Astronomical Statistics: Tutorial Questions 3

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

October 2012

1. An astronomer attempts to make a number of measurements of the distances and velocities of galaxies. Unfortunately, due to bad weather, only two measurements are possible. The velocities, from Doppler shifts, have negligible errors, but the distances are poorly known. The data are $v_1 = 0$, $r_1 = 1$ and $v_2 = 1$, $r_2 = 2$, in some suitable units. The measurements are independent, and have independent gaussian errors with $\sigma_1 = \sigma_2 = 1$. Use these data to do hypothesis testing, parameter estimation, and model selection, as follows:

   (i) Test the hypothesis that the distances are all zero. How many free parameters n_p does the model have, and hence how many degrees of freedom $n_{data} - n_p$ are there? Calculate χ^2 and compute the probability that χ^2 is *at least as high* as the observed value. You may use the result that the distribution of χ^2 for ν degrees of freedom is

   $$p(\chi^2|\nu) = \frac{e^{-\chi^2/2}}{2\Gamma(\nu/2)} \left(\frac{\chi^2}{2}\right)^{\nu/2-1}. \tag{1}$$

   (NB could also argue that the probability is double your answer, as a very *low* value of χ^2 would be just as remarkable).

   (ii) Parameter Estimation: Model the data by $r_i = mv_i + c$, where m and c are parameters to be determined. Find the most probable values of m and c, stating your assumptions. Estimate the conditional errors on m and c from the curvature of the posterior probability. From the second derivative matrix, estimate the marginal errors on m and c.

   (iii) Model selection: Two models are proposed. M_1 is that $r_i = c$, whereas M_2 is the model in (ii) above. What is the relative probability of the two models? For this you will have to make some choice of priors. Assume flat priors within allowed ranges of c and m which are both large, and express your answer in terms of the ranges Δc and Δm. If $\Delta m = 5$ which model is favoured, and by how much? Comment on your answer.

2. A set of independent measurements $\{x_i, y_i\}$ has independent gaussian errors in *both* x and y. We want to fit a straight line $y = mx + c$ and estimate parameters m and c.

   First consider a single data point x, y with errors σ_x, σ_y. Remember that we don't know the true value of x (call this z), so we have to consider the joint probability of x, y and z, and marginalise over z to get $p(x, y|m, c)$. Show that the likelihood is

   $$p(x, y|m, c) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} dz \exp\left[-\frac{(x-z)^2}{2\sigma_x^2}\right] \exp\left[-\frac{(y - \{mz + c\})^2}{2\sigma_y^2}\right]. \tag{2}$$

   For simplicity now assume that $\sigma_x = \sigma_y = 1$, and show that the probability of m and c is (assume uniform priors)

   $$p(m, c|\{x_i, y_i\}) \propto \prod_i (1 + m^2)^{-1/2} \exp\left[-\frac{(y_i - mx_i - c)^2}{2(1 + m^2)}\right]. \tag{3}$$

3. Explain what Malmquist bias is, and explain how it can fool the unwary astronomer into thinking that stellar or galaxy luminosities measured at different wavelengths can appear correlated, even if there is no true correlation.

4. An astronomical source emits photons with a Poisson distribution, at a rate of λ per second. A telescope detects the photons independently, with probability p. In time t, the source emits M photons, and N are detected. Show that the joint probability of N and M is

$$P(M,N) = \frac{\mu^M}{M!}\, e^{-\mu} \, \frac{M!}{N!(M-N)!} p^N q^{M-N} \tag{4}$$

where $\mu = \lambda t$ and $q = 1 - p$.

Marginalise over M to show that

$$P(N) = \frac{p^N q^{-N} e^{-\mu}}{N!} \sum_{M=N}^{\infty} \frac{(q\mu)^M}{(M-N)!} \tag{5}$$

Sum the series to show that N has a Poisson distribution with expectation value $p\mu$. Why could this have been anticipated?

Calculate the probability that the source has emitted M photons given that N have been detected, $P(M|N)$, for $M \geq N$, and deduce that $M - N$ also has a Poisson distribution, and compute the expectation value for $M - N$.