STAT 740 - Spring 2004 - Homework 7 *(A non-EM problem)*

Due: Monday, April 19th

For testing a variety of models it is often desired to generate Bernoulli random variables Y_1 and Y_2 that have a specified correlation coefficient ρ_B. For example the contingency table:

	$Y_2=1$	$Y_2=0$
$Y_1=1$	80	50
$Y_1=0$	48	37

has correlation 0.05048267. Given a particular table it is fairly easy to find the correlation, you simply need to make the vectors of zeros and ones and take the correlation, or better yet use the formula:

$$\hat{\rho}_B \frac{p_{12} - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} = \frac{\frac{80}{215} - \left(\frac{130}{215}\right)\left(\frac{128}{215}\right)}{\sqrt{\left(\frac{130}{215}\right)\left(\frac{85}{215}\right)\left(\frac{128}{215}\right)\left(\frac{87}{215}\right)}}.$$

It is a bit more complicated to go in the other direction and create a table that is expected to have a given correlation ρ_B. One method of doing this would be to generate a bivariate normal random variable with a certain correlation ρ_N and then convert the two normals to Bernoullis by comparing them to the appropriate cut-off values. The question then is, what is the relationship between ρ_N and ρ_B.

a) Verify that the above formula does indeed give you Pearson's product moment correlation for Bernoulli random variables.

b) Given an N(0,1) random number generator, how can you generate b(1,p) random variables?

c) Consider generating $(X_1,X_2)\sim$ bivariate normal ($\mu_1=0$, $\mu_2=0$, $\sigma_1=1$, $\sigma_2=1$, ρ_N) for various values of ρ_N and converting the resulting normals into Bernoulli random variables with p=0.6 and 0.7 respectively. Use a simulation to create a plot demonstrating the relationship between ρ_N =0.1, 0.2, … 0.9 and the resulting correlations (ρ_B's) of the Bernoullis. Provide appropriate error bars for your estimated ρ_B's.

d) It is possible to exactly calculate the ρ_B in (c) instead of using simulation. This can be done by finding the integrals of the bivariate normal that correspond to p_{12}, p_1, and p_2, and then applying the above formula. Replicate the plot in (c) using these exact values.

Note: The function `adapt` in `library(adapt)` will perform multivariate integration… you need to use the `Install package(s) from CRAN...` menu option to get it.

e) Your work in part (d) gave you a set of equations relating the two cut-off values and ρ_N to the two Bernoulli percentages and ρ_B. Create R-code (using `optim`?) to determine the ρ_N and two cut-off values for given Bernoulli percentages and ρ_B. (Do you need to put limits on the possible values of ρ_B?)

f) Choose a target value of p_1, p_2, and ρ_B and demonstrate that your function returns the appropriate cut-offs and ρ_N.

g) Why will the method you found also work for generating *n*-tuples of correlated Bernoullis instead of simply pairs.