Imperial College London

Department of Materials

MSE 101: Mathematics and Computer Programming

Introduction to Data Analysis

Prof. David Dye, Dr Sam Cooper

October 2016

Comments and corrections to david.dye@imperial.ac.uk or sjc08@ic.ac.uk

Lecture notes may be found on Blackboard (learn.imperial.ac.uk) or dyedavid.com/mse101

0.1 About this Lecture

This lecture is about data analysis. We briefly introduce how to handle uncertainty and errors, distinguish accuracy from precision, plot good scientific graphs and analyze the dimensions in an equation.

0.2 Course Support and Assessment

The course is primarily delivered through these notes, and the videos that can be found on YouTube via `dyedavid.com/mse101`. In the class session (lecture), we will use learning catalytics to support the learning process, by asking and then answering small conceptual problems. Prior to the session, students must read the notes and/or watch the videos, otherwise the class session won't work!

0.3 Further Resources

GL Squires, *Practical Physics*, 4th Ed., Cambridge University Press, 2001. This short book is an excellent guide to experimental procedure and data analysis, particularly for lectures 5-6 on the treatment of errors and linear regression. It is in the Imperial library at 530.028SQU.

0.4 Dimensional Analysis

0.4.1 Sanity Check

Dimensional analysis is one of the most useful tools for ensuring you don't make a fool of yourself. It's very common in science and engineering for someone to present you with an equation that you've never seen before and ask you to use it. This could even happen in your end of year exam!!

Looking at the dimensions of each of the terms in the equation not only gives you valuable insight into what they mean, but it also allows you to check that it makes logical sense. Consider the following equation:

$$\text{Speed of Impact} = \frac{\text{Distance of Shot} \times \text{Area of Target}}{\text{Density of Projectile}}$$

It might seem obvious to in this case that this equation doesn't make much sense, but using dimensional analysis allows us to specify what's wrong. Let now go through this equation replacing each term with sensible unit classes in square brackets (*N.B.* a meter is a unit as is and inch, but "length" is a class of units that contains both meters and inches).

$$\left[\frac{\text{Length}}{\text{Time}}\right] = \frac{[\text{Length}] \times [\text{Length}^2]}{\left[\frac{\text{Mass}}{\text{Length}^3}\right]}$$

If we now simplify the expression on the right and rearrange, we see that the expression we were given implies that

$$[\text{Time}] = \left[\frac{\text{Mass}}{\text{Length}^5}\right]$$

which I hope you are fairly confident is not true!

0.4.2 Unfamiliar terms

Another useful trick that dimensional analysis can help with is when you have an equation where you know the units of all the terms except one. For example, the famous heat equation, which describes the movement of heat through space and time, is given by

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$$

where T is the temperature; t is time; and x is distance. However, what if we don't know the units of α? To find out, we will once again replace each term in the equation with its unit classes and rearrange to make the units of α the subject.

$$\left[\frac{\text{Temperature}}{\text{Time}}\right] = [?] \times \left[\frac{\text{Temperature}}{\text{Length}^2}\right]$$
$$[?] = \left[\frac{\text{Length}^2}{\text{Time}}\right]$$

It is often more convenient to use units directly (rather than unit classes); however, be careful! If, the units of t were seconds and the units of x were meters, then the units of α would be [m^2/s], but if, instead, x had been measured in inches, then alpha would also need to be changed to [in^2/s].

0.5 Appropriate Precision

Often, you will see people write in lab reports that they measured, say, a concentration of hydrogen in a zirconium alloy of 243.4 parts per million. But, this is obviously wrong, since we can only measure hydrogen in zirconium to a precision of about 30 parts per million, so this is essentially 240 ppm, ±30. So here, we will consider how to handle precision and uncertainty correctly when writing lab reports.

Taking a tensile test specimen with a central section, the diameter of which I need to know, we could use a pair of digital callipers. But, how accurate is the measurement?

Well, I could try measuring it several times. If I do, I get 6.78 6.76 6.74 6.77 6.77 6.80 mm. On inspection, the variability looks like plus or minus 0.1 or 0.2 mm.

But, I could do better! I could find the average, the arithmetic mean of the measurements, and the standard deviation.

We define the mean $\overline{x}$ by taking the sum of all the measurements x_i, and dividing by the number of observations n.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

The standard deviation is defined as follows. We say that the standard deviation squared is the sum of the squares of the differences from the mean, divided by n;

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \tag{2}$$

Now, if I had a hundred observations from all the test samples ever made, I would divide by $n - 1$.

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \tag{3}$$

Usually, we are sampling, so then $(n-1)$ version is the correct thing to do. Now the uncertainty in the mean $\overline{x}$ is given by dividing that standard deviation by $\sqrt{n}$ again;

$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{4}$$

So for the data above, I will get a mean of 6.7717 mm, and a standard deviation of 0.0204 mm. I then obtain an uncertainty in the mean of 0.0083 mm. So I would write its diameter down as 6.772 ± 0.008 mm.

Now, note that I have written down the uncertainty to only 1 digit of precision, or one significant figure. And then the mean I have only quoted to that same level of precision, in this case 8 micrometers, less than 100th of the width of a human hair.

Now, if the did enough measurements we assume that they would follow a normal distribution. The frequency of occurrence of the observations are distributed around the mean. It falls off like the function e^{-x^2} away from the mean, and how fast it falls away is a measure of the standard deviation. In fact, and well show this later in the course, about 68% of the observations fall within plus/minus 1 standard deviation, 95% fall within 2 standard deviations, and 99.8% fall within three standard deviations.

Now, theres a couple of things to say here. To illustrate the point, if you roll a dice lots of times, it doen't initially look very random, but eventually you will obtain a distribution that is flat, with equal occurrence of all the possibilities.

I can calculate the mean, which will be 3.5. If had enough observations, n would be very large, the standard deviation would converge to a value of 1.71, and the uncertainty in the mean would eventually converge to zero.

But, because the data in this case are non-continuous - the dice only gives integer values, and because they arent normally distributed, then although we can calculate a standard deviation, then it isnt guaranteed that 2/3rds of the data are within plus/minus 1 standard deviation. The data being normally distributed is an assumption that isnt necessarily true.

Finally, we need to talk about levels of precision. If you dont have many measurements, then usually, if you assume its normally distributed, you should take two-thirds of the range as being the plus/minus precision you should quote, as a rule of thumb.

If you only have one measurement, you should estimate the level of precision; in my case with my calipers, it looks like 0.1mm. Then, if I quote a diameter of 6.77mm I would be quoting to a 10 micron precision, or 2 decimal places or d.p.

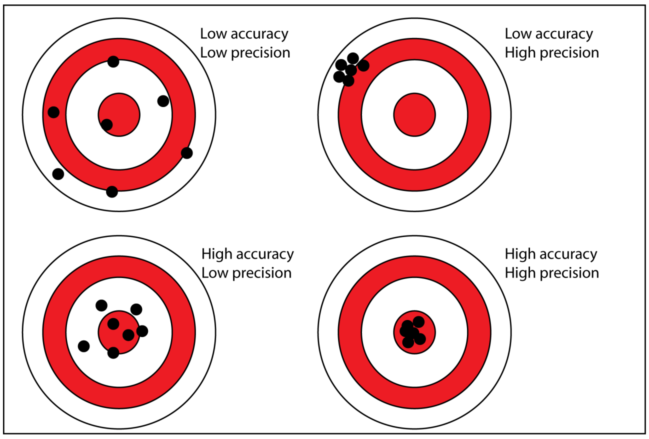
Another option people use are significant figures, s.f. Here we quote the number of numbers accuracy we use. If we measure a number of 1243, then to 3s.f. this rounds to 1240. A measurement of 1245 would round to 1250, for example. If we measured 1200.11, then to 3sf this is 1200. Theres a problem here, which is that unless I tell you, this looks like 2 s.f., but it would be the same number, 1200, to 4s.f. So you have to state your precision when writing down the number - *e.g.* “1200 (3sf).” Similarly, if I talked of a grain of sand having a diameter of 0.0345mm, then this would be 0.034 mm to 2sf, or 34 micrometers to 2 s.f.

So, if you are in a lab and need to report a number in your write-up, that’s how you do it. If possible, make multiple measurements and estimate an uncertainty in the mean. If you cant estimate the uncertainty and then write your result down appropriately. If you can't do that, then quote to a sensible level of precision, state what it is, and explain why.

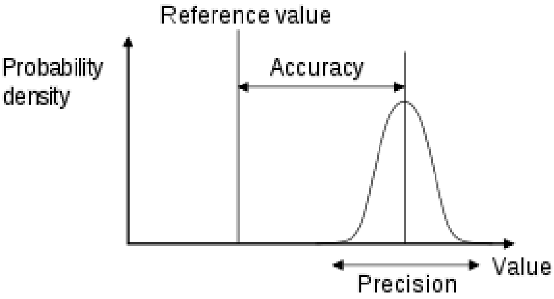
0.6 Precision and Accuracy

Precision and accuracy are similar, but not the same. Precision tells you about the relative location of each data points compared to the others in the set and can be thought of as a measure of "grouping". Accuracy compares the mean location of your data to a target value.

Another way to think of this is that precision measures the random errors, where as accuracy measures the systematic errors. The four target boards in the figure below concisely illustrate this difference.



This information can also be re-expressed in terms of probability density curves as in following figure.



Distinguishing between these two concepts is crucial for diagnosing and hopefully curing problems with the experiments and simulations that you'll be running over the course of your degree.

0.6.1 Reporting numbers

When reporting numbers, it is important to consider the degree of precision that is appropriate. This will be discussed in more detail in the propagation of errors topic in this course, but for it's good enough to just know how to apply the two main methods.

Significant figures

When handling numerical data one is often confronted with question about what is significant or not. For example, are the zeros in the number 0001.000 significant? What about 1,000? In the first case, the leading zeros are not significant but the trailing zeros are. More specifically, the trailing zeros establish the precision to which the number is reported while the leading zeros dont really add any meaning. In the second case the trailing zeros may or may not be significant depending on the context. There are some (general) rules for deciding what is significant:

The following figures *are* significant:

1. All non-zero digits
2. Zeros appearing between two non-zero digits
3. Trailing zeros in a number containing a decimal point

The following figures *are not* significant:

1. Leading zeros

The following figures *may be* significant:

1. Trailing zeros in a number without a decimal point. In these cases there should be some indication of what is significant, e.g. 1000 may indicate 3 significant digits, or 1000 (2 sf) can indicate 2 significant figures.

Decimal places

If a specific number of decimal places is specified, this simply requires you to report to that number of terms after the decimal point (even if they are all zeros), rounding to find the final term if necessary. So, 999.999 quoted to 2 decimal places is 1000.00, while 0 reported to two decimal places is 0.00.

Examples

The following table shows some numbers as well as their significant figures and decimal places using the rules above:

When performing analysis on experiment data it often happens that two sets of numbers of differing precision are combined. In these cases its important to keep in mind what is meaningful or significant. As a general rule, when combining numbers the number of decimal places in the result is the smaller of the number of decimal places in the terms being combined.

Number	Significant Figures	Decimal Places
1.00	3	2
0.100	3	3
1001.00001	9	5
0.000000023	2	9
1000.	4	0
1200 (3 sf)	3	0
15,0$\underline{0}$0	4	0

0.7 Plotting Data

Graphs are a way of plotting a lot of data in a way that visually illustrates the trends, in a way that the relationships between the variables on the axes become apparent. But, casual reading of any newspaper or website will show that often, people plot really poor graphs that are misleading or obscure more than they reveal. In science, we want to seek truth such that the graphs we plot really do illuminate our understanding so its very important that we plot graphs well.

Axes. Firstly, a good graph has axes that make it clear what we are plotting with axis labels that describe the variables and their units properly, preferably not just using symbols (that could have several meanings). The numbers on the axes should use appropriate precision.

Data points. The individual data points should, for preference, have appropriate uncertainties (error bars) plotted, so that the significance of any scatter can be appreciated. Use the simplest symbols you can. If the data are continuous you can use lines (and scatterbands if possible) instead.

Lines. Any line you draw through the points is illustrating a hypothesis about how you think the data are related. So, dont joint the dots! These imply that any noise in the data are real and that you can safely take any point on the line if you want to interpolate. Instead, fit a formula based on your hypothesis say, a straight line. Well return to this point later in the course. *In extremis*, it is permissible to draw a freehand guide to the eye to illustrate the trend in the absence of a quantitative hypothesis, but you should state in the caption that this is what you have done.

Colour and multiple data sets. The data should be clear, preferably if reproduced in black-and-white. So avoid the use of shading and colour where at all possible. Line thickness is a good way to introduce emphasis here.

Chartjunk. A good general rule is to maximise the amount of ink used on data and minimise that ink used on everything else. This means that bogus 3D shadows and other chart junk should be avoided.

Gridlines. Generally, these are to be avoided, as they violate the data-ink maxim. Instead, use major and minor tick marks on both sides of the axes in order to enable the reader to measure accurately from the graph without parallax error.

Ranges. Generally, to make the data trends apparent, the data range should fill as much of the chart as possible. But, if zero is physically significant, as in zero time on the horizontal axis or zero on the vertical axis for a straight line with zero intercept, then the axes should include zero. Similarly, if negative numbers arent physically significant, they shouldnt be plotted.

Legends. If possible, avoid the use of legends, and instead label the data groups, in order to enhance clarity.

Captions. Scientific figures have a numbered caption that explains the figure, usually of 20-50 words. For example, "Figure 2. Tensile behaviour of alloy IN718 during loading at room temperature. The sample was in the fully aged condition described in Chapter 2." Because you have a caption, the figure doesnt also need a title.

Log plots. Often, for example if we have an Arrhenius or power law relation, well want to use a log scale on one or both axes, in order to linearise our data. We could even plot the square or cube root of one of the axis variables in order to achieve this. But, if you plot $\log_{10} t$, where t is time in hours, then how many seconds does a value of 3.5 correspond to? Well $10^3.5 = 3162.8$, so that is 11,400 seconds. So often log scales are plotted with the tick marks and numbers on the axes corresponding to the values, *i.e.* writing 10^3 as 1000 etc. So when you measure distances, you have to raise to the power, but its relatively easy to read from the graph directly, by eye.

Extrapolation and Interpolation. If you have plotted a sensible line reflecting your hypothesis, its probably safe to interpolate between the data points. But, its often unsafe to extrapolate because the mechanisms might change. For example, if you measured the thermal expansion of a material between absolute zero and 700 C, then it might be unsafe to extrapolate this to 1200C because you dont know if the material might melt!

The best book on making graphs and tables ever written is by *Edward Tufte, The Visual Display of Quantitative Information*, which is very entertaining.

0.8 Combining Errors

Following on from the previous discussion on errors, we need to ask, when to quote the mean and uncertainty on the mean, and when to quote the population uncertainty. Consider a group of water bottles and measure how much fluid each contains. If I want to know how much the average bottle contains, I would use the uncertainty on the mean. If you want to know the variability between bottles how much the next bottle picked is likely to contain, you should use the population uncertainty. So most likely, you should use the population uncertainty.

A related question is how we should combine errors. For example, if you want to measure the volume of a cylinder then you can do this by finding the diameter d and height h, then assume that it is a regular, right cylinder and say that its volume V is given by $V = \pi d^2 h/4$. But, how to estimate the uncertainty in the volume? Well look at this properly later in the course, once you know how to do partial differentiation, but for now well simply give some rules.

For multiplication and division, the fractional uncertainties add. For addition and subtraction, the absolute uncertainties add.

So, for example, if we have three variables a, b and c, and associated uncertainties σ_a, σ_b and σ_c, we can say that the fractional uncertainties are $\frac{\sigma_a}{a}$, etc. Then if we have a function $f(a, b)$ given by, for example;

$$f(a, b) = ab \tag{5}$$

Then we can say that

$$\frac{\sigma_f}{f} = \frac{\sigma_a}{a} + \frac{\sigma_b}{b} \tag{6}$$

$$\sigma_f = \sigma_a b + \sigma_b a \tag{7}$$

Then if we have a function $g(a, b, c) = f(a, b) + c = ab + c$ we can say that

$$\sigma_g = \sigma_f + \sigma_c = \sigma_a b + \sigma_b a + \sigma_c \tag{8}$$

We can use this to consider what to do with powers as well. For example, if some function $h(a) = a^2 = a \times a$ then

$$\sigma_h = 2a\sigma_a \tag{9}$$

in fact, this is differentiation, so repeating this for $h(a) = a^n$ gives

$$\sigma_h = na^{n-1}\sigma_a \tag{10}$$

We will address more complicated cases later in the course.

Returning to the example of the cylinder of volume $V = \frac{\pi}{4}d^2h$, we can therefore say that

$$\sigma_V = \frac{\pi}{4}(2dh\sigma_d + d^2\sigma_h) \tag{11}$$

Other ways to find the volume of a cylinder would be to measure the volume directly using Archimedes method, or to measure its mass and then use knowledge of the density of the material composing the cylinder.

One thing to note here is that therefore subtraction of similar sized numbers results in large uncertainties, because the absolute values of the uncertainties add.

Another thing to note is that often, the uncertainty in the final result will be dominated by certain terms, such that some sources of uncertainty can largely be ignored. This is particularly important in the mathematical modelling of materials.