TYPES OF DATA AND THEIR INTERPRETATION

Comparing sets of data includes looking for patterns or associations between the sets. The method you choose to compare data sets depends on the type of data.

Data can be divided into two basic categories, numerical and categorical, both of which can be further divided into sub-categories.

Numerical data

Discrete data: Each piece of data is one of a set you can write down so that the number of possibilities can be counted, e.g. the number of children in a family.

Continuous data: The possible outcomes cannot be written down as all decimal places are possible. You usually have to measure the outcomes to a set number of decimal places, e.g. the time for the 100 m sprint at the Olympics is measured to the nearest hundredth of a second.

Categorical data (non-numerical)

Ordinal: Ordered in some way, e.g. rating a statement as strongly disagree, disagree, neutral, agree or strongly agree.

Nominal: There is no order, e.g. recording people's favourite search engine.

Comparing numerical variables

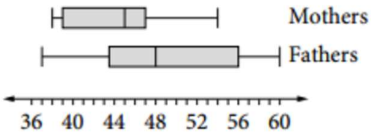
A visual display of this type of data can be done using a parallel box plot or a back-to-back stem-and-leaf plot.

For example, the same information concerning the age of Year 12 parents attending a Year 12 information evening is represented by both these displays.

Back-to-back stem-and-leaf plot:

Fathers	Stem	Mothers
9 7	3	8 8 9
8 7 5 3 2	4	2 3 5 6 6 7
8 8 4 3 1	5	2 4
0	6	

Parallel box plots:



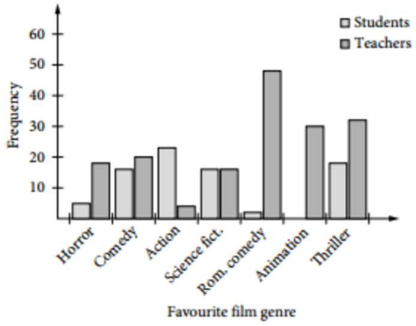
Parallel box plots can be used to compare more than two data sets. The back-to-back stem-and-leaf plot can only be used when there are two sets of data.

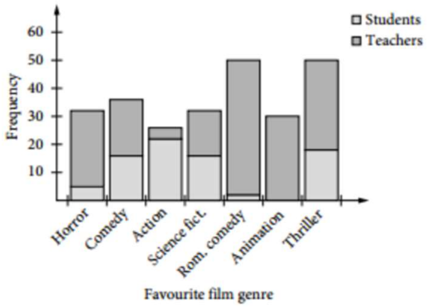
Comparing two categorical variables

The simplest way to compare two sets of nominal categorical data is to draw a back-to-back frequency table such as the following.

Students	Favourite film genre	Teachers
5	Horror	18
16	Comedy	20
22	Action	4
16	Science fiction	16
2	Romantic comedy	48
0	Animation	30
18	Thriller	32

This information can also be shown in a clustered column graph or a stacked column graph.





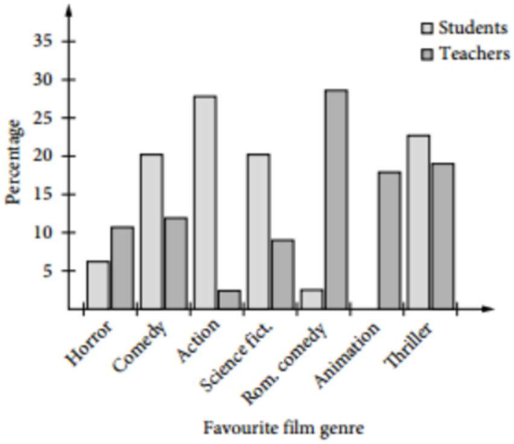
The unequal number of students and teachers in the survey (the total number of entries in each student and teacher column) makes it difficult to see any difference or similarity that may exist in the answers for the two groups.

In this case, converting the data to percentages may show any differences or similarities in the groups.

The table then becomes:

% Students	Favourite film genre	% Teachers
6.3	Horror	10.7
20.3	Comedy	11.9
27.8	Action	2.4
20.3	Science fiction	9.5
2.5	Romantic comedy	28.6
0	Animation	17.9
22.8	Thriller	19.0

And the column graph becomes:



The differences between the percentages are more clearly seen here. The students surveyed preferred 'Action' and 'Science fiction', whereas the teachers surveyed preferred 'Romantic comedy' and 'Animation'.

These graphs may also be drawn using a spreadsheet.

Example 1

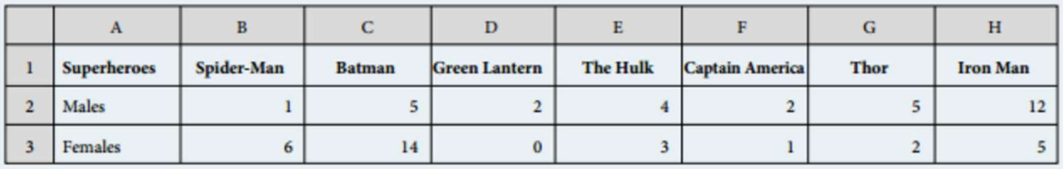
The favourite superheroes of a group of 62 people are shown in the table below.

	Spider-Man	Batman	Green Lantern	The Hulk	Captain America	Thor	Iron Man
Men	1	5	2	4	2	5	12
Women	6	14	0	3	1	2	5

Using a spreadsheet, draw the clustered column graph for the information given in the table.

Solution

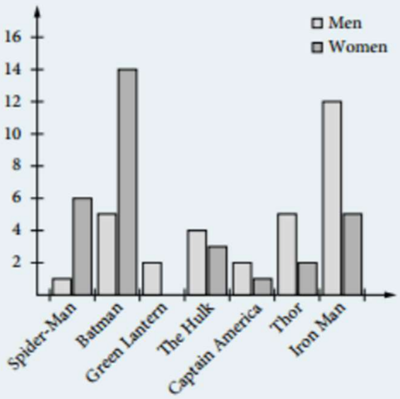
Step 1: Enter the data into the spreadsheet, keeping the same headings.



	A	B	C	D	E	F	G	H
1	**Superheroes**	**Spider-Man**	**Batman**	**Green Lantern**	**The Hulk**	**Captain America**	**Thor**	**Iron Man**
2	Males	1	5	2	4	2	5	12
3	Females	6	14	0	3	1	2	5

Step 2: Select Insert > Column > 2D cluster. (The exact selection may vary depending on spreadsheet software, but it should be available as a kind of column graph.)

Step 3: A graph like the following should be produced.



Note: There are other useful graphs (such as the segmented bar graph) that can be drawn by selecting a different graph output.

Two-way frequency tables

Consider the results of surveying 200 shoppers about a proposal to change the arrangements for childcare at a shopping centre.

	Male	Female	Total
Agree	25	88	113
Disagree	45	42	87
Total	70	130	200

This is called a two-way frequency table as it shows not just the ‘agree’ and ‘disagree’ totals but breaks down the figures into subcategories as well.

From the table it can be seen that of the 200 shoppers, 70 were male and 130 were female.

In total, 113 shoppers agreed with the proposal and 87 disagreed.

While it can be seen that more females (88) than males (25) agree with the proposal, the numbers disagreeing appear approximately the same.

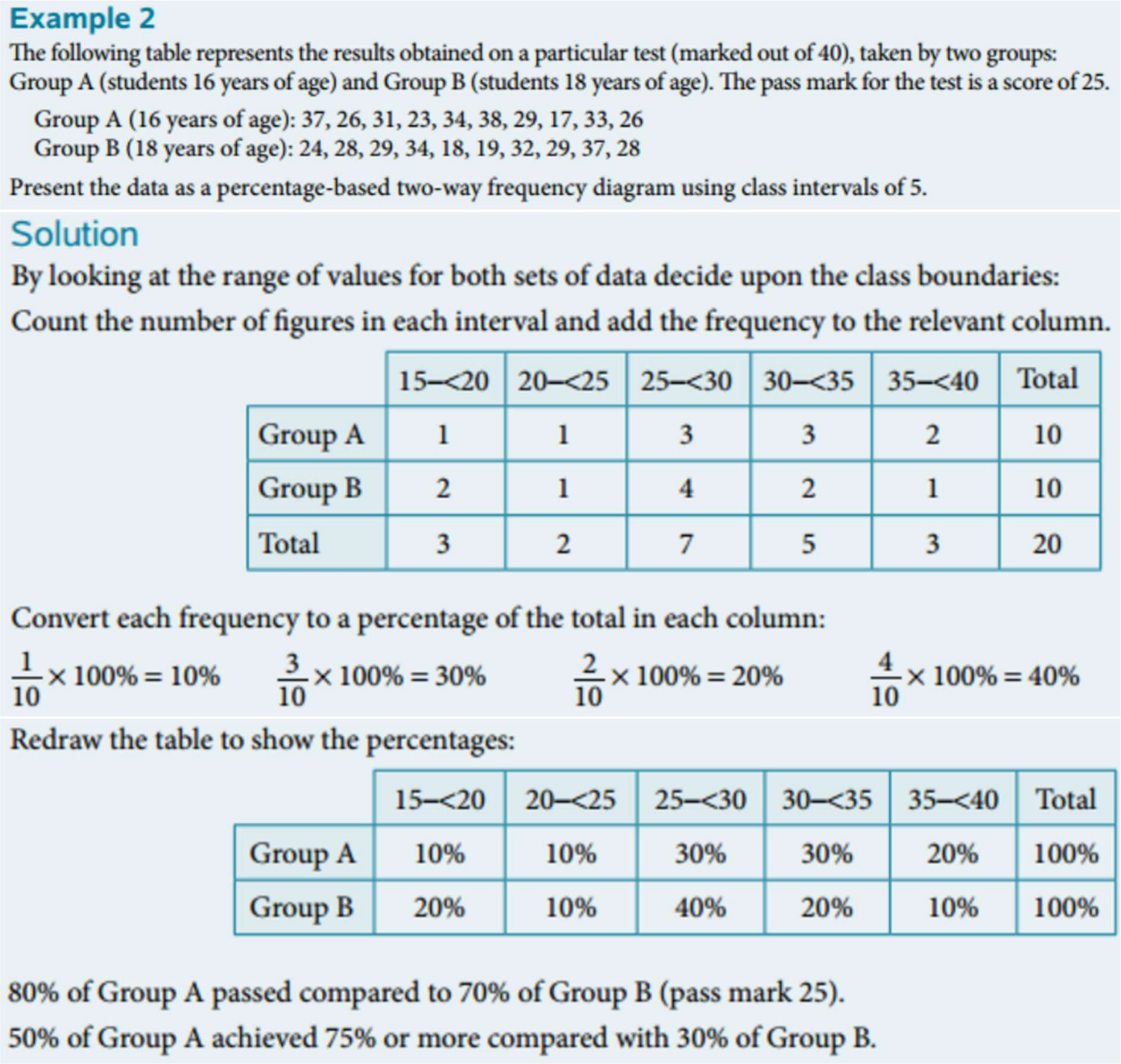
In this case, the gender of the person being interviewed may influence their response but a person’s response will not influence their gender. The variable influencing a response is called an independent variable. The independent variable is given in the vertical columns of the two-way table. The variable reacting to the independent variable is called the response variable, or dependent variable.

Since the number of males and females interviewed is different, the trend will become clearer if you convert the figures to percentages. Remove the last column (since this refers to the dependent variable), and convert each of the figures into a percentage of the total in the last row of the column.

	Number of men	Percentage	Number of women	Percentage
Agree	25	$\frac{25}{70} \times 100 = 35.7\%$	88	$\frac{88}{130} \times 100 = 67.7\%$
Disagree	45	64.3%	42	32.3%
Total	70	100%	130	100%

	Male %	Female %
Agree	35.7	67.7
Disagree	64.3	32.3
Total	100	100

Now the data shows more clearly that a woman is almost twice as likely as a man to agree with the proposal.



Example 2

The following table represents the results obtained on a particular test (marked out of 40), taken by two groups: Group A (students 16 years of age) and Group B (students 18 years of age). The pass mark for the test is a score of 25.

Group A (16 years of age): 37, 26, 31, 23, 34, 38, 29, 17, 33, 26
Group B (18 years of age): 24, 28, 29, 34, 18, 19, 32, 29, 37, 28

Present the data as a percentage-based two-way frequency diagram using class intervals of 5.

Solution

By looking at the range of values for both sets of data decide upon the class boundaries:

Count the number of figures in each interval and add the frequency to the relevant column.

	15–<20	20–<25	25–<30	30–<35	35–<40	Total
Group A	1	1	3	3	2	10
Group B	2	1	4	2	1	10
Total	3	2	7	5	3	20

Convert each frequency to a percentage of the total in each column:

$\frac{1}{10} \times 100\% = 10\%$ $\quad \frac{3}{10} \times 100\% = 30\%$ $\quad \frac{2}{10} \times 100\% = 20\%$ $\quad \frac{4}{10} \times 100\% = 40\%$

Redraw the table to show the percentages:

	15–<20	20–<25	25–<30	30–<35	35–<40	Total
Group A	10%	10%	30%	30%	20%	100%
Group B	20%	10%	40%	20%	10%	100%

80% of Group A passed compared to 70% of Group B (pass mark 25).

50% of Group A achieved 75% or more compared with 30% of Group B.