1.4 The meaning of \( \sigma \)

In this section we are going to consider in more detail what is meant by the error \( \sigma \) on a measurement. However, since this is related to the concept of the spread of values obtained from a set of repeated measurements, whose distribution will often resemble a Gaussian distribution, we will first have three mathematical digressions into the subjects of (a) distributions in general, (b) the mean and variance of a distribution, and (c) the Gaussian distribution.
Fig. 1.2(a) shows an example of a possible distribution of a continuous variable from a few observations of, say, the height h of 30-year-old men. Since only a few values are available, the data are presented by marking a bar along the h-axis for each measurement. In Fig. 1.2(b), the same data is presented as a histogram, where a fairly wide bin size for h is used and the vertical axis is labelled as n, the number of observations per cm interval of h, despite the fact that the bin size (Δh) used is 10 cm. The actual number of men in a given bin is nΔh, and the total number of men appearing in the histogram is \( \sum n \Delta h \). If 100 times more measurements were available, the number of entries in each bin of the histogram would increase by a large factor (Fig. 1.2(c)), but it would now become sensible to draw the histogram with smaller bins, in order to display the shape of the distribution with better resolution. Because we plot n(h) as the number of observations per cm, irrespective of bin size, the overall height of the histogram does not change much when we change the bin size (see Fig. 1.2(d)). Finally, for an even larger number of observations, we could make the bin size so small that the histogram would approximate to a continuous curve (see Fig. 1.2(e)); this could alternatively be viewed as a very good theoretical prediction about the numbers of men of different heights. Again n(h) is to be interpreted in the same way, but now the total number of men appearing in the histogram is \( \int n(h) \, dh \).

### 1.4.2 Mean and Variance

In order to provide some sort of description of a distribution such as shown in Fig. 1.2, we need measures of the x-value at which the distribution is centred, and how wide the distribution is. The mean \( \mu \) and the mean square deviation from the mean \( \sigma^2 \) (also known as the variance) are suitable for this. For a set of \( N \) separate measurements such as shown in Fig. 1.2(a), they are defined as†

\[
\bar{x} = \frac{\sum x_i}{N} \quad (1.2)
\]

and

\[
\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}. \quad (1.3)
\]

† We adopt the convention that the true mean and variance are denoted by \( \mu \) and \( \sigma^2 \), while the measured mean and variance of a sample are \( \bar{x} \) and \( s^2 \).
The summations in eqns (1.2) and (1.3) extend over the \( N \) values of the sample. In general, the true mean \( \mu \) is not known, and so eqn (1.3) cannot in fact be used to estimate the variance. Instead it is replaced by

\[
s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2. \tag{1.3'}
\]

Thus one measurement of a quantity does not allow us to estimate the spread in values, if the ‘true’ value is not known. (Other problems of using an estimated rather than a true value of a parameter appear in Sections 1.6 and 3.3.)

As an aside, we note that the factor \( 1/(N-1) \) in eqn (1.3') is required in order to make \( s^2 \) an unbiased estimate of the population’s variance \( \sigma^2 \), i.e. for a large sample, \( s^2 \) will tend to \( \sigma^2 \). If, however, we rewrite the equation as

\[
s^2 = \frac{1}{N+k} \sum (x_i - x)^2 \tag{1.3''}
\]

and set \( k = 0 \) rather than \(-1\), although the estimate \( s^2 \) is now biased, it turns out that its mean square spread from the correct value \( \sigma^2 \) is smaller\( \dagger \) than that for the unbiased estimate (1.3'). In fact, using \( k = +1 \) results in that estimate \( s^2 \) which has the smallest mean square spread from \( \sigma^2 \). (See problems 5.12 and 6.4.) However, we regard \( k = -1 \) as preferable. This is because we are usually interested in using the variance in order to judge the significance of the deviation of a measurement \( x \pm s \) from some preconceived value \( y \). We then calculate \( (x-y)^2/s^2 \), which should be distributed like Student's \( t \) distribution, but especially if the details of how \( s^2 \) was obtained are unavailable, we will be satisfied in assuming that it is Gaussian distributed (see Section 3.3). In that case, it is irrelevant to have a minimum spread estimate for \( s^2 \), since we are interested in \( 1/s^2 \), and it is unreasonably small estimates of \( s^2 \) which would give trouble. It is because the choices \( k = 0 \) and \(+1\) bias the estimates of \( s^2 \) downwards that we regard eqn (1.3') as preferable.

It is most important to realise that \( s \) is the measure of how spread out the distribution is, and is not the accuracy to which the mean \( \bar{x} \) is determined. This is known to an accuracy better by a factor of \( \sqrt{N} \). Thus

\[
d^2 = d^2 + b^2,
\]

where \( b \) is the bias, i.e. \( m - \bar{x} \). As \( k \) increases, \( d^2 \) decreases but \( b^2 \) increases, and there is an optimum value of \( k \) which minimises \( d^2 \).

\[
\text{Fig. 1.3. A histogram of a distribution in } x. m \text{ is the number of entries in each } x \text{ bin; there are 66 entries in all. The mean } \bar{x} \text{ and the variance } s^2 \text{ are estimated as 5.9 and (2.05)t respectively. The accuracy } u \text{ to which the mean } \bar{x} \text{ is determined is smaller than } s \text{ by a factor of } \sqrt{66}.
\]

by taking more and more observations of \( x \), the variance \( s^2 \) will not change (apart from fluctuations) since the numerator and denominator of eqn (1.3) or (1.3') grow more or less proportionally; this is sensible since \( s^2 \) is supposed to be an estimate of the variance of the overall population, which is clearly independent on the sample size \( N \). On the other hand, the variance of the mean \( (s^2/N) \) decreases with increasing \( N \); more data help locate the mean to higher accuracy. (We return to this point, and explain the origin of the 1/N factor, in Section 1.5.)

A minor computational point is worth noting. Formula (1.3') can be written

\[
s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2 = \frac{N}{N-1} [\bar{x}^2 - \bar{x}^2], \tag{1.4}
\]

where \( \bar{x}^2 \) is defined in analogy with eqn (1.2) as

\[
\bar{x}^2 = \sum x_i^2/N. \tag{1.5}
\]

Thus it is not necessary to loop over the data twice (first to calculate \( \bar{x} \) and then to obtain \( s^2 \) from (1.3')), but \( \bar{x}^2 \) and \( x \) can be calculated in the same loop.
1.4.3 Gaussian distribution

Since the Gaussian distribution is of such fundamental importance in the treatment of errors, we consider some of its properties now.

The general form of the Gaussian distribution in one variable \( x \) is

\[
y = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.14}
\]

The curve of \( y \) as a function of \( x \) is symmetric about the value of \( x = \mu \), at which point \( y \) has its maximum value. (See Fig. 1.5.) The parameter \( \sigma \) characterises the width of the distribution, while the factor \( \sqrt{(2\pi)\sigma} \) ensures that

\[
\int_{-\infty}^{+\infty} y \, dx = 1. \tag{1.15}
\]

The parameter \( \mu \) is the mean of the distribution, while \( \sigma \) has the following properties.

† With \( n_{\text{off}} \) set equal to the total number of events.
Fig. 1.5. The solid curve is the Gaussian distribution of eqn (1.14). The distribution peaks at the mean \( \mu \), and its width is characterised by the parameter \( \sigma \). The dashed curve is another Gaussian distribution with the same values of \( \mu \), but with \( \sigma \) twice as large as the solid curve. Because the normalisation condition (1.15) ensures that the area under the curves is the same, the height of the dashed curve is only half that of the solid curve at their maxima. The scale on the x-axis refers to the solid curve.

(i) The mean square deviation of the distribution from its mean is \( \sigma^2 \) (hence the factor of 2 within the exponent in eqn (1.14)).
(ii) The height of the curve at \( x = \mu \pm \sigma \) is \( 1/\sqrt{e} \) of the maximum value. Since

\[
1/\sqrt{e} = 0.607,
\]

\( \sigma \) is roughly the half width at half height of the distribution.
(iii) The fractional area underneath the curve and with \( \mu - \sigma \leq x \leq \mu + \sigma \)

is 0.68.
(iv) The height of the distribution at its maximum is \( (\sqrt{2\pi})^{-1} \). Thus as \( \sigma \) gets smaller, the distribution becomes narrower, and (to maintain the normalisation condition eqn (1.15)) higher at the peak.

By a suitable change of variable to

\[
x' = (x - \mu)/\sigma,
\]

any normal distribution can be transformed into a standardised form

\[
y = \frac{1}{\sqrt{(2\pi)}} \exp(-x'^2/2),
\]

with mean zero and unit variance.

One feature which helps to make the Gaussian distribution of such widespread relevance is the central limit theorem. One statement of this is that if \( x_i \) is a set of \( n \) independent variables of mean \( \mu \) and variance \( \sigma^2 \), then for large \( n \)

\[
y = \frac{1}{n} \sum x_i
\]

tends to a Gaussian distribution† of mean \( \mu \) and variance \( \sigma^2/n \). The distribution of the individual \( x_i \) is irrelevant. Furthermore, the \( x_i \) can even come from different distributions with different means \( \mu_i \) and variances \( \sigma_i^2 \) in which case \( y \) tends to a Gaussian of mean \( (1/n) \sum \mu_i \) and variance‡ \( \sum \sigma_i^2/n \).

The only important feature is that the variance \( \sigma^2 \) should be finite. If the \( x_i \) are already Gaussian distributed, then the distribution of eqn (1.18) is also Gaussian for all values of \( n \) from 1 upwards. But even if \( x_i \) is, say, uniformly distributed over a finite range, then the sum of a few \( x_i \) will already look Gaussian. (This forms the basis of a Monte Carlo technique for generating a Gaussian distribution (see Section 6.4.) Thus whatever the initial distributions, a linear combination of a few variables almost always degenerates into a Gaussian distribution.

The proof of the central limit theorem is given in most standard textbooks (see, for example, Brandt, p. 67). An example of the central limit theorem is given below in Section 1.5.

The Gaussian distribution is discussed again in Section 3.3.

Now that we have concluded our mathematical digressions, we return to our consideration of the treatment of errors.

For a large variety of situations, the result of repeating an experiment many times produces a spread of answers whose distribution is approximately Gaussian; and the approximation is likely to be good especially if the individual errors that contribute to the final answer are small. When this is true, it is meaningless to speak of a 'maximum possible error' of a given experiment since the curve in Fig. 1.5 remains finite for all values of \( x \); the 'maximum possible error' would be infinite, and although this would make it easy to calculate the 'error' on any experiment, it would not distinguish a precision experiment from an inaccurate one.

† For small \( n \), the approximation is better near the peak of the distribution than in the tails.
‡ Provided that the numbers of observations from each of the separate distributions are constant.
Experimental errors

It is thus customary to quote $\sigma$ as the accuracy of a measurement. Since $\sigma$ is not the maximum possible error, we should not get too upset if our measurement is more than $\sigma$ away from the expected value. Indeed, we should expect this to happen with about $\frac{1}{6}$ of our experimental results. Since, however, the fractional areas beyond $\pm 2\sigma$ and beyond $\pm 3\sigma$ are only 5% and 0.3% respectively, we should expect such deviations to occur much less frequently.

One word of warning is necessary. Although experimental distributions often approximate to Gaussians near their peaks, in many cases the experimental data will lie above the calculated Gaussian curve when we look out into the tails of the distribution. This can occur because in a whole series of measurements, a few can have lower accuracy than the majority, and hence be more spread out. Even if these lower accuracy measurements themselves lie on a Gaussian distribution with larger $\sigma$, the combination of two Gaussian distributions of the same mean but different variances does not produce another Gaussian. This can cause problems with attempts to deduce how unlikely is a deviation of several standard deviations from the expected value. As calculated from a Gaussian distribution fitted to the majority of the data, the probability is likely to be very low. But in reality, it could well correspond to a much more probable fluctuation from a subset of data with larger variance. Thus experiments, whose conclusions are based on interpreting the probability of an unlikely occurrence as being $10^{-8}$ rather than $10^{-4}$, are likely to be unreliable unless their measurement accuracies are very closely controlled (see problem 3.5).

1.5 Combining errors

We are often confronted with a situation where the result of an experiment is given in terms of two (or more) measurements. Then we want to know what is the error on the final answer in terms of the errors on the individual measurements.

1.5.1 Linear situations

As a very simple example, consider

$$a = b - c.$$  

(1.19)

† Note that this is a different type of problem from that discussed in the context of the central limit theorem on p. 15.

Combining errors

To find the error on $a$, first differentiate

$$\delta a = \delta b - \delta c.$$  

(1.20)

If we were talking about maximum possible errors, then we would simply add the magnitudes of $\delta b$ and $\delta c$ to get the maximum possible $\delta a$. But we have already decided that it is more sensible to consider the root mean square deviations. Then, provided that the errors on $b$ and $c$ are uncorrelated,† the rule is that we add the contributions $\delta b$ and $-\delta c$ in quadrature:

$$\sigma_a^2 = \sigma_b^2 + \sigma_c^2.$$  

(1.21)

Two points are worth noting

(i) If in a particular experiment we know that the measurements on $b$ and $c$ were incorrect by specific amounts $\delta b$ and $\delta c$, then the answer would be incorrect by an amount $\delta a$, given in terms of $\delta b$ and $\delta c$ by the formula (1.20). But the whole point is that in any given measurement we do not know the exact values of $\delta b$ and $\delta c$ (or else we would simply correct for them), and get the answer for $a$ exactly, but only know their mean square values $\sigma^2$ over a series of measurements. It is for these statistical errors that eqn (1.21) applies.

(ii) For linear combinations like eqn (1.19), it is the errors themselves which occur in eqn (1.21); percentage errors, which are useful for products (see below) are here completely irrelevant. Thus if you wish to measure your height by making independent measurements of the distances of your head and your feet from the centre of the earth, each to 1% accuracy, the final answer will not in general be within 1% of the correct answer; in fact, you may well get a result of $-40$ miles for your height.

Why do we use quadrature for combining these statistical errors? We look at this in several ways

(a) Mnemonic non-proof

The errors on $b$ and on $-c$ can be in phase with each other to give contributions which add up in $\delta a$; or they can be out of phase, so that they partially cancel in $\delta a$. So perhaps on average they are orthogonal to each other and hence Pythagoras theorem should be used for obtaining $\sigma_a^2$. (See Fig. 1.6.)

We stress that this is not a proof; in particular there is no obvious second dimension in which $\delta b$ and $\delta c$ can achieve orthogonality.

† The meaning of 'uncorrelated' becomes clearer later in this section.
(b) Formal proof

Given the relationship (1.19), then for the mean square errors\(\dagger\)

\[
\sigma_a^2 = \langle (a - \bar{a})^2 \rangle \\
= \langle [(b - c) - (\delta - \bar{c})]^2 \rangle \\
= \langle (b - \bar{b})^2 \rangle + \langle (c - \bar{c})^2 \rangle - 2\langle (b - \bar{b})(c - \bar{c}) \rangle \\
= \sigma_b^2 + \sigma_c^2 - 2\text{cov}(b, c).
\]

(1.22)

The last term involves the covariance of \(b\) and \(c\). This is to do with whether their errors are correlated or not. It can be positive, negative or, in the case where the errors are uncorrelated, it will be zero (see Fig. 1.7). Its value is related to the extent to which the exact value of \(\delta b\) in a particular experiment affects that of \(\delta c\). Some examples of correlations are given below.

If the value of \(\text{cov}(b, c)\) is zero, then (1.22) reduces to the expected formula (1.21) for uncorrelated variables.

\(\dagger\) We use both \(\langle x \rangle\) and \(\bar{x}\) to represent the average value of \(x\).

(c) The infinitesimal probability argument

We perform an experiment which consists of tossing an unbiased coin 100 times. We score 0 for each heads and 2 for each tails (i.e. the expectation is \(1 \pm 1\) each time we toss the coin). For the complete experiment, we expect on average to score 100. Now a final score of 0 or 200 is possible, so if we were interested in the maximum possible error, this would be \(\pm 100\). But the probability of achieving either of these scores is only \((\frac{1}{2})^{100}\). Thus if we had a team of helpers such that the experiment could be repeated once every second, we would expect to score 0 or 200 once every \(\sim 10^{24}\) years. Since the age of the earth is less than \(10^{10}\) years, we can thus reasonably discount the possibility of extreme scores, and thus consider instead what are the likely results.

The expected distribution for the final score follows the binomial distribution (see Section 3.1). For 100 throws, this is very like the Gaussian distribution, with mean 100 and \(\sigma \sim 10\). We thus have an example of the central limit theorem mentioned in Section 1.4.3; by combining a large number \(N\) of variables whose initial distribution consists of two \(\delta\)-functions, we end up with something very like a Gaussian distribution,\(\dagger\) the width of which increases only like \(\sqrt{N}\).

(d) Averaging is good for you

We know intuitively that it is better to take the average of several independent measurements of a single quantity than just to make do with a single measurement. This follows from the correct formula (1.21), but not from the incorrect one (1.20).

The average \(q\) of \(n\) measurements \(q_i\) each of accuracy \(\sigma\) is given by

\[
\bar{q} = \sum_i q_i.
\]

(1.23)

Then using (1.21) we deduce that the statistical error \(\delta\) on the mean is given by

\[
n\delta^2 = \sum_i \sigma_i^2 = n\sigma^2,
\]

whence

\[
\delta = \sigma / \sqrt{n}.
\]

(1.24)

Thus we have obtained the result quoted in Section 1.4.2 that the error on the mean is known more accurately than the error characterising the distribution by a factor \(\sqrt{n}\); this justifies our intuitive feeling that it is useful to average.

\(\dagger\) Provided, of course, that we don't look at it with too great a resolution, since this distribution is defined only for integral values, whereas the Gaussian is continuous.
1.5.2 Non-linear situations

So far we have been considering how to combine errors when the formula relating the answer to the measurements is linear. In other cases, the correct answer can be achieved by first differentiating, then collecting together the terms of each independent variable and finally adding these in quadrature, i.e. for \( y(x_1, x_2, \ldots, x_n) \),

\[
\sigma_y^2 = \sum_{i=1}^{n} \left( \frac{\partial y}{\partial x_i} \right)^2 \sigma_{x_i}^2
\]  

(1.33)

Thus, for example, if

\[ a = b^r c^s, \]

(1.34)

where \( r \) and \( s \) are known constants, then assuming the errors on \( b \) and \( c \) are uncorrelated

\[
\left( \frac{\sigma_a}{a} \right)^2 = r^2 \left( \frac{\sigma_b}{b} \right)^2 + s^2 \left( \frac{\sigma_c}{c} \right)^2,
\]

(1.35)

i.e. the fractional errors on \( b \) and \( c \) are combined to give the fractional error on \( a \).

† The reader is invited to think of a set of kinematic variables which possess the property (1.32).
1.6 Combining results of different experiments

When several experiments measure the same physical quantity and give a set of answers \( a_i \) with different errors \( \sigma_i \), then the best estimates of \( a \) and its accuracy \( \sigma \) are given by

\[
a = \frac{\sum (a_i/\sigma_i^2)}{\sum (1/\sigma_i^2)}
\]

(1.38)

and

\[
1/\sigma^2 = \sum (1/\sigma_i^2).
\]

(1.39)

Thus each experiment is to be weighted by a factor \( 1/\sigma_i^2 \). In some sense, \( 1/\sigma_i^2 \) gives a measure of the information content of that particular experiment. The proof of formulae (1.38) and (1.39) can be found in any standard textbook. (See, for example Orear, p. 8; Brandt, p. 97; or problem 5.7(a).)

Some comments on formulae (1.38) and (1.39) are in order.