Data

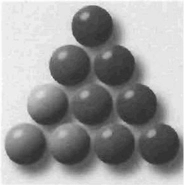
Data is information collected according to some principle.

We can make some statements about the 'quality' of our data in terms of our ability to carry out arithmetical operations.

Scales

Nominal Scale: this is where data is simply in terms of names or descriptions.

Example: colours red, blue, green would be nominal data.



Ordinal Scale: this is where data can be recognised as being in some order.

Example: a collection of names might be ordered alphabetically.

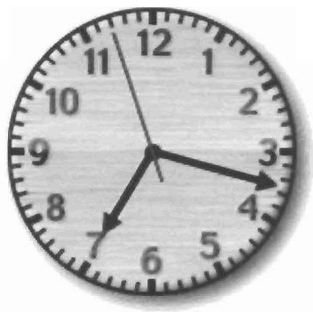
Interval Scale: this is where the gaps between whole numbers on the scale are equal. This permits the arithmetic operations of addition and subtraction.

Example: temperature

20° C is not twice as hot as 10° C because 0° is not an absolute zero.
It is the amount of heat beyond which water turns from solid to liquid.

Ratio Scale: a ratio scale permits full arithmetic operation.

Example: the time something takes is an example of using a ratio scale



If a train journey takes 2 hr and 35 min, then this is half as long as a journey which takes 5 hr and 10 min.

Discrete and continuous measures with a ratio scale

A full ratio scale can be modelled on a number line. A number line is **continuous** which means that there are no gaps between the numbers.

However, empirically we can only measure things with a degree of inaccuracy. That means there are gaps in our measure, and this we refer to as **discrete**.

Examples: continuous measures: time, height, weight
discrete measures: the number of cars in a car park, your last maths mark.

Frequency tables

Data are usually displayed on a **frequency table**. The observed different data are wrote on a column (row) and ordered if it is possible.

The number of times each datum $\mathbf{x_i}$ has been observed is called its **absolute frequency** $\mathbf{f_i}$ and it is wrote besides (below).

The number of observations or data is called **size of the sample n**.

Examples:

- A survey was done by asking n=200 students for their shoe size.
The results were:

39, 40, 41, 44, 36, 40, 42, 42, 39, 41, 42, 44, 41, 39, 40, 43, 42,
41, 37, 38 40, 40, 39, 42, 40, 41, 41, 38, 40, 39, 40, 37, 41, 40,
42, 39, 41, 40, 44, 4042, 44, 41, 39, 40, 43, 42, 41, 37, 38, 44,
40, 41, 44, 36, 40, 42, 42, 39, 41,40, 37, 41, 40, 42, 39, 40, 41,
44, 40, 36, 41, 39, 42, 40, 41, 41, 38, 40, 39,40, 43, 42, 41, 37,
38, 44, 40, 41, 41, 38, 40, 39, 40, 43, 42, 41, 39, 43, 44,43, 42,
41, 41, 38, 40, 39, 41, 43, 42, 41, 37, 38, 35, 40, 41, 42, 44, 39,
37,39, 40, 40, 44, 40, 36, 41, 39, 42, 36, 41, 41, 38, 43, 42, 39,
43, 37, 38, 36,40, 42, 42, 43, 44, 39, 37, 38, 35, 40, 41, 44, 42,
43, 41, 40, 42, 39, 40, 4341, 41, 38, 41, 38, 40, 39, 40, 43, 42,
41, 40, 43, 42, 41, 37, 38, 35, 40, 41,43, 39, 42, 44, 38, 39, 43,
41, 37, 44, 42, 39, 41, 38, 42, 37, 41, 39, 42, 38

We can organise our data by placing them in order.

Size (x_i)	35	36	37	38	39	40	41	42	43	44
Frequency (f_i)	3	6	17	12	25	37	40	29	16	15

- A frequency table on how n=100 students in a college found out about the Resources Centre

Source of information (x_i data)	Number of students (f_i absolute frequencies)
Class induction	49
Told by a teacher	20
Attended last year	10
LRC publicity	6
Told by friend	3
Other/no response	12

Note: A frequency table provides all the information but it takes a bit of work to see the main features, so it is useful to draw some sort of **chart or diagram**.

Grouped frequency

When data is collected in broad categories we need to fix a value to stand for the group, the **mid point** or **mid-interval** of it.

The precise values where you pass from one group into the next are called the **class boundaries.**

Examples:
- 109 recent graduates were asked about their annual incomes the data was recorded in 3000€ bands we can set the mid point of each band as 6500€, 9500€, 12500€ and so on

Income	Frequency (f_i)	Mid point (x_i)
5000-7999	3	6500
8000-10999	6	9500
11000-13999	14	12500
14000-16999	26	15500
17000-19999	35	18500
20000-23999	11	22000
24000-26999	8	25500
27000-29999	3	28500
30000-32999	2	31500
33000-35999	1	34500

- the distribution of weights of a bunch of 60 school kids:

Weight (kg)	31 — 40	41 — 50	51 — 60	61 — 70	71 — 80
Frequency	8	16	18	12	6

The class boundaries are 40.5, 50.5, 60.5 and 70.5
The mid-points are 35.5, 45.5, 55.5, 65.5

Cumulative frequency

Cumulative frequency is used to determine the number of observations that lie above (or below) a particular value in a data set.

The cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total.

Example: The number of daily visitors to a museum recorded over a 30-day period:
31, 49, 19, 62, 24, 45, 23, 51, 55, 60, 40, 35 54, 26, 57
37, 43, 65, 18, 41, 50, 56, 4, 54, 39, 52, 35, 51, 63, 42

Interval	Frequency (f)	Cumulative frequency
0-9	1	1
10-19	2	$1+2=3$
20-29	3	$3+3=6$
30-39	5	$6+5=11$
40-49	6	$11+6=17$
50-59	9	$17+9=26$
60-69	4	$26+4=30$

Measures of Central Tendency or Central Location

Descriptions of a set of data tend to suggest some notion of centrality. It's what is most, or in the middle. We use three main types of measure (all of them suitable for ratio scale data):

Mode (Mo): the value that occurs with the highest frequency (also suitable for nominal scale data).

Median (Me): the middle datum, once the data are arranged in order of size (also suitable for ordinal scale data). If there is an even number of values then the median is the mean of the middle pair. We can make use of our cumulative frequency table to find out the median.

Mean ($\bar{x}$): the average value or "equal shares" that is the sum of all the data divided by the number of them (also suitable for interval scale data). That is

$\bar{x} = \frac{1}{n}\sum_i x_i \cdot f_i$ being $n = \sum_i f_i$ the number of data.

Note: One of the problems with the mean is that it gives a result of a division, but does not necessarily exist in any real sense. It is only a measure, a descriptor of a set.
Note: Other disadvantage of the mean is that it can be "skewed" by large outliers.

The **weighted mean** is typically used when working out an index number. Data 'items' are multiplied by a 'weight', added, and divided by the total weight.

Examples: ■ A survey is carried out amongst n=60 school pupils about their preferred brand of trainers. The results are shown in the table:

Brand (x_i)	A	B	C	D	E
Frequency (f_i)	5	20	17	3	15

Brand B is the most popular. The mode is B.

■ A survey was done by asking n=200 students for their shoe size:

Size (x_i)	35	36	37	38	39	40	41	42	43	44
Frequency (f_i)	3	6	17	12	25	37	40	29	16	15
Cumulative (Fi)	3	9	26	38	63	100	140	169	185	200

The middle term will therefore lie between 100 and 101.
The 100th term is 40, the 101st term is 41. Half way between these is 40.5
The median is 40.5

■ Find the mean of these results obtained by throwing a dice.

score	1	2	3	4	5	6
frequency	18	17	23	20	24	18

$$\bar{x} = \frac{18\times1+17\times2+23\times3+20\times4+24\times5+18\times6}{18+17+23+20+24+18} = \frac{429}{120} = 3.575$$

In the example above, the 60th and 61st values are both 4 so the median is 4; the mode is 5.

- Estimate the mean value of h from the figures given in the table.
An estimate of the mean is given by

$$\bar{x} = \frac{755}{72} = 10.486\ldots$$

10.5 = to 1 d.p.

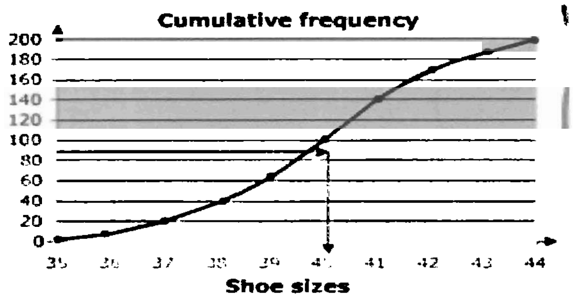
Interval	frequency (f_i)	midpoint (x_i)	$f_i \times x_i$
$0 < h \leqslant 5$	8	2.5	20
$5 < h \leqslant 10$	24	7.5	180
$10 < h \leqslant 15$	29	12.5	362.5
$15 < h \leqslant 20$	11	17.5	192.5
Totals	72		755

- A survey was done by asking n=200 students for their shoe size:

Size (x_i)	35	36	37	38	39	40	41	42	43	44
Frequency (f_i)	3	6	17	12	25	37	40	29	16	15
$x_i \cdot f_i$	105	216	444	646	975	1480	1640	1218	688	660

We can multiply each x_i by its frequency f_i, total these products and divide by 200. Mean shoe sizes = 8072 ÷ 200 = 40.36

We can graph the cumulative frequencies and use it to find out the median: 40



- Calculating the retail price index (the measure used for inflation)

Item	Bread	Milk	Eggs	Butter	Jam
Price	0.65€	0.32€	0.15€	0.93€	1.35€
Weight	7	16	12	2	1

The mean is then calculated by $\bar{x} = \frac{\sum x_i w_i}{\sum w_i} = \frac{7\cdot 0.65 + 16\cdot 0.32 + \ldots + 1\cdot 1.35}{7+16+\ldots+1} = 0.386$

The weighted mean price for the household basket is 39 cent.

Skew

The **lower quartile** is the data between the first and the second fourths of the data; the **upper quartile** is the data placed between the third and the last fourths of the data.

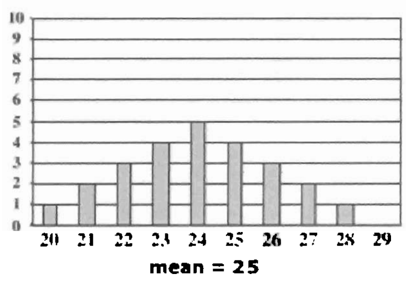
They can indicate whether the data values show **positive skew** (Q2–Q1<Q3–Q2) or **negative skew** (Q2–Q1>Q3–Q2).

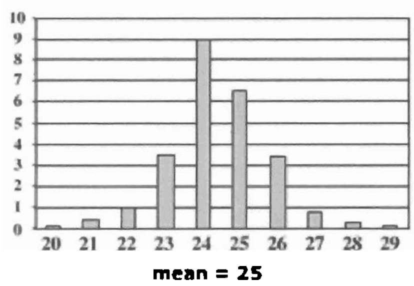
Note: When the quartile falls between two terms, the data is treated as though they were continuous and the value is the proportionate distance between the terms.

Note: The median and the interquartile range are usually represented by box and whisker diagrams.

Measures of Dispersion (Spread)

We use a measure of central tendency to give a description of our set of data. Different distributions may share the same mean or median but have a markedly different shape.





We now need to use new measures which takes account of the spread (or dispersion) of data around the central measure. They are **measures of spread.**

Range: difference between the highest value and the lowest value.

Inter-quartile range: the distance between the upper and lower quartiles. It is an indicator of the density of data in the middle 50% of the data when placed in order.

Mean deviation: $\frac{1}{n}\sum_i |x_i - \bar{x}| \cdot f_i$ the 'average' amount by which the data items differ from the mean.

Variance (σ^2): $\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot f_i$ being $n = \sum_i f_i$ the number of data, $\bar{x} = \frac{1}{n}\sum_i x_i \cdot f_i$ the mean. It is the mean of the squares of the deviations (deviations are the amounts the data differ from the mean). Notice that the units of the data have been squared in the process. Another expression for the variance is $\sigma^2 = \left(\frac{1}{n}\sum_i x_i^2 \cdot f_i\right) - (\bar{x})^2$

Standard deviation (σ): $\sigma = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot f_i}$. It the square root of the variance. It is measured in the same units as the data and is the value most commonly used to measure dispersion at this level.

Examples: ■ We can obtain the mean $\bar{x} = 4.75$ from this table of frequency

Measure (x_i)	1	2	3	4	5	6	7	8
Frequency (f_i)	1	2	3	5	8	5	3	1
Deviations $(\lvert x_i - \bar{x}\rvert)$	3.75	2.75	1.75	0.75	0.25	1.25	2.25	3.25
$\lvert x_i - \bar{x}\rvert \cdot f_i$	3.75	5.25	5.25	3.75	2	6.25	6.75	3.25

the sum of the previous is 36.5
the mean of the deviations is 36.5 ÷ 28 = 1.304 that is the mean deviation
the range is 8 – 1 = 7

■ We can obtain the mean $\bar{x} = 4.75$ from this table of frequency

Measure x_i	1	2	3	4	5	6	7	8
Frequency f_i	1	2	3	5	8	5	3	1
Deviat. $\lvert x_i - \bar{x} \rvert$	3.75	2.75	1.75	0.75	0.25	1.25	2.25	3.25
Sq. dev. $(x_i - \bar{x})^2$	14.0625	7.5625	3.0625	0.5625	0.0625	1.5625	5.0625	10.5625
$(x_i - \bar{x})^2 \cdot f_i$	14.0625	15.125	9.1875	2.8125	0.5	7.8125	15.1875	10.5625

the sum of the previous is 75.25

the mean of the squared deviations is $\sigma^2 = 75.25 \div 28 = 2.6875$ (the variance)

the square root of the variance is $\sigma = \sqrt{2.6875} = 1.6394$ (the standard deviation)

■ A survey was done by asking n=200 students for their shoe size:

Size (x_i)	35	36	37	38	39	40	41	42	43	44
Frequency (f_i)	3	6	17	12	25	37	40	29	16	15
Cumulative (Fi)	3	9	26	38	63	100	140	169	185	200

Lower quartile = ¼(200 + 1) = 50.25

That is the value a quarter of the way between the 50th and 51st term.

In this case the two terms are both 39, so the lower quartile is 39.

Upper quartile = ¾(200 + 1) = 150.75

That is the value three quarters of the way between the 150th and 151st term.

In this case the two terms are both 41, so the upper quartile is 41.

The interquartile range is 41 – 39 = 2

50% of the shoe sizes in this survey are between 39 and 41

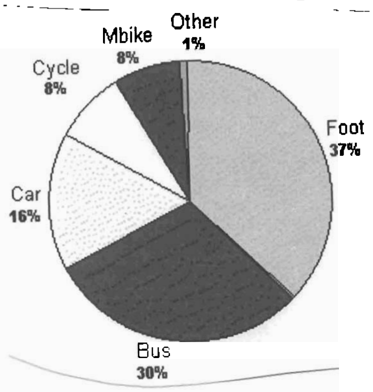
Statistical diagrams

The use of statistical diagrams may help in comparing distributions and revealing further information about the data.

A **pie chart** is used to display the proportion (i.e. percentage or fraction) of the data belonging to different categories. Pie charts should only be used for nominal (categorical, qualitative) data. If ordinal data are used it is important that the order of the sectors follows the natural order of the data.

Example: **Display the data below (main mode of travel of 120 first year students)**

Mode of Travel	Foot	Bus	Car	Cycle	Mbike	Other	Total
No. of students	45	36	19	10	9	1	120
% of students	37	30	16	8	8	1	100%

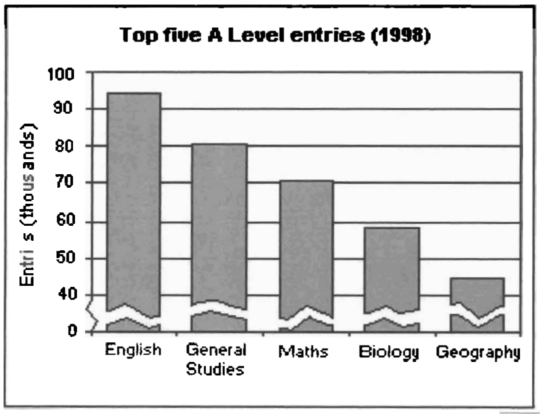


The **bar-chart** (or **column chart**) is the simplest and most versatile of statistical diagrams. In a vertical bar chart the height of a bar is used to represent the frequency. All bars must be the same width and evenly spaced along the x (category) axis. The gaps between the bars are used to emphasise the distinction between the categories or discrete values.

The ordering of the bars deserves careful consideration. If the categories have a natural order, then it must be maintained. If there is no natural order then the bars should be placed in order of height.

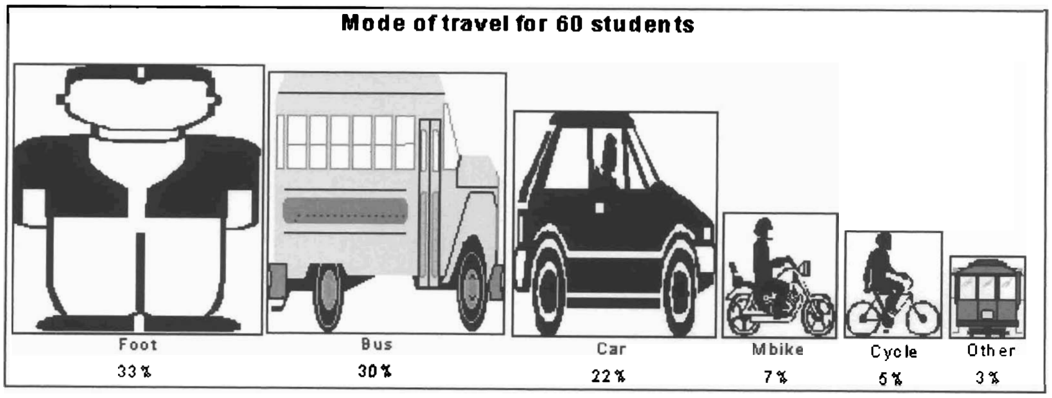
A popular trick in newspapers and magazines is to use a non-zero origin on the frequency axis to exaggerate the differences between bars. To avoid the risk of misleading the reader it is best if the zero is always included in a bar chart. If a non-zero origin is used, it should be very clearly indicated using a dramatic break in the bars and axis.

Example:

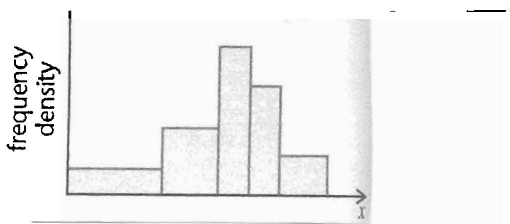


An attractive alternative to standard bar charts is to replace the bars with pictorial images that is called a **pictogram**.

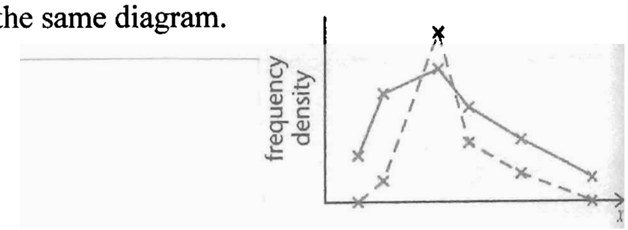
Example:



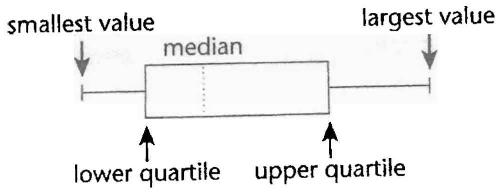
Histograms are useful for illustrating grouped continuous data. The area of a bar is proportional to the frequency in that interval. The height of the rectangle is given bay the **frequency density**, (frequency/width of the interval). The **modal class** is the interval with the greatest frequency density, the highest bar on the histogram.



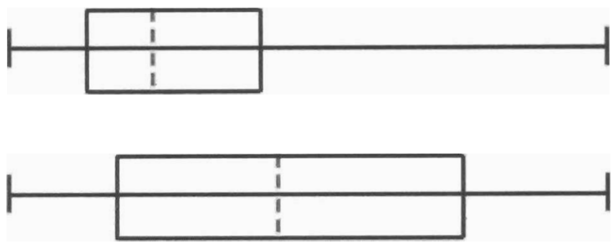
Joining with straight lines the mid points of the tops of the bars of a histogram gives a **frequency polygon**. Comparisons can be made by superimposing more than one frequency polygon on the same diagram.



A **box and whisker plot** shows the location and spread of a distribution at a glance.



Example:



The upper diagram shows a narrow grouping around the median, and a skew towards the lower end of the scale. The lower diagram shows the median is more central and the inter-quartile range much more spread out.

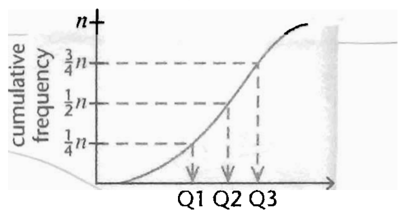
A **back-to-back stem and leaf diagram** allows direct comparison of two sets of data to be made. The diagram gives a sense of location and spread for each data set.

Example: The spread of marks is similar for the boys and the girls but the average for the girls is higher than for the boys.

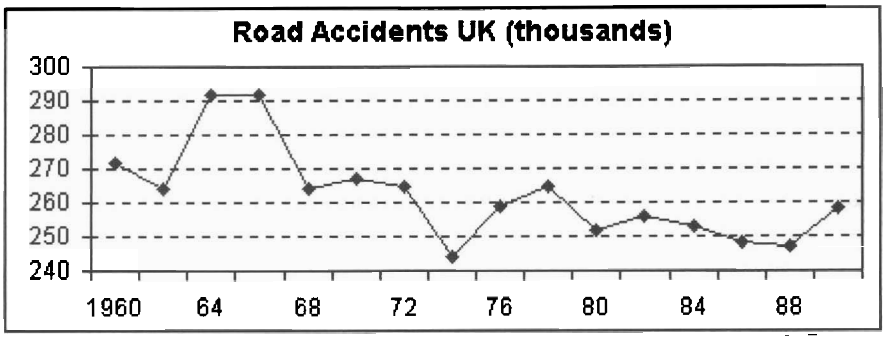
boys	coursework marks	girls
2	4	1 7
5 3	3	2 5 8 8
5 4 4 2 1	2	4 4 7 9
8 7 6 4	1	3 6 6
6 5	0	7

2 | 4 means 42%   4 | 1 means 41%

In a **cumulative frequency diagram**, cumulative frequencies (running totals) are plotted against upper class boundaries of the intervals. The median Q2 and quartiles Q1 and Q3 can be estimated from the graph. For "n" items, Q1 is the value 25% through the data (n/4), Q2 is the value 50% through the data (n/2), Q3 is the value 75% through the data (3n/4).



A **line chart** is used for displaying time series data. A **time series** is a set of observations recorded at successive points in time, normally at regular intervals.
Example:



Bivariate data. Scatter diagrams.

A **scatter diagram** may be used to represent bivariate data. The extent to which the points approximate to a straight line gives an indication of the strength of a linear relationship between the variables, known as the **linear correlation**.

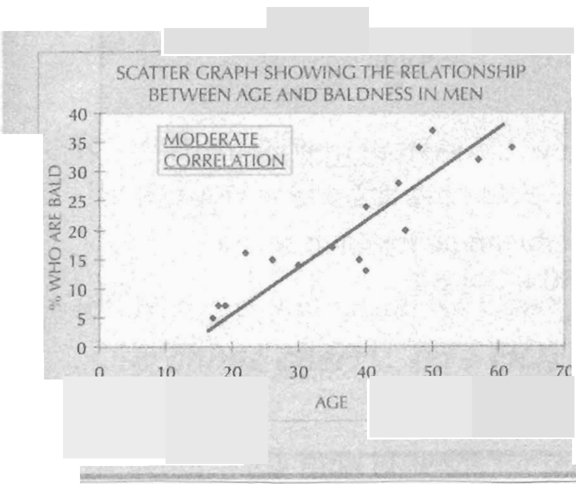
A good correlation (or strong correlation) means the points form quite a nice line, and it means the two things are closely related to each other.

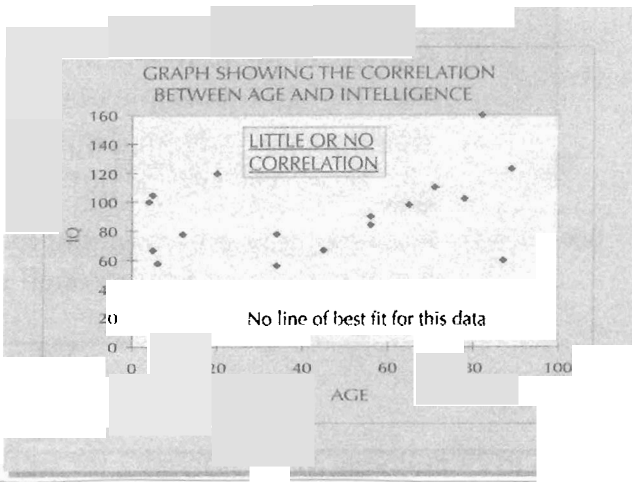
A poor correlation (or weak correlation) means the points are all over the place and so there's very little relation between the two things.

If the points form a line sloping uphill from left to right, then there is positive correlation, which just means that both things increase or decrease together.

If the points form a line sloping downhill from left to right, then there is negative correlation. That just means that as one thing increases the other decreases.

So when you're describing a scatter graph you have to mention both things, i.e. whether it's a strong / weak / moderate correlation and whether it's positive / negative.





EXERCISES

1) For the numbers 8, 14, 20, 1, 6, 6, 13, 12, 4,
(i) find (a) the mean (b) the standard deviation (c) the median
(d) the interquartile range,
(ii) draw a box and whisker plot.

2) Find the mean and standard deviation of this set of data:

x	0	1	2	3
f	4	6	7	3

3)

Mass (g)	$20 \leqslant m < 40$	$40 \leqslant m < 50$	$50 \leqslant m < 55$	$55 \leqslant m < 60$	$60 \leqslant m < 90$
Frequency	10	12	10	7	9

For the above data
(a) estimate the mean and the standard deviation,
(b) draw a histogram and state the modal class,
(c) draw a cumulative frequency curve and use it to estimate
(i) the median (ii) the interquartile range.

4) a) The mean of this data is 3. Find the value of y.

x	1	2	3	4	5
$f(x)$	5	y	13	14	3

b) Find the value of k if the mean of this data is 2.6:

x	0	1	2	3	4	5
f	2	3	5	4	k	3

c) The mean of this data is 7. Find the value of y.

x	5	6	7	8	9	10
f	3	y	8	4	2	1

d) Here is a table of the ages of students in a youth club:

Age	11	12	13	14	15
Frequency	3	3	x	x	$x+3$

(i) If the mean age is 13.5, find the value of x.
(ii) Name the mode.
(iii) How many students are in the club?

5) The marks of 15 students in a test were as follows:

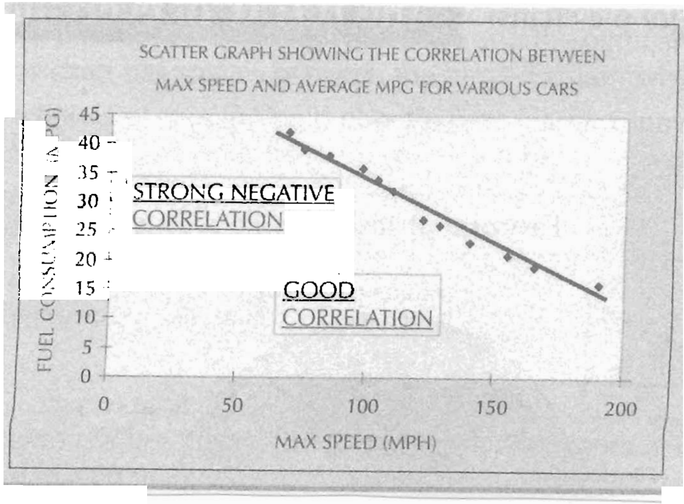
53 67 43 71 21
49 58 48 77 37
82 51 61 98 84

(i) Verify that the mean mark is 60.
(ii) Copy and complete the grouped frequency table below:

Mark	20–40	40–60	60–80	80–100
Frequency	2			3

(iii) Use mid-interval values to estimate the mean from this table. Is this value greater or less than the true mean?

Bivariate data. Correlation.



One way to arrive at a numerical measure of the correlation is to use the **product-moment correlation coefficient, r.**

For n pairs of (x, y) values:

$$S_{xx} = \sum x^2 - n\bar{x}^2 \qquad S_{yy} = \sum y^2 - n\bar{y}^2 \qquad S_{xy} = \sum xy - n\bar{x}\bar{y},$$

and the product–moment correlation coefficient is given by:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left\{\sum(x-\bar{x})^2\right\}\left\{\sum(y-\bar{y})^2\right\}}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

The Product-Moment Correlation Coefficient (PMCC, or r, for short) measures how close to a straight line the points on a scatter graph lie.

The PMCC is always between +1 and −1.
If all your points lie exactly on a straight line with a positive gradient (perfect positive correlation), r = +1.
If all your points lie exactly on a straight line with a negative gradient (perfect negative correlation), r = −1.

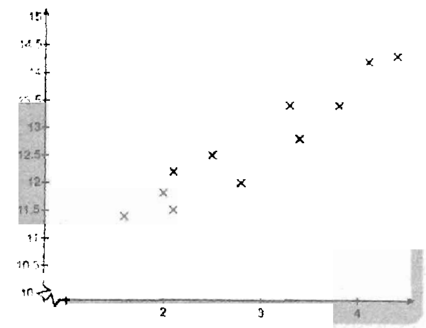
(In reality, you'd never expect to get a PMCC of +1 or −1 — your scatter graph points might lie pretty close to a straight line, but it's unlikely they'd all be on it.)

If r = 0 (or more likely, pretty close to 0), that would mean the variables aren't correlated.

Example: Illustrate the following data with a scatter diagram, and find the product-moment correlation coefficient (r) between the variables x and y.

x	1.6	2.0	2.1	2.1	2.5	2.8	2.9	3.3	3.4	3.8	4.1	4.4
y	11.4	11.8	11.5	12.2	12.5	12.0	12.9	13.4	12.8	13.4	14.2	14.3

1) The scatter diagram's the easy bit — just plot the points.



Now for the correlation coefficient. From the scatter diagram, the points lie pretty close to a straight line with a positive gradient — so if the correlation coefficient doesn't come out pretty close to +1, we'd need to worry...

2) There are 12 pairs of readings, so n = 12. That bit's easy — now you have to work out a load of sums. It's best to add a few extra rows to your table...

x	1.6	2	2.1	2.1	2.5	2.8	2.9	3.3	3.4	3.8	4.1	4.4	$35 = \Sigma x$
y	11.4	11.8	11.5	12.2	12.5	12	12.9	13.4	12.8	13.4	14.2	14.3	$152.4 = \Sigma y$
x^2	2.56	4	4.41	4.41	6.25	7.84	8.41	10.89	11.56	14.44	16.81	19.36	$110.94 = \Sigma x^2$
y^2	129.96	139.24	132.25	148.84	156.25	144	166.41	179.56	163.84	179.56	201.64	204.49	$1946.04 = \Sigma y^2$
xy	18.24	23.6	24.15	25.62	31.25	33.6	37.41	44.22	43.52	50.92	58.22	62.92	$453.67 = \Sigma xy$

Stick all these in the formula to get: $r = \dfrac{\left(453.67 - \frac{35 \times 152.4}{12}\right)}{\sqrt{\left(110.94 - \frac{(35)^2}{12}\right) \times \left(1946.04 - \frac{(152.4)^2}{12}\right)}} = \dfrac{9.17}{\sqrt{8.857 \times 10.56}} = \underline{0.948}$ (to 3 s.f.)

This is pretty close to 1, so there's a high positive correlation between x and y.

An alternative measure of correlation is given by **Spearman's rank correlation coefficient, r_s.**

The set of values for each variable must first be ranked from largest to smallest. Some care is needed with equal values:

The three equal values occupy positions 2, 3 and 4. The average positional value is given by:

$$\frac{2+3+4}{3} = 3.$$

x	rank
39	1
37	3
37	3
37	3
32	5

Each of the equal values is given a rank of 3.

At each data point, the difference in the rank of the two variables is denoted by d.

The rank correlation coefficient is given by $r_s = 1 - \dfrac{6\,\Sigma d^2}{n^3 - n}$.

This gives values of r_s between −1 (representing a perfect negative correlation between the rankings) and +1 (representing a perfect positive correlation between the rankings).

The two correlation coefficients given above are comparable but not necessarily equal.

Bivariate data. Regression.

Whereas correlation is determined by the *strength* of a linear relationship between the two variables, **regression** is about the *form* of the relationship given by the equation of a **regression line**. (Make sure you known the difference between correlation and regression.)

The purpose in establishing the equation of a regression line is to make predictions about the values of one variable (known as the **response variable**) for some given values of the other variable (known as the **explanatory variable**).

Predictions should only be made for values within the range of readings of the explanatory variable. **Extrapolation** for values outside this range is unreliable. Another factor affecting the accuracy of any predictions is the influence of **outliers** on the equation of the regression line.

Figure 1 illustrates y-**residuals** given by

d = (observed value of y) – (predicted value of y).

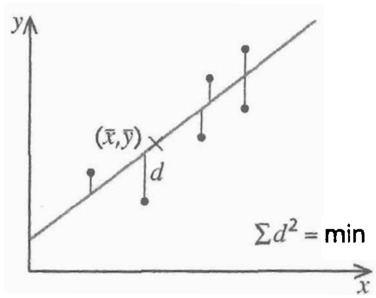
Figure 2 illustrates x-**residuals** given by

d = (observed value of x) – (predicted value of x).

A **least squares regression line** is a line for which the sum of the squares of either the x-residuals or y-residuals is minimised.

Both lines always pass through the point $(\bar{x}, \bar{y})$.

Figure 1



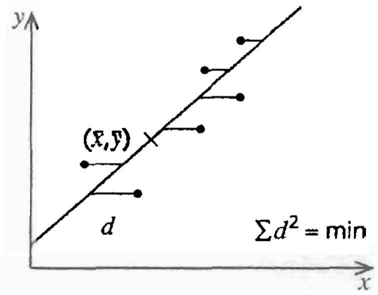
This gives the regression line of y on x as

$$y = a + bx,$$

where $b = \dfrac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$.

Use this equation to estimate values of y for given values of x when x is the explanatory variable and y is the response variable.

Figure 2



This gives the regression line of x on y as

$$x = c + dy,$$

where $d = \dfrac{S_{xy}}{S_{yy}}$ and $c = \bar{x} - d\bar{y}$.

Use this equation to estimate values of x for given values of y when y is the explanatory variable and x is the response variable.

Unless there is perfect correlation between the variables, the two regression lines will be different and you cannot rearrange one equation to obtain the other.

EXERCISES

6) Twenty people go on a historic tour of Kilkenny. Their ages are as follows:

15	14	25	23	33
45	13	51	58	48
19	57	47	56	44
11	38	46	21	16

(i) Verify that 34 is the mean of this data.

(ii) Copy and complete the following grouped frequency table for this data:

Age	0–20	20–40	40–60
Frequency			

Use this table to estimate the mean.

(iii) Find, to the nearest unit, the percentage error in the estimated mean, compared with the true mean.

(Percentage error $= \dfrac{\text{Error}}{\text{True answer}} \times \dfrac{100}{1}\%$)

7) Values of two variables x and y obtained from a survey are recorded in the table below.

x	1	2	3	4	5	6	7	8
y	0.50	0.70	0.10	0.82	0.64	0.36	0.16	0.80

Represent these data on a scatter diagram, and obtain the product-moment correlation coefficient (PMCC) between the two variables.

What does this tell you about the variables?

8) *Plot a scatter diagram and calculate the product-moment coefficient of correlation for the data below.*

Height (cm)	165	176	159	167	174	171	169	168	169	172
Weight (kg)	72	90	70	75	86	84	80	81	82	83

What does the value of the product-moment coefficient of correlation tell you about the data?

9) The summary data for 10 pairs of (x, y) values is as follows:

$$\Sigma x = 146, \quad \Sigma x^2 = 2208, \quad \Sigma y = 147, \quad \Sigma y^2 = 2247, \quad \Sigma xy = 2211.$$

1 Find the value of the product–moment correlation coefficient between x and y.

2 Find the equation of the least squares regression line of y on x.