematical details that are required to analyse the data collected.

The following are some example sets of data which can be used to apply the methods that are being explained.

3.1 Data Set 1: Tossing a Coin

A fair coin was tossed 100 times and the values on the top face were recorded. The data are recorded in "Data Set 1: Tossing a coin" (Table 1).

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

Table 1: Results of 100 tosses of a fair coin. H means that the coin landed heads-up and T means that the coin landed tails-up.

3.2 Data Set 2: Casting a die

A fair die was cast 100 times and the values on the top face were recorded. The data are recorded in "Data Set 2: Casting a die" (Section 3.2: Data Set 2: Casting a die).

3	5	3	6	2	6	6	5	5	6	6	4	2	1	5	3	2	4	5	4
1	4	3	2	6	6	4	6	2	6	5	1	5	1	2	4	4	2	4	4
4	2	6	4	5	4	3	5	5	4	6	1	1	4	6	6	4	5	3	5
2	6	3	2	4	5	3	2	2	6	3	4	3	2	6	4	5	2	1	5
5	4	1	3	1	3	5	1	3	6	5	3	4	3	4	5	1	2	1	2
1	3	2	3	6	3	1	6	3	6	6	1	4	5	2	2	6	3	5	3
1	1	6	4	5	1	6	5	3	2	6	2	3	2	5	6	3	5	5	6
2	6	6	3	5	4	1	4	5	1	4	1	3	4	3	6	2	4	3	6
6	1	1	2	4	5	2	5	3	4	3	4	5	3	3	3	1	1	4	3
5	2	1	4	2	5	2	2	1	5	4	5	1	5	3	2	2	5	1	1

Table 2: Results of 200 casts of a fair die.

3.3 Data Set 3: Mass of a Loaf of Bread

There are regulations in South Africa related to bread production to protect consumers. Here is an excerpt from a report about the legislation:

"The Trade Metrology Act requires that if a loaf of bread is not labelled, it must weigh 800g, with the leeway of five percent under or 10 percent over. However, an average of 10 loaves must be an exact match to the mass stipulated. - Sunday Tribune of 10 October 2004 on page 10"

We can use measurements to test if consumers getting value for money. An unlabelled loaf of bread should weigh 800g. The masses of 10 different loaves of bread were measured at a store for 1 week. The data are shown in Table 3.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
802.39	787.78	815.74	807.41	801.48	786.59	799.01
796.76	798.93	809.68	798.72	818.26	789.08	805.99
802.50	793.63	785.37	809.30	787.65	801.45	799.35
819.59	812.62	809.05	791.13	805.28	817.76	801.01
801.21	795.86	795.21	820.39	806.64	819.54	796.67
789.00	796.33	787.87	799.84	789.45	802.05	802.20
788.99	797.72	776.71	790.69	803.16	801.24	807.32
808.80	780.38	812.61	801.82	784.68	792.19	809.80
802.37	790.83	792.43	789.24	815.63	799.35	791.23
796.20	817.57	799.05	825.96	807.89	806.65	780.23

Table 3: Masses (in g) of 10 different loaves of bread, from the same manufacturer, measured at the same store over a period of 1 week.

3.4 Data Set 4: Global Temperature

The mean global temperature from 1861 to 1996 is listed in Table 4. The data, obtained from http://www.cgd.ucar.edu/stats/D
, was converted to mean temperature in degrees Celsius.

Year	Temperature	Year	Temperature	Year	Temperature	Year	Temperature
1861	12.66	1901	12.871	1941	13.152	1981	13.228
1862	12.58	1902	12.726	1942	13.147	1982	13.145
1863	12.799	1903	12.647	1943	13.156	1983	13.332
1864	12.619	1904	12.601	1944	13.31	1984	13.107
1865	12.825	1905	12.719	1945	13.153	1985	13.09
1866	12.881	1906	12.79	1946	13.015	1986	13.183
1867	12.781	1907	12.594	1947	13.006	1987	13.323
1868	12.853	1908	12.575	1948	13.015	1988	13.34
1869	12.787	1909	12.596	1949	13.005	1989	13.269
1870	12.752	1910	12.635	1950	12.898	1990	13.437
continued on next page							

[1] http://www.cgd.ucar.edu/stats/Data/Climate/

1871	12.733	1911	12.611	1951	13.044	1991	13.385
1872	12.857	1912	12.678	1952	13.113	1992	13.237
1873	12.802	1913	12.671	1953	13.192	1993	13.28
1874	12.68	1914	12.85	1954	12.944	1994	13.355
1875	12.669	1915	12.962	1955	12.935	1995	13.483
1876	12.687	1916	12.727	1956	12.836	1996	13.314
1877	12.957	1917	12.584	1957	13.139		
1878	13.092	1918	12.7	1958	13.208		
1879	12.796	1919	12.792	1959	13.133		
1880	12.811	1920	12.857	1960	13.094		
1881	12.845	1921	12.902	1961	13.124		
1882	12.864	1922	12.787	1962	13.129		
1883	12.783	1923	12.821	1963	13.16		
1884	12.73	1924	12.764	1964	12.868		
1885	12.754	1925	12.868	1965	12.935		
1886	12.826	1926	13.014	1966	13.035		
1887	12.723	1927	12.904	1967	13.031		
1888	12.783	1928	12.871	1968	13.004		
1889	12.922	1929	12.718	1969	13.117		
1890	12.703	1930	12.964	1970	13.064		
1891	12.767	1931	13.041	1971	12.903		
1892	12.671	1932	12.992	1972	13.031		
1893	12.631	1933	12.857	1973	13.175		
1894	12.709	1934	12.982	1974	12.912		
1895	12.728	1935	12.943	1975	12.975		
1896	12.93	1936	12.993	1976	12.869		
1897	12.936	1937	13.092	1977	13.148		
1898	12.759	1938	13.187	1978	13.057		

continued on next page

1899	12.874	1939	13.111	1979	13.154		
1900	12.959	1940	13.055	1980	13.195		

Table 4: Global temperature changes over the past 135 years. There has been a lot of discussion regarding changing weather patterns and a possible link to pollution and greenhouse gasses.

3.5 Data Set 5: Price of Petrol

The price of petrol in South Africa from August 1998 to July 2000 is shown in Table 5.

Date	Price (R/l)
August 1998	R 2.37
September 1998	R 2.38
October 1998	R 2.35
November 1998	R 2.29
December 1998	R 2.31
January 1999	R 2.25
February 1999	R 2.22
March 1999	R 2.25
April 1999	R 2.31
May 1999	R 2.49
June 1999	R 2.61
July 1999	R 2.61
August 1999	R 2.62
September 1999	R 2.75
October 1999	R 2.81
November 1999	R 2.86
December 1999	R 2.85
January 2000	R 2.86
February 2000	R 2.81
March 2000	R 2.89
April 2000	R 3.03
May 2000	R 3.18
June 2000	R 3.22
July 2000	R 3.36

Table 5: Petrol prices

4 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")

Exercise 1: Grouping Data ***(Solution on p. 25.)***

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

4.1 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 llll 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 6

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

Table 7

Click here for the solution[2]

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 8

Draw up a grouped frequency table using intervals 11-20, 21-30, 31-40, etc.
Click here for the solution[3]

[2] http://www.fhsst.org/l4k
[3] http://www.fhsst.org/l40

5 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 how the data values are spread and about the central values in the data set. Other methods of summarising and representing data will be covered in grade 11.

5.1 Measures of Central Tendency

5.1.1 Mean or Average

The mean, (also known as arithmetic mean), is simply the arithmetic average of a group of numbers (or data set) and is shown using the bar symbol . So the mean of the variable x is $\overline{x}$ pronounced "x-bar". The mean of a set of values is calculated by adding up all the values in the set and dividing by the number of items in that set. The mean is calculated from the raw, ungrouped data.

Definition 2: Mean
The mean of a data set, x, denoted by $\overline{x}$, is the average of the data values, and is calculated as:

$$\overline{x} = \frac{\text{sum of all values}}{\text{number of all values}} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \tag{1}$$

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 2: Mean *(Solution on p. 25.)*
What is the mean of $x = \{10, 20, 30, 40, 50\}$?

5.1.2 Median

Definition 3: 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

1. Order the data from smallest to largest or from largest to smallest.
2. Count how many data values there are in the data set.
3. Find the data value in the central position of the set.

Exercise 3: Median *(Solution on p. 25.)*
What is the median of $\{10, 14, 86, 2, 68, 99, 1\}$?

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.

Exercise 4: Median ***(Solution on p. 25.)***

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

5.1.3 Mode

Definition 4: 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.

Exercise 5: Mode ***(Solution on p. 25.)***

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

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.

Khan academy video on statistics

This media object is a Flash object. Please view or download it at
<http://www.youtube.com/v/uhxtUt_-GyM&rel=0>

Figure 3

5.2 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, percentiles and quartiles.

5.2.1 Range

Definition 5: 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

1. Find the highest value in the data set.
2. Find the lowest value in the data set.
3. Subtract the lowest value from the highest value. The difference is the range.

Exercise 6: Range *(Solution on p. 26.)*
Find the range of the data set $x = \{1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 8, 9, 10, 10\}$

5.2.2 Quartiles

Definition 6: 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
		$\downarrow$			$\downarrow$			$\downarrow$		
		Lower quartile			Median			Upper quartile		
		(Q_1)			(Q_2)			(Q_3)		

Table 9

Method: Calculating the quartiles

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

Exercise 7: Quartiles *(Solution on p. 26.)*
What are the quartiles of $\{3, 5, 1, 8, 9, 12, 25, 28, 24, 30, 41, 50\}$?

5.2.3 Inter-quartile Range

Definition 7: 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. $Q_3 - Q_1$.

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

Exercise 8: Medians, Quartiles and the Interquartile Range *(Solution on p. 26.)*
A class of 12 students writes a test and the results are as follows: 20, 39, 40, 43, 43, 46, 53, 58, 63, 70, 75, 91. Find the range, quartiles and the Interquartile Range.

5.2.4 Percentiles

Definition 8: 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

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

5.3 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.

5.4 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 4. The box can lie horizontally (as shown) or vertically. For a horizonatal 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.

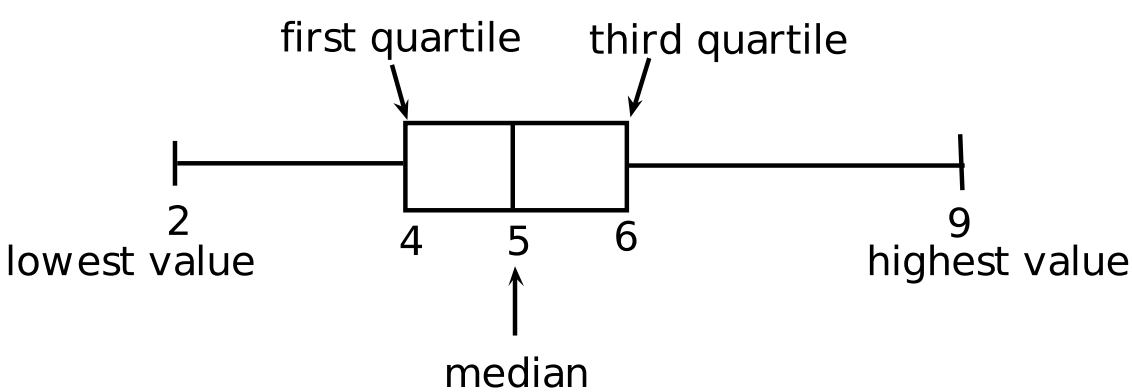


Figure 4: Main features of a box and whisker plot

Exercise 9 *(Solution on p. 27.)*

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

5.5 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

   Click here for the solution[4]
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?

   Click here for the solution[5]
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.
   Click here for the solution[6]
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?
   Click here for the solution[7]
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.
   Click here for the solution[8]
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.
   Click here for the solution[9]
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?

   Click here for the solution[10]

[4] http://www.fhsst.org/l48
[5] http://www.fhsst.org/l49
[6] http://www.fhsst.org/l4X
[7] http://www.fhsst.org/l4I
[8] http://www.fhsst.org/l2Q
[9] http://www.fhsst.org/l2U
[10] http://www.fhsst.org/l2P

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.

Exercise 10: Mean, Median and Mode for Grouped Data ***(Solution on p. 29.)***

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 10

5.5.1 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 11

Click here for the solution[11]

[11] http://www.fhsst.org/l45

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 12

Click here for the solution[12]

6 Bias and error in measurements

All measurements have some error associated with them. Random errors occur in all data sets and are sometimes known as non-systematic errors. Random errors can arise from estimation of data values, imprecision of instruments, etc. For example if you are reading lengths off a ruler, random errors will arise in each measurement as a result of estimating between which two lines the length lies. Bias is also sometimes known as systematic error. Bias in a data set is where a value is consistently under or overestimated. Bias can arise from forgetting to take into account a correction factor or from instruments that are not properly calibrated (calibration is the process of marking off predefined measurements). Bias leads to a sample mean that is either lower or higher than the true mean.

7 Data interpretation

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.

[12] http://www.fhsst.org/l4N

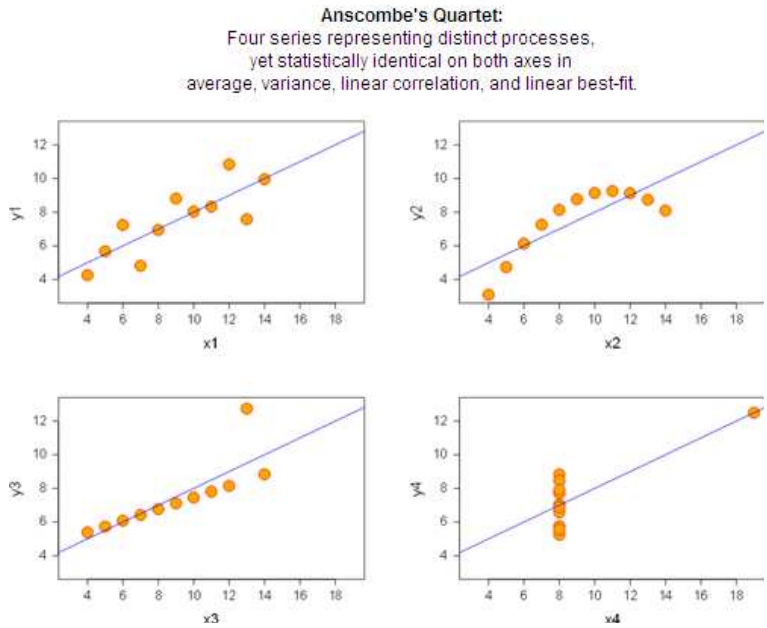


Figure 5: Anscombe's quartet

8 Misuse of Statistics - For enrichment, not in CAPS

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.

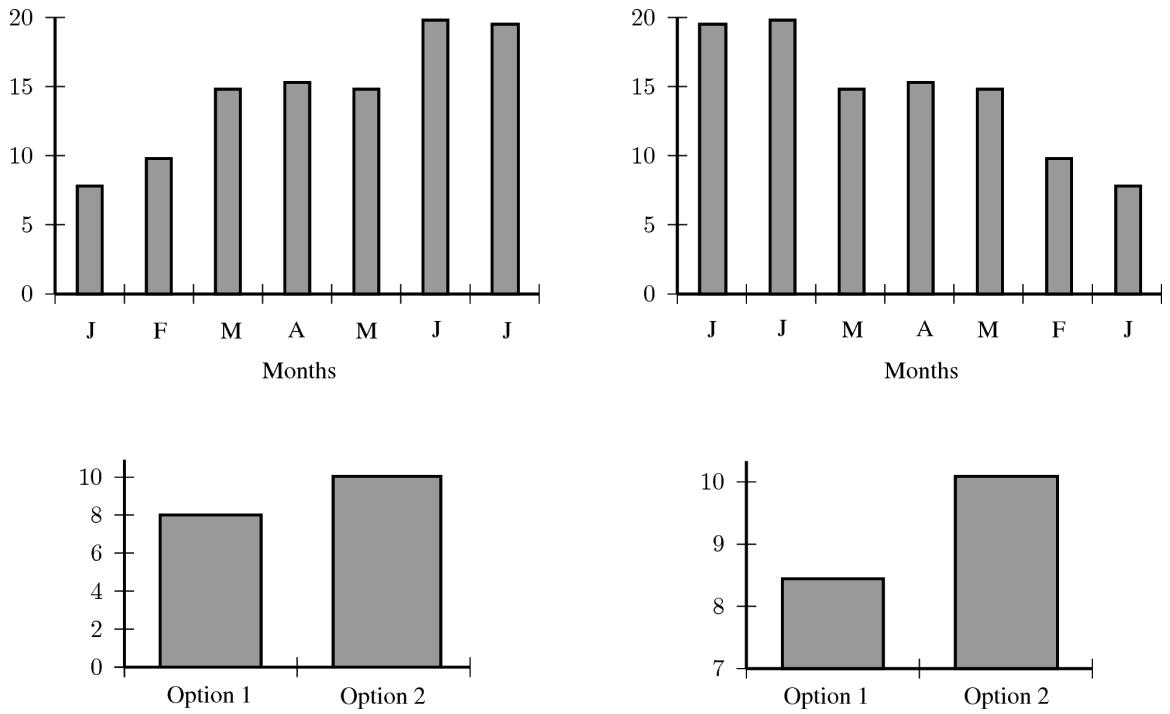


Figure 6

8.1 Exercises - Misuse 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.

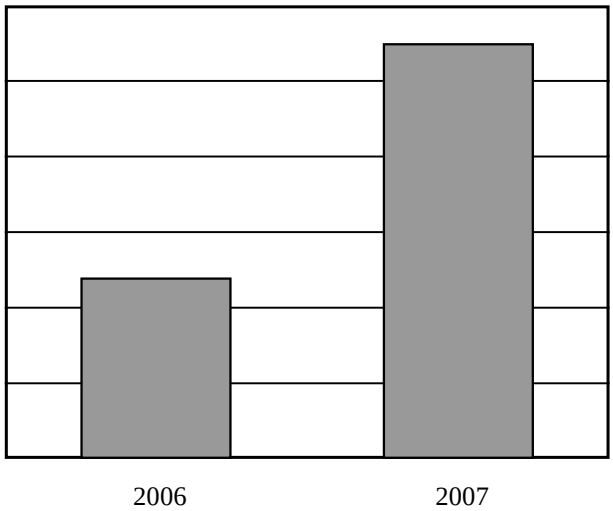


Figure 7

Click here for the solution[13]

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.

   Click here for the solution[14]
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.

[13] http://www.fhsst.org/l4R
[14] http://www.fhsst.org/l4n

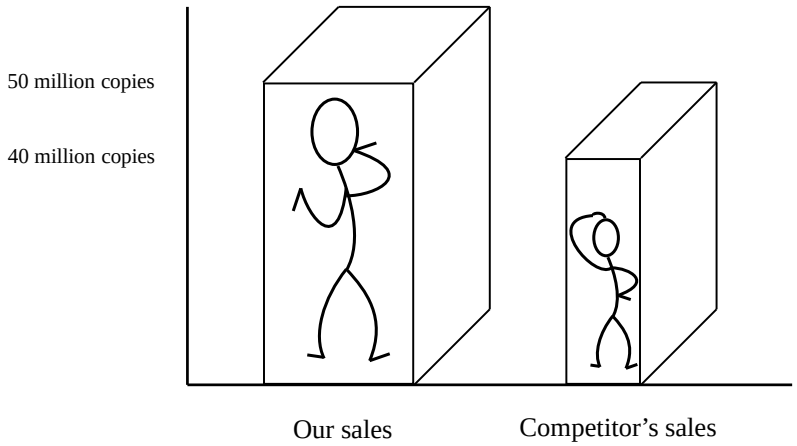


Figure 8

Click here for the solution[15]

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?

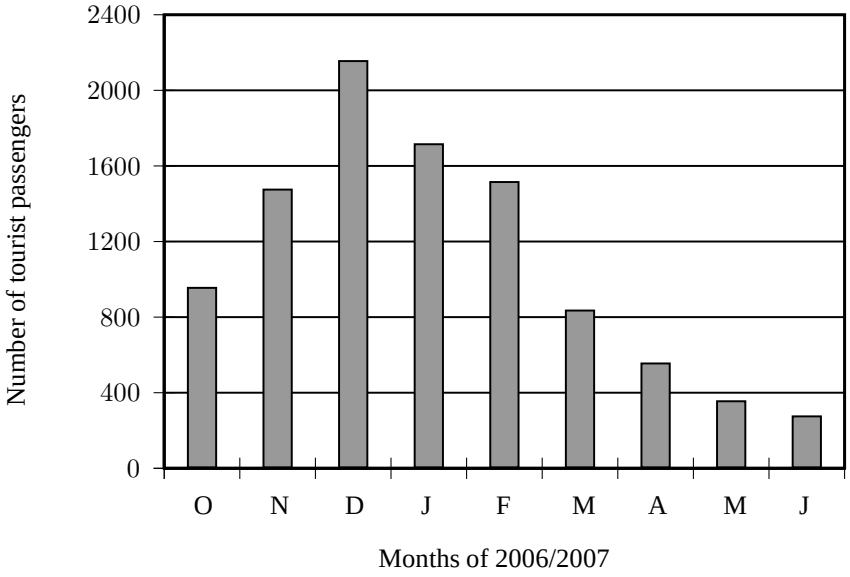


Figure 9

Click here for the solution[16]

[15] http://www.fhsst.org/l4Q
[16] http://www.fhsst.org/l4U

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.

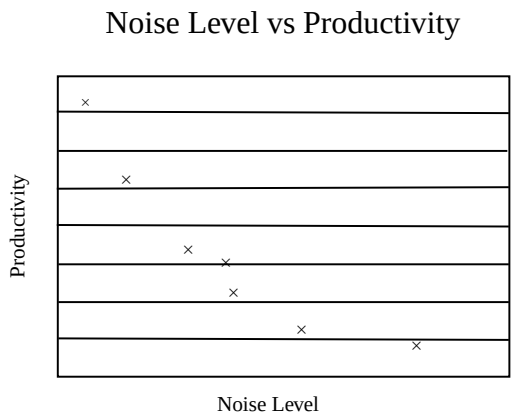


Figure 10

The following statement was then made: "If an office environment is noisy, this leads to poor productivity." Explain the flaws in this thinking.
Click here for the solution[17]

9 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:
  - The mean of a data set, x, denoted by $\overline{x}$, is the average of the data values, and is calculated as:

$$\overline{x} = \frac{\text{sum of values}}{\text{number of values}} \tag{2}$$

  - 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. $Q_3 - Q_1$. Half of this value is the semi-interquartile range.

[17] http://cnx.org/content/m38371/latest/ http://www.fhsst.org/l4P

- 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

10 Exercises

1. Calculate the mean, median, and mode of Data Set 3.
   Click here for the solution[18]
2. The tallest 7 trees in a park have heights in metres of 41, 60, 47, 42, 44, 42, and 47. Find the median of their heights.
   Click here for the solution[19]
3. The students in Bjorn's class have the following ages: 5, 6, 7, 5, 4, 6, 6, 6, 7, 4. Find the mode of their ages.
   Click here for the solution[20]
4. An engineering company has designed two different types of engines for motorbikes. The two different motorbikes are tested for the time it takes (in seconds) for them to accelerate from 0 km/h to 60 km/h.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Average
Bike 1	1.55	1.00	0.92	0.80	1.49	0.71	1.06	0.68	0.87	1.09	
Bike 2	0.9	1.0	1.1	1.0	1.0	0.9	0.9	1.0	0.9	1.1	

Table 13

   a. What measure of central tendency should be used for this information?
   b. Calculate the average you chose in the previous question for each motorbike.
   c. Which motorbike would you choose based on this information? Take note of accuracy of the numbers from each set of tests.

   Click here for the solution[21]
5. The heights of 40 learners are given below.

154	140	145	159	150	132	149	150	138	152
141	132	169	173	139	161	163	156	157	171
168	166	151	152	132	142	170	162	146	152
142	150	161	138	170	131	145	146	147	160

Table 14

[18] http://www.fhsst.org/l4E
[19] http://www.fhsst.org/l4m
[20] http://www.fhsst.org/l4y
[21] http://www.fhsst.org/l4p

a. Set up a frequency table using 6 intervals.
b. Calculate the approximate mean.
c. Determine the mode.
d. How many learners are taller than your approximate average in (b)?

Click here for the solution[22]

6. In a traffic survey, a random sample of 50 motorists were asked the distance they drove to work daily. This information is shown in the table below.

Distance in km	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
Frequency	4	5	9	10	7	8	3	2	2

Table 15

a. Find the approximate mean.
b. What percentage of samples drove
   i. less than 16 km?
   ii. more than 30 km?
   iii. between 16 km and 30 km daily?

Click here for the solution[23]

7. A company wanted to evaluate the training programme in its factory. They gave the same task to trained and untrained employees and timed each one in seconds.

Trained	121	137	131	135	130
	128	130	126	132	127
	129	120	118	125	134
Untrained	135	142	126	148	145
	156	152	153	149	145
	144	134	139	140	142

Table 16

a. Find the medians and quartiles for both sets of data.
b. Find the Interquartile Range for both sets of data.
c. Comment on the results.

Click here for the solution[24]

8. A small firm employs nine people. The annual salaries of the employers are:

R600 000	R250 000	R200 000
R120 000	R100 000	R100 000
R100 000	R90 000	R80 000

Table 17

[22] http://www.fhsst.org/l4d
[23] http://www.fhsst.org/l4v
[24] http://www.fhsst.org/l4w

a. Find the mean of these salaries.
b. Find the mode.
c. Find the median.
d. Of these three figures, which would you use for negotiating salary increases if you were a trade union official? Why?

Click here for the solution[25]

9. The marks for a particular class test are listed here:

67	58	91	67	58	82	71	51	60	84
31	67	96	64	78	71	87	78	89	38
69	62	60	73	60	87	71	49		

Table 18

Complete the frequency table using the given class intervals.

Class	Tally	Frequency	Mid-point	Freq × Midpt
30-39		34,5		
40-49		44,5		
50-59				
60-69				
70-79				
80-89				
90-99				
		Sum =		Sum =

Table 19

Click here for the solution[26]

[25] http://www.fhsst.org/l4f
[26] http://www.fhsst.org/l4G

Solutions to Exercises in this Module

Solution to Exercise (p. 9)

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 20

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

Solution to Exercise (p. 10)

Step 1.

$$10 + 20 + 30 + 40 + 50 = 150 \tag{3}$$

Step 2. There are 5 values in the data set.
Step 3.

$$150 \div 5 = 30 \tag{4}$$

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

Solution to Exercise (p. 10)

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 $\{1, 2, 10, 14, 68, 86, 99\}$.

Solution to Exercise (p. 11)

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

$$(11 + 14) \div 2 = 12,5 \tag{5}$$

Solution to Exercise (p. 11)

Step 1.

data value	frequency	data value	frequency
1	1	6	1
2	1	7	1
3	1	8	2
4	3	9	1
5	1	10	2

Table 21

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.

Solution to Exercise (p. 12)

Step 1. 10 is the highest value and 1 is the lowest value.
Step 2.

$$10 - 1 = 9 \tag{6}$$

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.

Solution to Exercise (p. 12)

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 3.

$$12 \div 4 = 3 \tag{7}$$

Step 4.

1	3	5	$\|$	8	9	12	$\|$	24	25	28	$\|$	30	41	50
			Q_1				Q_2				Q_3			

Table 22

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)

Solution to Exercise (p. 12)

Step 1.

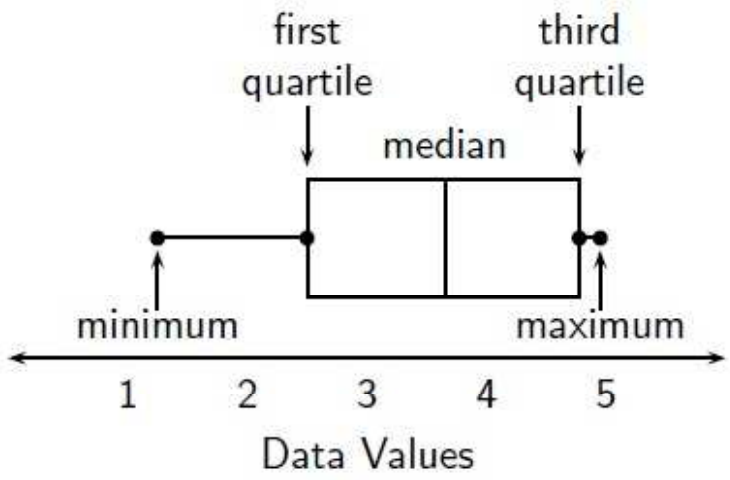
20	39	40	$\|$	43	43	46	$\|$	53	58	63	$\|$	70	75	91
			Q_1				M				Q_3			

Table 23

Step 2. The range = 91 - 20 = 71. This tells us that the marks are quite widely spread. (Remember, however, that 'wide' and 'large' are relative terms. If you are considering one hundred people, a range of 71 would be 'large', but if you are considering one million people, a range of 71 would likely be 'small', depending, of course, on what you were analyzing).
Step 3. i.e. $M = \frac{46+53}{2} = \frac{99}{2} = 49,5$
Step 4. i.e. $Q_1 = \frac{40+43}{2} = \frac{83}{2} = 41,5$
Step 5. i.e. $Q_3 = \frac{63+70}{2} = \frac{133}{2} = 66,5$
Step 6. The quartiles are 41,5, 49,5 and 66,5. These quartiles tell us that 25% of the marks are less than 41,5; 50% of the marks are less than 49,5 and 75% of the marks are less than 66,5. They also tell us that 50% of the marks lie between 41,5 and 66,5.
Step 7. The Interquartile Range = 66,5 - 41,5 = 25. This tells us that the width of the middle 50% of the data values is 25.
Step 8. The Semi-interquartile Range = $\frac{25}{2}$ = 12,5

Solution to Exercise (p. 13)

Step 1. Minimum $= 1,25$
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.

Figure 11

Solution to Exercise (p. 15)

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 24

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.