G476/576 Hydrology

Geomath Primer: Review of Basic Quantitative Techniques

Contents

Table 1.1 Lower case and upper case letters of the Greek alphabet. Greek letters are frequently used in mathematical expressions even though use of the equivalent roman letters would be equally valid

Greek characters	*Name*
α, A	alpha
β, B	beta
γ, Γ	gamma
δ, Δ	delta
ε, E	epsilon
ζ, Z	zeta
η, H	eta
θ, Θ	theta
ι, I	iota
κ, K	kappa
λ, Λ	lambda
μ, M	mu
ν, N	nu
ξ, Ξ	xi
o, O	omicron
π, Π	pi
ρ, P	rho
σ, Σ	sigma
τ, T	tau
υ, Y	upsilon
ϕ, Φ	phi
χ, X	chi
ψ, Ψ	psi
ω, Ω	omega

Table 1.2 Commonly used symbols and their usual meanings

Symbol	*Usual meaning*
z	Depth
T	Temperature
t	Time
x	Horizontal distance
ρ	Density
ϕ	Porosity and grain size
θ	An angle
P	Pressure
r	Radius
v	Velocity
σ	Stress

Table 1.4 Small numbers of various sizes expressed as a power of 10

Number	*Power of 10*
0.001	10^{-3}
0.0001	10^{-4}
0.000 01	10^{-5}
0.000 001	10^{-6}
1 billionth	10^{-9}

Table 1.5 List of prefixes used in the SI system for denoting very small and very large units. Each prefix represents a unit 1000 times larger than the prefix on the preceding line. Thus a millimetre is 1000 times larger than a micrometre and a femtogram is 1000 times smaller than a picogram

Multiple	*Prefix*	*Symbol*	*Example*
10^{-18}	atto	a	attometre (am)
10^{-15}	femto	f	femtometre (fm)
10^{-12}	pico	p	picometre (pm)
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)
10^{12}	tera	T	terametre (Tm)

Table 1.3 Large numbers of various sizes expressed as a power of 10

Number	*Power of 10*
1 000	10^{3}
10 000	10^{4}
100 000	10^{5}
1 000 000	10^{6}
1 billion	10^{9}

Table 2.1 The ages of buried sediments in a dried-out lake bed if sediments accumulated at a rate of 500 years for each metre of sediment and if the lake dried out 1 My ago

Depth (m)	*Age (My)*
0	1.00
20	1.01
40	1.02
60	1.03
80	1.04
100	1.05

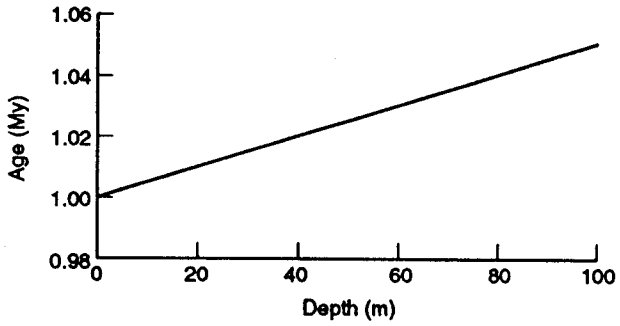


Figure 2.1 Graph of age versus depth data from Table 2.1.

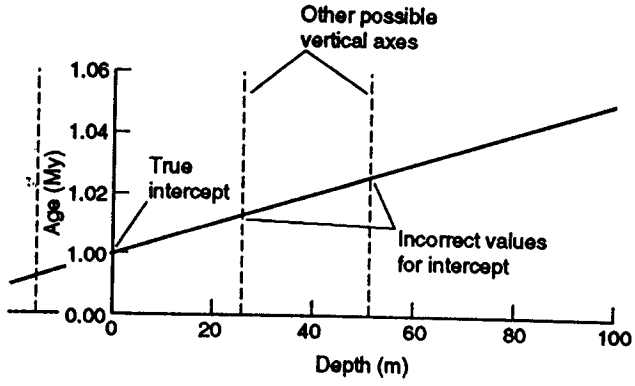


Figure 2.2 The intercept should always be given for a vertical axis which passes through the origin of the horizontal axis. Any other vertical axis will give a different value for the intercept.

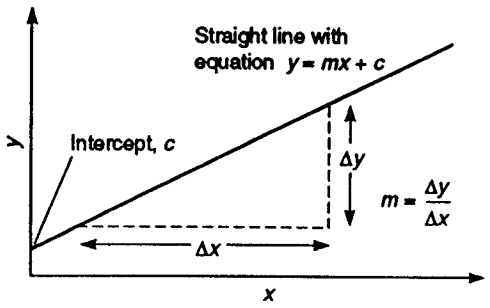


Figure 2.5 The general form of the equation of a straight line, $y = mx + c$.

Table 2.4 The result of raising 3 to the power of the integers between -2 and 2

n	3^n
-2	1/9
-1	1/3
0	1
1	3
2	9

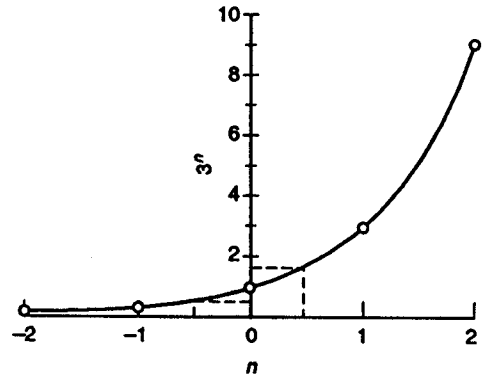


Figure 2.9 A smooth curve drawn through the points in Table 2.4. The dashed lines indicate the point where $n = 0.5$ giving a value for $3^{1/2}$ of approximately 1.7 as well as the point where $n = -0.5$.

Table 2.6 Ten raised to the power of the integers between −2 and 3

n	10^n
−2	0.01
−1	0.1
0	1
1	10
2	100
3	1000

Table 2.7 Rewriting the data in Table 2.6 as a logarithmic table by swapping the columns around

Number	*Logarithm*
0.01	−2
0.1	−1
1	0
10	1
100	2
1000	3

Table 2.8 Number of faults of length greater than or equal to a given size at a particular outcrop location, e.g. there are 11 faults of length 1 m or longer

Fault length (m)	*Number*
0.001	10 109
0.01	957
0.1	132
1	11
10	1

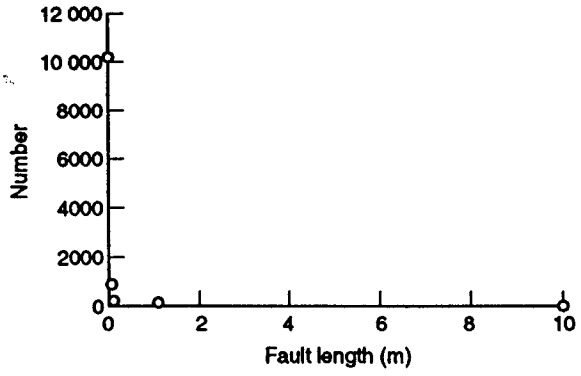


Figure 2.12 The result of plotting a graph of the data in Table 2.8.

Table 2.9 Result of taking logarithms of the columns in Table 2.8

Fault length (m)	*log(length)*	*Number*	*log(number)*
0.001	−3	10 109	4.00
0.01	−2	957	2.98
0.1	−1	132	2.12
1	0	11	1.04
10	1	1	0

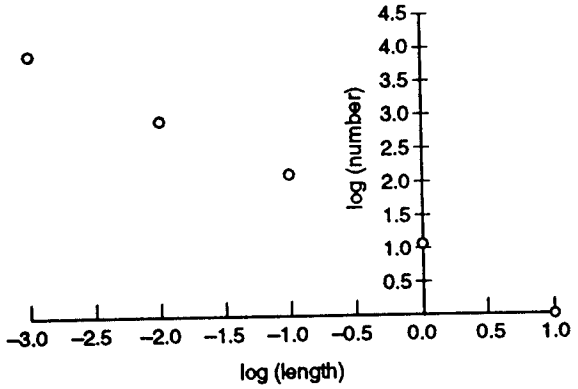


Figure 2.13 Graph of the logarithmic data from Table 2.9.

Table 2.10 The result of raising 6 to the power of various integers and the resulting table of logarithms obtained by swapping the columns around

n	6^n
−2	$1/6^2 = 1/36 = 0.0278$
−1	$1/6^1 = 1/6 = 0.167$
0	$6^0 = 1$
1	$6^1 = 6$
2	$6^2 = 36$

Hence:

x	$\log_6(x)$
0.0278	−2
0.167	−1
1	0
6	1
36	2

3.2 REARRANGING SIMPLE EQUATIONS

It is very obvious, but it is vitally important to appreciate, that an equation is a mathematical statement in which two expressions equal one another. Look again at the lake bed sediment example from Chapter 1:

$$\text{Age} = k \times \text{Depth}. \tag{1.1}$$

The left-hand expression is very simple, it contains 'age'. The right-hand side is also simple and is the product of k and 'Depth'. The point is that the left- and right-hand sides are stated to be equal and this is what makes (1.1) an equation. The reason that I labour this point is that the golden and unbreakable rule when manipulating equations is that, whatever you do, the left- and right-hand sides must remain equal to one another. This is simply achieved. Whenever you manipulate one side of an equation, you must perform exactly the same operation on the other side. Thus, if you add a constant to one side, you must add the same constant to the other side as well; if you double one side, you must double the other; and so on. For example, given equation (1.1), the following expressions are also true:

$$\begin{aligned} \text{Age} + 3 &= (k \times \text{Depth}) + 3 && \text{(i.e. add 3 to both sides)} \\ 2 \times \text{Age} &= 2k \times \text{Depth} && \text{(i.e. double both sides)} \\ \sqrt{\text{Age}} &= \sqrt{(k \times \text{Depth})} && \text{(i.e. square root both sides).} \end{aligned}$$

By combining suitable operations on the two sides of an equation, it is possible to rearrange an equation into another form. As an example, suppose that instead of an equation which tells us the age if we know the depth (i.e. equation (1.1)) we actually need an equation which tells us the depth we would need to dig to reach sediments of a specified age. How do we do this? We must manipulate equation (1.1) to give a new equation which has 'Depth = ' on the left-hand side rather than 'Age = '. The problem is that 'Depth' only appears in combination with 'k'; it does not stand on its own. We must, somehow, remove 'k'. Now, if '$k \times$ Depth' is divided by 'k' then we are left with 'Depth'. However, if we do this to the right-hand side of equation (1.1) we must also do this to the left-hand side. This gives

$$\text{Age}/k = \text{Depth}$$

which can obviously be rewritten as

$$\text{Depth} = \text{Age}/k \tag{3.1}$$

which is the expression that we wanted.

The above example is very simple and could probably have been done almost automatically by many readers of this book. However, it is important that you read very carefully through the logic of the above example.

3.3 COMBINING AND SIMPLIFYING EQUATIONS

If we know the mass of the earth and can also find its volume, the earth's average density could be calculated. The volume of the earth can be estimated using the standard formula for the volume, V, of a sphere of radius r. This is

$$V = 4\pi r^3/3. \tag{3.8}$$

Using the radius of the earth given above and $\pi \approx 3.142$ gives a volume of

$$\begin{aligned} V &= 4 \times 3.142 \times (6.37 \times 10^6)^3/3 \\ &= (4 \times 3.142 \times 6.37^3/3) \times 10^{18} \\ &= 1083 \times 10^{18} \\ &= 1.083 \times 10^{21}\,\mathrm{m}^3. \end{aligned}$$

The density (usually denoted by rho, ρ) is related to mass and volume by

$$\rho = M/V. \tag{3.9}$$

Thus, using the mass and volume already found, the average density of the earth is given by

$$\begin{aligned} \rho &= (5.97 \times 10^{24})/(1.083 \times 10^{21}) \\ &= (5.97/1.08) \times 10^3 \\ &= 5.51 \times 10^3 \\ &= 5510\,\mathrm{kg\,m^{-3}} \end{aligned}$$

which is more than five times the density of water (which is around $1000\,\mathrm{kg\,m^{-3}}$).

The above derivation was a little tortuous. It is possible to combine the three expressions (equations (3.7)–(3.9)) into a single expression for density. This means less numerical calculation is necessary and fewer errors will be made. Equation (3.9) tells us to divide mass by volume to give density. The mass is given by equation (3.7) whilst the volume is given by equation (3.8). Therefore we can immediately write down

$$\begin{aligned} \rho = M/V &= \text{equation (3.7)/equation (3.8)} \\ &= \frac{gr^2/G}{4\pi r^3/3}. \end{aligned} \tag{3.10}$$

In other words, it is always possible to replace an expression (e.g. M) by another equal expression (in this case gr^2/G). Since the initial and replacement expressions are equal, the right-hand side of the equation does not change its value and the equation remains true. We now have an expression for density which could be evaluated by substituting the known values for g, r and G. However, equation (3.10) looks a bit daunting. Evaluating it is not any easier than separately evaluating M and V as before. Fortunately, it is possible to simplify. First, we multiply both the top of the right-hand side and the bottom of the right-hand side by G. Since we are multiplying the whole expression then by G/G (which equals 1.0), this has no effect upon the left-hand side. The result is

$$\begin{aligned} \rho &= \frac{G(gr^2/G)}{G(4\pi r^3/3)} \\ &= \frac{gr^2}{4G\pi r^3/3} \end{aligned} \tag{3.11}$$

since the two Gs on the top cancel each other. Now we can do a similar trick to remove the division by 3. Multiplying top and bottom by 3 yields

$$\rho = \frac{3gr^2}{4G\pi r^3}. \tag{3.12}$$

In many textbooks you will see a slightly different way of doing dimensional analyses. In this approach, the **dimensions** of a unit are expressed in terms of mass, length and time (abbreviated to M, L and T). For example, acceleration has units of $\mathrm{m\,s^{-2}}$ which is a length divided by a time squared. Thus, acceleration has dimensions LT^{-2}. If this procedure is repeated for all items in equation (3.13), the result is

$$\begin{aligned}&\text{Dimensions of } \rho\\ &= \frac{\text{Dimensions of 3} \times \text{Dimensions of } g}{\text{Dimensions of 4} \times \text{Dimensions of } G \times \text{Dimensions of } \pi \times \text{Dimensions of } r}\\ &= \frac{LT^{-2}}{L^3M^{-1}T^{-2}L} = ML^{-3}\end{aligned} \tag{4.43}$$

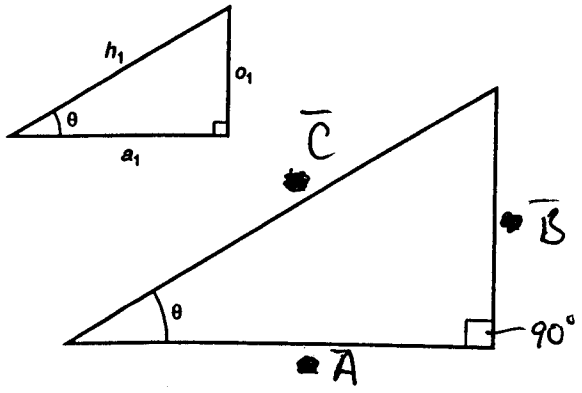
which balances since the dimensions of density are indeed mass divided by length cubed. In practice, it does not much matter which of these two approaches to dimensional analysis you use since they are entirely equivalent. However, I prefer the first in most situations since it is easier when quantities such as temperature are included and it also forces you to check that you have used consistent units for all quantities in the equation.

Trigonometry 5

5.1 INTRODUCTION

Trigonometry is the study of triangles. Triangles rather than, say, squares or hexagons because any other **polygon** (a closed shape with straight edges) can be constructed by adding triangles together (Figure 5.1). Thus, if the properties of triangles are understood, any other polygon can also be dealt with.

Triangles are ideal for purposes such as mapping since there is a simple set of rules relating the lengths of their sides to the size of their angles. Figure 5.2 illustrates the quantities which define a given triangle. This triangle has three sides of length a, b and c and three angles of size α, β and γ. Letters A, B and C are used here to denote the **vertices** (i.e. corners) of the triangle. This gives an alternate way of specifying angles. For example, angle α could be denoted angle BAC, i.e. the angle formed between line BA and line AC.



$\sin\theta = \frac{B}{C}$

$\cos\theta = \frac{A}{C}$

$\tan\theta = \frac{B}{A}$

Statistics

7

7.1 INTRODUCTION

This chapter is a very brief introduction to the subject of geological statistics. Statistics is probably the most intensively used branch of mathematics in the earth sciences. For this reason, even an introduction to the subject fills an entire book and there is a large number of such texts. I do not intend, therefore, to cover this topic in the depth it deserves but to give an introduction which I hope will help ease you into the subject and allow you to go on to other texts with some idea of what to expect.

A major problem with statistics is that it is very easy to mislead. A good example comes from the statisticians' favourite subject, coin-tossing. If a coin is tossed six times it is quite likely that there will be three heads. It is very unlikely that six heads will occur. If I then went on to state that it is more probable that the result will be HTHTHT than HHHHHH (where H represents heads and T tails) I would be seriously misleading you. Both of these events are equally unlikely! The reason that the most likely result is three heads and three tails is that there is a large number of ways of doing this (e.g. HHTHTT, HTHTTH and HHHTTT) whereas there is only one way to get six heads (i.e. HHHHHH). Any particular combination of heads and tails is as likely as any other. Other ways in which statistics can mislead are much more subtle and even experts can, and do, make very serious errors. However, do not let me put you off statistics! If you work carefully and thoughtfully statistics can produce results that could not be obtained in any other way.

Question 7.1 Write down all possible results of tossing a coin three times. Tabulate the results in terms of the number of different ways of obtaining 0 heads, 1 head, 2 heads or 3 heads

7.2 WHAT IS A STATISTIC?

I have been writing up to now as if everybody knows exactly what a statistic is. However, even this term is popularly misused. A statement such as 'in 1970

the oil refining capacity of Belgium was 32.6 million tonnes per year' is a fact, not a statistic. So what is a statistic? Let me start with an example of a situation in which statistics might be useful.

Consider a pebbly beach. How would you go about determining the typical composition, mass and length of the pebbles on this particular beach? If I were to pick up one pebble from this beach I would have a **specimen** from the beach. This would probably tell me the composition of some of the pebbles. However, this specimen might be very untypical. A better way to get information would be to pick up 100 or 1000 pebbles from random locations on the beach. I would then have a **sample** from the beach. This would give me a much better idea of the most common rocks the pebbles were produced from and their typical masses and lengths. Finally, I could examine (in principle) all the pebbles from the beach. This is the **population** of all pebbles from this beach and I could then make definitive statements about the composition of the beach. To recap, a specimen is one object, a sample is a number of objects and a population is all the relevant objects. Note that the world 'sample' is frequently used in geology to denote a specimen (e.g. 'a sample of sandstone' meaning a single piece of sandstone). This is confusing and I recommend that you use the world 'specimen' whenever possible.

Question 7.2 If I have 6 books from a library containing 10 000 books, do these 6 form a specimen from the library, a sample from the library or the library population?

Now, we can return to the idea of a statistic. Is the average mass of a pebble a statistic? This depends on whether this average is determined from a sample of pebbles or from the total population of pebbles. The average of the population is a **parameter** of the beach and is a simple fact (just like the Belgian oil refining capacity). The average of a sample, on the other hand, is a **statistic**, it is an attempt to estimate the average mass of all the pebbles by calculating the average mass of some of them. In other words, a statistic is an estimate of a parameter based upon a sample of the population. As another example, consider voting patterns in an election. The estimates of voting intentions obtained by polling organizations before the election itself are statistics (they are based on questioning a small minority of voters) whereas the final official result is a parameter of the election.

Returning to the beach example, the way in which the masses vary from pebble to pebble is described by many parameters in addition to the average mass. For example, the pebble masses may all be very close to one another or they may be widely different. One parameter which quantifies this is called the **standard deviation**. This will be defined more precisely in the next section. Another parameter is called the **skew** of the population and this tells us whether there are more pebbles which are heavier than the average or more pebbles which are lighter. All of these parameters would normally be estimated from a sample of the population. Each of the resulting estimates of a parameter is a statistic.

Whether or not these statistics are good estimates depends on how well the sampling was performed and also on the size of the sample. Two pebbles picked up from one place on the beach are unlikely to yield a good estimate of the average mass. One hundred pebbles picked up at random from all over the beach will give a much better estimate. Designing an experiment or fieldwork so that the information collected represents a good sample is an important part of a scientific approach to a problem.

7.3 COMMONLY ENCOUNTERED PARAMETERS AND STATISTICS

Table 7.1 shows the masses of 100 pebbles from a beach. From this sample we might wish to get an idea of the typical mass and also the spread of masses. Each of these can be quantified in several different ways.

Table 7.1 The weights of 100 pebbles sampled from a beach

Pebble No.	*Mass (g)*	*Pebble No.*	*Mass (g)*	*Pebble No.*	*Mass (g)*	*Pebble No.*	*Mass (g)*
1	822	26	375	51	483	76	666
2	355	27	134	52	369	77	871
3	909	28	204	53	496	78	714
4	632	29	161	54	414	79	783
5	706	30	160	55	1480	80	498
6	359	31	419	56	1115	81	539
7	881	32	147	57	618	82	597
8	284	33	68	58	1227	83	1834
9	607	34	91	59	751	84	915
10	1263	35	167	60	1349	85	418
11	290	36	459	61	658	86	191
12	795	37	151	62	360	87	898
13	1120	38	135	63	454	88	224
14	1154	39	80	64	325	89	1473
15	439	40	197	65	1645	90	994
16	229	41	233	66	1243	91	397
17	182	42	75	67	393	92	1158
18	383	43	115	68	648	93	294
19	719	44	314	69	1090	94	1752
20	509	45	414	70	476	95	625
21	322	46	83	71	1017	96	685
22	578	47	239	72	1814	97	1251
23	1336	48	146	73	508	98	499
24	686	49	145	74	636	99	479
25	488	50	62	75	541	100	1331

The typical mass can be described by the **mean**, $\bar{w}$,

$$\bar{w} = (\text{Total mass of the sample})/(\text{Number of pebbles}) \quad (7.1)$$
$$= 61\,018/100 \approx 610\,\text{g}$$

for the pebbles in Table 7.1. Frequently, this computation will be described using the following notation:

$$\bar{w} = \frac{1}{N}\left(\sum_{i=1}^{N} w_i\right) \quad (7.2)$$

where w_i is the mass of the ith pebble and N is the number of pebbles in the sample (i.e. w_1 is the first mass in Table 7.1 (822 g), w_5 is the fifth mass (706 g) and so on). The symbol $\sum$ (sigma) is an instruction to add together the w_i (i.e. add the masses of the pebbles together). The $i = 1$ below the $\sum$ and the N above indicate that all items numbered between 1 and N have to be added (i.e. the mass of 100 pebbles in our example). Finally, the result of this addition is divided by N (i.e. 100 in our case). Equation (7.2) may be rewritten in words as 'the average mass may be found by summing the masses of N pebbles and dividing by N'. This notation will be used repeatedly throughout this chapter so make sure that you understand it.

Table 7.2 Pebbles ranked according to decreasing weight

Rank	*Mass (g)*	*Rank*	*Mass (g)*	*Rank*	*Mass (g)*	*Rank*	*Mass (g)*
1	1834	26	822	51	496	76	284
2	1814	27	795	52	488	77	239
3	1752	28	783	53	483	78	233
4	1645	29	751	54	479	79	229
5	1480	30	719	55	476	80	224
6	1473	31	714	56	459	81	204
7	1349	32	706	57	454	82	197
8	1336	33	686	58	439	83	191
9	1331	34	685	59	419	84	182
10	1263	35	666	60	418	85	167
11	1251	36	658	61	414	86	161
12	1243	37	648	62	414	87	160
13	1227	38	636	63	397	88	151
14	1158	39	632	64	393	89	147
15	1154	40	625	65	383	90	146
16	1120	41	618	66	375	91	145
17	1115	42	607	67	369	92	135
18	1090	43	597	68	360	93	134
19	1017	44	578	69	359	94	115
20	994	45	541	70	355	95	91
21	915	46	539	71	325	96	83
22	909	47	509	72	322	97	80
23	898	48	508	73	314	98	75
24	881	49	499	74	294	99	68
25	871	50	498	75	290	100	62

An alternative way of quantifying the typical mass is to use the **median** value. This is obtained by **ranking** the pebbles from the heaviest to the lightest and taking the central value. If we had 5 pebbles, the central one would be the third heaviest pebble (also the third lightest). Similarly, for 99 pebbles the median would be the mass of the fiftieth heaviest pebble. However, for an even number of pebbles there is no central pebble. For example, for 4 pebbles the second and third are equally close to being central. In such cases the procedure is to average the two most central values. For 100 pebbles we must average the mass of the fiftieth and fifty-first pebble. Ranking of the beach sample in Table 7.1 results in Table 7.2. Pebbles 50 and 51 are 498 and 496 g respectively. Thus the median mass is the average of 498 and 496, i.e. 497 g. Note that this is quite different from the average mass.

So much for statistics which indicate the typical mass. What about other aspects of the distribution of pebble masses? For example, the pebbles might all have very similar masses or the masses might be very widely dispersed. What is needed is a measure of **dispersion**. A very simple way to indicate this would be to give the **range** of values, i.e. the lowest and highest masses in the sample. In the case of Table 7.1 (or better, Table 7.2) the range is from 62 to 1834 g. However, the heaviest and lightest pebbles might be very untypical. It would be better to use a measure of the spread of masses which is determined by all of the pebbles in the sample rather than a small minority. One such measure is the **mean square deviation from the mean**. This is also called the **variance**. For the total population of pebbles this is denoted by σ^2 and is defined as

$$\sigma^2 = \overline{(\text{Mass} - \text{Average mass})^2} \quad (7.3)$$

where the bar over the expression indicates that the average value of this quantity should be calculated. In other words, we first find the average pebble mass and then calculate the difference between this and the mass of each individual pebble. The result is then squared which gives the square deviation from the mean. Finally, the average value of this for all the pebbles is found. The deviation of each pebble from the mean is squared since some of the deviations are negative (mass less than average) and some are positive (mass higher than average) leading to an average deviation of zero. Squaring all the deviations from the mean ensures that the average of a series of positive numbers is found which will, of course, also be a positive number. Notice that if the masses are all very similar then they will all be very close to the mean leading to a small value for the variance. If, on the other hand, the masses differ widely from one another then some of these masses will be a long way from the average value and the variance will be much larger. The standard deviation, σ, which is simply the square root of equation (7.3), could also be used to indicate the range of values in the population.

However, it would be better to have a measure of distribution width based upon a sample rather than the entire population. An obvious candidate would be the **sample variance**, s^2, i.e. apply equation (7.3) to a sample rather than the population. In terms of the notation introduced earlier this gives

$$s^2 = \frac{1}{N}\left(\sum_{i=1}^{N} (w_i - \bar{w})^2\right) \tag{7.4}$$

where $\bar{w}$ is the average mass defined by equation (7.2). There is a slightly easier method for calculating s^2 since equation (7.4) can be rearranged to give

$$s^2 = \overline{w^2} - (\bar{w})^2 \tag{7.5}$$

i.e. the mean of the squared masses minus the square of the mean mass (proof of this is given as an exercise at the end of the chapter). Using the figures from Table 7.1, the mean square mass is $560\,117\,\text{g}^2$ whilst the square of the mean is $610^2 = 372\,100\,\text{g}^2$. Thus, the sample variance is $560\,117 - 372\,100 = 188\,017\,\text{g}^2$.

As a measure of the width of the distribution this number has several disadvantages. The first problem is that $\bar{w}$ itself has been estimated from the sample. In fact, the estimate of $\bar{w}$ obtained from equation (7.2) has the property that it is the value which minimizes the sample variance. If another value is used in equation (7.4) in place of $\bar{w}$ a larger value for s^2 is always obtained. Hence, if the true population mean were substituted into equation (7.4) instead of its estimate, $\bar{w}$, a larger value for s^2 would result. Equation (7.4) is therefore **biased** towards a smaller value than the true one. To counteract this effect the **unbiased estimate**, $\hat{s}^2$ is used instead, where

$$\hat{s}^2 = [N/(N-1)]s^2. \tag{7.6}$$

Note that for large sample sizes this increases the variance very slightly whereas for smaller sample sizes this variance estimate is significantly larger than s^2. A formal proof that $\hat{s}^2$ is a better estimate of the population variance than s^2 is beyond the scope of this chapter but you should be able to see that it has the effect of altering s^2 in the right direction (i.e. it increases it by an amount which depends upon the sample size). Table 7.1 has a value for N of 100; the figure calculated above for s^2 then produces an estimate of the population variance of $\hat{s}^2 = 100 \times 188\,017/99 = 189\,916\,\text{g}^2$.

Another problem with using variance to measure distribution width is that the final number (in this case $189\,916\,\text{g}^2$) is not very understandable. What does this result actually mean? Perhaps the simplest way to look at this is to use the variance just for comparison purposes. Take two samples of 100 pebbles from two different beaches. If the first beach has a larger variance for the masses than the second, the pebbles on the second beach tend to have more similar masses to each other than those from the first beach.

The 'interpretability' of variance is further improved by taking its square root. This gives an estimate, $\hat{s}$, of the population standard deviation σ. In the case of our data this yields an answer of $\sqrt{189\,916} = 436\,\text{g}$. For reasons that fall outside the scope of this chapter, this result implies that around 68% of all pebble weights should fall within 436 g of the mean value (610 g). Thus, 68% of all weights should fall between 174 and 1046 g. In fact, out of the 100 measurements in Table 7.1, 66 fall in this range which is of course 66% of the total. Thus the theoretical prediction that 68% of the pebble weights should fall in this range is not at all bad! Standard deviation is therefore a very simple way of describing the range of values in your data. A large standard deviation implies a wide spread of values whilst a small standard deviation implies a small spread.

Standard deviation is probably the most common measure of variability that you will encounter. If, in a particular case, no indication is given as to the measure used, you can reasonably assume that standard deviation is implied. For example, a geochemical analysis might quote the amount of lead present in a mineral as $5.2 \pm 0.5\%$. In this case, the statement implies that the average from a large number of measurements is 5.2% and the variation between measurements is described by a standard deviation of 0.5%. Note that the reason for variation from specimen to specimen in an example like this could be due to measurement errors or to real variability in the mineral composition.

Question 7.3 Using the first 10 values from Table 7.1, calculate:

(i) the sample mean,
(ii) the sample median,
(iii) the sample variance.

Also estimate:

(iv) the population variance and standard deviation.

Compare these results to those obtained above from the sample of 100 pebbles.

There are many other parameters and statistics which could be calculated for given populations and samples. However, the most important are undoubtedly the mean and the standard deviation and these are the ones which you should be most familiar with.

7.4 HISTOGRAMS

It is useful to have a method for displaying, for example, the distribution of pebble masses in Table 7.1 graphically. The simplest such method uses the **frequency histogram** which allows the general properties of the distribution to be visualized. To construct such a plot we must first count the number of occurrences of pebble masses within specified ranges. For example, there are

Table 7.3 Number of pebbles in Table 7.1 which fall into 200 g-wide classes

Range (g)	*Number of occurrences*
1–200	19
201–400	19
401–600	20
601–800	16
801–1000	7
1001–1200	6
1201–1400	7
1401–1600	2
1601–1800	2
1801–2000	2

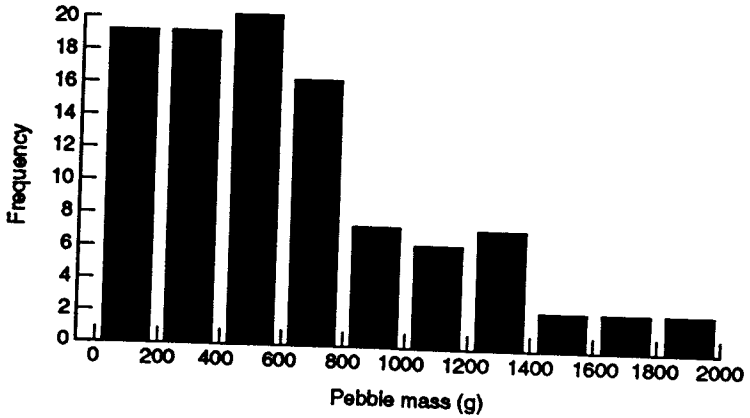


Figure 7.1 The frequency histogram resulting from counting the number of pebbles from Table 7.1 in each 200 g-wide class.

eight pebbles in Table 7.1 with masses between 201 and 300 g. Table 7.3 lists the number of pebbles with masses in 200 g-wide **classes** from 1 to 2000 g. Such a table is called a **frequency distribution**. If these figures are plotted as a bar chart the result is a frequency histogram (Figure 7.1). From this we can instantly see that the masses most commonly fall between 400 and 600 g and are heavily skewed towards the low end.

It was pointed out in Chapter 5 that, if your data are a function of a cyclic variable such as direction or longitude, they are best represented by a polar plot. The same is true for histograms. For example, cross-bedding and ripple marks in sandstones can be used to indicate the **palaeocurrent direction**, i.e. the direction of transport of ancient rivers or submarine currents. Such data will usually have considerable scatter due to uncertainties in measurement and the effect of local topographic features. Thus, if the general direction of

Table 7.4 Number of palaeocurrent measurements as a function of direction

Direction range (°E of N)	*Number of measurements*
1–30	43
31–60	23
61–90	10
91–120	11
121–150	14
151–180	20
181–210	10
211–240	4
241–270	15
271–300	20
301–330	40
331–360	36

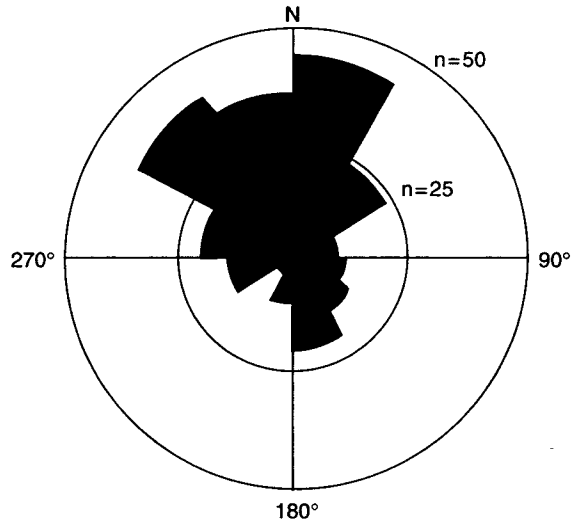


Figure 7.2 A 'rose diagram' in which frequency as a function of direction is represented by distance from the centre.

transport is required, the best course of action is to collect a large number of measurements and then to plot these on a histogram. Table 7.4 lists the frequency data from such a series of measurements.

Now, an obvious way to plot these data is as a histogram on polar graph paper. In other words, plot the frequency as a function of direction such that direction is represented by angle around the plot and the frequency is proportional to distance from the plot centre. This yields a **rose diagram** (Figure 7.2) from which it is very clear that the main current direction was roughly NNW.

Table 7.5 Estimates of the probability of a pebble picked at random being in a specific range. Data modified from Table 7.3

Range (g)	*Probability*
1–200	0.19
201–400	0.19
401–600	0.20
601–800	0.16
801–1000	0.07
1001–1200	0.06
1201–1400	0.07
1401–1600	0.02
1601–1800	0.02
1801–2000	0.02

7.5 PROBABILITY

Probability is a central concept in statistics. In essence, the idea is very simple. If I perform a very large number of measurements on field data or experimental data then I can determine how often a particular result is obtained. This will then allow me to predict the probability that a particular result will occur in any future measurement. Thus, if I toss a dice 1000 times and the number two occurs 400 times, I can predict that the probability is 0.4 of two being the result of my next throw of the dice. I can also conclude that the dice is probably loaded! Note that an event which has a probability of 1 is certain to occur whilst an event whose probability is zero will never occur.

Similarly, for the data shown in Table 7.3 and Figure 7.1, the most probable weight range (401–600 g) occurs in 20 cases out of 100, i.e. 20% of the time. In other words, an estimate of the probability of a pebble, picked up at random from the beach, being in the range 401–600 g is 0.2. Repeating this procedure for the entire table leads to Table 7.5 which is a **probability distribution**. The results can then be plotted in a new histogram (Figure 7.3). Note that the shape of this is identical to Figure 7.1 except that the vertical scale has been shrunk by a factor of 100 (i.e. divided by the size of the sample).

This probability distribution can be compared to various theoretical distributions. The most important of these is the **normal distribution** otherwise known as the **Gaussian distribution**. This has the form

$$P(x) = \frac{\exp[-(x-\bar{x})^2/2\sigma^2]}{(2\pi)^{1/2}\sigma} \qquad (7.7)$$

where $P(x)$ is called the **relative probability** of obtaining the value x, x is the average of all x-values and σ is the standard deviation of the distribution. A graph of this function is shown in Figure 7.4 for the case of mean equal to 5.0

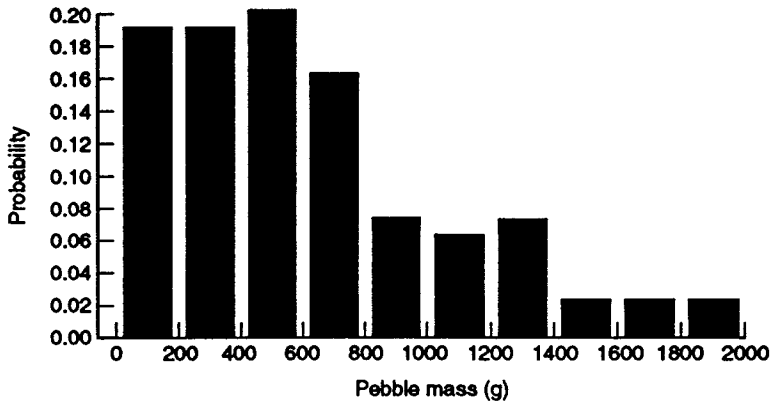


Figure 7.3 Probability distribution histogram for the pebble masses from Table 7.1. Note that the form of this is identical to Figure 7.1 except that the vertical scale has been shrunk by a factor of 100 (equal to the total number of pebbles in the sample).

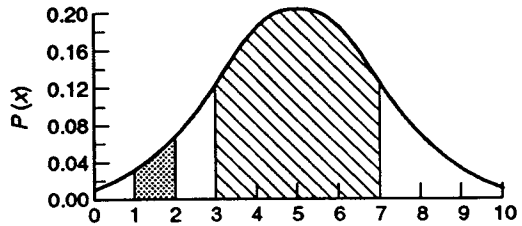


Figure 7.4 Gaussian probability distribution for a mean of 5.0 and standard deviation of 2.0. The diagonally shaded region is the area under the graph within one standard deviation of the mean. The area of the stippled region gives the probability of a result lying between 1.0 and 2.0. The total area under the graph is exactly 1.0.

and standard deviation equal to 2.0. Note that the maximum probability occurs at the mean value, that the curve is symmetrical about the mean and that a large fraction of the area under the graph occurs between $\bar{x}-\sigma$ and $\bar{x}+\sigma$ (i.e. the diagonally shaded area).

The probability of obtaining values within a specified range is governed by two things. Firstly, the higher the graph is within that range, the higher the probability is. Thus, given the graph shown in Figure 7.4, you are more likely to obtain a value between, say, 5 and 6 (where the graph is high) than you are between 1 and 2. Secondly, the probability of obtaining a value within a specified range increases as the width of the range increases. Thus, you are more likely to obtain a value between 5 and 7 than between 5 and 6 because more possibilities are included in the second case. In fact, the relative probability distribution is defined such that the probability of obtaining a

The area under the Gaussian curve as a function of multiples of standard deviation (s.d.) from the mean. For example, the area lying within two standard deviations of the mean is 0.954

No. of s.d.s	*Area*	*No. of s.d.s*	*Area*	*No. of s.d.s*	*Area*
0.00	0.000				
0.10	0.080	1.10	0.729	2.10	0.964
0.20	0.159	1.20	0.770	2.20	0.972
0.30	0.236	1.30	0.806	2.30	0.979
0.40	0.311	1.40	0.838	2.40	0.984
0.50	0.383	1.50	0.866	2.50	0.988
0.60	0.451	1.60	0.890	2.60	0.991
0.70	0.516	1.70	0.911	2.70	0.993
0.80	0.576	1.80	0.928	2.80	0.995
0.90	0.632	1.90	0.943	2.90	0.996
1.00	0.683	2.00	0.954	3.00	0.997

value within a given range is given by the area under the graph over that range. For example, the probability of obtaining a value between 1.0 and 2.0 is given by the area of the stippled region in Figure 7.4 whilst the (much greater) probability of obtaining a value between 3 and 7 is given by the area of the diagonally shaded region. Given this definition, the total area under the graph is 1.0 since the probability of obtaining some value is 1.0. This is indeed the case for the Gaussian distribution defined in equation (7.7).

So, how can such areas be found? The simplest method is to use a table showing area under the curve as a function of multiples of standard deviation from the mean (e.g. Table 7.6). From such a table it can be seen that the area of the diagonally shaded region in Figure 7.4 is 0.683. Similarly, the area under the curve within two standard deviations (i.e. between 1.0 and 9.0 in the example shown in Figure 7.4) is 0.954.

The table can also be used to find areas such as the stippled zone in Figure 7.4. To do this, you should note that 1.0 is two standard deviations from the mean whilst 2.0 is 1.5 standard deviations from the mean. From Table 7.6 it can be seen that the area within 1.5σ is 0.866 whilst the area within 2σ is 0.954. Thus, the area between 1.5σ and 2.0σ is $0.954 - 0.866 = 0.088$. However, there are two such zones, one between 1.0 and 2.0 and the other between 8.0 and 9.0 (Figure 7.5). The area between 1.5σ and 2.0σ is shared equally between these two regions and thus the area of the zone between 1.0 and 2.0 is half of 0.088 (i.e. 0.044). Thus, if a process has a probability distribution like that shown in Figure 7.4, the probability of obtaining a result between 1.0 and 2.0 is 0.044.

Thus, if a population variable has a Gaussian probability distribution, the probability of obtaining results within specified ranges can be calculated. However, the Gaussian distribution function is an idealized model. Real populations are often approximately Gaussian but never exactly so. How good is the Gaussian model at predicting the probabilities shown in Table 7.5

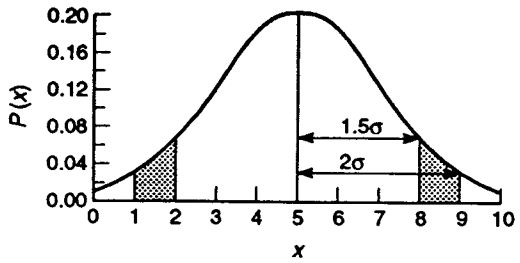


Figure 7.5 The area under the Gaussian curve between 1.5 standard deviations from the mean and 2.0 standard deviations from the mean. Note that there are two such zones and the area of each is therefore half that given using Table 7.6.

Table 7.7 Comparison of the probabilities estimated in Table 7.5 with the probabilities estimated assuming a Gaussian distribution

Range (g)	*Measured probability*	*Range (multiples of σ)*	*Gaussian probability*
1–200	0.19	−1.4 to −0.94	0.090
201–400	0.19	−0.94 to −0.48	0.14
401–600	0.20	−0.48 to −0.02	0.18
601–800	0.16	−0.02 to 0.44	0.18
801–1000	0.07	0.44 to 0.90	0.15
1001–1200	0.06	0.90 to 1.36	0.098
1201–1400	0.07	1.36 to 1.81	0.053
1401–1600	0.02	1.81 to 2.27	0.026
1601–1800	0.02	2.27 to 2.73	0.0059
1801–2000	0.02	2.73 to 3.19	0.0027

for the pebble weights on our beach? Now, the mean and standard deviation for the pebble weights have been estimated to be 610 and 436 g respectively. Using these values, Table 7.6 can be used to predict the probabilities in various ranges of weights. The results are shown in Table 7.7 together with the measured probabilities shown in Table 7.5 (in fact I have used a slightly more detailed table than 7.6).

Question 7.4 The range from 801 to 1000 g starts 0.44 standard deviations above the mean and ends 0.90 standard deviations above the mean. Using Table 7.6 and these values, estimate the Gaussian probability of a pebble weight lying in this range. N.B. To get the area under the curve within 0.44 standard deviations (call it $P_{0.44}$) assume that it is given by the expression

$$P_{0.44} \approx 0.6P_{0.4} + 0.4P_{0.5}$$

i.e. an average of the probabilities corresponding to 0.4σ and 0.5σ weighted towards the 0.4σ probability.

These results are perhaps best described as patchy. The probabilities measured and predicted are quite close in some cases (e.g. 601–800 g and 1401–1600 g) but in other ranges are not at all close (e.g. 1–200 g and 1801–2000 g). The mismatches have two possible causes. Firstly, the distribution may not be Gaussian at all. Secondly, sampling effects may cause the measured probabilities to be very inaccurate, i.e. other samples of 100 pebbles might give quite different results. Deciding whether the mismatch is a statistical fluctuation or a real difference is briefly covered in section 7.7.

7.6 'BEST FIT' STRAIGHT LINES

So much for statistical analysis of a single variable. What about analysis of two related variables? As discussed in Chapter 2, 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. Thus, a straight line must be found which passes as close to all the data points as possible. The problem of how to find this line is ideally suited to a statistical treatment.

First, we have to define what we mean by a 'best fit' straight line. The usual definition is that the mean square difference between the data and the straight line should be a minimum. Figure 7.6 illustrates this idea. The graph of y as a function of x consists of eight points and a straight line has been drawn which passes close to all of these. The deviation, Δy, of the last point from the straight line is also shown. The mean square deviation is found by calculating the square of this distance for all of the points and then finding the average. Now, if the line is a poor fit to the data, Δy for many of the points will be large and the average squared value will also be large. A good fit for the straight line will result in a much smaller average. The best fit straight line is defined

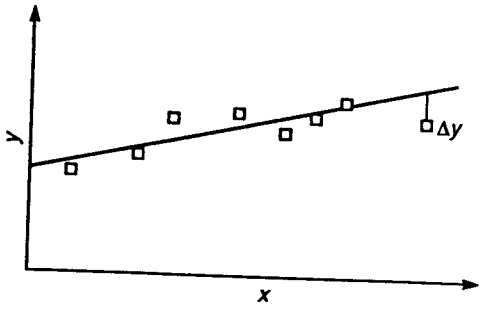


Figure 7.6 A straight line drawn through the x–y data such that it passes close to all of the points. The deviation, Δy, of the line from the last point is also shown. The deviation could be found for each of the other seven points as well. The best fit straight line is that which produces the smallest possible $(\Delta y)^2$ averaged over all of the points.

as that line which results in the smallest possible **mean square deviation**. The process of finding such a line is called **linear regression**.

We now need a method for estimating the gradient, m, and intercept, c, of this straight line. A formal proof is, again, beyond the scope of this chapter but the result is that the best gradient is given by

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}. \tag{7.8}$$

In this expression all summations are evaluated for n measurement pairs (x, y). The first summation, $\sum xy$, is simply the sum of all products x_1y_1, x_2y_2 up to x_ny_n. Similarly, $\sum x$ is the sum of all the x-values, $\sum y$ is the sum of all the y-values and $\sum x^2$ is the sum of all the x-values squared. The best estimate for the intercept then follows directly from the fact that the best fit line passes through the point $(\bar{x}, \bar{y})$, i.e. the point defined by the average x-value and the average y-value. Thus

$$\bar{y} = m\bar{x} + c$$

giving

$$c = \bar{y} - m\bar{x}. \tag{7.9}$$

Table 7.8 shows the age versus depth data used in question 2.11 but with the first point excluded. Now, given that the remaining data when plotted seemed to fit a reasonably good straight line, what is the best fit line through these points? First, we should construct all the summations given in equation (7.8). In this example, we need to plot age as a function of depth and therefore 'Depth' replaces x and 'Age' replaces y. The summations required by equation (7.8) are then

$$\sum \text{depth} = 407 + 545 + 825 + 1158 + 1454 + 2060 + 2263 = 8712\,\text{cm} \tag{7.10}$$

$$\sum \text{age} = 10\,510 + 11\,160 + 11\,730 + 12\,410 + 12\,585 + 13\,445 + 14\,685$$
$$= 86\,525\,\text{years} \tag{7.11}$$

Table 7.8 The age versus depth data used at the end of Chapter 2. This time the first data point has been excluded since it did not fall close to a straight line defined by these remaining points

Depth (cm)	*Age (years)*
407.0	10 510
545.0	11 160
825.0	11 730
1158.0	12 410
1454.0	12 585
2060.0	13 445
2263.0	14 685

$$\sum \text{depth} \times \text{age} = (407 \times 10\,510) + (545 \times 11\,160) + (825 \times 11\,730) + (1158 \times 12\,410) + (1454 \times 12\,585) + (2060 \times 13\,445) + (2263 \times 14\,685)$$
$$= 4\,277\,570 + 6\,082\,200 + 9\,677\,250 + 14\,370\,780 + 18\,298\,590 + 27\,696\,700 + 33\,232\,155$$
$$= 113\,635\,245 \text{ cm y} \tag{7.12}$$

$$\sum(\text{depth}^2) = 407^2 + 545^2 + 825^2 + 1158^2 + 1454^2 + 2060^2 + 2263^2$$
$$= 13\,963\,148 \text{ cm}^2. \tag{7.13}$$

From equations (7.10) and (7.11), together with the fact that there are seven measurements, the mean depth and mean age are

$$\overline{\text{Depth}} = 8712/7 = 1245 \text{ cm} \tag{7.14}$$

and

$$\overline{\text{Age}} = 86\,525/7 = 12\,361 \text{ years}. \tag{7.15}$$

Thus, substituting results (7.10)–(7.13) into equation (7.8) gives

$$m = \frac{7 \times 113\,635\,245 - 8712 \times 86\,525}{7 \times 13\,963\,148 - 8712 \times 8712}$$
$$= 1.91 \text{ y cm}^{-1}. \tag{7.16}$$

Substituting this into equation (7.9) gives the best estimate of the intercept as

$$c = \overline{\text{Age}} - (m \times \overline{\text{Depth}})$$
$$= 12\,361 - (1.91 \times 1245)$$
$$= 9983 \text{ years}. \tag{7.17}$$

The data from Table 7.8 are plotted in Figure 7.7 together with a straight line of gradient 1.91 y cm^{-1} and intercept 9983 years. As you can see, the fit is remarkably good. Note that, before starting this exercise, I deliberately excluded a point that did not fit the general linear trend. This means that the resultant line is not appropriate at ages close to this point (i.e. for sediments less than about 10 000 years old). However, it was important to exclude this point since fitting a straight line through points which do not have a simple linear trend would be a meaningless exercise.

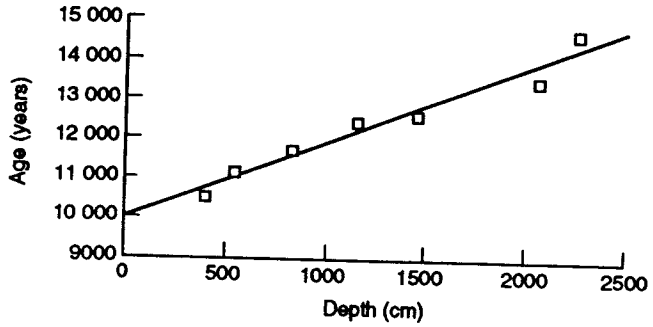


Figure 7.7 The data from Table 7.3 plotted for comparison with a best fit straight line calculated to have a gradient of 1.91 y cm^{-1} and an intercept of 9983 years.

Question 7.5 Calculate the best fit straight line through the data in Table 2.2.

7.7 HYPOTHESIS TESTING

This section is about methods for determining the probability that a statement is correct. Hypothesis testing is one of the most important parts of statistics to the geologist. However, a proper treatment of this subject requires an entire textbook to itself. In this section I attempt to give an introduction to the underlying concepts of this branch of statistics. The examples I have chosen are designed to illustrate these ideas rather than as realistic examples of the problems actually tackled using hypothesis testing.

A medical scientist attempting to establish whether there is a link between diet and a particular disease is a good example of hypothesis testing. This would be done by attempting to demonstrate that there is no link between diet and the disease. This is called a **null hypothesis**. A null hypothesis is usually a 'hypothesis of no difference' which, in this case, is that 'diet makes no difference to the incidence of the disease'. Other examples of null hypotheses will be given later.

The next step is to attempt to reject the null hypothesis. This is done by attempting to demonstrate that any apparent link shown by the data could not, with reasonable probability, have occurred purely as a result of random fluctuations. To do this, we simply calculate the frequency with which random fluctuations in the data would produce an apparent link at least as strong as that observed. If the probability of an apparent link being a random effect falls below a predefined **significance level**, the null hypothesis is rejected. In the example used here, if the probability of the observed link occurring by chance is 2% and a 5% significance level was chosen, the null hypothesis would be rejected. Thus, the link would be regarded as accepted.

The concepts of 'null hypothesis' and 'significance level' are very important so it is advisable to make sure you understand them.

Question 7.6 A palaeontologist is attempting to determine whether a particular fossil specimen is an *Orthoceras* by comparing its length to those of known *Orthoceras* specimens.

(i) What would be a good null hypothesis for this problem?

(ii) The specimen was 1.5 times as long as an average *Orthoceras* and the probability of this occurring for a true *Orthoceras* was given as 10%. Should the specimen be accepted as an *Orthoceras* if a 5% significance level and your null hypothesis are being used?

To illustrate hypothesis testing, I will use an example based upon the beach pebbles discussed in section 7.2. The sample of 100 pebbles was found to have a mean mass of 610 g based upon a sample of 100 pebbles. Suppose that there are some theoretical grounds for expecting the beach to have pebbles with a mean mass of 550 g and a standard deviation 400 g (perhaps based upon a study of tides and currents in the area). Thus, the observed mean is 60 g heavier than the value predicted by the theory. Now, statistics vary slightly from sample to sample so is the observed mean consistent with the predicted value or is the theory wrong?

The null hypothesis, in this example, is the statement that there is no difference between the population from which the sample of 100 pebbles was taken and a population with a mean of 550 g and a standard deviation of 400 g. We need to calculate the probability that a sample of 100 pebbles from such a population would have a sample mean deviating 60 g or more from the population mean. If this probability is low, the null hypothesis will be rejected as unlikely and we should conclude that the theory is wrong.

So, how much do sample means tend to differ from population means? Imagine collecting a large number of samples each consisting of n pebbles taken from the same population. The means of these samples will be slightly different and each will deviate from the true population mean. Populations with a large standard deviation will give rise to a wider scatter in the sample means than populations with a small standard deviation. In addition, the sample averages will tend to be closer to each other for larger sample sizes. The simplest way to specify the scatter in resultant mean weights is to give the standard deviation of the sample means. This figure is called the **standard error**, s_e, and is given by

$$s_e = \sigma/\sqrt{n} \tag{7.18}$$

where σ is the population standard deviation and n is the size of the sample (proof of this statement is well beyond the scope of this brief introduction). Note that this definition does indeed have the expected dependency on standard deviation and sample size. Thus, in the case of the problem we are considering, the standard error of the theoretical population is

$$s_e = 400/10 = 40\,\text{g} \tag{7.19}$$

and therefore, if the pebble weight theory is correct, 68% of all random samples of 100 pebbles should have a mean within 40 g of 550 g (cf. section 7.3 on standard deviation where it was stated that roughly 68% of all measurements should fall within 1 standard deviation of the mean).

Question 7.7 The pebble masses from Table 7.1 were drawn from a population whose mean and standard deviation (as well we could estimate them) were 610 and 436 g respectively. Assuming these figures are correct, what would you expect the standard error for a sample of 10 pebbles to be? Calculate the mean mass for any 4 samples of 10 pebbles drawn from Table 7.1. Hence, determine the deviation of these means from 610 g. How do the size of these deviations compare to the standard error (i.e. are they much bigger, much smaller or about the same size)?

In other words, 32% of all samples have means more than 40 g away from 550 g. Clearly, the deviation we actually have (observed mean – predicted mean = 60 g) is larger than the standard error and we can therefore conclude that considerably fewer than 32% of all samples would deviate by 60 g or more. In other words, the null hypothesis can be rejected at the 32% significance level. However, a 32% significance level is considerably higher than that normally used. It would be a rash gambler who concluded from this that the theory of pebble weights is wrong. So, what sort of significance level should be set?

When testing a hypothesis, there are two types of error that can be made. Firstly, we might accept as true a hypothesis which is actually false. Secondly, we might reject a hypothesis which is actually true. The best level of significance to set is determined by the relative importance of these two types of error. If it is very important that we do not accept a false null hypothesis then the significance level should be set high. This will ensure that the null hypothesis is relatively easily rejected. On the other hand, it might be more important that we do not reject a true null hypothesis. In this case the significance level should be small so that it relatively easy to accept the hypothesis. These ideas should be made a little clearer by Figure 7.8 which

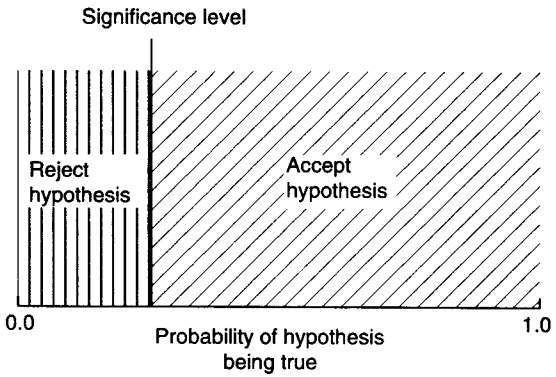


Figure 7.8 The relationship between hypothesis probability, significance level and whether or not the hypothesis is accepted.

of significance

Significance (%)	*Number of standard errors*
20	1.29
10	1.64
5	1.96
1	2.58
0.1	3.29

illustrates the relationship between probability, significance level and whether a hypothesis is accepted. The exact level to set is a question of judgement; there are no concrete mathematical rules which can be applied. However, significance levels of 1 or 5% are, perhaps, the most commonly used values. This implies that we are normally making it relatively easy to accept the null hypothesis. Thus, it is very important that a null hypothesis is chosen intelligently!

So, is the pebble weight theory acceptable at the 1 or 5% significance levels? Table 7.9 gives the multiple of standard error necessary to reach various levels of significance. Thus, for example, 20% of all sample means are beyond 1.29σ of the population mean. Our pebble sample deviates from the predicted mean by 1.5 standard errors ($1.5 \times 40 = 60$ g). Thus, the null hypothesis should be accepted at both the 1 and 5% levels. At this stage we would, provisionally, accept the pebble weight theory as consistent with the data.

Question 7.8 Would you also accept the pebble weight theory at the 0.1% significance level?

7.8 MORE ADVANCED HYPOTHESIS TESTING

The above example is actually a very simple case of hypothesis testing. In other cases we might wish to test whether a particular fossil is an *Orthoceras*, whether a sedimentary sequence is random or whether the distribution of pebble masses has a particular shape. In each of these cases a slightly different type of test is necessary. Statistics textbooks are full of many such tests and a large part of the skill in using statistical methods consists of deciding what you wish to prove, designing appropriate fieldwork or experiments and selecting the best statistical test. However, in nearly all cases the statistical procedure can be broken down as follows:

1. The first step in devising a test is to set up a null hypothesis that there is no difference between the population being investigated and some hypothetical population. Thus, for fossil identification, the null hypothesis is that the specimen is an *Orthoceras*. For the sedimentary sequence case, the null hypothesis is that the sequence is random. For the pebble mass case, the null hypothesis might be that the masses have a Gaussian distribution. Note that, associated with the null hypothesis, there is a theoretical population for which the null hypothesis is true. In the fossil identification case this is the population of all *Orthoceras* specimens. In the sedimentary sequence case, this is the population of all random sedimentary sequences. In the pebble distribution case, this is all possible pebbles drawn from a Gaussian distributed population.
2. Next, a level of significance is chosen. It is good practice to decide this in advance. You should never adjust the level of significance to aid acceptance of a pet theory!
3. Having set up a null hypothesis and a significance level we must now choose a statistical test. This determines the probability that the observed deviations from the null hypothesis could have occurred by chance if the sample really came from the theoretical population. If this probability is less than the significance level we have set then the null hypothesis is rejected and the sample is assumed to be from a different population. If this probability is higher than the significance level we set then the null hypothesis is accepted and the sample is assumed to be from the theoretical population.

FURTHER QUESTIONS

7.9 Measured dip directions for a particular bed at 20 different locations in a field area are given in Table 7.10.

(i) Use these data to calculate the frequency distribution and estimate the probability distribution using 45° wide classes.

Table 7.10 Dip directions for one bed at different locations within a field area

Location No.	*Orientation (°E of N)*	*Location No.*	*Orientation (°E of N)*
1	27	11	14
2	63	12	355
3	10	13	300
4	87	14	96
5	103	15	276
6	256	16	190
7	200	17	191
8	23	18	23
9	17	19	10
10	25	20	5