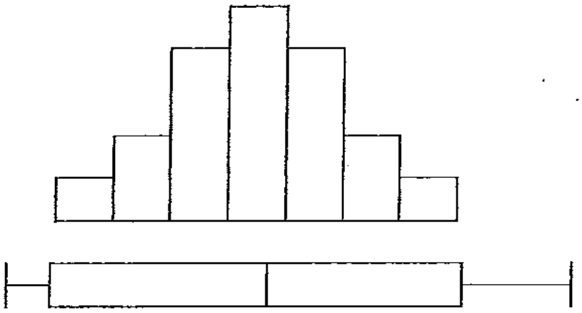
Data collection & statistics

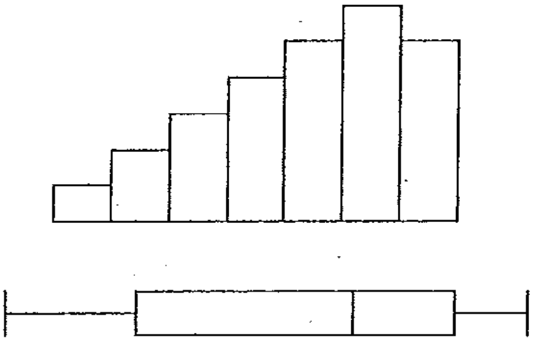
SKEWED DATA

The interpretation and analysis of data can be **comparative** (compared to other information) or **absolute** (with reference to itself).



Symmetrical Data

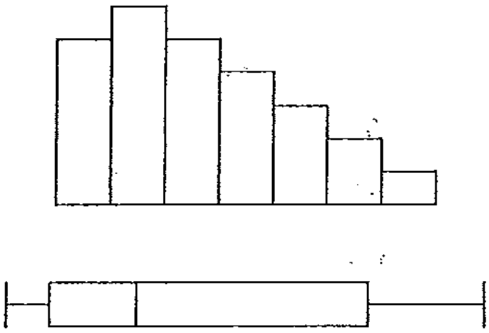
The median is in the middle of the data set.
In the box plot of the data the median is in the middle of the box.



Negative Skew

In the box plot the median is closer to the upper quartile and usually larger than the mean and less than the mode.

i.e. mean < median < mode



Positive Skew

In the box plot the median is closer to the lower quartile than the upper quartile.

Usually,
mode < median < mean

The length and position of the whiskers have no effect on the symmetry of the data.

STA

The n
the G
the st
devi

The r
scient
($\overline{x}$).
devia

STEP

Exam

Solut

Note

STANDARD DEVIATION

The most reliable measure of spread or dispersion is **standard deviation**, denoted by the Greek letter, σ. It tells us how the scores are spread around the mean. The higher the standard deviation, the more spread out the scores. The lower the standard deviation, the more densely grouped the scores are around the mean.

The method for calculating the standard deviation is rather complicated, however, scientific calculators calculate the standard deviation at the same time as the mean ($\bar{x}$). On a scientific calculator the symbol σ_n gives the **population standard deviation**, and σ_{n-1} gives the **sample standard deviation.**

STEPS:

(i) Switch calculator to Statistical Data Mode.

(ii) Ensure the calculator is clear of all statistics.

(iii) Enter each score using the $\boxed{\text{data}}$ key.

(iv) Pressing $\boxed{\bar{x}}$ gives the mean of the scores.

(v) Pressing $\boxed{\sigma_n}$ gives the population standard deviation.

(vi) Pressing $\boxed{\sigma_{n-1}}$ gives the sample standard deviation.

(vii) Pressing $\boxed{n}$ gives a check of the number of scores entered.

Example: Calculate the standard deviation for the following set of values

101 68 245 16 222 94 56 80 46

(a) When they are from an entire population,

(b) When they are a random sample of a population.

Solution: (a) Using the σ_n key on a scientific calculator, the population standard deviation is 73.930 (to 3 decimal places).

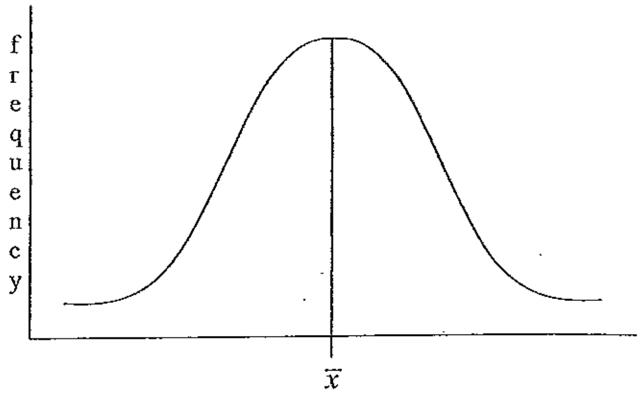
(b) Using the σ_{n-1} key, the sample standard deviation is 78.415 (to 3 decimal places).

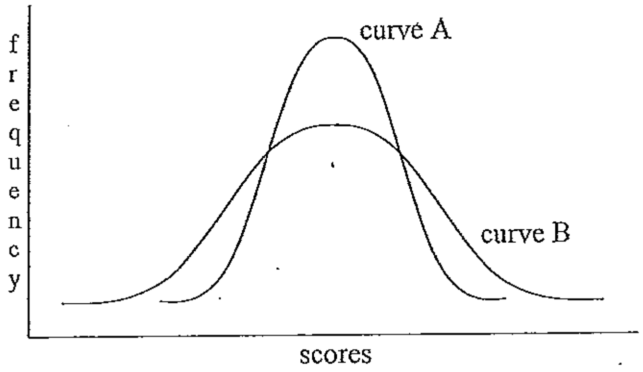
Note: σ_{n-1} provides a better estimate of the population's standard deviation, particularly when the number of scores is small. As the number of scores increases, the difference between σ_n and σ_{n-1} becomes negligible.

NORMAL DISTRIBUTION

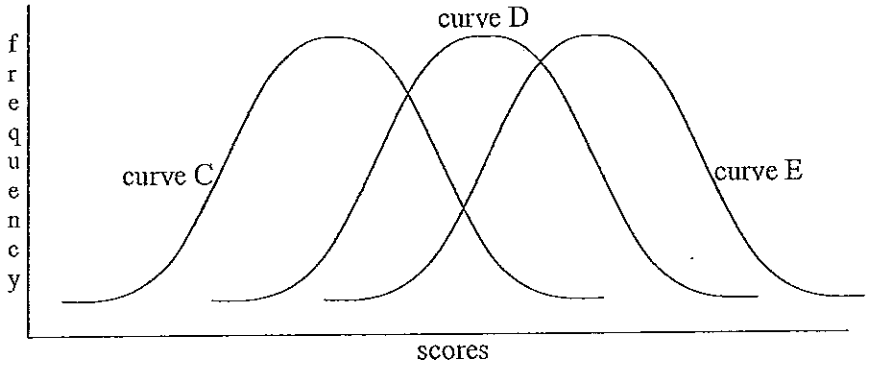
A **normal distribution** is one in which the scores are concentrated symmetrically around the mean. The mean, median and mode all have the same value. When a normal distribution of scores is represented in a frequency polygon, the graph has a bell-like shape and is known as a **bell curve**.

The **smaller** the standard deviation, the **thinner** and **taller** the bell is.
The larger the standard deviation, the wider and shorter the bell is.





The diagram shows two distributions with the same mean, median and mode, but curve B has a larger standard deviation than curve A.



In the above diagram, the curves have the same standard deviation, but the mean, median and mode for curve D are more than those for curve C but less than those for curve E.

For all normal distributions:

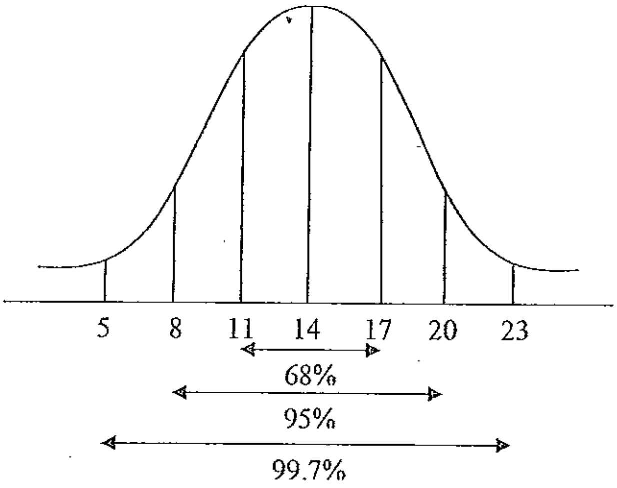
68% of the scores lie within 1 standard deviation of the mean
95% of the scores lie within 2 standard deviations of the mean
99.7% of the scores lie within 3 standard deviations of the mean

From the above, it can be seen that any score will ALMOST CERTAINLY lie within 3 standard deviations of the mean, and VERY PROBABLY lie within 2 standard deviations of the mean. MOST of the scores (more than two-thirds) lie within 1 standard deviation of the mean.

Example: A normal distribution has a mean of 14 and a standard deviation of 3. What percentage of scores:

(a) lie between 11 and 17
(b) lie between 8 and 20
(c) are greater than 17
(d) are less than 5

Solution: It always makes the question easier to answer if a normal curve is drawn first with the standard deviations as shown.



(a) between 11 and 17 there are 68% of the scores
as $14 - 3 = 11$ and $14 + 3 = 17$

(b) between 8 and 20 there are 95% of the scores
as $14 - (2 \text{ x } 3) = 8$ and $14 + (2 \text{ x } 3) = 20$

(c) 68% of scores lie between 11 and 17.
$\therefore$ 32% of scores lie either side of 11 and 17 (100%-68%)
$\therefore$ 16% of scores are greater than 17.

(d) 99.7% of scores lie between 5 and 23.
$\therefore$ 0.3% of scores lie either side of 5 and 23 (100%-99.7%)
$\therefore$ 0.15% of scores are less than 5.

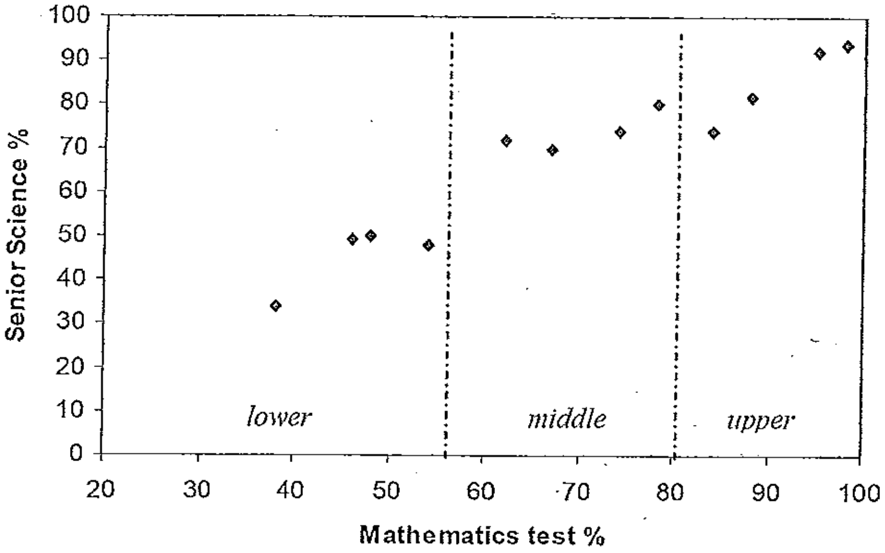
MEDIAN REGRESSION LINE

The recommended method for constructing a line of best fit is to draw a **median regression line.** This involves dividing the data into three groups and finding the median (or summary point) of each group.

Example: The following are the Mathematics and Senior Science test marks for a group of Year 12 students.

Maths	84	88	54	62	98	67	78	74	95	48	38	46
Science	74	82	48	72	94	70	80	74	92	50	34	49

Plot the data and draw the median line of regression.



To draw the Line of Regression:

Step 1: With vertical lines, split the data into three groups. $12 \div 3 = 4$points

Lower region:	(38, 34)	(46, 49)	(48, 50)	(54, 48)
Middle region:	(62, 72)	(67, 70)	(74, 74)	(78, 80)
Upper region:	(84, 74)	(88, 82)	(95, 92)	(98, 94)

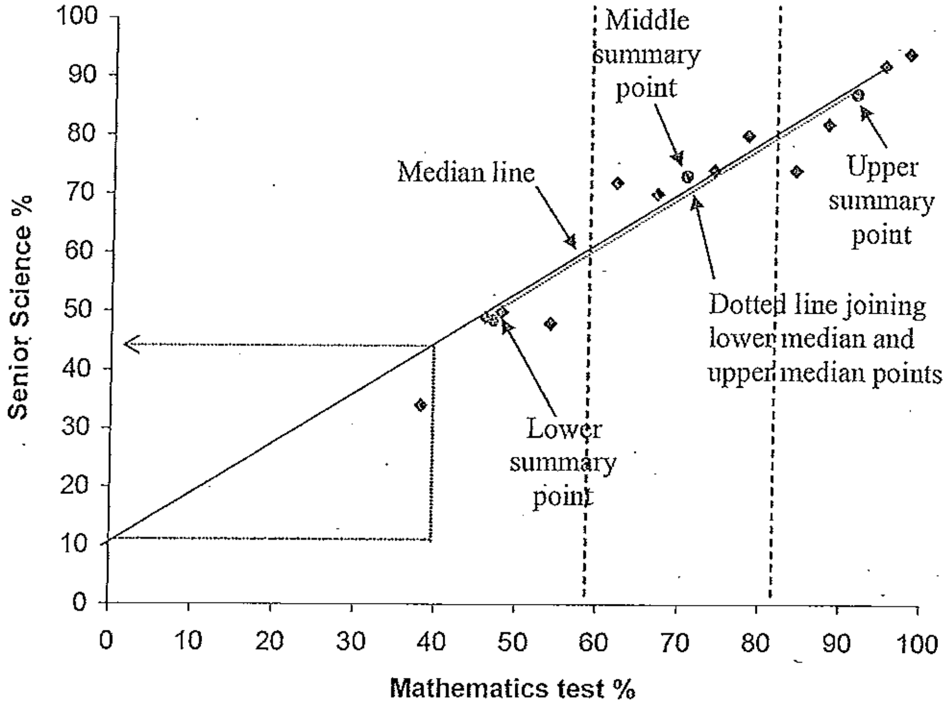
Step 2: Find the medians of the x-values and the y-values in each region.

Lower region x: 38 46 48 50 median of x is 47
(reordered) y: 34 48 49 50 median of y is 48.5 (47, 48.5)

Middle region x: 62 67 74 78 median of x is 70.5
(reordered) y: 70 72 74 80 median of y is 73 (70.5, 73)

Upper region x: 84 88 95 98 median of x is 91.5
y: 74 82 92 94 median of y is 87 (91.5, 87)

Step 3: Plot the three points. Lightly join the **lower summary point** to the **upper summary point** to form the line of regression.



Step 4: Move the line of regression ⅓ of the distance towards the middle median point.

Once the median line of regression has been constructed the y-intercept can be determined by extrapolating (extending) the line to the y-axis.

In this example, the y-intercept is 10.

Using a right-angled triangle, drawn below the median line of regression, the gradient can be calculated.

$$\text{Gradient} = \frac{\text{rise}}{\text{run}} = \frac{35}{40} = \frac{7}{8}$$

Equation of median line of regression: $y = mx + b \quad \Rightarrow \quad y = \frac{7}{8}x + 10$

REVIEW EXERCISE – LEVEL 1 DATA COLLECTION

1. Calculate mean, mode and standard deviation for the following sets of scores:

   (a) 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6

   (b) 85, 52, 62, 68, 79, 68, 46, 72, 81

   (c)

x	3	4	5	6	7	8
f	8	6	1	5	9	2

2. Represent the sets of scores in question 1 in separate box-and-whiskers plots.

3. (a) Organise these data into a frequency distribution table.

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

   (b) Calculate the mean and standard deviation.

   (c) Draw a frequency histogram and polygon.

4. The following were scores in a History test.

87	78	52	52	55	91	92	90	44	50	69	42
74	52	40	59	95	87	46	70	19	45	28	46

   (a) Using class intervals 0-9, 10-19, 20-29, … etc, organise the data into a frequency distribution table.

   (b) Use class centres to calculate the mean. Compare with the mean of the actual scores.

   (c) Draw a cumulative frequency ogive and find the median and interquartile range.

5. In a Maths test, the mean was 60 and the standard deviation was 12.

   (a) What mark corresponds to a z-score of: (i) 1 (ii) – 1.5 (iii) 3?

   (b) What z-scores correspond to marks of: (i) 84 (ii) 48 (iii) 66?

   (c) What percentage of scores, approximately, will be between 24 and 96?

   (d) What percentage of scores, approximately, are greater than 84?

6. Rebecca is averaging 60 in 4 English tests. What must she score in the fifth test to average 65?

REVIEW EXERCISE – LEVEL 2

1. Display the following sets of scores in a back-to-back stem-and-leaf diagram.

A:	12	15	18	19	19	20	24	26
B:	12	13	15	19	19	23	25	26

Which set of scores has the greatest: (i) standard deviation

(ii) interquartile range

(iii) range

2. When Real Estate agents talk about average house prices, they often use the median house price rather than the mean house price. Why?

3. The Smiths have 2 children whose average age is 8. The Brown's three children have an average age of 13. What is the average age of all 5 children?

4. (a) In the first innings of a cricket game, Nick has figures of 4 wickets for 252 runs. Dean has 1 wicket for 84 runs. Who has the better average?

(b) In the second innings, Nick takes 3 wickets for 84 while Dean takes 7 foe 252. Who has the better average?

(c) Which bowler has the better match figures (overall average)?

5. Ms. King decides that her students' scores on a test were too high.

(a) If she decides to reduce every mark by 10, what effect will this have on the mean and standard deviation?

(b) If she decides to halve each score, what effect will this have on the mean and standard deviation?

(c) Ms. King's top student's mark had a z-score of 2. If the mean was 75 and the standard deviation was 8.5, what did her top student score?

6. The following are the Chemistry and Physics test marks for a group of Year 12 students.

Chemistry	74	78	44	52	84	60	70	54	85	38	22	46
Physics	69	78	43	67	92	62	68	56	82	45	29	44

(a) Plot the data and draw the median line of regression.

(b) Determine the gradient and equation of the line of regression.

(c) Use the line of regression to estimate a mark for Joanna who missed the Physics test and scored 50 in the Chemistry test.