STA 214 Assignment 1 Spring 1997

Readings: Prior to Tuesday next ....

(a) Chapter 2 of Gelman et al - the basics of probability calculus in 'standard' Bayesian models. Appendix A to the book is a good catalogue of distributions, notation, etc. We'll get to multivariate distributions later.
(b) Mixtures: Titterington et al handout.
(c) S-Plus, manual, help-files, and ISDS intro document (from the Computer Help Web pages).

Please hand in solutions to the following to Yang Chen, the TA, next Thursday morning.

Exercise 1

Amongst diseased individuals ($\theta = 1$), measured blood levels x of a biochemical indicator are distributed as $(x|\theta = 1) \sim N(x|m_1, \tau^2)$. Amongst healthy individuals ($\theta = 0$), the distribution is $(x|\theta = 0) \sim N(x|m_0, \tau^2)$. Also $m_1 > m_0$, levels being higher amongst diseased people. The disease incidence rate is assumed known at 0.01, so that θ has pdf $p(\theta)$ given by $p(1) \equiv Pr(\theta = 1) = 0.01$ and $p(0) \equiv Pr(\theta = 0) = 0.99$.

For a randomly chosen individual, use Bayes' theorem to find, as a function of the test result x, the *posterior odds on disease* $O(1|x) = Pr(\theta = 1|x)/Pr(\theta = 0|x)$. For a fixed number $k \quad (>> 1)$, show that the posterior odds exceed k times the prior odds of 1:99 when

$$x > \bar{m} + \log(k)\tau^2/(m_1 - m_0),$$

where $\bar{m} = .5(m_0 + m_1)$. Interpret this result, and comment on how the practical effectiveness of the test depends on the size of $m_1 - m_0$. Graphs will help.

Exercise 2

Of a large population of manufactured items, an unknown (small) proportion π are defective. We want to estimate π based on a random sample of n items for the population. Our quality control system checks each item and records data x_i where $x_i = 1$ if the $i-$th item is classified as OK, $x_i = 0$ if classified as defective. Items are classified independently. If the system worked perfectly, we'd have independent Bernoulli trials with success probability $Pr(x_i = 1|\pi) = 1 - \pi$, and so a standard inference problem (first several pages of Chapter 2 of Gelman et al). But the system is fallible; it miss-classifies items 1% of the time.

(a) Prove that the x_i are independent Bernoulli trials with success probability

$$Pr(x_i = 1|\pi) = 0.99(1 - \pi) + 0.01\pi.$$

(b) Take a uniform prior $\pi \sim U(\pi|0, 1)$. Suppose we make just one test, $n = 1$. Write down the explicit form of the posterior density for π in each of the two possible cases $x_1 = 1$ and $x_1 = 0$.
(c) What is the form of the posterior density based on n independent items/observations?

Exercise 3

Parts (a) and (b) of Excercise 10 on p60 of Gelman et al. Use S-Plus (or Matlab, or whatever); it is easy in S-Plus (my code is below). I suggest a grid of $m = 500$ points, at least. Explore different m in building these kinds of basic 'discrete' numerical approximations. Part (b) is to be done by pretending that the discrete approximation to $p(\theta|y)$ from (a) is exact, and then simply randomly sampling from that discrete distribution. (In S-Plus, this is easy using one of several methods, including the `sample` function in S-Plus). Compare the histogram with the "true" posterior density computed in (a).

Repeat this exercise with new "data". First, generate a "true" value of θ from the uniform $U(\theta|0, 1)$ prior. Then generate a sample of 5 new observations from the Cauchy distribution with that θ value as mode. Use this new data to repeat the exercises (a) and (b). The S-Plus functions `runif` and `rcauchy` are useful.

I have written some S-Plus code to get you started on this. – the analysis with the newly generated data is below. Figure out what each step does (e.g., by trying it and checking the help pages). Run it a few times, learn more S-Plus and explore Cauchy models. Experiment with other distributions than Cauchy, e.g., normal, Student T, etc.

```
motif()
help.start()
truetheta← runif(1)
y←truetheta+rcauchy(5)   (or y←rcauchy(5,truetheta))
m←500
theta←seq(0,m)/m
post←rep(1,m+1)
for (i in 1:5) post←post/(1+(y[i]-theta)*(y[i]-theta))
x←sample(theta,1000,replace=T,p=post)
hist(x,nclass=25,probability=T)
lines(theta,m*post/sum(post),col=2)
```