GCE AS

eduqas
Part of WJEC

WJEC Eduqas GCE AS in

GEOLOGY

ACCREDITED BY OFQUAL
DESIGNATED BY QUALIFICATIONS WALES

Mathematical Guidance for GCE AS Geology

Teaching from 2017
For award from 2018

wjec
cbac

Contents

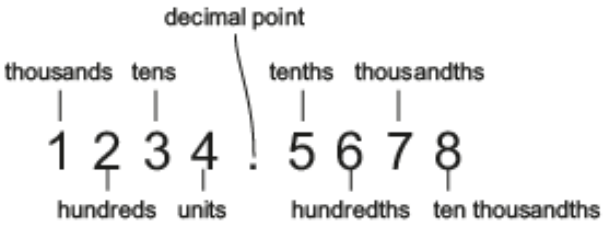
Introduction

Geology is often falsely described as a qualitative (i.e. purely descriptive) science. However, many of the topics covered in AS geology have important underlying mathematical concepts that really need to be understood before a thorough grasp of that subject area is possible. Questions like how do we measure earthquakes?, what is the absolute age of a mineral? and what is the size of the Earth's core? can only be answered by students possessing some quantitative skills. The following pages document some of the more important mathematical skills that are required of students following the WJEC Eduqas Geology AS course.

Students will require the use of a scientific calculator in their lessons and in the examinations.

Decimal and standard form

All the numbers we use to describe quantity, size etc. in geology can be written in **decimal form**, with an arbitrary number of decimal places. Decimal integers written to the right of the decimal point specify the number of tenths, hundredths, thousandths and so on. In the same way as the integers to the left of the decimal point indicate the number of units, tens, hundreds and so on.



$$1234.5678 = 1000 + 200 + 30 + 4 + \frac{5}{10} + \frac{6}{100} + \frac{7}{1000} + \frac{8}{10000}$$

In geology we often encounter both very large and very small numbers e.g. the age of the Earth and the diameter of a clay mineral. It is impractical to write the age of the Earth as 454 000 000 000 years old or the diameter of a clay mineral as 0.0 00039 m, so **standard index form** is used instead. This is the conventional way of writing both large and small numbers and has the additional advantage of simplifying calculations by enabling the use of the index laws. In standard index form the decimal form is expressed as a number between 1 and 10 multiplied by 10 to the appropriate index.

To convert large numbers to standard index form, count the zeros in the number. This gives the value of the index of 10. Then make any adjustment required so that the number in front lies between 1 and 10.

e.g. 4 540 000 000 years is 454 followed by 7 zeros = $454 \times 10^7 = 4.54 \times 10^9$ years

To convert small numbers to standard index form, count the zeros in the number, including the zero before the decimal point. This gives the value of the index of 10, which will be negative if the number is less than one. The digits in front of the index should be written to lie between 1 and 10.

e.g. 0.00000391 m is 6 zeros followed by 391 = 3.91×10^{-6} m

As an alternative to standard index form, in the **S.I. system** of units different names are introduced for each thousandfold increase or decrease in size. For example, the basic unit of length is the metre. The next unit up, one thousand times larger, is called the kilometre, and the next unit down, one thousand times smaller, is called the millimetre. The prefixes used in S.I. units are listed in the table below.

multiple	prefix	symbol	example of units
10^{-9}	nano	n	nanometre
10^{-6}	micro	μ	micrometre
10^{-3}	milli	m	millimetre
1	no prefix		metre (m)
10^{3}	kilo	k	kilometre (km)
10^{6}	mega	M	megametre (Mm)
10^{9}	giga	G	gigametre (Gm)

A clay mineral with a diameter of 0.00000391 m could therefore written as 3.91×10^{-6} m or 3.91 μm.

Significant figures and estimation

Significant figures are 'each of the digits of a number that are used to express it to the required degree of precision, starting from the first non-zero digit'. Numbers are often rounded to avoid reporting insignificant figures. For example, it would create false precision to express a measurement as 12.34500 kg (which has seven significant figures) if the scales only measured to the nearest gram and gave a reading of 12.345 kg (which has five significant figures).

Non-zero figures are always significant. Thus, 22 has two significant figures, and 22.3 has three significant figures. With zeroes, the situation is more complicated:

a. Zeroes placed before other figures are not significant; 0.046 has two significant figures.
b. Zeroes placed between other figures are always significant; 4 009 has four significant figures.
c. Zeroes placed after other figures but behind a decimal point are significant; 7.90 has three significant figures.
d. Zeroes at the end of a number are significant only if they are behind a decimal point as in (c). Otherwise, it is impossible to tell if they are significant. For example, in the number 8 200, it is not clear if the zeroes are significant or not. The number of significant figures in 8 200 is at least two, but could be three or four. To avoid uncertainty, use standard index form to place significant zeroes behind a decimal point:

8.200×10^3 has four significant figures; 8.20×10^3 has three significant figures; 8.2×10^3 has two significant figures

In a calculation involving multiplication, division, trigonometric functions etc. when asked to round to an appropriate level of accuracy, the number of significant figures in an answer should equal the least number of significant figures in any one of the numbers being multiplied, divided etc.

For example, the mass of a granite pebble is determined as 276.5 g (four significant figures) and its volume is 105 cm^3 (three significant figures). The density of the pebble is $\frac{276.5}{105} =$ 2.63 g/cm^3 (three significant figures).

When quantities are being added or subtracted, the number of decimal places (not significant figures) in the answer should be the same as the least number of decimal places in any of the numbers being added or subtracted.

Also when doing multi-step calculations, keep at least one more significant figure in intermediate results than needed in your final answer. For instance, if a final answer requires two significant figures, then carry at least three significant figures in calculations. If you round-off all your intermediate answers to only two significant figures, you are discarding the information contained in the third significant figure, and as a result the second significant figure in your final answer might be incorrect. (This phenomenon is known as a "rounding error.")

It is possible to quickly and easily work out the answer to any calculation by performing an estimate. This should always be done to check that an answer is reasonable. We don't want an estimate to take a long time (otherwise, we may as well do the full calculation), so the quickest idea is to round all numbers off to 1 significant figure.

For example estimate the answer to $4.2 + 9.8 \times 19.4$.

$4.2 + (9.8 \times 19.4) \approx 4 + (10 \times 20) \approx 4 + 200 \approx 200$

Order of magnitude calculations

Simply speaking an order of magnitude is how many powers of ten there are in a number. Orders of magnitude can be determined easily when a number is written in standard form. For example, 237 (2.37×10^2) and 823 (8.23×10^2) both have an order of magnitude of 2.

Comparing orders of magnitude is a useful way of estimating the difference between two numbers. For example, the permeability of rocks varies enormously, from 1 microdarcy (1×10^{-6} D) for shales and clays that form cap-rocks to several darcies for extremely good reservoir rocks. An exceptional reservoir rock has a permeability of 1 darcy (1×10^0 D) or more. Comparing the magnitude of these two numbers by dividing enables us to determine that exceptional reservoir rocks are 6 orders of magnitude (6 powers of ten) more permeable than a typical cap-rock;

$$= \frac{(1 \times 10^0)}{(1 \times 10^{-6})} = 10^6$$

A scanning electron microscope may produce a magnification of up 40 000 (4×10^4) whereas a typical optical laboratory microscope may produce a magnification of just 40 (4×10^1). A scanning electron microscope therefore produces an image 4 orders of magnitude bigger than the object and 3 orders of magnitude bigger than an optical microscope.

Ratios, fractions and percentage

Ratios, fractions and percentages are some of the most useful mathematical concepts in geology as they enable comparisons to be made between the sizes of many different geological phenomena.

Mineral/ hardness		*Common equivalent*
Diamond	10	
Corundum	9	
Topaz	8	
Quartz	7	
		← steel pin
Orthoclase feldspar	6	
Apatite	5	← copper coin
Fluorite	4	← finger nail
Calcite	3	
Gypsum	2	
Talc	1	

The above table of Mohs hardness scale can be used to show the link between ratios, fractions and percentages. For example, three of the minerals out of the ten can be scratched by a copper coin which as a fraction is $\frac{3}{10}$ or 30%. The proportion (a part to whole comparison) of minerals that can be scratched by a copper coin can be expressed as $\frac{3}{10}$ or 30% or 3 in 10. The ratio (a part to part comparison) of minerals that can be scratched by a copper coin to those that cannot is 3:7. In comparison five of the minerals out of the ten can be scratched by a steel pin which as a fraction is $\frac{5}{10}$ (which simplifies to $\frac{1}{2}$) or 50%. The ratio of minerals that can be scratched by a steel pin to those that cannot is 5:5 (which simplifies as 1:1).

Data and Statistical Analysis

Most geological phenomena are extremely complex in their inter-relationships and vast in their spatial distribution. Consequently an exact description of a geological system is rarely feasible and almost certainly uncertain. A principal problem is the need to sample a very small sub-section of a very large population and the need to make deductions from this data set. Designing a good experiment or fieldwork investigation so that the information represents a good sample and yields meaningful statistics (estimates of the population) is therefore extremely important. Consequently statistics, and their presentation, is probably the most intensively used branch of mathematics in geology.

Sampling methods

Sampling should be conducted in a way that will best represent the data being collected. There are three main types of sampling: random, systematic and stratified.

In **random sampling** every item has an equal chance of being selected. For many studies this is the most desirable approach as there is no bias. The most common way of random sampling is to use a random number table or generator. The twelve pieces of fault rupture length-earthquake magnitude data analysed later were sampled from a set of 58 by this approach.

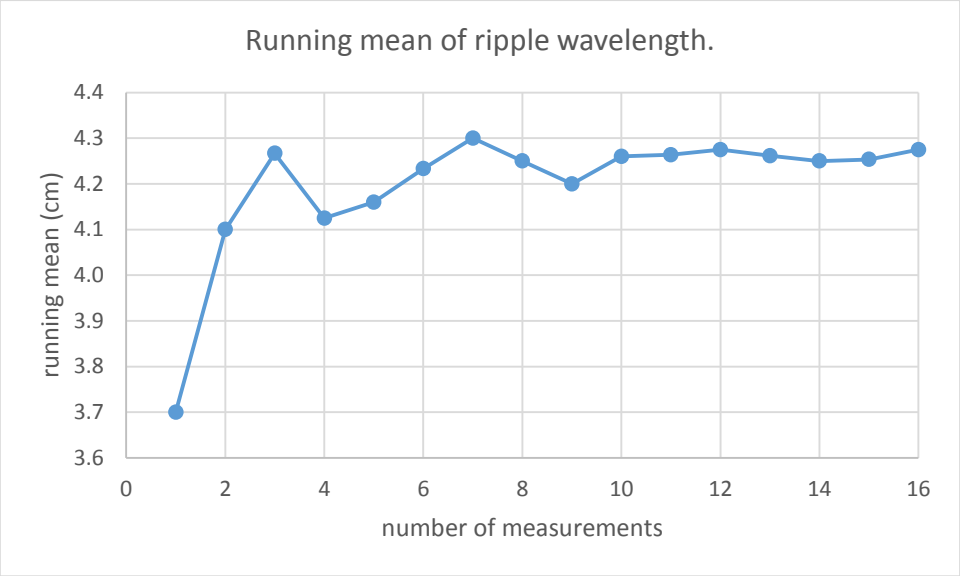
In **systematic sampling** there is some structure or underlying order to the way in which the data is selected. The fifty pieces of Schmidt hammer hardness data analysed later were sampled by use of a five by ten gridding system.

With **stratified sampling** the population is purposely split into separate groups/layers (strata). Then each group is further analysed by random or systematic sampling. Stratified sampling has the advantage of reducing the sample size required to produce the same precision as other techniques. This approach could be adopted to investigate the particle shapes in a sieved unconsolidated sediment. In this case the sample size from each layer would be proportional to the mass of each grain size fraction.

Once the sampling method has been selected, the next step is to decide the number of samples that should be taken to provide a reasonable estimate of the mean of the population. Although the more samples taken the more reliable estimate of the mean, there comes a stage when taking more readings is unlikely to be productive. One way of determining when enough measurements have been taken (the optimum sampling size) is to calculate the running mean as measurements are taken.

The data below was obtained by systematic sampling (line transect) along a large exposed limestone bedding plane in the lower Cretaceous Purbeck Formation near Swanage. Well exposed ripple marks on the bedding plane enable the ripple wavelength to be measured which is necessary to calculate the ripple index.

Number of measurements	ripple wavelength (cm)																running mean (cm)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	3.7																3.7
2	3.7	4.5															4.1
3	3.7	4.5	4.6														4.3
4	3.7	4.5	4.6	3.7													4.1
5	3.7	4.5	4.6	3.7	4.3												4.2
6	3.7	4.5	4.6	3.7	4.3	4.6											4.3
7	3.7	4.5	4.6	3.7	4.3	4.6	4.7										4.3
8	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9									4.2
9	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8								4.3
10	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8							4.3
11	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3						4.3
12	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3	4.4					4.3
13	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3	4.4	4.1				4.3
14	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3	4.4	4.1	4.1			4.3
15	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3	4.4	4.1	4.1	4.3		4.3
16	3.7	4.5	4.6	3.7	4.3	4.6	4.7	3.9	3.8	4.8	4.3	4.4	4.1	4.1	4.3	4.6	4.3



Scrutinising the graph above shows that when $n \geq 10$ the running mean tends to 4.3 (levelling off of the mean) suggesting that, in this case, 10 measurements may have been a suitable sample number to estimate the population mean.

Univariate analysis is the simplest form of analysing data as it deals with just one variable. Consequently univariate analysis doesn't describe relationships; its major purpose is to describe the characteristics of the sample. Looking at two variables at one time is termed bivariate analysis and three or more variables is multivariate analysis.

Univariate data analysis

There are several choices for displaying and analysing univariate data but these depend upon the type of variable being investigated.

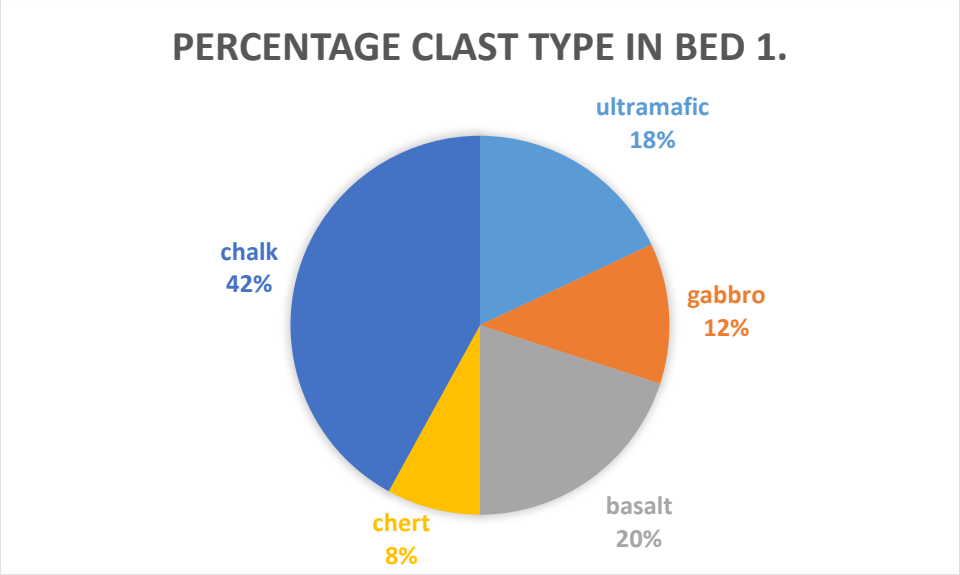
In the case of categorical variables (often called qualitative variables) frequency tables are constructed to produce pie charts and bar charts. In the case of numerical data (often called quantitative variables) frequency tables are constructed to produce:

i) bar charts or vertical line graphs for ungrouped discrete data.
ii) histograms and box-and-whisker plots (as well as a range of other graphs) for continuous data or grouped discrete data.

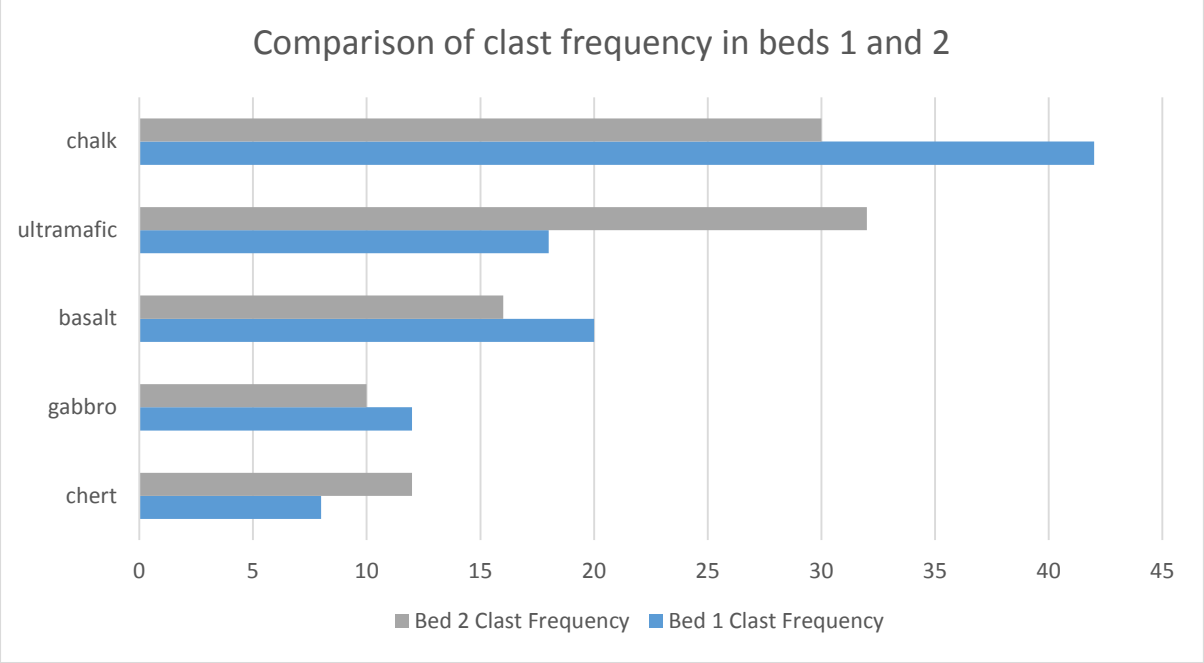
The following frequency table provides information on the lithology of clasts from two adjacent Pleistocene fluvial deposits near Paphos airport in south east Cyprus. The variable, clast type, is a categorical variable so the information in this table could be displayed either as a pie or bar chart.

	Bed 1		Bed 2	
Clast Type	Frequency (and %)	Relative frequency	Frequency (and %)	Relative frequency
ultramafic	18	0.18	32	0.32
gabbro	12	0.12	10	0.10
basalt	20	0.20	16	0.16
chert	8	0.08	12	0.12
chalk	42	0.42	30	0.30
total	100	1	100	1

The pie chart below shows the results of clast type variation for bed 1. Each category is represented by a slice of the pie where the area of the slice is proportional to the percentage of responses in the category.



Pie charts are effective when displaying the relative frequencies of a small number of categories (a bar chart is a better option for a large number of categories). Bar charts also are very powerful for comparing the distributions of two or more samples – see below. Note, whether the bars are vertical or horizontal depends on which is felt most visually informative.



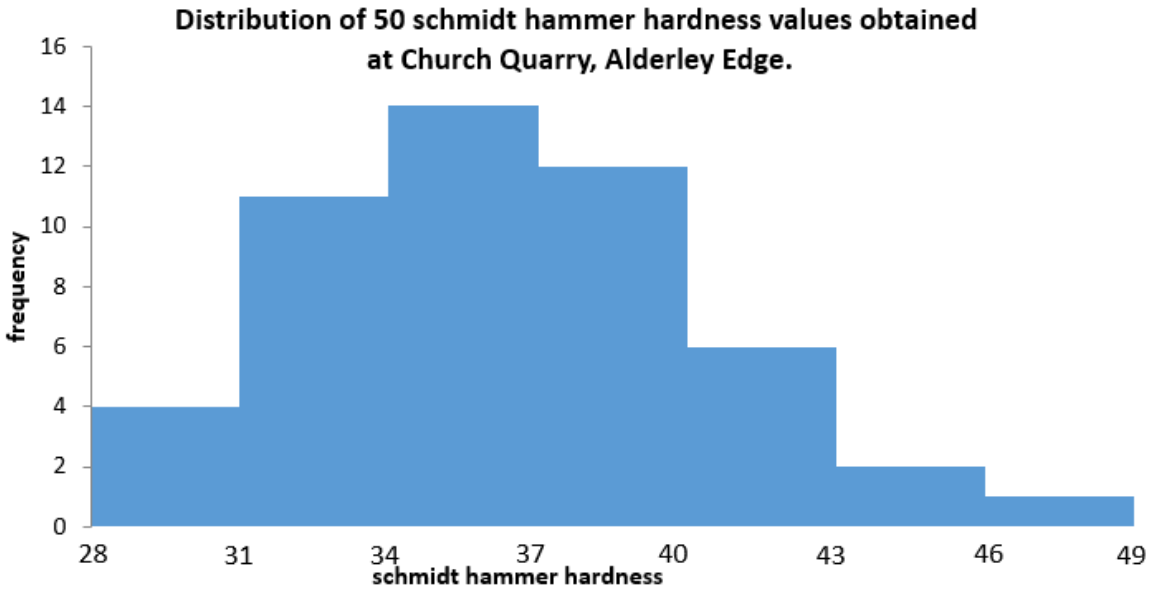
Note the gap between the variable categories and display of frequency of clasts on the bar chart.

Let us now consider an example with a continuous variable. The table below shows 50 Schmidt hammer hardness values (a proxy for uniaxial compressive strength) obtained from Triassic sandstones at Church Quarry, Alderley Edge, Cheshire. Although the practicalities of resolution preclude the Schmidt hammer measurement being truly continuous (a problem with all measurements) the value of Schmidt hammer hardness can range in a continuous scale from 0 to 100 and is not made up of discrete steps.

Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness
1	41	*11*	43	*21*	39	*31*	34	*41*	32
2	38	*12*	33	*22*	33	*32*	37	*42*	38
3	44	*13*	32	*23*	43	*33*	36	*43*	36
4	38	*14*	36	*24*	34	*34*	38	*44*	39
5	31	*15*	46	*25*	36	*35*	36	*45*	40
6	37	*16*	35	*26*	36	*36*	40	*46*	38
7	30	*17*	33	*27*	34	*37*	36	*47*	36
8	29	*18*	36	*28*	41	*38*	38	*48*	41
9	40	*19*	31	*29*	38	*39*	37	*49*	42
10	32	*20*	33	*30*	34	*40*	35	*50*	49

The first useful step in the interpretation of this data is to produce a frequency table by dividing the data into class intervals – customarily of the same width – to provide a count of the frequencies of the classes. This will then enable a visualisation of the data as a histogram (a graphical representation of a frequency table) and a cumulative frequency graph if required. A crude rule of thumb regarding the size of the classes is that there should be at least six and the number of classes should equal the square root of the number of points in the data set (variations on this theme exist) but common sense must be used and trial and error is advised. In this example there are 50 data values and therefore 7 class intervals seem initially prudent.

Schmidt hammer hardness (SHH) class		Frequency	Cumulative frequency, %
1	28<SHH≤31	4	8
2	31<SHH≤34	11	30
3	34<SHH≤37	14	58
4	37<SHH≤40	12	82
5	40<SHH≤43	6	94
6	43<SHH≤46	2	98
7	46<SHH≤49	1	100



Class intervals need not be of equal width. Indeed, when data are grouped together narrower class intervals may be prudent and conversely where the data are spread out wider class intervals could be used. Additionally, wider class intervals should be encouraged to avoid gaps in data. To draw a histogram for unequal class intervals, the height of the rectangles must be adjusted so that the area of the rectangle is proportional to the frequency. The height of the rectangle, called the frequency density, is found by dividing the frequency by the class width. It should be recognised here that statisticians would consider that a histogram should always be a plot of frequency density versus class interval. The advantage of this approach is that it enables the probability of a particular event occurring to be determined.

To demonstrate the construction of a variable class width histogram, the 50 data values collected at Alderley Edge have been regrouped again into 7 classes but this time with differing class intervals.

Schmidt hammer hardness (SHH) class		frequency	class width	frequency density $= \frac{\text{frequency}}{\text{class width}}$
1	28<SHH≤31	4	3	1.3·
2	31<SHH≤33	7	2	3.5
3	33<SHH≤35	6	2	3
4	35<SHH≤37	12	2	6
5	37<SHH≤39	9	2	4.5
6	39<SHH≤43	9	4	2.25
7	43<SHH≤49	3	6	0.5

Some ways in which patterns in univariate quantitative data can be described include:

- Measures of central tendency (mean, mode, median)
- Measures of dispersion (maximum, minimum, range, quartiles (including the interquartile range) variance and standard deviation)
- Measures of shape (coefficient of skewness)

Measures of central tendency

Mean: To find the mean, add up the values in the data set and then divide by the number of values that were added, i.e.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{41 + 38 + \ldots.. + 49}{50} = 37$$

Mode: The mode of a sample is the most frequently occurring value. When data is grouped into classes, the modal class is the class containing the greatest number of values. Mode helps identify the most common or frequent occurrence in a dataset. It is possible to have two modes (bimodal), three modes (trimodal) or more modes within larger sets of numbers. In the example the histogram clearly shows the data is unimodal with the modal class being 35-37 and the mode as 36. In the rare case that all the data only happened once, then the mode may not exist.

Median: The median of a sample is the value that evenly splits the number of observations into a lower half of smaller observations and an upper half of larger measurements. Hence determination of the median requires ranking of all the sample values (see rewritten table below). In the case of an even number of observations the median is the arithmetic mean of the two middle numbers. In the example the two middle numbers are both 36 (shaded dark grey on table below) so the median is 36.

Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness	Number	Schmidt hammer hardness
8	29	*22*	33	*26*	36	*4*	38	*45*	40
7	30	*24*	34	*33*	36	*29*	38	*1*	41
5	31	*27*	34	*35*	36	*34*	38	*28*	41
19	31	*30*	34	*37*	36	*38*	38	*48*	41
10	32	*31*	34	*43*	36	*42*	38	*49*	42
13	32	*16*	35	*47*	36	*46*	38	*11*	43
41	32	*40*	35	*6*	37	*21*	39	*23*	43
12	33	*14*	36	*32*	37	*44*	39	*3*	44
17	33	*18*	36	*39*	37	*9*	40	*15*	46
20	33	*25*	36	*2*	38	*36*	40	*50*	49

Measures of dispersion

Measures of dispersion give an idea of the spread of the data.

Extreme values: The extreme values are the maximum and minimum values in the sample. In the example the minimum value is 29 and the maximum value is 49, hence the range is 49 – 29 = 20.

Quartiles (including the interquartile range): The idea of the median splitting the ranked sample into two halves can be generalized to any number of partitions with equal numbers of observations. The partition boundaries are called quantiles or fractiles. The names for the most common quantiles are:
• Median, for 2 partitions
• Quartiles, for 4 partitions
• Deciles, for 10 partitions
• Percentiles, for 100 partitions
The number of boundaries is always one less than the number of partitions.

Quartiles: Three quartiles divide a list of numbers into four equal parts. The middle quartile (the median) has already been discussed. The lower and upper quartiles are calculated by dividing the both halves of data either side of the median into a lower quarter of smaller observations and an upper quarter of larger measurements. In the case of an even number of observations calculate the mean of the two middle numbers. In the example the lower quartile is 34 and the upper quartile is 39 (shaded light grey on table above).

Interquartile range (IQR): The interquartile range is the difference between the upper and lower quartiles thereby giving a measure of the central spread of the data. A practical rule of thumb is to regard any value deviating more than 1.5 times the IQR from the median as a mild outlier and any value deviating more than 3 times the IQR from the median as an extreme outlier. Outliers are values so markedly different from the rest of the sample that they raise the suspicion that they may be from a different population (e.g. value was measured in a nearby conglomerate horizon) or may be in error (e.g. incorrect application of the Schmidt hammer) but it is notoriously difficult to show that the values are anomalous.

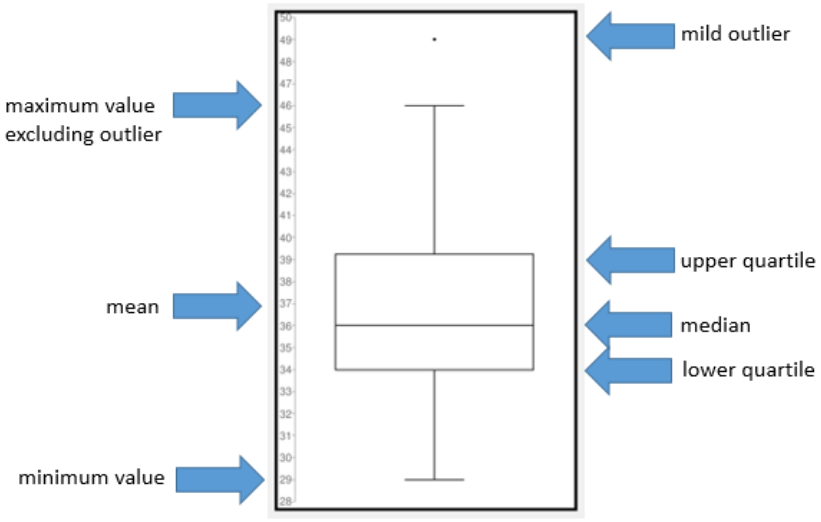
In the example above the IQR is 39 – 34 = 5. The mild outlier boundaries are

= median ± 1.5 IQR = 36 ± 1.5(5) = 29 and 44.

Therefore only two values (sample 15 and 50) may be considered as mild outliers with sample number 50 being the most extreme.

In comparison to variance/standard deviation (discussed below) the IQR is a more robust method for analysing the central spread of the measurements but, unlike variance/standard deviation, is insensitive to the lower and upper tails. Generally speaking if the median is thought to be the best way in which to describe the data average then the IQR is used as the measure of spread. Conversely if the mean is believed to be the best way in which to describe the data average then the standard deviation is utilised.

All the statistics calculated above can be graphically displayed on a box and whisker plot although variations on the specifics displayed abound.



Variance: The sample variance, s^2, is another method used to calculate how varied or spread out from the mean a sample is. Sample variance is mathematically defined as **the average of the squared differences from the mean**. To calculate variance, it is useful to break the calculation down into steps:
Step 1: Calculate the mean (previously discussed).
Step 2: Subtract the mean from each of the values and square the result.
Step 3: Divide by n – 1

In mathematical notation this is written as:

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

where s^2 is the sample variance

x_i is the individual value

$\bar{x}$ is the sample mean

n is the sample size

An incomplete summary table shows how this data could be laid out:

Value (x_i)	Mean ($\bar{x}$)*	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
41	36.88	4.12	16.97
38	36.88	1.12	1.254
44	36.88	7.12	50.69
38	36.88	1.12	1.254
31	36.88	− 5.88	34.57
….	….	….	….
			$\Sigma\left(x_i - \bar{x}\right)^2 = 837.3$
			$\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n-1} = 17$

* Note value used with 4 significant figures to avoid rounding errors.

The sample variance for the exemplar data is therefore 17. While this value is useful in a mathematical sense, the principal use of this calculation is to allow standard deviation to be determined.

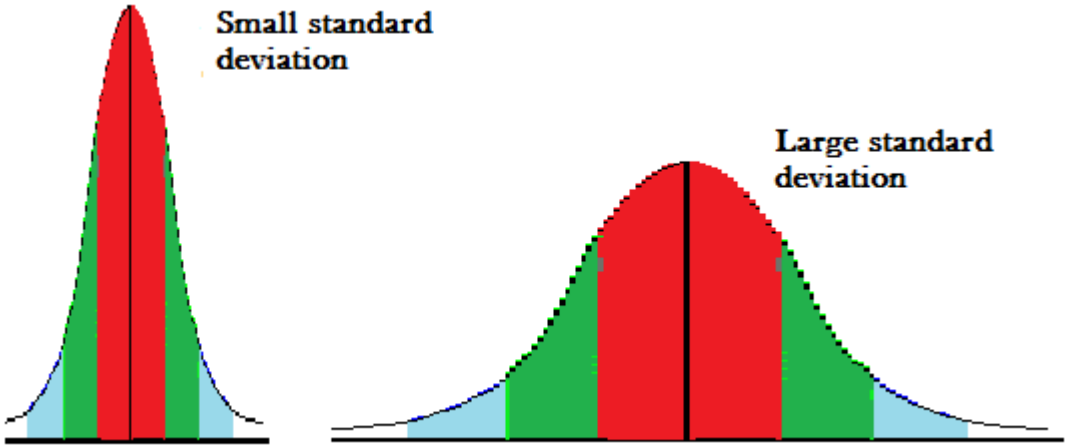
Standard deviation: The standard deviation is the positive square root of the variance.

In mathematical notation:

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

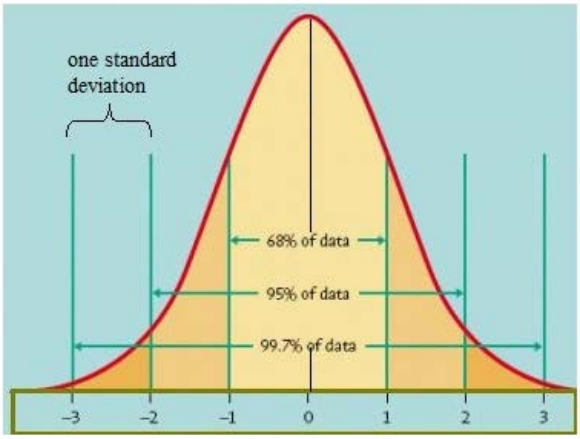
For the exemplar data the value of standard deviation, $s = \sqrt{17.09} = 4$ Schmidt hammer hardness units.

Standard deviation gives us a measure of how clustered the data are around the mean. A smaller value of standard deviation indicates that the data is tightly clustered around the mean and *vice-versa* (see below).



Small and large standard deviation; Statistics How To www.statisticshowto.com/

Observing the shape of these two curves shows that they are symmetrical about the centre. This type of curve is called a bell curve and shows that the data is normally distributed about the centre – the mean. In such a normal distribution the mean, mode and median are equal and exactly half the values are to the left of the centre and half the values are to the right. In the standard normal model about 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean and just over 99% of the data falls within three standard deviations of the mean.



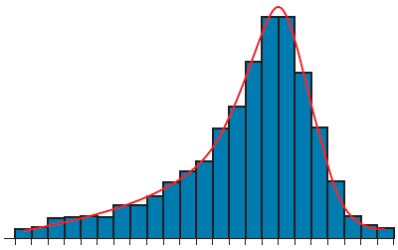
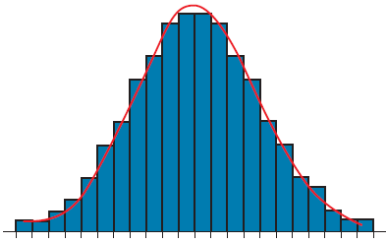
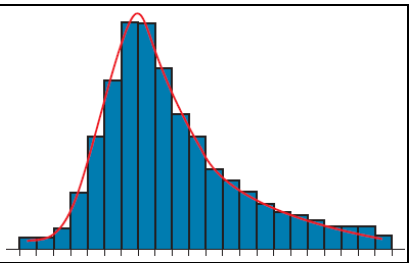
Standard normal model; This is believed to be in the public domain, however if there are omissions or inaccuracies please inform us so that any necessary corrections can be made

Therefore if the 50 Schmidt hammer hardness values obtained at Alderley Edge fit the standard normal model then we could say that 68% of the data lies between 37 ± 4, 95% of the data lies between 37 ± 8 and just over 99% of the data lies between 37 ± 12.

Although many large populations follow the standard normal model many samples of data do not. A quick comparison of the bell curve to the histogram produced earlier shows that this is the case with the collected Schmidt hammer hardness data in that the curve is not symmetrical and indeed the mean, mode and median are not coincident. A measure of how a sample set differs from the standard normal model can be made by calculating the coefficient of skewness.

Measures of shape

Skewness: is a term used to describe the degree of asymmetry of a set of data from the normal distribution. Whether a sample is symmetrical or skewed to the left (negative skew) or to the right (positive skew) is clearly shown in a histogram.

		
A negative skew. The tail of the data extends to the left (in a negative direction). In such a case the median is usually larger than the mean.	No skew. Data is perfectly symmetrical about the mean.	A positive skew. The tail of the data extends to the right (in a positive direction). In such a case the mean is usually larger than the median.

There are many different formulae for calculating skewness. A simple and convenient formula is:

$$\text{skew} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

For the Schmidt hammer data the coefficient of skew is $= \frac{3(36.88 - 36)}{4.13} = +0.64$

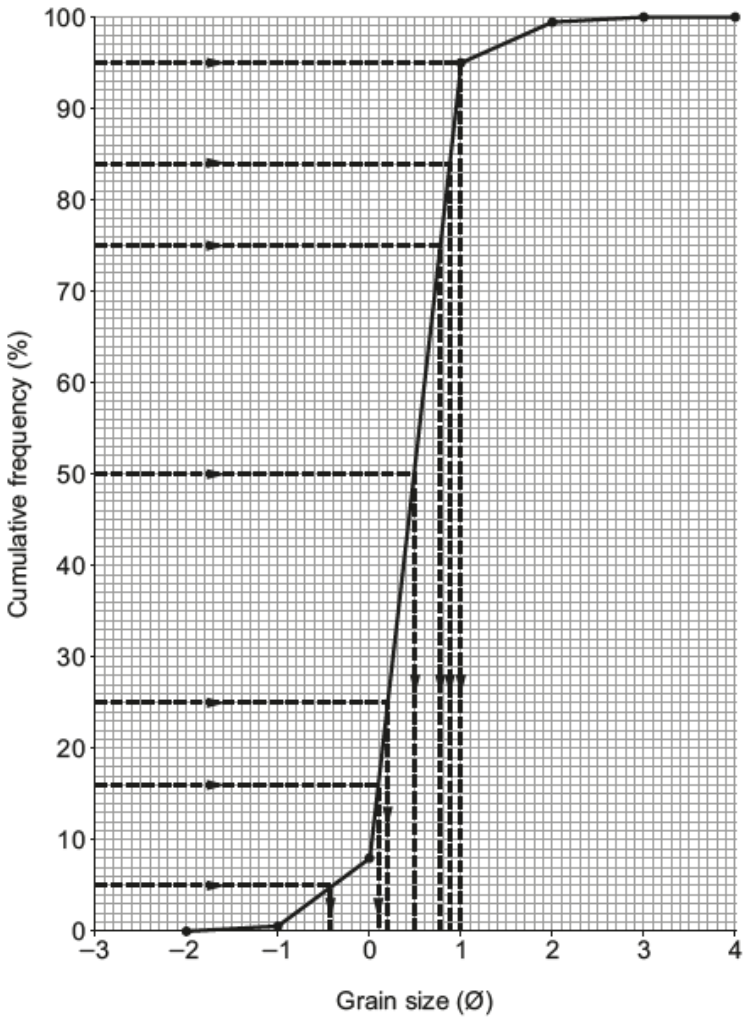
This positive skew can be seen by the obvious tail to the right in the data displayed on both the bar chart and histogram.

Example: statistical analysis of a sieved sediment

Sieving of unconsolidated sediment is a common practical exercise at A level. Presented below are the results for a sieved modern beach sand. These data illustrate the usefulness of constructing a percentage cumulative frequency graph.

Here the percentage cumulative mass is calculated by adding the percentage mass values as you go along. It is conventional (as the table shows) to add the coarsest sediment mass to subsequent finer class intervals to give the 'running total'. This is important to note because when plotting a percentage cumulative frequency graph (shown below) the points are plotted at the upper class boundary because the table gives the successive totals that are less than this upper class boundary.

grain size (mm)	grain size (ø)	mass (g)	mass (%)	cumulative mass (%)
4≤mm<8	−2≥ ø>−3	0.0	0.0	0.0
2≤mm<4	−1≥ ø>−2	0.3	0.3	0.3
1≤mm<2	0≥ ø>−1	7.0	7.8	8.1
0.5≤mm<1	1≥ ø>0	78.5	87.1	95.2
0.25≤mm<0.5	2≥ ø>1	4.0	4.5	99.7
0.125≤mm<0.25	3≥ ø>2	0.2	0.2	99.9
0.063≤mm<0.125	4≥ ø>3	0.1	0.1	100.0
0.032≤mm<0.063	5≥ ø>4	0.0	0.0	100.0



In the percentage cumulative frequency diagram above the points are joined by a straight line. Although there is no steadfast rule for whether this line (ogive) should be straight or curved in a percentage frequency diagram, the advantage of a straight line ensures greater consistency in reading extrapolated grain size values from stipulated percentile values when calculating the statistics of the grain size distribution as demonstrated below.

percentile	grain size (ø)
5	−0.40
16	0.10
25	0.20
50	0.50
75	0.75
84	0.85
95	1.00

Mode: the modal class is evident from the table – $0.5 \leq \text{mm} < 1$, i.e. the sediment is a coarse sand.

Median: the median is the phi value at the 50 percentile ($ø_{50}$), i.e. $ø_{50}$ = 0.50.

To convert this into mm, $2^{-0.50}$ = 0.71 mm.

Mean: the graphic mean is calculated from the formula,

$$\bar{x} = \frac{ø_{75} + ø_{50} + ø_{25}}{3} = \frac{0.75 + 0.50 + 0.20}{3} = 0.48ø$$

To convert this into mm, $2^{-0.48}$ = 0.72 mm.

Skewness: It is readily evident from the above values that the three averages are very close to each other and therefore it would be expected that this sediment will not be significantly skewed. The grain sizes of the sediment would therefore approximate to a normal distribution. This can be confirmed by calculating the graphic skewness of the sediment using the formula:

$$\text{skew} = \frac{(ø_{84} - ø_{50})}{(ø_{84} - ø_{16})} - \frac{(ø_{50} - ø_{5})}{(ø_{95} - ø_{5})} = \frac{(0.85 - 0.50)}{(0.85 - 0.10)} - \frac{(0.50 - (-0.40))}{(1.00 - (-0.40))} = 0.46 - 0.64 = -0.18$$

Using the table of descriptive terms for skewness shown below, then the sediment can be described as (slightly) negatively skewed i.e. there is a slightly coarse tail to the grain size distribution.

skewness descriptor	**graphical skewness value**
very negatively skewed	−1.0 to −0.3
negatively skewed	−0.3 to −0.1
symmetrical	−0.1 to 0.1
positively skewed	0.1 to 0.3
very positively skewed	0.3 to 1.0

Standard deviation: calculation of the standard deviation of the grain size distribution is used to calculate the sorting of the sediment using the following formula,

$$\bar{x} = \frac{\phi_{84} - \phi_{16}}{2} = \frac{0.85 - 0.10}{2} = 0.38$$

Using the table of descriptive terms for sorting shown below, then the sediment can also be described as well sorted.

sorting descriptor	graphical sorting value
very well sorted	<0.35
well sorted	0.35 - 0.50
moderately well sorted	0.50 - 0.70
moderately sorted	0.70 - 1.00
poorly sorted	1.00 - 2.00
very poorly sorted	2.00 - 4.00
extremely poorly sorted	>4.00

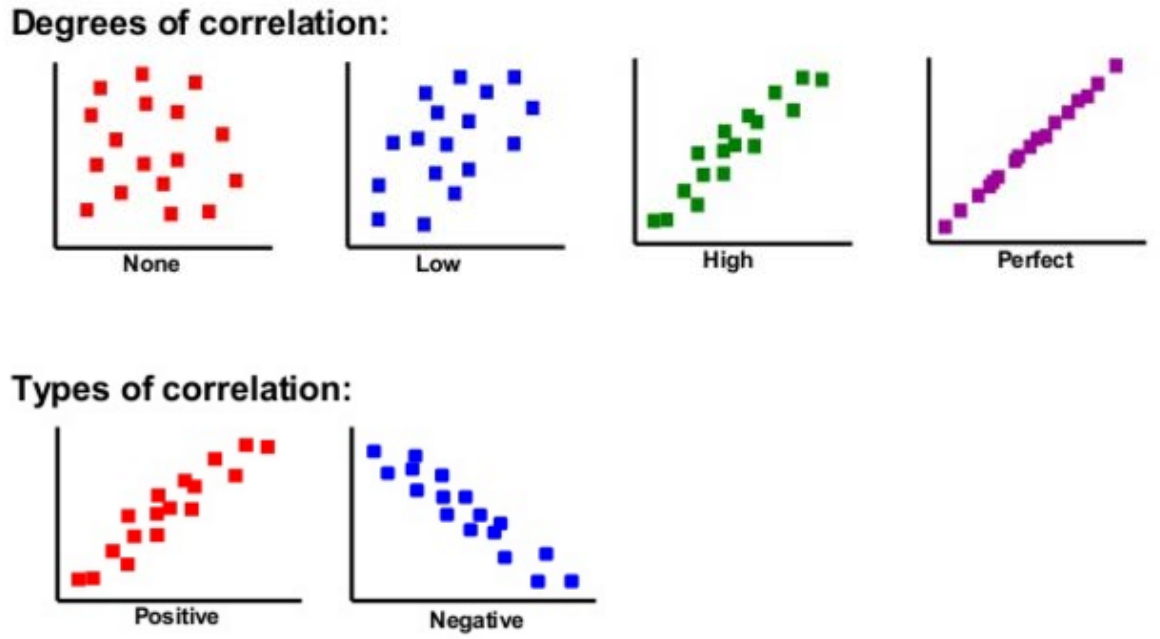
Bivariate data analysis

A dataset that contains two variables is termed bivariate data. In Geology there is often an interest in comparing two measurements made for the same site (e.g. in an outcrop – fracture spacing and permeability) or same object (e.g. in a hand specimen – porosity and density). Among the many commonly used graphical techniques used to analyse and display bivariate data perhaps the most frequently utilised is the scatter diagram.

Scatter diagrams

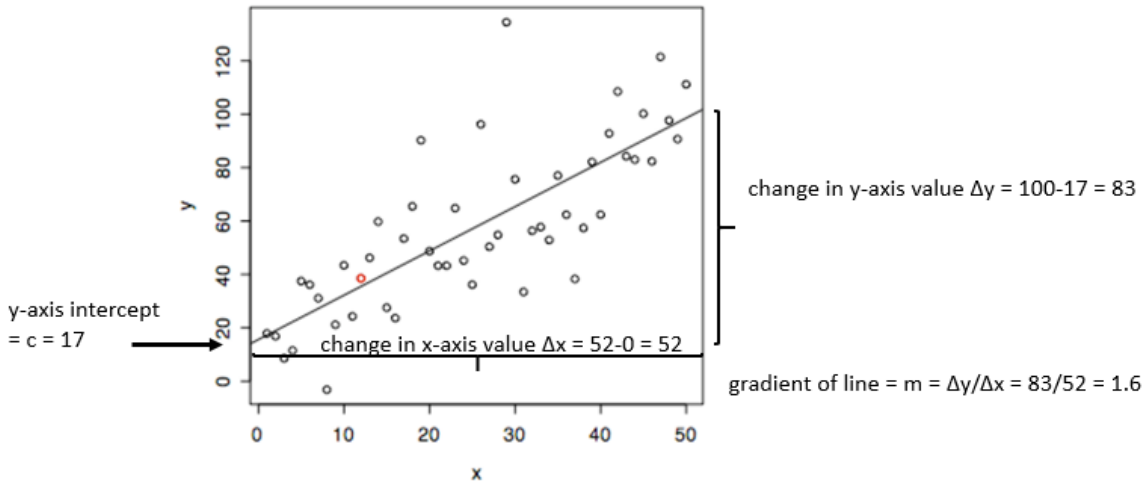
Scatter diagrams are used to show graphically the relationship between two variables. Two axes are drawn in the usual way with the variable that is believed to cause the change in the other (the so-called independent variable) plotted on the x-axis; the dependent variable is therefore plotted on the y-axis.

By studying the resulting pattern of the pairs of data on the scatter diagram the degree of correlation may be evident. Correlation gives an idea of how strong the linear relationship between the bivariate data is (e.g. for the curved data no correlation exists). Commonly encountered patterns include:



Scatter Diagram - correlation; ABB Group goo.gl/PhgWx4

It is very common for graphs of the relationship between pairs of geological variables to be well approximated by straight lines. However, the fit is never perfect. Despite this fact it may be possible to draw in by eye, a best fit straight line, which should appear to pass as close as possible to all the points plotted (with care taken to exclude obvious anomalies). The best fit straight line does not need to pass through the origin but it is good practice that the line of best fit should pass through the double mean point $(\bar{x}, \bar{y})$ i.e. the point that is the mean of x values: mean of y values. A generic example of a best fit straight line is shown on the next page.



equation of straight line; y = mx + c; y = 1.6x + 17

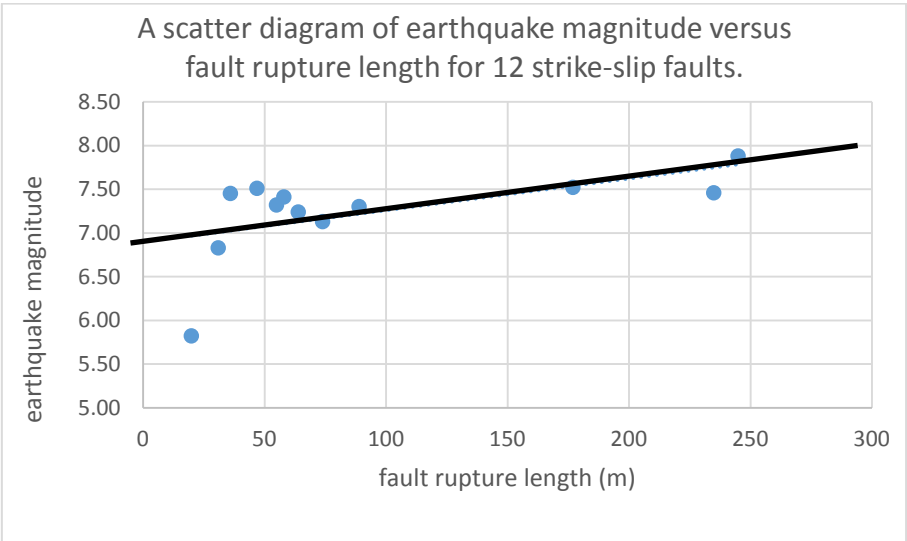
Once the best fit straight line is drawn in by eye it is possible to obtain predictions of unknown values. This may be undertaken directly from the graph or more accurately by obtaining the equation of the straight line. The general equation of a straight line is:

$$y = mx + c$$

where m is the gradient of the line and c is the y-axis intercept.

In the example above the equation of the best fit straight line is y = 1.6x +17, hence if a value of y at x = 42 is required then y = (1.6 x 42) + 17 = 84. Care must be exercised in the prediction of values outside of the graph area in that there may be a degree of uncertainty whether this mathematical relationship would still hold true.

The scatter diagram below represents a set of fault rupture length (independent variable) and magnitude (dependent variable) data for twelve globally-distributed strike-slip fault earthquakes.



Immediate inspection of the scatter graph indicates that at low values of fault rupture length the degree of correlation worsens and this may suggest a non-linear relationship or the influence of another variable.

A rate (r) is calculated by determining the amount of change (for example, distance travelled) and the time elapsed. To do this, we need two values for time (t_1 and t_2) and two corresponding values for the condition that is changing (d_1 and d_2). So for example:

$$r = \frac{d_2 - d_1}{t_2 - t_1}$$

where d_2 is the distance travelled at time t_2 and d_1 is the distance travelled at time t_1. The Greek letter Δ, "delta," means change, and you may often see it used in rate calculation problems. Written using delta, our example rate equation becomes:

$$r = \frac{\Delta d}{\Delta t}$$

This simple linear algebraic equation can be applied to calculate the rate (speed) of plate motion. In 2006, geologists working with the Plate Boundary Observatory Network began closely tracking the location of a GPS station west of the San Andreas Fault in California. The station, which is located on the Pacific Plate, is moving slowly northwest past the North America Plate. In May 2007, researchers recorded the station 33 mm northwest of its original position. In May 2012, they recorded it 195 mm northwest of its original position. To calculate the rate of motion of the station (and thus the Pacific Plate) between 2007 and 2012 firstly the total displacement (Δx) needs to be determined,

$$\Delta x = x_2 - x_1$$

$$\Delta x = 195 - 33 = 162 \text{ mm}$$

Since the time period of interest is between 2007 and 2012, we know that Δt = 5.00 years.

Therefore:

$$r = \frac{\Delta x}{\Delta t}$$

$$r = \frac{162}{5.00} = 32.4 \text{ mm/year}$$

So between 2007 and 2012 the Pacific Plate has a rate of motion (speed) of 32.4 mm/year or a velocity of 32.4 mm/year to the northwest.

If the plate continues to move at the same rate in the same direction, then it is possible to calculate how far it will be from its original (May 2006) position. Therefore, by May of 2050, rearranging the equation:

$$r = \frac{\Delta x}{\Delta t}$$

to make Δx the subject of the equation so that:

$$\Delta x = r \times \Delta t$$

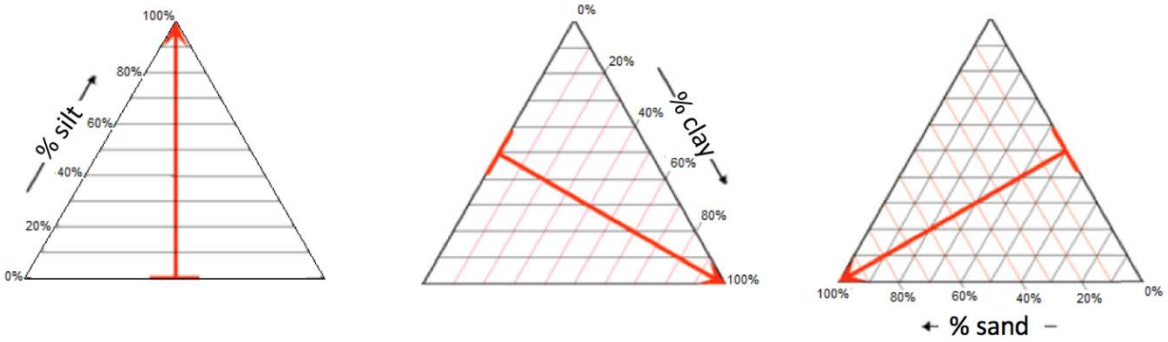
then, Δx = 32.4 × (2050 – 2006) = 32.4 × 44.00 = 1 430 mm

Therefore by May 2050, the station will have moved 1.43 m if the speed and direction remained constant.

Multivariate data analysis

Triangular plots

Triangular (ternary) diagrams have three axes instead of two and are useful for visualising the relative proportions of three components in a sample. Examples in geology of triangular plots include quartz-feldspar-rock fragments diagrams and sand-silt-clay composition diagrams in the study of sedimentary rocks.



With triangular graphs each axis is divided into 100 – representing percentages.
From each apex lines are drawn at an angle of 60^{o} to carry the values across the graph.
The data must be in the form of three percentage values and these values must add up to 100.

The examples above show how the relative proportions of silt, clay and sand vary with respect to each apex of the triangle. Examples are shown below of four samples (1, 2, 3 and 4) with different proportions of silt, clay and sand to illustrate how these values would plot on the triangular graph.

No	Silt	Clay	Sand
1	60%	20%	20%
2	10%	20%	70%
3	25%	35%	40%
4	0%	75%	25%