Homework for BDA chapter 8

1. Show that strong ignorability implies ignorability.

2. You are consulting for a company that makes a product called a widget, and sells it on Amazon. The company has sold widgets to 5138 customers on Amazon, and a subset of customers have posted satisfaction ratings on Amazon. Suppose the ratings are integer values: 1, 2, 3, 4, or 5. The company is interested in knowing the distribution of satisfaction ratings of its customers.

   (a) Explain, in words, why using only the posted satisfaction ratings (and ignoring the "data collection process") might be misleading.

   (b) Set up a probabilistic model for this problem. Assume each customer has a potential rating from 1 to 5, and given the rating, has some probability (depending on the rating) of posting it on Amazon. Put conditionally-conjugate priors on the parameters for the potential ratings and the posting probabilities (hint: use a Dirichlet and 5 Betas). For simplicity, assume the rating of each customer is accompanied by the customer index (1, 2, 3, ...).

   (c) You convince the company to conduct a survey to gather some more data. A subset of 20 customers are selected uniformly at random, and are asked (A) whether they posted a rating, and (B) what rating they would give (or did give) the widget on Amazon. To clarify the setup, attached (widgets.txt) is all the data: the first column contains the ratings, with 0 indicating an unobserved value, the second column contains indicators of which customers posted their ratings on Amazon, and the third column contains indicators of which customers were surveyed. Add random variables to your model for the survey indicators.

   (d) Consider the joint posterior distribution of all the unknown quantities—specifically, of (i) the probability of each rating, (ii) the posting probabilities, and (iii) all of the unobserved ratings. Derive the full conditionals required to implement a Gibbs sampler for this joint posterior.

   (e) Implement your Gibbs sampler and run it on the data. Make box plots of the posteriors of each of the rating probabilities, and (separately) each of the posting probabilities. On the plots, compare your results to the following

"true" values (these were used to generate the data):

$$\text{distribution of ratings} = (.2, .1, .1, .4, .2)$$
$$\text{posting probabilities} = (.9, .2, .2, .1, .4).$$

(f) Compare your results with the following two alternate approaches:

i. The naive approach of just using the ratings posted on Amazon and ignoring the data collection process.

ii. Just using the 20 ratings obtained from the survey.

Make box plots of the posteriors on the rating probabilities for these two approaches as well, and discuss.

3. BDA exercise 8.11, with the following additional instructions. Let m denote the number of fish tagged on the first fishing trip, and let L denote the number of fish that had to be caught in order to get k tagged fish on the second fishing trip (e.g., $m = 100$, and $L = 90$, and $k = 20$ is what occurred in the scenario described). We will treat m and k as fixed and known.

(a) There are two natural ways of defining "the" posterior on N in this problem. One is $N|L$. The other is the posterior given the particular binary sequence $t_1, \ldots, t_L$ indicating whether each fish caught was tagged or not.

i. Derive a closed form expression for the probability of observing a particular sequence $t_1, \ldots, t_L$ given N.

ii. Derive a closed form expression for the distribution of L given N. This distribution has a well-known form (although it might be new to you). What is it called?

iii. Are $N|t_1, \ldots, t_L$ and $N|L$ the same?

iv. What is the asymptotic behavior of $p(L|N)$ as $N \to \infty$? More precisely, what is the simplest function $f(N, L, k, m)$ you can find such that $p(L|N)/f(N, L, k, m) \to 1$ as $N \to \infty$? Your answer should work for any $L, m, k \geq 0$ such that $m \geq k$ and $L \geq k$.

v. Consider priors of the form $p(N) \propto (1/N^r)\mathbb{1}(N > 0)$. Using your result from the previous part, what is the smallest (nonnegative) integer value of $r = r(m, k)$ for which the posterior $p(N|L)$ is proper for any L?

(b) (Skip.)

(c) (To clarify, please give an expression for the posterior predictive probability that the next fish is tagged, given L.)

(d) Write down a reasonable model to handle this missing data problem, however, you do not need to give the posterior distribution. (I believe there is probably more than one way to model this—there is not one "right" answer).