Glen Cowan
RHUL Physics
15 December, 2011

Statistics Problems for the NIKHEF Onderzoekschool Subatomaire Fysica (part 2)

Exercise 4: Consider a likelihood $L(x|\theta)$, which gives the probability for the data x given a parameter θ. The Jeffreys prior for θ is given by

$$\pi(\theta) \propto \sqrt{I(\theta)}\ , \tag{1}$$

where

$$I(\theta) = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] \tag{2}$$

is the expected Fisher Information.

(a) Show that

$$-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]\ , \tag{3}$$

providing that the set of the allowed values of x does not depend on θ.

Hint: Write the right-hand side of (3) as

$$E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = \int \left[\frac{\partial}{\partial \theta}\left(L\frac{\partial \ln L}{\partial \theta}\right) - L\frac{\partial^2 \ln L}{\partial \theta^2}\right] dx \tag{4}$$

and use the fact that one can bring the derivative $\partial/\partial\theta$ outside of the integral as long as the region of integration does not depend on θ. Also use the fact the the integral of $L(x|\theta)$ over all x is equal to unity for any θ.

(b) Suppose one uses the Jeffreys prior for θ to obtain the posterior pdf

$$p(\theta|x) \propto L(x|\theta)\pi(\theta)\ , \tag{5}$$

Suppose now that one transforms to a new parameter $\eta(\theta)$, such that the posterior pdf for η is

$$p(\eta|x) = p(\theta|x)\left|\frac{\partial \theta}{\partial \eta}\right|\ . \tag{6}$$

Show that one arrives at the same posterior pdf as (6) by beginning directly from the Jeffreys prior for η. That is, inference made using the Jeffreys prior is invariant under a parameter transformation.

Exercise 5: Consider an experiment where one measures a number of events n, which is modeled as following a Poisson distribution with mean $s+b$, where s and b are the contributions from signal and background processes, respectively. To constrain the parameter b, one carries out a control measurement that counts a number of events m, which follows a Poisson distribution with mean τb, where τ is a known scale factor.

The problem thus contains a single parameter of interest, s, and a nuisance parameter b. The likelihood can be written

$$L(n,m|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b} \ , \tag{7}$$

Show that the Maximum Likelihood (ML) estimators for s and b are

$$\hat{s} = n - m/\tau \ , \tag{8}$$

$$\hat{b} = m/\tau \ , \tag{9}$$

and that the conditional ML estimator for b given s is

$$\hat{\hat{b}}(s) = \frac{m+n-(1+\tau)s + \sqrt{(m+n-(1+\tau)s)^2 + 4(1+\tau)ms}}{2(1+\tau)} \ . \tag{10}$$

The quantities $\hat{s}$, $\hat{b}$ and $\hat{\hat{b}}$ are what we require to compute the profile likelihood ratio

$$\lambda(s) = \frac{L(s,\hat{\hat{b}})}{L(\hat{s},\hat{b})} \ . \tag{11}$$

(b) Write a Monte Carlo program that generates values of n and m according to hypothesized values of s and b (e.g., use $b=20$, $\tau=1$ and $s=0$) and use these to evaluate the profile likelihood ratio. (See the program `runSigCalc_MC.cc` on the course webpage for examples of how to generate Poisson distributed values by using the ROOT class `TRandom3`.)

To carry out a test of $s=0$, we can use the statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} \geq 0 \ , \\ 0 & \hat{s} < 0 \ . \end{cases} \tag{12}$$

Generate the distribution of q_0 assuming $s=0$, $b=20$ and $\tau=1$. In the large sample limit, this should approach a "half-chi-square" distribution for one degree of freedom (a delta function at zero plus a chi-square distribution, each with a weight of one half). Check to what extent this holds for different values of b (e.g., $b=2,20,200$).

(c) Extend the Monte Carlo program from (b) to compute the distribution of q_0 by generating data with a nonzero value of s, e.g., take $b=20$, $\tau=1$, $s=10$. Find the median value of q_0 under assumption of this value of s, and thus find the median discovery significance for this s (i.e., the discovery sensitivity). Compare this to the value based on the Asimov data set,

$$\mathrm{med}[Z_0|s,b] \approx \sqrt{2\left((s+b)\ln(1+s/b) - s\right)} \ . \tag{13}$$