SCM Lecture 2 - teaching notes

Jeanne Wilson

August 16, 2010

1 Combining many measurements

Suppose we have a set of values without errors. I asked a bunch of people down the pub how many units of alcohol they drank per week. OK I'm joking - these are just some numbers I made up. But suppose I have done a survey and nobody gave me any errors on their estimate - lets face it, if I'm interviewing a bunch of drunks in the pub I wouldn't trust their errors anyway. The sensible thing to do is to estimate the uncertainty from the spread of the values we have collected.

Values we can calculate from a set of data:

1.1 Mean

The best estimate of x from a set of measurements.

$$\bar{x} =< x >= \frac{1}{N} \sum_{i=1}^{N} x_i$$

This is the best value to describe the quantity X from a set of measurements.

1.2 Median

The value for which half the measurements are below, half are above. ie. order the numbers and take the middle value. If you have an even number of data values, take the mean of the two central measurements.

1.3 Mode

The most frequent value. The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most ambiguous case occurs in uniform distributions, wherein all values are equally likely. For samples from a continuous distributions it isn't really of use with data in the raw form: (1.654, 1.687, 2.344, 2.346, 9.010) so you would need to histogram the data and determine the modal bin (see histograms later).

1.4 Range

The maximum spread of values

$$x_{\max} - x_{\min}$$

1.5 Residual

The distance of each measurement to the mean.

$$r = (x_i - \bar{x})$$

What is the average of the residuals?
- it is always zero - as we have calculated the mean so that there are values on either side of it.

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{N}{N}\bar{x} = \bar{x} - \bar{x} = 0$$

so this average doesn't tell us much about how spread out the distribution is.

1.6 Root Mean Square - RMS

We are interested in the *magnitude* of deviation from the mean, not the direction, so we take the average of the residuals squared.
The RMS describes the spread of values about the mean. We square the values so that deviation in either direction is positive and things don't cancel.

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2)}$$

1.7 Standard Deviation

For large N this is $\approx$ the RMS - replace N with $N - 1$. This describes the spread of values about the mean and best represents the *68% probable* range of a single measurement, x_i.

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2)}$$

This is the most useful to describe the spread of a set of measurements.

1.8 Standard Deviation of the mean

The mean value can be strongly affected by 1 or 2 measurements.
eg. 2.2, 2.3, 2.4, 2.3, 2.1
These five values have a mean = 2.3
but if some muppet takes another measurement:
eg. 2.2, 2.3, 2.4, 2.3, 2.1, 86
the mean becomes = 16.2!

The more measurements contributing to the mean, the less sensitive it is to one ridiculous value like that. Therefore, it is reasonable to assume the uncertainty in the mean is inversely proportional to N. It is infact described by:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

four times the number of measurements gives half the uncertainty on the mean.

1.9 Full Width Half Maximum

A full width at half maximum (FWHM) is an expression of the extent of a function, $f(x)$, given by the difference between the two x values at which $f(x)$is equal to half of its maximum value. *see slide*
For the normal distribution

$$\text{FWHM} = 2\sqrt{2\ln 2}\sigma \approx 2.35482\sigma$$

1.10 Skewness

Indicates how asymmetric a distribution is.

$$S = \frac{Mean - Mode}{\sigma_x}$$

What is the dimensionality of this? No dimensions.
skewness = 0 for a symmetric distribution.
If the distribution is very skewed, the standard deviation may not be a good estimate of the uncertainty - may need asymmetric errors.

1.11 Summary on assigning uncertainties to repeatable measured quantities

- If you only make one measurement, x with measurement accuracy Δs (eg. length with a ruler with mm gradings) quote $x \pm \frac{\Delta s}{2}$.
- If you have 2–3 values spreading $> \Delta s$, the range is an indication of the uncertainty. Quote the mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \pm \frac{x_{max} - x_{min}}{2}$. This uncertainty can be taken as both the error on the mean and on the single measurement.
- If you have N values with a spread large compared to Δs, again quote the mean, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. For the error on a single value quote the standard deviation, $\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2})$, but for the error on the *m*ean use $\pm\frac{\sigma_x}{\sqrt{N}}$

2 Graphs

Given 2 observables, x and y we often want to know

- What relationship exists between them?
- Parameters of that relationship.

Example:
Spring displacement, y, vs weight. x. We can take a series of measurements of how much the spring stretches as we place different weights on it *(diagram on board).*
What relationship would you expect?
Probably linear relationship - the more weight you put on the more it will stretch. Can test this by plotting y as a function of x. *Slide 1*

- Can see if relationship exists by eye - visually see a trend.
  Yes it looks linear - but what is going on there?
- Can spot mistakes or features. - eg. One point incorrect, measurement lies way below the trend.

If you find an anomalous point - CHECK it.

- Check your equipment
- Remeasure
- Take some other measurements close to/around the anomaly.

Slide 2 - first shot at line through data If the trend looks linear - draw a line so that points are scattered equally above and below. **Don't assume a line goes through the origin**. There may be a strong argument for this to be the case but think carefully. *Slide 3 - a better line through data - but what is going on with last point? Take more data - slide 4 shows that it is trending upwards there - too much weight breaking the spring?*

2.1 Straight line fits

Least squares?

2.2 Gradient

$$\text{Gradient} = \frac{\Delta y}{\Delta x}$$

To assign an uncertainty use the steepest and flattest lines - this gives you $\Delta x_{\max}$ and $\Delta x_{\min}$ and $\Delta y_{\max}$ and $\Delta y_{\min}$. *Draw on board*
Take the mean gradient $= \frac{\frac{\Delta y_{\max}}{\Delta x_{\max}} + \frac{\Delta y_{\min}}{\Delta x_{\min}}}{2}$ and quote the range/2 $= = \frac{\frac{\Delta y_{\max}}{\Delta y_{\max}} - \frac{\Delta y_{\min}}{\Delta x_{\min}}}{2}$ as the uncertainty in it.

$$\text{Gradient} = \frac{\Delta y}{\Delta x}$$

3 Making things easier to plot

What if things aren't linearly related? eg. quadratic. If you know the expected dependency you can redefine things so you are plotting a straight line.

- Example 1 - pendulm

$$T = 2\pi\sqrt{\frac{L}{g}}$$

We want to find g, acceleration due to gravity so we measure the length of the pendulum, L and the period T. We can rewrite relationship as

$$T^2 = L\frac{4\pi^2}{g}$$

So if we plot T^2 against L we should get a straight line with gradient $\frac{4\pi^2}{g}$ (draw it).

- Example 2 - refractive index

$$n = A + \frac{B}{\lambda^2}$$

we measure n and λ and want to find the constants, A and B - what should we plot?
plot n vs $\frac{1}{\lambda^2}$ (draw). This will give a gradient of B and a y-intercept of A.

- Example 3 - radioactive decay

$$N = N_0 \exp^{-\frac{t}{\lambda}}$$

Here we would measure the number of decays, which gives us N in a given time, t in an effort to find the half-life, *lambda*. Tricky to plot exponentials by hand so we can linearise it by taking logs:

$$\log_{10} N = \log_{10} N_0 - \frac{t}{\lambda}\log_{10} e$$

We use $\log_{10}$ rather than natural logarithms because we're human and we like to think in base 10 and use base 10 log-linear graph paper. (need slide). So what is the gradient? $(-\frac{\log_{10} e}{\lambda})$
And the intercept? $(\log_{10} N_0)$

- Example 4 - power laws.

$$V = kI^{3/2}$$

Again we can take logarithms to make things easier to plot:

$$\log_{10} V = \log_{10} k + \frac{3}{2}\log_{10} I$$

3.1 How to plot on log-linear paper

x-axis is normal linear scale but for y axis bold divisions indicate factors of 10. As always - you need to look at the data you have to plot to decide where to start the axes.
eg. - if our smallest data point is 0.34 and the largest is 5465 we could choose the first point on this axis to be 0.1 and the last to be 10000 with 5 divisions: 0.1, 1, 10, 100, 1000, 10000.
(Why don't we start with zero? - $\log 0 = \inf$
Now these subdivisions divide the larger scale into 10. So taking between 0.1–1 we have 9 divisions marked: 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.
So what do we have between 100–1000?
Learning to read off a log scale - show plots of data on log and linear scale.

3.2 The Good News

PhysPlot does this for you - you can easily change the scale linear to logarithmic, you can use it to fit lines to your data.

4 Histograms

Slide showing Luccio's data
For the previous examples we had sets of data where we measured one y value for each x value - so we could easily plot one against the other. What happens if we just measure the same thing a number of times. For example - I could measure the height of everybody in this room. Then the ideal way to display the data would be in a histogram.
The trick in plotting histograms is to choose suitable binning - ie the scale on the x-axis. The aim is to show a smooth distribution/structure.

- If the bins are too wide we lose information
- if the bins are too narrow you don't see the trend of the data because you are too sensitive to statistical fluctuations.

4.1 Normalisation

To normalise a histogram you divide the contents of each bin by the total number of measurements, N. For example, if I took 500 measurements and summed the contents over all the bins in the histogram I would really expect:

$$\sum_{i=1}^{nbins} n_i = 500$$

otherwise something has gone wrong. So I divide every bin by 500 and get:

$$\frac{1}{N} \sum_{i=1}^{nbins} n_i = \frac{500}{500} = 1$$

Slide - A good example of plotting data - discuss

- Title
- Axis labels
- Units
- Axis values
- Sensible scale
- Legend (if more than one set of data shown)
- Error bars