Additional notes for PHYS2020

Michael Ashley—June 2004

1 Numerical integration

1.1 Midpoint method

Please see the lecture notes.

1.2 4th order Runge-Kutta

Please see the lecture notes.

2 Statistical analysis of data

2.1 Moments of distributions

Consider a number of independent measurements of a physical quantity x. Let the individual estimates be $x_0, \ldots, x_{n-1}$. Then, we can define various *moments* of the distribution. For example:

2.1.1 The zeroeth moment

The zeroeth moment is simply n, the number of samples.

2.1.2 The first moment—mean

The first moment is the *mean*, and can be calculated as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n-1} x_i$$

2.1.3 The second moment—variance

The second moment is the *variance*:

$$\text{var} = \frac{1}{n-1}\sum_{i=0}^{n-1}(x_i - \bar{x})^2$$

The variance is a measure of how closely the data is clustered around the mean. The term $\frac{1}{n-1}$ should be replaced by $\frac{1}{n}$ if the value of the mean is known in advance rather than being determined from the data $x_0, \ldots, x_{n-1}$. This makes sense if you think about how the value of the mean moves as n is increased.

Closely related to the variance is the *standard deviation* which is simply

$$\sigma = \sqrt{\text{var}}$$

for a Gaussian distribution, 68% of the measurements will lie within $\pm\sigma$ of the mean; 95% will lie within $\pm 2\sigma$ of the mean; 99.7% will lie within $\pm 3\sigma$ of the mean. However, in real physics situations we often have glitches that cause data values to lie in statistically unlikely places.

Calculating the variance has some subtleties associated with it. **Don't be tempted to make the following simplification:**

$$\text{var} = \frac{1}{n-1}\sum_{i=0}^{n-1}(x_i - \bar{x})^2 = \frac{1}{n-1}\left\{\left(\sum_{i=0}^{n-1} x_i^2\right) - n\bar{x}^2\right\}$$

while this leads to some computational saving (since the $\bar{x}$ and $\sum x_i^2$ can be calculated using a single pass through the data), it can lead to significant rounding errors if the variance is much smaller than the mean. It is much better to use a two-pass algorithm, calculating $\bar{x}$ first, and then calculating the variance using the definition given earlier. A further refinement is to use the following algorithm:

$$\text{var} = \frac{1}{n-1}\left\{\sum_{i=0}^{n-1}(x_i - \bar{x})^2 - \frac{1}{n}\left[\sum_{i=0}^{n-1}(x_i - \bar{x})\right]^2\right\}$$

which on the face of it looks strange, since the second term is mathematically zero. However, this term is not zero when only a limited number of digits of precision are being used in the calculations, and, in fact, the term does a good job of correcting for rounding errors in the first term.

2.1.4 The third moment—skew

The third moment is the *skew*:

$$\text{skew} = \frac{1}{n}\sum_{i=0}^{n-1}\left[\frac{x_i - \bar{x}}{\sigma}\right]^3$$

The skew measures the degree of symmetry of the distribution around the mean. If the distribution has a long tail above the mean, the skew is positive, else it is negative. A symmetrical distribution (such as a Gaussian) has a zero skew. Note that since we divide by σ in calculating the skew, the resultant number is dimensionless (unlike the mean and variance).

2.1.5 The fourth moment—kurtosis

The fourth moment is the *kurtosis*:

$$\text{kurt} = -3 + \frac{1}{n}\sum_{i=0}^{n-1}\left[\frac{x_i - \bar{x}}{\sigma}\right]^4$$

The kurtosis measures how "pointy/flat" the distribution is. The mysterious -3 in the definition ensures that a Gaussian has a zero kurtosis. A positive kurtosis means that the distribution rises to a point more rapidly than a Gaussian. A negative kurtosis means that the distribution is flatter than a Gaussian. Like skew, kurtosis is dimensionless.

2.2 Using the moments in practice

The zeroeth through second moments are generally useful (and, in fact, are all that is necessary to completely specify a Gaussian distribution). The higher order moments, such as skew and kurtosis, are less likely to be meaningful, and should be approached with caution.

2.3 Robust estimators

Real-life data often has *outliers*, i.e., points that are anomalously far (many standard deviations) from the mean. This can happen, e.g., in astronomical images

when cosmic rays hit the CCD, thereby generating a spike of noise that is not truly representative of the underlying distribution. It can also result from defects in the instrumentation used to obtain the data, e.g., a high-order bit in an analog-to-digital converter might occasionally flip on, adding 1024 (or some power of two) to the normal data value.

Outliers can have a large effect on measures such as the variance, which are sensitive to the square of the distance from the mean. The mean itself can be pulled away from its true value. There are various techniques for coping with outliers, some of which are detailed below.

2.3.1 Sigma-clipping

The idea of sigma-clipping is to do a first pass through the data, calculating the mean and standard-deviation, and then do a second pass where data points that are more than a certain number of standard deviations away from the mean are ignored. Additional passes can be performed. The motivating idea behind sigma-clipping is to prevent outlying points from affecting the statistical moments.

The high and low clipping thresholds need not be the same (which is useful, e.g., with cosmic ray spikes on a CCD image, which are always positive). Care needs to be taken that the first estimate of the mean hasn't been two wildly affected by outliers to allow the subsequent passes to recover. Don't be tempted to choose a very low threshold (like one standard deviation) since this will remove many valid data samples.

2.3.2 Average deviation

The *average deviation* can be calculated from

$$\mathrm{adev} = \frac{1}{n}\sum_{i=0}^{n-1} |x_i - \bar{x}|$$

and is less affected by outliers than the variance.

2.3.3 Median

The *median* is the data value at which half the samples lie below, and half above. If the number of samples, n, is odd, the median will be the middle sample when

the samples are sorted numerically. If n is even, the median is the average of the two middle values.

Note that the median is entirely insensitive to outliers, e.g., if there is a data point that is 100σ away from the mean, it will not affect the median more than if the point was, say, 3σ away from the mean. This is why the median is regarded as a robust estimator.

The simplest way of calculating the median is to sort the entire array of data samples into ascending order, and then find the middle value(s). However, this is computationally wasteful, and there are more efficient algorithms that you should explore if the time taken for this calculation is of importance to you.

2.3.4 Percentile points

A generalisation of the concept of the median is to define, e.g., the 5% and 95% points in the distribution. These are, respectively, the points at which 5% and 95% of the distribution lie below them.

3 Modelling data

Very often in physics there is a need to compare measurements with a model of some sort. The measurements will all have associated uncertainties (we usually assume that the samples come from a Gaussian distribution, with a well-defined variance). The model will depend on various parameters. We wish to determine the model parameters from our observations, to estimate the errors in the parameters, and to determine whether the model is a good fit to the data.

3.1 Chi-square minimisation

Suppose we have an observed quantity y, which is a function of some input quantity x. Further assume that x can be set precisely, and that the resulting measurements of y follow a Gaussian distribution with standard deviation σ. Further, suppose we have a model function $y(x; p_0 \dots p_{m-1})$ which attempts to predict y given x using a model with m parameters, $p_0 \dots p_{m-1}$. Then, we can define a quantity called "chi-square" as follows:

$$\chi^2 = \sum_{i=0}^{n-1} \left(\frac{y_i - y(x_i; p_0 \dots p_{m-1})}{\sigma_i} \right)^2$$

It is clear that χ^2 will be zero if the model fits perfectly. Our aim is to choose values for the parameters $p_0 \dots p_{m-1}$ such that χ^2 is minimised (note that it is never negative).

3.2 Least-square minimisation

For the simple case where the σ for each y_i are the same, minimising χ^2 is equivalent to minimising

$$\sum_{i=0}^{n-1} (y_i - y(x_i; p_0 \dots p_{m-1}))^2$$

which is just the sum of the squares of the differences between the model and the measurements. Hence, *least-square minimisation* is a special case of chi-square minimisation.

3.3 Fitting a straight line to data

One of the simplest models is a straight line fit

$$y = y(x; a, b) = a + bx$$

Using chi-square minimisation, the solution can be derived analytically to be

$$a = \frac{S_{xx}S_y - S_x S_{xy}}{SS_{xx} - (S_x)^2}$$
$$b = \frac{SS_{xy} - S_x S_y}{SS_{xx} - (S_x)^2}$$

where

$$S = \sum_{i=0}^{n-1} \frac{1}{\sigma_i^2}; \qquad S_x = \sum_{i=0}^{n-1} \frac{x_i}{\sigma_i^2}; \qquad S_y = \sum_{i=0}^{n-1} \frac{y_i}{\sigma_i^2}$$
$$S_{xx} = \sum_{i=0}^{n-1} \frac{x_i^2}{\sigma_i^2}; \qquad S_{xy} = \sum_{i=0}^{n-1} \frac{x_i y_i}{\sigma_i^2}$$

and we can also derive the variances in our estimates of a and b:

$$\sigma_a^2 = \frac{S_{xx}}{SS_{xx} - (S_x)^2}$$
$$\sigma_b^2 = \frac{S}{SS_{xx} - (S_x)^2}$$

and if we are really clever we can estimate whether the data is well-fitted by a straight line. We do this by calculating

$$Q = \mathrm{gammaq}\left(\frac{N-2}{2}, \frac{\chi^2}{2}\right)$$

where gammaq is the *incomplete gamma function*. If Q is between about 0.1 and 1, then the data is well-fit by a straight line. If Q is less than about 0.001, then the line is a poor fit.

NOTE: the above formulae are not ideal for numerical computation, since they are susceptable to round-off errors due to subtracting quantities that can have almost the same value. Rearranging the formulae, and using a two-pass algorithm as we did for calculating the variance, can improve the situation. See *Numerical Recipies in C* for details.

3.4 Fitting polynomials

You need at least as many data points as you have coefficients in the polynomials. E.g., a second-order polynomial (a quadratic) has three coefficients, and will give an exact fit to any three (x, y) pairs.

You should choose the smallest number of coefficients that reasonably fit the data. Using more coefficients is likely to result in oscillation and instabilities in regions where the model is not well constrained. This is particularly problematic when trying to extrapolate beyond the available data.

3.5 Fitting arbitrary functions