Astronomical Statistics: Tutorial Questions 2

John Peacock
C20, Royal Observatory; jap@roe.ac.uk

October 2012

1. A model with 10 parameters describes the statistics of the pattern of temperature fluctuations of the Cosmic Microwave Background. We fit this model to 50 published data points of the variance of the fluctuations as a function of angular scale and find $\chi^2 = 53.3$. We then set 5 of the model parameters to zero and fit the model again, this time getting $\chi^2 = 54$. Were we justified in using all 10 parameters in the model?

2. A paper claims to measure the matter density parameter as $\Omega_m = 0.24 \pm 0.015$. A competing paper published by different authors claims that $\Omega_m = 0.30 \pm 0.025$. Assuming that the data used are independent, what value for Ω_m and its uncertainty $\Delta\Omega_m$ should we adopt by combining these results?

   A year later, both groups publish a refined analysis, in which their claimed values for Ω_m are unchanged, but their errors have declined by a factor 2. Should we believe these claims, and what would now be a sensible estimate of the combined result and its error?

3. At early times, the density of the universe is weakly perturbed, so that $\rho(\mathbf{x}) + \bar{\rho}[1 + \delta(x)]$, where $\delta(x) \ll 1$ is a field that has a nearly Gaussian probability distribution with mean zero and variance σ^2. The field cannot be exactly Gaussian, because the density cannot be negative. Consider the lognormal transformation $1 + \delta(\mathbf{x}) = \exp[G(\mathbf{x}) + c]$, where G is an exactly Gaussian zero-mean field and c is a normalization constant. What is the pdf of δ in this model, and how are the variances in δ and in G related? Thus show that a suitable choice of c allows δ to describe a density fluctuation, in which δ approaches G in the limit $\sigma \to 0$.

4. Show that the pdf for the sum of two random variables, x and y, is the convolution of the pdfs $p(x)$ and $p(y)$. Repeat the derivation to deduce an expression for the pdf of $z \equiv xy$, the product of two random variables, x and y (take care over the signs of x & y). Hence show that, for the case where x and y are uniformly distributed between 0 and 1, $p(z) = \ln(1/z)$.

[PTO]

5. State the central limit theorem, and the conditions under which it is valid. The pdf for the sum of two random variables, x and y, is the convolution of the pdfs $p(x)$ and $p(y)$; use this fact to show that the sum of two Gaussian-distributed variables also has a Gaussian pdf.

   Explain why the Lorentzian pdf $p(x) = A[1 + (x/\sigma)^2]$, with $-\infty < x < \infty$ violates the Central Limit Theorem. What is the constant A? Given two measurements, x and y, drawn from this pdf, calculate the pdf for the sum using Fourier methods – you may assume that the characteristic function for the Lorentzian is $\phi(k) = \exp(-|k|\sigma)$. Calculate the 95% confidence range for $x + y$ and compare this with the 95% confidence range for x and y separately.

6. A model for the integral redshift distribution in a galaxy survey is $P(> z) = \exp(-az^b)$, where a and b are free parameters. Suppose we have only a single redshift, measured at $z = 1$. Show that the maximum-likelihood solution is $a = 1$, $b = \infty$.

   Now suppose that we measure a second redshift, $z = 2.71828\ldots$. Show that the likelihood function based on these two redshifts is

$$\mathcal{L} = a^2 b^2 \exp\left[b - 1 - a - ae^b\right].$$

   Show that the ML solution has to be obtained numerically, and is approximately $(a, b) = (0.1664, 2.3994)$. Prove that the Hessian matrix is

$$\mathbf{H} = -\begin{pmatrix} 2/a^2 & e^b \\ e^b & 2/b^2 + ae^b \end{pmatrix}.$$

   and thus show that the errors on a and b are approximately $(\sigma_a, \sigma_b) = (0.2456, 1.4137)$, with correlation coefficient $r = -0.8779$. Compare with the conditional errors on a and b and draw a sketch to explain why the latter are smaller.