Harley Wood Winter School 2012

Practical Statistics for Astrophysicists

Tamara Davis, University of Queensland

1 Introduction

Handling statistics forms a crucial part of any data analysis, but it is something that is somewhat neglected in our formal education, or taught more formally than practically. This lecture will try to bridge that gap, and outline the basic data analysis techniques that are most commonly used in astrophysics, including some common pitfalls and tips-of-the-trade. To focus the discussion we will use cosmology as a case study, and look at the constraints on our cosmological model parameters that can be derived from things like supernovae (SNe), baryon acoustic oscillations (BAO), and the cosmic microwave background (CMB). In other words, how one generates the contours seen in Fig. 1 from the raw cosmological data we measure.

This is not meant to be a rigorous mathematical derivation of the statistics. Just a practical guide on how to implement them.

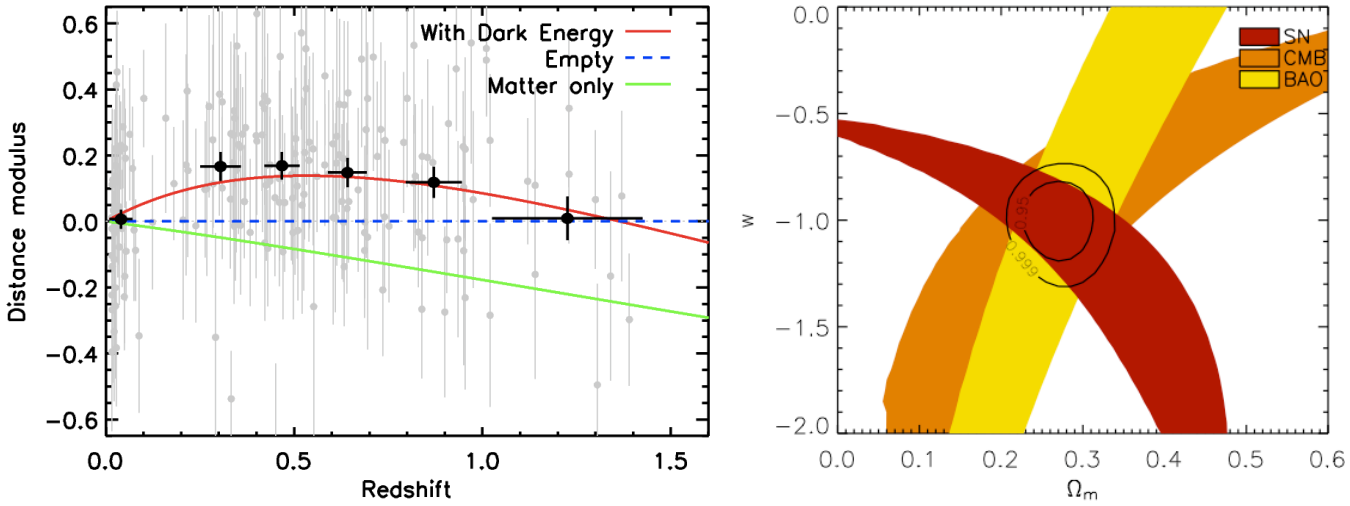

Figure 1: The aim of this lecture is to reveal some of the subtleties involved in creating these types of contour plots for cosmology. How do you go from the supernova distance modulus vs redshift data on the left, to the red 1σ confidence contour on the right? Once you've got there, how do you combine that data with the other data sets including possible correlations, nuisance parameters, and multi-parameter models. Although we're using cosmology as a case study, the techniques are applicable to data analysis in general.

Models and Parameters:

It is important to distinguish up-front the difference between testing models and finding the best fit parameters within those models. Different models could be based on different gravitational theories – for example General Relativity or $f(R)$ gravity – or they could simply be different parameter combinations within a single theory of gravity – for example General Relativity with a cosmological constant and cold dark matter (ΛCDM) or General Relativity with more general dark energy and cold dark matter (wCDM). When you assume the universe is flat, that reduces the number of parameters you need to fit for and therefore constitutes a different model as well. So testing flat-ΛCDM is different to testing ΛCDM.

Let $\mathcal{P}$ denote a set of cosmological parameters. For ΛCDM the parameter set would be

$$\mathcal{P}_{\Lambda\mathrm{CDM}} = (H_0, \Omega_\mathrm{M}, \Omega_\Lambda), \tag{1}$$

where $\Omega_{\rm M}$ and Ω_Λ are the present-day energy density parameters of matter and the cosmological constant respectively, and H_0 is the Hubble parameter. Given these parameters and general relativity we know the expansion history of the universe. Similarly, in other models we would have,

$$\mathcal{P}_{\rm flat-\Lambda CDM} = (H_0, \Omega_{\rm M}), \tag{2}$$
$$\mathcal{P}_{\rm flat-}w{\rm CDM} = (H_0, \Omega_{\rm M}, w), \tag{3}$$
$$\mathcal{P}_{w{\rm CDM}} = (H_0, \Omega_{\rm M}, \Omega_{\rm x}, w). \tag{4}$$

The first part of this lecture will address how to measure the best-fit parameters within a particular model. Comparing between models will be addressed in the final section (Sect. 5).

1.1 Data sets:

Supernovae:

We use supernovae to trace the expansion history of the universe by measuring their apparent magnitude and redshift. Apparent magnitude is a function of redshift, the cosmological parameters of your model, and the absolute magnitude of the supernovae, M, according to,

$$m(\mathcal{P}, z) = 5\log_{10}[D_L(\mathcal{P}, z)] + 25 + M, \tag{5}$$

where D_L is the luminosity distance. The distance modulus is the difference between the apparent and absolute magnitudes,

$$\mu(\mathcal{P}, z) = m(\mathcal{P}, z) - M, \tag{6}$$

or in other words,

$$\mu(\mathcal{P}, z) = 5\log_{10}[D_L(\mathcal{P}, z)] + 25. \tag{7}$$

The luminosity distance is dependent on the cosmological parameters, and has an annoying factor of H_0 that can be factored out by defining D'_L to be the H_0-independent part of D_L, i.e. $D_L = (c/H_0)D'_L$.

In the luminosity distance we can separate out the part that depends on H_0 and just consider the part that depends on the other parameters such as matter density and dark energy. Using $D_L = \frac{c}{H_0}\ D'_L$ the distance modulus can be rewritten as,

$$\mu(\mathcal{P}, z) = 5\log_{10}[D'_{\rm L}(\mathcal{P}', z)] + 5\log_{10}[c/H_0] + 25, \tag{8}$$

where $\mathcal{P}'$ is just $\mathcal{P}$ excluding H_0. This is useful because H_0 is degenerate with another observational issue we have so far ignored. Any uncertainty in the absolute magnitude of the supernovae also enters as an additive constant. This can be incorporated by adding a term, $\sigma_{\rm M}$, that represents how far this arbitrary absolute magnitude is from the correct value,

$$\mu(\mathcal{P}, z) = 5\log_{10}[D'_{\rm L}(\mathcal{P}', z)] + 5\log_{10}[c/H_0] + 25 + \sigma_{\rm M}, \tag{9}$$
$$= 5\log_{10}[D'_{\rm L}(\mathcal{P}', z)] + \mathcal{M}, \tag{10}$$

where now all of the additive terms, including uncertainty in Hubble's constant and SN Ia absolute magnitude, have been collected in the parameter, $\mathcal{M}$. When trying to measure dark energy, we don't care about H_0 and σ_M, and since they have relatively large uncertainties we would prefer not to fit for them. Instead we treat them as nuisance parameters and marginalise over them (see Sect. 3).

2 Comparing data to model

2.1 χ^2, $\Delta\chi^2$, reduced χ^2, and combining data sets

To calculate how well a model matches the data, we start with a χ^2 test. The lower the χ^2 the better the fit. The value of χ^2 for a particular data/model combination is given by,

$$\chi^2_0 = \sum_i \left(\frac{\mu_{\rm model} - \mu_i}{\sigma_i} \right)^2, \tag{11}$$

where $\mu_{\rm model}$ is the value predicted by your cosmological model and $\mu_i \pm \sigma_i$ is the ith data point and its uncertainty.

Given a value of χ^2 you can calculate the likelihood of that model. We show how to calculate likelihoods in Sect. 2.2, but you can also calculate your uncertainties directly from the χ^2 values. These are appropriate in the case of Gaussian distributed likelihoods. (The key is not whether the data points are Gaussianly distributed about the model, but rather whether the resulting distribution of likelihoods is Gaussian.)

What is the best fit, and the uncertainty?: The parameter values that give you the lowest χ^2 are the best fit parameters of your model. To calculate the uncertainties on those best fits you compare the other models to the χ^2 for the best fit,

$$\Delta\chi^2 = \chi^2 - \min(\chi^2). \tag{12}$$

The larger the $\Delta\chi^2$ the worse the fit. There are standard thresholds of $\Delta\chi^2$ that correspond to uncertainties of one, two, and three standard deviations ($\sigma_1, \sigma_2, \sigma_3$). The meaning of the standard deviations is that simply, 68.27%, 95.45%, and 99.73% of the data should lie within the one, two, and three σ limits, respectively. Note that these are not exactly round numbers, and the 95% confidence interval is slightly different to the 2σ limit.[1]

The values of $\Delta\chi^2$ corresponding to $[\sigma_1, \sigma_2, \sigma_3]$ differ depending on the number of parameters you're fitting. When you're fitting one parameter $\Delta\chi^2 = [1, 4, 9]$ correspond to the first, second, and third standard deviations, respectively. When you're fitting two parameters the first three standard deviations are given by $\Delta\chi^2 = [2.30, 6.18, 11.83]$. When you're fitting more parameters I recommend you use likelihoods rather than the raw χ^2 value anyway. We come to that in the next section.

Table 1: $\Delta\chi^2$ values corresponding to different standard deviations ($\sigma_1, \sigma_2, \sigma_3$) for different numbers of parameters (one and two).

$\Delta\chi^2$ for	σ_1	σ_2	σ_3
one parameter	1	4	9
two parameters	2.30	6.18	11.83

How good is the fit?: When fitting models to data it is not enough to find out which parameters give the best fit, you also need to establish whether that best fit is actually a good fit. To get an approximate measure of goodness of fit use the **reduced** χ^2. To get the reduced χ^2 divide χ^2 by the number of degrees of freedom: $n_{\rm dof} = n_{\rm data} - 1 - n_{\rm params}$, where $n_{\rm data}$ is the number of data points and $n_{\rm params}$ is the number of parameters you are fitting for. For a good fit the result should be close to 1. This is not rigorously defined rule, but rather a rule of thumb. A reduced χ^2 much less than 1 indicates that the error bars are probably overestimated (too large), while a reduced χ^2 of much larger than 1 indicates that the data aren't a good fit to your model. However, care should be taken because a high reduced χ^2 could actually indicate a number of other things as well. A high reduced χ^2 could mean,

[1] The standard deviation values are actually defined as $\sigma_z = {\rm erf}(z/\sqrt{2})$, where erf (the error function) is the cumulative sum of a gaussian distribution,

$${\rm erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{13}$$

- the model is inadequate and a better model is needed;
- the uncertainties on the data points have been underestimated;
- there is a systematic error in the data.

Including priors: If we have prior information about a particular parameter that goes into this model, then we want to weight the χ^2 value so it prefers parameter values close to our prior. We do this by adding an extra term in the χ^2 equation. Say we have a prior on a parameter, x, such that we know $x = x_{\rm prior} \pm \sigma_{\rm prior}$. For example $x_{\rm prior}$ could be $\Omega_{\rm M} = 0.27 \pm 0.03$ or a combination such as $\Omega_{\rm M} + \Omega_\Lambda = 1.00 \pm 0.02$. The prior contributes,

$$\chi^2_{\rm prior} = \left(\frac{x_{\rm model} - x_{\rm prior}}{\sigma_{\rm prior}} \right)^2 . \tag{14}$$

Then the total is simply the sum of those χ^2 values, $\chi^2 = \chi^2_0 + \chi^2_{\rm prior}$.

Including other independent data sets: Including other data sets is simple, when the data sets are independent. You simply repeat the above for whichever data set you are interested in, calculate the χ^2 for that new data set, and add it to the χ^2 total as though it were another prior. The $\Delta\chi^2$ thresholds for $\sigma_{1,2,3}$ don't change as you add extra data sets, because you still have the same number of parameters in your model. We deal with correlated data in Sect. 4.

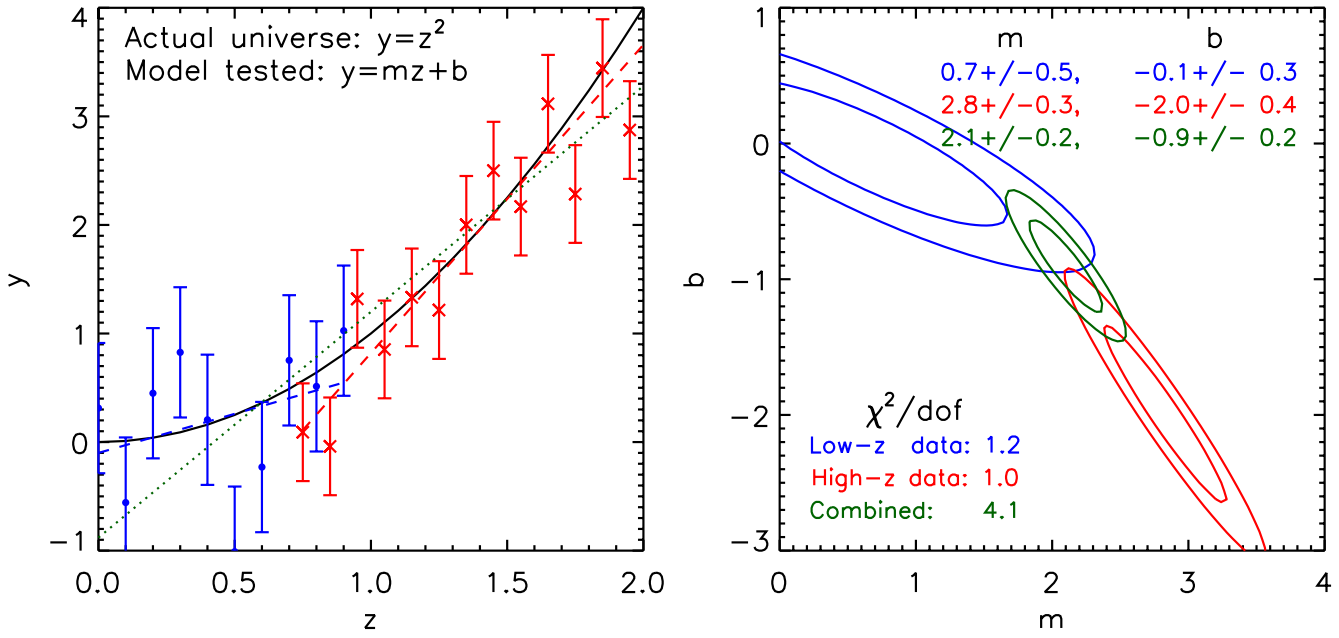

Figure 2: Contour plot showing two badly aligned data sets. How good is the fit and how tight are the constraints? The real model from which the data is generated is a parabola $y = z^2$. The model we're testing is a linear fit, $y = mz + b$. The two parameters we fit for are the slope, m, and the y-intercept, b. I've split the data into two overlapping data sets. The low-z data in blue and the high-z data in red. The linear model is a good fit to each data set, given the uncertainties I've given the fake data points. I've given the high-z data slightly tighter error bars and more data points, which results in slightly tighter contours. The best linear fits are shown in dashed lines on the plot on the left, with the dotted line showing the best fit to the two data sets combined. The plot on the right shows the likelihood contours in the m vs b plane, and the best fit values are given in the upper right for each of the data sets (red and blue), and the data sets combined (green). The χ^2 per degree of freedom is given in the bottom left. The point of this plot is to show that even though both data sets are good fits to the linear model (χ^2/dof$\sim$ 1), once you combine them the total indicates a bad fit (χ^2/dof$\sim$ 4). Nevertheless, you often see people quoting an amazingly precise result, like the green contours here, when all that is actually showing is that the model is a bad one. Beware the wrath of the referee if you make that error.

2.2 Likelihoods

Converting a χ^2 value into a likelihood $\mathcal{L}$ is simple,

$$\mathcal{L} = e^{-\chi^2/2}, \tag{15}$$

from which you can see why χ^2 is often referred to as the 'log-likelihood',

$$-2\ln\mathcal{L} = \chi^2. \tag{16}$$

Thanks to the logarithm you can see that whereas you *add* χ^2 values from independent experiments to get the total χ^2 you *multiply* likelihoods to get the total likelihood.

Once you have a likelihood distribution for a parameter, it is straightforward to integrate under the curve to find the value of the likelihood, within which 68.27% of the likelihood area is enclosed. The values of the parameter that match that likelihood are your $\pm 1\sigma$ uncertainties.

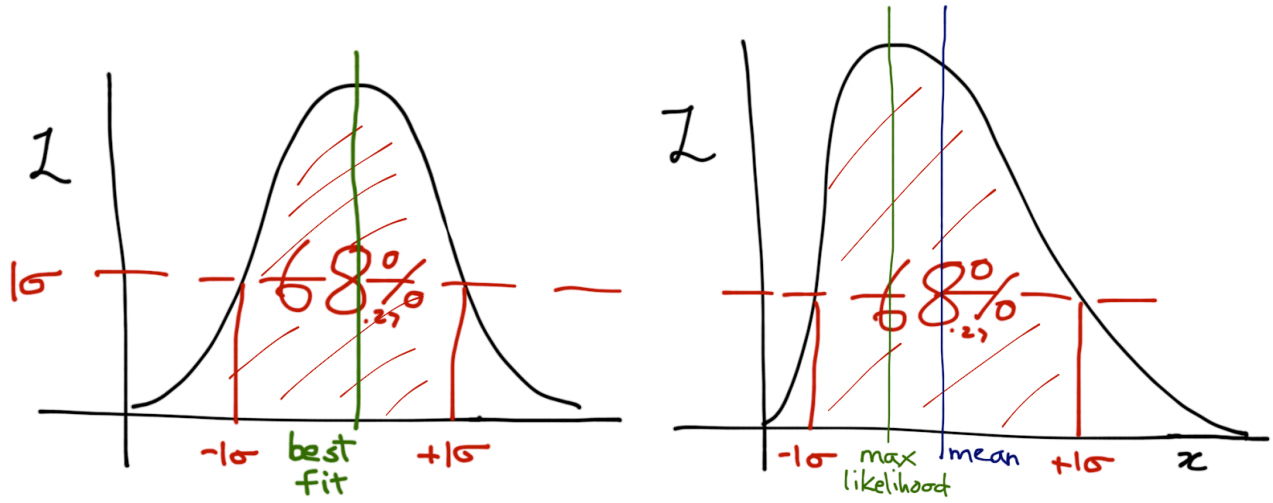

Figure 3: Demonstration of 1σ range under a Gaussian likelihood distribution (left) and a skewed likelihood distribution (right). Note that in the skewed distribution the maximum likelihood point is not the mean of the distribution.

The same procedure works whether or not the likelihood surface is Gaussian.

However, in the case of a non-Gaussian distribution, picking your best fit value and uncertainties becomes tricky. The maximum likelihood value is no longer in the centre of the distribution, and if there is a long tail in one direction or the other, then the majority of the likelihood can end up sitting far to one side or the other of the actual best fit. Therefore, another way to choose the best fit value is to choose the value for which 50% of the likelihood is above and 50% below. In other words, take the mean of the posterior distribution (if you like Bayesian language) or find where the cumulative likelihood hits 0.5.

When the likelihood distribution is Gaussian, these two are equivalent, but in the case where the likelihood distribution is skewed, the two can differ substantially (see Fig. 4). It has become good practise to quote both results. For example, the Wilkinson-Microwave-Anisotropy-Probe (WMAP) seven-year cosmology results for the ΛCDM model (Komatsu et al., 2011, Table 1) give two slightly different constraints on Ω_Λ. Their maximum likelihood value is $\Omega_\Lambda = 0.227$, while the mean of the posterior distribution is $\Omega_\Lambda = 0.249^{+0.056}_{-0.057}$.

The other instance in which the mean and the maximum likelihood will not coincide, is when a parameter has a hard cut-off on one side, and that cutoff is within the realistic limits of your measurement. For example, matter density can't be less than zero, so if the data allow for zero matter density then the likelihood value will be truncated.

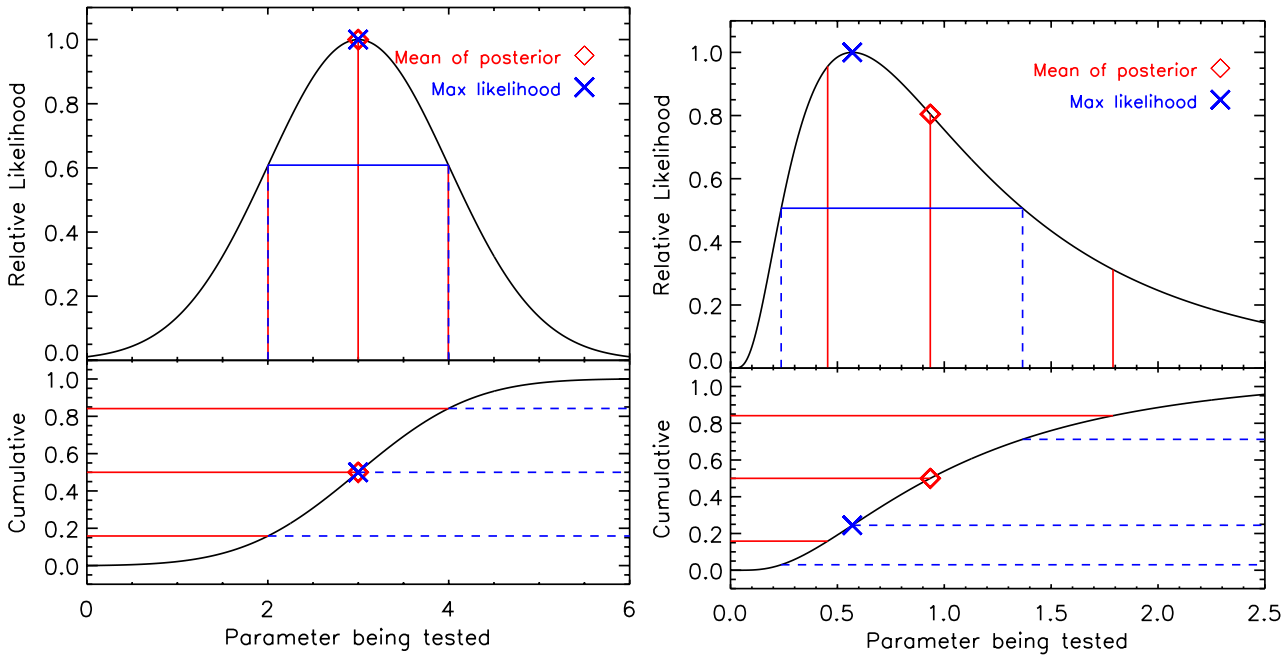

Figure 4: Example of a Gaussian likelihood distribution on the left, and a skewed likelihood distribution on the right. The relative likelihood appears in the top panel of each with the cumulative likelihood below. In the case of a Gaussian likelihood distribution there is no difference between the mean and max-likelihood methods for determining the best fit. However, in the skewed case the values are very different. The maximum likelihood method would give $0.57^{+0.80}_{-0.33}$, whereas the mean of the posterior would give $0.93^{+0.86}_{-0.48}$. The maximum likelihood makes more sense when looking at the relative likelihood plot, as the point chosen has the maximum likelihood (thus the name), and the $\pm 1\sigma$ limits have the same relative likelihood (horizontal blue line). On the other hand, the mean makes more sense when looking at the cumulative likelihood plot, as the point chosen is in the middle of the probability distribution and the $\pm 1\sigma$ limits enclose equal amounts of the cumulative probability. Which you use is a matter of taste.

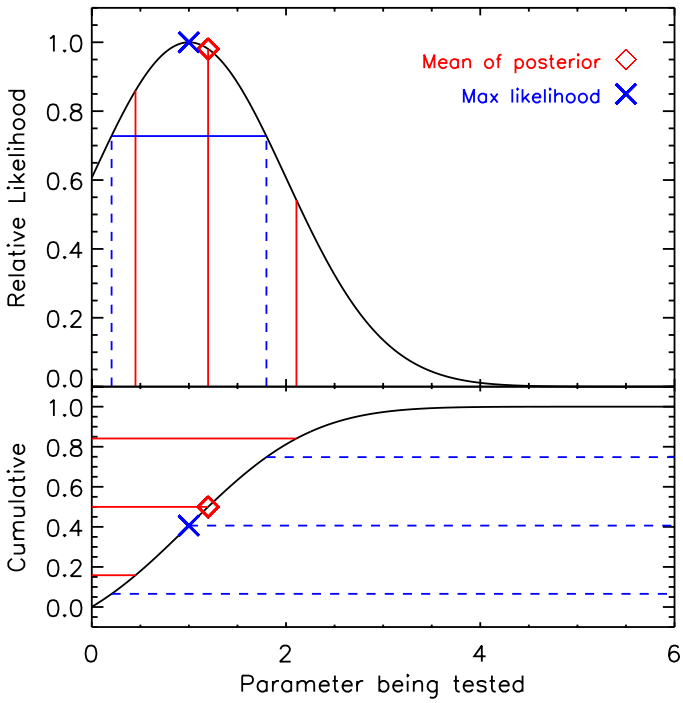

Figure 5: A truncated likelihood like this one also results in the mean and max-likelihood methods differing. This occurs where the parameter range being tested either (1) doesn't include the full range that the data says is possible, or (2) butts up against a hard limit, such as the limit of $\Omega_M \geq 0$ needed because we can't have a negative matter density.

3 Marginalisation

Marginalisation is primarily needed in two circumstances. Firstly, when you have a nuisance parameter (like $\mathcal{M}$ for SNe), and secondly when you have a multi-parameter fit for which you would like to quote results for a single parameter (or plot a 2D contour when you've tested more than two parameters). Marginalisation just reduces the dimensions of your array of likelihoods.

At its heart marginalisation is a very simple procedure. Say you're testing Ω_M with SN data, within the flat-ΛCDM model. The parameter Ω_M is the only free parameter of the model, but due to uncertainties in the absolute magnitude of the SNe and H_0, which both shift the magnitude redshift curve up and down, you also have to vary the nuisance parameter $\mathcal{M}$. So for each value of Ω_M you test, you have an array of χ^2 values, corresponding to all the different $\mathcal{M}$ values you tested. To marginalise over $\mathcal{M}$ you just convert those χ^2 values to likelihoods and add up the total likelihood for each Ω_M. (Note that in the case of the distance modulus data from SNe there is an analytical way to marginalise, but I'm just going into the general principles applicable to all data sets here.)

When you have a Gaussian likelihood distribution, you can use a shortcut and rather than summing over the likelihoods, you can just use the χ^2 values for the best-fit $\mathcal{M}$ for each Ω_M.

When using a Monte-Carlo Markov Chain (MCMC) technique, rather than a grid, to test your parameters, which is recommended for large data sets, the theory is the same but the procedure is slightly different. The MCMC generates points for each model tested, and it is the density of points in each Ω_M bin that gives you the likelihood of that Ω_M value. So finding the maximum likelihood in an MCMC chain involves finding where the density of points is highest. Marginalising over extra parameters is just as simple in the MCMC case – you just add up the number of points, N, in the unwanted parameters and contribute them to the corresponding Ω_M you're testing (and normalise by the total number of points in the chain).

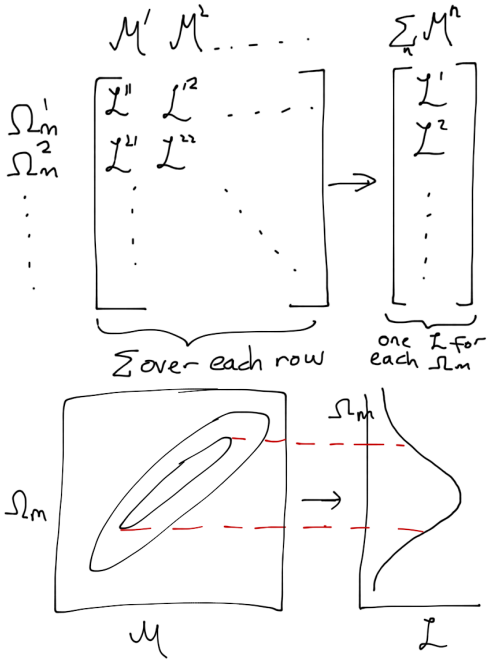

Figure 6: Marginalising over a nuisance parameter $\mathcal{M}$ to extract constraints on Ω_M. For each Ω_M being tested you simply add the likelihoods for all possible values of $\mathcal{M}$, to get the total relative likelihood for that Ω_M. On the left this is shown as reducing a matrix of likelihoods to a 1D array. On the right is the equivalent pictorial representation, reducing a contour to a 1D likelihood curve.

$$\mathcal{L}(\Omega_M) = \int \mathcal{L}(\Omega_M, \mathcal{M}) d\mathcal{M}, \quad \text{in theory,} \tag{17}$$

$$= \sum_i \mathcal{L}(\Omega_M, \mathcal{M}_i), \quad \text{over a grid,} \tag{18}$$

$$= \mathcal{L}(\Omega_M, \mathcal{M}_{\rm max}), \quad \text{when Gaussian,} \tag{19}$$

$$= \sum_i N(\Omega_M, \mathcal{M}_i)/N_{\rm tot}, \quad \text{in MCMC.} \tag{20}$$

Even when there are no 'nuisance' parameters, you often need to marginalise to present the results of a multi-parameter fit. For example, when fitting for BAO in ΛCDM you need to fit for the set $\mathcal{P}_{\Lambda CDM} = (H_0, \Omega_M, \Omega_\Lambda)$. How do you generate the contour plots for $(\Omega_M, \Omega_\Lambda)$? You marginalise over H_0. In other words, $\mathcal{L}(\Omega_M, \Omega_\Lambda) = \sum_i \mathcal{L}(H_{0,i}, \Omega_M, \Omega_\Lambda)$. In general, for higher numbers of parameters, you just keep repeating the sum over the unwanted parameters until you get down to the parameter you're interested in.

Since you generally are just looking for relative likelihoods, normalising the likelihood surface is often not necessary, but it is good practise to make sure that the total likelihood adds up to one.

Is your best fit still a good fit? Because you sum likelihoods, that's equivalent to multiplying χ^2 values. Don't attempt to look at the χ^2 values after marginalisation. Just look at the lowest χ^2 in your unmarginalised grid to see whether that is a good fit.

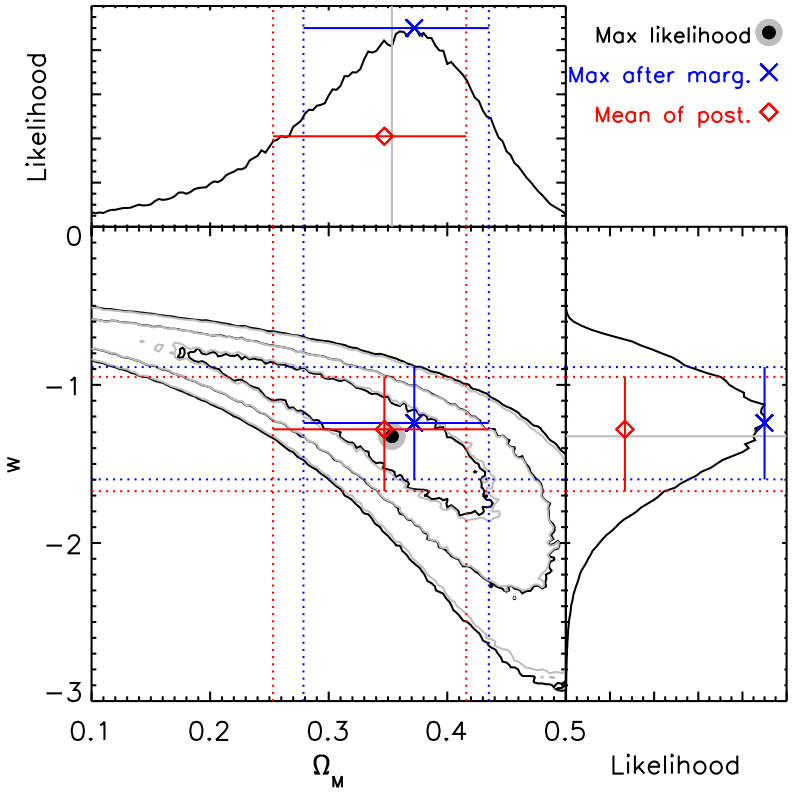

Figure 7: Likelihood contours for the flat-*w*CDM model, using the SN data from the Union 2 compilation. Black contours show the 1, 2, and 3, sigma χ^2 limits for 2 degrees of freedom. Gray contours show the likelihoods for the same. They differ slightly because this test has been done on a grid and the grid truncates the edge of the distribution a little. The upper and right panels show the marginalised likelihood distributions for Ω_M and w respectively. The maximum likelihood point before marginalisation is shown as the grey/black bullseye in the contour plot. The maximum likelihood of the 1D distributions after marginalisation are shown as the blue crosses and error bars. The mean of the posterior is shown as the red diamond and error bars. Note that the maximum likelihood of the marginalised distribution does not find the maximum of the total distribution, but the mean of the posterior comes closer to picking the true best fit.

Mean of the posterior vs maximum likelihood

In Figure 7 I've demonstrated the marginalisation procedure for the flat-*w*CDM model with supernova data. The two parameters are w and Ω_M, and the 2D likelihood contours are shown in the bottom left, with the 1D marginalised likelihood distributions for Ω_M and w in the upper and right-hand panels respectively. The peak of the entire likelihood distribution is shown as the black and grey 'bullseye' at the centre of the contours. Notably, this maximum likelihood is *not* the maximum likelihood of either parameter after marginalisation. If you marginalise, and *then* choose the maximum likelihood for each parameter, you get the blue cross. On the other hand, if you take the mean of the marginalised likelihood distribution you get the red diamond.

When you look at the marginalised distributions you would think "Why would one ever use the mean of the posterior, since it obviously doesn't pick the best fit value of this parameter?" However, the reason becomes clear when you go back to the full likelihood distribution and see that the mean after marginalisation is closer to the maximum likelihood point of the whole distribution than the maximum likelihood after marginalisation is. (The red diamond is closer to the black bulls-eye than the blue cross is.) In general, the mean of the posterior is a better representation of the overall likelihood distribution than the maximum likelihood point after marginalisation.

If we know the best fitting point overall (the bullseye), then why go to all the trouble of marginalising anyway? Well, that bullseye point does not have any uncertainty estimates. The real reason we go to the trouble of marginalising is to figure out the size of our error bars. Again, you can see in the WMAP papers (Komatsu et al., 2011) that their maximum likelihood value is not given any error bars, while the mean of the posterior distribution is given error bars. So if you're using WMAP to fix parameters of a model, they recommend you use their maximum likelihood values, but if you're varying that parameter and using WMAP as additional data, you'd use the mean of the posterior with the uncertainties.

4 Correlated data sets

Whether combining with the CMB, or taking the ratios of two BAO measurements, there are many potential sources of correlation between data sets. In this section we attempt to present in a straightforward manner how to account for them in your statistical analysis.

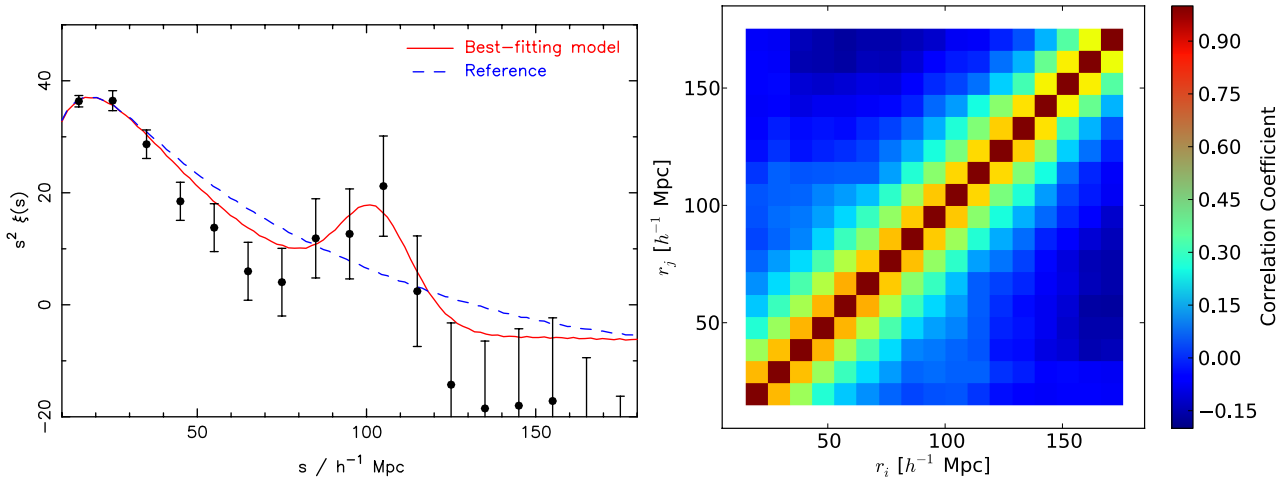

Figure 8: An example of correlated data from Blake et al. (2011). This is the correlation function measured for the galaxies discovered by the WiggleZ survey on the Anglo-Australian Telescope. The correlation function tells you the overdensity of galaxy pairs as a function of separation. In this case we're measuring the correlation between galaxy position, but that's not the correlation I want to point out. Rather, I want to point out that the data points on this plot are correlated. That's because the volumes in which we're measuring our galaxy pairs are overlapping, so we're sampling the same density field. The covariance matrix is shown on the right. The axes are both separation (corresponding to s on the left panel). Red is correlated, blue is uncorrelated or anti-correlated. (To be precise, we plot the amplitude of the cross-correlation, $C_{ij}/\sqrt{C_{ii}C_{jj}}$.) The correlation explains why the data points aren't as scattered as you would expect for the size of the error bars. The error bars are just the diagonal part of the covariance matrix.

When the two data sets are not independent you can't simply add the χ^2 values together, because that would double-count the correlated part of the data and give falsely tight, and potentially misleading constraints. The χ^2 technique described above, is actually just a simplification of the full analysis taking correlations into account. To do it properly you should use the covariance matrix, which is simply the matrix of uncertainties.

In the simple case of uncorrelated data, the equation for χ^2 can be rewritten as a matrix product. Say you have two data points with theoretical expectation $x_{\rm thry}$ and $y_{\rm thry}$ and measured values $x_{\rm obs} \pm \sigma_x$ and $y_{\rm obs} \pm \sigma_y$. The covariance matrix is just the matrix of the variances, which in the uncorrelated case is simply,

$$\mathbf{V} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}. \tag{21}$$

Writing the difference between observation and theory as a vector, $\mathbf{X}^{-1} = (x_{\rm thry} - x_{\rm obs},\ y_{\rm thry} - y_{\rm obs}) = (\Delta x, \Delta y)$, the equation for χ^2 is then

$$\chi^2 = \mathbf{X}^{-1}\mathbf{V}^{-1}\mathbf{X}, \tag{22}$$

or writing it out longhand,

$$\chi^2 = [\Delta x, \Delta y] \begin{bmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{23}$$

Correlated data means that if one data point lies above the expected curve, the other is likely to do so too. So if you don't take that into account it can (a) lead to skewed results and (b) give you falsely tight constraints. The way to take into account correlations is to add uncertainty to the measurement. Another way to put it is that when data points are correlated, adding extra data reduces the uncertainty by less than the usual $\sqrt{n}$ amount (where n is the number of data points).

In general, when two parameters, with theoretical expectation $x_{\rm thry}$ and $y_{\rm thry}$ and measured values $x_{\rm obs} \pm \sigma_x$ and $y_{\rm obs} \pm \sigma_y$ are correlated by correlation coefficient ρ_{xy} (not to be confused with density) the covariance and inverse covariance matrices are given by (Wall & Jenkins, 2003, Eq. 4.4),

$$\mathbf{V} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}^2 \sigma_x \sigma_y \\ \rho_{xy}^2 \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix} \quad ; \quad \mathbf{V}^{-1} = \frac{1}{1-\rho_{xy}^2} \begin{bmatrix} \frac{1}{\sigma_x^2} & \frac{-\rho_{xy}}{\sigma_x \sigma_y} \\ \frac{-\rho_{xy}}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix}. \tag{24}$$

For three parameters, the covariance matrix (before inversion) is

$$\mathbf{V} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}^2 \sigma_x \sigma_y & \rho_{xz}^2 \sigma_x \sigma_z \\ \rho_{xy}^2 \sigma_x \sigma_y & \sigma_y^2 & \rho_{yz}^2 \sigma_y \sigma_z \\ \rho_{xz}^2 \sigma_x \sigma_z & \rho_{yz}^2 \sigma_y \sigma_z & \sigma_z^2 \end{bmatrix}. \tag{25}$$

At this point I stop writing out the inverse covariance matrices and recommend you get your computer to calculate those for you. ⌣ There are various ways to calculate what the correlation is between two measurements, but we leave that for another day.

Slightly more complex correlations:

Sometimes you need to combine data from different sources before calculating χ^2. If the data you're combining is correlated, there are some subtleties in how to address that. For example, the s_z parameter used in BAO studies is actually a ratio between the sound horizon scale at last scattering, as measured by the CMB ℓ_A to that seen in the galaxies d_z. (They're actually defined in opposite senses, so the ratio becomes the product, ... details details ...).

When working with WiggleZ, we had three measurements of d_z and one measurement of ℓ_A that you combine to create three measurements of $s_z = \ell_A d_z / \pi$. Two of the d_z values were correlated to each other, while all of the d_z measurements were correlated with the ℓ_A. How do you calculate the covariance matrix for that combination?

Start with the vector $\mathbf{s} = [s_1, s_2, s_3] = [d_1 \ell_A, d_2 \ell_A, d_3 \ell_A]$. The Jacobian is a handy matrix made up of the partial derivatives of the components of s with respect to each of d_1, d_2 ,d_3, and ℓ_A:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial s_1}{\partial d_1} & \frac{\partial s_1}{\partial d_2} & \frac{\partial s_1}{\partial d_3} & \frac{\partial s_1}{\partial \ell_A} \\ \frac{\partial s_2}{\partial d_1} & \frac{\partial s_2}{\partial d_2} & \frac{\partial s_2}{\partial d_3} & \frac{\partial s_2}{\partial \ell_A} \\ \frac{\partial s_3}{\partial d_1} & \frac{\partial s_3}{\partial d_2} & \frac{\partial s_3}{\partial d_3} & \frac{\partial s_3}{\partial \ell_A} \end{bmatrix} = \begin{bmatrix} \ell_A & 0 & 0 & d_1 \\ 0 & \ell_A & 0 & d_2 \\ 0 & 0 & \ell_A & d_3 \end{bmatrix}. \tag{26}$$

Imagine that d_1 and d_2 are correlated (as in the SDSS points at $z = 0.2$ and $z = 0.35$) but d_3 is uncorrelated (for example, a WiggleZ point at $z = 0.6$), then the covariance matrix for the original four data points looks like

$$\mathbf{C} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & 0 & 0 \\ \sigma_{12}^2 & \sigma_{22}^2 & 0 & 0 \\ 0 & 0 & \sigma_{33}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\ell_A \ell_A}^2 \end{bmatrix}. \tag{27}$$

We need to convert that to a new covariance matrix for the three components of $\mathbf{s}$. That is what the Jacobian is used for. The covariance matrix for $\mathbf{s} = [s_1, s_2, s_3]$ is,

$$\mathbf{C}^{\rm new} = \mathbf{J}\,\mathbf{C}\,\mathbf{J}^T = \begin{bmatrix} \sigma_{11}^2 \ell_A^2 + \sigma_{\ell_A \ell_A} d_1^2 & \sigma_{12}^2 \ell_A^2 + \sigma_{\ell_A \ell_A} d_1 d_2 & \sigma_{\ell_A \ell_A} d_1 d_3 \\ \sigma_{12}^2 \ell_A^2 + \sigma_{\ell_A \ell_A} d_1 d_2 & \sigma_{22}^2 \ell_A^2 + \sigma_{\ell_A \ell_A} d_2^2 & \sigma_{\ell_A \ell_A} d_2 d_3 \\ \sigma_{\ell_A \ell_A} d_1 d_3 & \sigma_{\ell_A \ell_A} d_2 d_3 & \sigma_{33}^2 \ell_A^2 + \sigma_{\ell_A \ell_A} d_3^2 \end{bmatrix}. \tag{28}$$

A similar procedure then needs to be followed to calculate the correlation between the results of this measurement and any other correlated data, such as the CMB-$\mathcal{R}$ parameter.

5 Model comparison

So far we've been trying to figure out what the best fit parameters are within a model. We've seen that when a model is a poor fit, you get a χ^2 per degree of freedom significantly greater than one. That often indicates that a more complex model is needed – for example, one with an extra parameter, or a new model entirely. Some people have tried to quantify how bad the χ^2 must be before a new model, or extra parameter, is justified. This is by no means an exact science, and is to a great extent just quantifying your common sense.

In general, when one model is nested within another model, as in the case of flat-ΛCDM being a special case of ΛCDM, the model with extra parameters is *guaranteed* to be a better fit. So when comparing between models, you can't just prefer the model with the lowest χ^2 because add another parameter, and you'll get a lower χ^2 again. So there has to be some threshold of improvement that you get in χ^2 to justify the addition of the extra parameter.

There are various methods used to penalise the model with the extra parameters. Bayesian model selection is a popular way to select which model is the preferred model. You will often see simplified concepts such as the Akaike Information Criterion (AIC) and Baysean Information Criterion (BIC) used. Basically these give a threshold by which the more complex model must improve on the simpler one before the extra parameter is considered justified by the data. Thus, the aim is to quantify Occam's razor and prefer the simpler model unless the data show a more complex model is necessary to get a good fit.

The BIC is also known as the Schwarz information criterion (Schwarz, 1978) and is given by,

$$\mathrm{BIC} = -2\ln\mathcal{L} + k\ln N, \tag{29}$$

where $\mathcal{L}$ is the maximum likelihood, k is the number of parameters, and N is the number of data points used in the fit. You can see that in the case of Gaussian errors the difference in the BIC values between two models just become $\Delta\mathrm{BIC} = \Delta\chi^2 + \Delta k \ln N$. A difference in BIC of 2 is considered positive evidence against the model with the higher BIC, while a ΔBIC of 6 is considered strong evidence (Liddle, 2004). Meanwhile the AIC (Akaike, 1974) is given by,

$$\mathrm{AIC} = -2\ln\mathcal{L} + 2k. \tag{30}$$

This gives results similar to BIC, although the AIC is somewhat more lenient on models with extra parameters.

You can test these yourself by creating data scattered about a quadratic, like I did in Fig. 7. Fit a cubic to that data and you will improve the fit. That doesn't mean that the cubic is the right model, and you can see if you use AIC or BIC that the extra parameter isn't justified.

I mention these so that you know what their motivation is when you see them, but in general just recognise that models with more parameters will tend to have lower χ^2 values, and that doesn't necessarily mean that they are a better model.

That's it! Hope that helped.

References

Akaike, H. 1974, IEEE Transactions on Automatic Control, 19, 716

Blake, C. et al. 2011, MNRAS, 415, 2892, 1105.2862

Komatsu, E. et al. 2011, ApJS, 192, 18, 1001.4538

Liddle, A. R. 2004, MNRAS, 351, L49, astro-ph/0401198

Schwarz, G. 1978, The Annals of Statistics, 6, 461

Wall, J. V., & Jenkins, C. R. 2003, Practical Statistics for Astronomers (Cambridge, UK: Cambridge University Press)