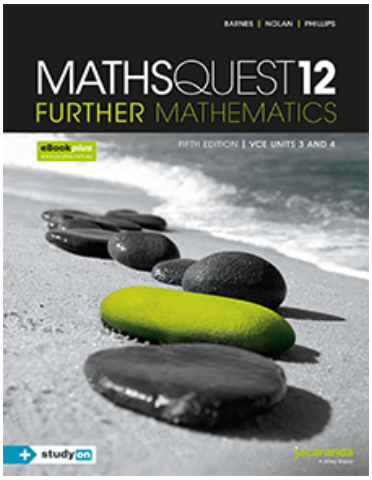
Point Cook Senior



Further Mathematics 2016
Core: DATA ANALYSIS
Chapter 1 - Univariate Data

Extract from Study Design

Key knowledge

- Types of data: categorical (nominal and ordinal) and numerical (discrete and continuous)
- Frequency tables, bar charts including segmented barcharts, histograms, stem plots, dot plots, and their application in the context of displaying and describing distributions
- log (base 10) scales, and their purpose and application
- Five-number summary and boxplots (including the designation and display of possible outliers)
- Mean $\bar{x}$ and standard deviation s_x
- Normal model and the 68–95–99.7% rule, and standardised values (*z*-scores)

Key skills

- Construct frequency tables and bar charts and use them to describe and interpret the distributions of categorical variables
- Answer statistical questions that require a knowledge of the distribution/s of one or more categorical variables
- Construct stem and dot plots, boxplots, histograms and appropriate summary statistics and use them to describe and interpret the distributions of numerical variables
- Answer statistical questions that require a knowledge of the distribution/s of one or more numerical variables
- Solve problems using the *z*-scores and the 68–95–99.7% rule

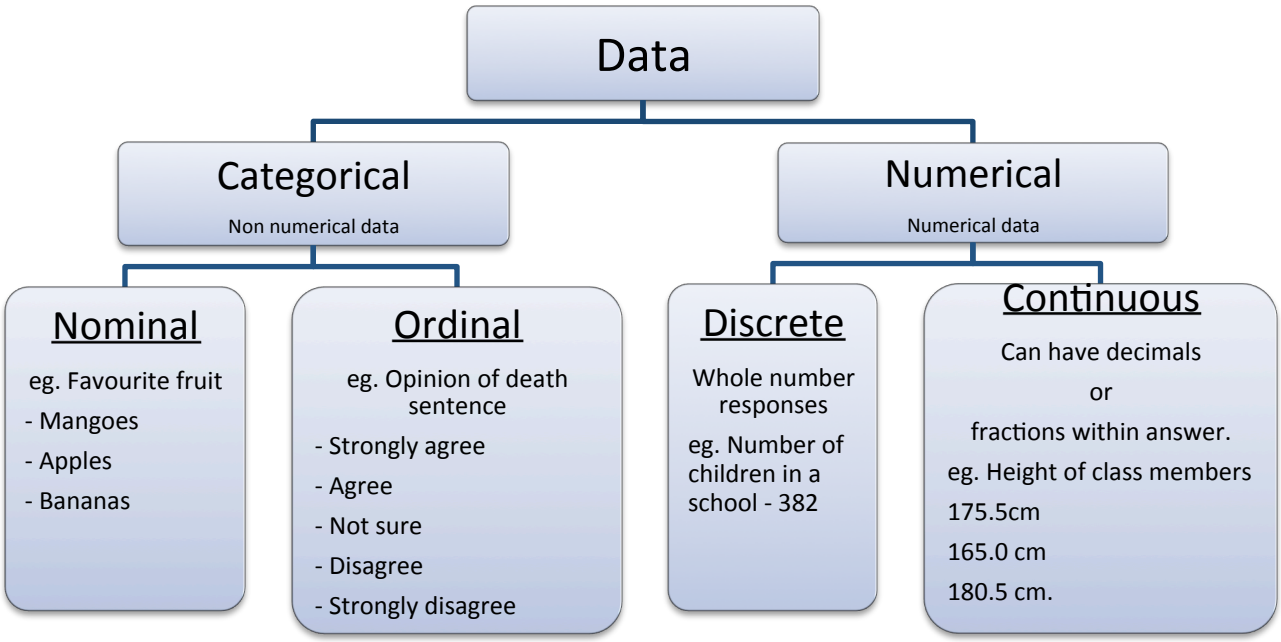
Chapter Sections	Questions to be completed
1.2 Types of data	10, 11, 12, 13, 14, 16
1.3 Stem plots	1, 10, 12, 13, 17
1.4 Dot plots, frequency tables and histograms and bar charts	1, 5, 7, 9, 10, 11, 12
1.5 Describing the shape of stem plots and histograms	6, 7, 9, 11, 12
1.6 The median, the interquartile range, the range and the mode	10, 11, 13, 15, 17
1.7 Boxplots	9, 10, 12, 14, 18, 19
1.8 The mean of a sample	2, 5, 6, 7, 14, 16
1.9 Standard deviation of a sample	6, 9, 14
1.10 Populations and simple random samples	2, 4, 7, 8, 9, 10, 11
1.11 The 68-95-99.7% rule and z-scores	2, 4, 6, 8, 10, 11, 14, 15, 16, 17, 18, 20, 21

More resources available at

http://pcsfurthermaths.weebly.com

Table of Contents

1.2 Types of Data



Numerical Data – data involving quantities which are *measurable* or *countable*.
Eg height, test marks, ages salaries.

Categorical Data – data which are divided into categories or groups.
Eg genders (sexes), football teams, finish positions in a race (1^{st} 2^{nd} 3^{rd} 4^{th} etc), ratings 1-5, age groups (1-9, 10-19, 20-29 etc), and hair colour.

Discrete Data – fixed values whole numbers.
Eg number of children in a house, number of matches in a match box.

Continuous Data – value between 2 values (any value within a range) .
Eg heights of students in a school (60.3cm), weight, age, length and time (3hr 26nim 2 sec)

Different ways to display data include:

- Stem plots
- Frequency histograms
- Bar charts
- Segmented bar charts
- Dot plots

Worked Example 1

Which of the following is *not* numerical data?

A. Maths test results
B. Ages
C. AFL football teams
D. Heights of students in a class
E. Lengths of bacterium

Worked Example 2

Which of the following is *not* discrete data?

A. Number of students older than 17.5 years old
B. Number of girls in a class
C. Number of questions correct in a multiple choice test
D. Number of students above 180 cm in a class
E. Height of the tallest student in a class

1.3 Stem Plots (stem-and-leaf plot)

A *stem-and-leaf plot*, or **stem plot** for short, is a way of ordering and displaying a set of data, with the advantage that all of the raw data is kept. Since all the individual values of the data are being listed, it is only suitable for smaller data sets (up to about 50 observations).

A stem plot is constructed by splitting the numerals of a record into two parts – the stem, which in this case is the first digit, and the leaf, which is always the last digit.

- Used for up to 50 data points
- Consists of a stem and a leaf
- Last digit is always the leaf eg. 501, 512, 511

stem	leaf
50	1
51	1 2

- Must be in order from lowest to highest
- Must have a key
- If bunched, break stem into halves or fifths (i.e. if Leaf is too big

Example

Plot the following data in a stem and leaf plot: (a) Halves, (b) Fifths

50 51 53 53 54 55 55 56 56 57 59

(a) Halves

Stem	Leaf	
5	0 1 3 3 4	←0-4
5*	5 5 6 6 7 9	←5-9

(b) Fifths

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

Example

For the month of August, a small country hospital recorded the birth weights of babies (in kilograms).

3.1, 3.3, 4.2, 3.5, 4.1, 3.6, 3.5, 3.3, 4.6, 3.9, 4.0, 3.7, 3.8, 3.0, 3.8

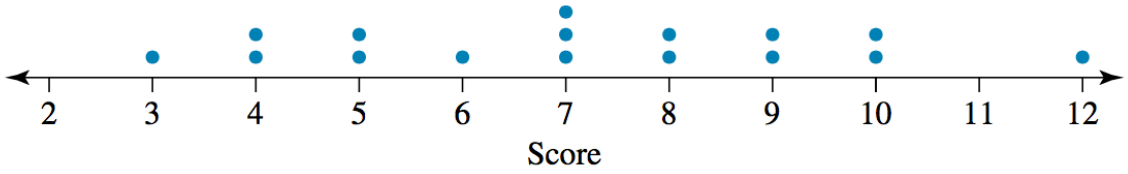
Produce a stem plot for the above data by splitting the data into:

a) halves b) fifths

1.4 Dot Plots, Frequency Tables and Histograms and Bar Charts

Dot Plots

- Display discrete and categorical data
- Single dot is used to represent each data value



Worked Example 6
The number of hours per week spent on art by 18 students is given below. Display the data on a dot plot.

4 0 3 1 3 4 2 2 3
4 1 3 2 5 3 2 1 0

The mode

One of the features of a data set, revealed by dot plots; frequency tables; histograms and bar charts is the **mode** or **modal** category. The mode is the most frequently occurring value or category.
In dot plots it is the vale or category with the most dots (highest number). We will see that in frequency tables and histograms it is the value that occurs to most, and in bar charts it is the highest or longest bar.

Frequency Tables and Histogram

- Used for large sets of data (over 50) but can be used for smaller sets
- Should construct a frequency table first
- Frequency on *y*-axis / Class interval on *x*-axis

Guidelines for commenting on Dot plots/Frequency Tables/Histograms and Bar charts:

- Summarise the context in which the data was collected, including the number of data values
- If there is a clear mode, ensure that it is mentioned
- Included numbers or frequencies or percentages in the comment
- If there are lots of categories it is not necessary to mention each one, but the modal category should be mentioned.

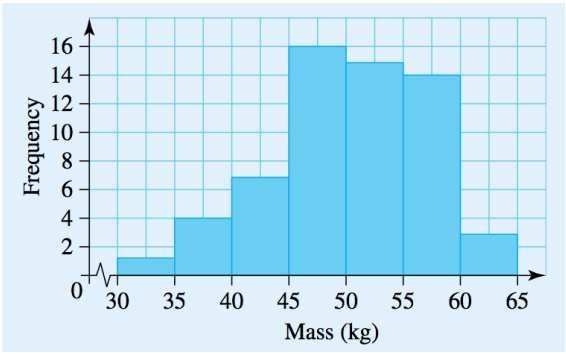
Worked Example 7

The data below show the distribution of masses (in kilograms) of 60 students in Year 7 at Northwood Secondary College. Construct a frequency histogram to display the data and **Use the information** in the histogram to comment on the data.

45.7 45.8 45.9 48.2 48.3 48.4 34.2 52.4 52.3 51.8 45.7 56.8 56.3 60.2 44.2 53.8 43.5 57.2 38.7 48.5 49.6 56.9 43.8 58.3 52.4 54.3 48.6 53.7 58.7 57.6 45.7 39.8 42.5 42.9 59.2 53.2 48.2 36.2 47.2 46.7 58.7 53.1 52.1 54.3 51.3 51.9 54.6 58.7 58.7 39.7 43.1 56.2 43.0 56.3 62.3 46.3 52.4 61.2 48.2 58.3

1. First construct a frequency table Determine the lowest and highest value.
2. Divide the data into class intervals, usually between about 5 and 15
3. Put one mark for each value against the appropriate interval
4. Add up the tally marks and put it in the frequency column
5. Construct a histogram from the table

Class interval	Tally	Frequency
30–	|	1
35–	||||	4
40–	𝍸 ||	7
45–	𝍸 𝍸 𝍸 |	16
50–	𝍸 𝍸 𝍸	15
55–	𝍸 𝍸 ||||	14
60–	|||	3
	Total	60



Comment:

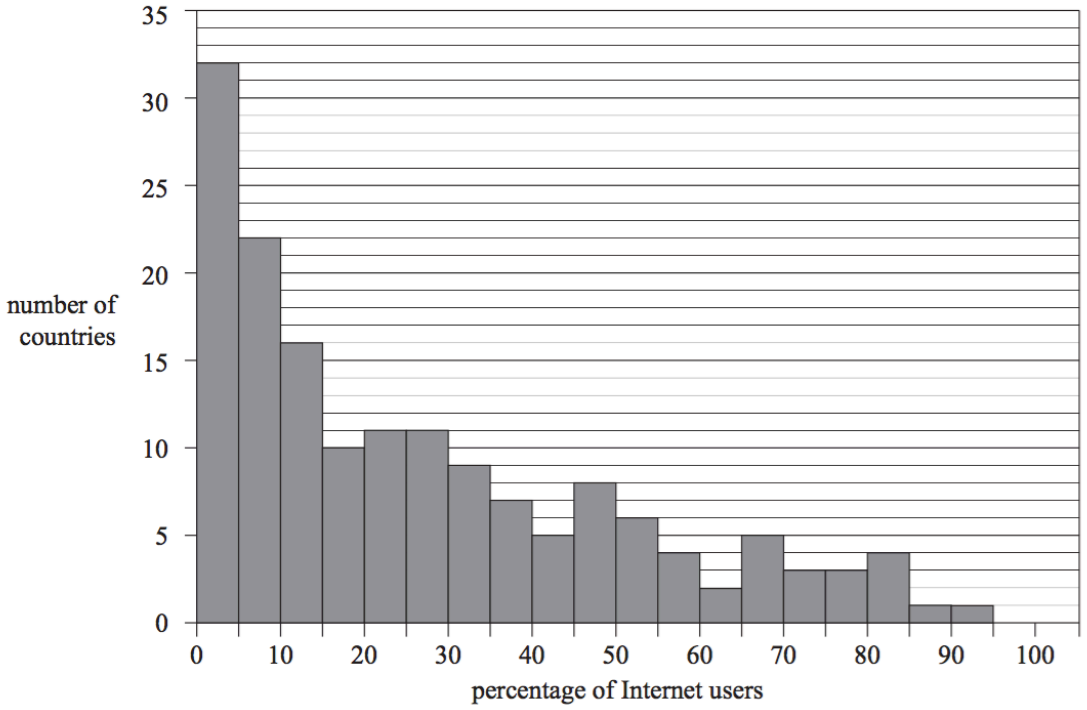
When 60 students in year 7 at Northwood SC were weighed their weights were in the range 30 to 65kg. The modal category was 45-49.9kg class group* with a frequency of 16, closely followed by the 50- and 55- kg class groups. From the histogram it is clear that the majority of the students were in the 45 to 60kg class intervals, with only 15 of the 60 students falling outside of these intervals.

*Note: this is grouped data so we need to talk about groups/intervals we have chosen.

PAST EXAM QUESTION (Exam 1 2011)

Use the following information to answer Questions 1, 2 and 3.

The histogram below displays the distribution of the percentage of Internet users in 160 countries in 2007.



Based on data obtained from: www.data.un.org

Question 1

The shape of the histogram is best described as

A. approximately symmetric.
B. bell shaped.
C. positively skewed.
D. negatively skewed.
E. bi-modal.

Question 2

The number of countries in which less than 10% of people are Internet users is closest to

A. 10
B. 16
C. 22
D. 32
E. 54

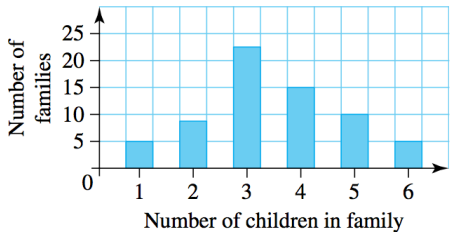
Question 3

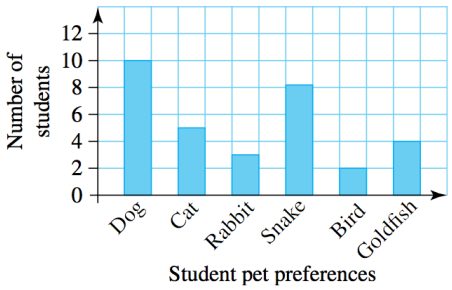
From the histogram, the median percentage of Internet users is closest to

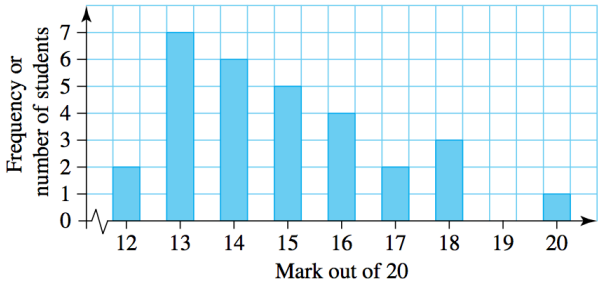
A. 10%
B. 15%
C. 20%
D. 30%
E. 40%

Bar chart

- Usually used to display categorical data
- Horizontal or vertical bars that are separated by small spaces







Segmented bar charts

A **segmented bar chart** is a single bar, which is used to represent all the data being studied. It is divided into segments, each segment representing a particular group of the data. Generally, the information is presented as percentages and so the total bar length represents 100% of the data.

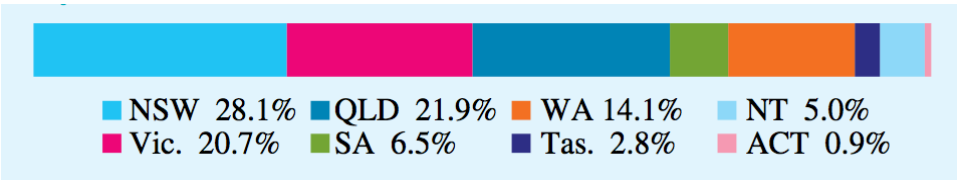
Worked Example 9 Segmented percentage bar chart

The table shown represents fatal road accidents in Australia. Construct a segmented bar chart to represent this data.

	Accidents involving fatalities								
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	Aust.
2008	376	278	293	87	189	38	67	12	**1340**

1. To draw a segmented bar chart the data needs to be converted to percentages.
2. To draw the segmented bar chart to scale decide on its overall length, let's say 100 mm.
3. Therefore NSW = 28.1%, represented by 28.1 mm. Vic = 20.7%, represented by 20.7 mm and so on.
4. Draw the answer and colour code it to represent each of the states and territories.

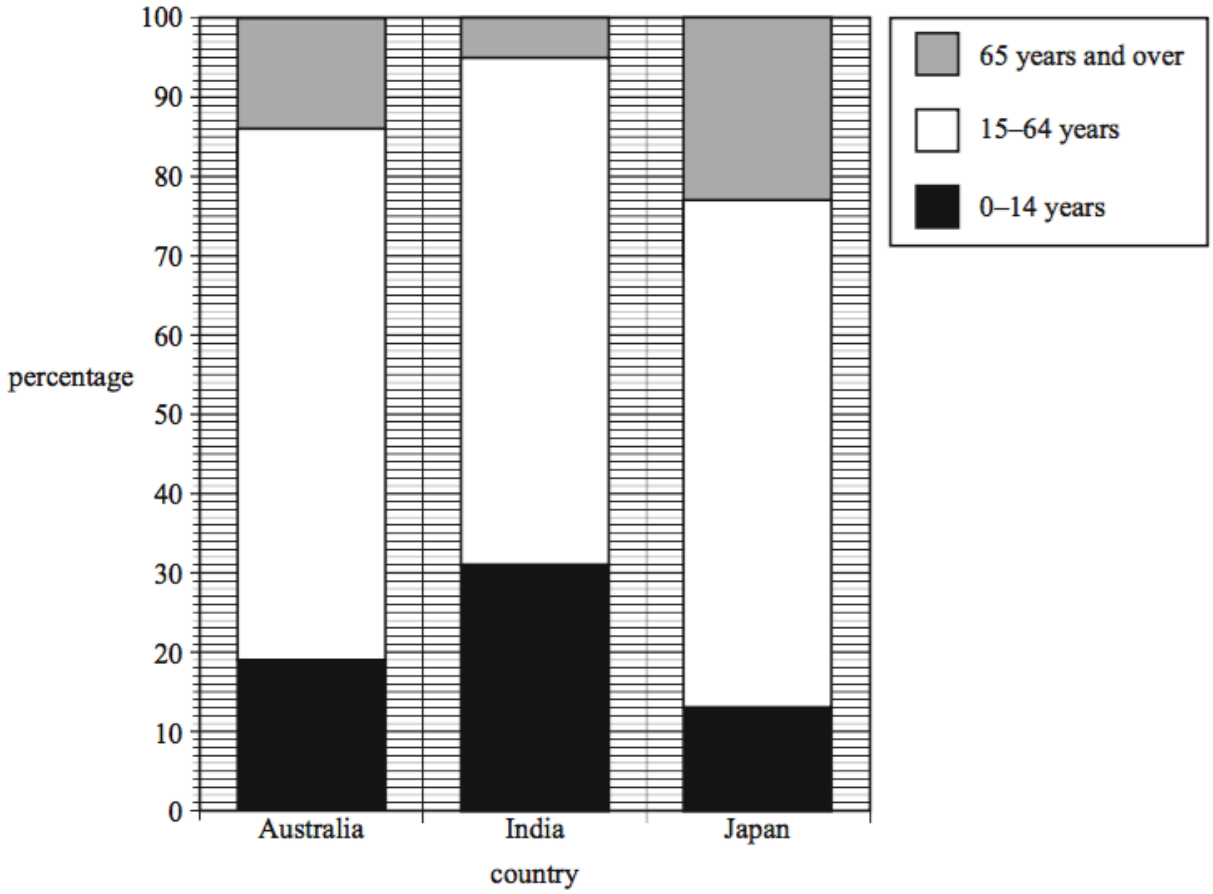
State	Number of accidents	Percentage
NSW	376	376 ÷ 1340 × 100% = 28.1%
Vic.	278	278 ÷ 1340 × 100% = 20.7%
Qld	293	293 ÷ 1340 × 100% = 21.9%
SA	87	87 ÷ 1340 × 100% = 6.5%
WA	189	189 ÷ 1340 × 100% = 14.1%
Tas.	38	38 ÷ 1340 × 100% = 2.8%
NT	67	67 ÷ 1340 × 100% = 5.0%
ACT	14	12 ÷ 1340 × 100% = 0.9%



PAST EXAM QUESTION (Exam 2 2014)

Question 1 (3 marks)

The segmented bar chart below shows the age distribution of people in three countries, Australia, India and Japan, for the year 2010.



Source: Australian Bureau of Statistics, *3201.0 – Population by Age and Sex, Australian States and Territories*, June 2010

a. Write down the percentage of people in Australia who were aged 0–14 years in 2010.
Write your answer, correct to the nearest percentage. 1 mark

b. In 2010, the population of Japan was 128 000 000.

How many people in Japan were aged 65 years and over in 2010? 1 mark

c. From the graph above, it appears that there is no association between the percentage of people in the 15–64 age group and the country in which they live.

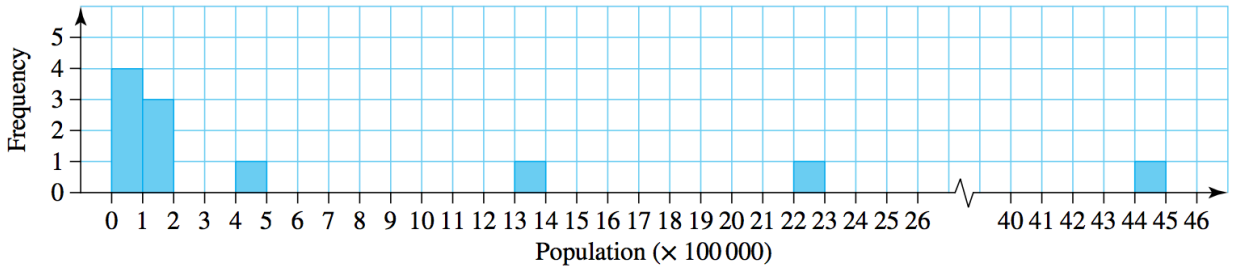
Explain why, quoting appropriate percentages to support your explanation. 1 mark

Using a log (base 10) scale

Sometimes a data set will contain data points that vary so much in size that plotting them using a traditional scale becomes very difficult.
For example: If we are studying the population of different cities in Australia we might end up with the following data points:

City	Population
Adelaide	1 304 631
Ballarat	98 543
Brisbane	2 274 460
Cairns	146 778
Darwin	140 400
Geelong	184 182
Launceston	86 393
Melbourne	4 440 328
Newcastle	430 755
Shepparton	49 079
Wagga Wagga	55 364

A histogram splitting the data into class intervals of 100 000 would then appear as follows:



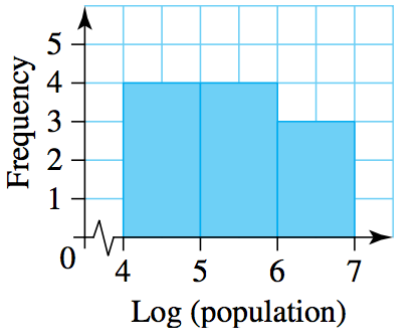
A way to overcome this is to write the numbers in a **logarithmic (log) form.** The log of a number is the power of 10 which creates this number.

$\log_{10}(10) = \log(10^1) = 1$
$\log_{10}(100) = \log(10^2) = 2$
$\log_{10}(1000) = \log(10^3) = 3$

$\log(10^n) = n$

Note: when log is written without the subscript 10 it always refers to $\log_{10}$

Applying a logarithm scale to the above graph transforms to:



Examples of where the logarithmic scale is used in real life: Richter scale measuring strength of an earthquake and sound or noise decibels

Worked Example 10

The following table shows the average weights of 10 different adult mammals. Display the data in a histogram using a log base 10 scale.

Mammal	Weight (kg)
African elephant	4800
Black rhinoceros	1100
Blue whale	136000
Giraffe	800
Gorilla	140
Humpback whale	30000
Lynx	23
Orang-utan	64
Polar bear	475
Tasmanian devil	7

1. Using CAS, calculate the logarithmic values of all of the weights, e.g.:

log(4800) = 3.68 (correct to 2 decimal places)

log(1100) = 3.04 (correct to 2 decimal places)

etc…

Note: The calculator needs to be in approx. mode to get the decimal results.

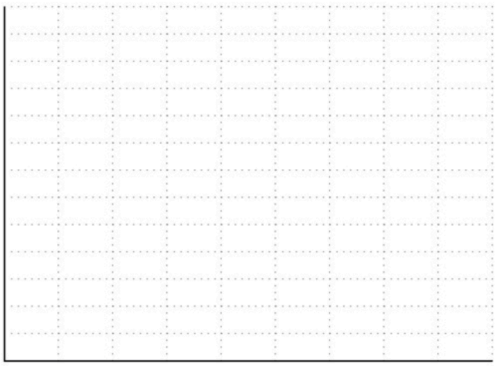
	A weight	B logweight	C
=		=log(a[])	
1	4800.	3.68124	
2	1100.	3.04139	
3	136000.	5.13354	
4	800.	2.90309	
5	140.	2.14613	
6	30000.	4.47712	
7	23.	1.36173	
8	64.	1.80618	
9	475.	2.67669	
10	7.	0.845098	
11			

=3.6812412373756

2. Group the logarithmic weights into class intervals and create a frequency table for the groupings.

Log (weight(kg))	Frequency
0 – 1	
1 – 2	
2 – 3	
3 – 4	
4 – 5	
5 - 6	

3. Construct a histogram of the data set.



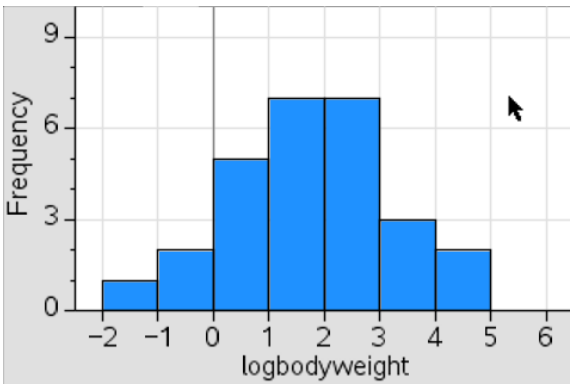
Interpreting log (base 10) values

If we are given values in logarithmic form, by raising 10 to the power of the logarithmic number we can determine the conventional number.

For example, the number 3467 in log (base 10) form is 3.54, and $10^{3.54} = 3467$. We can use this fact to compare values in log (base 10) form, as shown in the following example.

Example

The histogram shows the distribution of the weights of 27 animal species plotted on a log scale.



a) What bodyweight (in kg) is represented by the number 4 on the log scale?

b) How many of these animals have body weights more than 10 000 kg? More than 1 000 kg?

c) If the log(weight) of an elephant is 3.4, Determine the weight of an elephant (in kg) to the nearest whole number.

Worked Example 11

The Richter Scale measures the magnitude of earthquakes using a log (base 10) scale. How many times stronger is an earthquake of magnitude 7.4 than one of magnitude 5.2? Give your answer correct to the nearest whole number.

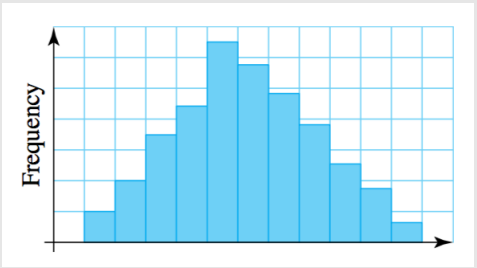
1.5 Describing the shape of stem plots and histograms and boxplots

The shape can be described as: Symmetric or positively skewed or negatively skewed

Symmetric distribution

- The **mean and median** will be similar
- Can use either to measure the centre

stem	leaf
0	1 3
1	2 4 7 9
2	3 3 8 9 9 9
3	0 1 6 8
4	0 2



Stem & Leaf and Histogram: Single peak, tails off on both sides

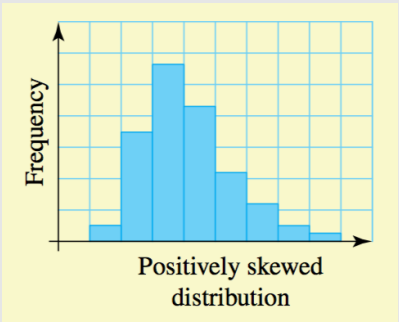
Boxplot: whiskers are about the same length, median close to the middle of the box.

Skewed distributions

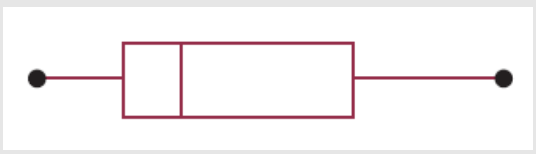
Positively skewed distribution

- The mean will be greater than the median
- The **median** will give a better indication of the centre.

stem	leaf
2	0 3 4 7 8
3	0 2 5 6 6 7 8 9 9
4	1 3 3 5 5 8 8
5	0 5 6 6 7
6	3 5 5
7	0



Stem & Leaf and Histogram: Peaks at the start, then tails to the right.

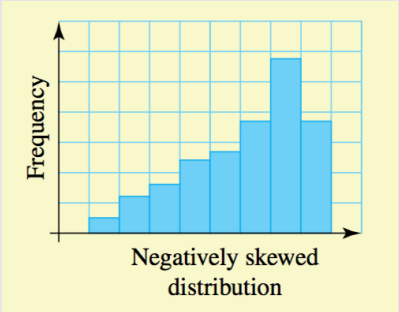


Boxplot: left whisker is shorter than the right, median towards left.

Negatively skewed distribution

- The mean will be less than the median
- The **median** will give a better indication of the centre.

stem	leaf
0	8
1	5 7 7
2	0 5 7 8 8
3	3 4 6 6 7 8 8
4	2 2 3 5 6 7 8 9
5	0 1 2 5 7 7 8



Stem & Leaf and Histogram: tails at the start, and peaks to the right.

Boxplot: left whisker larger than right, median towards the right.

1.6 Median/Interquartile Range/Range and the Mode

Summary statistics/Five number summary

Are the following values

- median
- interquartile range
- range
- mode

Note: the data must be ***ordered*** before they these values can be calculated. (that is they must be in increasing order)

Median (or Q_2)

Median is the **midpoint** (middle) of a set of data. This is easily found as the middle value for an odd number of data values, however, it is more difficult for an even number of values. In either case, the median can be located as the position.

$$median\ position = \left(\frac{n+1}{2}\right)$$

Where n = number of data points or values

Example: If we have 10 data points then the median is located at the

Worked Example 13

Consider the stem plot at below which contains 22 observations. What is the median?

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

Key: 3 | 4=34

Interquartile range (IQR)

The interquartile range is the middle 50% of the values in a set of data.

Consider, the following 16 numbers

1, 2, 3, 4, 5, 6, 7, 8 9, 10, 11, 12, 13, 14, 15, 16

Obviously, the median (Q_2) is between the 8 and 9 and therefore $Q_2 = 8.5$

If we now divide the bottom 8 numbers in half, we have the 'position' of Q_1 and if we divide the top 8 in half also we have the 'position' of Q_3. (In this case Q_1=4.5 and Q_3=12.5)

The difference between Q_3 and Q_1 is the interquartile range. Also, given by: **$IQR = Q_3 - Q_1$**

In summary, to locate Q_1 and Q_3 follow these steps:

1. Order data from lowest to highest.
2. Locate median.
3. Q_1 (lower quartile) is the middle number of the 1st half of data. (Don't include Q_2)
4. Q_3 (upper quartile) is the middle number of the 2nd half of the data. (Don't include Q_2)

Example: Find the interquartile range of the following data sets.

Case 1: Even number of data values (odd number halves)

3 6 10 12 15 21

The data values are already ordered. The median is 11.

Consider the lower half of the set, which is 3 6 10. The middle score is 6, so $Q_1 = 6$.

Consider the upper half of the set, which is 12 15 21. The middle score is 15, so $Q_3 = 15$.

Case 2: Odd number of data values (odd number halves)

4 9 11 13 17 23 30

The data values are already ordered. The median is 13.

Consider the lower half of the set, which is 4 9 11. The middle score is 9, so $Q_1 = 9$.

Consider the upper half of the set, which is 17 23 30. The middle score is 23, so $Q_3 = 23$.

Case 3: Even number of data values (even number halves)

1 3 9 10 15 17 21 26

The data are already ordered. The median is 12.5.

Consider the lower half of the set, which is 1 3 9 10. The middle score is 6, so $Q_1 = 6$.

Consider the upper half of the set, which is 15 17 21 26. The middle score is 19, so $Q_3 = 19$.

Case 4: Odd number of data values (odd number halves)

2 7 13 14 17 19 21 25 29.

The data are already ordered. The median is 17.

Consider the lower half of the set, which is 2 7 13 14. The middle score is 10, so $Q_1 = 10$.

Consider the upper half of the set, which is 19 21 25 29. The middle score is 23, so $Q_3 = 23$.

Worked Example 15 - Using the CAS calculator to calculate the IQR.

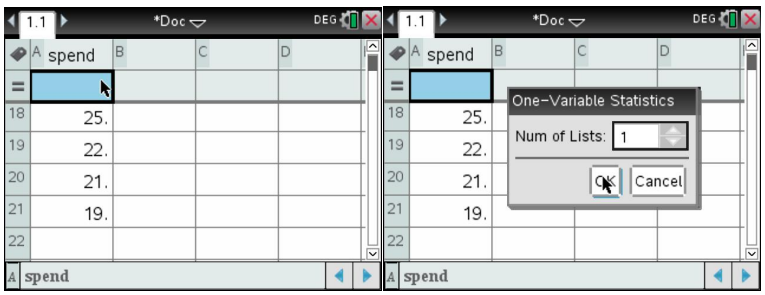
Parents are often shocked at the amount of money their children spend. The data below give the amount spent (to the nearest whole dollar) by each child in a group that was taken on an excursion to the Royal Melbourne Show.

15, 12, 17, 23, 21, 19, 16, 11, 17, 18, 23, 24, 25, 21, 20, 37, 17, 25, 22, 21, 19

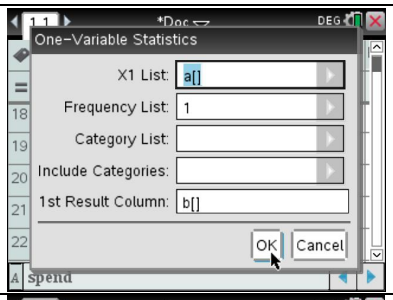
Calculate the Interquartile range for this data.

On a List & Spreadsheet page, Enter the data values in column A. Then press

- Menu [menu]
- 4: Statistics [4]
- 1: Stat Calculations [1]
- 1: One-Variable Statistics [1]

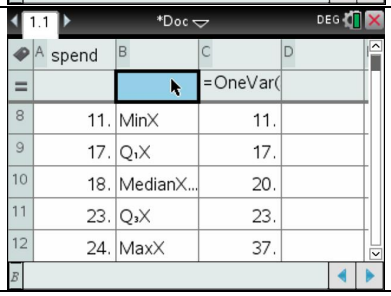


To generate the five number summary, select the number of lists as 1 and then complete the table as shown.
Note: Press Tab [tab] to move.



Scroll down through the column of summary statistics to find the five number summary. In this case we need the first (Q_1X) and third (Q_3X) quartiles to calculate the Interquartile Range (IQR)

IQR = 23 - 17 = 6



The Range

$$\textbf{Range} = \textbf{highest value} - \textbf{lowest value}$$
$$= X_{max} - X_{min}$$

The range gives us some idea about the spread of the data.

Example
For the following sets of data find the range

a) 2 3 5 8 9 11

b) 12 14 17 19 22 25

Mode

Is the number that occurs the most frequently/most often.

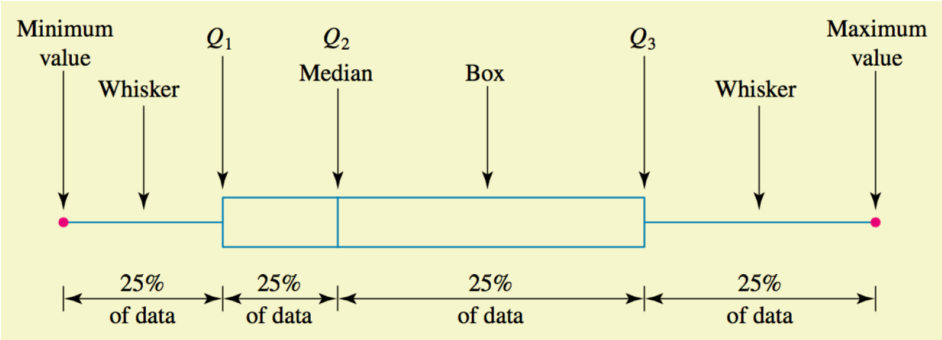
Example
Find the mode for each set of data

a) 2, 2, 2, 4, 4, 5, 5, 7, 8, 8, 8, 8, 8, 8, 9, 9, 9 mode =

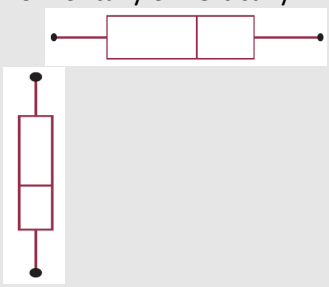
In certain situations there can be two numbers that occur most often, in this case the data is considered to be BIMODAL (two modes).

b) 4, 4, 6, 6, 6, 6, 6, 7, 7, 8, 8, 9, 9, 9, 9, 9, 10, 11, 11, 12 mode =

1.7 Boxplots (box and whisker)



Boxplots can be drawn horizontally or vertically.

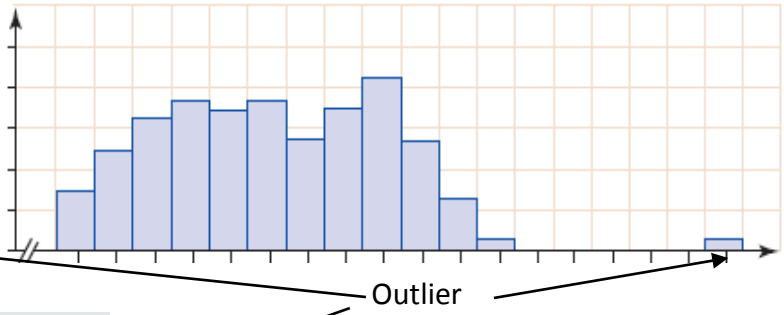


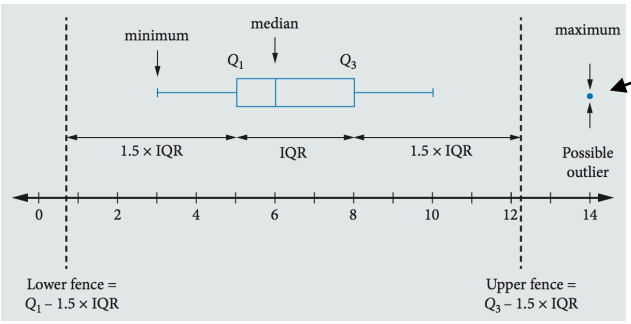
- Divides the data into 4 sections (25% in each section, hence the term quartiles)
- Shows the five-figure summary statistics
- Made up of a **box** with straight lines (**whiskers**) extending from opposite sides of the box.
- Must have a labelled scale.
- The **length of the box is** given by the **interquartile range**.

Outlier

A Data set may also have a data value that lies well away from other data. This is called an outlier

stem	leaf
1	9
2	0 5
3	2 3 7 9
4	5 7
5	8
6	
7	9
8	





An outlier is either:

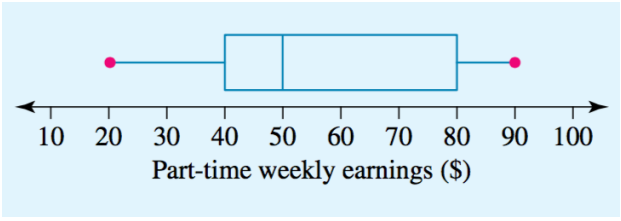
Less than the Lower boundary/fence: $Q_1 - (1.5 \times IQR)$

Or

Greater than the Upper boundary/fence: $Q_3 + (1.5 \times IQR)$

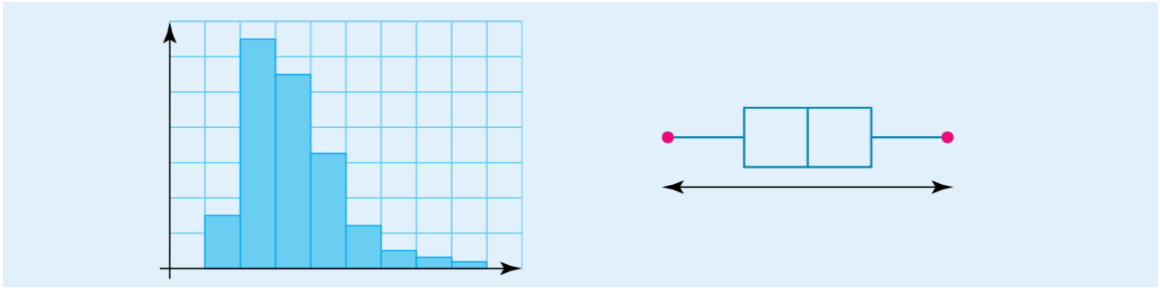
Worked Example 16

The boxplot at right shows the distribution of the part-time weekly earnings of a group of Year-11 students. Write down the range, the median and the interquartile range for these data.



Worked Example 17

Explain whether or not the histogram and the boxplot shown below could represent the same data.

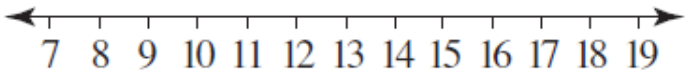


Worked Example 18

The results (out of 20) of oral tests in a Year-12 Indonesian class are:

15 12 17 8 13 18 14 16 17 13 11 12

Display these data using a boxplot and discuss the shape obtained.



Worked Example 18 (CAS Calculator)

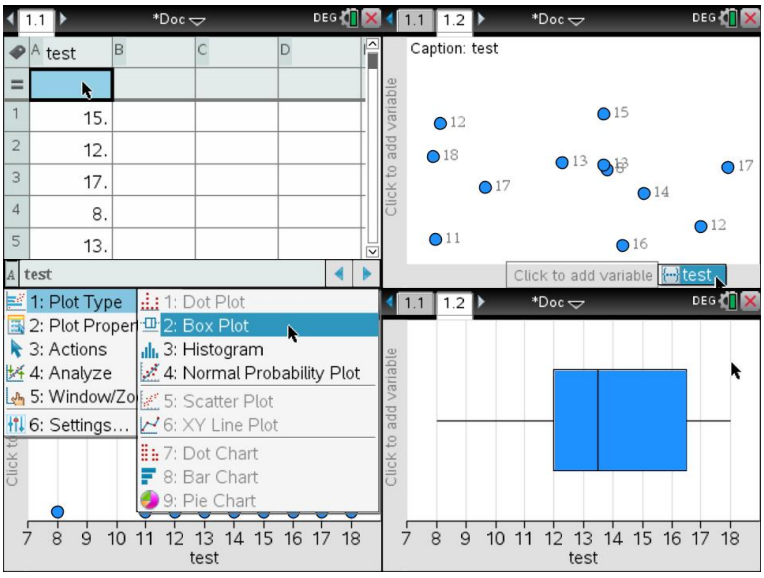
On a List & Spreadsheet page, Enter the data values in column A. Then press

- ctrl doc and add a data & statistics page

Click to add variable and choose the labelled column "test"

Now press

- menu 1 2 for a boxplot



Worked Example 19

The times (in seconds) achieved by the 12 fastest runners in the 100-m sprint at a school athletics meeting are listed below.

11.2 12.3 11.5 11.0 11.6 11.4 11.9 11.2 12.7 11.3 11.2 11.3

Draw a boxplot to represent the data, describe the shape of the distribution and comment on the existence of any outliers

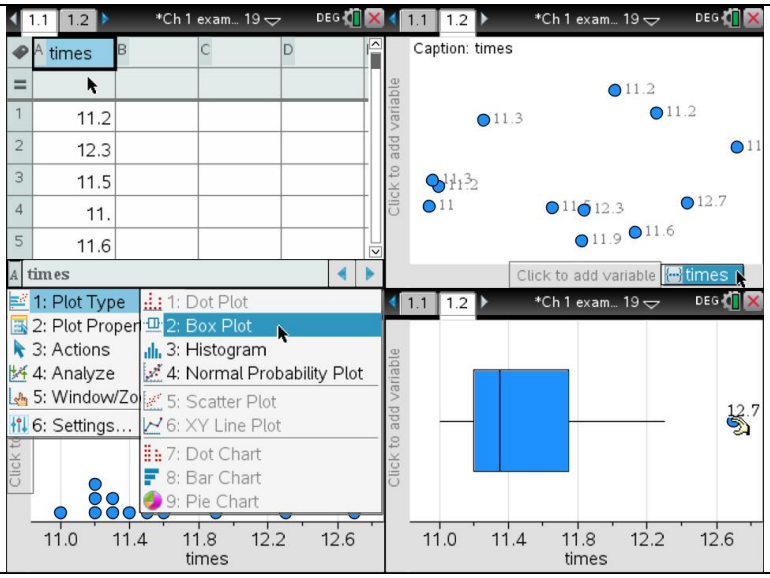
On a List & Spreadsheet page, Enter the data values in column A. Then press

- ctrl doc and add a data & statistics page

Click to add variable and choose the labelled column "test"

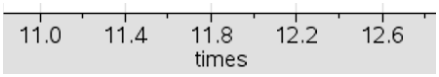
Now press

- menu 1 2 for a boxplot



Transfer the boxplot

Comment:



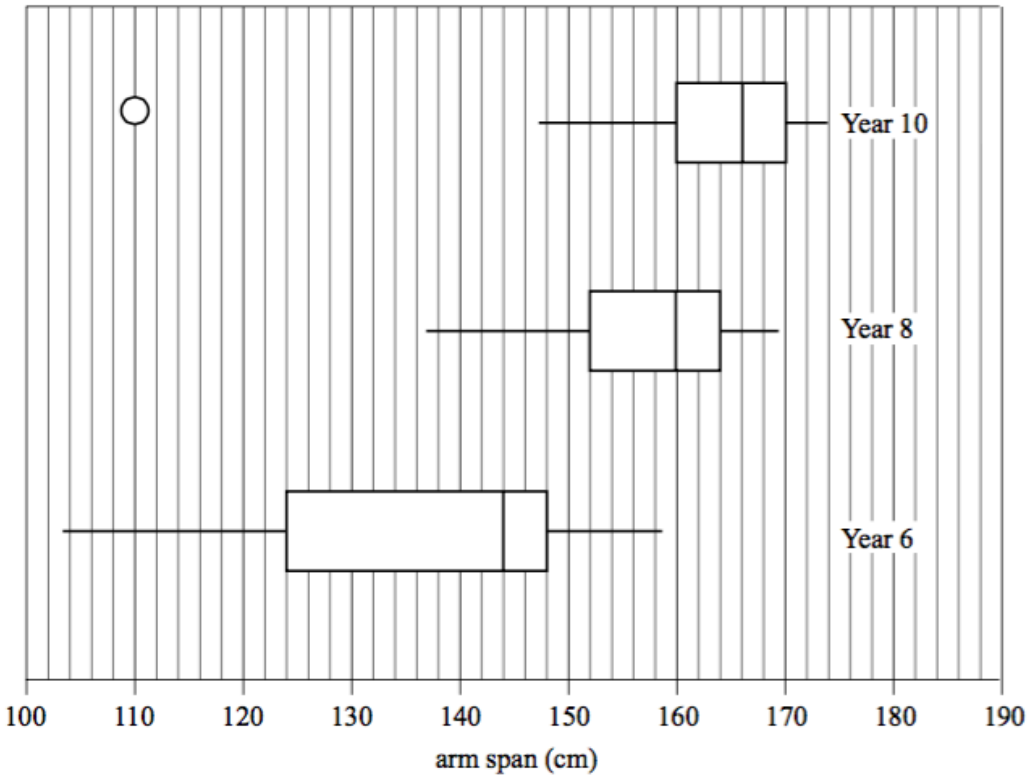
Which display do you use

Display	Type of Data	Guidelines
Bar Chart	Categorical data	Categories can be represented on the horizontal or vertical axis
Segmented bar chart	Categorical data	Should have no more than five segments
Histogram	Numerical data	Best if data has been grouped between 5 and 15 intervals
Boxplot	Numerical data	Best if you want to easily read the five-number summary
Dot Plot	Numerical data	Best used with a maximum of 50 data values and when the data values are not too spread out
Stem Plot	Numerical data	Best used with a maximum of 50 data values

Past exam question - Exam 2 2008

Question 3

The arm spans (in cm) were also recorded for each of the Years 6, 8 and 10 girls in the larger survey. The results are summarised in the three parallel box plots displayed below.



a. Complete the following sentence.

The middle 50% of Year 6 students have an arm span between ☐ and ☐ cm.

1 mark

b. The three parallel box plots suggest that arm span and year level are associated.
Explain why.

1 mark

c. The arm span of 110 cm of a Year 10 girl is shown as an outlier on the box plot. This value is an error. Her real arm span is 140 cm. If the error is corrected, would this girl's arm span still show as an outlier on the box plot? Give reasons for your answer showing an appropriate calculation.

2 marks

1.8 The Mean (average) of a sample

The mean (or average) is also a *summary statistic* and is a measure of the **centre** of a distribution.

The formal definition of the mean is:

$$\bar{x} = \frac{\sum x}{n}$$

where $\sum x$ represents the sum of all of the observations in the data set and n represents the number of observations in the data set.

Note that the symbol Σ, is the Greek letter, sigma, which represents 'the sum of'.

The mean is the point about which the distribution 'balances'.

Shape:

- Symmetric – $mean \approx median$
- Positively skewed – $mean > median$
- Negatively skewed – $mean < median$

- *When an outlier is present the mean becomes less reliable and the median is a better measure of the centre of such a distribution.*
- *If the data is skewed, the mean is less reliable as a measure of centre.*

Example

Case 1

Consider the masses of 7 potatoes, given in grams, below.

100 120 130 145 160 170 190

Median = Mean =

The distribution is _____

Case 2

Consider the masses of a different set of 7 potatoes, given in grams below.

100 105 110 115 120 160 200

Median = Mean =

The distribution is _____

Case 3

Consider the data below, showing the weekly income (to the nearest $10) of 10 families living in a suburban street.

$600 $1340 $1360 $1380 $1400 $1420 $1420 $1440 $1460 $1500

Median = Mean =

The distribution is _____

Worked Example 20

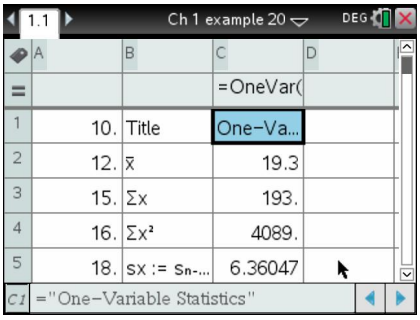
Calculate the mean of the set of data below using a CAS calculator.

10, 12, 15, 16, 18, 19, 22, 25, 27, 29

On a Lists & Spreadsheet page, enter the data into column A and label it *data*.

- Highlight the column and then press:
- Menu [menu], [4] Statistics, [1] Stat Calculations, [1] One-Variable Statistics

Press ENTER twice to generate the summary statistics.



Mean for grouped data

- Must find the midpoint of each interval
- Midpoint is multiplied by the frequency

To find the mean for grouped data,

$$\bar{x} = \frac{\sum(f \times m)}{\sum f}$$

where f represents the frequency of the data and m represents the midpoint of the class interval of the grouped data.

Worked Example 21

The ages of a group of 30 people attending a superannuation seminar are recorded in the frequency table below. Calculate the mean age of those attending the seminar.

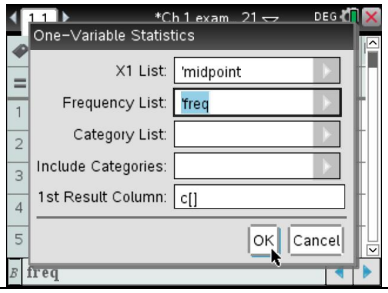
Age (class intervals)	Frequency, f	Midpoint of class interval, m	$f \times m$
20 – 29	1		
30 – 39	6		
40 – 49	13		
50 – 59	6		
60 – 69	3		
70 – 79	1		

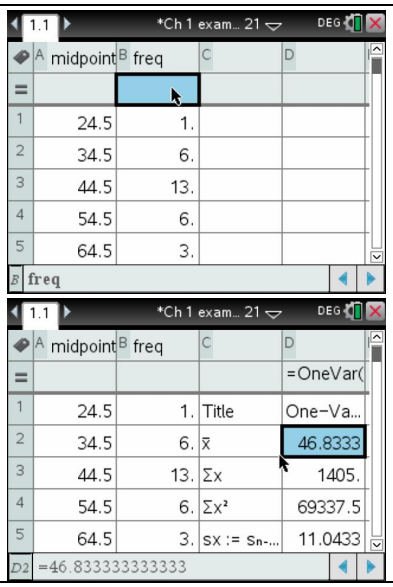
Calculate the mean using a CAS calculator.

On a Lists & Spreadsheet page, enter the midpoints into column A and label it *midpoints*.
Enter the frequencies into column B and label it *freq*.
To calculate the summary statistics for this set of data, press:

- MENU
- Statistics
- Stat Calculations
- One-Variable Statistics

The number of lists is 1 so select OK.
Complete the table as shown and select OK.





Rounding to a given number of significant figures

Significant figures are a method of simplifying a number by rounding it to a base 10 value. Questions relating to significant figures will require a number to be written correct to ***x*** number of significant figures. In order to complete this rounding, the relevant significant figure(s) needs to be identified.

Let's have a look at an example. Consider the number 123.456 789.

- This value has 9 significant figures, as there are nine numbers that tell us something about the particular place value in which they are located.
- The most significant of these values is the number 1, as it indicates the overall value of this number is in the hundreds.
- If asked to round this value to 1 significant figure, the number would be rounded to the nearest hundred, which in this case would be 100.
- If rounding to 2 significant figures, the answer would be rounded to the nearest 10, which is 120.
- Rounding this value to 6 significant figures means the first 6 significant figures need to be acknowledged, 123.456. However, as the number following the 6th significant figure is above 5, the corresponding value needs to round up, therefore making the final answer 123.457.

Rounding hint: If the number after the required number of significant figures is 5 or more, round up. If this number is 4 or below, leave it as is.

Significant figures and Zeros

Zeros present an interesting challenging when evaluating significant figures and are best explained using examples.

- 4056 contains 4 significant figures. The zero is considered a significant figure as there are numbers on either side of it.
- 4000 contains 1 significant figure. The zeros are ignored as they are place holders and may have been rounded.
- 4000.0 contains 5 significant figures. In this situation the zeros are considered important due to the zero after the decimal point. A zero after the decimal point indicates the numbers before it are precise.
- 0.004 contains 1 significant figure. As with 4000, the zeros are place holders.
- 0.0040 contains 2 significant figures. The zero following the 4 implies the value is accurate to this degree.

The following examples show how these rules work:

0.00**3561** – leading digits are ignored – 4 significant figures

70.036 – zeros between other digits are significant – 5 significant figures

5.320 – zeros included after decimal digits are significant – 4 significant figures

450000 – trailing zeros are not significant – 2 significant figures

78000.0 – the zeros after the decimal point are significant, so the zeros between other numbers are significant – 6 significant figures

As when rounding to a given number of decimal places, when rounding to a given number of significant figures consider digit after the specified number of figures. If it is five or more, round the final digit up; if it is four below, keep the final digit as it is.

5067.37 – rounded to 2 significant figures is **51**00

3199.01 - rounded to 4 significant figures is **3199**

0.004931 - rounded to 3 significant figures is 0.00**493**

1020004 - rounded to 2 significant figures is **10**000000

1.9 Standard Deviation of a given sample

The standard deviation is a measure of the spread of data from the mean.

In summary,

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

where s represents sample standard deviation
$\sum$ represents 'the sum of'
x represents an observation
$\bar{x}$ represents the mean
n represents the number of observations.

Symbol	Definition
	mean of the data
	sum of the data
	sum of the data squared
	sample standard deviation
	population standard deviation
	number of data points

Example:
Calculate the standard deviation using the following data.

8 10 11 12 12 13 $\quad \bar{x} = 11$

Data , x	Deviation from mean, $(x - \bar{x})$	$(x - \bar{x})^2$
8	8 – 11 = -3	
10		
11		
12		
12		
13		
	Total	

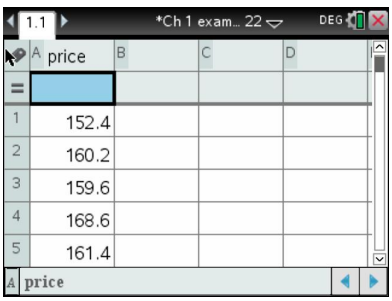
Worked Example 22

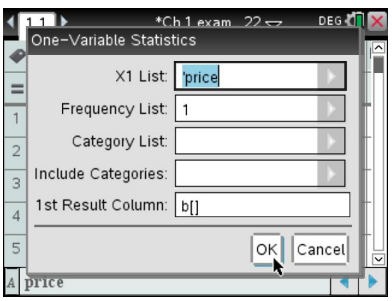
The price (in cents) per litre of petrol at a service station was recorded each Friday over a 15-week period. The data are given below.

152.4, 160.2, 159.6, 168.6, 161.4, 156.6, 164.8, 162.6, 161.0, 156.4, 159.0, 160.2, 162.6, 168.4, 166.8

Calculate the standard deviation for this set of data, correct to 2 decimal places.

On a Lists & Spreadsheet page, enter the data into Column A and label it *price.*
To calculate the summary statistics press:
- MENU [menu]
- [4] Statistics
- [1] Stat Calculations
- [1] One-Variable Statistics

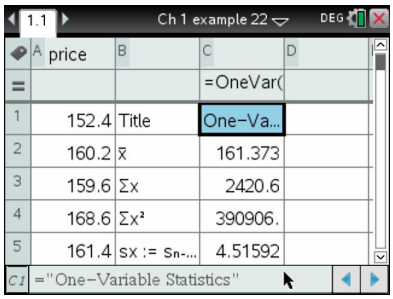




The number of lists is 1 so select OK.
Complete the table as shown and then press OK.

Scroll down to standard deviation sx

S_x = 4.52 cents/litre



Worked Example 23

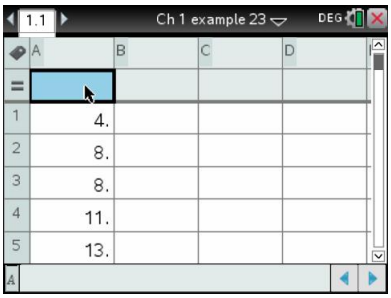
The number of students attending SRC meetings during the term is given in the stem plot shown. Calculate the standard deviation for this set of data, correct to 4 significant figures.

Stem	Leaf
0	4
0*	8 8
1	1 3 4
1*	5 8
2	3
2*	5

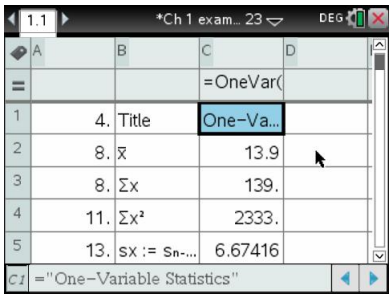
Key: 1|4 = 14 students

On the CAS proceed as above:

1) Enter data in Lists & Spreadsheets

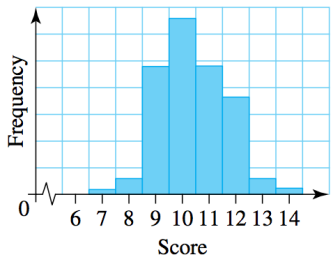


2) Calculate Sumary Staistics

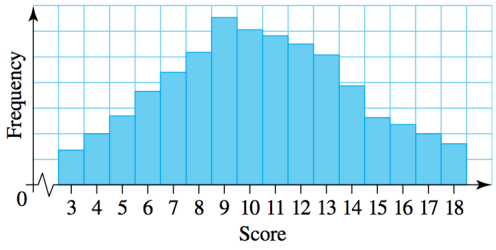


s_x = 6.674 (correct to 4 significant figures)

The standard deviation is a measure of the spread of data from the mean. Consider the two sets of data shown below.



Mean = 10
Standard dev. = 1



Mean = 10
Standard dev. = 3

Therefore we can say; _____

_____

_____

1.10 Populations and simple random samples

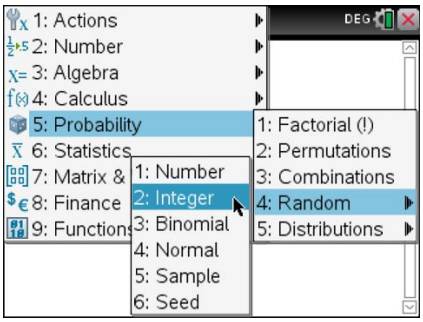
A group of Year 12 decide to investigate how much money all Year 12s spend on birthday presents. It would take a long time to survey all 200 students. So a smaller group known as a *sample* is taken from the total *population* of Year 12.

The mean of a data set which represents a population is μ.
The mean of a data set which represents a sample is $\bar{x}$.
The standard deviation of a data set which represents a population is σ.
The standard deviation of a data set which represents a sample is s.

Worked Example 24
Generate 5 random numbers (integers) between 1 and 50.

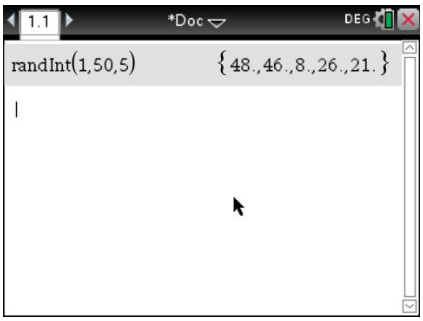
To generate random integers using a CAS calculator, open a Calculator page and press:

- MENU menu
- 5: Probability 5
- 4: Random 4
- 2: Integer 2



To generate 5 random numbers between 1 and 50, complete the entry line as: randInt(1, 50, 5).

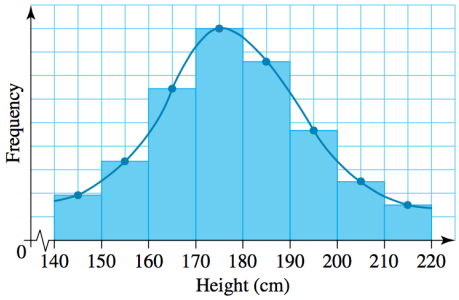
Then press ENTER enter.



1.11 The 68–95–99.7% rule and z-scores

The 68–95–99.7% rule

This rule can only be used for a **bell** shaped curve (symmetric distribution).



A bell shaped curve or symmetric distribution is referred to as a **normal distribution.**

Peaks in the middle and tails off either side.

The 68–95–99.7% rule for a bell-shaped curve states that approximately:
1. 68% of data lie within 1 standard deviation either side of the mean
2. 95% of data lie within 2 standard deviations either side of the mean
3. 99.7% of data lie within 3 standard deviations either side of the mean.

Figure 1

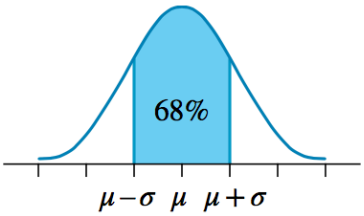


Figure 2

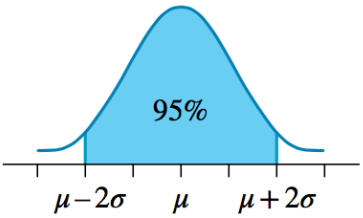
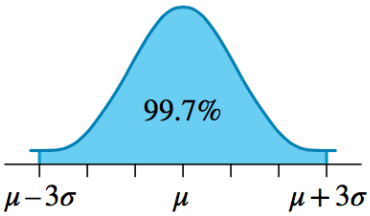
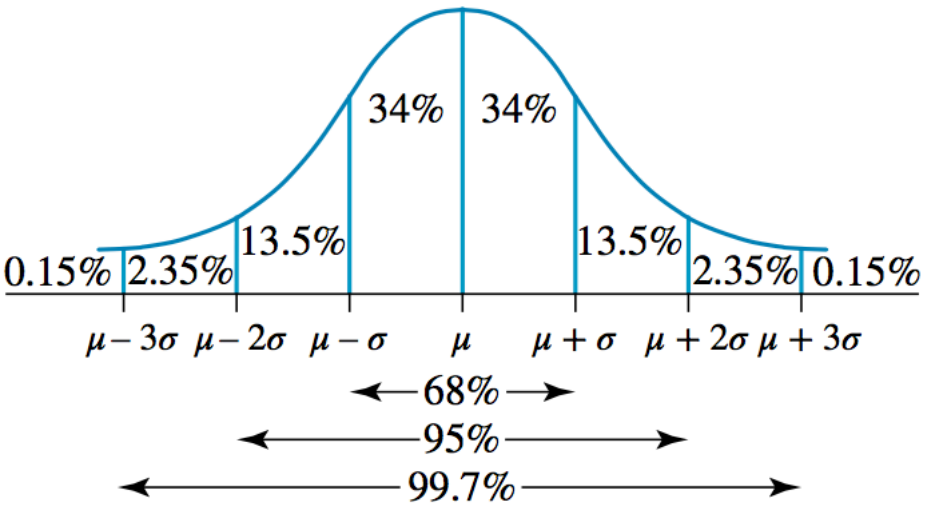


Figure 3



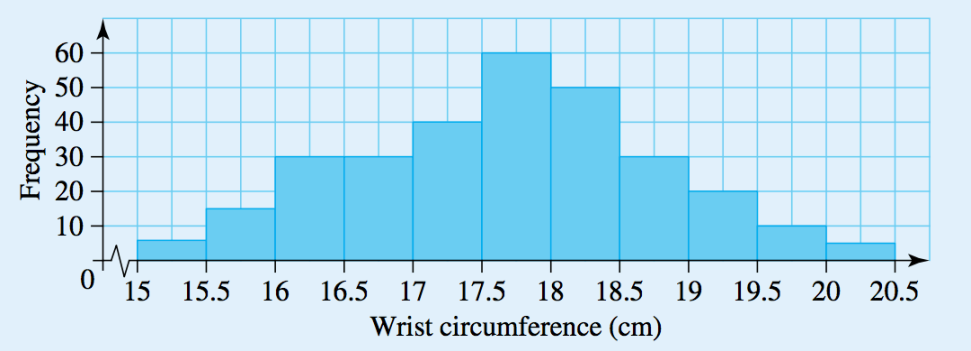
Using the 68–95–99.7% rule, we can work out the various percentages of the distribution which lie between the mean and 1 standard deviation from the mean and between the mean and 2 standard deviations from the mean and so on.

This diagram summarises this.



Worked Example 25

The wrist circumferences of a group of people were recorded and the results are shown in the histogram below. The mean of the set of data is 17.7 and the standard deviation is 0.9. Write down the wrist circumferences between which we would expect approximately:



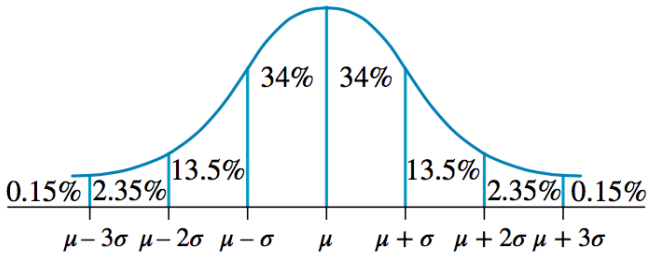
a) 68% of the group to lie

b) 95% of the group to lie

c) 99.7% of the group to lie.

Worked Example 26

The distribution of the masses of packets of 'Fibre-fill' breakfast cereal is known to be bell-shaped with a mean of 250 g and a standard deviation of 5 g. Find the percentage of Fibre-fill packets with a mass which is:



a) less than 260 g

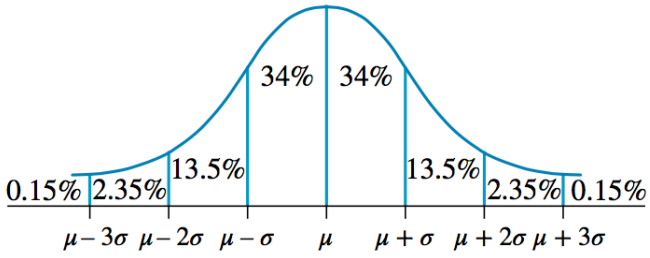
b) less than 245 g

c) more than 240 g

d) between 240 g and 255g.

Worked Example 27

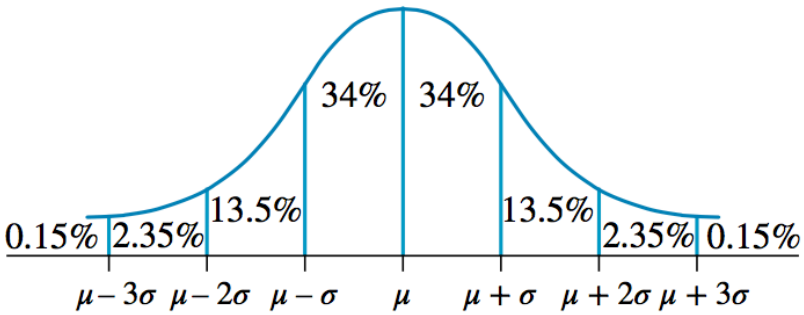
The number of matches in a box is not always the same. When a sample of boxes was studied it was found that the number of matches in a box approximated a normal (bellshaped) distribution with a mean number of matches of 50 and a standard deviation of 2. In a sample of 200 boxes how many would be expected to have more than 48 matches?



Standard z-scores

The ***z*-score** (also called the **standardised score**) is used to measure the position of a score in a data set relative to the mean.

A *z*-score of 0 indicates that the score obtained is equal to the mean.
The *z*-score measures the distance from the mean in terms of the standard deviation.
A score that is exactly one standard deviation above the mean has a *z*-score of 1.
A score that is exactly one standard deviation below the mean has a *z*-score of – 1.



To calculate a *z*-score we use the formula:

$$z = \frac{x - \mu}{\sigma}$$

where x = the score, μ = the mean of the population and σ = the standard deviation of the population.

Scores can be compared by their *z*-scores as they compare the score with the mean and the standard deviation.

When comparing scores, read the question carefully to see if a higher or lower *z*-score is a better outcome.

Worked Example 28

In an IQ test, the mean IQ is 100 and the standard deviation is 15. Dale's test results give an IQ of 130. Calculate this as a *z*-score.

Now calculate the *z*-score if Dale's IQ was 88

Example

To obtain the average number of hours study done by Year 12 students per week, Kate surveys 20 students and obtains the following results.

12 18 15 14 9 10 13 12 18 25 15 10 3 21 11 12 14 16 17 20

a) Calculate the mean and standard deviation (correct to 2 decimal places).

b) Robert does 16 hours of study each week. Express this as a *z*-score based on the above results. (Give your answer correct to 2 decimal places.)

An important use of *z*-scores is to compare scores from different data sets.

Suppose that in your maths exam your result was 74 and in English your result was 63. In which subject did you achieve the better result?

Worked Example 29

Janine scored 82 in her physics exam and 78 in her chemistry exam. In physics, $\mu = 62$ and $\sigma = 10$, while in chemistry, $\mu = 66$ and $\sigma = 5$.

a) Write both results as a standardised score.

b) Which is the better result? Explain your answer.

NORMAL DISTRIBUTION DIAGRAMS FOR USE IN YOUR EXAMS