1B40 Practical Skills

('The result of this experiment was inconclusive, so we had to use statistics' – overheard at an international conference)

In this section we shall look at a few statistical ideas. However, to quote L.Hogben, "the experimental scientist does not regard statistics as an excuse for doing bad experiments".

The Frequency Distribution

If a large number of measurements e.g. n = 500, are taken and a histogram plotted of their frequency of occurrence in small intervals we may get a distribution as in Fig 1.

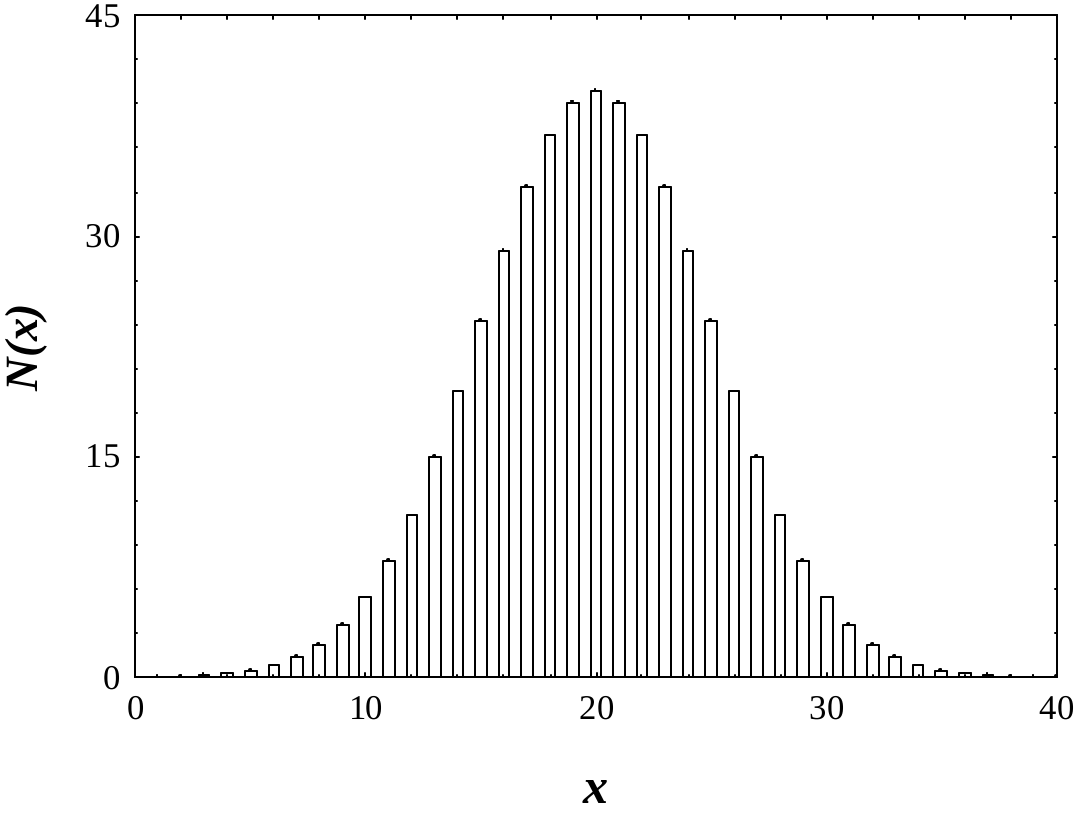

Figure 1. Frequency histogram of measurements

In order to provide some sort of description of this distribution we need measures of the x-value at which it is centred and how wide it is, i.e. some measure of the scatter or dispersion about this mean. The mean μ and mean square deviation s^2 (also called the variance) serve this purpose.

For a set of n measurements the mean and variance are defined, respectively, by

$$\overline{x} \equiv \langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m\right)^2,$$

where we use the symbol μ to denote the true mean of the infinite sample.

The histogram above is a close approximation to, what is called, a continuous frequency distribution $f\left(x\right)$ which would have been obtained for an infinite number of measurements. The quantity $f\left(x\right)dx$ is the probability of obtaining a measurement of x lying between x and $x + dx$. The sum over all probabilities must be unity so the probability distribution must satisfy the integral relation

$$\int_{-\infty}^{\infty} f\left(x\right)dx = 1.$$

The mean $\langle x \rangle$ of the distribution is given by

$$\langle x \rangle = \int_{-\infty}^{\infty} x\, f\left(x\right)dx.$$

Since the number of measurements in the distribution is large and is (assumed to be) free of systematic error $\langle x \rangle$ may be taken as equal to the true value of x. The variance is

$$s^{2} = \int_{-\infty}^{\infty}\left(x - m\right)^{2} f\left(x\right)dx.$$

The Normal (Gaussian) distribution

The frequency distribution may take many forms. One very common one is the *Normal* or *Gaussian* distribution. This follows the functional form

$$f\left(x\right) = \frac{1}{s\sqrt{2p}}\exp\left[-\frac{\left(x-m\right)^2}{2s^{2}}\right],$$

where μ is the (unknown) true value of x that is being measured, and σ is the **standard deviation** and defines the width of the curve. Figure 2 shows a normal distribution with mean value 20, standard deviation 5.

The Gaussian function is relevant to many but not all random processes. The counting of the arrival rate of particles in atomic and nuclear physics is better described by the Poisson distribution. (There is a fuller discussion of these distributions in the lectures accompanying the second-year laboratory course).

Note for the Gaussian distribution μ is also

- the mode -- the most probable value of x i.e. where *f(x)* is a maximum,
- the median -- that x such that the area under the curve for x > μ is equal to that for x < μ, i.e. there is equal probability that a measurement will be greater or less than μ.

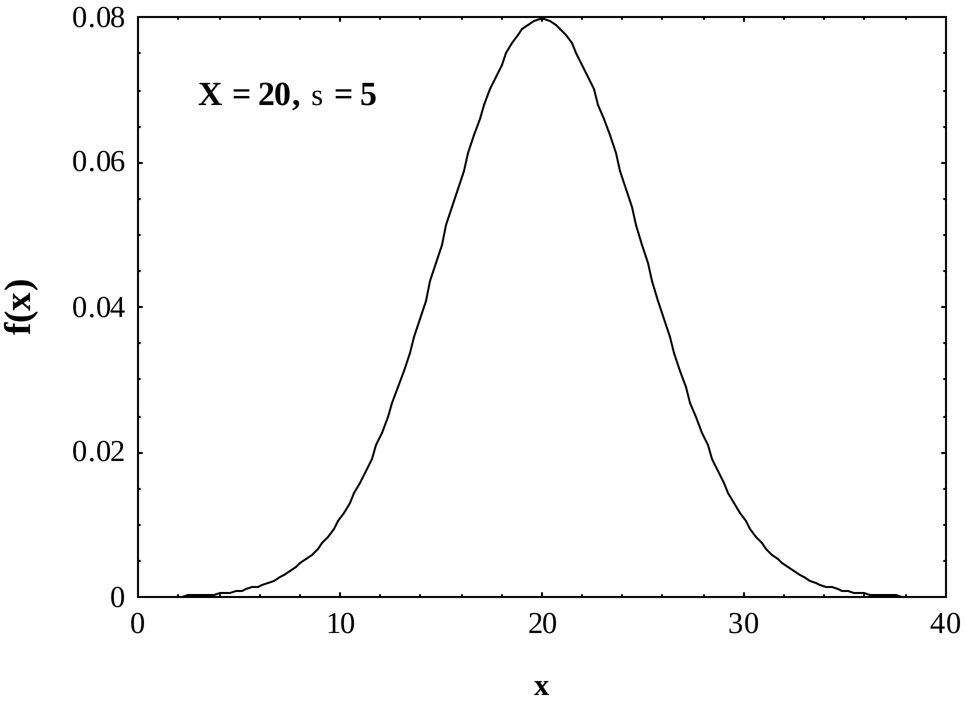

Figure 2. Normal distribution with m = 20 and s = 5.

Best estimate of the true value and the precision for a finite sample

For a set of *n* measurements the mean and variance are defined earlier by

$$\overline{x} \equiv \langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m\right)^2,$$

where we use μ to denote the true mean of the infinite sample.
Since, in general, the true mean is not known we estimate it by $\overline{x}$ and so write the variance as

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2,$$

where the residual, d_i, is the departure of each measurement from the mean,

$$d_i = x_i - \langle x \rangle.$$

It seems plausible, and can be proved, that given a finite set of *n* measurements each of equal quality, the larger we make *n* the nearer the mean $\overline{x}$ approaches μ.

The best estimate that can be made for σ, the (unknown) standard deviation of the infinite population, would be expected to be given by the standard deviation s_n of the *n* readings:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2},$$

with the expectation that s_n approached σ as n becomes large. (This quantity is calculated by the Excel function STDEVP – the standard deviation of the population.) However if only one measurement is made then this leads to s_n being zero which is unreasonable. The better estimate of the standard deviation of the unknown parent distribution from which the sample of n values of x_i are drawn is given by

$$s_{\ n} = s_n\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}.$$

For $n = 1$ this gives 0/0 which is acceptably indeterminate as we have no knowledge of σ from one measurement, $x_{1\ ,}$ alone. Thus one measurement does not allow an estimate of the spread in values if the true value is not known. The expression for the variance above is calculated by the Excel function STDEV- the standard deviation of a sample.

It is worth noting for computational purposes that the variance formula may be written as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2 = \frac{n}{n-1}\left[\langle x^2\rangle - \langle x\rangle^2\right],$$

where $\langle x^2\rangle$ is defined by

$$\left\langle x^2\right\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i^2 .$$

So it is not necessary to loop over the data twice, first to calculate $\left\langle x\right\rangle$ then to obtain s^2, but $\left\langle x\right\rangle$ and $\langle x^2\rangle$ can be calculated in the same loop.

It is important to realise that s_n is a measure of how spread out the distribution is. It is not the accuracy to which the mean value is known. This is known to an accuracy improved by a factor √n as will be shown later, thus the more observations that are taken the better.

The Standard Deviation s and the Standard Error on the Mean s$_{\mathbf{m}}$.

The best estimate for the standard deviation, $s_{\ n}$, of the Gaussian distribution from which a sample of n measurements is drawn is **not** the quantity we need to convey the uncertainty in an experiment, as it does not tell us how well the mean value is known. This Gaussian distribution is the distribution of **single** measurements of the quantity. As $n \to \infty$ then $s_{\ n} \to s$, the width of the distribution obtained for an infinite number of measurements, and $\overline{x} \to m$, but s does not represent the uncertainty on the result of the experiment as expressed by the **mean of the readings**.

What we need to know is how the mean of our sample, comprised of n measurements of the quantity x, would vary if we were to repeat the experiment a large number of times, taking n readings each time and calculating the mean each time. The results for the mean value would be slightly different. We could construct a frequency distribution of the mean values – not that of the individual measurements which s_n measures – and determine the standard deviation of this distribution. This quantity is s_m – the standard error on the mean. The derivation in the appendix shows that

$$s_m = \frac{s}{\sqrt{n}}.$$

The standard uncertainty on the mean, $s_{m,}$ reduces as the square root of the number of measurements. Hence an increase in the number of readings lowers the uncertainty on the mean!

Strictly speaking we don't know the value of σ for the infinite sample. But we can make a reasonable estimate for σ using

$$s = s_n\sqrt{\frac{n}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^2}.$$

Uncertainty on the standard deviation

The standard deviation has itself been estimated from a finite number of measurements, n and so is subject to an uncertainty. It may be shown by considering the Gaussian distribution of the standard deviations that the standard error on the standard deviation is $s/\sqrt{2(n-1)}$. Thus if $n = 10$, $ds/s \simeq 0.24$ and this implies that the estimate we can make of the errors is itself only known to 1 part in 4. Even with $n = 40$, the estimate is still only known to 1 part in 10. Hence **it is almost never valid to quote more than one significant figure when stating uncertainties**.

Summary

If in an experiment we taken n measurements of a quantity x whose unknown value is μ ,

1. the best estimate of μ is the mean

$$\overline{x} = \langle x\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

2. the standard deviation, s_n, of this (large) sample of n measurements is

$$s_n^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/n = \sum_{i=1}^{n} d_i^2/n,$$

where the residual, $d_i = x_i - \langle x\rangle$, is the departure of each measurement from the mean.

3. when n is small a better estimate for the standard deviation of the sample is

$$s_n = s_n\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{(n-1)}}.$$

4. the standard error on the mean, s_m is given by

$$s_m = \frac{s}{\sqrt{n}} \simeq \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n(n-1)}}.$$

To use these results in practice find

- the mean of your readings as in step 1,
- the standard deviation of the residuals (step 2 or 3) provided you have enough readings to make this sensible, say 8 or more. (Otherwise estimate the uncertainty from the precision of the instrument),
- the standard error on the mean from step 4.
-

Estimating the Errors in your Measurements

The standard error on the mean is a simple function of the standard deviation *s*. However, what are we to do if we have too few measurements to make a significant estimate of *s*? The best that can be done is to estimate the standard deviation from some identifiable intrinsic limit of uncertainty of the equipment. For example, suppose you measure a length of 10 mm with a rule graduated in millimetres with no further subdivisions. You would quote the length as 10 ± 1 mm, assuming that you would get a standard deviations on a single measurement of 1mm were you to repeat the measurement a large number of times. (Note that if only a **single** measurement is made any estimate of the error may be widely wrong).

Thus,

- for a small number of measurements, say 1 to 3, estimation of the error will be given by the precision of the apparatus,
- if you make a significant number of measurements of the same variable you should quote:
  variable = average value ± standard error on the mean (s_m).

Appendix

Estimating s and s m from finite samples

The derivation below is given for completeness. Its reading may be omitted if desired!
If we take *n* measurements yielding $x_1, x_{2,} \ldots x_n$ (a random sample of the distribution) the error, *E*, on the mean $\overline{x}$ (of this set of results) from the (unknown) true value μ is given by

$$E = \overline{x} - m = \frac{1}{n}\sum_{i=1}^{n} x_i - m = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m\right) = \frac{1}{n}\sum_{i=1}^{n} e_i\,,$$

where $e_i = x_i - m$ are the individual errors.
The value of E^2 is given by

$$E^2 = \frac{1}{n^2}\left[\sum_{i=1}^{n} e_i^2 + \sum_{i\neq j} e_i e_j\right].$$

If the *n* measurements are repeated a large number of times *N* the e_i would be different for each sample and we would obtain a different value for E^2 from each set of *n* measurements. The average value of E^2 for the *N* data sets would be the standard deviation on the mean, s_m, - the quantity we seek. It is given by

$$\left\langle E^2 \right\rangle = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n} e_{ik}^2 + \sum_{i\neq j} e_{ik} e_{jk}\right)\right],$$

where e_{ik} is the error in reading *i* from the true value *m*. for set *k* from the set *N*. The average of the $e_{ik}\, e_{jk}$ terms will be zero as they are just as likely to be positive or negative, this yields

$$\left\langle E^2 \right\rangle = \frac{1}{N}\frac{1}{n^2}\sum_{k=1}^{N}\sum_{i=1}^{n} e_{ik}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{N}\sum_{k=1}^{N} e_{ik}^2\right).$$

This is wholly equivalent to a sum of *n* data sets with (large) *N* readings (each measurement is a random sample of the distribution). The quantity

$$\frac{1}{N}\sum_{k=1}^{N} e_{ik}^2 = s^{\,2}$$

can be used to measure the standard deviation of the sample and we have *n* of these so

$$\left\langle E^2 \right\rangle = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{N}\sum_{k=1}^{N} e_{ik}^2\right) = \frac{1}{n^2}\left(ns^{\,2}\right) = \frac{s^{\,2}}{n}.$$

Now $\left\langle E^2 \right\rangle = s_m^{\,2}$ and hence we have finally that

$$s_m = \frac{1}{\sqrt{n}}s\,.$$

The best estimate for σ, which is not known, is the standard deviation of the subset of results we have for our *n* measurements:

$$s_n = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m\right)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2,$$

though where μ is not known. However instead of the errors e_i from the unknown true mean m ,we have the deviations (residuals d_i) of each x_i from the mean of the sample,

$$e_i = x_i - m,$$
$$d_i = x_i - \overline{x}.$$

The error, E on the mean is given by

$$E = \overline{x} - m \Rightarrow \overline{x} = E + m.$$

Combining these we have

$$d_i = x_i - \left(E + m\right),$$
$$d_i = e_i - E.$$

Now s_n the standard deviation of the sample is given by

$$s_n^2 = \frac{1}{n}\sum_i \left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_i \left(e_i - E\right)^2,$$
$$s_n^2 = \frac{1}{n}\sum_i e_i^2 - 2E\frac{1}{n}\sum_i e_i + E^2.$$

Since

$$\frac{1}{n}\sum_i e_i = \frac{1}{n}\sum_i \left(x_i - m\right) = \overline{x} - m.$$

we then have

$$s_n^2 = \frac{1}{n}\sum_i e_i^2 - 2E^2 + E^2 = \frac{1}{n}\sum_i e_i^2 - E^2.$$

This is the standard deviation for one set of n measurements. As before we take the average of this over a large number N of sets in the distribution and get

$$\left\langle s_n^2 \right\rangle = \left\langle \frac{1}{n}\sum_i e_i^2 \right\rangle - \left\langle E^2 \right\rangle = s^2 - s_m^2.$$

We have shown that $s_m^2 = s^2/n$ so

$$\left\langle s_n^2 \right\rangle = s^2\left(1 - \frac{1}{n}\right) = s^2\frac{\left(n-1\right)}{n},$$

giving

$$s^2 = \frac{n}{n-1}\left\langle s_n^2 \right\rangle, \quad s_m^2 = \frac{1}{\left(n-1\right)}\left\langle s_n^2 \right\rangle.$$

Strictly the quantity $\left\langle s_n^2 \right\rangle$ obtained by averaging over a large number of sets of data is unknown. The best estimate for this is $s_n^{2.}$ Substituting this we obtain the following *approximate* relations

$$s \approx \left(\frac{n}{n-1}\right)^{\frac{1}{2}} s_n \quad \text{and} \quad s_m \approx \left(\frac{1}{n-1}\right)^{\frac{1}{2}} s_n.$$

We now have expressions for σ and σ_m in terms of an experimentally measurable quantity. **Note that s_m may be reduced by taking more precise measurements or more readings - OR BOTH.**