Bayesian Statistics
Exercises 3.

1. The Dirichlet distribution is a generalization of the Beta distribution. A k-tuple $(\theta_1, \theta_2, \ldots, \theta_k)$ is said to have a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ if its (joint) probability density function is

$$p(\theta_1, \theta_2, \ldots, \theta_k) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1},$$

provided $\theta_1 + \cdots + \theta_k = 1$, and $p(\theta_1, \theta_2, \ldots, \theta_k) = 0$ otherwise.

   i) If $(\theta_1, \theta_2, \ldots, \theta_k)$ has a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, find $\mathrm{E}(\theta_i)$, $\mathrm{var}(\theta_i)$, and $\mathrm{covar}(\theta_i, \theta_j)$ for $i, j = 1, 2, \ldots, k$, $i \neq j$. For fixed $\mathrm{E}(\theta_i)$, $i = 1, 2, \ldots, k$, how do the qualitative properties of the distribution depend on its parameters?

   ii) Let $X_1, \ldots, X_n$ denote independent and identically distributed random variables on $\{1, \ldots, k\}$, with $\Pr(X_i = k) = \theta_k$, Assume that the prior distribution for $\theta = (\theta_1, \ldots, \theta_k)$ is Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

      a) Derive the posterior distribution for $\theta | X_1, \ldots, X_n$.

      b) Derive the predictive distribution $p(X_n | X_1, \ldots, X_{n-1})$.

2. Using the above results on the Dirichlet, we now return to the exercise in `exercises/seeb/train_test.R`. In that exercise you were told, after step 1, to "Assume for the remainder of this exercise that these allele frequencies from the training set are the "true" frequencies in each population." Some of you may have noticed that zeros in allele counts in the training set then create problems, and perhaps solved this problem by adding pseudo counts (in which case you have already made a good start on this exercise). In this exercise you are now asked to revisit the exercise using a more complete Bayesian procedure, using a Dirichlet$(\alpha, \ldots, \alpha)$ prior distribution for the frequencies of the alleles at each marker/locus. Explicitly:

   a) In step 2, for each sample j in the test set, implement and apply a method to compute (given α) the posterior probability

$$\Pr(\text{sample } j \text{ came from population } k | G_j, \text{Training data}, \alpha),$$

   where G_j denotes the genetic data on test sample j.

b) Tabulate and compare the error rate (step 3) for different values of α. You should consider $\alpha = 0.5$ and $\alpha = 1$, as well as a "very small" and a "very big" value for α.

c) Derive the likelihood for α given the training set data, $L(\alpha) :=$ Pr(Training data$|\alpha$), and obtain the maximum likelihood estimate for α for these training data.

d) Compute the error rate (step 3) using the maximum likelihood estimate of α (effectively this is the "Empirical Bayes" approach), and compare it with the other values you obtained.

e) The Empirical Bayes approach could be criticized for failing to fully reflect uncertainty in α. An alternative would be to put a prior distribution on α, and perform fully Bayesian inference, integrating over the uncertainty in α. Perform this analysis, using a uniform discrete prior on $\alpha \in \{0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$. Again, compare the error rate for this "fully Bayes" analysis with the other approaches.

[Note that the above approaches, consider classifying test data one at a time given *only the information from the training data*. In practice, if we had access to multiple test samples (of unknown origin) we should also use any information they give us when attempting to classify samples. That is, we would compute

Pr(sample j came from population $k|$All test data genotypes, Training data genotypes and p

However, this full analysis is beyond us for now.]

3. Read the chapter by Krebs entitled "Independence, Exchangeability, and de Finetti's Theorem". Do exercise 1 at the end.

4. Consider throwing a thumb tack repeatedly. Find a thumb tack, but *do not throw it yet*. With heads and tails as described in Krebs' chapter (tail means point uppermost), what is your prior for θ, the long run proportion of heads?

   Sketch a graph of your prior.

   The appropriate conjugate prior here is the family of Beta distributions. Choose one, or a mixture of two, Beta distributions which (roughly) approximate your prior. (Don't waste time trying to get a very good

approximation.) In the remainder of the question I will refer to this as your approximate prior.

Sketch (or plot) a graph of your approximate prior.

Toss your tack 100 times under similar conditions and record the results.

Using your approximate prior as the prior, calculate and then sketch the posterior after $n = 1$, 10, 50, and 100 tosses.

What is your predictive probability, based on the first n tosses, that the $(n+1)$st toss would result in a tail, for these values of n?

My prior is an equal mixture of the Beta(2,4) and the Beta(4,2) distributions. Using my prior as the prior, calculate and then sketch the posterior after $n = 1$, 10, 50, and 100 tosses.

What is the predictive probability, based on the first n tosses, that the $(n+1)$st toss would result in a tail, for these values of n?

Comment.

5. Verify the claim in lectures that an exchangeable sequence X_1, X_2, with

$$\mathrm{P}(X_1 = 1, X_2 = 0) = \mathrm{P}(X_1 = 0, X_2 = 1) = \frac{1}{2},$$

cannot be embedded in an exchangeable sequence of length three.

6. Show that if $X_1, X_2, \ldots$ is an infinite exchangeable sequence of 0-1 valued random quantities, then for any pair i, j, $\mathrm{Cov}(X_i, X_j) \geq 0$.

In fact this result is true for any infinite exchangeable sequence. Prove this also.