Stat 111 Spring 2025 Week 9: Bayesian Inference

1. **Binary Decision Problems**

   a) "The two envelopes paradox" is a classic example of a binary decision problem (Blitzstein 9.1.6). Imagine two wealthy-looking visitors recruit you and a friend to give them a tour of the college. As a reward, the visitors present you with two envelopes and tell you that one envelope contains twice as much money as the other. You and your friend flip a coin to assign the envelopes and you see that yours contains \$100. You realize your friend has either \$50 or \$200, and because you flipped a coin to decide, you figure the expected value of your friend's amount is $(1/2)(50) + (1/2)(200) = \$125$. So you're thinking you'd like to switch envelopes.
   Meanwhile, your friend hasn't looked yet, but she figures if she has X dollars, then you have $X/2$ or $2X$, for an expected value of $(1/2)X/2 + (1/2)(2X) = 1.25X > X$. So she'd like to switch envelopes too. Explain the flaw in this reasoning, using θ to represent the larger amount and X to represent the amount in your envelope (and Y the amount in your friend's envelope).

   b) Give a Bayesian solution to the 2-envelopes paradox. How would the prior change if this were a gameshow prize rather than a tip? Identify the prior specification that results in the paradox of both friends wanting to switch, then switch back, and so on. Consider Exponential and Uniform as possible prior distributions for θ, and work out the posterior probabilities for θ given the observed value $X = x$. Discuss the difference between deciding based on the higher probability or based on the posterior mean of your friend's envelope value.

   c) It would be a paradox if you could decide to switch without learning X, the value in your envelope. With this information, you could make a decision in the following way: Independent of your observed X, choose a random value T from any distribution with support that includes some or all of the interval $(\theta/2, \theta)$ (e.g., Exponential). If you observe $T > X$, then switch, and otherwise stay. Prove that this strategy gives probability greater than 0.5 of obtaining the larger value (and note how this would work for any two values $\theta_1 < \theta_2$).

2. **Exact Sequential Posterior Simulation**
   In some very simple situations, your posterior distribution has a CDF and inverse CDF you can compute, in which case you can find any posterior probabilities explicitly. More often, Bayesian inference relies on simulation. Consider the Normal hierarchical model with equal and known level-1 variances V.

$$\text{level-1:} \qquad Y_i \,|\, \theta_i, \sigma^2 \overset{\text{indep}}{\sim} N(\theta_i, V), \quad i = 1, \ldots, k$$

$$\text{level-2:} \qquad \theta_i \,|\, \mu, A \overset{\text{i.i.d.}}{\sim} N(\mu, A)$$

   a) The marginal distribution of the Y_i's is $N(\mu, V + A)$. Write out the log-likelihood function for μ and the shrinkage factor $B = \frac{V}{V+A}$. Show that Jeffreys' prior for μ is constant, and for B it is $1/B$.

   b) Assuming $p(\mu, B) \propto 1/B$, $0 < B < 1$, give the marginal posterior distribution of $B|y$, the conditional posterior distribution of $\mu|y, B$, and of θ_i, for $i = 1, \ldots, k$. Show what would be involved in working out the marginal posterior densities for $\mu|y$ or $\theta_1, \ldots, \theta_k|y$. The alternative is to simulate a large sample from the joint posterior density, Demonstrate for the WNBA or wordle data. Along with estimating posterior means, try commenting on the rankings. For example, what is the probability the Minnesota Lynx have the highest θ_i, or that CRANE has the lowest θ_i?

c) Often our Y_i's are group averages, and we have $V = \frac{\sigma^2}{n}$. Suppose we have a within-groups variance estimate $s^2|\sigma^2 \sim \text{Gamma}(m, m/\sigma^2)$ that is independent of the Y_i's. Assume the prior distribution $p(\sigma^2) \propto 1/\sigma^2$ and show how the simulation algorithm can be expanded to allow for an unknown σ^2. Demonstrate and show how the posterior intervals change.

d) Explain why things become much more complicated if the V's are unequal, and how equal V's requires equal sample sizes as well as equal σ's.

3. **Empirical Bayes**

The James-Stein estimate is an example of an empirical Bayes estimate: it approximates the posterior mean of the θ_i's using frequentist estimates for the hyper-parameters μ and B. This approach generalizes to more complicated problems. Consider the Normal hierarchical model with level-1 variances $V_i = \frac{\sigma^2}{n_i}$. Assume a constant prior on μ, A and $\log(\sigma^2)$.

a) Give the conditional distribution of $\theta_i|y_i, \mu, \sigma^2$ and note how we need multiple shrinkage factors B_i. Show how we could estimate these using the MSE and an estimate of A. Write out the likelihood function and show that the MLE/posterior mode for A is not available in closed form, even after assuming a value for σ^2.

b) The Newton Raphson algorithm allows you to locate a point where a function has a zero derivative (e.g., a maximum). Explain how this algorithm works by updating a current estimate to the value that would maximize a quadratic function based on the first and second derivative at the current estimate.

c) For the unequal n WNBA data, estimate μ as the overall average and σ^2 as the mean square error, then use the NR algorithm to find the posterior mode for A, treating μ and σ^2 as known. Use these to estimate the k posterior means.

d) Make a shrinkage graph showing the y_i's and estimated θ_i's, and how the ordering changes due to crossover in the shrinkages.

4. **Brute force posterior simulation**

Suppose we observe a discrete data summary y from a data distribution with parameter θ. If we assume prior pdf $p_\theta(\theta)$ then the joint distribution of Y and θ is given by

$$P(Y = y \,|\, \theta)p_\theta(\theta)d\theta \;\approx\; P(Y = y,\, \theta \in [\theta,\, \theta + d\theta))$$

If $p_\theta(\theta)$ represents a proper distribution, we can simulate from the joint distribution of Y and θ by first drawing $\theta \sim p_\theta(\theta)$ and then drawing Y according to the pmf $P(Y = y \,|\, \theta)$. Demonstrate this procedure for women's soccer goal data:

```
x=c(0,1,3,1,1,2,5,0,3,3); n=10; y=sum(x)
```

a) Recall that the sum of the counts Y is a sufficient statistic for iid Poisson data, and that Y also follows a Poisson distribution. Suppose $\theta \sim \text{Gamma}(2, 1)$, which gives prior mean 2 goals but is rather dispersed (not very informative). Generate 1000 pairs $\{\theta, y\}$ and make a scatterplot of θ vs. y. Explain how the posterior distribution of θ for a given data value y (e.g., for $y = 19$ as in our data) is represented in this graph.

b) Repeat the simulation in a, but use N_{sim} much larger than 1000. Take the subset of simulated θ values for which $y = 19$ and make a histogram of these values. Explain why this represents an iid sample from $f_{\theta|y}(\theta|Y = 19)$.

c) Estimate the posterior probability $P(\theta > 2 \,|\, Y = 19)$ using your sample. Compare this with the exact probability found using pgamma with the exact posterior distribution.

d) Use your sample and the quantile function to find a 95% credible interval for θ. Compare to the exact interval using qgamma with the exact posterior distribution.

e) Now suppose you assume prior distribution $\theta \sim F^*_{(2,3,2)}$ instead of Gamma(2,1). Show that you can still get the simulation estimates for c and d, but the exact answers are not easy to compute. To simulate $\theta \sim F^*_{(a,b,c)}$, draw $V_1 \sim \text{Gamma}(a,1)$ independent of $V_2 \sim \text{Gamma}(b,1)$ and set $\theta = c\frac{V_1}{V_2}$.

5. **Rejection Sampling**

We wish to simulate $\theta \sim f_{\theta|y}(\theta|y)$ and we know how to simulate $\theta \sim f_o(\theta)$ for a pdf f_o such that $\max_\theta \left(\frac{f_{\theta|y}(\theta|y)}{f_o(\theta)}\right) = M < \infty$. We say that $f_o(\theta)$ *envelopes* the pdf $f_{\theta|y}(\theta|y)$ because $Mf_o(\theta) \geq f_{\theta|y}(\theta|y)$.

a) Draw a candidate $\theta \sim f_o(\theta)$ independent of $U \sim \text{Unif}(0,1)$. Show that the conditional pdf for θ given $U < \frac{f_{\theta|y}(\theta|y)}{Mf_o(\theta)}$ is $f_{\theta|y}(\theta)$. Explain how this defines a rejection sampling algorithm that accepts each candidate θ with probability $\frac{f_{\theta|y}(\theta|y)}{Mf_o(\theta)}$. See Rice 3.5.2 example D.

b) Draw a graph representing an arbitrary $f(\theta|y)$ and $Mf_o(\theta)$ plotted against θ. Use this graph to give intuition for how selective rejections correct for the difference between the candidate pdf $f_o(\theta)$ and the target pdf $f_{\theta|y}(\theta|y)$. Also give intuition from the graph for why the overall acceptance probability is $1/M$.

c) Explain why, if θ is defined for a bounded interval (e.g. $\theta \in [-1,1]$) with $f(\theta) < \infty$, then a proper Uniform distribution over this interval may be used as an envelope density for rejection sampling.

d) Demonstrate rejection sampling for the following example. Suppose for $i = 1, \ldots, n$, variables X_i and Y_i are standard bivariate Normal (both have $N(0,1)$ marginal distributions) with unknown correlation θ. We observe the following $n = 4$ pairs (x_i, y_i):

i	1	2	3	4
x_i	-1.5	-0.25	0.25	1.5
y_i	-0.25	-1.5	1.5	0.25

The joint pdf is $f_{xy}(x_i, y_i \,|\, \theta) = \frac{1}{2\pi\sqrt{1-\theta^2}}\, e^{-\frac{x_i^2 - 2\theta x_i y_i + y_i^2}{2(1-\theta^2)}}$

Assume prior pdf $p_\theta(\theta) = 0.5I(-1 < \theta < 1)$ and write out $f_{\theta|x,y}(\theta|x,y)$ up to a constant multiplier. Evaluate for a fine grid of θ values (e.g., seq(-.999,.999,.001)) and identify the approximate posterior mode (and maximum likelihood estimate).

e) Make a graph of $f_{\theta|x,y}(\theta|x,y)$ against θ, scaled to have maximum value 1. Draw in a horizontal line to represent a Uniform envelope density. Generate a large sample from the posterior distribution for the correlation θ and make a histogram. Identify the middle 95% of the simulated values to represent a 95% credible interval.