Worksheet 5 — Data Analysis

1 Introductory Notes

1.1 Some Statistical Definitions

For a set of N measurements of a variable represented by x_i where $i = 0, 1, \ldots, N-1$ the *mean*

$$\begin{aligned} \mu_x &= \frac{1}{N}\sum_{i=0}^{N-1} x_i \\ &= \langle x \rangle \end{aligned} \tag{1}$$

and the *variance*

$$\begin{aligned} \sigma_x^2 &= \frac{1}{N}\sum_{i=0}^{N-1} (x_i - \mu_x)^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{2}$$

where σ_x is the *standard deviation*. The mean calculated in this way is only an estimate of the true mean because only a finite number of measurements is used. The error on μ_x is given by:

$$\sigma_{\mu_x} = \frac{\sigma_x}{\sqrt{N}}. \tag{3}$$

Again, this is only an estimate of the error on the mean. Dividing by $N - 1$ in Equation 3 rather than N improves the estimate by making it unbiased but does not change the fact that it is an estimate.

For two sets of N measurements represented by x_i and y_i the *covariance*

$$\begin{aligned} \mathrm{cov}[x, y] &= \frac{1}{N}\sum_{i=0}^{N-1} (x_i - \mu_x)(y_i - \mu_y) \\ &= \langle xy \rangle - \langle x \rangle \langle y \rangle. \end{aligned} \tag{4}$$

The *correlation coefficient* of x and y is given by:

$$\rho[x, y] = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}. \tag{5}$$

The correlation coefficient is a number between -1 and 1 that tells us to what extent two variables are statistically dependent on each other. If $\rho[x, y]$ is zero then x and y are independent *i.e.* there is no tendency for measurements of x and y to fluctuate in sympathy. If the correlation coefficient is non-zero then there is some dependency and in the extreme case, when $\rho[x, y] = \pm 1$, the two variables are completely dependent and values of x and y fluctuate together.

1.2 Error Calculations with Correlated Variables

This section is for your information and is not required reading for the exercises.

It is often the case that the physical quantity we are interested in is not measured directly but is calculated from one of more variables that have associated uncertainties. The problem is then to estimate the error on the quantity of interest from the errors on the input variables. You may be familiar with how to do this when the variables are uncorrelated but possibly less familiar with the case when they are correlated. Having to deal with correlated input variables is quite common. For example, if the variables were measured from a common data set it is possible that there is some degree of correlation.

Let us say the quantity of interest is a and is calculated using:

$$a = f(x, y),$$

where f is an arbitrary function and x and y are variables with associated errors σ_x and σ_y. In the case when x and y are uncorrelated the square of the error on a

$$\sigma_a^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2.$$

If we have two correlated variables x and y, the square of the error on a becomes:

$$\sigma_a^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\mathrm{cov}[x, y].$$

1.3 Histograms and Binning

Measurements always have a characteristic probability distribution. Making a histogram is a way of estimating the distribution from a finite number of measurements. If the quantity we wish to histogram is x we decide on intervals with endpoints x_0, x_1, ..., x_M which define a set of contiguous 'bins'. The measurements of x are then 'binned' by counting the number that fall within each of the bins. The result is a set of M data bin values $d_{0\ldots M-1}$. The number of entries in a bin will be random

and approximately follow a Gaussian with standard deviation $\sqrt{d}$ if d is about 10 or larger. If the mean number of entries is smaller than about 10 then the Poisson distribution is a better description.

1.4 Fitting (Regression)

Measurements of a quantity are often expected to follow a certain distribution. The form of the distribution is known and described by a function but there are parameters in the function with values which are not known. In this situation the data are 'fitted' with the function by changing the parameters until the 'best' fit is found. By doing this the parameters are measured.

You have probably used a *least squares* fit to fit a straight line through a set of data points in the past. This is one example of the *chi-squared* fitting method. In a chi-squared fit the measurements are first binned to give a set of M values $d_{0\ldots M-1}$ which estimate the distribution. A function is then fitted to the data by minimising the sum of the squared differences between each of the bin values and the expected value predicted by the function. To take account of possible differences in the error on the bin values each 'difference squared' term is divided by the error squared. The chi-squared can therefore be written as:

$$\chi^2 = \sum_{j=0}^{M-1} \frac{(d_j - f_j(\alpha, \beta, \ldots))^2}{\sigma_{d_j}^2} \tag{6}$$

where d_j are the data bin values, σ_{d_j} are their errors and $f_j(\alpha, \beta, \ldots)$ are the predicted values calculated from the function. The function depends on the parameters $\alpha, \beta, \ldots$ and it is these parameters that are adjusted to minimise χ^2. For a straight line fit this can be done analytically but in general it is necessary to iterate until the chi-squared is minimised.

Once the best values of the parameters have been found it is possible to make a quantitative evaluation of how good the fit is from the minimum chi-squared value. If the predicted values are compatible with the data values you should find that

$$\chi^2 \sim N_{\mathrm{df}} = M - N_{\mathrm{c}}$$

where N_{df} is the number of *degrees of freedom* given by the number of data values, M, minus the number of *constraints*, N_{c}. Every parameter that is varied to minimise the chi-squared constitutes a constraint and there may be others such as the sum of data values being constrained to a certain value. It is customary to convert the chi-squared into the *chi-squared per degree of freedom* by dividing the minimum chi-squared by the number of degrees of freedom. The minimum chi-squared will vary a lot from experiment to experiment but the chi-squared per degree of freedom should always be close to one if there are many degrees of freedom. Note that the an

excellent way of quickly determining if a fit has worked is to plot the data with error bars and the fitted function and compare by eye! The errors on the fit parameters are determined by finding the change in each of the parameters that causes the chi-squared to increase by 1. This is done in turn for each parameter while the other parameters are fixed at their fitted value.

The *maximum likelihood* method is another way of performing a fit. The probability, or likelihood, of making a single measurement x is proportional to $f(x)$ where $f(x)$ describes the data probability distribution. Hence, the probability of making the set of measurements $x_{0\ldots N-1}$ is proportional to

$$l = f(x_1)f(x_2)f(x_3)\ldots f(x_{N-2})f(x_{N-1}) \tag{7}$$

If $f(x)$ accurately describes the data distribution then we expect l to be larger than when it does not accurately describe the distribution. This means we can find the best values of the parameters which determine the shape of $f(x)$ by maximising l.

The N multiplications in Equation 7 can lead to very large or very small numbers so it is easier to maximise the log likelihood

$$L(\alpha, \beta, \ldots) = \sum_{i=0}^{N-1} \ln(f(x_i; \alpha, \beta, \ldots)) \tag{8}$$

where we have dropped any constant terms which come from constant multiplicative factors. The errors on the fitted parameters are found in a similar way to the chi-squared method by finding the change in the parameter that causes the log likelihood to decrease by 0.5.

The maximum likelihood method has a number of advantages over the chi-squared method. The data does not have to be binned so there is no loss of information and no assumptions have to be made about the size of errors. The difference becomes most important when statistics are low because large bins have to be used to ensure that each bin has a reasonable number of entries. Another advantage becomes clear if you consider what would happen if there was some (perhaps experimental) effect that depends on x and modifies the data distribution. The distribution becomes $A(x)f(x)$ where $A(x)$ describes the x dependent effect. In the chi-squared method we would need to know $A(x)$ to perform the fit but in the maximum likelihood method the likelihood is simply multiplied by a constant factor $A(x_1)A(x_2)\ldots A(x_{N-1})$ and the position of the maximum is not affected. The main advantage the chi-squared method over the maximum likelihood method is that it provides a quantitative measure of the goodness of fit.

2 Exercises

Week 7, Session 1

2.1 One Random Variable

Generate a set of random values according to a normal distribution with mean 0 and variance 1 using Mathcad's *rnorm* function. *Calculate the mean and variance of the data using the formulae given above.* Try selecting the equation that generates the random data and use the function key F9 to repeatedly generate new data to see how the estimated mean and variance fluctuate. You can use this technique at any time to get a feel for how the numbers and distributions change when you analyse a different data set.

Find out how to use the *hist* function to make a histogram. Use *hist* to histogram the data so that you can see how they are distributed. Make it is easy to change the number of bins, the value of the lower edge of the first bin and upper edge of the last bin. This will then make a useful template for later in the exercises when you will need to repeat this process. Start by using the 'bar' or 'point' plotting option and make sure the bars or points are plotted at the centre of the bins. Experiment with the size of the bins and choose an appropriate bin width for the number of values you are generating. Find out how to plot error bars in Mathcad. The error on the number of entries in a bin is approximately $\pm\sqrt{n}$ where n is the number of entries. Plot error bars on your histogram.

Calculate the error on the mean and check your error using a Monte Carlo approach. The way to do this is to simulate doing a large number of 'experiments' where each experiment involves generating a fixed number of data values and calculating the mean. The error on the mean is just the standard deviation of the mean. You will need to create a large array and calculate the mean for sub-sections of it. Make a histogram of the mean so that you can see its distribution.

Try changing to data generated according to a flat distribution with mean 0 and variance 1. The Mathcad expression $rnd(\sqrt{12}) - \sqrt{12}/2$ returns a random number that fits this specification. Look for any significant changes in the quantities you have calculated and pay particular attention to the error on the mean and the distribution of the mean. *Comment on what does or does not change.*

2.2 Two Correlated Random Variables

Mathcad does not have a built-in mechanism for generating random numbers with a certain correlation, so we will have to do it ourselves. The standard method is to generate two uncorrelated random values and then make orthogonal linear combinations to give two random values with the required degree of correlation. If a and b are uncorrelated random variables with mean 0 and variance 1 then correlated random variables x and y with mean 0 and variance 1 can be calculated using:

$$x = \frac{1}{\sqrt{2}}\left[(\sqrt{1+\rho})a + (\sqrt{1-\rho})b\right] \quad (9)$$

$$y = \frac{1}{\sqrt{2}}\left[(\sqrt{1+\rho})a - (\sqrt{1-\rho})b\right] \quad (10)$$

where ρ is the required correlation coefficient.

Use Equations 9 and 10 to generate two sets of correlated random variables with the normal distribution. Make an XY scatter plot of the data and *describe what happens as you vary ρ between -1 and $+1$. Finally, calculate the correlation coefficient from the data and see how it compares with the expected value.*

2.3 The Chi-Squared Distribution

It is instructive to investigate the shape of the chi-squared distribution for various numbers of degrees of freedom. This will give a feel for the answers we might expect when trying to fit data. Unfortunately, the chi-squared distribution is a rather complex function. However, any good mathematical package should have functions such as this defined, and Mathcad is no exception. This function is very easy to use: dchisq(x, d) gives the probability distribution for a chi-squared with d degrees of freedom.

Try plotting dchisq for values of the number of degrees of freedom between 1 and 10, and *note down the general features of these distributions* (eg mean and distribution). It probably easier to see what is going on if we plot the distribution of chi-squared per degree of freedom. To use the dchisq function to plot this, you will need to plot:

$$d \times \text{dchisq}(d \times x, d)$$

the initial d being there to preserve a normalisation of 1 under the curve. Try plotting this for values of d of 3, 5, 10, 30 and 100. This should give you a better idea how the expected chi-squared value will vary when you increase the number

of degrees of freedom — *note down what you observe. Use this plot to say how you would interpret a value of 1.5 for an observed chi-squared per degree of freedom when fitting data in each of the cases d = 3, 5, 10, 30 and 100.*

Note that although you have been looking at the distribution of the actual chi-squared values, often we refer to the chi-squared probability distribution as the probability (for a certain degree of freedom) that the chi-squared value will lie *below* a definite value, x. This is given by the integral from zero to x of the distributions you have been using.

Week 8, Session 1,2

2.4 Chi-Squared Fitting

In this exercise we are going to fit data with an exponential curve, as would be required in a nuclear decay experiment, for example. However, before we can start to do some fitting we need to generate some data to play with.

The standard version of Mathcad does not include a random number generator for an exponential so we will have to do it ourselves. The need to generate random numbers according to a specific distribution is very common when doing Monte Carlo work so it is worth demonstrating the general technique. The idea is to generate a random number between 0 and 1 from a flat distribution and transform this into a number from the required distribution. We can find the expression that does this by equating the probability of getting a value less than or equal to r from a flat distribution with the probability of getting a value less than or equal to R from the distribution of interest *i.e.*

$$\begin{aligned} \int_0^r dx &= \int_0^R f(x)dx \\ r &= \int_0^R f(x)dx \end{aligned} \tag{11}$$

where $f(x)$ is the normalised function which describes the distribution we want. If we do the integration and solve for R in terms of r we end up with an expression which performs the required transformation.

Find the expression that transforms a value from a flat distribution into a value from the normalised exponential:

$$nexp(t) = \frac{1}{T}\exp(-t/T) \tag{12}$$

where T is the parameter that determines the shape of the distribution (this would be the lifetime in a decay experiment). Use this to generate $N = 1000$ values

according to an exponential with $T = 2$. Histogram the data using 10 bins over the range $0 < t < 10$.

Define a function that calculates the chi-squared for an exponential fit to the data with the normalisation of the exponential, A, and T as parameters. You can assume that the error on a bin is the square root of the number of entries. With A fixed at ($N\times$ bin width), make a plot of the chi-squared as a function of T. *Concentrate on the minimum and describe what happens as you repeatedly generate new data using* the F9 function key. (Using global assignments will allow you to move the data generating expression next to the plot.) *Now increase the value of N and comment on the changes you see. Also try decreasing the number of bins and see how this affects the chi-squared.*

Now make a contour plot of the chi-squared as you vary both T and A. Finding the minimum in the chi-squared by eye is clearly a tedious thing to do if there is more than one parameter. We obviously want the computer to do the hard work here. Mathcad has a variety of built-in fitting functions but they are useless for serious fitting work because there is no way to specify the errors on the data and consequently no way for Mathcad to return the error on the fitted parameters. However, there is a way we can get Mathcad to automate the fitting process. Mathcad has a function called *Minerr* that solves equations using an iterative technique and will return the best solution it can find even if there is no exact solution. We can use *Minerr* to find the best parameter values by asking it to solve the equation $\chi^2(A, T) = 0$. Use *Minerr* in this way and use it to fit for A and T. Make a plot of the data with error bars and the fitted function to check your fit looks good. We can also use *Minerr* to get the error on a fitted parameter by asking it to find the value of the parameter that gives a chi-squared larger than the minimum by 1. *Estimate the error on A and T using this approach.*

2.5 Log Likelihood Fitting (Optional)

Define a function that calculates the log likelihood for an exponential fit to the data with T as a parameter. Make a plot of the log likelihood as T is varied. Again concentrate on the maximum and try regenerating the data to see what happens. Do a maximum likelihood fit for T using *Minerr*. Note that you will have to choose a value of the log likelihood safely above the maximum but not too far above because *Minerr* may have problems. Use *Minerr* to find the error on T.

Think about the pros and cons of using the chi-squared fit and maximum likelihood fit in this case, and try to find ways that either fit might be improved. Finally, which fitting method would you choose to use if there was only a fairly small number of data?

References:

R.J. Barlow , Statistics, Wiley, 1988.
G. Cowan , Statistical Data Analysis, Oxford Univ Press, 1998.
S.L. Meyer , Data Analysis for Scientists and Engineers, 1975.
You can also find statistical information on a number of web sites including

http://www.statsoftinc.com/textbook/stathome.html