GCE A LEVEL

eduqas
Part of WJEC

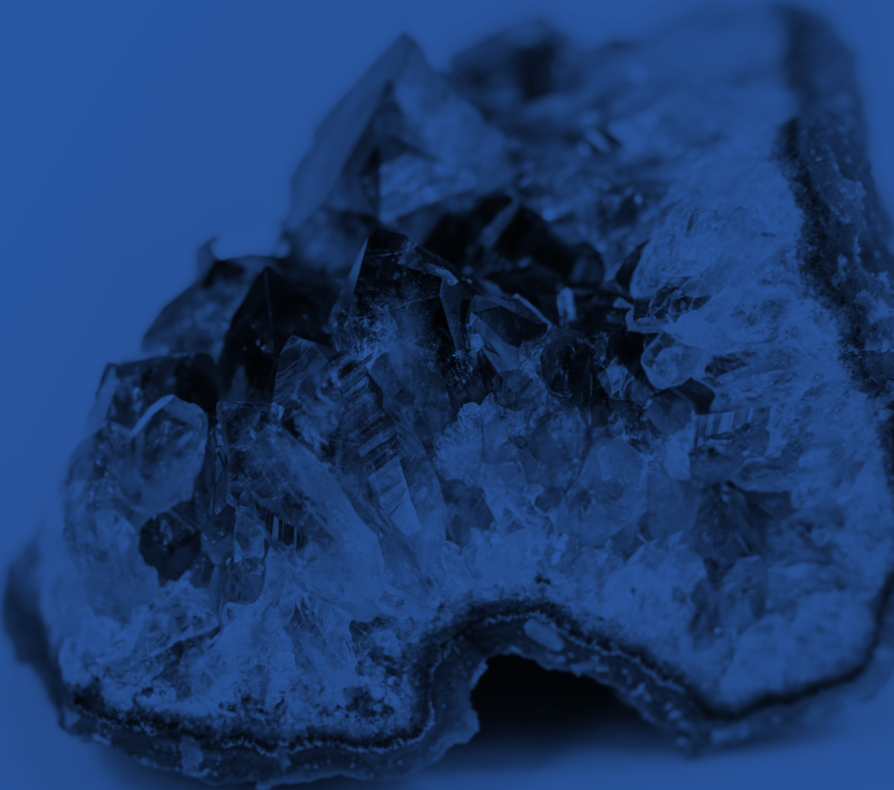
WJEC Eduqas GCE A LEVEL in

# GEOLOGY

ACCREDITED BY OFQUAL
DESIGNATED BY QUALIFICATIONS WALES

# Mathematical Guidance for GCE A LEVEL Geology

Teaching from 2017
For award from 2019



wjec
cbac

## Contents

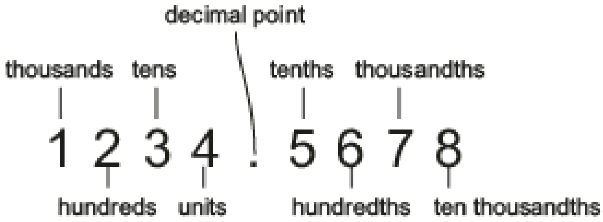
## Introduction

Geology is often falsely described as a qualitative (i.e. purely descriptive) science. However, many of the topics covered in A level 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 A level 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 = $3.91 \times 10^{-6}$ 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 (nm) |
| $10^{-6}$ | micro | µ | micrometre (µm) |
| $10^{-3}$ | milli | m | millimetre (mm) |
| 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 $3.91 \times 10^{-6}$ 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:

$8.200 \times 10^3$ has four significant figures; $8.20 \times 10^3$ has three significant figures;
$8.2 \times 10^3$ 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 $\text{cm}^3$ (three significant figures). The density of the pebble is $\frac{276.5}{105} =$ 2.63 $\text{g/cm}^3$ (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 ($2.37 \times 10^2$) and 823 ($8.23 \times 10^2$) 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 ($1 \times 10^{-6}$ 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 ($1 \times 10^0$ 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 ($4 \times 10^4$) whereas a typical optical laboratory microscope may produce a magnification of just 40 ($4 \times 10^1$). 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.

# Uncertainty in measurements

Uncertainty in measurements is unavoidable and estimates the range within which the answer is likely to lie. This is usually expressed as an absolute value, but can be given as a percentage.

## Single and repeat measurements

The normal way of expressing a measurement $x_0$, with its uncertainty, $u$, is $x_0 \pm u$. This means that the true value of the measurement is likely to lie in the range $x_0 - u$ to $x_0 + u$.

When working with a single reading it is recommended that the uncertainty is taken to be ± the smallest measuring division which of course depends on the scale of the measuring instrument. Suggestions for everyday measuring instruments ±1 mm for a rule, ±1 g for a standard mass, ± 1 °C for a laboratory thermometer and ± 1 cm³ for a typical measuring cylinder but professional judgement is important. Indeed sometimes the experimenter may feel that the uncertainty of a single reading can be made to ½ the smallest division of the instrument, however, more commonly the uncertainty will be larger than this because of other factors e.g. measuring the dip magnitude of an undulating bedding plane in a force 9 gale! Occasionally the uncertainty in a single value may be provided e.g. the absolute age of a particular rock unit e.g. the radiocarbon $^{14}$C age of 11 750 ± 120 year B.P. for an organic lake sediment.

Percentage uncertainty is a measure of the uncertainty of a measurement compared to the size of the measurement, expressed as a percentage. The calculation is derived by dividing the uncertainty of the experiment into the total value of the measurement and multiplying it by 100. The percentage uncertainty in the radiocarbon $^{14}$C age of 11 750 ± 120 year B.P. for an organic lake sediment is therefore:

% uncertainty, $p = \dfrac{120}{11750} \times 100 = 1.0\%$

In geology, it is of course best practice to undertake repeat measurements whenever possible. Suppose the value of a quantity $x$ is measured several times and a series of different values obtained: $x_1$, $x_2$, $x_3$……..$x_n$. Unless there is reason to suspect that one of the results is anomalous, the best estimate of the true value of $x$ is the arithmetic mean of the readings:

mean value $\bar{x} = \dfrac{x_1 + x_2 + ........x_n}{n}$

A reasonable estimate of the uncertainty in this set of readings is ½ the range:

i.e. $u = \dfrac{x_{\max} - x_{\min}}{2}$ , where $x_{\max}$ is the maximum and $x_{\min}$ the minimum reading of $x$.

For example, the following results were obtained for the thickness of a sedimentary bed; 4.5 cm, 4.8 cm, 4.6 cm, 5.1 cm, 5.0 cm. The best estimate of the true thickness is given by the mean which is:

$$\bar{t} = \frac{4.5 + 4.8 + 4.6 + 5.1 + 5.0}{5} = 4.8\text{ cm}$$

The uncertainty, $u$, is given by: $u = \dfrac{5.1 - 4.5}{2} = 0.3\text{ cm}$

The final answer and uncertainty should be quoted, with units, to the same no. of decimal places, i.e. bed thickness = 4.8 ± 0.3 cm. The percentage uncertainty in the mean bed thickness is therefore 6.3%.

## Combining uncertainties

Very frequently in geology, the values of two or more quantities are measured and then these are combined to determine another quantity; e.g. the density of a material is determined using the equation: $\rho = \dfrac{m}{V}$

To do this the mass, $m$, and the volume, $V$, are first measured. Each has its own estimated uncertainty and these must be combined to produce an estimated uncertainty in the density. In most cases, quantities are combined either by multiplying or dividing and this will be considered first. The percentage uncertainty in a quantity, formed when two or more quantities are combined by either multiplication or division, is the sum of the percentage uncertainties in the quantities which are combined.

The following results were obtained when measuring the density of a perfect pyrite cube with a top pan balance, resolution 0.1 g, and 30 cm rule with a hand lens, resolution 0.05 cm (i.e improved to half a subdivision).

Mass = $6.3 \pm 0.1\text{g}$ so % uncertainty, $\rho_m = \dfrac{0.1}{6.3} \times 100 = 1.6\%$

Length of cube side $= 1.1 \pm 0.05\text{cm}$, so % uncertainty, $\rho_l = \dfrac{0.05}{1.1} \times 100 = 4.5\%$

The volume of the specimen is length$^3$ = 1.1$^3$ = 1.3 cm$^3$

The percentage uncertainty in $x^n$ *is* $n$ times the percentage uncertainty in $x$. Therefore the percentage uncertainty in the volume $= 3 \times 4.5 = 14\%$. So the volume of the specimen is 1.3 cm$^3$ ± 14%. The density of the pyrite cube is $\rho = \dfrac{m}{V}$ The percentage uncertainty in the density is therefore = 1.6 + 14 = 15%. So $\rho = \dfrac{m}{V} = \dfrac{6.3}{1.1^3} = 4.7 \pm 15\%$ or 4.7 ± 0.7 g/cm$^3$.

If two or more quantities are added or subtracted the absolute uncertainties are added. If we consider the equation for relative density where we are only required to use a top pan balance:

$$\text{Relative density} = \frac{\text{weight in air}}{(\text{weight in air} - \text{weight in water})} \approx \frac{\text{mass in air}}{(\text{mass in air} - \text{mass in water})}$$

(Note here that as weight = mass × gravitational field strength and $g$ is constant then $g$ cancels through and values of mass can be directly inserted into the revised equation. Additionally the units of the denominator and numerator (g) cancel so relative density is dimensionless).

An irregular sample of impure galena had a mass in air, $m_a$, = 130.5 ± 0.1 g and a mass in water, $m_w$, = 112.8 ± 0.1 g hence the difference in the mass of the sample measured in air and water = 17.7 ± 0.2 g. This is a percentage uncertainty of: $\frac{0.2}{17.7} \times 100 = 1.1\%$ Additionally the percentage uncertainty in the mass in air of the sample is $\frac{0.1}{130.5} \times 100 = 0.1\%$

The density of the impure galena sample is therefore $\frac{130.5}{17.7} = 7.37 \pm 1.2\%$ or $7.37 \pm 0.09 \text{g} / \text{cm}^3$

## 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.

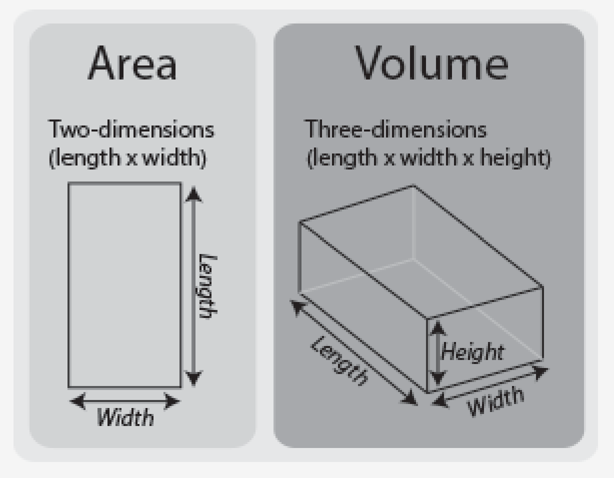
| *Mineral/ hardness* | | *Common equivalent* |
|---|---|---|
| Diamond | 10 | |
| Corundum | 9 | |
| Topaz | 8 | |
| Quartz | 7 | |
| Orthoclase feldspar | 6 | ← steel pin |
| Apatite | 5 | |
| Fluorite | 4 | ← copper coin |
| Calcite | 3 | ← finger nail |
| Gypsum | 2 | |
| Talc | 1 | |

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).

## Calculating circumference, surface area and volumes of regular shapes

The ability to calculate the perimeter, circumference, area, surface area and volume of various types of regular shapes is needed in many areas of geology and is especially important in resource geology where a quantification of exploitable reserves is essential. Formulae for calculating these parameters are given below.

**Rectangular-based solids.**



Area formula; Skills You Need goo.gl/wmvkv3

Volume formula; Skills You Need goo.gl/wmvkv3

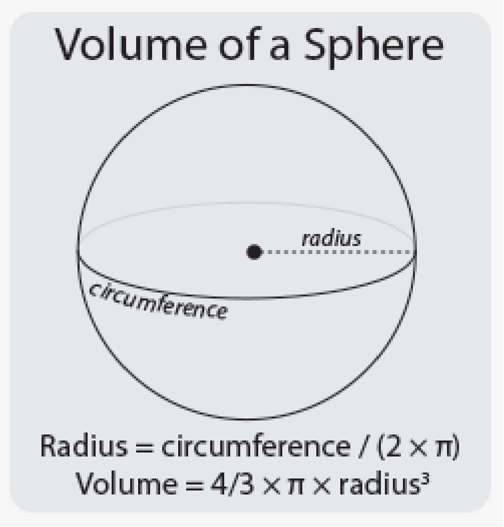
The perimeter of the rectangle = 2 × (length + width)

The area of the rectangle = length × width

The volume of the cuboid = length × width × height

The surface area of the cuboid = 2 × ((length × width) + (width × height) + (length × height))

## Circles and spheres



Volume of a Sphere; Skills You Need goo.gl/wmvkv3

The circumference of the circle $= 2 \times \pi \times \text{radius}$

The area of the circle $= \pi \times (\text{radius})^2$

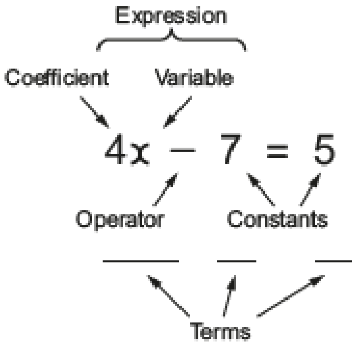
The volume of the sphere = $\frac{4}{3} \times \pi \times (\text{radius})^3$

The surface area of the sphere = $4 \times \pi \times (\text{radius})^2$

The surface area and volume of more complex regular solids (e.g. closed cylinder and cones) and indeed some irregular solids can often be determined by combining some of the above relationships. A website that documents examples of such calculations is www.skillsyouneed.com/num/volume.html

# Manipulating algebraic equations

An equation is a statement that the values of two mathematical expressions are equal. There are numerous parts to an equation as exemplified in the simple example below:



## Linear algebraic equations

To solve a simple linear algebraic equation it is necessary to isolate the variable by performing the same inverse operations to both sides of the equation to make the variable the subject of the equation.

e.g. $4x - 7 = 5$

$4x = 12$

$x = 3$

A geological example of a simple linear algebraic equation is determining geothermal gradient ($G$; $^{o}$C/m) from borehole temperature data. Here:

$$G = (T_d - T_s) \div d$$

(where $T_s$ is the mean surface temperature in $^{o}$C and $T_d$ the temperature at a depth $d$ below the surface in $m$) has more variables but can be rearranged for $T_d$ in a similar fashion.

$$G \times d = T_d - T_s$$

$$T_d = (G \times d) + T_s$$

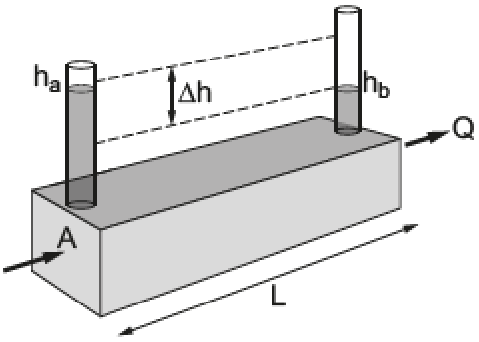
If values are known, e.g. $T_s$ = 11 $^{o}$C, $G$ = 0.046 $^{o}$C/m and $d$ = 450 m then substituting we get the predicted subsurface temperature, $T_d$ as:

$$= (0.046 \times 450) + 11 = 32\,^{o}\text{C}$$

Another equation regularly used in various branches of geology is Darcy's Law. In its simplest form the Law which relates the flow of a fluid through a permeable medium can be written as:

$$Q = -KA\left(\frac{h_a - h_b}{L}\right)$$

where $Q$ is the total fluid discharge in m$^3$/s, $A$ is the cross-sectional area in m$^2$, $h_a$ and $h_b$ record the heights, in m, the flowing water reaches in thin piezometric tubes inserted into the sand to measure the pressure difference (hydraulic head) over length $L$, in m along which the fluid passes. The term in brackets is known as the hydraulic gradient and is dimensionless and $K$ is a constant called the hydraulic conductivity with units m/s. The negative sign satisfies a convention such that the direction of flow is toward lower hydraulic head.



It is important to note that the constant $K$ is not the permeability. It represents both the properties of the permeable medium as well as the properties of the fluid flowing through the permeable medium.

In order to calculate the value of hydraulic conductivity it is firstly necessary to rearrange Darcy's Law making $K$ the subject of the equation (note the negative sign has been dispensed with to avoid confusion):

$$K = \frac{(Q \times L)}{A(h_a - h_b)}$$

If we substitute exemplar figures from a very well sorted gravel in a tube, where $Q$ is measured as 0.000032 m$^3$/s and $A = \pi(0.050)^2$ (diameter of tube = 0.100 m) with the length of tube being 0.500m and the drop in height also being 0.500 m we obtain a value:

$$\mathrm{K} = \frac{(0.000032 \times 0.500)}{\pi(0.050)^2 \times (0.500)} = 4.1 \times 10^{-3} \text{ m s}^{-1}$$

## Algebraic equations involving indices

A number multiplied by itself is called a *square* and is written in the form *Number*$^{index}$. In this example, the value of the index is 2. If the value of the index was 3, we would be finding the *cube* of a number. When we cube a number, we are *raising it to the power of* 3. The *square root* of a number must be squared (that is multiplied by itself or raised to the power 2) to give the number. The *cube root* of a number must be cubed (that is raised to the power 3) to give the number. So the *cube root* of 64 is 4. We can write this as $^3\sqrt{}64$. The square and square root as well as cube and cube root are inverse operators.

An equation encountered in seismology is for $v_p$, the velocity of a P wave,

$$v_p = \sqrt{\frac{k + \frac{4}{3}\mu}{\rho}}$$

Where:
$\rho$ is the density
$k$ is the bulk modulus
$\mu$ is the shear (or rigidity) modulus

Making $k$ the subject of this formula, for example, involves:

1. Squaring both sides, to eliminate the square root on the RHS of the equation

   $v_p{}^2 = \dfrac{\left(k + \frac{4}{3}\mu\right)}{\rho}$ square and square root are inverse operations.

2. Multiplying both sides by ρ, to transfer ρ to the LHS of the equation:

   $v_p{}^2\rho = \left(k + \frac{4}{3}\mu\right)$ multiplication and division are inverse operations.

3. Subtracting $\frac{4}{3}\mu$ from both sides, to transfer $\frac{4}{3}\mu$ to the LHS of the equation

   so $v_p{}^2\rho - \frac{4}{3}\mu = k$ addition and subtraction are inverse operations.

There are several laws that are useful for the further manipulation of formulae involving powers (the index laws) but these will not be further discussed here.

## Algebraic equations involving logarithms

The logarithm of a number is the power to which the base must be raised in order to get the number:

Log of 64 if base of 2 is used $\log_2 64 = 6$ (because $2^6 = 64$)

Log of 64 if base of 4 is used $\log_4 64 = 3$ (because $4^3 = 64$)

Log of 64 if base of 8 is used $\log_8 64 = 2$ (because $8^2 = 64$)

If we write $10^2 = 100$, we can also say 'the logarithm of 100 to base 10 is 2'. We write this as $\log_{10} 100 = 2$.

Because we count in base 10, $\log_{10}$ is usually shortened to just 'log' (this is what appears on your calculator button) and is known as a 'common logarithm'.

We can write logarithms in bases other than 10. For example, in sedimentology the phi scale is used which uses base 2.

The number $e$ is 2.7183 (to 4 d.p).

If we use logs with a base of $e$, $\log_e$ , we write them as ln (not $\log_e$).
These are called natural logarithms.
So the natural log of 100 is ln100 = 4.61
(given by a calculator) because $e^{4.61} = 100$

Logarithms and natural logarithms are used in many geological calculations (e.g. radiometric dating).

Logarithms solve the problem of how to arrange an equation of the form $y = a^x$ into an equation for $x$ in terms of $y$.

$$y = a^x$$

(take the logs of both sides)

$$\log y = \log a^x$$

(use the law of logs: $\log m^n = n \log m$)

$$\log y = x \log a$$

$$\frac{\log y}{\log a} = x$$

The phi scale uses base 2 logarithms. The strict definition of the phi scale is $ø = -\log_2 d$, where $ø$ is the grain size in phi units and $d$ is the grain size in mm. Because the negative logarithm is used, the coarsest sizes have the lowest $ø$ values and because base 2 is used, each one $ø$ increase in size represents a halving of the diameter in mm. A simple conversion chart is shown below.

e.g. diameter = 32 mm

$$ø = -\log_2 32$$

$$= -5 \text{ (because } 2^5 = 32)$$

| mm | phi | Name | |
|---|---|---|---|
| | | Boulders | |
| 256 | -8 | | |
| 128 | -7 | | |
| 64 | -6 | Cobbles | |
| 32 | -5 | | Gravel<br>Conglomerate |
| 16 | -4 | | |
| 8 | -3 | Pebbles | |
| 4 | -2 | | |
| | | Granules | |
| 2 | -1 | | |
| | | Very coarse sand | |
| 1 | 0 | | |
| | | Coarse sand | |
| 0.5 | 1 | | |
| | | Medium sand | Sand<br>Sandstone |
| 0.25 | 2 | | |
| | | Fine sand | |
| 0.125 | 3 | | |
| | | Very fine sand | |
| 0.063 | 4 | | |
| | | Coarse slit | |
| 0.031 | 5 | | |
| | | Medium slit | |
| 0.0156 | 6 | | |
| | | Fine slit | Mud<br>Mudrock |
| 0.0078 | 7 | | |
| | | Very fine slit | |
| 0.0039 | 8 | | |
| | | Clay | |

However, say our value of $ø$ is not an integer as above? How do we calculate the equivalent grain size in mm? Using the equation $ø = -\log_2 d$, to make $d$ the subject of the equation:

$$ø = -\log_2 d$$

(multiply both sides by −1)

$$-ø = \log_2 d$$

(use log definition)

$$2^{-ø} = d$$

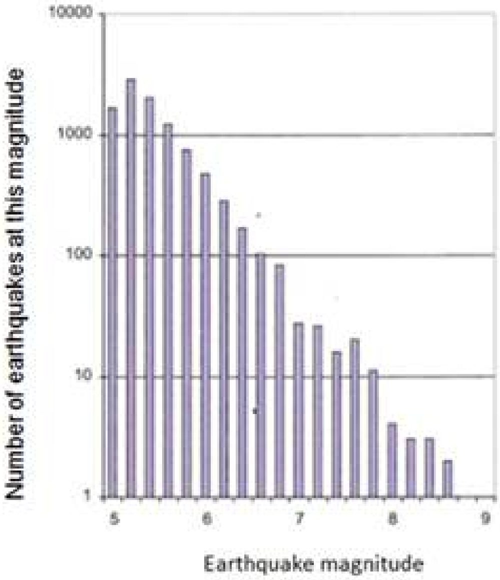
So if we take a value of $ø = 0.74$ then:

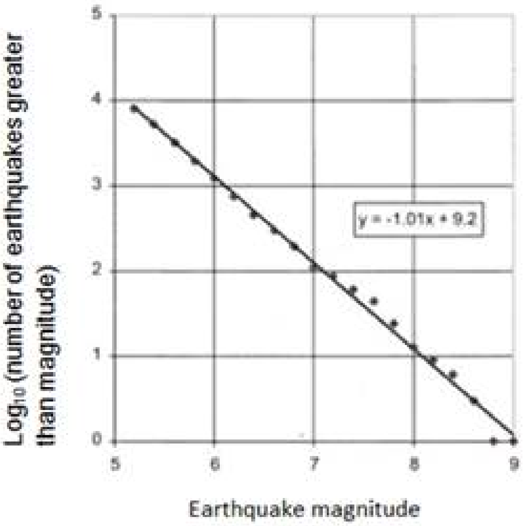
$$d = 2^{-0.74} = 0.60\,\text{mm}$$

There are two other principal uses of logarithms in geology:

- reducing exponential functions to simple straight line graphs of the form $y = mx + c$
- and compressing data sets that range over many orders of magnitude.

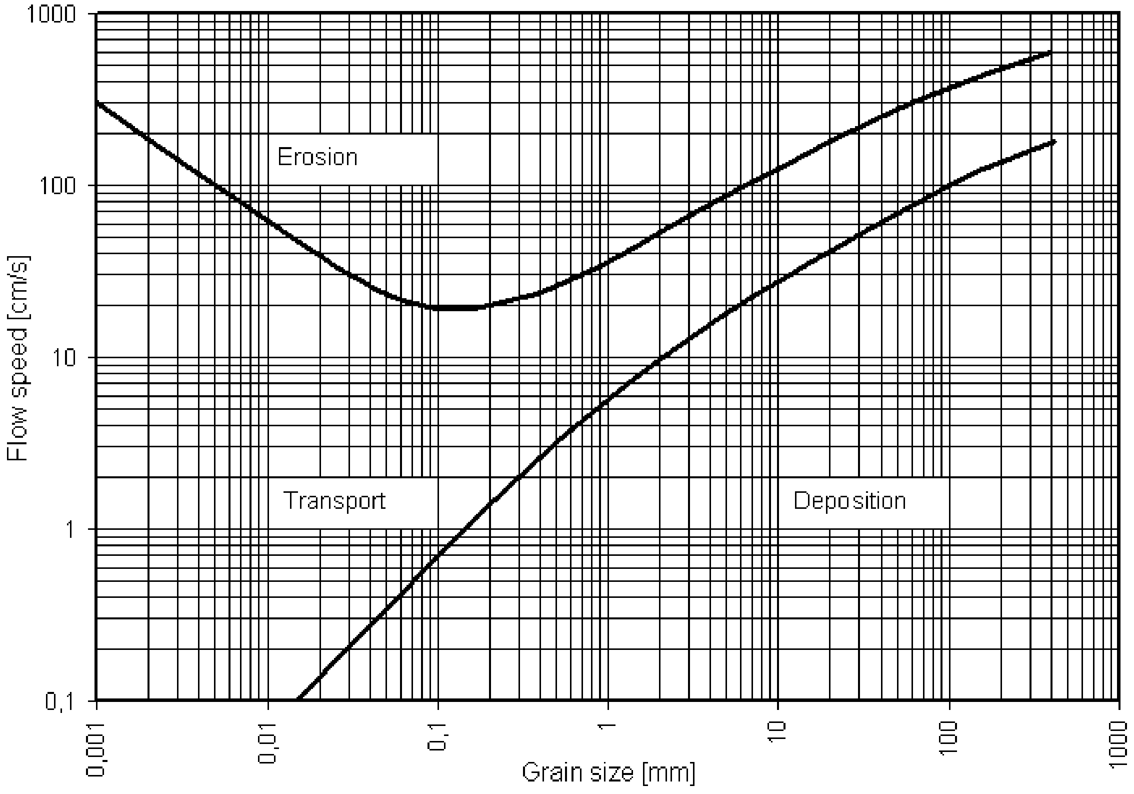
Both of these benefits are exemplified below by the bar chart and cumulative scatter graph for global earthquakes of magnitude 5 and over recorded by the USGS between 2000-2006.





The second graph above is known as a log ($y$-axis) normal ($x$-axis) graph. To read a value off the $y$-axis you have to read down the axis and take the inverse logarithm. For example, in the graph above if we want to calculate the number of seismic events of magnitude 6 or above, we obtain $\log_{10}$ (number of events) = 3.1, so the number of events is $10^{3.1} = 1\,300$.

In cases where both variables to be plotted have large ranges, log-log plots are produced – a classic example is the Hjlustrom plot shown below.



Hjulstrom curve; Wikimedia Creative Commons goo.gl/UWvzm1

The exponential function and logarithms are used in the determination of the absolute age of rock and mineral specimens. The fundamental starting point for this process is the decay rate equation:

$$N = N_0 e^{-\lambda t}$$

Where $N$ = number of unstable nuclei at time $t$

$N_0$ = number of unstable nuclei at time $t = 0$

$t$ = millions of years (My)

And $\lambda$ = the decay constant; a constant which depends on the radio-nucleide being used e.g. $K$, $U$ etc

The decay constant has a simple relationship with the half-life $\left(T_{1/2}\right)$ where $T_{1/2} = \frac{\ln 2}{\lambda}$

Substituting the equation for half-life into the decay rate equation and performing some simple rearrangements enables us to rewrite the decay rate equation as:

$$t = \left(\frac{T_{1/2}}{\ln 2}\right)\ln\left(\frac{N_d}{N_p}+1\right) \text{ or } \left(\frac{T_{1/2}}{0.693}\right)\ln\left(\frac{N_d}{N_p}=1\right)$$

where $t$ is the age of the specimen in millions of years:

$\frac{N_d}{N_p}$ = the ratio of daughter atoms $(N_d)$ to parent atoms $(N_p)$ at time $t$

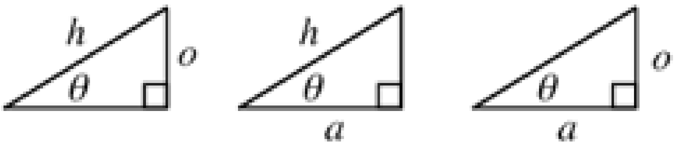
The validity of this equation can be seen by considering the case of a rock dated by the U-235 Pb 207 dating technique with a $T_{1/2}$ = 704 My. If four half-lives have elapsed the rock is obviously 4 × 704 = 2 820 My old. If we substitute the values into the equation we need to know the ratio $\frac{N_d}{N_p}$ after four half-lifes which is of course 15:1 or 15. Therefore:

$$t = \left(\frac{704}{0.693}\right) \times \ln(15+1) = 2\,820 \text{ My}$$

## Algebraic equations involving trigonometric functions

Trigonometry is concerned with the study of triangles and hence any other polygonal shape which can be constructed from triangles. An understanding of trigonometry is essential in geological map work where an appreciation of the length (outcrop width, true thickness) and angle (dip and strike) is necessary. However, trigonometry is applicable in all areas of geology.

The starting point for an appreciation of trigonometry is a right-angled triangle with the longest side (hypotenuse) always opposite the right angle and the two remaining sides being adjacent (next to) or opposite the angle of interest. This allows for the definition of three trigonometric functions; sine (sin), cosine (cos) and tangent (tan).



$$\sin\,\theta = \frac{o}{h} \qquad \cos\,\theta = \frac{a}{h} \qquad \tan\,\theta = \frac{o}{a}$$

$$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}.$$

These relationships enable the calculation of the true thickness of a layer from a map.

On flat ground, the relationship between true bed thickness ($t$), vertical thickness ($Vt$) (as would be measured in a borehole) and the angle of dip of the beds ($\theta$) can be used to calculate each variable.

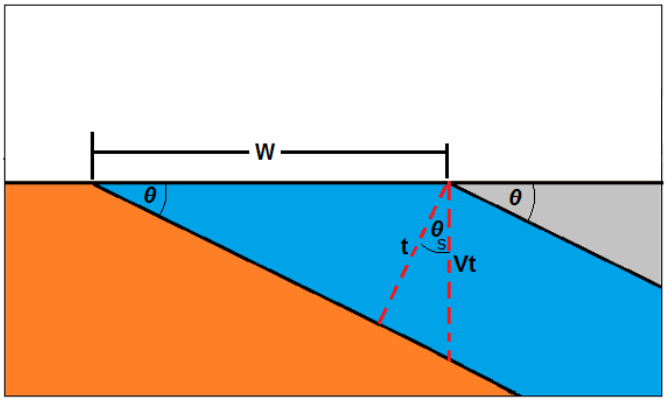
$\sin \theta = \frac{t}{w}$

so $t = w \sin \theta$

and

$\cos \theta = \frac{t}{Vt}$ and $t = Vt \cos \theta$

(as $\theta$ and $\theta_s$ are similar angles)



These can also be determined when the ground surface is not horizontal.

Given that the dip of the bed is $\theta$, the width of the outcrop on the map is $W$, and the difference in height between the outcrop of the top and bottom bedding planes is h then:

$Vt = h + (w \tan \theta)$

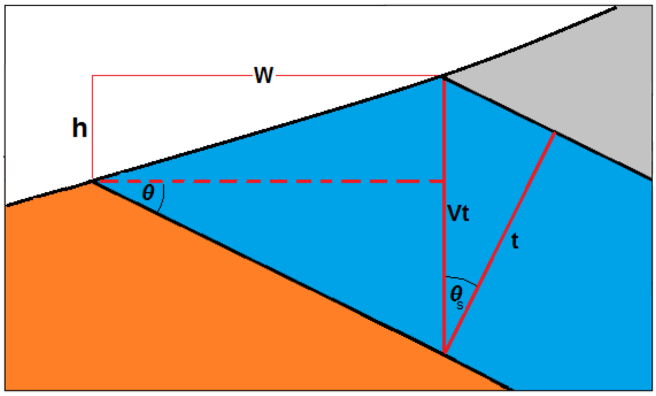
and

$t = Vt \cos \theta$

(as $\theta$ and $\theta_s$ are similar angles)

If the bed is vertical ($\theta = 90$) then

$Vt = w$



In the example on the right:

$\theta = 18^{o}$,

$h = 244 - 83 = 161$ m;

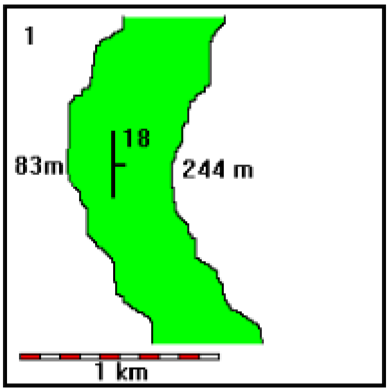
$W = 520$ m.

$t = VT \cos \theta$

Hence, true thickness =

(161 + (520 × tan18)) × cos18

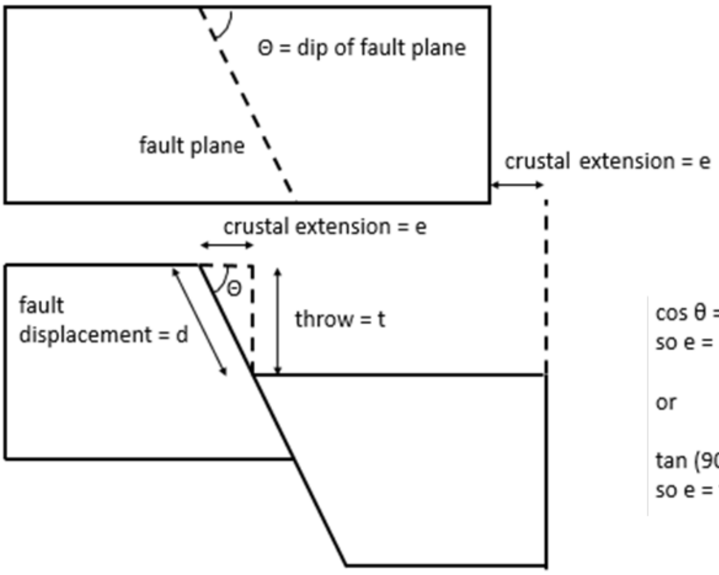
= (161 + 169) × 0.95 = 313.5 or 314 m.



Thickness of a layer from a map data; Steven Dutch, Natural and Applied Sciences goo.gl/XOpflV

The inverse trigonometric functions allow for the calculation of an angle from a sine, cosine or tangent. These inverse functions are called arcsine, arccosine and arctangent and are usually denoted in calculations as $\sin^{-1}$, $\cos^{-1}$ and $\tan^{-1}$.

Trigonometric functions can also be used to calculate the amount of crustal extension along a fault as demonstrated on the diagram below.



$\cos\theta = e/d$
so $e = d \times \cos\theta$

or

$\tan(90 - \theta) = e/t$
so $e = t \times \sin(90 - \theta)$

## 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.

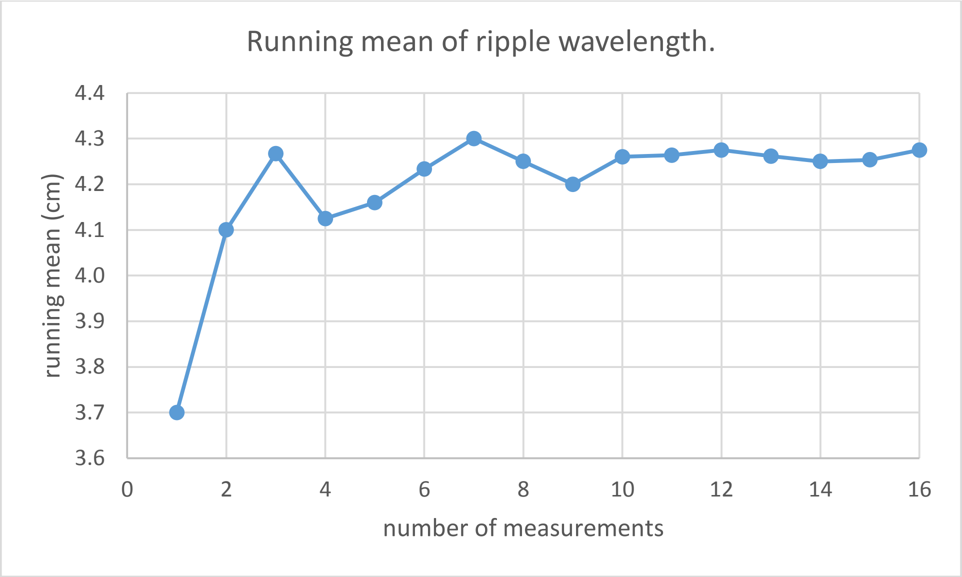
In **systematic sampling** there is some structure or underlying order to the way in which the data is selected. The fifty pieces of Schmidt hammer hardness data analysed later were sampled by use of a five by ten gridding system.

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.

| | ripple wavelength (cm) | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of measurements | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | running mean (cm) |
| 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.

# Data Analysis

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.

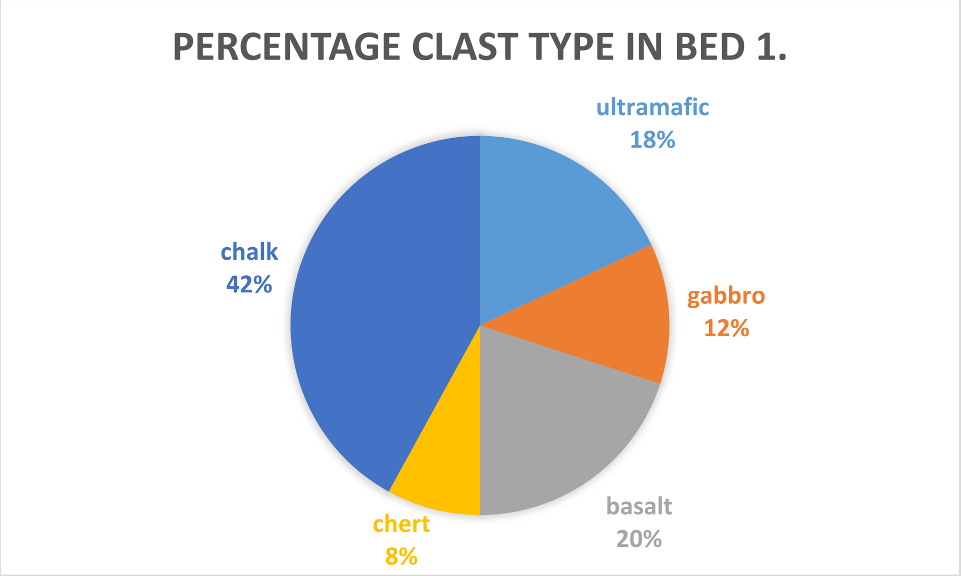
In the case of categorical variables (often called qualitative variables) frequency tables are constructed to produce pie charts and bar charts. In the case of numerical data (often called quantitative variables) frequency tables are constructed to produce:

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.

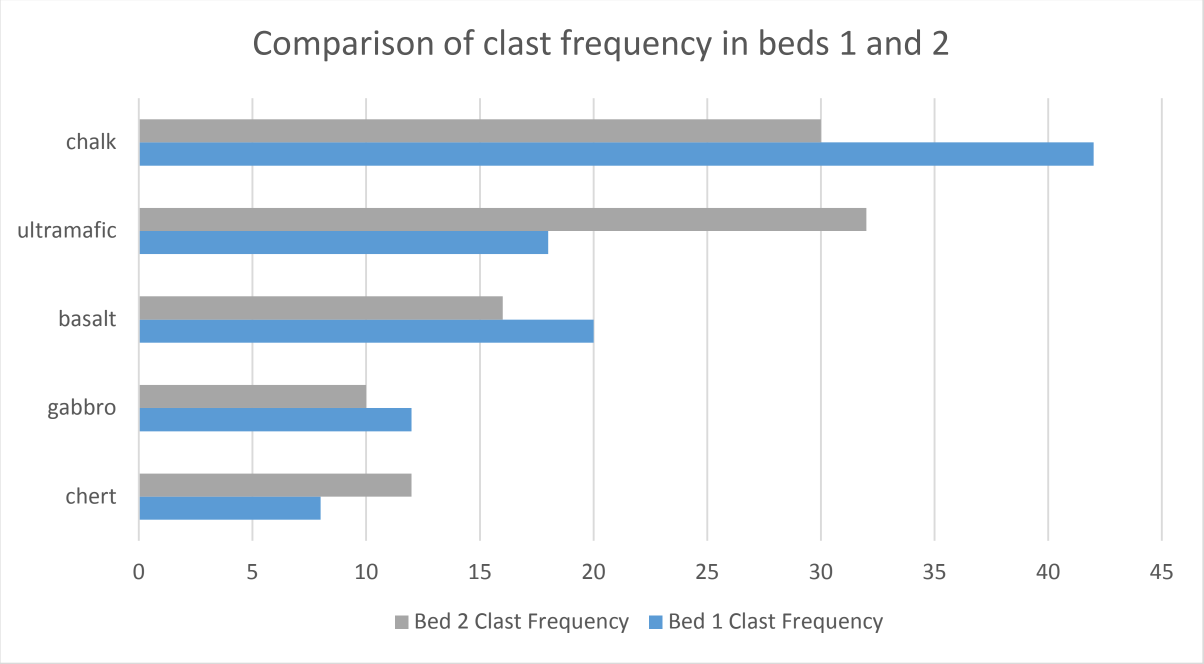
The following frequency table provides information on the lithology of clasts from two adjacent Pleistocene fluvial deposits near Paphos airport in south east Cyprus. The variable, clast type, is a categorical variable so the information in this table could be displayed either as a pie or bar chart.

| | 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.



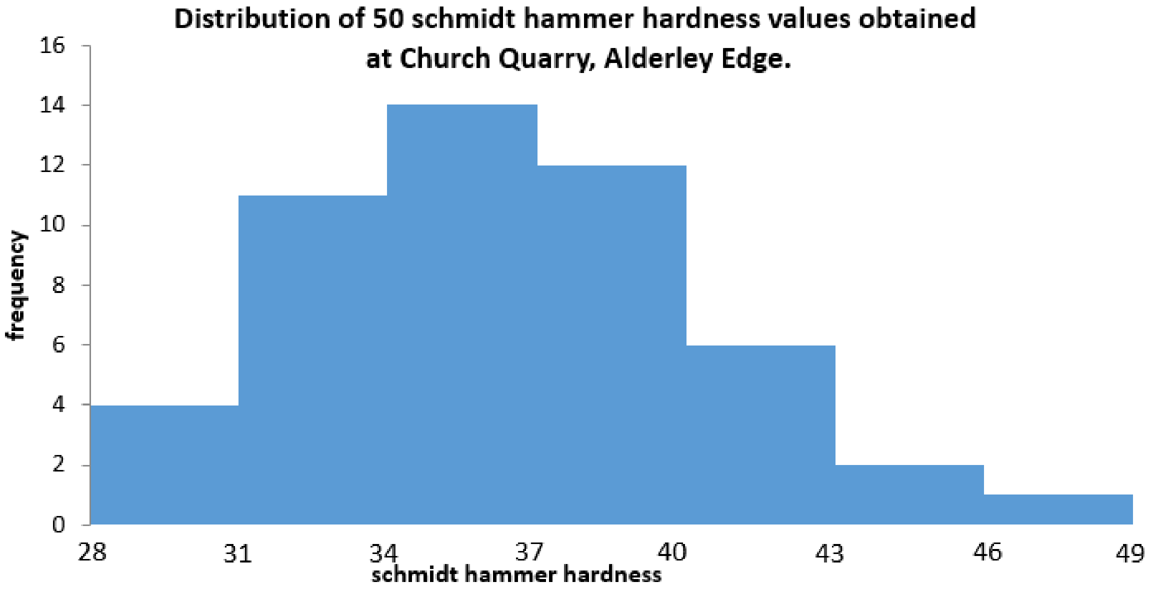
Note the gap between the variable categories and display of frequency of clasts on the bar chart.

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.

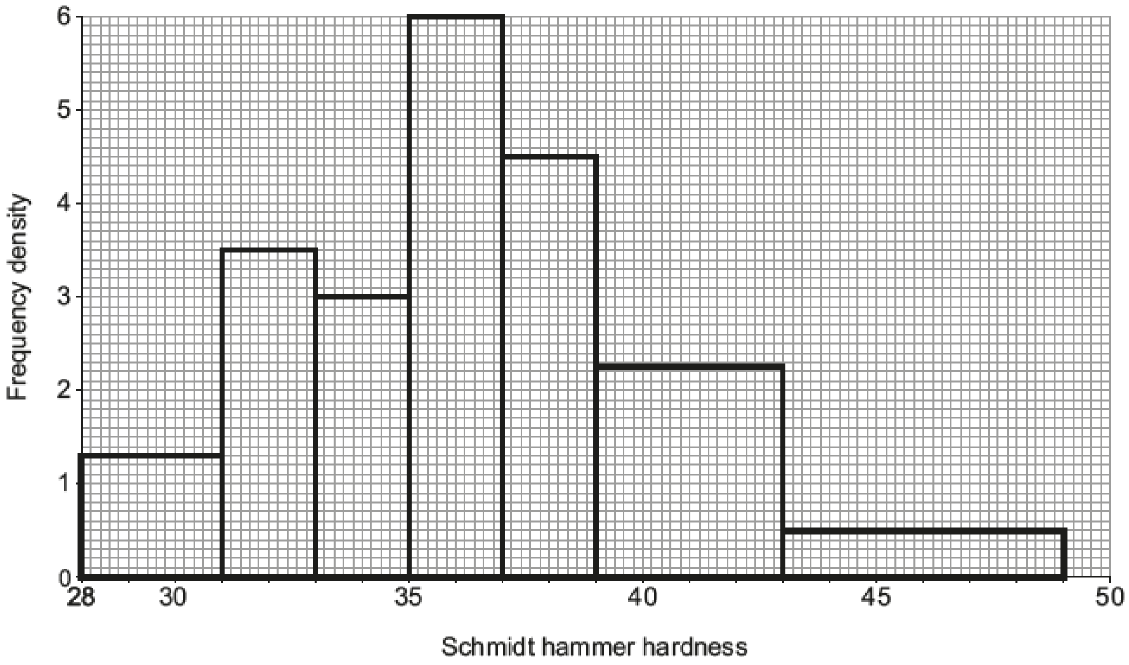
| Schmidt hammer hardness (SHH) class | | Frequency | Cumulative frequency, % |
|---|---|---|---|
| 1 | 28<SHH≤31 | 4 | 8 |
| 2 | 31<SHH≤34 | 11 | 30 |
| 3 | 34<SHH≤37 | 14 | 58 |
| 4 | 37<SHH≤40 | 12 | 82 |
| 5 | 40<SHH≤43 | 6 | 94 |
| 6 | 43<SHH≤46 | 2 | 98 |
| 7 | 46<SHH≤49 | 1 | 100 |


Distribution of 50 schmidt hammer hardness values obtained at Church Quarry, Alderley Edge.

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 $= \dfrac{\text{frequency}}{\text{class width}}$ |
|---|---|---|---|---|
| 1 | 28<SHH≤31 | 4 | 3 | $1.\dot{3}$ |
| 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.

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

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

**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 of 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.

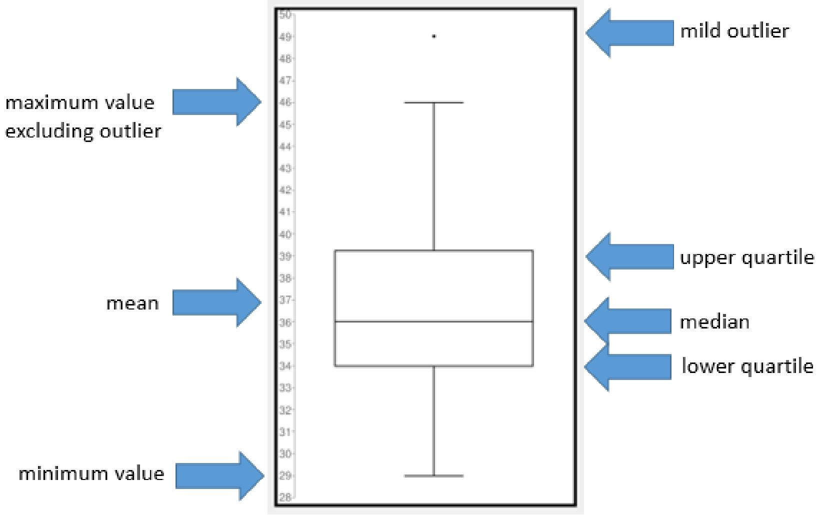
In the example above the IQR is 39 – 34 = 5. The mild outlier boundaries are

= 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 |
| …. | …. | …. | …. |
| | | | $\Sigma\left(x_i - \bar{x}\right)^2 = 837.3$ |
| | | | $\dfrac{\Sigma\left(x_i - \bar{x}\right)^2}{n - 1} = 17$ |

* 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.

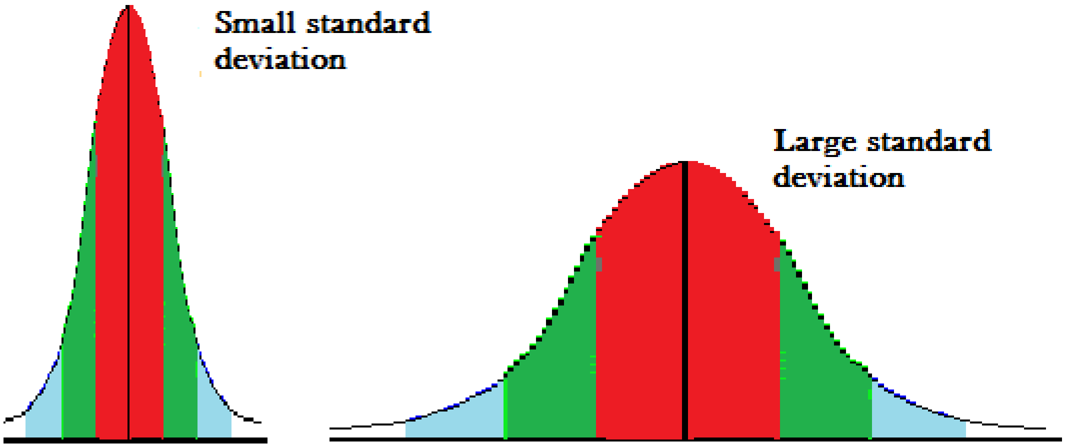
**Standard deviation**: The standard deviation is the positive square root of the variance.

In mathematical notation:

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

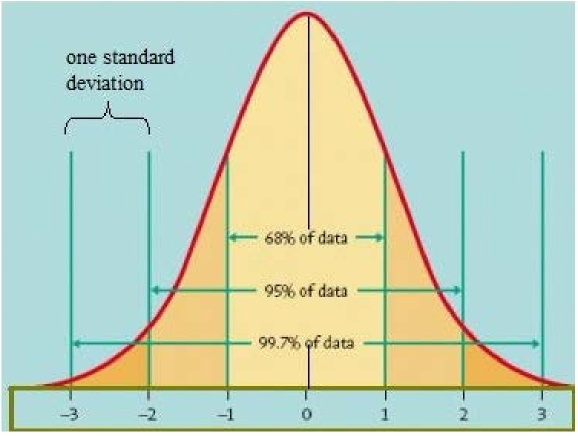
For the exemplar data the value of standard deviation, $s = \sqrt{17.09} = 4$ Schmidt hammer hardness units.

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.



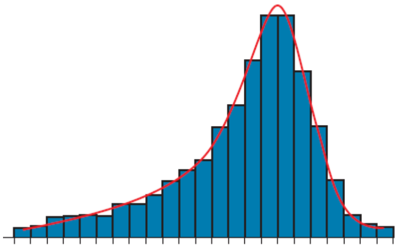
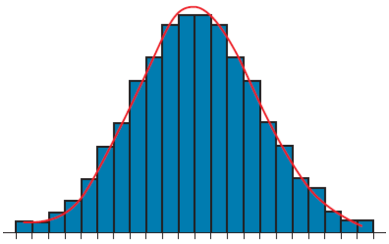
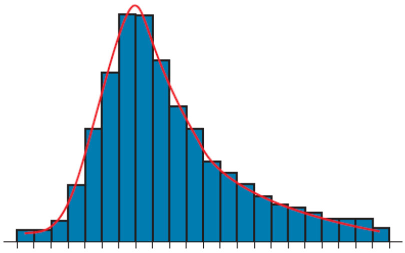
Standard normal model; This is believed to be in the public domain, however if there are omissions or inaccuracies please inform us so that any necessary corrections can be made

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 $= \dfrac{3(36.88 - 36)}{4.13} = +0.64$

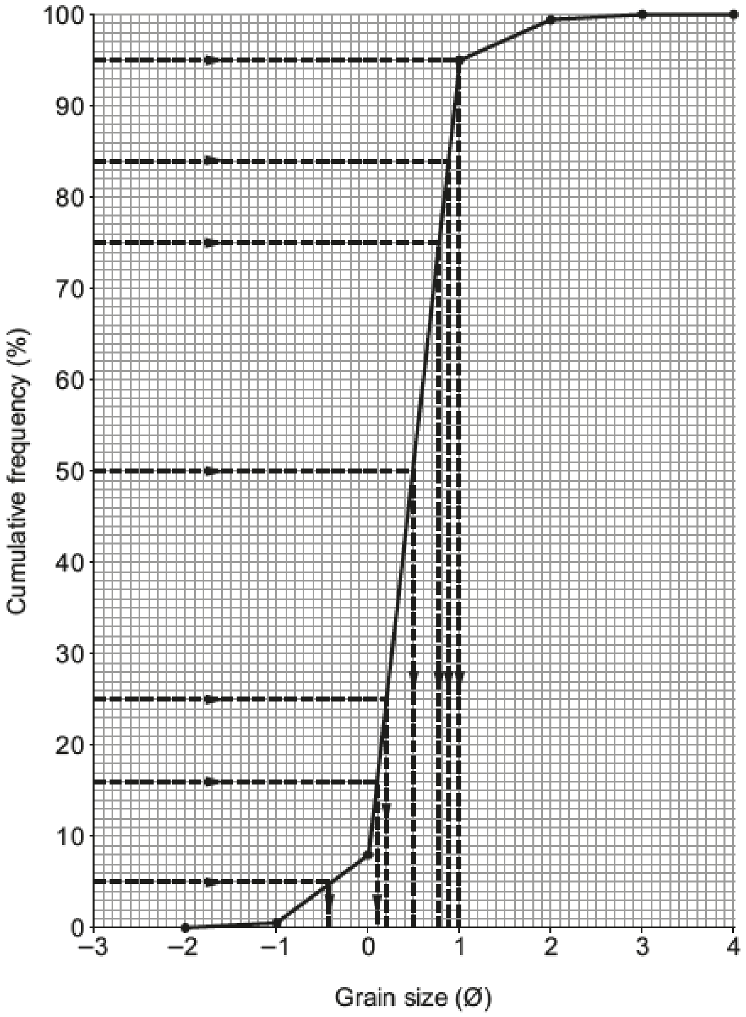
This positive skew can be seen by the obvious tail to the right in the data displayed on both the bar chart and histogram.

## Example: statistical analysis of 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, $2^{-0.50} = 0.71\,\text{mm}$.

**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, $2^{-0.48} = 0.72\,\text{mm}$.

**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,

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

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 |

## Probability

The probability of an event is the measure of the chance that the event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty). The higher the probability of an event, it is more likely that the event will occur. Probability theory is studied in many areas of geology but is crucial in risk analysis in hazard geology.

Estimating the probability of a geological hazard occurring is notoriously difficult. Such estimates are, at best, very approximate because geological hazards are triggered by the interplay of complex processes that themselves are poorly understood. Additionally, crucial to any probability estimation is an accurate knowledge of the past frequency of the hazard which, because of the constraints of absolute dating, again at times has a high level of uncertainty.

Crucial to any estimate of the probability of a geological hazard occurring is the calculation of the return period. The return period is a statistical measurement based on historical data denoting the average recurrence interval over an extended period of time.

When there is a magnitude associated with the data (such as discharge with a flood or seismic moment with an earthquake) the return period ($T$) can be calculated from:

$$T = \frac{(n+1)}{m}$$

Where $n$ is the number of years of the record and $m$ is the rank of the magnitude.

Such an approach is useful where well constrained data is available. This example below catalogues return periods for peak river discharge on a river in the ten year period between 2000 and 2009 ($n$ = 10).

| Year | River discharge (cubic metres per second) | Rank | Return period (years) |
|---|---|---|---|
| 2000 | 2316 | 8 | 1.4 |
| 2001 | 3483 | 2 | 5.5 |
| 2002 | 2147 | 9 | 1.2 |
| 2003 | 3172 | 4.5 | 2.4 |
| 2004 | 2823 | 6 | 1.8 |
| 2005 | 3228 | 3 | 3.7 |
| 2006 | 1625 | 10 | 1.1 |
| 2007 | 2693 | 7 | 1.6 |
| 2008 | 4163 | 1 | 11 |
| 2009 | 3172 | 4.5 | 2.4 |

Where there is a partial or no magnitude record, the return period $(T)$ is the number of years in the record $(N)$ divided by the number of events $(n)$.

$$T = \frac{N}{n}$$

For example, between 1832 and 1984 there have been 38 eruptions on Mauna Loa in Hawaii. The mean return period is therefore $\frac{(1984 - 1832)}{38}$ = 4 years.

Knowing the mean return period and assuming that the eruptions are random events, one way to calculate the probability of future eruptions of Mauna Loa is by using the following equation:

$$P = 1 – e^{(-t/a)}$$

where $P$ = probability of event occurring

$t$ = time period of interest

$a$ = mean return period

The probability of an eruption occurring on Mauna Loa during the next 1, 5 and 10 years, for example, can then be calculated.

| $t$ (years) | probability |
|---|---|
| 1 | $1 – e^{(-t/a)} = 1 – e^{(-1/4)} = 0.22$ |
| 5 | $1 – e^{(-t/a)} = 1 – e^{(-5/4)} = 0.71$ |
| 10 | $1 – e^{(-t/a)} = 1 – e^{(-10/4)} = 0.92$ |

## Circular data

Orientated (vector) data is an important category of geological information. It is possible to distinguish between two types of orientated observations:

i) those that are distributed on a circle where only an azimuth(bearing) is measured. Examples of such two dimensional data include features such as sole structures on bedding planes and axial plane traces on maps.

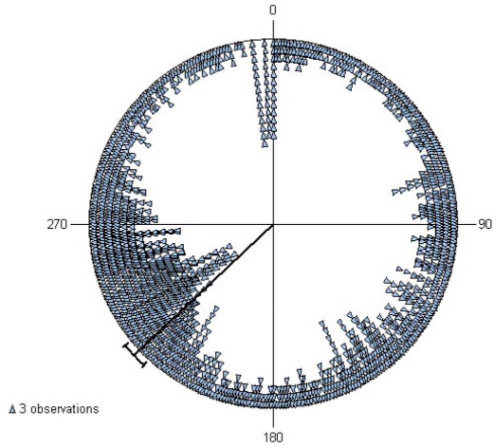
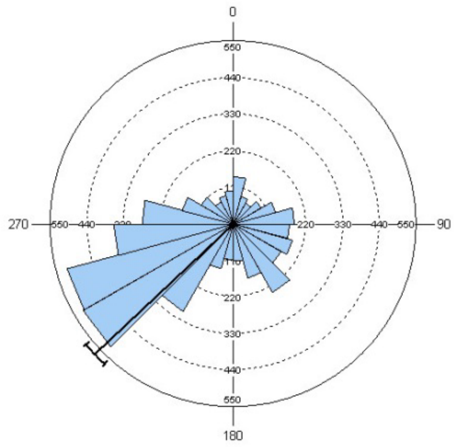
ii) those that are distributed on a sphere where both an azimuth (bearing) and an inclination are measured. Examples of such three dimensional data include various types of planar surfaces (bedding planes, cleavage planes) where both strike and dip are recorded.

In both cases an additional distinction can be made between oriented and directed features. The difference between the two can be understood by imagining a car travelling in a north-bound direction on a north-south orientated motorway – the car has a direction, the motorway an orientation (no direction). Flute casts and bedding planes (which have a way-up) are directed features whereas axial plane traces and cleavage planes are not.

All circular data can be successfully represented on circular plots. In all such graphs the azimuth (a horizontal bearing between 000-360$^{o}$ measured clockwise from North) is plotted around the circumference of the circle. Where only an azimuth is measured the results tend to be plotted on circular bar charts/histograms - rose diagrams. Such graphs avoid the disadvantage of conventional bar charts/histograms in that values that lie close together (e.g. 359$^{o}$ and 001$^{o}$) are plotted close together. Where both an azimuth and an inclination is measured, data tend to be plotted on stereographic projections (stereonets).

### Rose diagrams

The simplest circular graph will show the distribution of raw data i.e. individual bearings as raw data plots. The graph below on the left, for example, shows the direction of dip of cross-bedding fore-sets involving a large data set.

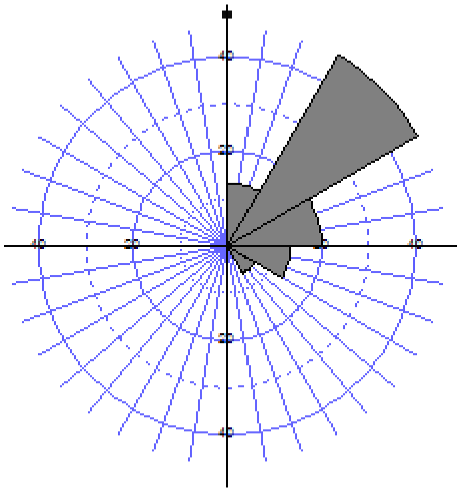
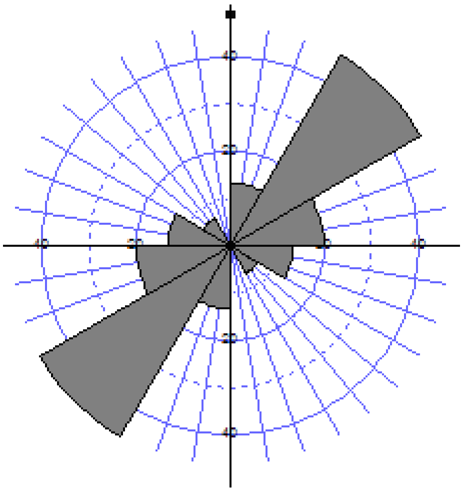
| | |
|---|---|
|  |  |
| Raw data plot for the dip direction of cross-bedding fore-sets. | Equivalent rose diagram for 15$^{0}$ grouped data. |

Rose diagram and raw data plots; Kovach Computing Services https://kovcomp.co.uk/oriana/

However, it is more common for directional data to be grouped (see the same data grouped with an azimuthal class interval of 15$^{o}$ presented to the right of the raw data plot. The bar shows the vector mean of this data).

Consequently if the circle is subdivided into segments, and the number of values within each segment is counted, the results can be drawn as a rose diagram, or circular histogram, with a number of petals whose radius is proportional to the class frequency.

It is also worth noting here how oriented and directed data differ on rose diagrams as it is the convention for oriented data to duplicate each petal on opposite sides of the rose diagram to produce a 'bow tie' shape. Directed features commonly therefore have a unimodal pattern whereas oriented features are often (bipolar) bimodal.

| | |
|---|---|
|  |  |
| Rose diagram for directed data | Rose diagram for oriented data |

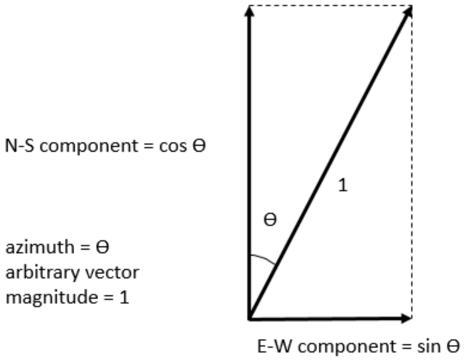
Rose diagram; John W. F. Waldron goo.gl/r2dUVm

Just as with standard univariate statistics it is possible to calculate measures such as the mean, mode and median for circular data (correctly called the vector mean etc.). However, special care needs to be taken when calculating the vector mean as it is not the same as the mean. For example, the vector mean direction of 090° and 180° is 135°, which coincidentally is equal to the mean. But what is the mean direction of 359° and 001°? They are both pointing almost due north (±1°), and clearly the true mean direction is exactly due north. However, the arithmetic mean gives us 180° which is due south and exactly the opposite of the correct answer!

One possible way in which to calculate the vector mean direction is described below for a set of systematically sampled data from Carboniferous (planar) cross bedded fluvial sandstones located at Ramshaw Rocks in the Peak District. Note that the vector mean can only be calculated for such directed features. Orientated data (bipolar bimodal) will always yield a vector mean of zero!

| Raw data (dip direction of cross bedding fore-sets- pointing in the palaeo-current direction) | | | |
|---|---|---|---|
| 355 | 020 | 008 | 346 |
| 038 | 014 | 010 | 022 |
| 358 | 004 | 015 | 350 |

The most widely applicable method to calculate the vector mean is to resolve each individual azimuth into an E-W and N-S component; a technique well known to A level maths and physics students. For example, in the diagram below a unit vector with an azimuth of $\theta$ = 030° can be resolved into a N-S component = cos 30 = 0.866 and a E-W component = sin 30 = 0.500.



If all the individual N-S and E-W components are summed to make a resultant N-S and E-W vector and then these two resultants are combined it is possible to calculate the vector mean of the circular data. This is done by using the following equation where the vector mean is

$$= \arctan \frac{\Sigma \sin \theta}{\Sigma \cos \theta}$$

where $\theta$ is the azimuth reading and $n$ the number of observations.

| azimuth ($\theta$) | sin$\theta$ | cos$\theta$ |
|---|---|---|
| | | |
| 346 | −0.242 | 0.970 |
| 350 | −0.174 | 0.985 |
| 355 | −0.087 | 0.996 |
| 358 | −0.035 | 0.999 |
| 004 | 0.070 | 0.998 |
| 008 | 0.139 | 0.990 |
| 010 | 0.174 | 0.985 |
| 014 | 0.242 | 0.970 |
| 015 | 0.259 | 0.966 |
| 020 | 0.342 | 0.940 |
| 022 | 0.375 | 0.927 |
| 038 | 0.616 | 0.788 |
| | $\Sigma \sin \theta = 1.678$ | $\Sigma \cos \theta = 11.514$ |

For the 12 cross bedding results shown in the table:

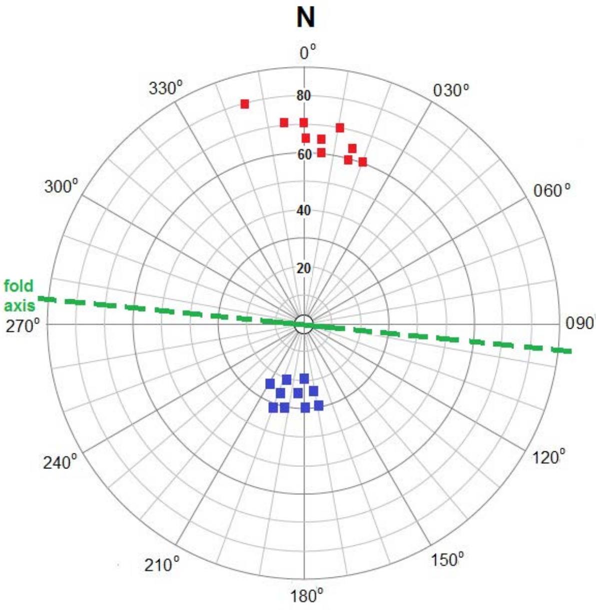
$$\text{vector mean} = \arctan\left(\frac{1.678}{11.514}\right) = 008.3^{\circ}$$

## Polar equal area "stereonets"

An appropriate way to present data that involve both direction (azimuth) and dip plotted on the same diagram, e.g. variation in dip angle and direction of a bedding surface, fractures or aligned fragments in a superficial deposit, is to use a polar equal ‘stereonet’. Similar to a rose diagram, this involves a simplified ‘stereonet’ showing polar plots only and not projections or great circles.

Data is collected on both dip angle and direction of dip in the field or from a secondary source (e.g. geological map). The dip angle is plotted as a point from the centre of the graph in the direction of dip (azimuth) according to the radial scale (zero – 90 degrees). It gives a visual display of trends and amounts but has a drawback for folded strata. If this method is used to plot dip directions on either side of a fold or series of folds, the type of fold (antiform or synform) will not be obvious. Although the dip directions of the limbs are shown, there is no indication as to whether the limbs are dipping towards or away from each other unless further annotation is given. An approximation of the orientation of the fold axis can be determined and plotted by eye or calculated from field measurements.

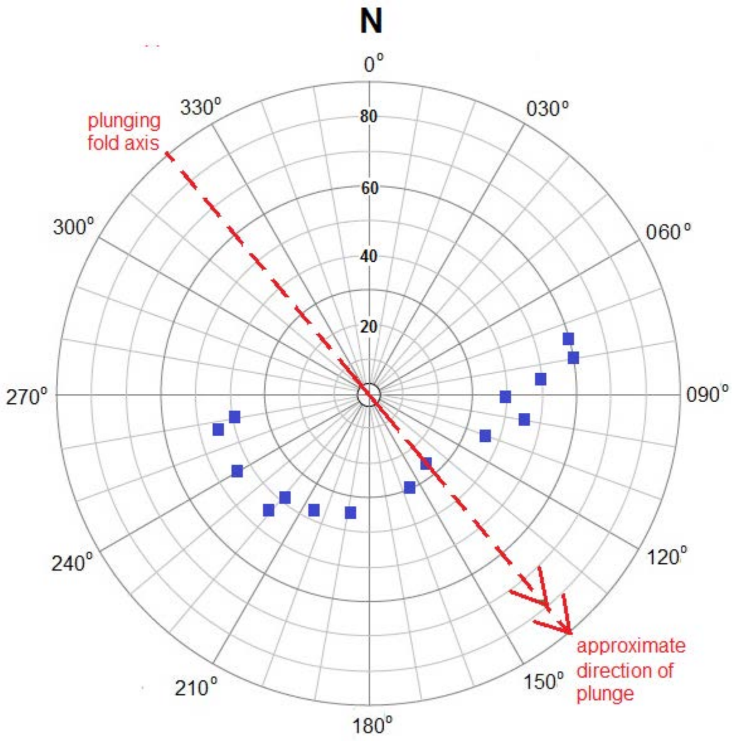
| Azimuth/Dip angle ($^{o}$) | Azimuth/Dip angle ($^{o}$) |
|---|---|
| North dipping limb | South dipping limb |
| 000/71 | 169/30 |
| 001/65 | 171/24 |
| 005/65 | 179/30 |
| 007/60 | 180/19 |
| 011/70 | 185/24 |
| 015/60 | 193/30 |
| 016/64 | 198/20 |
| 020/60 | 199/26 |
| 345/80 | 200/32 |
| 353/72 | 210/25 |



The plot above shows a fold with an axis trending approximately east-west with one limb dipping steeply to the north (between 60$^{o}$ – 80$^{o}$) and another dipping less steeply to the south (~ 20$^{o}$-30$^{o}$). It is either an antiform or synform but this can’t be determined from the diagram.

The plot below of a single bedding surface shows a fold plunging to the SE. One limb is dipping at a maximum angle of 45$^o$ to the W-SW whilst the other limb dips at a maximum of 60$^o$ to the E – NE. Although the fold can be seen to be plunging to the SE from the pattern of the limb dip data, the fold type cannot be determined from this diagram alone.

| Azimuth ($^o$) | Dip angle ($^o$) |
|---|---|
| 073 | 60 |
| 080 | 60 |
| 085 | 50 |
| 090 | 40 |
| 100 | 45 |
| 110 | 36 |
| 140 | 27 |
| 156 | 30 |
| 188 | 35 |
| 206 | 38 |
| 220 | 39 |
| 222 | 45 |
| 240 | 44 |
| 258 | 45 |
| 260 | 39 |



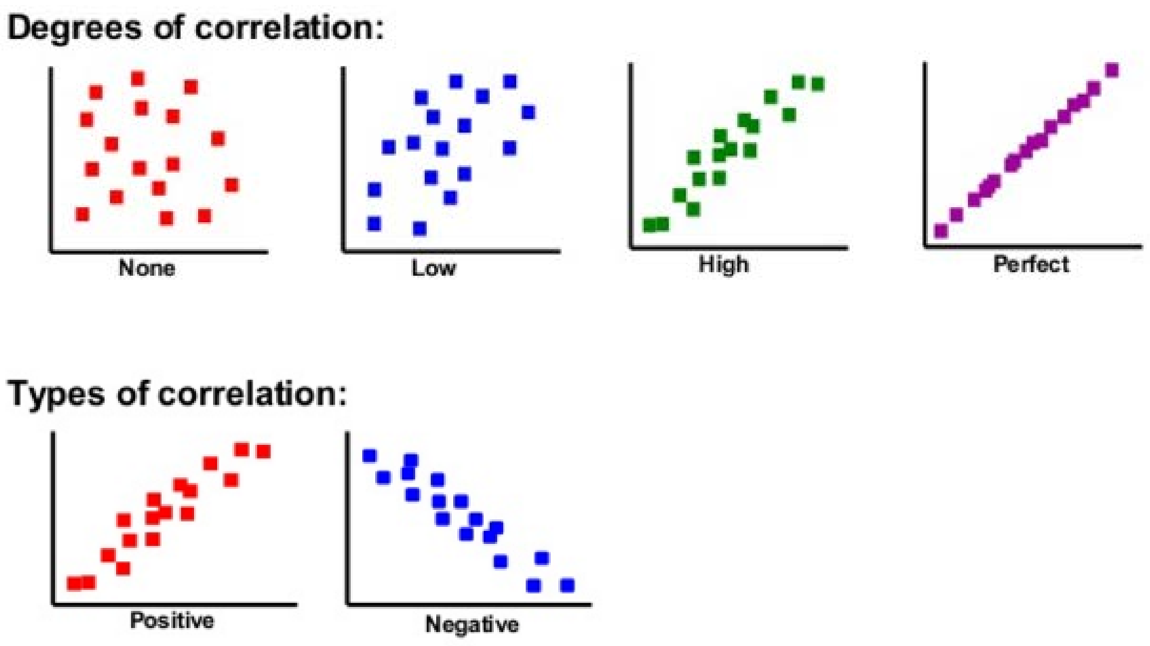
# Bivariate data analysis

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

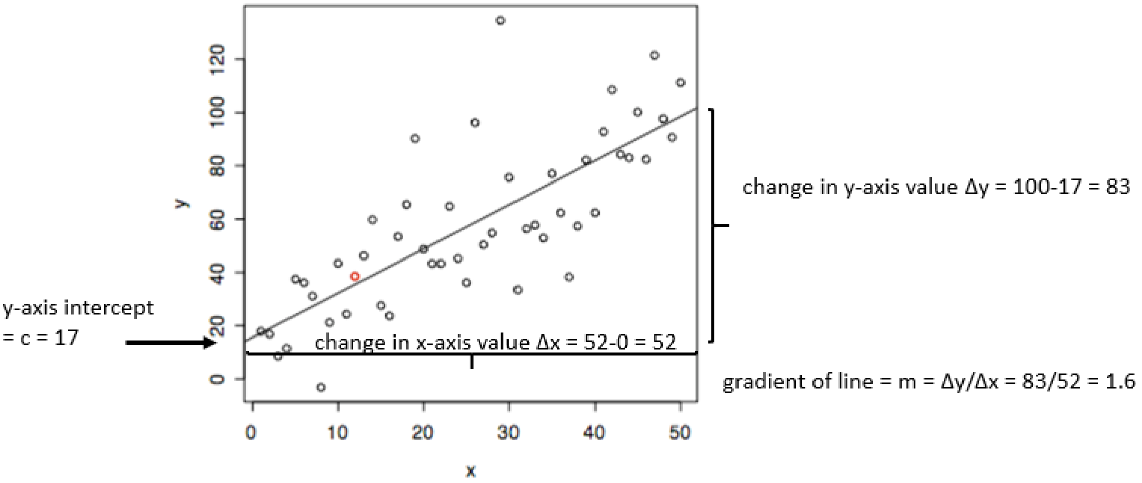
Scatter diagrams are used to show graphically the relationship between two variables. Two axes are drawn in the usual way with the variable that is believed to cause the change in the other (the so-called independent variable) plotted on the $x$-axis; the dependent variable is therefore plotted on the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

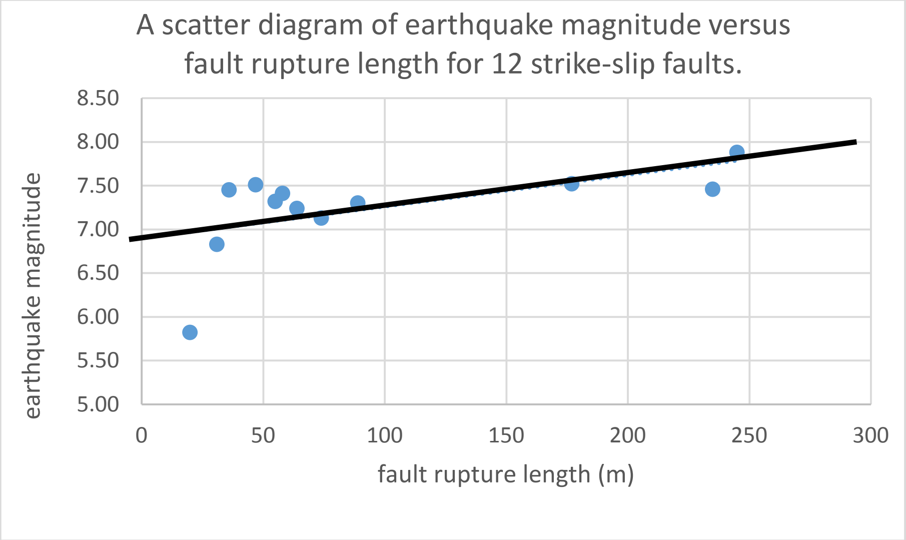
Once the best fit straight line is drawn in by eye it is possible to obtain predictions of unknown values. This may be undertaken directly from the graph or more accurately by obtaining the equation of the straight line. The general equation of a straight line is:

$$y = mx + c$$

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 $y$ = 1.6$x$ +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.


A scatter diagram of earthquake magnitude versus fault rupture length for 12 strike-slip faults.

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. It is often good practice to next test the degree of correlation between the two variables using a test such as Spearman's Rank Correlation.

## Spearman’s Rank Correlation Coefficient ($r_s$)

This is a technique which can be used to test the strength and direction (negative or positive) of a linear relationship between samples of two variables. The result will always be between 1 and minus 1. This is a test that needs at least ten pairs of data to proceed.

To undertake Spearman’s test a scatter graph would generally be constructed first to see whether some correlation exists between the two variables (see above). Next a null hypothesis ($H_o$) would be written. A null hypothesis states that there is no relationship between the two variables being tested i.e. ‘$H_o$: there is no linear relationship between fault rupture length and earthquake magnitude’.

The first step in the process is to calculate the value of Spearman’s Rank Correlation Coefficient ($r_s$). Create a table from the data. Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest value in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking. This should be done for both sets of measurements. Tied scores are given the mean (average) rank.

Find the difference in the ranks ($d$): this is the difference between the ranks of the two values on each row of the table. The rank of the second value (earthquake magnitude) is subtracted from the rank of the first (fault rupture length). Square the differences ($d^2$) to remove negative values and then sum them ($\sum d^2$).

| Fault rupture length (m) | Rank | Earthquake magnitude | Rank | Difference ($d$) | Difference squared ($d^2$) |
|---|---|---|---|---|---|
| 177 | 3 | 7.52 | 2 | 1 | 1 |
| 64 | 6 | 7.24 | 9 | −3 | 9 |
| 245 | 1 | 7.88 | 1 | 0 | 0 |
| 36 | 10 | 7.45 | 5 | 5 | 25 |
| 58 | 7 | 7.41 | 6 | 1 | 1 |
| 31 | 11 | 6.83 | 11 | 0 | 0 |
| 74 | 5 | 7.13 | 10 | –5 | 25 |
| 47 | 9 | 7.51 | 3 | 6 | 36 |
| 89 | 4 | 7.30 | 8 | −4 | 16 |
| 20 | 12 | 5.82 | 12 | 0 | 0 |
| 235 | 2 | 7.46 | 4 | −2 | 4 |
| 55 | 8 | 7.32 | 7 | 1 | 1 |
|  |  |  |  |  | $\sum d^2 = 118$ |

Calculate the coefficient ($r_s$) using the formula below. The answer will always be between 1.0 (a perfect positive correlation) and −1.0 (a perfect negative correlation). When written in mathematical notation, the formula looks like this:

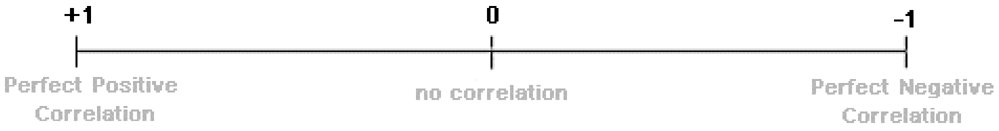
$$r_s = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

where $n$ is the number of pairs of data.

Now insert all the values into the above formula.

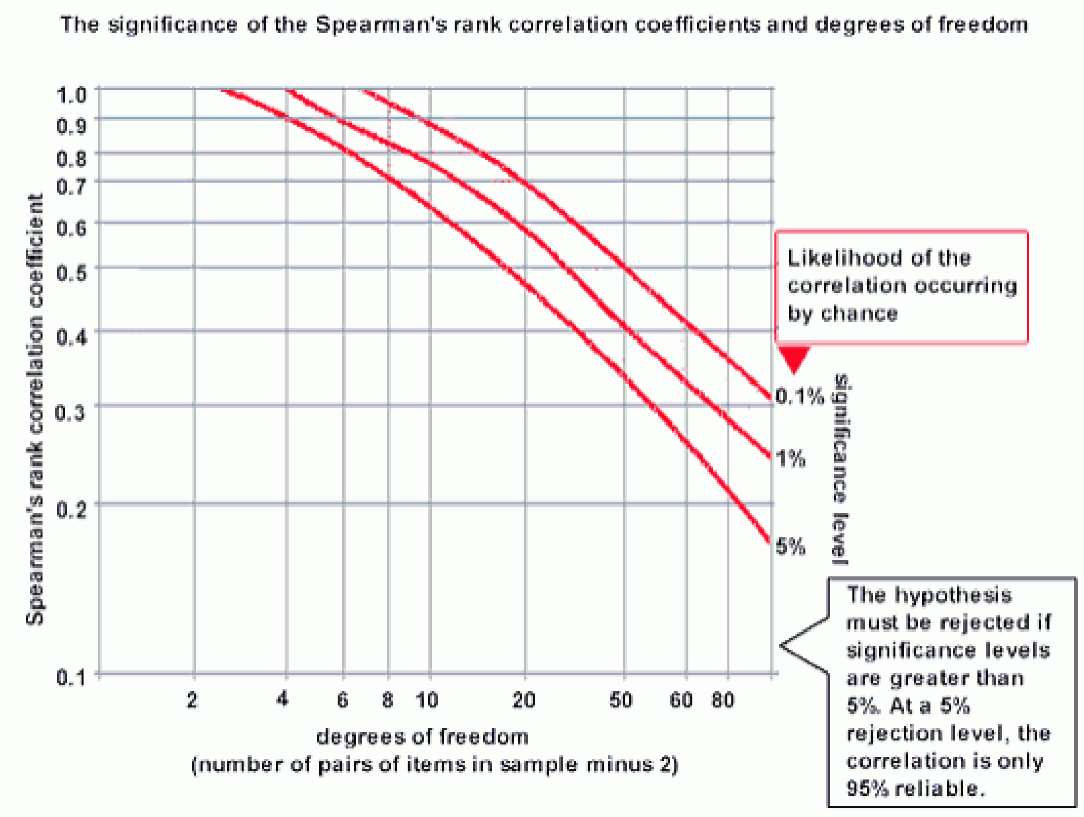
$$r_s = 1 - \left(\frac{708}{1716}\right) = 0.59$$

But what does this value mean? The closer $r_s$ is to +1 or –1, the stronger the likely correlation. The $r_s$ value of 0.59 suggests a moderate positive relationship.



The next step is to test the significance of this correlation because there is a possibility that the correlation is not meaningful and just happened by chance. In other words, if a different sample had been taken, there might have been completely different results. If the correlation is truly meaningful (i.e. it doesn't happen by chance) similar results would be generated no matter which sample was taken. By default, the minimum “level of certainty” or “confidence level” required is 95% (i.e. there should be at least a 95% chance that the correlation is NOT a coincidence). It's the same as saying, in reverse, that the “significance level” must be no more than 100% − 95% = 5% (i.e. there should be no more than 5% chance that the correlation was a coincidence).

To test the significance of this correlation, first of all the ‘degrees of freedom’ of the test is calculated, which is equal to the number of pairs of data in the sample minus 2 i.e. ($n$ – 2) = 12 − 2 = 10. Next the $r_s$ value of 0.59 and ‘degrees of freedom’ value of 10 is plotted on a Spearman's Rank Significance Graph.



Speaman's rank correlation coefficients; Contributions to https://greenfieldgeography.wikispaces.com/ are licensed under a Creative Commons Attribution Non-Commercial 3.0 License

If the result is BELOW the 5% line, the confidence level is too low: this means that you cannot reliably say that the correlation is really meaningful or just happened by chance. If you took a different sample of data, you might obtain different results. If the result is ABOVE the 5% line (or even better above the 1% or 0.1% line!), it means that the correlation is significant with less than a 5% margin of error (i.e. a level of certainty of 95% or more) and hence the null hypothesis ($H_o$) would be rejected.

The results above do not meet the required confidence level of 95%. Consequently it is neither possible to reject the null hypothesis, nor is it possible to conclude that there is no linear correlation in the sample or whether this correlation is just due to chance. As data reliability is related to the size of the sample, the incorporation of new data into the sample may enable a more confident conclusion to be made.

## Other statistical tests

It is possible to undertake a variety of additional statistical tests to investigate the relationships between two or more sets of data and to compare observed measurements with theoretical distributions. One such test, the chi-squared ($X^2$) test, can perform both these tasks for all levels of data i.e. categorical (qualitative) and numerical (quantitative). It is this versatility that makes the chi-squared ($X^2$) test so very useful, especially for categorical data. In contrast, the Mann Whitney U-test, can only be used to compare the relationships between quantitative data but, being a more sophisticated test, it is generally used in preference to the chi-squared ($X^2$) test.

### Chi-squared ($X^2$) test

The chi-squared test is the most efficient test available to analyse numerical (quantitative) data. This test can only be used on data which has the following characteristics:

i) The data must be in the form of frequencies counted in a number of groups (% cannot be used).

ii) The total number of observations must be > 20.

iii) The observations must be independent (i.e. one observation must not influence another).

iv) The expected frequency in any one category must not normally be < 5. It may be necessary therefore to combine groups.

The following data provide information on the lithology of clasts from two adjacent Pleistocene fluvial deposits near Asprokremmos dam in south east Cyprus. Preliminary field investigations (cross-bedding directions) suggest that the two deposits were derived from different sources i.e. their provenance differs. To investigate this a systematic quadrat survey was undertaken to count the number and type of clasts in the breccio-conglomerates to see whether their populations differ. The results of clast lithology for the two deposits are shown in the table below.

| Clast Lithology type | Observed Bed 1 | Observed Bed 2 | Row Total |
|---|---|---|---|
| Ultramafic | 18 | 32 | 50 |
| Gabbro | 12 | 10 | 22 |
| Basalt | 20 | 16 | 36 |
| Chert | 8 | 12 | 20 |
| Chalk | 42 | 30 | 72 |
| Column Total | 100 | 100 | 200 |

The null hypothesis (*Ho*) states that 'there is no significant difference in the composition of clasts sampled in the two deposits'.

The "expected" number of clasts of each lithology in each of the deposits, on the assumption that both deposits have the same proportion of lithologies, is given by the equation

$$\text{Expected value} = \frac{\text{column total} \times \text{row total}}{\text{Overall total}}$$

e.g. the expected value of ultramafic clasts in bed 1 $= \dfrac{100 \times 50}{200}$

$$= \frac{5000}{200}$$

$$= 25$$

An "expected" clast table should be completed (see below).
Note it is not necessary for the number of samples from the two sites to be the same.

| Clast Lithology type | Expected Bed 1 | Expected Bed 2 | Row Total |
|---|---|---|---|
| Ultramafic | 25 | 25 | 50 |
| Gabbro | 11 | 11 | 22 |
| Basalt | 18 | 18 | 36 |
| Chert | 10 | 10 | 20 |
| Chalk | 36 | 36 | 72 |
| Column Total | 100 | 100 | 200 |

Next the value of $X^2$ needs to be calculated using the formula:

$$X^2 = \frac{\Sigma(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$X^2 = \frac{(18-25)^2}{25} + \frac{(12-11)^2}{11} + \frac{(20-18)^2}{18} + \frac{(8-10)^2}{10} + \frac{(42-36)^2}{36} + \frac{(32-25)^2}{25} + \frac{(10-11)^2}{11} +$$

$$\frac{(16-18)^2}{18} + \frac{(12-10)^2}{10} + \frac{(30-36)^2}{36}$$

$$X^2 = 1.96 + 0.09 + 0.22 + 0.40 + 1.00 + 1.96 + 0.09 + 0.22 + 0.40 + 1.00$$

$$X^2 = 7.3$$

It is clear that the greater the difference between the two samples, the greater will be the value of $(O - E)$, the higher the value of $X^2$. A 'high' value of $X^2$ represents a significant difference between the two samples and vice-versa. A more precise definition of 'high' and 'low' $X^2$ values is provided by comparing the computed $X^2$ value with tabulated critical values using the table below.

| Degrees of Freedom | Probability | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.95 | 0.90 | 0.80 | 0.70 | 0.50 | 0.30 | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |
| 1 | 0.004 | 0.02 | 0.06 | 0.15 | 0.46 | 1.07 | 1.64 | 2.71 | 3.84 | 6.64 | 10.83 |
| 2 | 0.10 | 0.21 | 0.45 | 0.71 | 1.39 | 2.41 | 3.22 | 4.60 | 5.99 | 9.21 | 13.82 |
| 3 | 0.35 | 0.58 | 1.01 | 1.42 | 2.37 | 3.66 | 4.64 | 6.25 | 7.82 | 11.34 | 16.27 |
| 4 | 0.71 | 1.06 | 1.65 | 2.20 | 3.36 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 | 18.47 |
| 5 | 1.14 | 1.61 | 2.34 | 3.00 | 4.35 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 | 20.52 |
| 6 | 1.63 | 2.20 | 3.07 | 3.83 | 5.35 | 7.23 | 8.56 | 10.64 | 12.59 | 16.81 | 22.46 |
| 7 | 2.17 | 2.83 | 3.82 | 4.67 | 6.35 | 8.38 | 9.80 | 12.02 | 14.07 | 18.48 | 24.32 |
| 8 | 2.73 | 3.49 | 4.59 | 5.53 | 7.34 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 | 26.12 |
| 9 | 3.32 | 4.17 | 5.38 | 6.39 | 8.34 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 | 27.88 |
| 10 | 3.94 | 4.86 | 6.18 | 7.27 | 9.34 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 | 29.59 |
| | Non-significant | | | | | | | | Significant | | |

To use this chart the number of degrees of freedom must firstly be calculated.

Degrees of Freedom (*df*) = (number of rows –1) × (number of columns –1) = (5 –1) × (2 –1) = 4

In this chart the row headings are that of the degrees of freedom (1-10) and the column headings are level of significance – 0.01 means a 99% confidence level and 0.05 means a 95% confidence level. Again if we assume that a 95% confidence level is required to show that the null hypothesis should be rejected, then the computed value of $X^2$ should be is greater than the critical value obtained from the table. In this example, the computed value of $X^2$ = 7.3 is less than the critical value of 9.49. Consequently there is no reason to reject the null hypothesis '$H_O$: there is no significant difference in the composition of clasts sampled in the two deposits' at the 5% significance level.

**Mann-Whitney *U*-test**

The Mann-Whitney $U$-test is another test which is used to analyse the difference between two samples of independent data: more specifically the medians of the two datasets. The data should be quantitative and there should be >5 but ≤ 20 pieces of data in each sample. Additionally, there need not be the same number of observations in each sample. Like other statistical tests, the starting point is a null hypothesis of the form '$H_o$: there is no significant difference between the two samples'.

The example discussed below was conducted to investigate the working hypothesis that the transport of both chalk and basalt clasts would lead to the chalk clasts becoming more rounded than the basalt clasts as they are softer. Therefore if the research hypothesis is valid, the chalk should have higher values of Cailleux's roundness index (the higher the value the more rounded the clast) than the basalt. Because we are assuming that the chalk clast roundness will be greater than the basalt clast roundness, this is a one-tailed test. If we were assuming that either the basalt or chalk would have a greater roundness, then this would be a two-tailed test. This is important when interpreting the critical values of $U$ table to assess the confidence level of our conclusion. By contrast, the chi-squared ($X^2$) test can merely tell us that a difference does or does not exist between the sets of data, not the direction, hence the critical values of the $X^2$ table relate to what essentially is a two-tailed test.

Measurements of Cailleux's roundness index were systematically sampled along a line transect on the beach at Agio Yeoryios Alamanos near Limassol, southern Cyprus until ten chalk (sample $a$) and ten basalt (sample $b$) clasts had been measured. The results are presented below.

| Cailleux's roundness index | | rank | |
|---|---|---|---|
| chalk ($n_a$) | basalt ($n_b$) | $R_a$ | $R_b$ |
| 780 | 650 | 1 | 9 |
| 640 | 620 | 10 | 11 |
| 690 | 570 | 5 | 16 |
| 710 | 700 | 3 | 4 |
| 550 | 610 | 17 | 12.5 |
| 670 | 490 | 7 | 20 |
| 720 | 520 | 2 | 18 |
| 660 | 600 | 8 | 14 |
| 510 | 590 | 19 | 15 |
| 610 | 680 | 12.5 | 6 |
| | | $\sum R_a = 84.5$ | $\sum R_b = 125.5$ |

The measurements are then ranked (highest to lowest) out of the total set of data i.e. they are the rankings of the values not the rankings within the samples as per Spearman's rank. Where two (or more) equal values occur, the mean rank is used (e.g. the two clasts with a roundness of 610 are the 12th and 13th ranked clasts, so both are given the mean rank of 12.5 and there is no rank of 12 or 13). The sum of the ranks is then calculated for chalk and basalt clasts and this information is put into the formulae below to calculate the $U$ values for sample 1 and sample 2:

$$U_a = n_a n_b + \frac{n_a(n_a+1)}{2} - \Sigma R_a$$

and

$$U_b = n_a n_b + \frac{n_b(n_b+1)}{2} - \Sigma R_b$$

where

$U_a$ and $U_b$ are the Mann Whitney scores for samples $a$ and $b$ respectively

$n_a$ and $n_b$ are the number in samples $a$ and $b$ respectively

and $\sum R_a$ and $\sum R_b$ are the sums of the ranks for samples $a$ and $b$ respectively

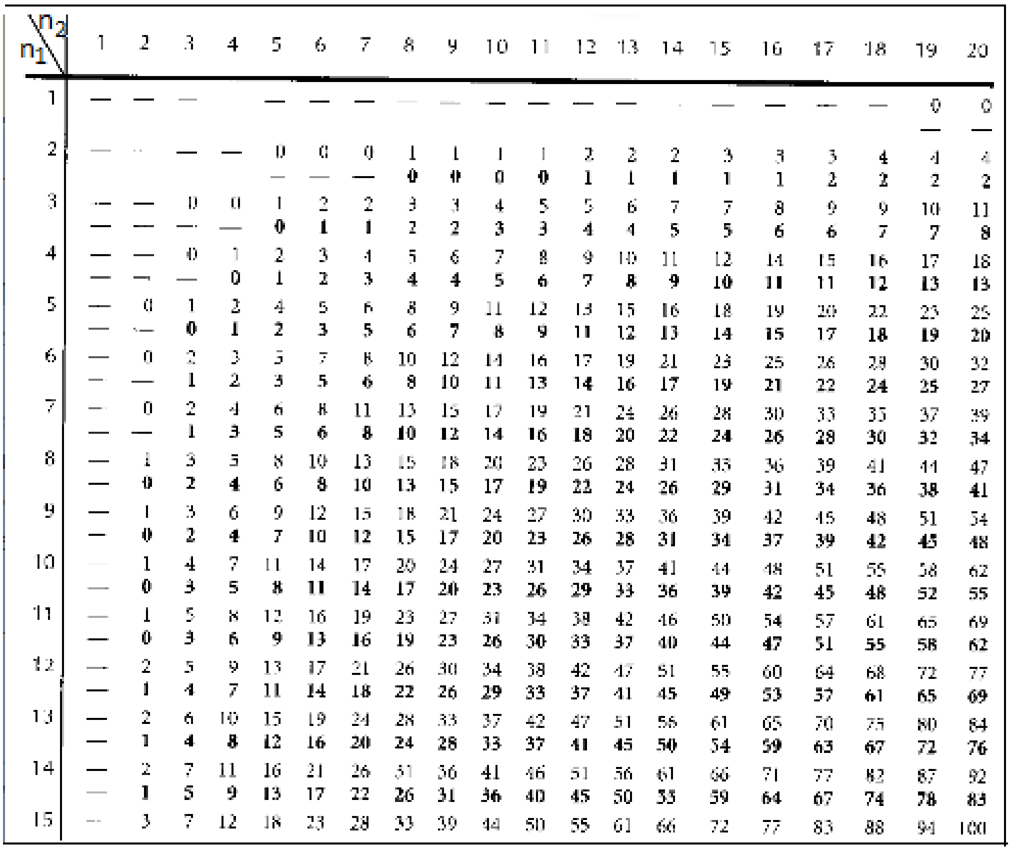
$$U_a = (10 \times 10) + \frac{10(10+1)}{2} - 84.5 = 70.5$$

$$U_b = (10 \times 10) + \frac{10(10+1)}{2} - 125.5 = 29.5$$

A quick check of the accuracy of the calculations is possible as $U_a + U_b = n_a \times n_b = 100$.

The lower value of $U_a$ or $U_b$ is used to assess the significance of any difference between the two sets of samples. The lower value is $U_b = 29.5$. The greater the difference between the two samples, the smaller will be the lower value of $U$. Thus if the computed value of $U$ ($U_b = 29.5$) is less than the critical value, the null hypothesis is rejected at that significance level. Conversely if the computed value of $U$ is greater than the tabulated value, then the null hypothesis is accepted and it must be assumed that the expected difference between the two samples does not exist at that significance level.

## Critical values of the Mann-Whitney U Test



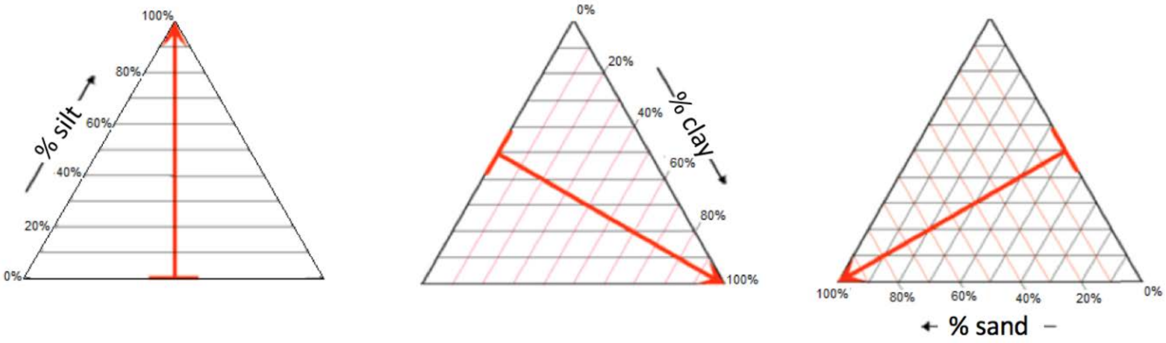
| $n_1$ \ $n_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | — | — | — |  | — | — | — | — | — | — | — | — | — |  | — | — | — | — | 0 | 0 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | — | — |
| 2 | — | — | — | — | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 |
|  |  |  |  |  | — | — | — | **0** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **2** | **2** | **2** | **2** |
| 3 | — | — | 0 | 0 | 1 | 2 | 2 | 3 | 3 | 4 | 5 | 5 | 6 | 7 | 7 | 8 | 9 | 9 | 10 | 11 |
|  | — | — | — | — | **0** | **1** | **1** | **2** | **2** | **3** | **3** | **4** | **4** | **5** | **5** | **6** | **6** | **7** | **7** | **8** |
| 4 | — | — | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 17 | 18 |
|  | — | — | — | **0** | **1** | **2** | **3** | **4** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **11** | **12** | **13** | **13** |
| 5 | — | 0 | 1 | 2 | 4 | 5 | 6 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 18 | 19 | 20 | 22 | 23 | 25 |
|  | — | — | **0** | **1** | **2** | **3** | **5** | **6** | **7** | **8** | **9** | **11** | **12** | **13** | **14** | **15** | **17** | **18** | **19** | **20** |
| 6 | — | 0 | 2 | 3 | 5 | 7 | 8 | 10 | 12 | 14 | 16 | 17 | 19 | 21 | 23 | 25 | 26 | 28 | 30 | 32 |
|  | — | — | **1** | **2** | **3** | **5** | **6** | **8** | **10** | **11** | **13** | **14** | **16** | **17** | **19** | **21** | **22** | **24** | **25** | **27** |
| 7 | — | 0 | 2 | 4 | 6 | 8 | 11 | 13 | 15 | 17 | 19 | 21 | 24 | 26 | 28 | 30 | 33 | 35 | 37 | 39 |
|  | — | — | **1** | **3** | **5** | **6** | **8** | **10** | **12** | **14** | **16** | **18** | **20** | **22** | **24** | **26** | **28** | **30** | **32** | **34** |
| 8 | — | 1 | 3 | 5 | 8 | 10 | 13 | 15 | 18 | 20 | 23 | 26 | 28 | 31 | 33 | 36 | 39 | 41 | 44 | 47 |
|  | — | **0** | **2** | **4** | **6** | **8** | **10** | **13** | **15** | **17** | **19** | **22** | **24** | **26** | **29** | **31** | **34** | **36** | **38** | **41** |
| 9 | — | 1 | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 24 | 27 | 30 | 33 | 36 | 39 | 42 | 45 | 48 | 51 | 54 |
|  | — | **0** | **2** | **4** | **7** | **10** | **12** | **15** | **17** | **20** | **23** | **26** | **28** | **31** | **34** | **37** | **39** | **42** | **45** | **48** |
| 10 | — | 1 | 4 | 7 | 11 | 14 | 17 | 20 | 24 | 27 | 31 | 34 | 37 | 41 | 44 | 48 | 51 | 55 | 58 | 62 |
|  | — | **0** | **3** | **5** | **8** | **11** | **14** | **17** | **20** | **23** | **26** | **29** | **33** | **36** | **39** | **42** | **45** | **48** | **52** | **55** |
| 11 | — | 1 | 5 | 8 | 12 | 16 | 19 | 23 | 27 | 31 | 34 | 38 | 42 | 46 | 50 | 54 | 57 | 61 | 65 | 69 |
|  | — | **0** | **3** | **6** | **9** | **13** | **16** | **19** | **23** | **26** | **30** | **33** | **37** | **40** | **44** | **47** | **51** | **55** | **58** | **62** |
| 12 | — | 2 | 5 | 9 | 13 | 17 | 21 | 26 | 30 | 34 | 38 | 42 | 47 | 51 | 55 | 60 | 64 | 68 | 72 | 77 |
|  | — | **1** | **4** | **7** | **11** | **14** | **18** | **22** | **26** | **29** | **33** | **37** | **41** | **45** | **49** | **53** | **57** | **61** | **65** | **69** |
| 13 | — | 2 | 6 | 10 | 15 | 19 | 24 | 28 | 33 | 37 | 42 | 47 | 51 | 56 | 61 | 65 | 70 | 75 | 80 | 84 |
|  | — | **1** | **4** | **8** | **12** | **16** | **20** | **24** | **28** | **33** | **37** | **41** | **45** | **50** | **54** | **59** | **63** | **67** | **72** | **76** |
| 14 | — | 2 | 7 | 11 | 16 | 21 | 26 | 31 | 36 | 41 | 46 | 51 | 56 | 61 | 66 | 71 | 77 | 82 | 87 | 92 |
|  | — | **1** | **5** | **9** | **13** | **17** | **22** | **26** | **31** | **36** | **40** | **45** | **50** | **55** | **59** | **64** | **67** | **74** | **78** | **83** |
| 15 | — | 3 | 7 | 12 | 18 | 23 | 28 | 33 | 39 | 44 | 50 | 55 | 61 | 66 | 72 | 77 | 83 | 88 | 94 | 100 |

Using a significance level of 0.05 (95% confidence level) and with a sample size of $n_a$ = 10 and $n_b$ = 10, the critical value table for $U$ for a one tailed test (shown in bold on the above table) yields a critical value of 23. The computed value of $U$ = 29.5 which is greater than the critical value. Consequently, there is no reason to reject the null hypothesis and it is accepted that the expected difference in Cailleux's roundness index between the two samples does not exist at the 5% significance level.

# Multivariate data analysis

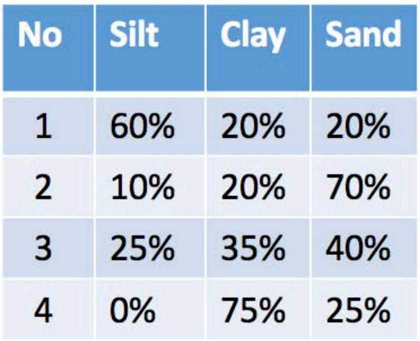
## Triangular plots

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.

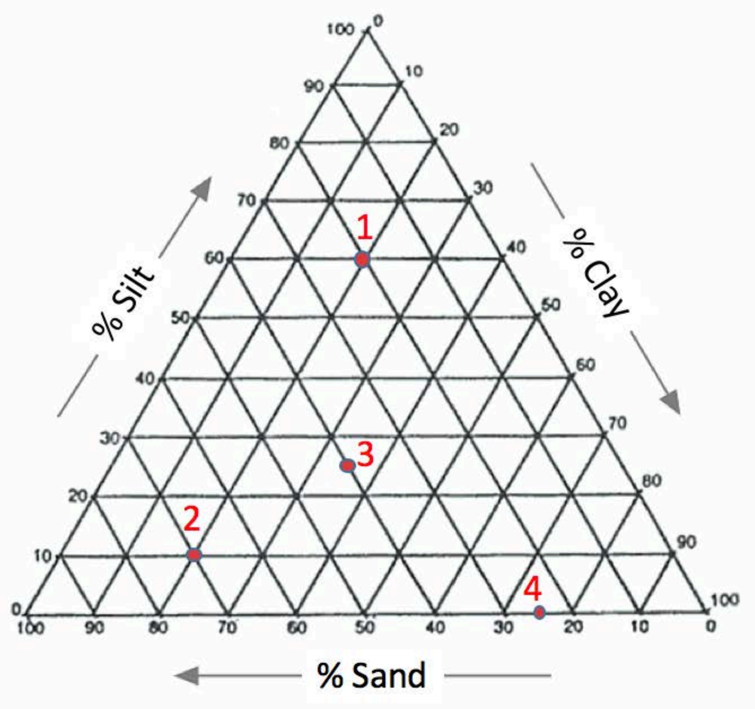


With triangular graphs each axis is divided into 100 – representing percentages.
From each apex lines are drawn at an angle of 60$^o$ to carry the values across the graph.
The data must be in the form of three percentage values and these values must add up to 100.

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% |


% Silt
% Clay
% Sand