Session 2010-11

SUPA Advanced Data Analysis: Problems Sheet

With thanks to Graham Woan and Phil Gregory

1. (a) Explain the circumstances in which the Poisson distribution

$$p(r) = \frac{\mu^r e^{-\mu}}{r!}$$

correctly describes the probability of r events occurring, and identify the meaning of the parameter μ in this equation.

(b) There is a constant probability that a fire will occur at any time in Glasgow, and there are two fires per day on average. Write down an expression for the probability, $p(n)$, of n fires occurring on any one day.

(c) To deal with any fire requires the presence of a fire engine for a whole day. What is the minimum number of fire engines required in Glasgow so that the probability of all fires on a given day are attended to is better than 99%?

2. The distribution of gamma ray bursts has been shown to be uniform on the sky – i.e. the probability of a gamma ray burst occurring in solid angle $d\Omega$ is simply proportional to $d\Omega$, where the constant of proportionality is independent of direction on the sky.

(a) Given that $d\Omega = \cos\beta d\beta d\alpha$, where α and β are galactic longitude and latitude respectively, determine the marginal pdfs in galactic longitude and galactic latitude of the gamma ray burst distribution.

(b) There exists a function $y = f(\beta)$, of galactic latitude such that the pdf of y is a uniform distribution between -1 and $+1$. What is the function $f(\beta)$?

3. The number r of events detected per hour by a particle physics detector is modelled as a Poisson variable with pdf

$$p(r|\mu) = \frac{\mu^r\, e^{-\mu}}{r!}$$

where μ is the average number of events per hour.

(a) By differentiating the natural logarithm, ℓ, of the likelihood function, show that – if r events are detected in a given hour of operation – the maximum likelihood estimate of μ based on these data is simply $\hat{\mu}_{\rm ML} = r$.

(b) The detector operates continuously for n hours, with the number of events detected in each hour denoted by $r_i\,;\ i = 1, ..., n$. Show that the maximum likelihood estimate of μ, based on the entire n hour run, is

$$\hat{\mu}_{\rm ML} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

stating clearly any assumptions that you make.

4. The distribution of X = log temperature for a population of spectroscopic sources is modelled to be Gaussian in form. X is measured for a sample of 16 sources with the following results (in suitably scaled units)

$$\sum x_i = 51.2 \qquad \sum x_i^2 = 243.19$$

Test the hypothesis that μ, the population mean log temperature, is equal to 4.0:

(a) assuming that σ, the population standard deviation, is known to be 1.9

(b) when σ is not known *a priori* and must be estimated from the sample data.

Suggest why (a) is the better hypothesis test *if* σ is known.

5. The sampled distribution of 100 background radiation measurements in a radioactively contaminated site are given in the table below. Construct a χ^2 goodness of fit test to see if you can reject the hypothesis at the 95% confidence level that the counts have a Poisson distribution.

count obtained	0	1	2	3	4	5	6	7	8	9	10	11	12
no. of occurrences	1	6	18	17	23	10	15	4	4	1	0	0	1

6. N observations, $x_k; k = 1, ..., N$, of the flux density of a quasar are affected by interstellar scintillation which introduces Gaussian errors of (unknown) variance σ^2.

(a) Explain what is meant by the *likelihood* of these data and show that, if the measurements are independent, the likelihood is

$$p(x_k|\mu, \sigma, I) = \left(\sigma\sqrt{2\pi}\right)^{-N} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k - \mu)^2\right],$$

where μ is the true flux density of the quasar.

(b) Explain the importance of the joint posterior pdf of μ and σ for parameter estimation. What is the meaning of the *marginal* posterior pdf for μ alone? Show that if the priors for μ and σ are uniform for positive values, and zero otherwise, the marginal posterior pdf for μ is

$$p(\mu|x_k, I) \propto \int_0^\infty t^{N-2} \exp\left[-\frac{t^2}{2}\sum_{k=1}^{N}(x_k - \mu)^2\right] \mathrm{d}t$$

where $t = 1/\sigma$.

(c) Determine the un-normalised value of this integral (i.e. without evaluating the constant of proportionality) given the standard result

$$\int_0^\infty x^n \exp\left(-ax^2\right) \mathrm{d}x \propto a^{-(n+1)/2}.$$

(d) By examining the maximum of $L = \ln\left[p(\mu|x_k, I\right]$, show that the maximum likelihood estimate for μ is

$$\mu_0 = \frac{1}{N}\sum_{k=1}^{N} x_k,$$

and that the uncertainty in this estimate is $S/\sqrt{N}$ where

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_0)^2 .$$

(e) Comment on how this result compares to the situation where σ is known.

7. The following data are used to fit a linear model to the relationship between the variables x and y:

x	2.71	2.05	2.67	2.23	2.36	2.52	2.91	2.43	2.27	2.84
y	-21.1	-19.2	-20.6	-19.4	-20.0	-20.2	-21.5	-19.8	-19.2	-20.9

(a) Use the method of least squares to determine the equation of the best-fit straight line under the linear model:

$$y_i \quad = \quad a + b\,x_i + e_i \tag{1}$$

assuming that the residuals, e_i, are normally distributed with mean zero and dispersion, $\sigma = 0.17$.

(b) Determine errors for the least squares estimates of a and b.

(c) Construct a χ^2 test to determine if these data give an acceptable fit to the linear model

8. A coin is tossed n times and a binomial model is adopted to describe the probability of obtaining r heads, i.e. the data are described by the likelihood:

$$p(r|\theta) \propto \theta^r\,(1-\theta)^{n-r}; \qquad 0 < \theta < 1$$

where θ is the probability of obtaining a head on any given toss of the coin and the constant of proportionality does not depend on θ.

(a) Write down an expression for the natural logarithm, $\ell(\theta)$, of the likelihood.

(b) By differentiating $\ell(\theta)$, show that the **maximum likelihood** estimate of the parameter θ is $\hat{\theta}_{\rm ML} = \frac{r}{n}$.

(c) A sequence of coin tosses is analysed within a Bayesian framework to make inferences about the value of θ, using Bayes' formula in the form

$$p(\theta|r) \propto p(r|\theta)p(\theta)$$

where $p(\theta)$ is a distribution describing our prior assumptions about the value of θ. Explain why, if we adopt a *uniform* prior for θ over the range $0 < \theta < 1$, then the maximum of the posterior probability distribution function for θ is again equal to r/n.

(d) If the coin is 'fair' one should expect that $\theta = 0.5$. Suppose we have a strong prior belief that our coin is fair, and we adopt a prior of the form:

$$p(\theta) \propto \left[1 - 4(\theta - 0.5)^2\right]$$

By writing down the natural logarithm of the posterior, and differentiating with respect to θ, show that with the above prior the maximum posterior probability occurs at a value of θ that is a solution of the following equation

$$r(1-\theta)\left[1-4(\theta-0.5)^2\right]-(n-r)\theta\left[1-4(\theta-0.5)^2\right]-8(\theta-0.5)\theta(1-\theta)=0.$$

(e) Assuming $r=1$ and $n=4$, make a plot (e.g. with excel) showing how the above equation changes as a function of θ. Show that a zero occurs at $\theta\sim 0.33$.

(f) Make the same plot but now for $r=248$, $n=1000$, and show that the zero now occurs at $\theta\sim 0.25$ – i.e. in agreement with the maximum likelihood estimate from part (b).

(g) Can you explain why the maximum of the posterior agrees with the maximum likelihood estimate in part (f), but not in part (e)?

9. Fit a straight line model to the data given in the table below, where $\overline{y}_i$ is the average of n_i data values, each measured at x_i. The probability of the individual y_i measurements is assumed to be normal with $\sigma=4.0$, regardless of the value of x_i.

x_i	$\overline{y}_i$	n_i
10	0.387	14
20	5.045	3
30	7.299	25
40	6.870	2
50	16.659	3
60	13.951	22
70	16.781	5
80	20.323	2

(a) Give the slope and the intercept of the best-fit line, together with their errors.

(b) Plot the best-fit straight line together with the data values and their error bars.

(c) Give the parameter covariance matrix

(d) Repeat (a) and (c), but this time use the average x-coordinate as the origin. Comment on the differences between the covariance matrices in (c) and (d).

10. An X-ray telescope observes a 'blank' area of sky, in order to estimate the X-ray background rate, and counts n X-ray photons in a time T. The likelihood of this observation follows the Poisson distribution,

$$p(n|b,I)=\frac{(bT)^n e^{-bT}}{n!}.$$

(a) Taking b to be a scale parameter, assign it an appropriate prior and determine the normalised posterior for b. You will need to use the standard integral

$$\int_0^\infty x^m e^{-ax}\mathrm{d}x=\frac{m!}{a^{m+1}}\quad(a>0;m=0,1,2,...)\,.$$

(b) Show that the mean of this posterior distribution is n/T, and that its standard deviation is the mean divided by $\sqrt{(n)}$.

(c) Repeat this analysis using a uniform prior for b. Do the two results differ substantially?