7 Statistics

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 are 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 are 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 more subtle and even experts can, and do, make very serious errors. However, don't 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 four times. Tabulate the results in terms of the number of different ways of obtaining 0 heads, 1 head, 2 heads, 3 heads or 4 heads. What is the most likely number of heads to get?

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 one hundred or one thousand 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 word 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 word 'specimen' whenever possible.

Question 7.2 If I have six books from a library containing 10 000 books, do these six 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 the pebbles. 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 masses of 100 pebbles sampled from a beach.

Pebble number	Mass (g)	Pebble number	Mass (g)	Pebble number	Mass (g)	Pebble number	Mass (g)
1	374	26	294	51	284	76	287
2	389	27	256	52	403	77	340
3	358	28	359	53	341	78	401
4	395	29	352	54	435	79	422
5	371	30	330	55	307	80	369
6	334	31	269	56	420	81	379
7	224	32	355	57	342	82	432
8	335	33	283	58	331	83	368
9	256	34	301	59	331	84	338
10	340	35	346	60	331	85	327
11	374	36	393	61	290	86	433
12	423	37	386	62	383	87	370
13	338	38	338	63	370	88	343
14	373	39	380	64	302	89	450
15	342	40	357	65	394	90	318
16	242	41	326	66	329	91	384
17	318	42	403	67	324	92	355
18	454	43	317	68	283	93	366
19	346	44	301	69	355	94	324
20	408	45	394	70	311	95	353
21	403	46	407	71	265	96	277
22	384	47	350	72	364	97	359
23	397	48	375	73	322	98	400
24	307	49	303	74	283	99	314
25	409	50	384	75	367	100	389

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

$$\begin{aligned}\overline{w} &= \text{(Total mass of the sample)/(number of pebbles)} \\ &= 35\,018/100 = 350.18 \text{ g}\end{aligned} \tag{7.1}$$

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

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

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 (374 g), w_2 is the 2nd mass (389 g) and so on). The symbol 'Σ' (sigma) is an instruction to add together the w_i's (i.e. add the masses of the pebbles together). The $i = 1$ below the Σ, 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. Note also that drawing a line over a letter (e.g. $\overline{w}$), to denote an average, is a very common convention.

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 then 353 g and 352 g, respectively. Thus, the median mass is the average of 353 and 352, i.e. 352.5 g. Note that this is different to 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 224 g to 454 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

Table 7.2 Pebbles ranked according to decreasing mass.

Rank	Mass (g)	Rank	Mass (g)	Rank	Mass (g)	Rank	Mass (g)
1	454	26	384	51	352	76	318
2	450	27	384	52	350	77	318
3	435	28	383	53	346	78	317
4	433	29	380	54	346	79	314
5	432	30	379	55	343	80	311
6	423	31	375	56	342	81	307
7	422	32	374	57	342	82	307
8	420	33	374	58	341	83	303
9	409	34	373	59	340	84	302
10	408	35	371	60	340	85	301
11	407	36	370	61	338	86	301
12	403	37	370	62	338	87	294
13	403	38	369	63	338	88	290
14	403	39	368	64	335	89	287
15	401	40	367	65	334	90	284
16	400	41	366	66	331	91	283
17	397	42	364	67	331	92	283
18	395	43	359	68	331	93	283
19	394	44	359	69	330	94	277
20	394	45	358	70	329	95	269
21	393	46	357	71	327	96	265
22	389	47	355	72	326	97	256
23	389	48	355	73	324	98	256
24	386	49	355	74	324	99	242
25	384	50	353	75	322	100	224

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:

$$\sigma^2 = \overline{(\text{mass} - \text{average mass})^2} \tag{7.3}$$

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 Eqn. 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 Eqn. 7.3 to a sample rather than the population. In terms of the notation introduced earlier this gives

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

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

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

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 square of the pebble 1 mass is 374 × 374 = 139 876 g^2. Repeating this for all pebbles and taking the average then gives a mean square mass of 124 876 g^2. The square of the mean mass, on the other hand, is 350.18^2 = 122 626 g^2. Thus, using Eqn. 7.5, the sample variance is 124 876 – 122 626 = 2250 g^2.

As a measure of the width of the distribution this number has several disadvantages. The first problem is that $\overline{w}$ itself has been estimated from the sample. In fact, the estimate of $\overline{w}$ obtained from Eqn. 7.2 has the property that it is the value which minimizes the sample variance. If another value is used in Eqn. 7.4, in place of $\overline{w}$, a larger value for s^2 is always obtained. Hence, if the true population mean were substituted into Eqn. 7.4 instead of its estimate, $\overline{w}$, 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 $\hat{s}^2 = 100 \times 2250/99 = 2273\ \text{g}^2$.

Another problem with using variance to measure distribution width is that the final number (in this case 2273 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 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 $\sqrt{2273} = 47.7$ g. For reasons that are covered in more detail later (see Section 7.5), this result implies that around 68% of all pebble weights should fall within 47.7 g of the mean value (350.2 g). Thus, 68% of all weights should fall between 302.5 g and 397.9 g. In fact, out of the 100 measurements in Table 7.1, 65 fall in this range which is, of course, 65% 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.

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;
(v) the 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 with which you should be most familiar.

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

Range (g)	Number
201–250	2
251–300	12
301–350	35
351–400	36
401–450	14
451–500	1

Table 7.3 Number of pebbles in Table 7.1 which fall into 50-g-wide classes.

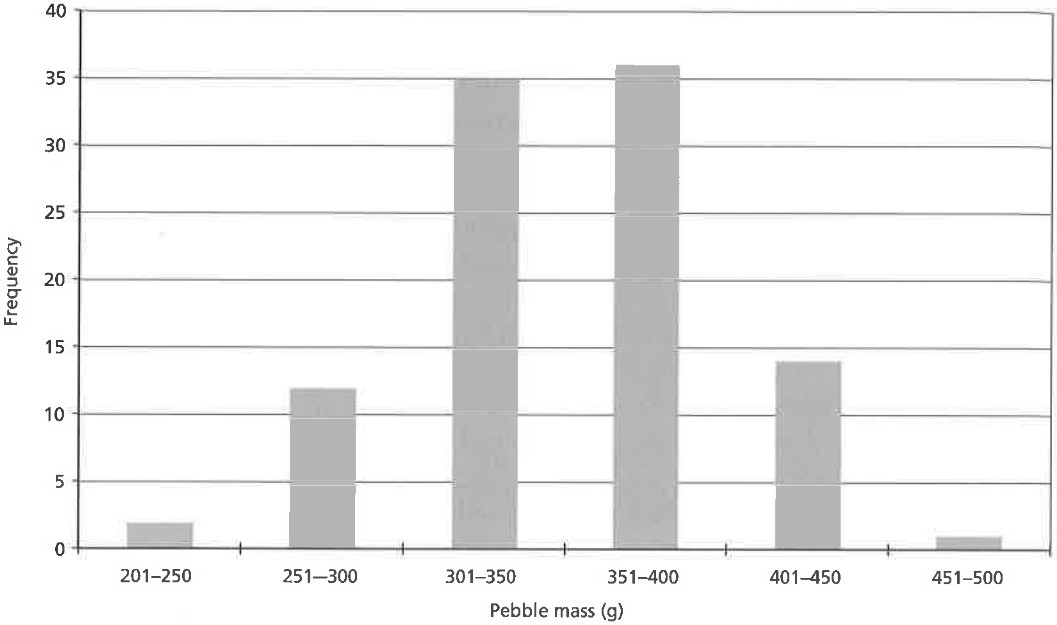

Fig. 7.1 The frequency histogram resulting from plotting the data from Table 7.3.

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 2 pebbles in Table 7.2 with masses between 201 g and 250 g. Table 7.3 lists the number of pebbles with masses in 50-g-wide **classes** from 201 g to 500 g. Such a table is called a **frequency distribution**. If these figures are plotted as a bar chart, the result is a frequency histogram (Fig. 7.1). From this we can instantly see that the masses most commonly fall between 300 g and 400 g.

It was pointed out in Chapter 6 that, if your data is a function of a cyclic variable such as direction or longitude, it is 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

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

usually have considerable scatter due to uncertainties in measurement and the effect of local topographic features. Thus, if the general direction of 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 this 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** (Fig. 7.2) from which it is very clear that the main current direction was roughly NNW. Spreadsheet *Rose.xls* can be used to plot this, and other similar, data.

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 one is certain to occur whilst an event whose probability is zero will never occur.

Similarly, for the data shown in Table 7.3 and Fig. 7.1, the most probable weight range (351–400 g) occurs in 36 cases out of 100, i.e. 36% 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 351–400 g is 0.36. Repeating this

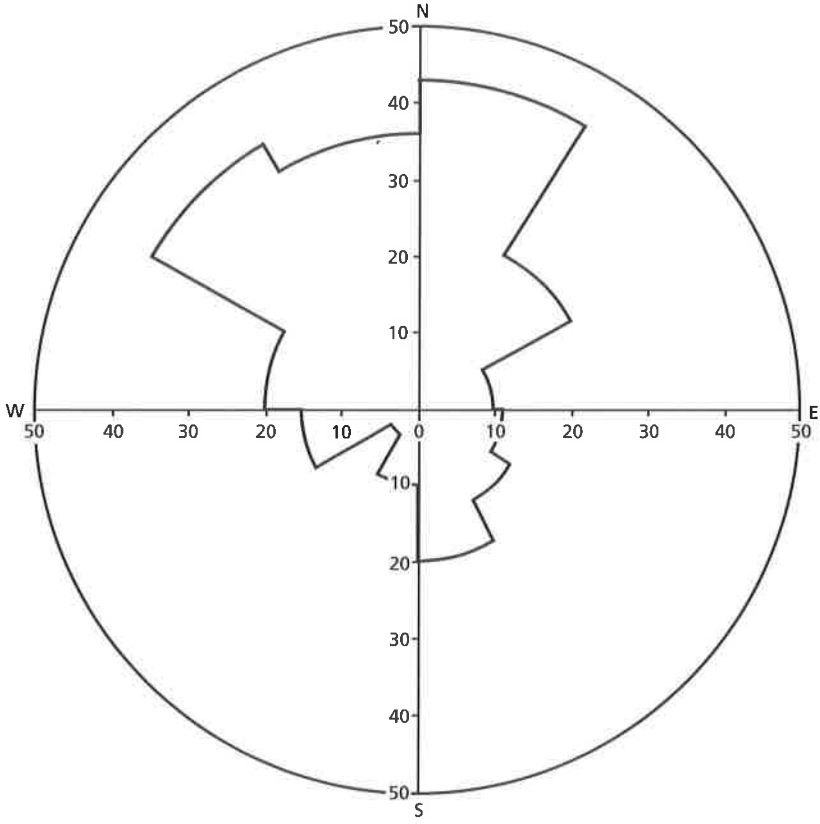

Fig. 7.2 Rose diagram of the data from Table 7.4.

Table 7.5 Estimates of probability for each pebble mass range using the data from Table 7.1. Note that results are simply those from Table 7.3 divided by the number of specimens (100).

Range (g)	Probability
201–250	0.02
251–300	0.12
301–350	0.35
351–400	0.36
401–450	0.14
451–500	0.01

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 (Fig. 7.3). Note that the shape of this is identical to Fig. 7.1 except that the vertical scale has been shrunk by a factor of 100 (i.e. divide 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

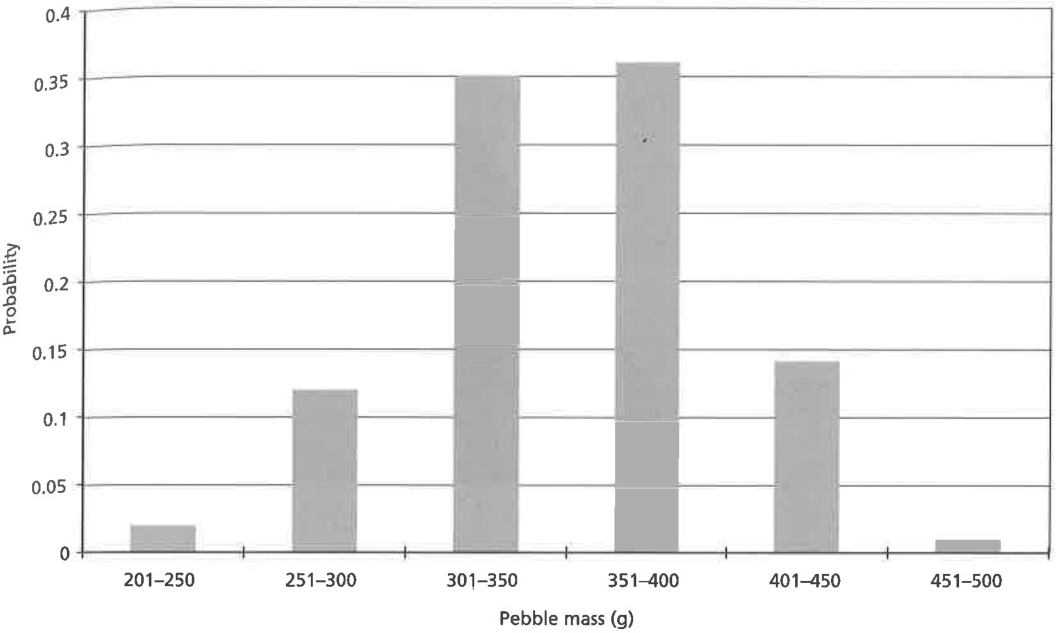

Fig. 7.3 Probability distribution histogram for the data from Table 7.5.

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

where $P(x)$ is called the *relative probability* of obtaining the value x,
$\bar{x}$ is the average of all x values,
σ is the standard deviation of the distribution.
A graph of this function is shown in Fig. 7.4 for the case of mean equal to 5.0 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 dark grey 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 Fig. 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 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

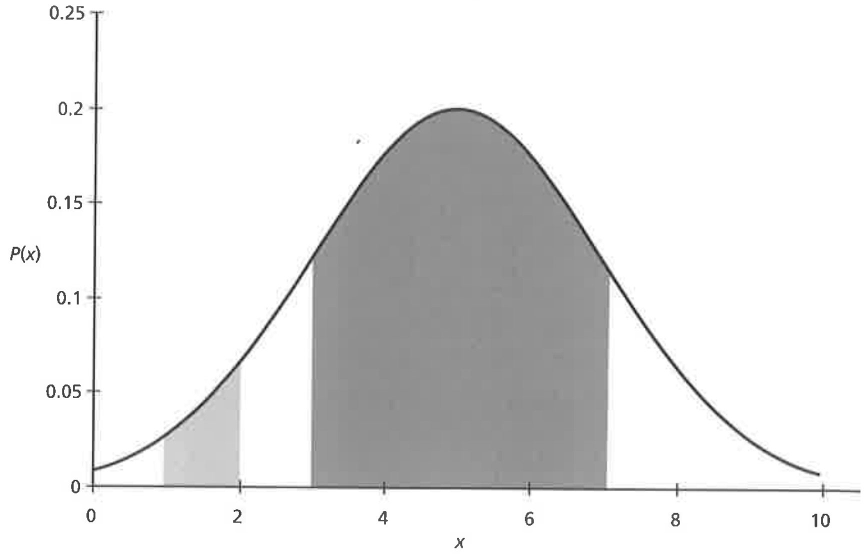

Fig. 7.4 Gaussian probability distribution for a mean of 5.0 and standard deviation of 2.0. The area of the light grey shaded region gives the probability of a specimen lying in the range 1–2. Similarly, the dark grey area gives the probability of a specimen lying between 3–7. The total area under the graph is 1.0.

given by the light grey area shown in Fig. 7.4 whilst the (much greater) probability of obtaining a value between 3 and 7 is given by the dark grey area. 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 Eqn. 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 dark grey region in Fig. 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 Fig. 7.4) is 0.954.

The table can also be used to find areas such as the grey zone in Fig. 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 1.5 σ is 0.866 whilst the area within 2 σ is 0.954. Thus, the area between 1.5 σ and 2.0 σ 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 (see Fig. 7.5). The area between 1.5 σ and 2.0 σ 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 Fig. 7.4, the probability of obtaining a result between 1.0 and 2.0 is 0.044.

Table 7.6 Area under the Gaussian curve as a function of number of standard deviations from the mean. For example the probability of lying within 2 sd of the mean is 0.954.

Number of standard deviations	Area	Number of standard deviations	Area	Number of standard deviations	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

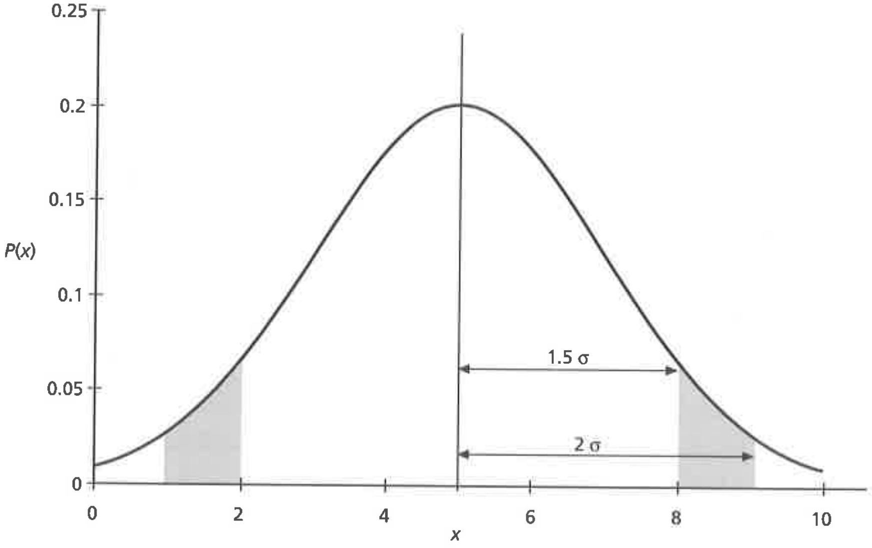

Fig. 7.5 The area under the Gaussian curve between 1.5 and 2.0 standard deviations from the mean. Note that there are two such zones.

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 for the pebble weights on our beach? Now, the mean and standard deviation

Table 7.7 Comparison of the probabilities estimated in Table 7.5 with a Gaussian distribution having the same mean and standard deviation.

Range (g)	Measured probability	Range (multiples of σ)	Gaussian probability
201–250	0.02	−3.10 to −2.06	0.019
251–300	0.12	−2.06 to −1.02	0.134
301–350	0.35	−1.02 to 0.02	0.354
351–400	0.36	0.02 to 1.06	0.347
401–450	0.14	1.06 to 2.10	0.127
451–500	0.01	2.10 to 3.13	0.017

for the pebble weights have been estimated to be about 350 g and 48 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 Table 7.6). For this data set the Gaussian model seems to be pretty good.

Question 7.4 The range from 401 g to 450 g starts 1.06 standard deviations above the mean and ends 2.10 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 1.06 standard deviations (call it $P_{1.06}$) assume that it is given by the expression

$$P_{1.06} \approx 0.6\, P_{1.1} + 0.4\, P_{1.0}$$

i.e. an average of the probabilities corresponding to 1.0 σ and 1.1 σ weighted towards the 1.1 σ probability. Check the result using spreadsheet *Gauss.xls*.

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

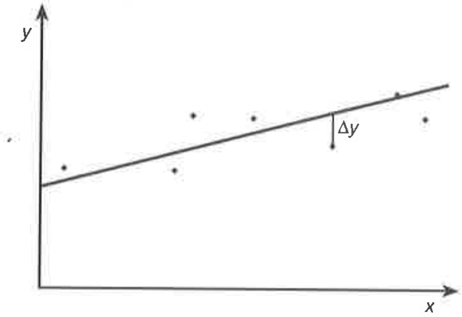

Fig. 7.6 A straight line drawn through some *x–y* data such that it passes close to all of the points. The deviation, Δ*y*, of one of the points from the line is also shown.

line should be a minimum. Figure 7.6 illustrates this idea. The graph of y as a function of x consists of seven points and a straight line has been drawn which passes close to all of these. The deviation, Δy, of one 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 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

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

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

$$c = \bar{y} - m\bar{x} \qquad (7.9)$$

Table 7.8 shows the age versus depth data used in question 2.11 of Chapter 2 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

Depth (cm)	Age (years)
407	10 510
545	11 160
825	11 730
1158	12 410
1454	12 585
2060	13 445
2263	14 685

Table 7.8 Age versus depth data from question 2.11 but with first data point excluded.

through these points? First, we should construct all the summations given in Eqn. 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 Eqn. 7.8 are then

$$\begin{aligned}\Sigma\, depth &= 407 + 545 + 825 + 1158 + 1454 + 2060 + 2263 \\ &= 8712 \text{ cm}\end{aligned} \tag{7.10}$$

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

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

$$\begin{aligned}\Sigma\, (depth^2) &= 407^2 + 545^2 + 825^2 + 1158^2 + 1454^2 + 2060^2 + 2263^2 \\ &= 13\ 963\ 148 \text{ cm}^2\end{aligned} \tag{7.13}$$

From Eqns. 7.10 and 7.11, together with the fact that there are seven measurements, the mean depth and mean age are

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

and

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

Thus, substituting results 7.10 to 7.13 into Eqn. 7.8 gives

$$\begin{aligned}m &= \frac{7 \times 113\ 635\ 245 - 8712 \times 86\ 525}{7 \times 13\ 963\ 148 - 8712 \times 8712} \\ &= 1.91 \text{ years cm}^{-1}\end{aligned} \tag{7.16}$$

Substituting this into Eqn. 7.9 gives the best estimate of the intercept as

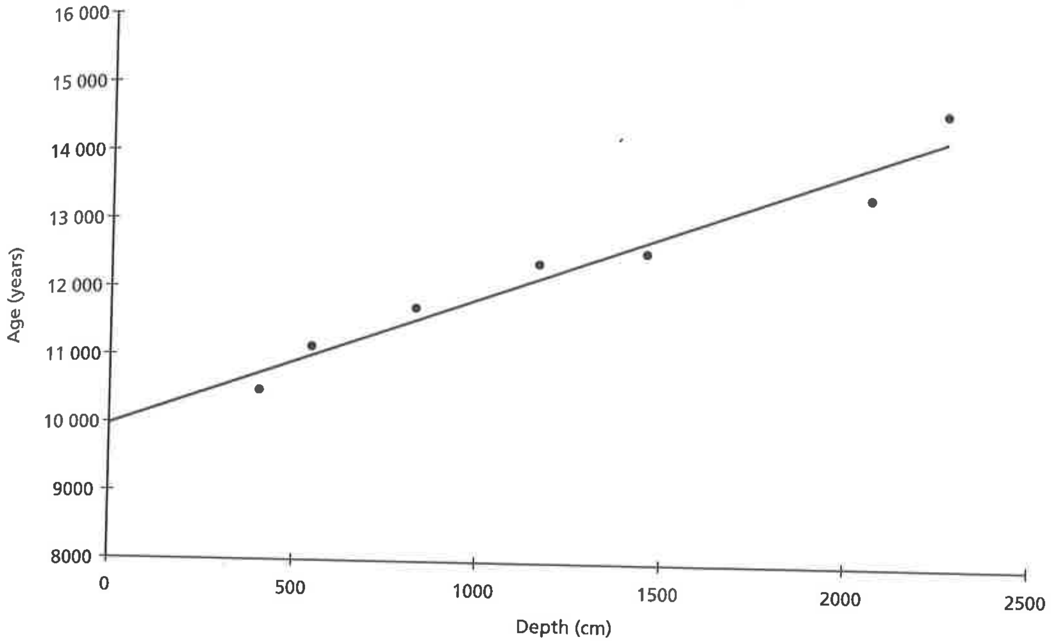

Fig. 7.7 The data from Table 7.8 together with a best fit straight line.

$$\begin{aligned} c &= \overline{age} - (m.\overline{depth}) \\ &= 12\ 361 - (1.91 \times 1245) \\ &= 9983 \text{ years} \end{aligned} \tag{7.17}$$

The data from Table 7.8 is plotted in Fig. 7.7 together with a straight line of gradient 1.91 yr 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.

Question 7.5 Calculate the best fit straight line through the data in Table 2.2 of Chapter 2.

The principles behind linear regression can be applied to other functions. For example, spreadsheet *Bfit.xls* is set up to calculate best fit polynomials through datasets and can be used for fitting curves of the form of Eqn. 2.9 (see Section 2.4) for $n = 0, 1, 2, 3$ and 4. Note that for $n = 0$ it will simply calculate the mean of the data, for $n = 1$ it will find the best fit straight line and for $n = 2$ it will fit a quadratic function.

7.7 The importance of error estimates

All data values are wrong! It is impossible to measure anything with infinite precision and so any measured quantity must differ slightly from its true value. A pebble mass may be measured to the nearest gram but is unlikely to really weigh an exact number of grams. Pebble 1 in Table 7.1 is quoted as weighing 374 g but it might be 374.126 547 g or 373.556 321 678 g or any other mass between 373.5 g and 374.5 g.

To make matters worse, many measurements are hard to make even as accurately as the measuring instrument can theoretically manage. Dip and strike measurements, for example, could in theory be measured with a standard geologist's compass clinometer to the nearest degree. In practice it is hard to make measurements this accurately and most field measurements are probably in error by several degrees.

An additional problem is that data values themselves may be inherently variable. The dip of a planar bed will change slightly between nearby locations because no real sedimentary surface is perfectly flat. Hence, any single measurement on that bed is likely to give a dip that differs at least a little from the average bed dip.

Table 7.9 shows dip and strike measurements taken by 20 different students at two different field locations. You can see that, for either of the two locations, the individual dip values vary by as much as 10° whilst the strike values are even more variable. These variations occur for the reasons given above, namely that the measurements are hard to do accurately and, in any case, the dip and strike depend upon the exact spot at which the measurements were taken.

Data without error estimates are useless. Take student number 1's measurements in Table 7.9. An interesting question for the student to ask might be 'is the dip higher at location B than at location A?' From his dips of 22° at location A and 25° at location B it is impossible to say since the increase may be real or may be entirely due to measurement error. Dips differing by 60° are clearly different but what if they differ by 1°? Most geologists would agree that a 1° difference is not significant, but why would they say this? The reason is that 60° is much bigger than any likely uncertainty in the measurement whilst a 1° difference is a lot less than variations caused by measurement error or unevenness of the bed.

Carefully taken dip measurements should be accurate to about 3° (I'll explain how to estimate this in Section 7.8). Hence, with student number 1's data, it is very hard to be sure if the dip has increased between locations. The difference between the dips is uncomfortably close to the uncertainty in the measurements.

Table 7.9 Dip and strike measurements from two locations measured by 20 different students.

	Location A		Location B	
Student	Strike	Dip	Strike	Dip
1	51	22	60	25
2	45	23	54	22
3	44	21	69	22
4	61	24	59	19
5	37	21	58	27
6	11	23	59	16
7	26	23	66	18
8	54	21	62	22
9	42	22	48	23
10	29	22	62	16
11	48	22	53	16
12	34	18	72	19
13	50	26	62	21
14	61	24	69	21
15	36	26	70	26
16	47	30	41	20
17	30	21	59	16
18	54	25	76	22
19	48	21	54	22
20	55	21	64	12
Average	43.15	22.80	60.85	20.25
Standard deviation	12.70	2.57	8.41	3.80
Standard error	2.84	0.57	1.88	0.85

Despite these types of difficulty, most geological data values are given without error estimates. This might be excusable for dip and strike measurements since most geologists have a pretty good (perhaps subconscious) idea of how accurate such measurements are. However, it is good practice to give error estimates whenever possible. Don't trust anyone's numbers if the person responsible cannot give you at least a rough idea of how accurate they might be.

Question 7.6 Cores taken from two different wells, but from the same sedimentary bed, have sand/shale ratios of 0.50 and 0.51, respectively. If it is hard to determine this ratio with an accuracy better than about 0.02, do you think it is safe to infer that the sand/shale ratio is higher in the second well?

7.8 Quantitative estimates of error

Returning to the data in Table 7.9, the fact that there are 20 independent estimates of the bed strike and dip allows much more to be said about the problem of whether the bed has changed orientation between the two locations. To start with, the methods given in Section 7.3 allow average strike/dip and associated standard deviations to be estimated. The results are given in Table 7.9.

As should be clear from Section 7.3, the standard deviation is an estimate of the typical deviation of any given measurement from the mean value. Hence, each individual student's strike measurements have an error of around 10° (strike standard deviations are 12.7° and 8.4° at A and B, respectively) and a dip error of about 3° (dip standard deviations are 2.6° and 3.8° at A and B, respectively). Note that student 1's estimates of the dips at locations A and B (22° and 25°, respectively) differ from each other by an amount which is similar to the measurement uncertainty and so, from student 1's measurements alone, it is impossible to say which dip is larger.

However, the average of the 20 measurements should be much more accurate than the individual values and, in general, should improve as the number of measurements increases. Note that, unlike the impression given by student 1's measurements, the average dip at B is actually less than at A. An estimate of the typical deviation of the sample mean from the true mean is given by the standard error, s_e, where

$$s_e = \hat{s}/\sqrt{N} \tag{7.18}$$

where $\hat{s}$ is the standard deviation and N the number of observations. The resulting s_e estimates are also given in Table 7.9.

The exact meaning of the standard error depends upon the sample size. This is because estimates of a population mean have a ***t*-distribution**. For large sample sizes the t-distribution is similar to the Gaussian distribution introduced earlier but for smaller sample sizes it is different. Figure 7.8 shows the multiple of the standard error within which there is a 95% probability of the true value lying. This depends upon the **degrees of freedom** (explained below) which, for the problem considered in this section, is given simply by $N - 1$. Hence for our sample size of $N = 20$ there are 19 degrees of freedom implying that the true mean has a 95% chance of lying within $2.1s_e$ of our mean estimate. At location A, for example, the true strike has a 95% chance of lying within $2.1 \times 2.9 = 6.1$ degrees of 43.2°. Similarly, the 95% confidence interval for the strike at B is $60.9 \pm 4.0°$. Hence, this analysis implies that the strike values at the two locations are clearly different.

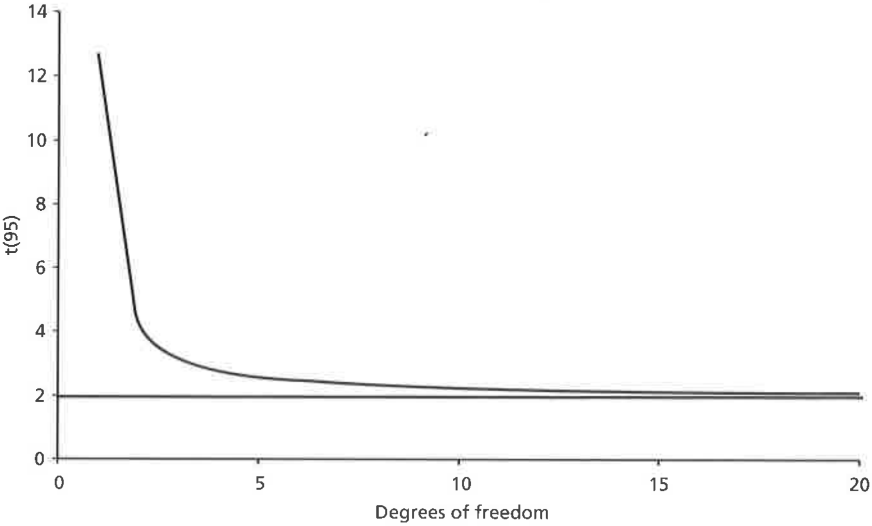

Fig. 7.8 Multiple of standard deviations from the sample mean within which there is a 95% chance of the true mean lying.

Question 7.7 Using Fig. 7.8, find the 95% confidence limits for the dips at locations A and B. Check your answer using spreadsheet *Sterr.xls*.

You should have found, from your answer to question 7.7, that the possible range of dips at locations A and B just overlap implying that, even after averaging 20 measurements, we still cannot say which location has the higher dip. A more sophisticated statistical approach (a ***t*-test**, not covered here) would just allow these particular two values to be distinguished. However, the simple approach used here, i.e. comparing 95% confidence ranges to determine whether two values are significantly different, will suffice for many purposes. Note that it is inherently conservative in the sense that only very well separated values will be distinguished as clearly different.

For completeness, a few words need to be said concerning the concept of degrees of freedom. Generally, if there are N specimens in a sample then there is freedom to change N numbers (i.e. there are N degrees of freedom for N measurements). However, in the problems discussed above, we first used the N measurements to determine a mean value. Once a mean has been set it is only possible to change $N - 1$ values since the Nth value is forced to be a particular number if the mean is to be retained. Hence, after estimating both a mean and a standard error, the confidence intervals are determined from the t-distribution for $N - 1$ degrees of freedom. Don't worry too much about this rather abstract point. In practice, statistics books give quite explicit instructions on how many degrees of freedom are appropriate for a particular problem.

Table 7.10 Dip directions for one bed at 20 different locations.

Location	Direction	Location	Direction
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

7.9 Further questions

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

(i) Use this data to calculate the frequency distribution and estimate the probability distribution using 30° wide classes.

(ii) Plot the frequency distribution on a suitable type of histogram.

(iii) Is there an overall trend for this data?

7.9 In Section 7.3 it was stated that

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

can be rearranged to give

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

Verify this by:

(i) Multiply out $(w_i - \bar{w})^2$ in Eqn. 7.4

(ii) Split the result into three separate summations using the following relationship:

$$\sum (a + b + c + \ldots) = \sum a + \sum b + \sum c + \ldots$$

(iii) Simplify each of the resulting summations using the following result:

$\sum ka = k \sum a$ where k is a constant.

Table 7.11 The percentage weight of calcium carbonate and organic carbon in the top 100 cm of core RAMA 44PC. Data from Keigwin, L. Jones, G. and Froelich, P. (1992). A 15 000 year palaeoenvironmental record from Meiji Seamount, far north-western Pacific. *EPLS*, **111**, 425–40.

Calcium carbonate (%)	Organic carbon (%)
6.10	0.35
5.30	0.27
5.30	0.28
6.70	0.35
9.00	0.42
7.20	0.43
3.20	0.22
14.30	0.39
13.40	0.48
15.30	0.68
9.00	0.70
3.40	0.87
6.30	0.86
10.20	0.95
10.50	0.95
13.40	1.23
15.10	1.22
5.70	1.25
1.90	1.05
2.00	0.98

(iv) Finally, use the definition of mean and mean square to simplify further and obtain the required result.

7.10 Skewness has been mentioned several times in this chapter and is a measure of the symmetry of a distribution. One of several possible definitions is

$$Skew = (\Sigma\, (w_i - \bar{w})^3)/Ns^3$$

which will equal zero if and only if the distribution is symmetric.
Evaluate this expression for the same 10 pebbles you used in question 7.3.

7.11 Table 7.11 lists the calcium carbonate and organic carbon weight percentages obtained at various points within the top 100 cm of a core from a seamount in the north-western Pacific Ocean.
(i) Calculate a linear regression for this data.
(ii) Plot the resulting best fit straight line together with a scatter plot of the original data (i.e. plotted as individual points not joined together).
(iii) How well do you think a linear regression works in this case?

7.12 The amounts of subsidence at two nearby locations in the south Pyrenean foreland during the mid-Eocene were as follows:

Age (Ma)	Subsidence at Puig d'Olena (m)	Subsidence at Tona (m)
46	12	15
42.5	183	102
40.1	340	381
38.4	491	641
35.6	742	788

Calculate, using linear regression, the best fit gradient and intercept for a plot of subsidence at Tona versus subsidence at Puig d'Olena.

7.13 The SiO_2 contents for samples taken from two adjacent locations in the Yilgarn Craton of Western Australia were:

Mount Monger	Emu
65.08	60.40
67.51	66.50
59.52	70.43
61.60	61.27
63.09	64.07
63.49	66.33
64.31	67.10
62.70	63.88
67.49	66.72
	69.38

Calculate the mean, standard deviation, standard error and 95% confidence limits for the SiO_2 contents at these two locations. Is there a significant difference between the mean SiO_2 content at Mount Monger and that at Emu?
Data from Smithies, R. and Champion, D. (1999). Late Archaean felsic alkaline igneous rocks in the Eastern Goldfields, Yilgarn Craton, Western Australia: a result of lower crustal delamination? *J Geol Soc* **156**, 561–76.

7.14 Use spreadsheet *Bfit.xls* to check your answers to 7.5, 7.11 and 7.12.

7.15 Use spreadsheet *Bfit.xls* to fit the temperature versus depth data given below. Try fitting it with $n = 0, 1, 2, 3$ and 4.

z (km)	T (°C)
100	1150
400	1500
700	1900
2 800	3700
5 100	4300
6 360	4300
7 620	4300
9 920	3700
12 020	1900
12 320	1500
12 620	1150