Section 3.1.7 – Mean & Standard Deviation of a Sample

VCAA "Dot Points"
Investigating data distributions, including:

- *population and sample, random numbers and their use to draw simple random samples from a population or randomly allocate subjects to groups, the difference between population parameters (e.g., μ and σ), sample statistics (e.g., x and s).*

The Mean of a Sample

The **mean of a population** is represented by the Greek letter **μ.**
The **mean of a sample is** represented by $\bar{x}$**.**

Both μ and $\bar{x}$ are calculate in the same way.
NB: We will use $\bar{x}$ in this section of notes.

The mean ($\bar{x}$) is commonly referred to as the **average**.

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

Where $\bar{x}$ represents the mean
Σx represents the sum of all the values/numbers in the set
n represents the number of values/numbers in the set

Example 1
A survey of 10 students was carried out. Each participant was asked the following question: "How many people are in your family?"

The responses were as follows:

2 3 3 4 4 4 5 5 6 6

What is the mean of this distribution?
What is the median of this distribution?

$\bar{x}$ = ?
Σx = 2 + 3 + 3 + 4 + 4 + 4 + 5 + 5 + 6 + 6 = 42
n = 10

$$\bar{x} = \frac{\Sigma x}{n}$$
$$= \frac{42}{10}$$
= 4.2 people

Median (Q_2) = ?
Median score = $\frac{10+1}{2}$ = 5.5th score
= $\frac{4+4}{2}$ = 4 people

NB: there is very little difference in the median and mean for this particular set of data as it is a relatively symmetrical distribution.

Example 2

Nine (9) family members gathered for a family birthday. The ages of those in attendance, including the great grandfather, were as follows:

5 8 10 12 15 36 40 42 92

What is the mean of this distribution?
What is the median of this distribution?

$\bar{x}$ = ?

Σx = 5 + 8 + 10 + 12 + 15 + 36 + 40 + 42 + 92 = 260

n = 9

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

$$= \frac{260}{9}$$

= 28.9 years

Median (Q_2) = ?

Median score = $\frac{9+1}{2}$ = 5^{th} score

= 15 years

NB: there is a large difference in the median and mean for this particular set of data.

When the data sample are relatively symmetrically distributed, the mean can be used as a reasonable measure of central tendency.

However, when the data is skewed or effected by an obvious outlier (ie. 92 year old great grand father), the mean does not provide a good measure of the centre of distribution. In such data samples, the median should be used to provide a measure of central tendency.

Calculating the Mean of Grouped Data

When presented with **grouped data** in the form of a frequency table, the actual raw data is unknown and as such our previous technique for finding the mean of the sample cannot be used. For grouped data the following approach is used:

$$\boxed{\bar{x} = \frac{\Sigma(f \times \mathrm{m})}{\Sigma f}}$$

Where $\bar{x}$ represents the mean for the grouped data
f represents the frequency of the data
m represents the midpoint of the class interval of the grouped data

Example 3

Twenty five (25) teachers were surveyed about travel times to school by car. The distribution of the driving times (in minutes) from home to school for the teachers is shown in the table below:

Driving Times (minutes)	Number of Teachers
0 to less than 10	3
10 to less than 20	10
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean travel time for the 25 teachers. Give your answer to 3 significant figures.

Driving times (minutes)	Frequency (f)	Midpoint class interval (m)	$f \times m$
0 -	3	5	15
10 -	10	15	150
20 -	6	25	150
30 -	4	35	140
40 -	2	45	90
	Σf =25		$\Sigma(f \times m) = 545$

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

$$= \frac{545}{25}$$

= 21.8 minutes (3 significant figures)

The Standard Deviation of a Sample

The standard deviation (***s***) provides a measure of the data's spread. It is a numerical measure of how far each observation "**deviates**" from the mean.

- A **low standard deviation** indicates that the data tend to be **close to the mean** of the set
- A **high standard deviation** indicates that the data is **spread out over a wider range** of values

First let's consider the calculation of a quantity known as the **variance (*s*^{2})**.

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

Where s^2represents the variance
Σ represents the "sum of"
x represents an observation
$\bar{x}$ represents the mean
n represents the number of observations

NB: This equation is typically used when the data is a sub-set, or sample, of a larger population.

The standard deviation (s) = $\sqrt{s^2}$

Therefore the standard deviation (s) can be calculated using the following equation:

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

Example 4

Calculate the standard deviation (s) of the data from example 1.

The data was as follows: 2 3 3 4 4 4 5 5 6 6

$\bar{x}$ = 4.2

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	(2 - 4.2) = -2.2	4.84
3	(3 - 4.2) = -1.2	1.44
3	(3 - 4.2) = -1.2	1.44
4	(4 - 4.2) = -0.2	0.04
4	(4 - 4.2) = -0.2	0.04
4	(4 - 4.2) = -0.2	0.04
5	(5 - 4.2) = 0.8	0.64
5	(5 - 4.2) = 0.8	0.64
6	(6 - 4.2) = 1.8	3.24
6	(6 - 4.2) = 1.8	3.24
$\Sigma(x - \bar{x})^2$		15.6

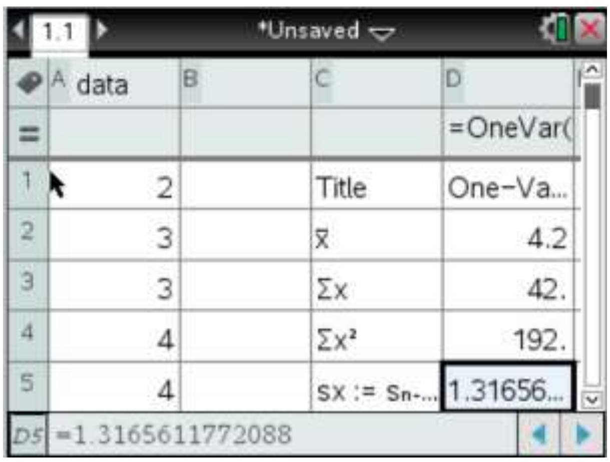
$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$$s = \sqrt{\frac{15.6}{10 - 1}}$$

= 1.31656

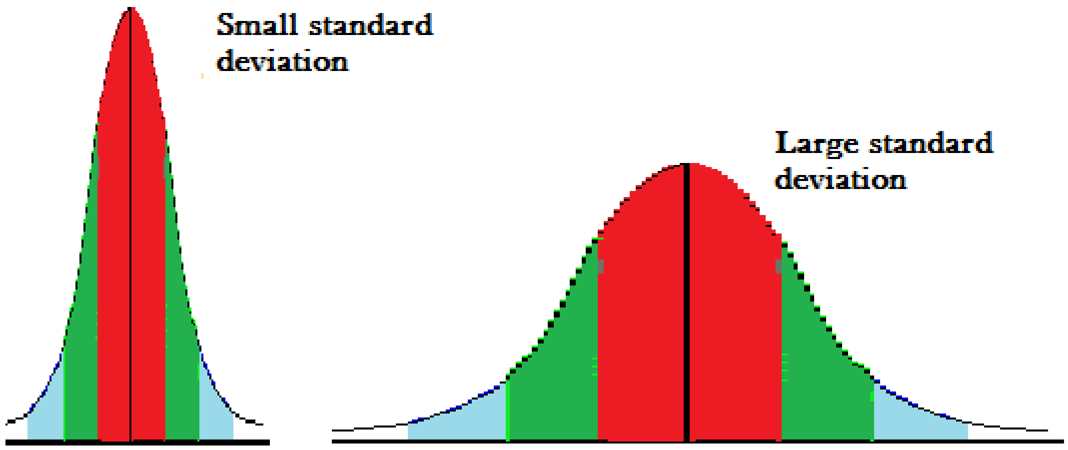
= 1.32 (3 significant figures)

Alternatively one can simply use the TI-nspire CX CAS calculator. Lists & Spreadsheet, menu option: Statistics/Stat Calculations/One-Variable Statistics



Comparing Standard Deviation

A **small standard deviation** indicates a **tightly clustered data**. Whereas, a **large standard deviation** a widely spread data. This can be visually demonstrated in the below diagrams.



Populations and Random Samples

Whenever statistical research is carried out the number of people surveyed has to be considered. For example, if the school wanted to survey students feedback upon the standard of canteen food served at schools, they may choose to survey the entire school **population**.

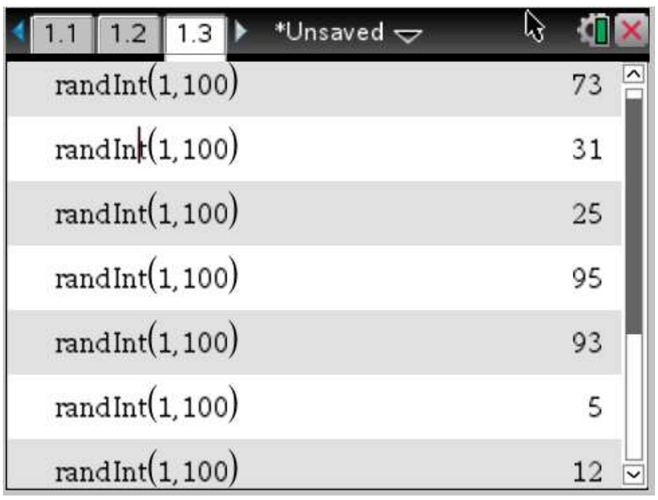
Alternatively, they may not have the resources necessary to survey the entire student population. In which case they may wish to survey a single campus, year level or indeed a class. Each of these three scenarios is an example of a **sample**.

Most surveys require a **random sample**. So selecting a sample of just one year level, gender or friendship group may produce a biased set of data. To generate a truly random sample from the school they may choose to draw a sample of 20 names from a hat.

Alternatively each student could be allocated a number and a calculator used to generate a **random number**, thereby selecting a 20 random students from the entire population.

Example 5
The function used by your TI-nspire CAS CX to generate a collection of random numbers from 1 to 100 is as follows:

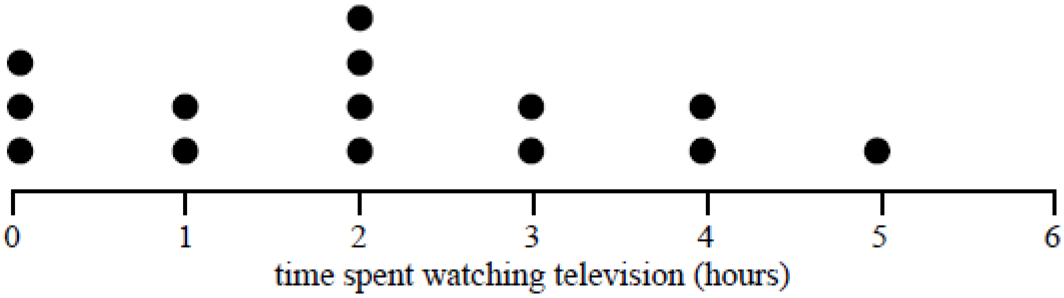
randInt(1,100)



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

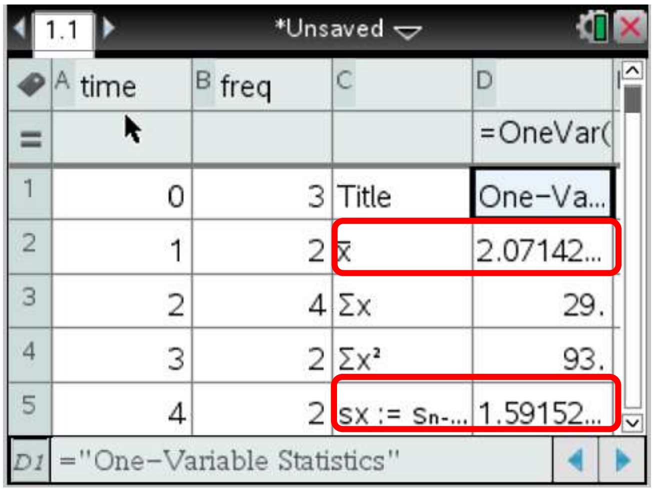
Example 6

A sample of 14 people were asked to indicate the time (in hours) they had spent watching television on the previous night. The results are displayed in the dot plot below.



Correct to one decimal place, the mean and standard deviation of these times are respectively

A. $\bar{x}$ = 2.0 s = 1.5
B. $\bar{x}$ = 2.1 s = 1.5
C. $\bar{x}$ = 2.1 s = 1.6
D. $\bar{x}$ = 2.6 s = 1.2
E. $\bar{x}$ = 2.6 s = 1.3

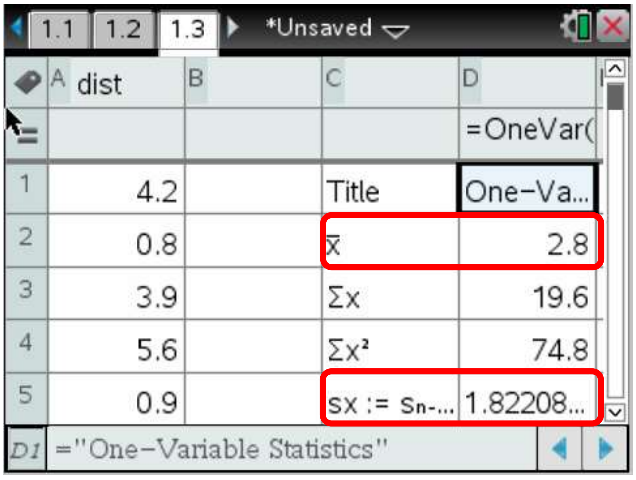


C

Example 7

The following table shows the distance that each driver travelled to the supermarket from their home.

Distance (km)
4.2
0.8
3.9
5.6
0.9
1.7
2.5



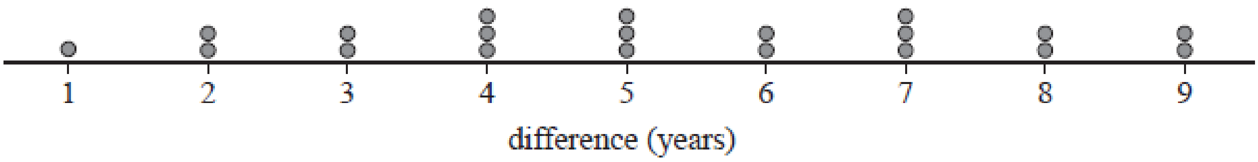
The mean, $\bar{x}$, and the stand

A. $\bar{x}$ = 2.5 s_x = 3.3
B. $\bar{x}$ = 2.8 s_x = 1.7
C. $\bar{x}$ = 2.8 s_x = 1.8
D. $\bar{x}$ = 2.9 s_x = 1.7
E. $\bar{x}$ = 3.3 s_x = 2.5

C

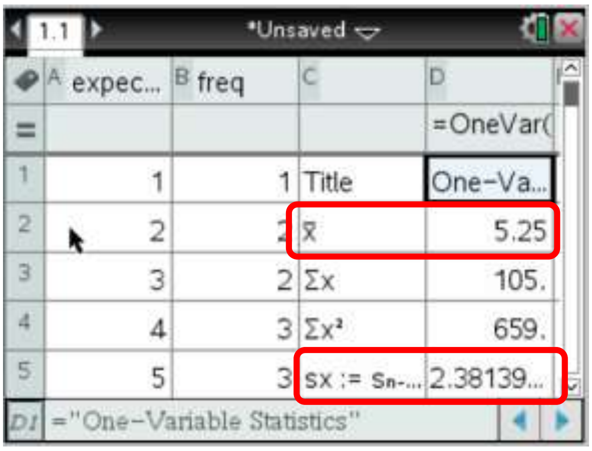
Example 8

The dot plot below displays the difference between female and male life expectancy, in years, for a sample of 20 countries.



The mean (x) and standard deviation (s) for this data are

A. mean = 2.32 standard deviation = 5.25
B. mean = 2.38 standard deviation = 5.25
C. mean = 5.0 standard deviation = 2.0
D. mean = 5.25 standard deviation = 2.32
E. mean = 5.25 standard deviation = 2.38



E

Exam Styled Questions (current study design) – Multiple Choice

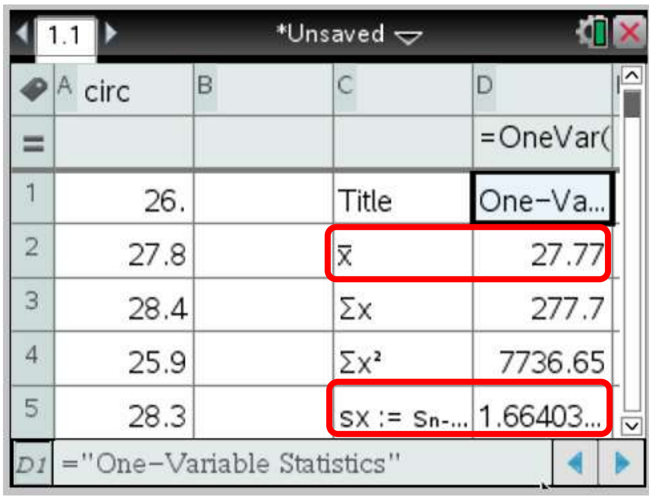
Question 1
(2017 Exam 1 Section A – Qn 3)

The table below shows the forearm circumference, in centimetres, of a sample of 10 people selected from this group of 252 people.

Circumference	26.0	27.8	28.4	25.9	28.3	31.5	28.2	25.9	27.9	27.8

The mean, $\overline{x}$, and the standard deviation, s_x, of the forearm circumference for this sample of people are closest to

A. $\bar{x}$ = 1.58 s_x = 27.8
B. $\bar{x}$ = 1.66 s_x = 27.8
C. $\bar{x}$ = 27.8 s_x = 1.58
D. $\bar{x}$ = 27.8 s_x = 1.66
E. $\bar{x}$ = 27.8 s_x = 2.30



D