Advanced methods in statistical data analysis

Examples for Thursday

Part A: Inference

1. Bayes theorem states

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

After a particularly long party session, you wake up with no idea what day it is. You recall that your friend always goes to the pub on Saturdays and Sundays, but only once during the week on a random day. You glance through the pub window and see them sitting there. What is the probability that it is the weekend?

How is your conclusion changed if you feel certain the day must be Thursday, Friday or Saturday?

2. A friend says she will throw a regular die (with sides numbered one to six as usual) out of your sight, but doesn't tell you how often. She then tells you that the sum of the throws is 5. What is the most probable number of times the die was thrown? How much more probable is this number of throws than the next best answer?

3. [From lecture] You've been to the doctor's with a broken wrist. She decides to give you a blood test. It comes back indicating you are HIV-positive. The doctor tells you the test is 95% reliable. What is your chance of having HIV?

How would your answer change had you been to the doctor's because of swollen lymph glands?

Part B: Parameter estimation

4. In running a Metropolis–Hastings MCMC calculation, the user has to specify the *proposal function*, which determines the probability of jumping to any other point. (NB: this is distinct from the comparison of old and new likelihood, which decides whether the proposed jump is accepted or not.) The efficiency of an MCMC calculation depends on a good choice for this function. Discuss possible consequences of a poor choice of proposal function. Supposing that in advance you know nothing about the shape of the likelihood function, what strategy would you suggest?

5. The Metropolis–Hastings algorithm involves the following steps:

(i) Select a random starting point in parameter space. Obtain the posterior probability $\mathcal{P}_{\rm old}$.

(ii) Propose a random jump to a new location using a suitably defined *proposal function*, and measure the likelihood $\mathcal{P}_{\rm new}$.

(iii) If $\mathcal{P}_{\text{new}} > \mathcal{P}_{\text{old}}$, accept the jump. Otherwise accept the jump with probability $p = \mathcal{P}_{\text{new}}/\mathcal{P}_{\text{old}}$. If the jump is declined, stay at the original point and duplicate its entry in the Markov chain.

(iv) Take the current point's posterior as $\mathcal{P}_{\text{old}}$ and return to step (ii), continuing until satisfied (or bored).

If the priors are uniform, then the posterior can be replaced by the likelihood.

A model features parameters x and y, which have uniform priors in the range [0,10]. The likelihood is a gaussian of unit width in each direction, centred at $(x, y) = (3, 4)$. Write a Metropolis–Hastings sampler to explore the posterior distribution, bearing in mind your conclusions on the proposal function in Q4. [The easiest would be to take the proposal function to be a uniform distribution; if you feel more ambitious you could try a gaussian.] Having run a chain, discard the first 100 elements to try and eliminate 'burn-in'.

Using your code, provide estimates of

a) How many steps are needed to find a point whose likelihood is within 10% of the maximum.

b) The number of chain elements necessary to estimate the best-fitting x to an accuracy of 0.2 or better.

b) The accuracy of the best-fit value of x obtained from a chain of 10^4 elements.

c) The number of chain elements required to obtain an accuracy of 0.1 on the *uncertainty* of x.

If you have time after undertaking this task, you could consider investigating the dependence of one or more of these answers on the width of your proposal function.