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 $\rho_B$. For example the contingency table:

<table>
<thead>
<tr>
<th></th>
<th>$Y_2=1$</th>
<th>$Y_2=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1=1$</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>$Y_1=0$</td>
<td>48</td>
<td>37</td>
</tr>
</tbody>
</table>

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:

$$\hat{\rho}_B = \frac{p_{12} - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}$$

$$= \frac{80}{215} - \frac{(130)(128)}{215}$$

$$= \frac{215}{215} - \frac{215}{215}$$

It is a bit more complicated to go in the other direction and create a table that is expected to have a given correlation $\rho_B$. One method of doing this would be to generate a bivariate normal random variable with a certain correlation $\rho_N$ 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 $\rho_N$ and $\rho_B$.

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 $(\mu_1=0, \mu_2=0, \sigma_1=1, \sigma_2=1, \rho_N)$ for various values of $\rho_N$ 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 $\rho_N = 0.1, 0.2, \ldots 0.9$ and the resulting correlations ($\rho_B$’s) of the Bernoullis. Provide appropriate error bars for your estimated $\rho_B$’s.

d) It is possible to exactly calculate the $\rho_B$ 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 $\rho_N$ to the two Bernoulli percentages and $\rho_B$. Create R-code (using optim?) to determine the $\rho_N$ and two cut-off values for given Bernoulli percentages and $\rho_B$. (Do you need to put limits on the possible values of $\rho_B$?)

f) Choose a target value of $p_1$, $p_2$, and $\rho_B$ and demonstrate that your function returns the appropriate cut-offs and $\rho_N$.

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