Bayesian Statistics
Exercises 3.

1. Jim Berger introduced the following medical diagnosis problem. Let D denote the event that a patient has the disease. Consider a patient sampled from a population where the disease prevalence is p_0, so $p_0 = \Pr(D)$. Consider a diagnostic test which can be positive (P) or negative (N). Let $p_1 := \Pr(P|D)$ and $p_2 := \Pr(P|notD)$. Let θ denote the probability that the patient has the disease given that the test comes out positive $\theta := \Pr(D|P)$.

   (a) Show that
   $$\theta = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}. \tag{1}$$

   (b) Berger considers the situation that the p_i are unknown, but estimated from (independent) data $X_i \sim Bin(n_i, p_i)$. Assume a $Beta(a, a)$ prior for p_i. Implement an R function that peforms Berger's following procedure for sampling from the posterior of θ given $n_1, n_2, n_3, X_1, X_2, X_3$ and a, and uses it to return an equal-tailed $100(1 - \alpha)\%$ Credible Interval (CI) for θ.

      i. sample p_i from its posterior given X_1, X_2, X_3, a.

      ii. compute θ from (1).

      iii. repeat this lots of times and use upper and lower percentiles of generated θ to produce the CI.

   (c) Use your procedure with $a = 0.5$ to try to replicate Berger's Table 1.

$n_0 = n_1 = n_2$	(x_0, x_1, x_2)	95% CI
20	(2,18,2)	(0.107,0.872)
20	(10,18,0)	(0.857,1.000)
80	(20,60,20)	(0.346,0.658)
80	(40,72,8)	(0.808,0.952)

Table 1: Berger's Table 1

(d) Assess sensitivity to a by producing the same table with $a = 1, 2$.

(e) Berger claims that the credible intervals from this procedure, with $a = 0.5$, also have good frequentist coverage properties.

i. Write a function that uses simulation to check frequentist coverage for any given n and p vectors. Show results for $n_0 = n_1 = n_2 = 20$ and several (at least three) different values for (p_0, p_1, p_2).

ii. Use your function to compare frequentist coverage properties for $a = 1, 2$ with $a = 0.5$.

iii. Explore how coverage is affected if p_i is at or near the boundaries (0,1)?

2. 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 $\text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

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

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

3. Suppose x has a Poisson distribution with unknown mean θ. Determine the conjugate prior, and associated posterior distribution, for θ. Determine the Jeffreys prior π^J for θ.

4. If $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ has a multinomial distribution with parameters n (fixed and known) and $\mathbf{p} = (p_1, \ldots, p_{k-1})$ (which is unknown), find the Jeffreys prior for $\mathbf{p}$.

5. 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$(\alpha, \dots, \alpha)$ 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, $L(\alpha) := \Pr(\text{Training data}|\alpha)$. Implement a method to compute the likelihood for any give α, and use numerical methods to 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 what is known as the "Empirical Bayes" approach, which we will study later), 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

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

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