UNIT: DATA ANALYSIS

(i) Interpreting Sets of Data

Measures of Location

Measures such as the **mean** and **median** give us information about the location of scores. For example, if we know that the mean of a set of exam marks is 92% we immediately have a sense of what sort of marks the students in the class scored.

The ***mean*** of a set of scores is commonly known as the ***average***. It is found by adding all the scores together, then dividing this sum by the number of scores.

Example 1:
Find the mean of 3, 7,9,15 and 21.

ANS: 11

The symbol $\bar{x}$ is used for the **mean.**

We can use the formula $\bar{x} = \frac{\Sigma x}{n}$ where x is the individual score and n is the number of scores.

Σ is the Greek letter sigma. It means **'sum'.**

The ***median*** is the middle score when the scores are **arranged in order from lowest to highest** (or highest to lowest).

If there is an **even number of scores**, there will be **two middle scores.** The median is the **average of those two middle scores.**

Example 2:
Find the median of 7, 5, 3, 2 and 8.

ANS: 5

Example 3:
Find the median of 2, 9, 1, 8, 4, 1, 7, 8, 5 and 6.

ANS: 5.5

The ***mode*** is the score with the highest frequency. ***(The score that occurs most often.)***

Example 4:
Find the mode of the scores: 3, 6, 8, 2, 6, 1, 2, 8, 6, 5.

ANS: 6

Further calculations with the mean

Example 5:
The mean of a set of eight scores is 26. Another score of 32 is added to the set. What is the new mean?

ANS: $26\frac{2}{3}$

A ***frequency distribution*** table can be used to simplify working when there are many scores. The *fx* column is where each score (*x*) is multiplied by its frequency (*f*).

The mean will therefore be given by $\bar{x} = \frac{\Sigma fx}{\Sigma f}$

Example 6:
The following frequency distribution table has been prepared from the goals scored in each game in a particular soccer team's last season.

x	f	fx
0	5	
1	4	
2	6	
3	3	
4	1	
Total		

(a) Complete the table.
(b) Find the mean number of goals scored per game to one decimal place.

ANS: 1.5

Example 7:
The following table has been drawn up to show the number of people who scored different marks in a short quiz.

Mark	4	5	6	7	8	9	10
Frequency	1	3	7	6	9	5	3

(a) What is the mode?
(b) How many people took part in the quiz?
(c) Draw up a CUMULATIVE frequency distribution table and use it to find the mean, to one decimal place.
(d) What is the median mark?

ANS: (a) 8 (b) 34 (c) 7.4 (d) 7.5

Example 8:

The table shows the number and price of tickets sold for a concert.

Price of tickets	Number sold
$10	42
$20	37
$50	15
$100	6

(a) How many tickets were sold altogether?
(b) What is the mean price of tickets sold?
(c) What is the modal price?
(d) What is the median?
(e) What percentage of tickets were sold for less than the mean?
(f) A report after the show stated the median price but not the mean. Why might this have been?

ANS: (a) 100 (b) $25.10 (c) $10 (d) $20 (e) 79%

(f) The median gives the middle value. The mean is higher because of the small number of higher priced tickets. The median is more representative of the price paid by most people.

Measures of spread

The ***range, interquartile range*** and ***standard deviation*** are ***measures of spread.*** They give us information about ***whether the scores are close together or spread far apart.***

The **range** of a set of scores is the **difference between the highest and lowest score.**

Example 9:
Find the range of these scores:
7, 2, 6, 8, 5, 3, 8, 4, 9, 5.

ANS: 7

Interquartile range

Quartiles divide the data into **four** parts. The **lower quartile** is the number for which **25% of the scores are lower**. The **upper quartile** is the number for which **25% of the scores are higher.** The median is the middle quartile.

The median divides the scores into two halves.

The lower quartile is the median of the bottom half and the **upper quartile is the median of the top half.**

The interquartile range is the difference between the upper and lower quartiles.

It gives us information about the spread of the middle 50% of the scores.

Interquartile range = upper quartile - lower quartile

Example 10:
The lower quartile of a set of scores is 19 and the upper quartile is 41. What is the interquartile range?

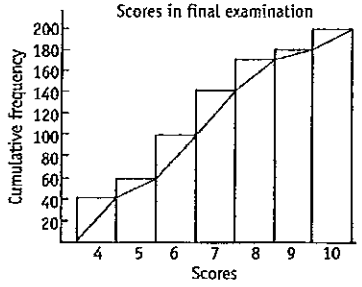
ANS: 22

Example 11:
For the scores 3 3 4 4 4 5 6 6 6 7 8 9
find:
(a) the lower quartile
(b) the upper quartile
(c) the interquartile range.

ANS: (a) 4 (b) 6.5 (c) 2.5

Quartiles can be read from an ogive (cumulative frequency polygon).

Example 12:
A cumulative frequency histogram and polygon have been drawn for a set of scores.



(a) What is the median?
(b) What is the upper quartile?
(c) What is the lower quartile?
(d) What is the interquartile range?

ANS: (a) 6.5 (b) 8 (c) 5 (d) 3

Standard Deviation

The ***standard deviation*** is another measure of the spread of scores. It gives us information ***about the position of most scores in relation to the mean.*** The **greater** the standard deviation, the ***further the scores are spread from the mean.***

The standard deviation can be calculated with a calculator after entering the scores in exactly the same way as is done to find the mean. The standard deviation varies slightly depending on whether the scores are for the whole population or for just a sample of the population.

The **population** standard deviation is σ_n.

The **sample** standard deviation is σ_{n-1}. [We usually use this]

Example 13:

An exam was given to all the students in a particular **year**. The marks scored by the students in **Ms Hope's class** are:

78 56 94 59 65 73 81 75 86 69 73

What is the standard deviation (σ_{n-1}) to two decimal places?

ANS: **11.25**

Example 14:

The members of a **class** were given a quiz and all the marks are given below:

8, 7, 6, 7, 5, 6, 9, 8, 10, 9
6, 6, 7, 5, 6, 4, 8, 8, 7, 9

What is the standard deviation (σ_n) to one decimal place?

ANS: **1.5**

Example 15:

Two classes both sat for the same test. The results, out of 50, were:

Class X: 37, 48, 39, 41, 27, 43, 44, 46
32, 35, 45, 38, 29, 44, 33, 40, 31

Class Y: 42, 36, 36, 45, 40, 41, 39, 45, 29
32, 36, 35, 37, 43, 30, 36, 38, 41

(a) Find the mean and standard deviation (σ_{n-1}) for both sets of scores. (Give the answers correct to one decimal place.)

(b) In which class were the scores closer to the mean?

ANS: (a) **CLASS X Average = 38.4, S.D = 6.3** **CLASS Y Average = 37.8, S.D = 4.7**

(b) **Class Y has scores closer to the mean because S.D is lower for class Y than for class X.**

Outliers

Outliers are **scores** that are much **higher** or much **smaller** than **all the rest of the scores.**

Example 16:

The scores (out of 25) gained by the members of a class in a test are given below:

1 16 17 19 19 20 21 21 21
21 22 22 22 23 23 24 24 25

(a) What is the mean to one decimal place?
(b) What is the standard deviation (σ_{n-1}) to one decimal place?
(c) What is the mode?
(d) What is the median?
(e) What is the range?
(f) John was ill and had to leave the room just after the test started and only scored one mark. Recalculate the mean, standard deviation, mode, median and range if John's score is ignored.

ANS: **(a) 20.1 (b) 5.3 (c) 21 (d) 21 (e) 24 (f) new mean = 21.2 new σ_{n-1} = 2.4**

From the above question we can see that, as you would expect, *the measures of spread (range and standard deviation), are affected strongly by an outlier.*

The *mode* will not be affected by an outlier, and the *median* will only be marginally affected, if at all. *The mean, however, can be greatly affected*.

Example 17:

The costs, in dollars, of casual labour to a small business are recorded for a number of weeks:

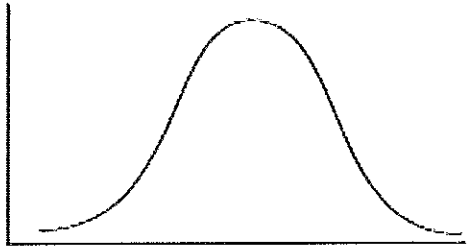
2015 1827 1955 2006 1978 1850
2019 1896 1921 9373 1945 1997

Ignoring any outliers, find the mean weekly cost of labour.

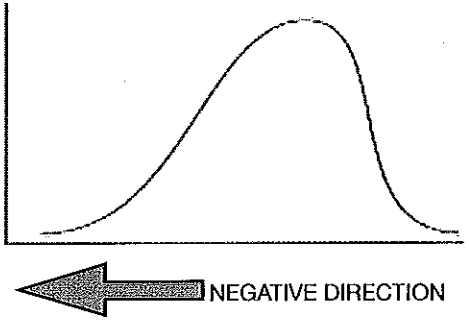
ANS: **Outlier = 9373** **The mean weekly cost of labour is $1946, to the nearest dollar.**

The Shape of the Display

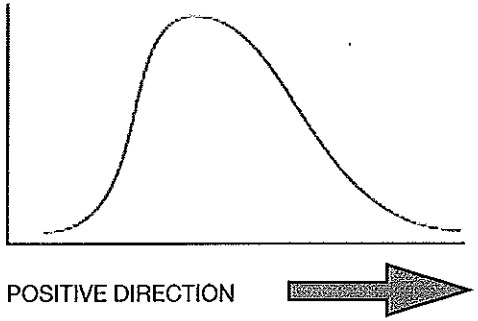
Type i: Normal Distribution (Not Skewed) / A symmetric bell curve.



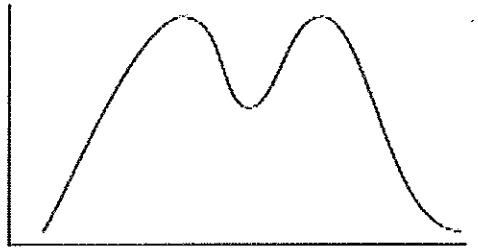
Type ii: Negatively skewed bell curve



Type iii : Positively skewed bell curve.

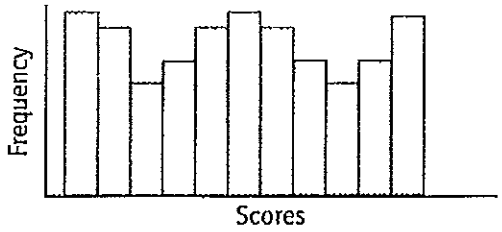


Type iv : bimodal curve (has two high points)



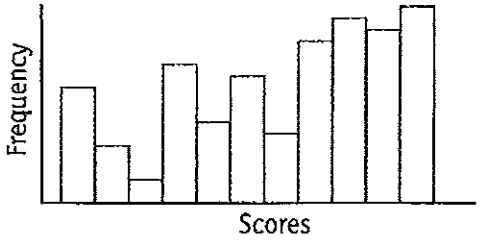
Example 18:
Briefly describe the shape of each of the histograms in terms of smoothness, symmetry and number of modes.

(a)



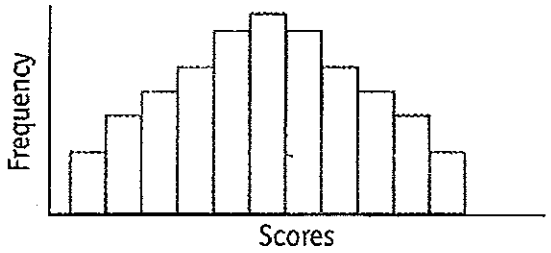
ANS: **The graph is reasonably smooth, and is symmetrical. There are three modes.**

(b)



ANS: **The graph is not smooth or symmetrical. It is skewed with more higher scores at the right, (negatively skewed). There is just one mode.**

(c)

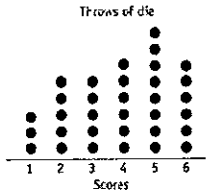


ANS: **This graph is smooth, symmetrical and has just one mode (at the mean).**

Example 19:
Ethan tossed a die a number of times and the results are given below.

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

(a) Draw a dot-plot of the results.
(b) What information can be gained about the data from the dot-plot?



ANS: **The mode is 5. The scores are negatively skewed. There are more high scores than low ones.**

Stem-and-leaf plots

A stem-and-leaf plot is a display that allows us to show all the scores in a data set in a simple way even if the range of the data is quite large. The stem is the first part of a number and the leaf is the last part.

Example 20:

A stem-and-leaf plot has been drawn to show the results of a class test.

Test results

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

(a) What is the range?
(b) What is the mode?
(c) What is the median?
(d) Briefly comment on the features of the display.

ANS: **(a) 48** **(b) 77** **(c) 71.5** **(d) The display is fairly smooth and symmetrical.**

Double (back-to-back) stem-and-leaf

In a double stem-and-leaf plot, **two** data sets are shown at the same.This allows us to easily observe some of the similarities and differences in the data.

Example 22:

Two classes sat for the same test (out of 50) and the Test results results are shown in the back-to-back stem-and-leaf.

Test results

Class N		Class Q
9	0	
9 6	1	7
7 7 4 2	2	3 8 9
9 9 8 5 2 0 0 0	3	1 4 4 4 6 7
7 6 4 3 1	4	0 2 2 5 6 8 8 9
	5	0

(a) Which class had the greater range of scores?
(b) What is the median score for each class?
(c) Alan is in class Q. He said, (my class did better than class N.' Do you agree? justify your answer.

ANS:
(a) class N=38 class Q= 33 **(b) class N=31 class Q= 37**
(c) Yes, Alan is correct. The scores in both classes are skewed but class Q has more higher scores and the median mark is higher.

Example 23:

Forty students sat for an exam and the results (out of 100), for boys and for girls are listed below.

Boys: 61 78 65 45 46 53 59 64 68 73
82 62 66 65 58 49 54 64 71 57

Girls: 56 68 65 62 59 75 70 66 74 45
78 82 69 61 71 68 77 72 63 74

(a) Draw up a back-to-back stem-and-leaf plot of the results.

(b) The teacher believed there was a significant difference between the results for the boys and for the girls. Do you agree? justify your answer.

(a)

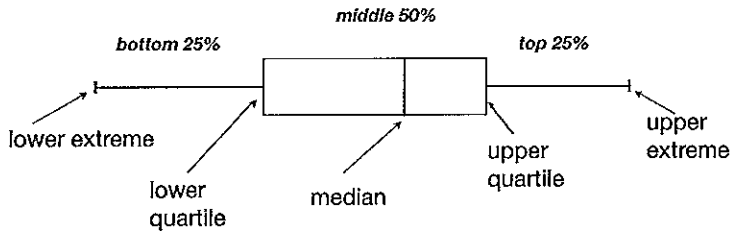
Exam results

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

(b) On the whole the girls do appear to have performed slightly better on the test than the boys. The range of scores is the same for both groups but the median mark for the girls is 68.5, 5.5 marks higher than the median mark for the boys (63). The girls marks are skewed towards the top whereas the boys marks are more symmetrical.

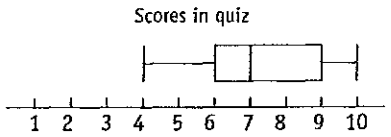
Box-and-whisker plots

A box-and-whisker plot gives an indication of the spread of scores. It uses the **five-number summary** (lower extreme, lower quartile, median, upper quartile and upper extreme).



Example 24:

A box-and whisker plot has been drawn to illustrate the scores out of ten by pupils in a quick quiz.

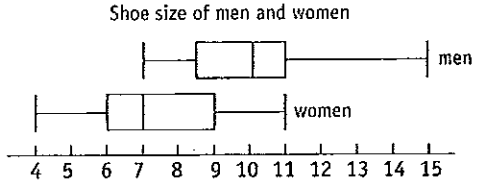


(a) What was the median mark?
(b) What was the range of marks?
(c) What is the interquartile range?
(d) Approximately what percentage of the pupils scored 6 or more in the quiz?

ANS: **(a) 7** **(b) 6** **(c) 3** **(d) 75%**

Example 25:

Information was collected about the shoe size of the men and women attending a conference. The results are depicted in the box-and-whisker plots below.



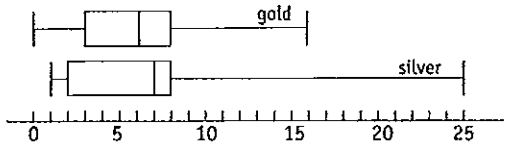
ANS: What similarities and differences are there between the two data sets? The sizes for men are generally larger than for women. The range for men is greater than for the women and the median is 3 sizes larger. The sizes for women are fairly symmetrically placed while those for men are a little skewed.

Example 26:
From London in 1948 until Sydney in 2000, the number of gold and silver medals gained by Australia at the Olympic games was studied. The five number summary was given for each type of medal.

Gold: [0, 3, 6, 8, 16]
Silver: [1, 2, 7, 8, 25]

(a) Draw two box-and-whisker plots on the same scale for the data.
(b) Briefly comment on any similarities and differences between the two data sets.

(a)



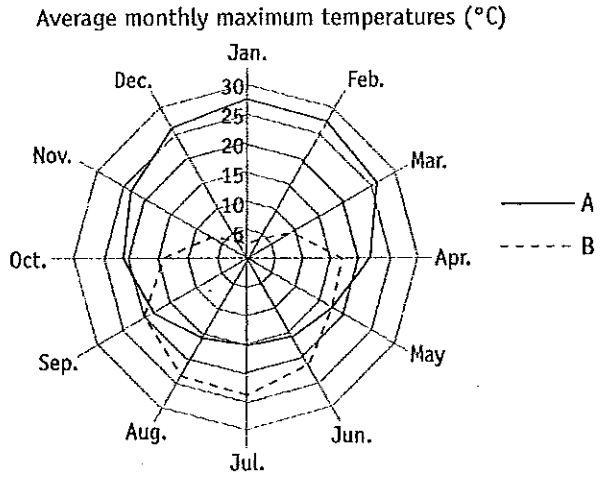
(b) Silver has a greater range and a greater interquartile range. The upper quartile is the same for both. The gold medals are more symmetrical. The silver medals are skewed. The bottom 75% are more closely spaced than the remaining 25% and the median is much closer to the upper quartile than the lower quartile.

Radar Charts

A radar chart is a graph drawn on a **circular** grid.

Example 27:

This radar chart shows the average monthly maximum temperatures for two cities A and B.



(a) What is the average monthly maximum temperature for city A in March?
(b) In what month are the average monthly maximum temperatures of the two cities the same?
(c) One city is in the southern hemisphere and one is in the northern hemisphere. Which city is in the northern hemisphere? Justify your answer.
(d) What other information can readily be seen from the graph?

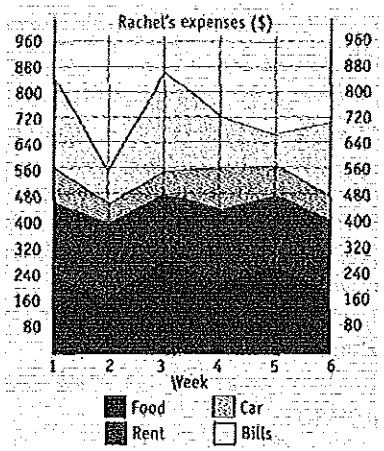
ANS: **(a) 26 C (b) May**
(c) City B is in the northern hemisphere. It has its coldest average monthly maximum temperatures in December, January and February whereas city A has its hottest average monthly maximum temperatures then.
(d) City A has a much milder climate than city B. City A has average monthly maximum temperatures ranging between 15 C and 28 C while city B has average monthly maximum temperatures that range between 2 C and 24C.

Area charts

An area chart is another display that allows us to quickly see information and compare different data sets.

Example 28:

Rachel is trying to work out a budget and recorded the amount she spent on food, rent, her car and the bills she had to pay for six weeks. She showed the results in an area chart.



(a) In which week did Rachel spend the least on the total of these expenses?
(b) If Rachel earns $960 per week, how much would she have had left after paying these expenses in week 1?
(c) One expense was the same every week. What is that expense and how can we see that it is the same?
(d) What do you notice about the food expenses?

ANS:
(a) Wk 2
(b) Total expenses in week 1 = $840 Left over = $960 - $840 = $120 Rachel would have $120 left over in week 1.
(c) Rent
(d) The food expenses are between about $150 and $260 per week. The food expenses tend to fluctuate, up one week, down the next.