1. Introduction to Data Analysis and Error Handling

Introduction and Definitions

In any measurement, the numerical value obtained from an experimental instrument is always somewhat different from the true value of the physical quantity being measured. The goal of the experimentalist is to understand this fact of life, and characterise the numerical quantities accordingly – in other words, to estimate both the value of the physical quantity and the error in its measurement. Some useful concepts relating to this issue are given below.

Systematic Error: A systematic error is one which occurs with the same value when we use the instrument in the same way. It may be caused by imperfect calibration of the instrument, imperfect methods of observation, or interaction of the environment with the measurement process. If the source of the systematic error can be identified, it can usually be eliminated.

Random Error: This is always present in any measurement, and is due to the inherently unpredictable fluctuations in the readings of a measuring instrument, or in the experimenter's interpretation of those readings. Random errors show up as different results for what is apparently the same experimental measurement. They can be estimated and reduced by taking multiple measurements and averaging them.

Accuracy: The accuracy of a measurement is the difference between the measured value and the true value of a physical quantity. It involves a description of the systematic errors in the measurement.

Precision: The precision of a measurement is the degree to which repeated measurements under the same conditions give the same results. It involves a descripion of the random errors in the measurement.

In simple terms, a set of data points from repeated measurements of the same quantity are *precise* if their values are close to each other, and *accurate* if their average is close to the true value of the quantity being measured.

Statistics is the branch of mathematics dealing with the collection, organisation, analysis, intepretation and presentation of data. We will use statistical techniques to extract useful information from experimental data.

Random Errors

To understand and manipulate random errors, we assume that measurements, x_i, of a physical quantity come from a probability distribution $p(x)$ which has mean μ and standard deviation σ. The value of the physical quantity is then reported as $\mu \pm \sigma$ or $x \pm \delta x$; in other words, the error δx is the standard deviation of the distribution $p(x)$. The goal of *estimation theory* is to estimate the values of μ and σ given a set of repeated measurements $x_i, i = 1 \ldots N$, of the physical quantity.

The *expectation value* of a function $f(x)$, where x comes from the probability distribution $p(x)$, is defined by

$$E\big[f(x)\big] = \int_a^b f(x)p(x)dx.$$

The *mean* of the distribution is the expectation value of x,

$$\mu = E\big[x\big] = \int_a^b xp(x)dx.$$

The *variance*, σ^2, a measure of the spread of the distribution, is the expectation value of $(x-\mu)^2$, the squared distance of x from the mean μ,

$$\sigma^2 = E\big[(x-\mu)^2\big] = \int_a^b (x-\mu)^2 p(x)dx.$$

σ is known as the *standard deviation*. From the linearity property of expectation values,

$$E\big[c_1 f(x) + c_2 g(x)\big] = c_1 E\big[f(x)\big] + c_2 E\big[g(x)\big],$$

it follows that

$$\begin{aligned}\sigma^2 &= E\big[x^2 - 2\mu x + \mu^2\big] = E\big[x^2\big] - 2\mu E\big[x\big] + \mu^2 E\big[1\big] \\ &= E\big[x^2\big] - \mu^2 = E\big[x^2\big] - E\big[x\big]^2.\end{aligned}$$

Sample Mean and Variance

To estimate the *population mean*, μ, and the *population variance*, σ^2, of the underlying probability distribution, we calculate the *sample mean*, $\overline{x}$, and *sample variance*, s^2, from the measurements, x_i,

$$\begin{aligned}\overline{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i, \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \frac{N}{N-1}\left[\overline{x^2} - \overline{x}^2\right].\end{aligned}$$

The reason for the average in s^2 involving $N-1$ instead of N is due to the fact that $\overline{x}$ is not the same as the actual mean μ. The quantities $\overline{x}$ and s^2 are *unbiased estimators* for μ and σ^2, in that their expectation values give the correct results, $E\big[\overline{x}\big] = \mu$ and $E\big[s^2\big] = \sigma^2$.

To show this result, we note that the *joint probability distribution* for the set of measurements $\{x_1, x_2 \dots x_N\}$ is simply the product of the individual probability distributions for the x_i as the measurements are statistically independent of each other,

$$P(x_1, x_2 \dots x_N) = p(x_1)p(x_2)\dots p(x_N).$$

It follows that

$$E\left[\overline{x}\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[x_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu,$$

so that $\overline{x}$ is an unbiased estimator for μ. The variance of the mean is found from

$$\begin{aligned} E\left[\overline{x}^2\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} E\left[x_i x_j\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} E\left[x_i^2\right] + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i} E\left[x_i\right]E\left[x_j\right] \\ &= \frac{1}{N}\left(\mu^2+\sigma^2\right) + \frac{N-1}{N}\mu^2 = \mu^2 + \frac{\sigma^2}{N}, \end{aligned}$$

so that

$$\operatorname{var}\left[\overline{x}\right] = E\left[\overline{x}^2\right] - E\left[\overline{x}\right]^2 = \frac{\sigma^2}{N}.$$

It follows that the standard deviation, or *error of the mean*, is reduced by the factor $\sqrt{N}$ from the error of an individual measurement. Hence taking the average of N measurements gives a more precise estimate of the sample mean μ. The expectation value of the sample variance is

$$\begin{aligned} E\left[s^2\right] &= \frac{1}{N-1}\sum_{i=1}^{N}\left(E\left[x_i^2\right] - 2E\left[x_i\overline{x}\right] + E\left[\overline{x}^2\right]\right) \\ &= \frac{N}{N-1}\left(\mu^2+\sigma^2 - 2E\left[x_i\overline{x}\right] + \mu^2 + \frac{\sigma^2}{N}\right). \end{aligned}$$

We therefore finally calculate

$$\begin{aligned} E\left[x_i\overline{x}\right] &= \frac{1}{N}\sum_{j=1}^{N} E\left[x_i x_j\right] = \frac{1}{N}E\left[x_i^2\right] + \frac{1}{N}\sum_{j\neq i} E\left[x_i\right]E\left[x_j\right] \\ &= \frac{1}{N}\left(\mu^2+\sigma^2\right) + \frac{N-1}{N}\mu^2 = \mu^2 + \frac{\sigma^2}{N}, \end{aligned}$$

which finally gives the result

$$E\left[s^2\right] = \frac{N}{N-1}\left(\sigma^2 - \frac{\sigma^2}{N}\right) = \sigma^2.$$

Given the set of measurements $\{x_1, x_2 \ldots x_N\}$, the experimental result would be quoted as $\overline{x} \pm s/\sqrt{N}$, since $\overline{x}$ and s^2 are the unbiased estimators for μ and σ^2, and the error of $\overline{x}$ is $\sigma/\sqrt{N}$.

Median and Mode

There are two other measures of the average value of a set of data besides the mean.

The *median* is the value which has an equal number of x_i greater and smaller than it. For an odd number of measurements we simply take the value of the x_i which is in the middle of the list of measurements when ordered by magnitude, $x_{\frac{1}{2}(N+1)}$. For an even number of measurements, we take the average of the two middle numbers in the list, $\frac{1}{2}(x_{\frac{1}{2}N}+x_{\frac{1}{2}N+1})$.

The *mode* is the most commonly occuring value.

For any truly random error, the median, mode and mean of a set of measurements should all be very close to each other. In physics we always work with the mean, as this is the quantity which has the best statistical properties.

Large Data Sets

If the number of measurements, N, is very large, it is obviously impractical to understand the data by examining each x_i one by one. It is much better to look at the *frequency distribution* in the form of a *histogram.*

First the data x_i is organised into *bins* of width a. The size of a most relevant to the problem will depend upon the spread of the measured values x_i.

With discrete data we may group by choosing convenient sets of values. For example, in recording the number of runs scored by cricketers in a Test Match, we might use the intervals 0 to 9, 10 to 19, 20 to 29 etc.

Since continuous data may, in theory, take any real value, we must use appropriate inequalities to ensure that the bin intervals are mutually exclusive. For example, in recording the height h (in metres) for a group of students, we might use the intervals $1 \leq h < 1.2$, $1.2 \leq h < 1.4$, $1.4 \leq h < 1.6$ etc, for which the bin width $a = 0.2$m.

Having decided upon bins, we now determine how many of the measured x_i fall into each bin, which yields the *frequency distribution* f_α, where α is the bin label. The sum over all the bins of f_α will give N, the total number of measurements made. The *normalised frequency distribution* is defined by $p_\alpha = f_\alpha/N$, and gives the probability that a measurement will fall into bin α. If the midpoint of bin α is x_α, then the sample mean and variance are given by

$$\overline{x} = \sum_\alpha p_\alpha x_\alpha, \qquad \overline{x^2} = \sum_\alpha p_\alpha x_\alpha^2, \qquad s^2 = \overline{x^2} - \overline{x}^2.$$

If we label each bin by its midpoint x_α, then we may obtain a discrete version of a *probability density function* $p(x_\alpha) = p_\alpha/a$; this definition still works if we use bins of variable size.

The exact form of $p(x_\alpha)$ will depend upon the bin size a. If a is too large then only one or two of the $p(x_\alpha)$ will be non-zero; if a is too small, then there will be almost no bins with frequency greater than 1. A balance is needed to give the most useful information.

Exercise 1: *An optical system measures length by counting electronically the number of light fringes passing a slit. Three series of measurements of the same length were as follows:*

No of fringes counted	*No. of counts*		
(in excess of 17000*)*	A	B	C
243	0	3	16
244	1	7	60
245	2	15	135
246	2	15	180
247	3	18	194
248	1	14	187
249	0	12	138
250	1	8	64
251	0	4	18
252	0	4	8

Plot these results in normalised form, and calculate the mean, median, mode and standard deviation for each series. Write down the experimental result with error for each series. Draw a smooth curve to represent series C and describe its general form.

Exercise 2: *Twenty-five measurements of the volume of a container were (all in cm^3):*

17.462	17.517	17.483	17.490	17.464
17.485	17.520	17.476	17.478	17.492
17.501	17.488	17.497	17.510	17.509
17.485	17.473	17.515	17.466	17.507
17.483	17.491	17.476	17.505	17.490

Present these results as a normalised histogram (i) with intervals of $0.01cm^3$, (ii) with intervals of $0.02cm^3$. Which is the better choice for the interval and why? Calculate the mean, median, mode and standard deviation of this data. Write down the experimental result with error for the volume of the container.