Second Year Statistics of Measurement - Lecture 7
The chi-squared estimation method

Mark Richards, 22 Oct 2018

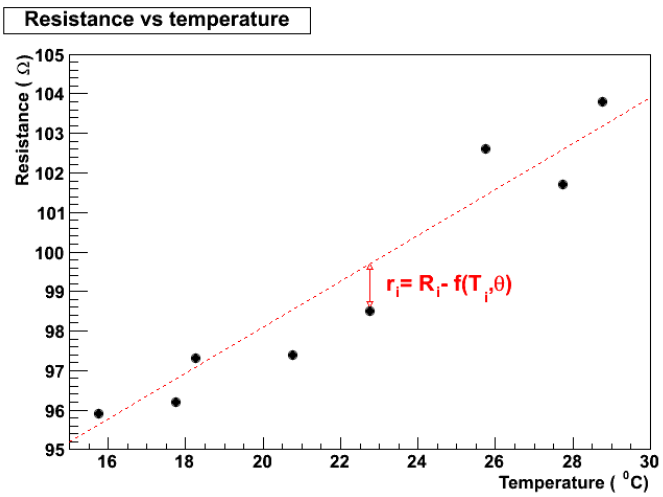
1 Introduction

We saw the maximum likelihood method in the previous lecture. Today, we will look at the chi-squared method, which is applicable in the case of measurements made from an underlying Gaussian distribution. The chi-squared is closely related to the least squares method, so we will start with that.

2 Least squares

The chi-squared method is a more powerful generalisation of the method of least squares. You will have met the method of least squares last year and probably will have used code able to calculate results using this method.

The most common application of least squares is when we have a set of measured data values y_i which were taken while another variable x_i was varied. The idea is that the true values of y depend on x according to some functional form $y = f(x;\theta_1,\theta_2,\ldots) = f(x;\theta_j)$ and that we want to find the parameters θ_j of this function, where j runs over the number of parameters. One example would be to estimate the linear temperature dependence of a resistor. For this, the function we would use is a straight line $f = \theta_1 + x\theta_2$ and so there are two parameters. Measurements of the resistance are taken for various temperatures and the results of the resistance as a function of temperature are plotted below.



To do this estimation, we use the "residual" r which is the difference of each measured value from the function value

$$r_i = y_i - f(x_i;\theta_1,\theta_2,\ldots) = y_i - f(x_i;\theta_j)$$

We then take the sum of the squares of the residuals

$$S(\theta_j;y_i) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} [y_i - f(x_i;\theta_j)]^2$$

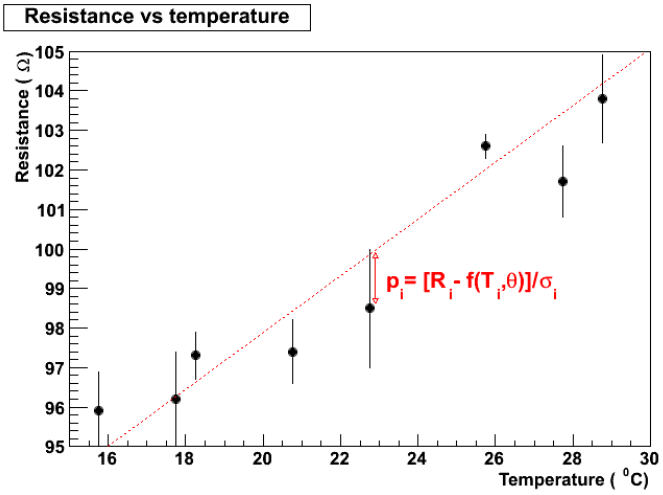
The value of S is considered as a function of the parameters θ_j given that we have observed data values y_i. Note, the x_i are not observed values; these we take as known values with no (or negligible) uncertainties. We then adjust the parameters so as to minimise this sum; mathematically we need to solve the simultaneous equations

$$\frac{\partial S}{\partial \theta_j} = 0$$

where there is one equation for each parameter θ_j. This gives us the estimates $\hat{\theta}_j$. Clearly this minimisation will tend to make the function agree with the measured values as far as possible, so that the differences, i.e. the residuals, are small. Hence, it is intuitive that this will give a reasonable method for estimating the θ_j.

3 Chi-squared method

The chi-squared method is a more rigorous approach than least squares but is fundamentally similar. The difference is that it incorporates the uncertainties on the measurements y_i, here denoted by σ_i, which can be different for each point. Clearly, measurements with large uncertainties should have less power than those which are more accurate and have small uncertainties. Note, this of course requires that the uncertainties are already known.



The chi-squared method starts by dividing the residual by the uncertainty to form a quantity called the "pull" p_i

$$p_i = \frac{r_i}{\sigma_i} = \frac{y_i - f(x_i;\theta_j)}{\sigma_i}$$

If all the data points are independent, the sum of the squares of the pulls is then called the chi-squared

$$\chi^2(\theta_j; y_i) = \sum_{i=1}^{N} p_i^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;\theta_j)}{\sigma_i}\right]^2$$

and we again want the minimum, i.e. to solve

$$\frac{\partial \chi^2}{\partial \theta_j} = 0$$

Although it is not obvious yet, it turns out that dividing the residuals by σ_i (as opposed to σ_i^2 or some other function) gives a good estimation method if the y_i have Gaussian distributions and this is where this assumption is implicitly needed.

Note, if all the uncertainties on the points are the same, i.e. $\sigma_i = \sigma$, then the chi-squared becomes

$$\chi^2(\theta_j; y_i) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; \theta_j)}{\sigma} \right]^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} [y_i - f(x_i; \theta_j)]^2 = \frac{S}{\sigma^2}$$

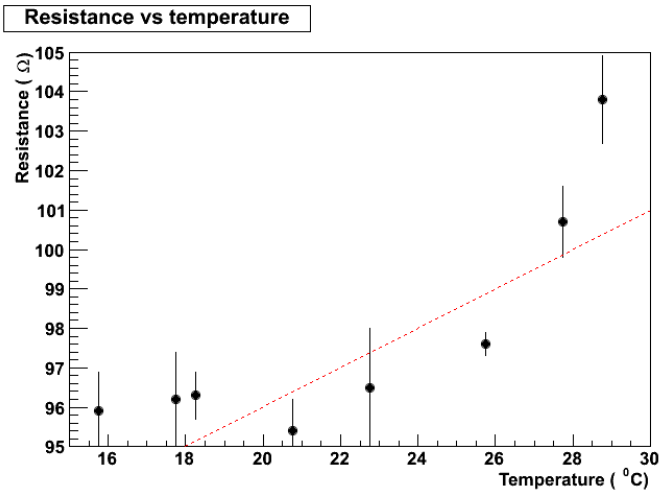
Hence the chi-squared is equal to the squares value S multiplied by a constant factor. This means the values of the parameters θ_j which minimise the chi-squared will be identical to those resulting from the least squares method. Hence, the estimates of the parameters from the two methods will be the same in this particular case.

Solving for the minimum can be done analytically for simple functions, but in most cases must be done numerically on a computer. Such a process of finding the values of parameters which minimise (or maximise) a function is generally called 'fitting' the parameters. Both the chi-squared and the maximum likelihood methods are therefore examples of fitting the parameters.

4 Goodness of fit

Besides allowing for differing sizes of uncertainties on the y_i, the chi-squared method is more useful even if all the σ_i have the same value (when it gives identical estimates for the θ_j as the simple least squares method). This is because the minimum chi-squared value itself gives further information, specifically on what is called the "goodness of fit". This tells us if the function we have assumed is a sensible one or not.

It is clear that if the function *is* correct, then the values of the y_i and the function $f(x_i; \theta_j)$ for each x_i should not be too different. Conversely, a bad fit will have the points far from the function; see an example below.



Specifically, for a good fit, we would expect the values of the y_i to differ from the fitted function by amounts similar to the σ_i as they are (assumed to be) Gaussian distributed. Specifically, if the fitted function is sensible, it should be close to the mean for that x_i and hence the residual (i.e. the fluctuation around that mean) would be expected to have a standard deviation around the fitted function of roughly σ_i. Hence, each term in the chi-squared sum would be expected to be $(r_i/\sigma_i)^2 \sim 1$ and hence the chi-squared at the minimum should have a value of

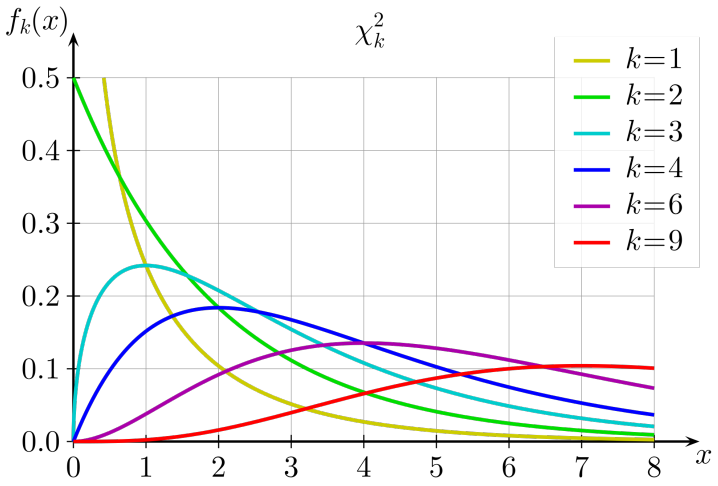
$\chi^2_{\text{min}} \sim N_{\text{data}}$. A much bigger value would imply that the wrong function is being used and so can help to identify if an underlying theory is incorrect.

To be more precise, it is not actually correct to compare the chi-squared value to N_{data}, but a correction is needed. To see this, consider an extreme example; a straight line fit to two data points will always give a minimum chi-squared of exactly zero. This does not mean the straight line fit is good; this is just saying there is a solution, not a fit, if the number of data points is equal to the number of parameters. It turns out the critical measure is the "number of degrees of freedom" N_{dof}, which is defined to be the difference between the number of data points being used in the fit and the number of parameters being estimated in the fit

$$N_{\text{dof}} = N_{\text{data}} - N_{\text{paras}}$$

If there are more data points than parameters, i.e. $N_{\text{dof}} > 0$, then there are more constraints than there are parameters to be determined and a fit is needed. The χ^2_{min} value then gives information on the goodness of fit. If the number of data points is equal to the number of parameters, i.e. $N_{\text{dof}} = 0$, then the equations can be solved rather than fitted, and the chi-squared will be identically zero so we have no information on the goodness of fit. Finally, if the number of data points is less than the number of parameters, i.e. $N_{\text{dof}} < 0$, the the system is underconstrained and no unique determination of the parameters is possible. In terms of the minimum value of the chi-squared, it turns out that for $N_{\text{dof}} > 0$ we would expect $\chi^2_{\text{min}} \sim N_{\text{dof}}$ for a good fit, while values much bigger than this would imply the function being fitted does not actually describe the data well.

The chi-squared is yet another example of a statistic and hence is a random variable with a PDF. At a basic level, given that all the measurement distributions are Gaussian, then we can calculate this PDF for the chi-squared value. It turns out this only depends on the number of degrees of freedom, not the number of data points or parameters separately, i.e. $\rho(\chi^2; N_{\text{dof}})$. The PDF has an expectation value of N_{dof} and has a long tail out to infinity, as shown in the plot below for various N_{dof} (taken from Wikipedia).

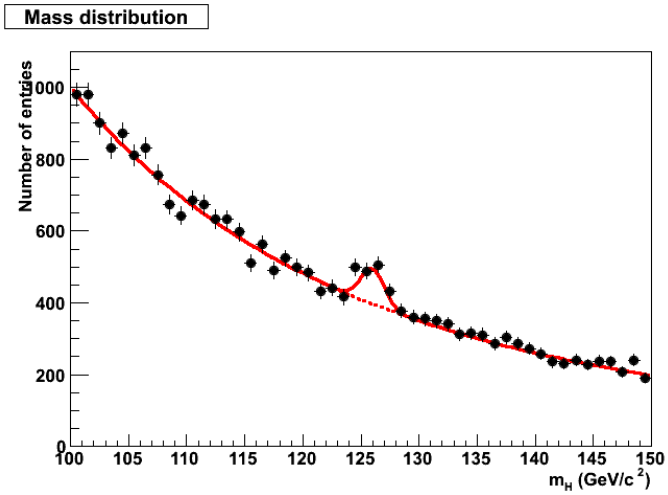


The goodness of fit is therefore effectively a statement about whether our fit gives us a chi-squared value near the PDF peak or in the tail at higher chi-squared. This is often quantified in terms of a "p-value" which gives the probability of getting the observed or a worse chi-squared value. There are tables of p-values in statistics books and on the web, and software packages may have functions to calculate the p-value from a given chi-squared and number of degrees of freedom. However, as the p-value corresponds to integrating the chi-squared distribution from the observed value up to infinity, then it corresponds exactly to what we previosuly called the

one-sided significance level, set by the data value. Hence, stating that the goodness of fit is poor is precisely a hypothesis test, i.e. we say we reject the hypothesis that the data agree with the function chosen for the fit if the goodness of fit is bad.

5 Chi-squared method for binned distributions

A very useful approximation can be made to allow us to use a chi-squared fit in a very common situation. We are often trying to find the parameters of a function which describes the shape of some data distribution. By histogramming the data into small ranges ("bins"), we can count the number of times the variable falls into each bin. An example is the Higgs; this was discovered in 2012 by finding the mass of its decay products. There is a large background of random combinations which gave a smooth background shape, while the Higgs itself gives a narrow Gaussian peak in mass. By fitting for the background shape parameters, the background-only hypothesis could be rejected due to a poor goodness of fit; in the Higgs case the significance level was 5σ which is very strong evidence for rejection. Knowing that the background was not sufficient, including parameters for the mean (i.e. the Higgs mass) and width of the Gaussian allowed its existence to be established and its mass to be determined.



This lends itself to a chi-squared treatment if the numbers of events in each bin are large, i.e. $n_i \gg 1$ for each bin i. This is because the probability distribution for the number n_i in a single bin is (to a good approximation) a Poisson, where the mean of the Poisson is just the total amount of the function integrated over the bin. Specifically

$$\mu_i(\theta_j) = \int_{\text{Bin } i} f(x;\theta_j)\, dx$$

It is also common in practise that μ_i is approximated to

$$\mu_i(\theta_j) \approx f(x_i;\theta_j)\, \Delta x$$

where x_i is the value of x at the centre of each bin and Δx is the bin width. Given this, n_i has a distribution of $P[n_i;\mu_i(\theta_j)]$ Hence, because of the Central Limit Theorem for large n_i as seen in Lecture 4, this is approximately Gaussian, with $\sigma_i \approx \sqrt{n_i}$. Hence the chi-squared can be approximated by

$$\chi^2(\theta_j;\underline{n}) = \sum_{i=1}^{N} \frac{[n_i - \mu_i(\theta_j)]^2}{n_i}$$

where the sum is over all the bins. This method is used very widely but it is important it is only done when all bins have a large number of entries, i.e. $n_i \gg 1$, which is typically at least 20 entries.

6 Connection to maximum likelihood estimation

Let's consider the case of a set of measurements y_i of Gaussian distributed random variables, each with a different mean μ_i and width σ_i. We will assume the means depend on some parameters θ_j so $\mu_i(\theta_j)$, but that the widths are known. The likelihood for one such measurement is

$$L_i = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(y_i-\mu_i)^2/2\sigma_i^2}$$

so the log-likelihood is

$$\ln(L_i) = -\frac{(y_i-\mu_i)^2}{2\sigma_i^2} - \ln(\sigma_i\sqrt{2\pi})$$

Hence, for all the measurements, the total log-likelihood is

$$\ln(L) = \sum_i \ln(L_i) = -\sum_i \frac{(y_i-\mu_i)^2}{2\sigma_i^2} - \sum_i \ln(\sigma_i\sqrt{2\pi}) = -\sum_i \frac{(y_i-\mu_i)^2}{2\sigma_i^2} + C$$

where C is a term which does not depend on the μ_i and hence also not on the θ_j parameters. This means

$$2C - 2\ln(L) = \sum_i \frac{(y_i-\mu_i)^2}{\sigma_i^2}$$

The term on the right can be seen to be what we defined as the chi-squared, i.e.

$$\chi^2 = 2C - 2\ln(L)$$

When we considered the chi-squared, we discussed the values as being described by some function but effectively all this does is to define a different mean for each measurement, i.e.

$$\mu_i(\theta_j) = f(x_i;\theta_j)$$

In these terms, then we would maximise the log-likelihood to find the estimates $\hat{\theta}_j$. However, since they differ by a sign, *maximising* the log-likelihood is equivalent to *minimising* the chi-squared and hence the estimates from either method will be identical. In fact, it is common to work with $-2\ln(L)$ and find a minimum rather than a maximum, as this is then directly comparable to the chi-squared in the Gaussian case.

In general however, there is no goodness-of-fit measure for a likelihood fit, unlike the chi-squared method. In the Gaussian case, the constant C above is calculable and so the chi-squared can be found from the maximum likelihood. However, for non-Gaussian cases (which are exactly when the maximum likelihood is most useful), the above form is not found and there is no such goodness-of-fit measure.

7 Non-examinable: the error matrix

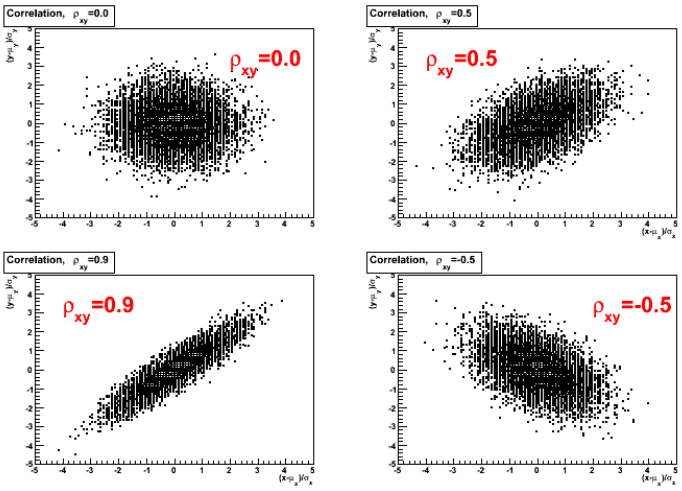
The general formulation for uncertainties with more than one variable is using the 'error matrix' $\underline{\underline{E}}$, sometimes called the 'covariance matrix'. For independent (so-called 'uncorrelated') variables,

the error matrix is diagonal and equal to

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where σ_i is the uncertainty on variable i.

This can be generalised; the zeros off-diagonal indicate that there are no correlations between the measurements. However, conversely, if the variables *are* correlated, then these off-diagonal elements are no longer zero, but have the form $\rho_{ij}\sigma_i\sigma_j$, where ρ_{ij}, the 'correlation coefficient', must be in the range $-1 \leq \rho_{ij} \leq 1$. Clearly, $\rho_{ij} = 0$ corresponds to the uncorrelated case. For non-zero ρ_{ij}, then if one variable fluctuates higher than the average, then the other will also tend to fluctuate higher (if they are correlated, i.e. $\rho_{ij} > 0$) or lower (if they are anticorrelated, i.e. $\rho_{ij} < 0$), as shown in the figure below.



Hence, the general error matrix is

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \dots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \dots \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Knowing about the error matrix allows us to write down the most general form for the propagation of errors formula. If a quantity z depends on several variables x_i, i.e. $z = z(\underline{x})$, then by forming a 'derivative vector' $\underline{d}$

$$\underline{d} = \begin{pmatrix} \partial z/\partial x_1 \\ \partial z/\partial x_2 \\ \partial z/\partial x_3 \\ \vdots \end{pmatrix}$$

the error on z is given by

$$\sigma_z^2 = \underline{d}^T \underline{\underline{E}}\, \underline{d}$$

For example, with two variables $x_1 = x$, $x_2 = y$, then generally

$$\sigma_z^2 = \begin{pmatrix} \partial z/\partial x & \partial z/\partial y \end{pmatrix} \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \begin{pmatrix} \partial z/\partial x \\ \partial z/\partial y \end{pmatrix}$$

$$
\begin{aligned}
&= \begin{pmatrix} \partial z/\partial x & \partial z/\partial y \end{pmatrix} \begin{bmatrix} (\partial z/\partial x)\sigma_x^2 + (\partial z/\partial y)\rho_{xy}\sigma_x\sigma_y \\ (\partial z/\partial x)\rho_{xy}\sigma_x\sigma_y + (\partial z/\partial y)\sigma_y^2 \end{bmatrix} \\
&= \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial z}{\partial x}\right)\left(\frac{\partial z}{\partial y}\right)\rho_{xy}\sigma_x\sigma_y
\end{aligned}
$$

This clearly reduces to the form given previously for the propagation of errors formula in the uncorrelated case, i.e. when $\rho_{xy} = 0$.

8 Non-examinable: chi-squared method for correlated uncertainties

There is an even more complete version of the chi-squared method, which occurs when the measurements y_i are correlated. The general formulation for uncertainties with more than one variable is using the error matrix $\underline{\underline{E}}$, as discussed in the previous section.

The expression for the chi-squared can be written in terms of vectors and matrices. Defining the "residual vector" $\underline{r}$ as

$$
\underline{r} = \begin{bmatrix} y_1 - f(x_1;\theta_j) \\ y_2 - f(x_2;\theta_j) \\ y_3 - f(x_3;\theta_j) \\ \vdots \end{bmatrix}
$$

so its transpose is

$$
\underline{r}^T = \begin{bmatrix} y_1 - f(x_1;\theta_j) & y_2 - f(x_2;\theta_j) & y_3 - f(x_3;\theta_j) & \dots \end{bmatrix}
$$

then the chi-squared can be written as

$$
\chi^2 = \begin{bmatrix} y_1 - f(x_1;\theta_j) & y_2 - f(x_2;\theta_j) & y_3 - f(x_3;\theta_j) & \dots \end{bmatrix} \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \dots \\ 0 & 1/\sigma_2^2 & 0 & \dots \\ 0 & 0 & 1/\sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{bmatrix} y_1 - f(x_1;\theta_j) \\ y_2 - f(x_2;\theta_j) \\ y_3 - f(x_3;\theta_j) \\ \vdots \end{bmatrix}
$$

The matrix in the middle is called the "weight matrix" $\underline{\underline{W}}$ and the chi-squared can be written compactly as

$$
\chi^2 = \underline{r}^T \underline{\underline{W}}\, \underline{r}
$$

The inverse of the weight matrix is obviously

$$
\underline{\underline{E}} = \underline{\underline{W}}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}
$$

From above, this will be recognised as the form of the error matrix for uncorrelated variables. As discussed previously, the more general error matrix, allowing for possible correlations, is

$$
\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \dots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \dots \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}
$$

Inverting this matrix to get the weight matrix then allows the chi-squared to be expressed in the same way as before

$$
\chi^2 = \underline{r}^T \underline{\underline{W}}\, \underline{r}
$$

This is the general form; even for the correlated case, then combining the residual vectors and weight matrix as above gives the chi-squared.

Second Year Statistics of Measurement - Lecture 8
Confidence intervals and parameter uncertainties

Mark Richards, 23 Oct 2018

9 Introduction

We have discussed estimating parameters in the last two lectures, using either the chi-squared or the maximum likelihood methods. However, we have not discussed how to find the uncertainty on the parameters we determine. We have also not defined precisely what we mean by an uncertainty on a parameter. This is where the real differences between frequentist and Bayesian definitions of probability arise. We shall discuss the frequentist approach here and then discuss Bayesian estimation in the next lecture.

The first thing to do is understand how to find the uncertainties on the parameters when using the two methods we have studied. In principle there is nothing new here. Both methods can be considered as giving an equation for the parameter(s) θ as a function of the data measurements x_i

$$\hat{\theta} = M_\theta(x_1, x_2, \ldots, x_N)$$

so, at least in a linear approximation, we can just apply the propagation of errors formula to get the uncertainty on each θ_i, i.e.

$$\sigma^2_{\theta_i} = \sum_i \left(\frac{\partial M_\theta}{\partial x_i}\right)^2 \sigma^2_{x_i}$$

However, in practical terms this can often be inconvenient to calculate, particularly when the maximum or minimum is only found numerically. Also, the propagation of errors equation is only itself an approximation. Let's consider each of the two methods in turn.

10 Maximum likelihood parameter uncertainties

The central limit theorem says for large N, the mean and variance of the sum of repeated measurements of any data quantity x are distributed according to a Gaussian of mean $\mu = \sum_i E_i$ and width $\sigma^2 = \sum_i V_i$. With enough measurements, many PDFs approximate to Gaussians. It turns out that the likelihood dependence on the parameters is then also Gaussian. It is therefore very common to take an approximation that the likelihood is close to a Gaussian.

For simplicity, assume there is only one parameter θ. We do a Taylor expansion of the log-likelihood in terms of θ around the maximum, which is where the parameter is $\theta = \hat{\theta}$. The Taylor expansion then gives

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] + \left.\frac{d\ln(L)}{d\theta}\right|_{\hat{\theta}} (\theta - \hat{\theta}) + \left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}} \frac{(\theta-\hat{\theta})^2}{2!} + \ldots$$

Because $\hat{\theta}$ is defined to be at the maximum, the first derivative is zero. The second derivative is negative (since it is a maximum) but is some constant when evaluated at $\hat{\theta}$; let this be

$$\frac{1}{\Sigma^2} = -\left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}}$$

where the negative sign means Σ^2 is positive and hence Σ is real. In terms of Σ, then

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] - \frac{(\theta-\hat{\theta})^2}{2\Sigma^2} + \ldots$$

If the higher order terms are now ignored, then taking exponentials gives

$$L(\theta) \approx L(\hat{\theta}) e^{-(\theta-\hat{\theta})^2/2\Sigma^2}$$

which is clearly a Gaussian function for θ. It is also seen that the quantity Σ generally is the width of the Gaussian, which is therefore the uncertainty on $\hat{\theta}$. Hence, in this approximation (or exactly if the function is truely a Gaussian), then the uncertainty is given by the equation above involving the second derivative, evaluated with $\theta = \hat{\theta}$. Warning: the above equation for $1/\Sigma^2$ is only valid for the one parameter case. For more parameters, this gives the inverse of a matrix, which has to be inverted to get the uncertainties (see the appendix).

Example: We have a large number N of measurements of an exponential random variable. We will express the exponential PDF as $e^{-x/a}/a$, where the parameter a (to be estimated) is the average, as shown in Lecture 3. The likelihood and hence log-likelihood for a single measurement is

$$L_i(a) = \frac{1}{a} e^{-x_i/a} \qquad \text{so} \qquad \ln[L_i(a)] = -\ln(a) - \frac{x_i}{a}$$

Hence, the total log-likelihood is

$$\ln[L(a)] = \sum_i \ln[L_i(a)] = -N \ln(a) - \frac{1}{a} \sum_i x_i$$

The derivative is

$$\frac{d \ln(L)}{da} = -\frac{N}{a} + \frac{1}{a^2} \sum_i x_i$$

so the estimate is given by this being zero, for which

$$\frac{N}{\hat{a}} = \frac{1}{\hat{a}^2} \sum_i x_i \qquad \text{so} \qquad \hat{a} = \sum_i x_i / N$$

as would be expected. The second derivative is

$$\frac{d^2 \ln(L)}{da^2} = \frac{N}{a^2} - \frac{2}{a^3} \sum_i x_i \qquad \text{so} \qquad \left. \frac{d^2 \ln(L)}{da^2} \right|_{\hat{a}} = -\frac{N}{\hat{a}^2}$$

and so is negative as required for a maximum. Therefore, the uncertainty on the estimate is approximately

$$\sigma_{\hat{a}} \approx \frac{1}{\sqrt{-d^2 \ln(L)/da^2|_{\hat{a}}}} = \frac{\hat{a}}{\sqrt{N}}$$

The same result would be obtained from using the propagation of errors formula on the result for $\hat{a}$, remembering that the standard deviation of an exponential random variable is a. Explicitly

$$\frac{\partial \hat{a}}{\partial x_j} = \frac{1}{N}$$

so

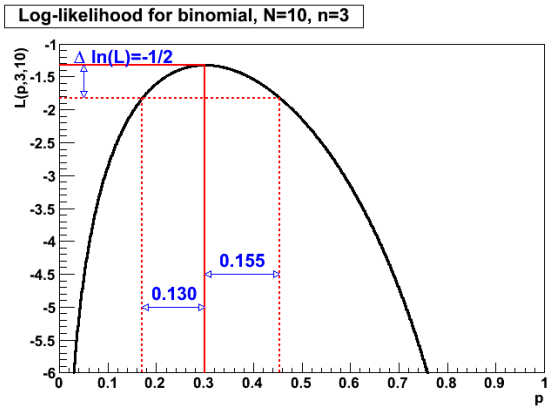
$$\sigma_{\hat{a}}^2 = \sum_j \left(\frac{\partial \hat{a}}{\partial x_j} \right)^2 \sigma_{x_j}^2 = \sum_j \frac{1}{N^2} \hat{a}^2 = \frac{\hat{a}^2}{N}$$

giving the same result for $\sigma_{\hat{a}}$.

What if the PDF is not even approximately Gaussian or/and propagation of errors is not a good approximation? The uncertainty can still be evaluated but by using a different method. For the Gaussian case, then the value of the log-likelihood at $\theta = \hat{\theta} \pm \Sigma$ is

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] - \frac{(\hat{\theta} \pm \Sigma - \hat{\theta})^2}{2\Sigma^2} = \ln[L(\hat{\theta})] - \frac{\Sigma^2}{2\Sigma^2} = \ln[L(\hat{\theta})] - \frac{1}{2}$$

Hence, the uncertainty is the range which makes the log-likelihood change by $-1/2$. It turns out this is true for non-Gaussian likelihoods too; the uncertainty can be evaluated by changing the parameter and seeing what range gives $\Delta \ln(L) \geq -1/2$. (In principle, this is only exact in the large N limit, but it is normally a very good approximation anyway.) In general, this can require a different shift in θ when going up compared to down, and so this gives asymmetric uncertainties. An example for a binomial measurement is shown below. In this case, you will see them written as e.g. $0.30^{+0.15}_{-0.13}$ for this example.



11 Chi-squared parameter uncertainties

As discussed in lecture 6, the chi-squared is a function of the parameters

$$\chi^2(\theta; y_i) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;\theta)}{\sigma_i} \right]^2$$

and the estimates of the parameters are taken to be the values that give the minimum chi-squared.

However, consider expanding the chi-squared around the minimum using a Taylor expansion in the parameters. We will again consider only one parameter for clarity. Since it is a minimum, there is no linear term and so we know immediately it must be of the form

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2} + \ldots$$

Since the best-fit chi-squared is a minimum, then the second derivative must be positive. Hence, let us express the second derivative as

$$\left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}} = \frac{2}{\Sigma^2} \qquad \text{i.e.} \qquad \frac{1}{\Sigma^2} = \frac{1}{2} \left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}}$$

such that we can write

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{\Sigma^2}$$

We can interpret this as being some sort of 'residual' of the parameters around the best estimate of $\hat{\theta}$. With this interpretation, then it is clear the uncertainty on the best estimate is Σ. It can be

shown (see appendix) that Σ is indeed the uncertainty on the parameter that would be obtained from using the propagation of errors formula, i.e.

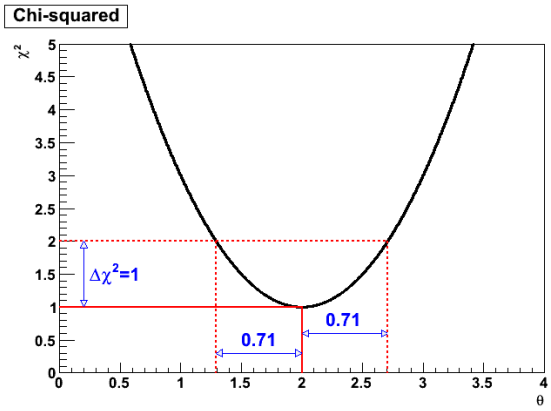
$$\Sigma^2 = \sum_i \left(\frac{\partial\theta}{\partial y_i}\right)^2 \sigma_i^2$$

so that our interpretation is indeed correct.

It is sometimes the case that the second derivative is easy to calculate and so the uncertainty can be found using the above method. However, the chi-squared method can be used for arbitrarily complicated functions $f(x_i;\theta)$ and in many cases it is impracticable to find the uncertainty this way. Luckily, there is a simple numerical method, which is to change the parameter from its best estimate and see what happens to the value of the chi-squared. If we shift the parameter θ from $\hat{\theta}$ (the value which gives the minimum chi-squared) to $\hat{\theta} \pm \Sigma$, then the chi-squared becomes

$$\chi^2(\hat{\theta} \pm \Sigma) \approx \chi^2(\hat{\theta}) + \frac{\Sigma^2}{\Sigma^2} \approx \chi^2(\hat{\theta}) + 1$$

Hence, if we numerically evaluate the chi-squared for various different values of θ, then the range for which the chi-squared changes by up to one unit gives the uncertainty on $\hat{\theta}$. An example is shown below, where the result would be quoted as $\hat{\theta} = 2.0 \pm 0.7$.



We showed in the last lecture that for the Gaussian case, the chi-squared is related to the log-likelihood by $\chi^2 \sim -2\ln(L)$. Hence, changing the chi-squared by $+1$ is equivalent to changing the log-likelihood by $-1/2$ and so these two uncertainty estimates are in agreement.

12 Confidence intervals for random variables

Back in lecture 1, we discussed the amount of the probability contained within certain ranges of the Gaussian distribution, e.g. the region $\pm 1\sigma$ of the mean contained 68.3% of the integral

$$\int_{\mu-\sigma}^{\mu+\sigma} G(x;\mu,\sigma)\,dx = \frac{2}{\sqrt{\pi}} \int_0^{1/\sqrt{2}} e^{-y^2}\,dy = \mathrm{erf}(1/\sqrt{2}) = 0.683$$

The same could be done for other ranges; e.g. the probability within $\pm 2\sigma$ is

$$\int_{\mu-2\sigma}^{\mu+2\sigma} G(x;\mu,\sigma)\,dx = \frac{2}{\sqrt{\pi}} \int_0^{2/\sqrt{2}} e^{-y^2}\,dy = \mathrm{erf}(2/\sqrt{2}) = 0.954$$

Considering the $\pm 1\sigma$ case, this means that a random measurement of x arising from a Gaussian distribution has a probability of 68.3% of lying within this range. This is referred to as the

'confidence interval', i.e. we have 68.3% confidence that the value will lie in the interval within $\pm 1\sigma$ of the mean. Hence, if we take many such measurements, then 68.3% of them (which is roughly 2/3) will be within $\pm 1\sigma$ of the mean.

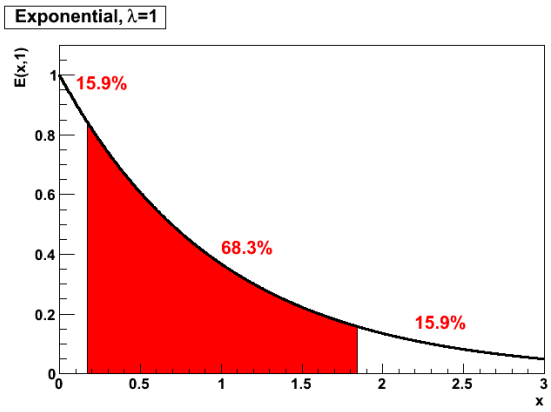
Ranges can also be defined for non-Gaussian distributions, of course. Even though the specific value of 68.3% arises from the integration of the Gaussian, it is still very common to use a range containing this amount of the integral, even for non-Gaussian distributions. It is even common to refer to this 68.3% confidence range as being the 'one sigma' range, even though there may be no such parameter σ in the distribution. However, even in trying to calculate the 68.3% range, there is complication, namely non-Gaussian distributions can be asymmetric. In this case, there is no unique answer as to how to define this range. We need

$$\int_a^b \rho(x)\, dx = 0.683$$

but this only gives one constraint on the two limits a and b. For example, for an exponential

$$\int_a^b \lambda e^{-\lambda x}\, dx = [-e^{-\lambda x}]_a^b = e^{-\lambda a} - e^{-\lambda b} = 0.683$$

This is a choice but a common way to handle this is to say that the amount outside the range, which is $100\% - 68.3\% = 31.7\%$, has to be equally divided above and below the range, so there is 15.9% on either side. This is shown for an exponential below.



13 Confidence intervals for estimated parameters

The meaning of the above is unambiguous since the quantity being considered is a random variable. This is fundamentally what we mean by the uncertainty on a random variable. This is uncontentious and not affected by frequentist or Bayesian definitions. However, we have seen that we use the results of experiments to determine parameters. These are *not* random variables but have a exact value, even if this is unknown to us. E.g. the electron mass is a fixed value, even if we are trying to measure it in an experiment. In the frequentist approach, we should look at the average of the values for a very large number of experiments, but each experiment must have the same initial conditions and this would include the electron mass being at its fixed value. Hence, there is no spread and no frequentist interpretation of what we would mean by saying a parameter has a probability of lying within a confidence interval of 68.3%.

However, the earlier part of the lecture discussed how to find the uncertainties on parameters, e.g. by finding $\Delta \ln(L) = -1/2$ or $\Delta\chi^2 = +1$, so what do these really mean? What we are in

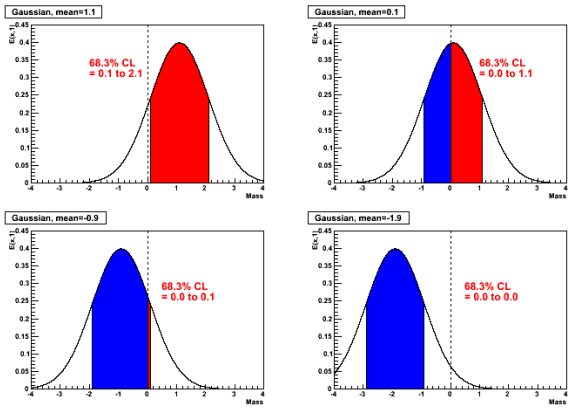
fact doing is giving a range. Clearly, the true value is either within this range or is not, and so any particular experiment may be right or wrong about the true value being within the range. However, the range chosen is such that 68.3% of experiments will contain the true value within the range on average. This is the meaning of a confidence interval for a parameter in the frequentist interpretation and this is what we actually mean when we quote the 'uncertainty' on a parameter. This is subtlely different from that for a random variable and can take a while to understand. In practise, we handle the parameter errors in the same way as random variable errors, e.g. using the propagation of errors formula, but there is an underlying conceptual difference.

Furthermore, sometimes we are only interested in one side of the confidence interval, e.g. if we are trying to determine a rate of a very rare process, we know a rate cannot be negative so we only want to set an upper limit. In this case, it is common to use the 90% or, more usually, 95% confidence level, but only on one side of the range. Hence, just like hypothesis testing, we can set one-sided or two-sided confidence intervals.

14 Physical constraints

One of the problems with frequentist methods is that we can end up with results outside the physical region. For example, we want to measure the mass of some powder of around 1 g using a set of weighing scales with 1% accuracy. The dish containing the powder is known to weigh 100 g so the uncertainty on the weight is 1 g. If we happened to measure 101.1 g then in the frequentist approach, we would set the 68.3% confidence interval for the mass of the powder to be 0.1 to 2.1 g, which is fine. However, it is not unlikely that the measurement could give something like 100.1 g, which would give a range from -0.9 to 1.1 g. However, the mass cannot be negative so the 68.3% confidence interval would have to be quoted as 0.0 to 1.1 g, which looks like it is more accurate, even though it uses the same apparatus. A bigger problem arises if the measurement happens to yield 99.1 g; the range would be -1.9 to 0.1 g so the apparent confidence interval would be 0.0 to 0.1 g. However, saying the mass is below 0.1 g when the scales are only accurate to 1 g is clearly not right. A final example might be a fluctuation where the measurement yields 98.1 g; the range would be -2.9 to -0.9 g which means there is no 68.3% confidence interval in the physical region. These cases are illustrated in the plots below.

We know these low measurements must have been 'unlucky' in some sense. Hence, we would expect them to be more likely to be some of the 31.7% of confidence intervals which do *not* contain the true value. However, it is very hard to quantify this in a consistent way in the frequentist approach. We will see how the Bayesian approach handles this in the next lecture.



15 Non-examinable: Appendix

15.1 Parameter uncertainty from chi-squared

Consider a simple chi-squared case with one parameter θ, where the function is linear in the parameter (although it can be arbitrarily complicated in x), so

$$f(x;\theta) = a(x) + b(x)\theta$$

for any functions $a(x)$ and $b(x)$. Hence,

$$\chi^2(\underline{\theta};\underline{y}) = \sum_{i=1}^{N}\left[\frac{y_i - f(x_i;\theta)}{\sigma_i}\right]^2 = \sum_{i=1}^{N}\left[\frac{y_i - a(x_i) - b(x_i)\theta}{\sigma_i}\right]^2$$

and the estimate of the parameter is given by solving

$$\left.\frac{d\chi^2}{d\theta}\right|_{\hat{\theta}} = 0 = -2\sum_{i=1}^{N}\frac{[y_i - a(x_i) - b(x_i)\hat{\theta}]b(x_i)}{\sigma_i^2}$$

This gives

$$\sum_{i=1}^{N}\frac{[y_i - a(x_i)]b(x_i)}{\sigma_i^2} = \hat{\theta}\sum_{j=1}^{N}\frac{b(x_j)^2}{\sigma_j^2}$$

so

$$\hat{\theta} = \frac{\sum_{i=1}^{N}[y_i - a(x_i)]b(x_i)/\sigma_i^2}{\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2}$$

Note, the σ_i^2 terms do *not* cancel out here as they are inside separate summations.

The uncertainty on $\hat{\theta}$ arises as the y_i have uncertainties σ_i. Hence, we can use propagation of errors to find this uncertainty. From above

$$\frac{\partial\hat{\theta}}{\partial y_i} = \frac{b(x_i)/\sigma_i^2}{\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2}$$

so that by propagation of errors

$$\sigma_{\hat{\theta}}^2 = \sum_{i=1}^{N}\left(\frac{\partial\hat{\theta}}{\partial y_i}\right)^2\sigma_i^2 = \sum_{i=1}^{N}\frac{b(x_i)^2/\sigma_i^4}{\left(\sum_{i=j}^{N} b(x_j)^2/\sigma_j^2\right)^2}\sigma_i^2 = \frac{\sum_{i=1}^{N} b(x_i)^2/\sigma_i^2}{\left(\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2\right)^2} = \frac{1}{\sum_{i=1}^{N} b(x_i)^2/\sigma_i^2}$$

However, it is claimed in the main part of the lecture that this is equal to the uncertainty Σ given by the second derivative

$$\frac{1}{\Sigma^2} = \frac{1}{2}\left.\frac{d^2\chi^2}{d\theta^2}\right|_{\hat{\theta}}$$

The second derivative is

$$\frac{d^2\chi^2}{d\theta^2} = 2\sum_{i=1}^{N}\frac{b(x_i)^2}{\sigma_i^2} = \frac{2}{\sigma_{\hat{\theta}}^2}$$

and so $\Sigma = \sigma_{\hat{\theta}}$ as claimed. Effectively any function can be approximated to being linear in its parameter by doing a Taylor expansion around the estimate, so the above holds approximately in general, with the approximation being good for small uncertainties.

This means that for the chi-squared, and for the maximum likelihood in the Gaussian approximation, the uncertainty on a single parameter is given by either

$$\frac{1}{\sigma_{\hat{\theta}}^2} = \frac{1}{2}\left.\frac{d^2\chi^2}{d\theta^2}\right|_{\hat{\theta}} \qquad \text{or} \qquad \frac{1}{\sigma_{\hat{\theta}}^2} = -\left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}}$$

respectively.

15.2 Error matrix for more than one parameter

In the Gaussian approximation for more than one parameter, then we have to use the error matrix $\underline{\underline{E}}$, or more specifically the weight matrix $\underline{\underline{W}} = \underline{\underline{E}}^{-1}$ mentioned in the notes for Lecture 7. The generalisation of the above expressions is in terms of the elements of the weight matrix W_{ij}, where

$$W_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}} \qquad \text{or} \qquad W_{ij} = - \left. \frac{d^2 \ln(L)}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}}$$

Second Year Statistics of Measurement - Lecture 9
Bayesian estimation

Mark Richards, 25 Oct 2018

16 Introduction

We saw in the previous lecture that a frequentist treatment of parameter uncertainties interprets them differently from random variables. As we will see in this lecture, the Bayesian approach allows us to handle these two cases in the same fashion, but at a cost of introducing a subjective element.

17 Bayesian probabilities

As mentioned briefly in Lecture 1, the Bayesian concept of a probability is subjective and depends on the available information. There is no need in the Bayesian framework to define multiple experiments which are repeatable. Hence, the Bayesian concept can be applied more widely.

In particular, we can then consider the parameters of our experiments as probabilistic variables. We are not saying e.g. that the true electron mass is really smeared out; what we mean is that the probability we assign to its possible values reflects the degree of our (lack of) knowledge. While this seems a minor change of philosophy, it means we can apply all the probability theory we have already learned for random variables directly to the parameter also. In particular, once we have the parameter probability distribution, we can find our 68.3% confidence interval (and hence uncertainties) just as we did for random variables.

In particular, we can apply Bayes' theorem of conditional probabilities to the parameters as well as the measurements. We saw Bayes' theorem in Lecture 1

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

where $P(x|y)$ is the probability for x given the value of y. Now, if we say x corresponds to the results from the experiment (the data) we have performed and y corresponds to the parameter θ, then this reads

$$P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

This allows us to calculate the conditional probability of the parameters given the data we have measured (the left-hand side) from the conditional probability of the data given the parameters (the numerator on the right-hand side), which we know from our study of probability distributions. In fact, this term is just equivalent to the likelihood, which we met in Lecture 6.

There are of course two other terms on the right-hand side. $P(\theta)$ is the probability for the parameter, independent of the outcome of the experiment; this means before the experiment is done. (This is sometimes difficult to evaluate and this issue is discussed further below.) Finally, the denominator is the probability of seeing the data we measured, independent of any particular values of the parameters. This is sometimes called the 'evidence' but it effectively acts as a normalisation constant, i.e. for a discrete parameter

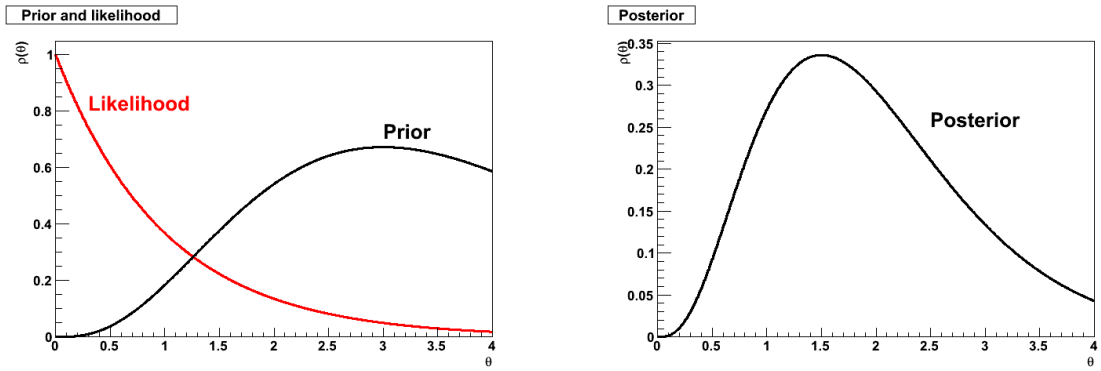
$$P(\text{Data}) = \sum_i P(\text{Data}, \theta_i) = \sum_i P(\text{Data}|\theta_i)P(\theta_i)$$

and for a continuous parameter

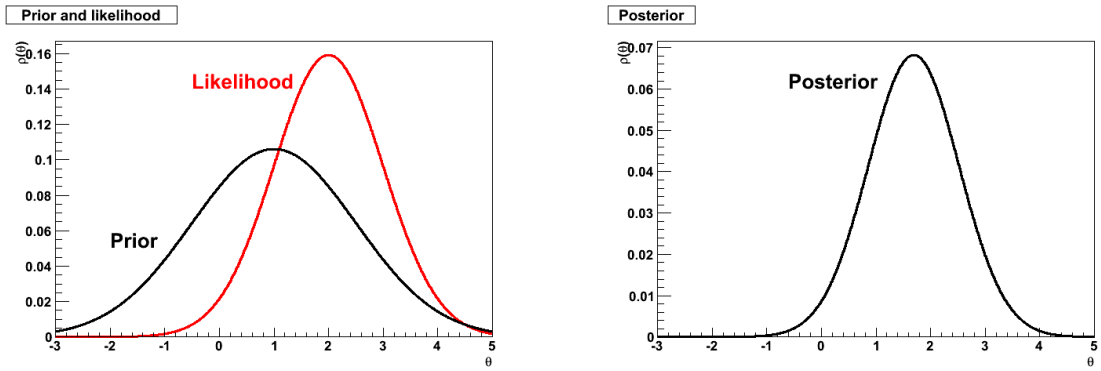
$$P(\text{Data}) = \int P(\text{Data}, \theta)\, d\theta = \int P(\text{Data}|\theta) P(\theta)\, d\theta$$

so it is simply the sum or integral of the numerator, which ensures that $P(\theta|\text{Data})$, the resulting probability distribution for θ, is normalised correctly. This property holds generally and is called the 'marginalisation rule'.

You can think of the Bayesian equation as taking our existing knowledge of the parameter probability distribution $P(\theta)$, and updating it with an experiment to give a new probability distribution $P(\theta|\text{Data})$. The probability $P(\theta)$ therefore expresses our knowledge of the parameter before the experiment and hence is often called the 'prior', while the resulting probability $P(\theta|\text{Data})$ is called the 'posterior' and expresses our knowledge after the experiment. The likelihood can be thought of as updating our knowledge from the prior to give the posterior.



A common case (due to the CLT) is that a previous measurement often results in a Gaussian prior. A further Gaussian measurement then results in a Gaussian likelihood. It is not obvious, but two Gaussian multiplied together result in a further Gaussian, so the posterior is also Gaussian distributed.



The posterior is the probability distribution for the parameter after having seen the results of the experiment. If we did a second experiment later, then the posterior of the first experiment encapsulates the state of our knowledge about the parameter before the second experiment. Hence the first posterior becomes the second prior. This can be repeated many times.

Let's see an example. Your dodgy flatmate asking you to bet on heads or tails but insists you use his coin and that he will bet heads. You suspect he will cheat and he has a double-headed coin. You play the game to determine if this is true or not (at the cost of some money). The parameter to be determined is the number of heads on the coin, i.e. whether it is single-headed

(SH) or double-headed (DH). You have to assign a prior probability for him cheating, i.e. for the coin being SH or DH. There is no 'right' numerical answer; this is the subjective Bayesian prior and so it is up to your judgement. Let's take the initial probability you estimate for him to be cheating to be 20%, which means

$$P(\text{DH}) = 0.2 \qquad \text{so} \qquad P(\text{SH}) = 0.8$$

You play twice and he wins with two heads in a row. The probability of the coin being double-headed (DH) is given by

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})}$$

For the right-hand side, then the likelihood is

$$P(\text{Two heads}|\text{DH}) = 1$$

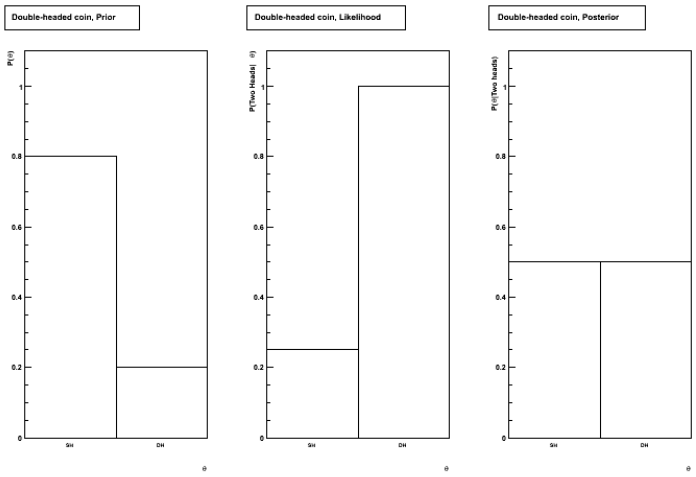
as a double-headed coin will always give two heads in two tosses. For the denominator, using the marginalisation rule

$$\begin{aligned} P(\text{Two heads}) &= P(\text{Two heads}|\text{DH})P(\text{DH}) + P(\text{Two heads}|\text{SH})P(\text{SH}) \\ &= 1 \times 0.2 + 0.25 \times 0.8 = 0.4 \end{aligned}$$

where the probability of getting two heads when using a normal (single-headed) coin is $P(\text{Two heads}|\text{SH}) = (1/2) \times (1/2) = 1/4$. Applying Bayes' theorem then gives

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})} = \frac{1 \times 0.2}{0.4} = 0.5$$

and so your initial guess (the prior) of 0.2 is modified to a posterior of 0.5 given the observed data, i.e. two heads. This is shown graphically below. You begin to suspect him of cheating.



You also can think of this in terms of a probability table, as we discussed in Lecture 1 as we have two random variables, the data (number of heads seen) and the parameter (number of heads on the coin). The table is

		Single-headed	Double-headed
		0.8	0.2
Two heads	0.4	0.2	0.2
Not two heads	0.6	0.6	0.0

and clearly, the two variables are correlated as the joint probabilities are not simply the products of the overall probabilities.

Hence, when you observe two heads (given by the upper row of the central part of the table), then the probabilities of single-headed or double-headed are the same. Quantitatively, we can use the expression for the joint probability to find the conditional probability. This gives

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{DH, Two heads})}{P(\text{Two heads})} = \frac{0.2}{0.4} = 0.5$$

as before.

You are not sure about your flatmate yet, so you decide to play for another two coin tosses, i.e. you repeat the experiment and (surprise) the result is another two heads. The prior for the second experiment is the posterior we just calculated. The likelihood is identical, so the evidence becomes

$$\begin{aligned} P(\text{Two heads}) &= P(\text{Two heads}|\text{DH})P(\text{DH}) + P(\text{Two heads}|\text{SH})P(\text{SH}) \\ &= 1 \times 0.5 + 0.25 \times 0.5 = 0.625 \end{aligned}$$

Applying Bayes' theorem then gives

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})} = \frac{1 \times 0.5}{0.625} = 0.8$$

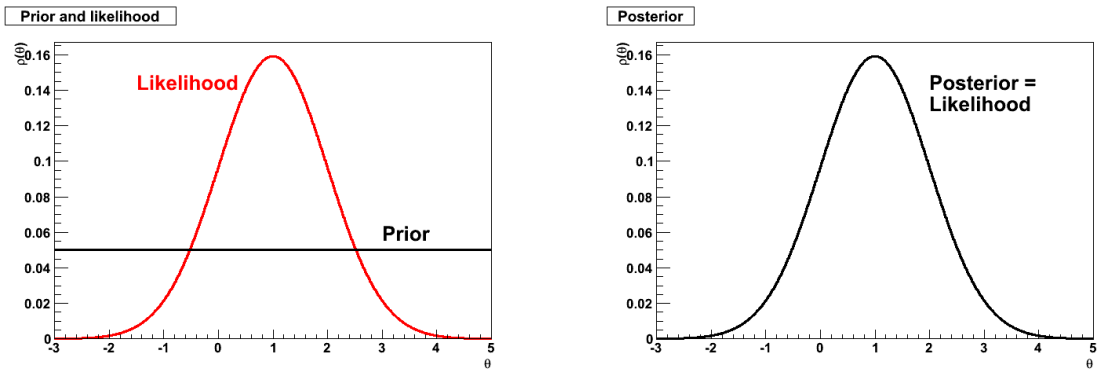
You should definitely suspect him of cheating.

One other note; if you calculated directly from your initial prior of $P(\text{DH}) = 0.2$, with the observed data of four heads (i.e. both experiments together), you would get exactly the same final answer.

18 Priors

It is with the prior that the problem with the Bayesian approach arises; we must have some knowledge of the parameter already, before we can do any estimation from the results of an experiment. Sometimes this is OK; we may not be the first people to measure the electron mass. However, this is not always the case. In addition, it means we cannot give the result 'purely' from our own experiment to compare with other people's experiments.

If we can't, or don't want to, use previous results, then it is standard to use a 'flat' prior, i.e. a uniform distribution for $P(\theta)$. If we do choose a flat prior, then the posterior is the same shape as the likelihood.

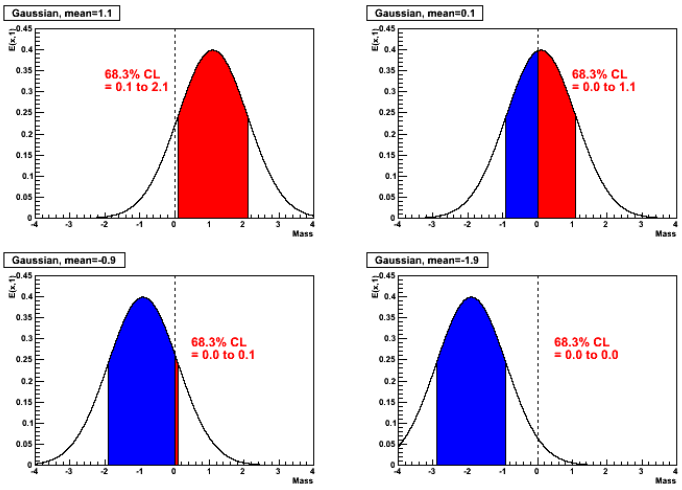


However, this is not as reasonable as it might first seem; in Lecture 3, we saw that a change in variables can change the shape of a distribution, so if we happened to use the square of the

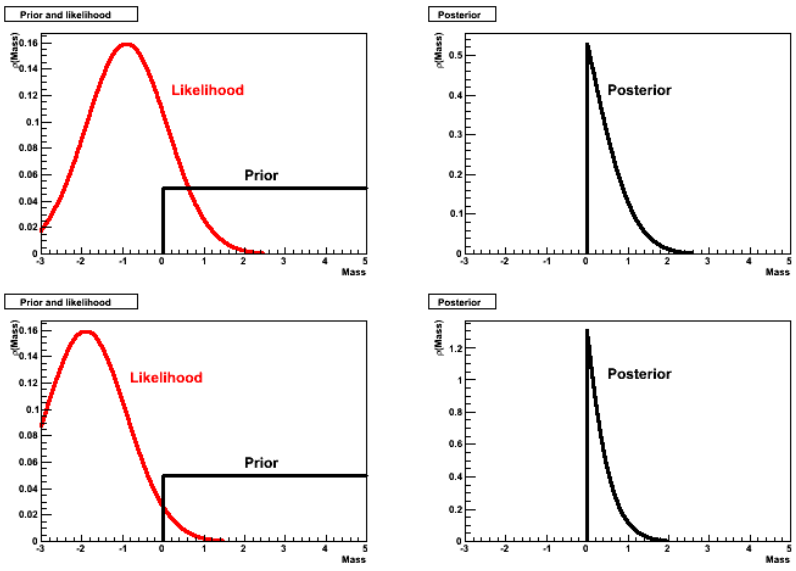
parameter instead, we would have to choose a different distribution in the parameter to get a flat distribution for the square of the parameter. Hence, there is no unique choice for an unknown prior.

19 Physical constraints

We saw that frequentist methods have problems if they end up with results outside the physical region. We discussed an example of measuring the mass of a powder in the last lecture.



How does the Bayesian approach tackle this issue? The prior, whether flat or not, is the key; the probability of the mass genuinely being negative is of course zero, so that we simply make the prior be zero anywhere that is unphysical. This might leave only a small amount of the probability distribution in the physical region, but this is countered by the normalisation constant in the denominator (i.e. the evidence), which will also be small. Hence, it gives a sensible result overall, with the probability distribution for the parameter only being non-zero in the physical region. This is illustrated for the 99.1 g and 98.1 g mass measurements below.



20 Frequentist or Bayesian?

This is a question which has vexed many people for a long time. Discussions can sometimes approach the intensity of religious wars; they are similar in as far as there is never any possibility of 'proving' one is better than the other.

Much of the time it doesn't matter. A Gaussian distribution well away from any physical constraints gives the same result in both frameworks. Since the Central Limit Theorem says many distributions look Gaussian, then in most cases the difference is minimal. Even when there is a difference, it is perfectly correct to use either method, as long as you state clearly what you did.

One of the articles in the reading list is called "Why isn't every physicist a Bayesian?". This argues that the advantages of handling unphysical regions means that the Bayesian approach is superior. I think the answer to the question posed in the title is that many physicists don't like to depend on an undefinable prior to interpret their experiments; they prefer to deal only with hard numbers. In the end, it is up to you to choose!