

Summarising numerical data: the mean and the standard deviation

- What is the mean, how is it calculated and what does it tell us?
- What is the relationship between the mean and the median?
- What is the standard deviation, how is it calculated and what does it tell us?
- What is the 68–95–99.7% rule and how do we use it?
- What are standard scores and how are they used?
- What is a simple random sample and how is it formed?

In the last chapter, we looked at one way of defining the centre and spread of a data distribution. This introduced us to the median, interquartile range and range. In this chapter, we will look at an alternative way of looking at centre and spread through the mean and the standard deviation.

3.1 The mean

The mean of a set of data is what most people call the ‘average’. The mean of a set of data is given by:

$$\text{mean} = \frac{\text{sum of data values}}{\text{total number of data values}}$$

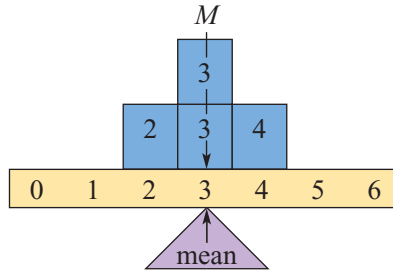
For example, consider the set of data:

$$2 \quad 3 \quad 3 \quad 4$$

The mean of this set of data is given by:

$$\text{mean} = \frac{2 + 3 + 3 + 4}{4} = \frac{12}{4} = 3$$

From a pictorial point of view, the mean is the **balance** point of a distribution.



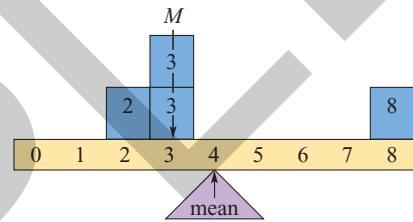
Note that in this case, the mean and the median coincide; the balance point of the distribution is also the point that splits the distribution in half, that is, there are two data points to the left of the mean and two to the right. This is a general characteristic of **symmetric** distributions.

However, consider the data set:

2 3 3 8

The median remains at $M = 3$, but:

$$\text{mean} = \frac{2 + 3 + 3 + 8}{4} = \frac{16}{4} = 4$$



Note that the mean is affected by changing the largest data value but that the median is not.

Some notation

Because the rule for the mean is relatively simple, it is easy to write in words. However, later you will meet other rules for calculating statistical quantities that are extremely complicated and hard to write out in words. To overcome this problem, we will introduce a shorthand notation that enables complex statistical formulas to be written out in a compact form. In this notation, we use:

- the Greek capital letter sigma, Σ , as a shorthand way of writing ‘sum of’
- a lower case x to represent a data value
- a lower case x with a bar, \bar{x} (pronounced ‘ x bar’), to represent the mean of the data values
- an n to represent the total number of data values

The rule for calculating the mean then becomes:

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

Example 1

Calculating the mean from the formula

The following is a set of reaction times (in milliseconds):

38 36 35 43 46 64 48 25

Write down the values of

a n **b** Σx **c** \bar{x}

correct to one decimal place

Solution

a n is the number of data values

$$n = 9$$

b Σx is the sum of the data values

$$\begin{aligned}\Sigma x &= 38 + 36 + 35 + 43 + 46 + 64 + 48 + 25 \\ &= 335\end{aligned}$$

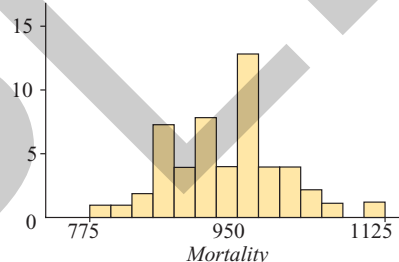
c \bar{x} is the mean. It is defined by $\bar{x} = \frac{\Sigma x}{n}$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{335}{9} = 37.2$$

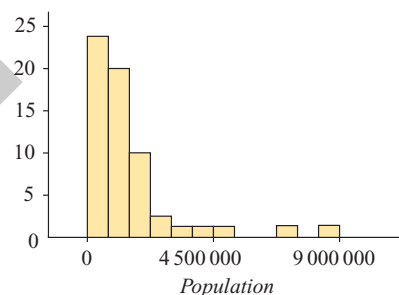
The relationship between the mean and the median

Whereas the **median** lies at the **midpoint** of a distribution, the **mean** is the **balance point** of the distribution. For approximately symmetric distributions, both the median and mean will be approximately equal in value.

An example of a **symmetric distribution** is the distribution of mortality rates for 60 US cities shown opposite. Calculations reveal that the mean mortality rate for the cities is 940 per 100 000 while the median mortality rate is 944 per 100 000. As expected, the mean and median are approximately equal in value.



An example of a highly **skewed distribution** is the population distribution of these cities shown opposite. This distribution is clearly positively skewed with two outliers. The mean population is 1.4 million, while the median population is 0.9 million. They are quite different in value. The mean has been pulled away from the body of the data by the extreme values in the tail and no longer represents the typical city.

**When to use the median rather than the mean**

Because the value of the **median** is relatively unaffected by the presence of extreme values in a distribution, it is said to be a **resistant** statistic. For this reason, the median is frequently used as a measure of centre when the distribution is known to be clearly **skewed** and/or likely to contain **outliers**. For example, median house prices are used to compare housing prices between capital cities in Australia because the distribution of house prices tends to be positively skewed. There are always a small number of very expensive houses sold for amounts that are an order of magnitude higher than the prices of the rest of houses sold (outliers). Likewise, the quartiles Q_1 and Q_3 are resistant statistics and are used as indicators of spread in such situations.

However, if a distribution is symmetric, there will be little difference in the value of the mean and median and we can use either. In such circumstances, the mean is often preferred because:

- it is more familiar to most people
- more can be done with it theoretically, particularly in the area of statistical inference

(which you will learn about at a later stage if you go on to study statistics at university)
Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMennamin

Choosing between the mean and the median

The **mean** and the **median** are both measures of the **centre** of a distribution. If the distribution is:

- **symmetric** and there are no outliers, either the **mean** or the **median** can be used to indicate the centre of the distribution
- clearly **skewed** and/or there are **outliers**, it is more appropriate to use the **median** to indicate the **centre** of the distribution

Exercise 3A

1 For each of the following data sets:

a 2 5 2 3 **b** 12 15 20 32 25 **c** 2 1 3 2 5 3 5

write down the value of n , the value of Σx and hence evaluate \bar{x} .

2 Calculate the mean and locate the median and modal value(s) of the following scores:

a 1 3 2 1 2 6 4 5 4 3 2 **b** 3 12 5 4 3 2 6 5 4 5 5 6

3 **a** Which statistic, the median or the mean, always divides a distribution in half?

b In what shaped data distributions do the mean and median have the same value?

c Which of the median or the mean is most affected by outliers?

d Which would be the most appropriate measure of the typical salary of adult workers in Australia: the mean salary or the median salary? Why?

4 The temperature of a hospital patient (in degrees Celsius) taken at six-hourly intervals over two days was as follows:

35.6 36.5 37.2 35.5 36.0 36.5 35.5 36.0

a Calculate the patient's mean and median temperature over the two-day period.

b What do these values tell you about the distribution of the patient's temperature?

5 The amounts (in dollars) spent by seven customers at a corner store were:

0.90 0.80 2.15 16.55 1.70 0.80 2.65

a Calculate the mean and median amount spent by the customers.

b Does the mean or the median give the best indication of the typical amount spent by customers? Explain your answer.

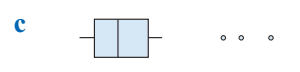
6 For which of the following variables might you question using the mean as a measure of the centre of the distribution? Justify your selection.



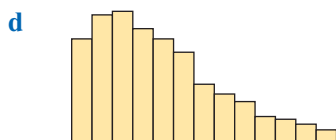
Life expectancy in Asia (years)



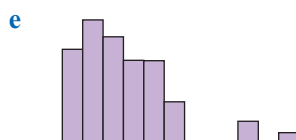
Life expectancy in Europe (years)



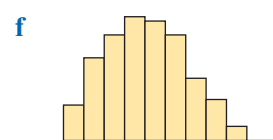
Fuel consumption of cars



Age distribution in a country



Urban car accident rates



Blood cholesterol levels

7 The stem plot shows the distribution of weights (in kg) of 24 footballers.

- a** From the shape of the distribution, which measure of centre, the mean or the median, do you think would best indicate the typical weight of these footballers?
- b** Calculate both the mean and median to check your prediction.

Weight (kg)	
6	9
7	0 2 4
7	6 6 7 8
8	0 0 1 2 3 3 4
8	5 5 5 6 9
9	1 2
9	8
10	3

8 The stem plot shows the distribution of life expectancies of 23 countries.

- a** From the shape of the distribution, which measure of centre, the mean or the median, do you think would best indicate the typical life expectancy in these countries?
- b** Calculate both the mean and median to check your prediction.

Life expectancy (years)	
5	2
5	5 6
6	4
6	6 6 7 9
7	1 2 2 3 3 4 4 4 4
7	5 5 6 6 7 7

3.2 Measuring the spread around the mean: the standard deviation

To measure the spread of a data distribution around the **median** (M) we use the **interquartile range** (IQR).

To measure the spread of a data distribution about the **mean** (\bar{x}) we use the **standard deviation** (s).

The standard deviation

The formula for the standard deviation, s , is:

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

where n is the number of data values (sample size) and \bar{x} is the mean.

Although not easy to see from the formula, the standard deviation is an average of the squared deviations of each data value from the mean. We work with the squared deviations because the sum of the deviations around the mean (the balance point) will always be zero. Finally, for

technical reasons that are not important in this course, we average by dividing by $n - 1$ not n . In practice this is not a problem, as dividing by $n - 1$ compared to n makes very little difference to the final value except for very small samples.

Calculating the standard deviation

Normally, you will use your calculator to determine the value of a standard deviation. However, to understand what is involved when your calculator is doing the calculation for you, it is best that you know how to calculate the standard deviation from the formula first.

How to calculate the standard deviation from the formula

Use the formula to calculate the standard deviation of the data set: 1 3 7 5 4
Give your answer correct to one decimal place.

Steps

- Write down the data and the value of n . $1, 3, 7, 5, 4 \quad n = 5$
- Use $\bar{x} = \frac{\sum x}{n}$ to find the value of the mean. $\bar{x} = \frac{\sum x}{n} = \frac{1+3+7+5+4}{5} = \frac{20}{5} = 4$
- To calculate s , it is convenient to set up a table with columns for:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
1	-3	9
3	-1	1
7	3	9
5	1	1
4	0	0
Sum	0	20

 - x the data values
 - $(x - \bar{x})$ the deviations from the mean
 - $(x - \bar{x})^2$ the squared deviations

The sum, $\sum (x - \bar{x})^2$, can then be found by adding the values in the $(x - \bar{x})^2$ column.
- Substitute the required values into the formula $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ and evaluate. $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{20}{5 - 1}} = 2.236 \dots$
- Write down your answer as required. *The standard deviation is $s = 2.2$, correct to one decimal place.*

As you can see, the calculation of the standard deviation from the formula is rather complicated and time consuming. Fortunately, you can use your graphics calculator to do the calculations for you. However, before we use any technology to do calculations for us, we need to have some way of checking whether the answers it gives are reasonable. To do this, we need to have some way of estimating the values of the quantities we are calculating.

Estimating the standard deviation from the range

Later in this chapter you will learn that for many data distributions, around 95% of data values lie within two standard deviations of the mean. We can use this information to **estimate** the value of the standard deviation by pushing this approximation a bit further.

Let us say that all the data values lie within two standard deviations of the mean. If this is true then

the range \approx four standard deviations (\approx means ‘approximately equals’)

$$\therefore \text{one standard deviation} \approx \frac{\text{range}}{4} \quad \text{or} \quad s \approx \frac{R}{4}$$

A rule for estimating the standard deviation

$$\text{standard deviation} \approx \frac{\text{range}}{4} \quad \text{or} \quad s \approx \frac{R}{4}$$

Example 2

Estimating the standard deviation

Estimate the value of the standard deviation of the data set 2 4 3 7 5 9 4 5 4 using the rule $s \approx \frac{R}{4}$.

Solution

- Determine the value of the range, R .
- Substitute the value of R in the formula $s \approx \frac{R}{4}$.
- Write down your answer, in this case rounding to the nearest whole number. **Remember, you are only estimating.**

$$R = 9 - 2 = 7$$

$$\therefore s \approx \frac{R}{4} = \frac{7}{4} = 1.75$$

The estimated value of the standard deviation is 2.

Note: The true value is 2.1, correct to one decimal place.

Now that we have a way of checking the reasonableness of our results, we can feel confident about using a calculator to calculate standard deviations.

How to calculate the mean and standard deviation using the TI-Nspire CAS

The following are all heights (in cm) of a group of women:

176 160 163 157 168 172 173 169

Determine the mean and standard deviation of the women’s heights. Give your answers correct to 2 decimal places.

Steps

- 1 Start a new document by pressing $\text{ctrl} + \text{N}$.
- 2 Select **3:Add Lists & Spreadsheet**.
Enter the data into a list named *height*, as shown.
- 3 Statistical calculations can be done in either the **Lists & Spreadsheet** application or the **Calculator** application (used here).

Press $\text{ctrl} + \text{I}$ and select **1:Add Calculator**.

- a Press $\text{menu} / 6:\text{Statistics}/1:\text{Stat Calculations}/1:\text{One-Variable Statistics}$.

Keystrokes: $\text{menu} \ 6 \ 1 \ 1$

This will generate the pop-up screen shown opposite.

- b As we only require the mean and standard deviation for one set of data

- i. Press enter to generate a second pop-up screen, as shown opposite.
- ii. To complete this screen, use the \blacktriangledown arrow and enter to paste in the list name *height*. Pressing enter exits this screen and generates the results screen shown next.

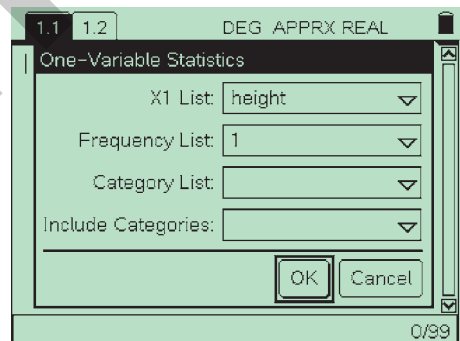
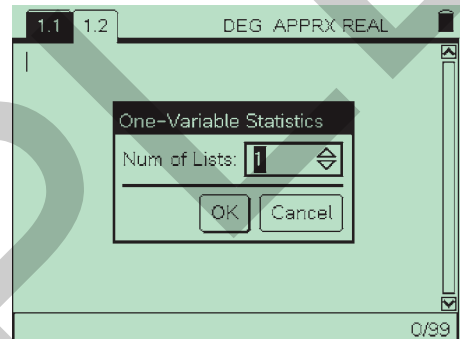
4. Write down the answers to the required degree of accuracy (i.e. 2 decimal places).

The mean height of the women is $\bar{x} = 167.25 \text{ cm}$ and the standard deviation is $s = 6.75 \text{ cm}$.

Notes:

- 1 The sample standard deviation is **sx**.
- 2 Use the $\blacktriangle \blacktriangledown$ arrows to scroll through the results screen to obtain values for additional statistical values (i.e. Q1, median, Q3 and maximum value) if required.

A	B	C	D
height			
1	176.		
2	160.		
3	163.		
4	157.		
5	168.		
6	172.		
Alt	176		



OneVar height, 1: stat.results

"Title"	"One-Variable Statistics"
" \bar{x} "	167.25
" Σx "	1338.
" Σx^2 "	224092.
"sx := s _{n-1} x"	6.67083
"sx := s _n x"	6.23999
"n"	8.
"Min X"	157.

How to calculate the mean and standard deviation using the ClassPad

The following are all heights (in cm) of a group of women:

176 160 163 157 168 172 173 169

Determine the mean and standard deviation of the women's heights. Give your answers correct to 2 decimal places.

Steps

- 1 Open the **Statistics** application and enter the data into the column labelled **height**.

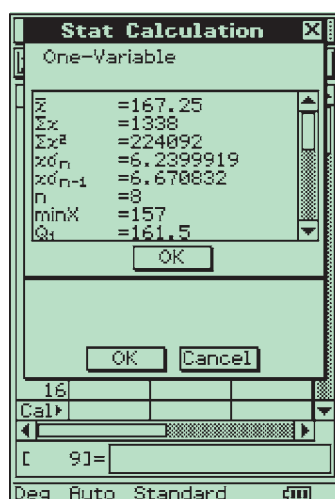
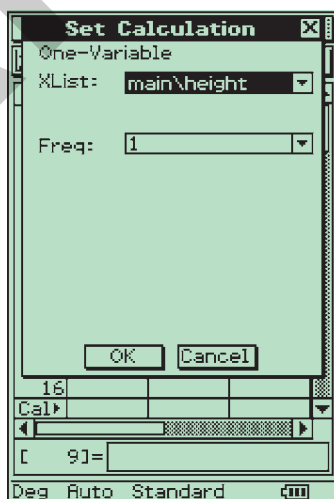
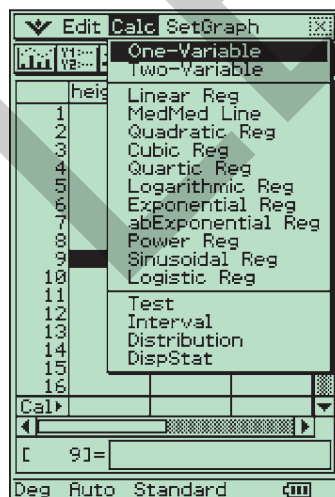
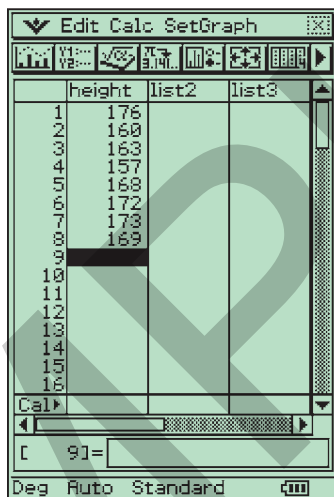
Your screen should look like the one shown.

- 2 To calculate the mean and standard deviation, select **Calc** from the menu bar and tap **One-Variable** from the drop-down menu to open the **Set Calculation** dialog box shown below (left).

- 3 Complete the dialog box as given below. For

- **XList:** select **main \ height**
- **Freq:** leave as **1**

- 4 Tap **OK** to confirm your selections and calculate the required statistics, as shown.



The mean height of the women is $\bar{x} = 167.25$ cm
and the standard deviation is $s = 6.75$ cm.

Notes:

- 1 The value of the standard deviation is given by $\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$.
- 2 Use the \blacktriangle \blacktriangledown side-bar arrows to scroll through the results screen to obtain values for additional statistical values (i.e. median, Q3 and the maximum value) if required.
- 5 Write down the answers to the required degree of accuracy (in this case, 2 decimal places).

Exercise 3B

- 1 Use the formula to calculate the value of the mean and standard deviation for each of these data sets.
 - a 1 2 2 4 6
 - b 10 10 8 6 16
 - c 5 5 5 5 5
- 2 Estimate the value of the standard deviation for each of the following data sets. Check the closeness of your estimates by calculating.
 - a 4 8 21 9 15 6 7 3 14 17 10
 - b 101 115 114 99 106 112 119
 - c 2.1 2.0 1.9 1.6 2.5 2.2 2.0
- 3 Which measure of spread:
 - a **always** incorporates 50% of the scores?
 - b uses only the smallest and largest scores in the distribution?
 - c gives the average variation around the mean?
- 4 For which of the following variables does it **not** make sense to calculate a mean or standard deviation?
 - a speed (in km/h)
 - b sex
 - c age (in years)
 - d year level (in a school)
 - e neck circumference (in cm)
 - f weight (underweight, normal, overweight)
- 5 A sample of 10 students were given a general knowledge test with the following results:

20 20 19 21 21 18 20 22 23 17

 - a Calculate the mean and standard deviation of the test scores correct to one decimal place.
 - b The median test score is 20, which is similar in value to the mean. What does this tell you about the distribution of test scores?

- 6 Calculate the mean and the standard deviation for each of the variables in the table. Give your answers correct to the nearest whole number for cars and TVs, and one decimal place for alcohol consumption.

Number of TVs/1000	Number of cars/1000	Alcohol consumption (litres)
378	417	17.6
404	286	12.5
471	435	16.0
354	370	24.1
539	217	9.9
381	357	9.5
674	530	14.6

7 The table below lists the pulse rates of 23 adult females and 23 adult males.

<i>Pulse rate (beats per minute)</i>	
<i>Females</i>	<i>Males</i>
65 73 74 81 59 64 76 83 95 70 73 79	80 73 73 78 75 65 69 70 70 78 58 77
64 77 80 82 77 87 66 89 68 78 74	64 76 67 69 72 71 68 72 67 77 73

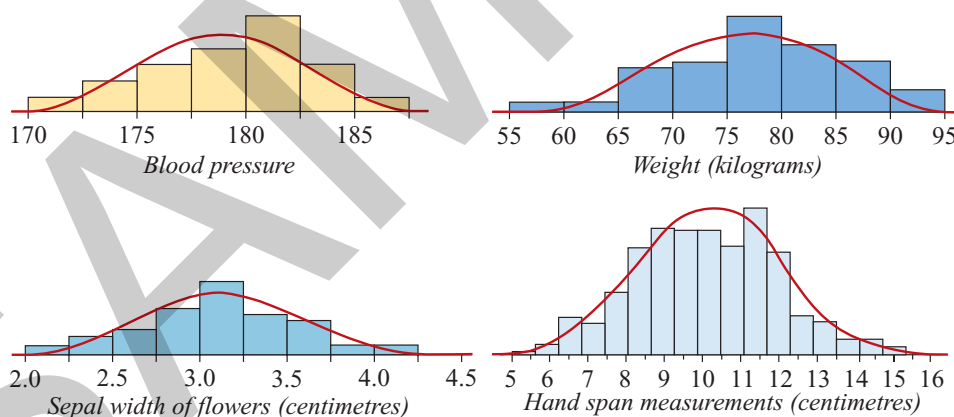
Calculate the mean and standard deviation for the male and female pulse rates (to 1 decimal place).

3.3 The normal distribution and the 68–95–99.7% rule: giving meaning to the standard deviation

We know that the interquartile range is the spread of the middle 50% of the data set. Can we find some similar way in which to interpret the standard deviation? It turns out that we can, but we need to restrict ourselves to a particular type of distribution known as the **normal** distribution. However, in practice, this is not a very limiting restriction, as many of the data distributions we work with in statistics (but not all) can be well approximated by this type of distribution.

The normal distribution

Many data sets that arise in practice are roughly symmetrical and have an approximate bell shape as shown in the four examples on the next page.



Data distributions that are approximately bell shaped can be modelled by a **normal** distribution.

The 68–95–99.7% rule

In normal distributions, the percentage of observations that lie within a certain number of standard deviations of the mean can always be determined. In particular, we are interested in the percentage of observations that lie within one, two or three standard deviations of the mean. This gives rise to what is known as the **68–95–99.7% rule**.

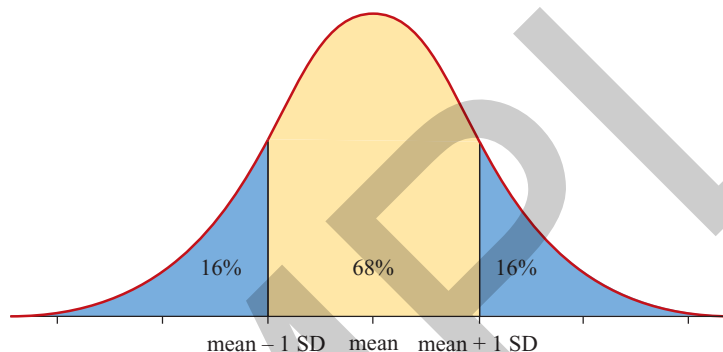
The 68–95–99.7% rule

For a **normal** distribution, approximately:

- **68%** of the observations lie within **one** standard deviation of the mean
- **95%** of the observations lie within **two** standard deviations of the mean
- **99.7%** of the observations lie within **three** standard deviations of the mean

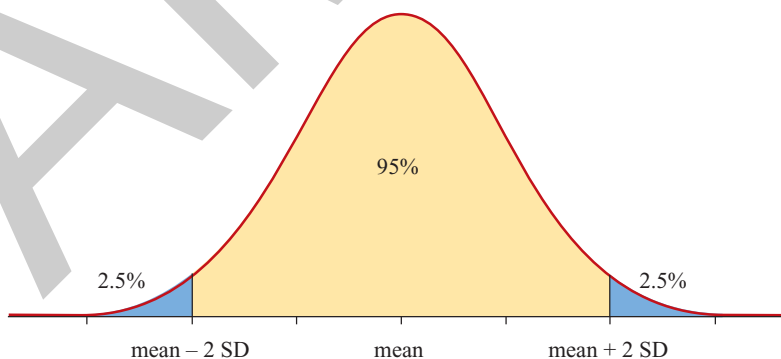
It is helpful to view this rule graphically. If a data distribution can be regarded as being approximately normal, then:

- around **68%** of the data values will lie within **one standard deviation (SD)** of the mean



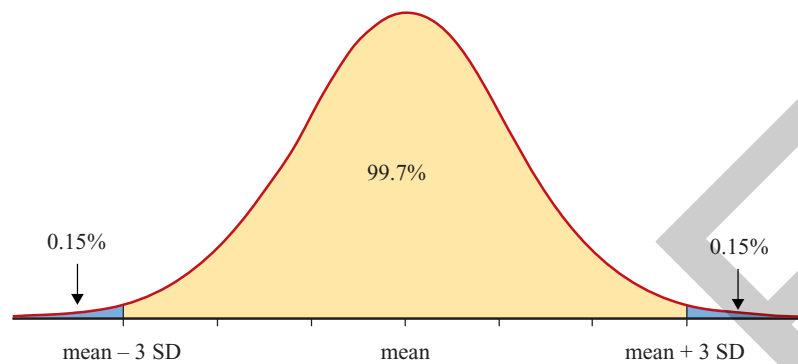
This also means that 32% of values lie outside this region. As the distribution is approximately symmetric, we can also say that around 16% of values lie in each of the tails (shaded blue).

- around **95%** of the data values will lie within **two standard deviations** of the mean



This also means that 5% of values lie outside this region. As the distribution is symmetric, we can also say that around 2.5% of values lie in each of the tails (shaded blue).

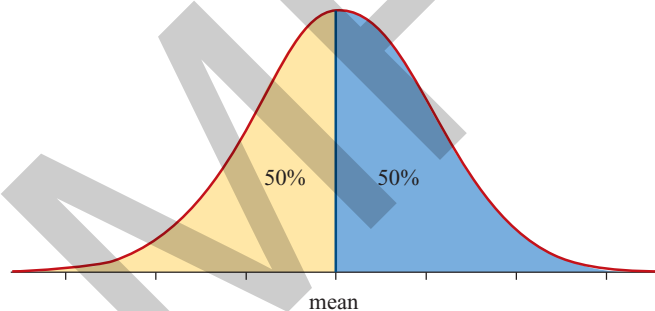
- around **99.7%** of the data values will lie within **three standard deviations** of the mean



This also means that 0.3% of values lie outside this region. As the distribution is symmetric, we can also say that around 0.15% of values lie in each of the tails (shaded blue).

In addition, because the **normal distribution** is **symmetric**, the mean and the median coincide so that:

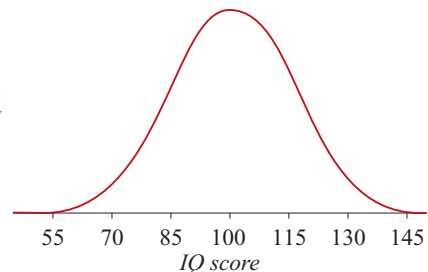
- **50%** of the data values will lie **above** the mean and **50%** of values will lie **below** the mean.



Applying the 68–95–99.7% rule

The distribution of IQ scores can be approximated by a normal distribution with a mean of 100 and a standard deviation of 15. This distribution is shown opposite.

With this information, the **68–95–99.7% rule** enables us to make the following statements about IQ scores and percentages.



- About **68%** of the IQ scores will lie between 85 ($= 100 - 1 \times 15$) and 115 ($= 100 + 1 \times 15$). This also implies that 16% of IQ scores will lie *below* 85 and 16% of IQ scores will lie *above* 115.
- About **95%** of the IQ scores will lie between 70 ($= 100 - 2 \times 15$) and 130 ($= 100 + 2 \times 15$). This also implies that 2.5% of IQ scores will lie *below* 70 and 2.5% of IQ scores will lie *above* 130.
- About **99.7%** of the IQ scores will lie between 55 ($= 100 - 3 \times 15$) and 145 ($= 100 + 3 \times 15$). This also implies that 0.15% of IQ scores will lie *below* 55 and 0.15% of IQ scores will lie *above* 145.

A strategy for solving problems involving the 68–95–99.7% rule

The key to solving problems involving the use of the **68–95–99.7%** rule is to determine whether the values involved are one, two or three standard deviations from the mean.

If the specified values:

- lie **between** one, two or three standard deviations from the mean, then use the **68–95–99.7%** rule directly
- are **more than** one, two or three standard deviations **above** or **below** the mean, use the tail percentages, **16%**, **2.5%** and **0.15%**

Finally, because the normal distribution is **symmetric**, **50%** of values lie **above** the mean and **50%** lie **below** the mean.

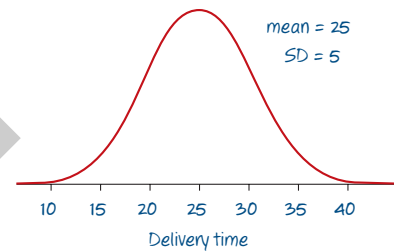
Example 3**Applying the 68–95–99.7% rule**

The distribution of delivery times for pizzas made by Pizza House is approximately normal, with a mean of 25 minutes and a standard deviation of five minutes.

a What percentage of pizzas have delivery times between 15 and 35 minutes?

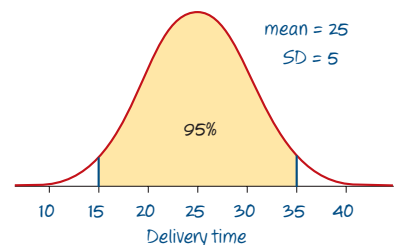
Solution

1 Draw, scale and label a normal distribution with a mean of 25 and a standard deviation of 5.



2 Shade in the region under the normal curve representing delivery times between 15 and 35 minutes.

3 Note that delivery times between 15 and 35 minutes lie within two standard deviations of the mean.
($15 = 25 - 2 \times 5$ and $35 = 25 + 2 \times 5$)



4 Recall that 95% of values lie within two standard deviations of the mean. Write down your answer.

95% of pizzas will have delivery times between 15 and 35 minutes

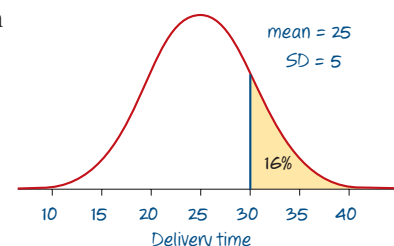
b What percentage of pizzas have delivery times greater than 30 minutes?

Solution

1 As before, draw, scale and label a normal distribution with a mean of 25 and a standard deviation of 5.

Shade in the region under the normal curve representing delivery times greater than 30 minutes.

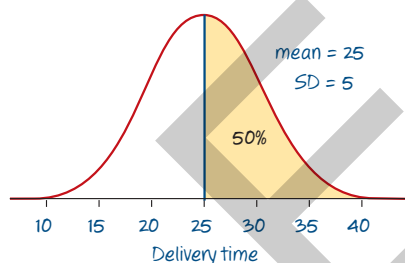
2 Note that delivery times greater than 30 minutes lie more than one standard deviation above the mean. ($30 = 25 + 1 \times 5$)



- 3 Recall that 16% of values lie more than one standard deviation above the mean. Write down your answer. *16% of pizzas will have delivery times greater than 30 minutes*
- c What percentage of pizzas have delivery times less than 25 minutes?

Solution

- 1 Draw, scale and label a normal distribution with a mean of 25 and a standard deviation of 5. Shade in the region under the normal curve representing delivery times less than 25 minutes.
- 2 Note that a delivery time of 25 minutes corresponds to the mean delivery time.
- 3 Recall that, in a normal distribution, 50% lie below the mean. Write down your answer.



- d In one month, Pizza House delivers 2000 pizzas. How many of these pizzas are delivered in less than 10 minutes?

Solution

- 1 Write down the total number of pizzas delivered.
- 2 Delivery times of less than 10 minutes are more than three standard deviations below the mean. ($10 = 25 - 3 \times 5$). Recall that 0.15% of values are more than one standard deviation below the mean. Write this down.
- 3 Therefore, the number of pizzas delivered in less than 10 minutes is 0.15% of 2000. Write this down and evaluate

50% of pizzas will have delivery times less than 25 minutes

Total number = 2000
Percentage delivered in less than 10 minutes = 0.15%

Number of pizzas delivered in less than 10 minutes

$$\begin{aligned}
 &= 0.15\% \text{ of } 2000 \\
 &= \frac{0.15}{100} \times 2000 = 3
 \end{aligned}$$



Exercise 3C

- 1** The distribution of blood pressure readings for executives is known to be approximately normally distributed with a mean blood pressure of 134 and a standard deviation of 20. From this information it can be concluded that:
- a** about 68% of the executives have blood pressures between and
 - b** about 95% of the executives have blood pressures between and
 - c** about 99.7% of the executives have blood pressures between and
 - d** about 16% of the executives have blood pressures above
 - e** about 2.5% of the executives have blood pressures below
 - f** about 0.15% of the executives have blood pressures below
 - g** about 50% of the executives have blood pressures above
- 2** The average weight of a bag of ten blood plums picked at U-Pick Orchard is approximately normally distributed with a mean of 1.88 kg and a standard deviation of 0.2 kg. From this information we can conclude that, from this orchard, the percentage of the bags of ten plums that weigh:
- a** between 1.68 and 2.08 kg is approximately %
 - b** between 1.28 and 2.48 kg is approximately %
 - c** more than 2.08 kg is approximately %
 - d** more than 2.28 kg is approximately %
 - e** less than 1.28 kg is approximately %
 - f** more than 1.88 kg is approximately %
- 3** The distribution of times taken for walkers to complete a circuit in a park is approximately normal, with a mean of 14 minutes and a standard deviation of 3 minutes.
- a** What percentage of walkers complete the circuit in:
 - i** between 8 and 20 minutes? **ii** less than 11 minutes? **iii** more than 20 minutes?
 - iv** less than 14 minutes? **v** less than 5 minutes? **vi** between 11 and 17 minutes?
 - b** In a week, 1000 walkers complete the circuit. How many of these are expected to take less than 8 minutes?
- 4** The distribution of heights of 19-year-old women is approximately normal, with a mean of 170 cm and a standard deviation of 5 cm.
- a** What percentage of these women have heights:
 - i** between 155 and 185 cm? **ii** more than 175 cm? **iii** more than 170 cm?
 - iv** less than 160 cm? **v** less than 165 cm? **vi** between 160 and 180 cm?
 - b** In a sample of 5000 of these women, how many are expected to have heights greater than 175 cm?
- 5** The distribution of rest pulse rates of 20-year-old men is approximately normal, with a mean of 66 beats/minute and a standard deviation of 4 beats/minute.

- a** What percentage of these men have pulse rates of:
- i** less than 66? **ii** more than 70? **iii** between 62 and 70 beats/minute?
iv less than 62? **v** between 58 and 74 **vi** less than 70?
- b** In a sample of 2000 of these men, how many are expected to have pulse rates between 54 and 78 beats/minute?

3.4 Standard scores

The **68–95–99.7%** rule makes the standard deviation a natural measuring stick for normally distributed data. For example, a person who obtained a score of 112 on an IQ test with a mean of 100 and a standard deviation 15 has an IQ score less than one standard deviation from the mean. Her score is typical of the group as a whole, as it lies well within the middle 68% of scores. In contrast, a person who scores 133 stands out; her score is more than two standard deviations from the mean and this puts her in the top 2.5%.

Because of the additional insight provided by being able to relate standard deviations to percentages, it is common to transform normally distributed data into a new set of units which show the number of standard deviations each data value lies from the mean of the distribution. This is called **standardising** and these transformed data values are called **standard** or **z-scores**.

Calculating standard (z) scores

To obtain a standard score for a data value, subtract the mean from the data value and then divide the result by the standard deviation.

That is:

$$\text{standard score} = \frac{\text{data value} - \text{mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{x - \bar{x}}{s}$$

Let us check to see that the formula works. We already know that an IQ score of 115 is one standard deviation above the mean, so it should have a standard or z-score of 1. Substituting in the above formula we find, as we had predicted:

$$z = \frac{115 - 100}{15} = \frac{15}{15} = 1$$

Standard scores can be both positive and negative:

- a **positive** z-score indicates the data value it represents lies **above** the mean
- a **zero** standardised score indicates that the data value is **equal** to the mean
- a **negative** z-score indicates that the data value lies **below** the mean

Example 4**Calculating standard scores**

The heights of a group of young women have a mean of $\bar{x} = 160$ cm and a standard deviation of $s = 8$ cm. Determine the standard or z -scores of a woman who is:

- a** 172 cm tall **b** 150 cm tall **c** 160 cm tall

Solution

- 1** Write down the data value (x), the mean (\bar{x}) and the standard deviation (s).

- 2** Substitute the values into the formula

$$z = \frac{x - \bar{x}}{s}. \text{ Evaluate.}$$

a $x = 172, \bar{x} = 160, s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{172 - 160}{8} = \frac{12}{8} = 1.5$$

b $x = 150, \bar{x} = 160, s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{150 - 160}{8} = -\frac{10}{8} = -1.25$$

c $x = 160, \bar{x} = 160, s = 8$

$$z = \frac{x - \bar{x}}{s} = \frac{160 - 160}{8} = \frac{0}{8} = 0$$

Using standard scores to compare performance

Standard scores are also useful for making comparisons across data distributions which have different means and/or standard deviations. For example, consider a student who obtained a mark of 75 in her psychology exam and a mark of 70 in her statistics exam. In which subject did she do better?

We could take the marks at face value and say that she did better in psychology because she got a higher mark in that subject. The assumption that underlies such a comparison is that the marks for both subjects have the same distribution with the same mean and standard deviation. However, in this case the two subjects have very different means and standard deviations as shown in the table below.

<i>Subject</i>	<i>Mark</i>	<i>Mean</i>	<i>Standard deviation</i>
Psychology	75	65	10
Statistics	70	60	5

If we assume that the **marks** are **normally distributed**, then **standardisation** and the **68–95–99.7%** rule gives us a way of resolving this issue.

Let us standardise the marks.

Psychology: standardised mark $z = \frac{75 - 65}{10} = 1$

Statistics: standardised mark $z = \frac{70 - 60}{5} = 2$

What do we see? The student obtained a higher score for psychology than statistics. However, relative to her classmates she did better in statistics.

- Her mark of 70 in statistics is equivalent to a z-score of 2. This means that her mark was two standard deviations above the mean, placing her in the top 2.5% of students.
- Her mark of 75 for psychology is equivalent to a z-score of 1. This means that her mark was only one standard deviation above the mean, placing her in the top 16% of students. This is a good performance, but not as good as for statistics.

Example 5**Applying standard scores**

Another student studying the same two subjects obtained a mark of 55 for both psychology and statistics. Does this mean that she performed equally well in both subjects? Use standardised marks to help you justify your answer and give a rating of her performance.

Solution

- 1 Write down her mark (x), the mean (\bar{x}) and the standard deviation (s) for each subject and compute a standardized score for both subjects.

Evaluate and compare.

- 2 Write down your conclusion.

$$\text{Psychology: } x = 55, \bar{x} = 65, s = 10$$

$$z = \frac{x - \bar{x}}{s} = \frac{55 - 65}{10} = \frac{-10}{10} = -1$$

$$\text{Statistics: } x = 55, \bar{x} = 60, s = 5$$

$$z = \frac{x - \bar{x}}{s} = \frac{55 - 60}{5} = \frac{-5}{5} = -1$$

Yes, her standardised score, $z = -1$, was the same for both subjects.

In both subjects she finished in the bottom 16%.

Exercise 3D

- 1 A set of scores has a mean of 100, and a standard deviation of 20. Standardise the following scores.

- a** 120 **b** 140 **c** 80 **d** 100 **e** 40
f 110 **g** 90 **h** 125 **i** 85 **j** 50

- 2 The table below contains the scores a student obtained in a practice test for each of his VCE subjects. Also shown is the mean and standard deviation for each subject.

- a** Calculate his standard score for each subject.
- b** Use the standard score to give a rating of his performance in each subject assuming a normal distribution of marks and using the 68–95–99.7% rule.

<i>Subject</i>	<i>Mark</i>	<i>Mean</i>	<i>Standard deviation</i>
English	69	60	4
Biology	75	60	5
Chemistry	55	55	6
Further Maths	55	44	10
Psychology	73	82	4

3.5 Populations and samples

Consider the following scenario: You have come up with a training strategy that you believe enables participants to score more highly on an IQ (Intelligence Quotient) test. You decide to set up a study to test the effectiveness of your training strategy. If your participants seem to benefit from the program, you would like to be able to say more than just that the program worked for these particular people. You would like to be able to generalise and say that it would be effective for everyone.

Likewise, if you have a treatment for the common cold that works for one group of patients, you would like to be able to conclude that it will work for all such patients. In statistics, the larger group of people or things that you would like to extend your conclusions to is called a **population**.

An early step in your study should be to determine exactly what you want your population to be. This is not always easy to do. For example, if you only wish to confine your study to students in your school, then your population is well defined. However, if you wanted to generalize to **all** students, you will have a much more difficult problem. Do we mean all students in the world, or just Australia, or even just Victoria? Are you interested in students at all levels, or just senior students, such as VCE students? With sufficient thought you are usually in a position to define with some precision the population of interest. However, unless the population is highly specialised, for example the VCE students in your school, it is either impossible and/or impractical to involve everyone in the population in your study. Imagine giving your intelligence improvement course to every VCE student in Victoria or even Melbourne!

What we do in practice is to conduct the study with a group of subjects known to belong to the population of interest. The group of subjects that you choose to work with in your study is called the **sample**.

Samples





























You can select a sample from a population in many different ways. Unfortunately, if you want to be able to generalise your results, not all samples are equally useful. The results of a study conducted with a group of VCE students, no matter how chosen, would not be appropriate if you wanted to make statements about the effectiveness of your intelligence training program for everybody. To be able to generalise, the sample you work with must be chosen in such a way that it can be regarded as being representative of the population as a whole.

Simple random sample (SRS)

The simplest way of obtaining a representative sample from a population is to select a **simple random sample (SRS)**. In an SRS, each member of the population has an equal chance of being selected, that is, no member of the population is systematically excluded from the sample, nor are any particular members of the population more likely to be included. Each member of the sample is also selected **independently**, that is, the selection of a member in no way influences the selection of another member.

Choosing a simple random sample (SRS)

To demonstrate the process of selecting an SRS we will focus on a hypothetical population of 100 VCE students who are studying at a secondary school in Victoria. The population is shown schematically below. Also shown for each student is the score obtained on a recently administered intelligence (IQ) test and an identity (ID) number, a two-digit number in the range 00 to 99. In this population there are 33 male students and 67 female students.

 00 IQ: 112	 01 IQ: 97	 02 IQ: 120	 03 IQ: 117	 04 IQ: 102	 05 IQ: 107	 06 IQ: 114	 07 IQ: 114	 08 IQ: 105	 09 IQ: 100
 10 IQ: 92	 11 IQ: 95	 12 IQ: 108	 13 IQ: 104	 14 IQ: 116	 15 IQ: 111	 16 IQ: 105	 17 IQ: 108	 18 IQ: 117	 19 IQ: 89
 20 IQ: 129	 21 IQ: 112	 22 IQ: 109	 23 IQ: 113	 24 IQ: 109	 25 IQ: 104	 26 IQ: 117	 27 IQ: 120	 28 IQ: 99	 29 IQ: 116
 30 IQ: 122	 31 IQ: 103	 32 IQ: 110	 33 IQ: 110	 34 IQ: 90	 35 IQ: 116	 36 IQ: 105	 37 IQ: 126	 38 IQ: 108	 39 IQ: 116
 40 IQ: 113	 41 IQ: 109	 42 IQ: 100	 43 IQ: 112	 44 IQ: 100	 45 IQ: 136	 46 IQ: 112	 47 IQ: 97	 48 IQ: 103	 49 IQ: 97
 50 IQ: 112	 51 IQ: 102	 52 IQ: 114	 53 IQ: 109	 54 IQ: 94	 55 IQ: 132	 56 IQ: 108	 57 IQ: 122	 58 IQ: 100	 59 IQ: 98
 60 IQ: 108	 61 IQ: 115	 62 IQ: 118	 63 IQ: 113	 64 IQ: 101	 65 IQ: 111	 66 IQ: 107	 67 IQ: 106	 68 IQ: 111	 69 IQ: 110
 70 IQ: 106	 71 IQ: 121	 72 IQ: 127	 73 IQ: 116	 74 IQ: 105	 75 IQ: 113	 76 IQ: 110	 77 IQ: 107	 78 IQ: 113	 79 IQ: 112
 80 IQ: 120	 81 IQ: 106	 82 IQ: 124	 83 IQ: 119	 84 IQ: 106	 85 IQ: 124	 86 IQ: 122	 87 IQ: 137	 88 IQ: 105	 89 IQ: 117
 90 IQ: 98	 91 IQ: 129	 92 IQ: 102	 93 IQ: 115	 94 IQ: 86	 95 IQ: 105	 96 IQ: 96	 97 IQ: 103	 98 IQ: 110	 99 IQ: 120

Population of 100 students, comprising 33 male and 67 female students, with their ID numbers and the scores they obtained on an IQ test

Selecting a sample

We could of course run the training program with all the students in the population of interest, but the training program is intensive so we decide to restrict the study to six students. The problem is to select at random six students from the population. One way of doing this is to write each person's ID number down on a piece of paper, put all the pieces of paper in a large bin, thoroughly mix them up, and then draw out six numbers. The six students whose numbers have been chosen would then constitute an SRS chosen from a population made up of the 100 VCE students.

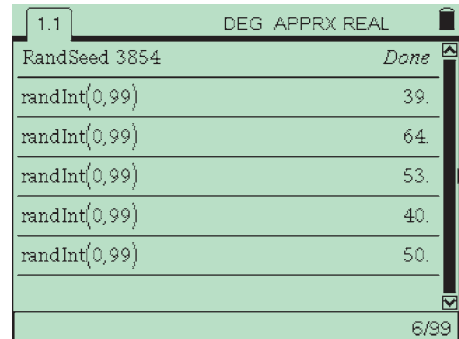
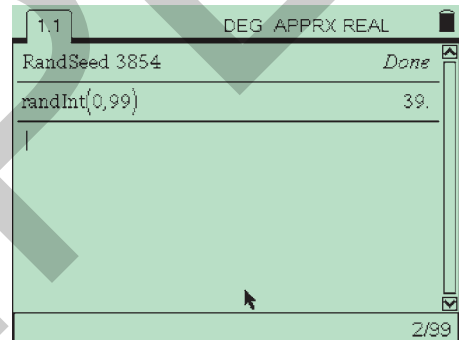
Another way is to use a graphics calculator to generate a set of six 2-digit random numbers.

How to generate a sequence of random integers using the TI-Nspire CAS

Generate a set of six random numbers between 00 and 99.

Steps


- 1 Start a new document: $\text{ctrl} + \text{N}$.
- 2 Select **1:Add Calculator**.
 - a Set the seed: press $\text{menu}/5:\text{Probability}/4:\text{Random}/6:\text{Seed}$ and type in any integer. (You could use the last four digits of your mobile number.) Press enter .
 - b Use the **randInt(·)** command to generate two-digit random integers between 0 and 99 (including 0 and 99).
Press: $\text{menu}/5:\text{Probability}/4:\text{Random}/2:\text{Integer}$ and type **0,99** inside the brackets, as shown. Press enter .
If your seed is different to the example shown, it is unlikely that your random integer will be the same as the one shown on the screen.
 - c Continue pressing enter to generate a sequence of two-digit random integers between 0 and 99.

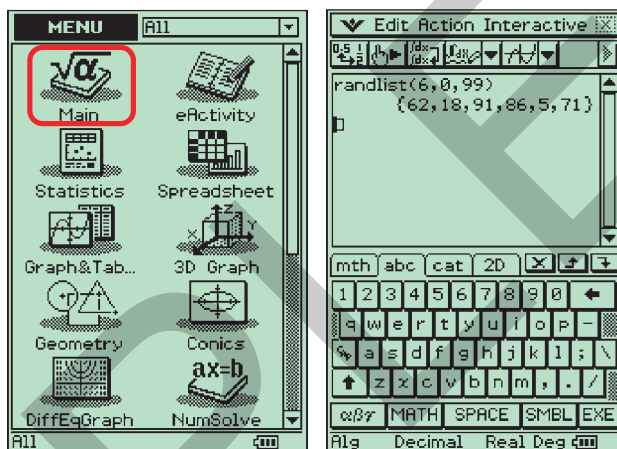


How to generate a sequence of random numbers using the ClassPad

Generate a set of six random numbers between 00 and 99.

Steps

- 1 Open the **Main** application by tapping  from the icon panel if it is not already visible.
- 2 To generate a list of random numbers:
 - a Press **Keyboard**, tap the **abc** tab and type **randlist(**.
 - b Complete the statement by typing:
 - **6**, (the number of random numbers)
 - **0**, (the starting integer value)
 - **99** (the finishing integer value), as shown.









Pressing **EXE** generates the required set of random numbers.

Note: By default a different list of random numbers is generated each time the command is executed.

Suppose we used a graphics calculator to generate six random numbers between 00 and 99, with the following result: 25 79 72 69 80 76

We could then use these numbers as ID numbers and use them to select at random six students from the population of 100 VCE students. The result is as shown below.

 25 IQ: 104	 79 IQ: 112	 72 IQ: 127	 69 IQ: 110	 80 IQ: 120	 76 IQ: 110
---	---	---	---	---	---

Sample 1

Population parameters and sample statistics

Population parameters

The mean IQ of the entire group of 100 VCE students is an example of a population parameter.

Population parameter

A **population parameter** is a number describing a population.

In this instance, because it is possible to list the IQ of all members of the population, it is also possible to calculate the population mean. This is not generally possible, and usually we either have to estimate or theorise the mean of the population we are studying. Greek symbols are used to represent population parameters. For the population mean we use the symbol μ (mu). In the population with which we have been working, $\mu = 110$. The key thing to note about a population parameter, such as the population mean, is that its value is **fixed** for **that** population.

Sample statistics

The mean IQ of six students chosen at random from a population is an example of a sample statistic. Roman letters are used to represent sample statistics. We use the familiar symbol \bar{x} to indicate we are working with a sample mean.

Sample statistic

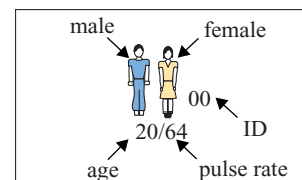
A **sample statistic** is a number that can be calculated from sample data.




























































Whereas the value of the population mean is fixed for a given population, the value of the sample mean \bar{x} will vary from sample to sample. For example, the mean IQ of the students in Sample 1 on the previous page is:

$$\bar{x} = \frac{104 + 112 + 127 + 110 + 120 + 110}{6} = 113.8$$

Exercise 3E

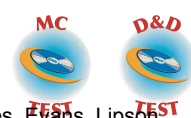
- Use a graphics calculator to select an SRS of ten people from the population displayed on page 74.
 - Work out the mean IQ score for this sample.
 - If you took a second sample of this size, would you expect the mean to be the same? Explain your answer.
- The population on page 77 is of 100 people, comprising 33 males and 67 females, who are shown with their ID numbers, ages (in years) and pulse rates (in beats per minute).



 00 20/64	 01 17/82	 02 23/65	 03 22/71	 04 29/74	 05 30/76	 06 23/73	 07 34/84	 08 18/69	 09 20/79
 10 27/80	 11 27/70	 12 34/72	 13 25/68	 14 30/74	 15 35/78	 16 19/73	 17 31/74	 18 26/79	 19 25/79
 20 24/66	 21 28/83	 22 21/67	 23 24/69	 24 30/78	 25 25/75	 26 29/76	 27 24/76	 28 24/69	 29 19/68
 30 25/73	 31 19/70	 32 23/70	 33 26/87	 34 23/73	 35 26/72	 36 27/74	 37 28/79	 38 23/69	 39 25/59
 40 19/78	 41 33/68	 42 28/73	 43 28/75	 44 23/70	 45 26/74	 46 32/78	 47 30/71	 48 29/89	 49 35/82
 50 26/68	 51 24/68	 52 28/83	 53 33/67	 54 21/84	 55 17/66	 56 31/75	 57 28/72	 58 20/68	 59 31/75
 60 34/76	 61 29/69	 62 22/65	 63 32/79	 64 25/87	 65 30/69	 66 35/78	 67 29/69	 68 18/60	 69 33/58
 70 25/70	 71 35/79	 72 30/58	 73 30/67	 74 20/68	 75 20/61	 76 23/69	 77 25/70	 78 34/78	 79 19/60
 80 23/76	 81 24/67	 82 26/75	 83 34/78	 84 20/57	 85 34/74	 86 29/69	 87 26/72	 88 25/85	 89 32/88
 90 25/70	 91 23/58	 92 29/67	 93 27/84	 94 24/78	 95 32/80	 96 18/65	 97 17/89	 98 29/84	 99 28/73

Use a table of random numbers to select an SRS of 12 people from the population displayed above.

- Record their ID number, sex, age and pulse rate (PR) in a table.
- For this sample, work out the:
 - mean age of the males
 - mean and standard deviation of the female pulse rates
 - median age and pulse rate of the males
 - percentage of females
- If you took a second sample of this size would you expect the values of the statistics you calculated above to be the same or different? Explain your answer.
- What is the mean age of the males in the population?
 - What symbol should you use to designate this mean, \bar{x} or μ ?



Key ideas and chapter summary

Summary statistics

Summary statistics are used to give numerical values to special features of a data distribution such as centre and spread.

The mean

The **mean**, the balance point of a data distribution, is a summary statistic which can be used to locate the **centre** of a **symmetric** distribution.

The mean \bar{x} is given by $\bar{x} = \frac{\Sigma x}{n}$, where Σx is the sum of the data values.

If the distribution is clearly skewed or there are outliers, the median is preferred to the mean as a measure of centre. This is because the value of the mean can be strongly influenced by one or two extreme values.

The standard deviation

The **standard deviation** is a summary statistic that measures the **spread** of the data values around the **mean**.

The **standard deviation** is given by $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

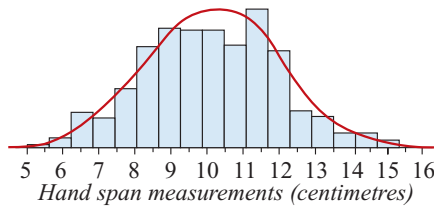
The value of the standard deviation is **estimated** by $\frac{\text{range}}{4}$

Means and standard deviations are usually evaluated using a calculator.

If the distribution is highly **skewed** or there are **outliers**, the **IQR** is preferred to the **standard deviation** as a measure of spread.

The normal distribution

Data distributions that have a bell shape can be modelled by a **normal** distribution.

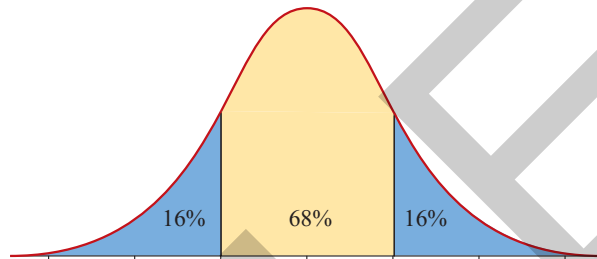


For normal distributions, the **68–95–99.7%** rule can be used to relate the mean and standard deviation to percentages in the distribution.

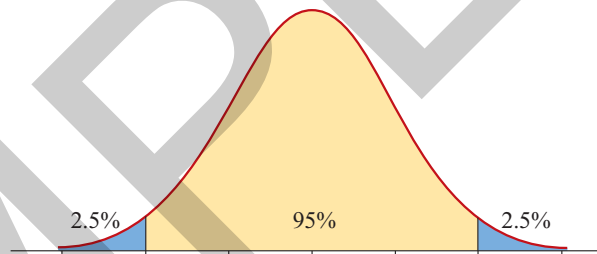
The 68–95–99.7% rule

The **68–95–99.7% rule** says that, for a normal distribution, approximately

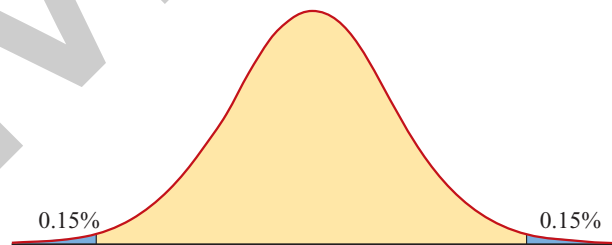
- 68% of the data values lie within 1 standard deviation of the mean.



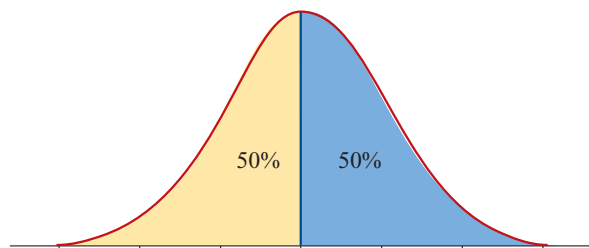
- 95% of the data values lie within 2 standard deviations of the mean.



- 99.7% of the data values lie within 3 standard deviations of the mean.

**Symmetry property**

Because the normal distribution is **symmetric**, **50%** of values lie **above** the mean and **50%** lie **below** the mean.



Standard (z) scores

Standardised or **z-scores** are calculated by subtracting the mean from each data value and then dividing by the standard deviation.

The formula for calculating standard scores is $z = \frac{x - \bar{x}}{s}$

Example: A student obtains a mark of 76 in a subject where the mean mark is 60 and the standard deviation is eight. The standardised score is:

$$z = \frac{x - \bar{x}}{r} = \frac{76 - 60}{8} = \frac{16}{8} = 2$$

The value of the standard score gives the **distance** and **direction** of a data value from the **mean** in **standard deviations**.

For example, if a data value has a standardised score of:

- $z = 2.1$ the data value is **two** standard deviations **above** the mean.
- $z = 0$ the data value is **equal** to the mean.
- $z = -1$ the data value is **one** standard deviation **below** the mean.

In combination with the 68–95–99.7% rule, standard scores can be used to give a measure of the level of performance.

For example, a student whose standardised score in a subject was

- $z = 2$ was in the top 2.5% of students in that subject
- $z = 0$ was exactly ‘average’ in that subject.
- $z = -1.2$ was in the bottom 16% of students in the subject.

Simple random sample (SRS)

In a **simple random sample** each member of the population has an **equal chance** of being selected.

Skills check

Having completed this chapter you should be able to:

- calculate the mean and standard deviation of a data set
- estimate the size of the standard deviation of a data set using: standard deviation $\approx \frac{\text{range}}{4}$, and use this estimate as a check when determining the standard deviation using a calculator
- understand the difference between the mean and the median as measures of centre and be able to identify situations where it is more appropriate to use the median
- know and be able to apply the 68–95–99.9% rule for bell-shaped distributions
- calculate standard or z-scores and use them to compare performance

Multiple-choice questions

The following information relates to Questions 1 to 3

The following is a set of test marks: 11, 1, 10, 15, 16, 25, 8, 10, 12

- 1 The mean value is:
 A 10 B 11 C 12 D 12.5 E 13
- 2 An estimate of the standard deviation (based on the range) is:
 A 2 B 3 C 4 D 5 E 6
- 3 Correct to two decimal places, the actual value of the standard deviation is:
 A 4.01 B 6.15 C 6.50 D 6.51 E 6.52
- 4 The mean of a data distribution is best described as:
 A the average B the middle value C the central value
 D the balance point E the middle 50% of values
- 5 It would not be appropriate to determine the mean and standard deviation of a group of people's:
 A age B phone numbers C height D weight E family size
- 6 The median is a more appropriate measure of the centre of a distribution than the mean when the distribution is:
 A symmetric B symmetric with no outliers C bell shaped
 D clearly skewed and/or there are outliers E normal
- 7 A student's mark on a test is 50. The mean mark for their class is 55 and the standard deviation is 2.5. Their standard score is:
 A -2.5 B -2.0 C 0 D 2 E 2.5
- 8 The 68–95–99.7% rule applies when a distribution is:
 A symmetric B positively skewed C negatively skewed
 D bell shaped E all of the above

In Questions 9 to 12, SD is used as an abbreviation for standard deviation

- 9 In a normal distribution, approximately 68% of values lie:
 A within one SD of the mean B within two SDs of the mean
 C within three SDs of the mean D more than one SD above the mean
 E more than two SDs below the mean
- 10 In a normal distribution, approximately 99.7% of values lie:
 A within one SD of the mean B within two SDs of the mean
 C within three SDs of the mean D more than one SD above the mean
 E more than two SDs below the mean
- 11 In a normal distribution, approximately 2.5% of values lie:
 A within one SD of the mean B within two SDs of the mean
 C within three SDs of the mean D more than one SD above the mean
 E more than two SDs below the mean

- 12** In a normal distribution, approximately 16% of values lie:
- A** within one SD of the mean **B** within two SDs of the mean
C within three SDs of the mean **D** more than one SD above the mean
E more than two SDs below the mean

The following information relates to Questions 13 to 15

The ages of a group of 500 first-year university students are approximately normally distributed with a mean of 18.4 and a standard deviation of 0.3 years.

- 13** The percentage of students with ages between 17.8 and 19.0 years is:
A 5% **B** 16% **C** 50% **D** 68% **E** 95%
- 14** The number of students with ages more than 18.4 years is:
A 25 **B** 80 **C** 250 **D** 340 **E** 475
- 15** The number of students with ages less than 18.1 years is:
A 25 **B** 80 **C** 160 **D** 340 **E** 475

Extended-response questions

- 1** The stemplot opposite shows the distribution of urbanisation rates (percentage) for 23 countries.
- a** From the shape of the distribution, which measure of centre, the mean or the median, do you think would best indicate the typical urbanisation rate in these countries?
- b** Calculate both the mean and median and check your prediction.

Urbanisation	
0	3 3 6 9 9 9
1	2 2 6 7
2	0 2 2 5 7 8 9
3	1 5
4	
5	4 6
7	
8	
9	9
10	0

- 2 a** The lifetimes (in hours) of 15 batteries were measured with the following results:
- 30 34 31 39 58 31 36 34 61 37 31 44 43 35 65
- What is a typical lifetime of the batteries measured? (Construct an appropriate stem plot to help you decide which measure of centre to use.)

- b** The following data was collected in an investigation of the typical amount of soft drink dispensed by an automatic filling machine.

<i>Fill number</i>	1	2	3	4	5	6	7	8	9	10	11
<i>Amount (millilitres)</i>	204	206	194	210	198	204	200	198	205	200	199

From this data, what would you say is the typical amount of drink dispensed by the machine? (Construct an appropriate stem plot to help you decide which measure of centre to use.)

- 3** The foot lengths (in cm) of a random sample of 13 students are shown below:

30.9 32.1 31.8 30.7 31.9 29.4 31.6 33.3 30.7 31.6 30.8 31.2 32.2

- a** Estimate the standard deviation for the foot lengths.
- b** Calculate the mean and standard deviation of the foot lengths (to 2 decimal places).
- c** Determine the median foot length. Compare the median foot length with the mean foot length. What does this comparison tell you about the distribution of foot lengths?
- 4** The average amount of life insurance sold per month by each salesman at a large company is \$100 000 with a standard deviation of \$17 500. If the distribution of the amount of insurance sold is known to be approximately normal then we can conclude:
- a** about 68% of the salesmen sold between and
- b** about 95% of the salesmen sold between and
- c** about 99.7% of the salesmen sold between and
- d** about 16% of the salesmen sold more than
- e** about 2.5% of the salesmen sold less than
- f** about 0.15% of the salesmen sold more than
- 5** The average number of cigarettes smoked per week by smokers in a certain state is 120 with a standard deviation of 10. If the distribution of the number of cigarettes smoked is known to be approximately normally distributed then we can conclude:
- a** about 68% of smokers smoked between and cigarettes
- b** about 95% of smokers smoked between and cigarettes
- c** about 99.7% of smokers smoked between and cigarettes
- d** about 2.5% of smokers smoked fewer than cigarettes
- e** about 50% of smokers smoked more than cigarettes
- f** about 16% of smokers smoked fewer than cigarettes

- 6 Some IQ tests are set so that on average, people taking the test score 100 points with a standard deviation of 15 points. IQ scores from this test are known to be approximately normally distributed. From this information we can conclude that:
- a almost all people taking the test will score between and
 - b if you scored 90 points your score would be above/below average
 - c if you scored between 85 and 115 you would be in the middle % of people taking the test
 - d 50% of people taking the test will score more than
 - e 99.85% of people taking the test will score more than
 - f 84% of people will score less than
- 7 The amount of time taken to serve a customer in a shop is approximately normally distributed with a mean of 18 seconds and a standard deviation of three seconds. From this information we can conclude that:
- a 99.7% (or almost all) customers will take between and seconds to serve
 - b % of customers will take less than 12 seconds to serve
 - c % of customers will take more than 21 seconds to serve
 - d % of customers will take more than 27 seconds to serve
 - e around two thirds of customers will take between and seconds to serve
 - f % of customers will take more than 15 seconds to serve
 - g % of customers will take less than 24 seconds to serve
 - h if a customer takes 20 seconds to serve, then he or she has taken above/below the average time to serve