Stat 238, Fall 2025
Homework 1

Alexander Strang
Due: by **11:00 am** Tuesday, February 11, 2025

Submission Instructions

Homework assignments may have a written portion and a code portion. Please follow the directions here Homework Guidelines Ed post when submitting, and check the link for homework policies.

This Homework contains two main parts. A series of relatively short book exercises, then a series of problems establishing key characteristics of the beta-binomial model. It is designed, in part, to help you think through the bias-variance trade-off inherent in posterior inference using informative priors.

Problems eligible for spot grading are marked with an asterisk. These are problems BDA 2.9, and 6 - 9. Please mark at least two problems for spot grading. Note, these will be graded for effort, not accuracy, so you should select the problems that show your best problem-solving effort. The goal is to target our attention to give you meaningful feedback on your work. All problems will be checked for completion and basic accuracy.

Book Problems

Please complete the following problems from *Bayesian Data Analysis* (Third Edition).

- Chapter 1: problems 3 and 7.
- Chapter 2: problems 1 and 9*.

Problem 5: Moments of the Beta-Binomial

Suppose that $\Theta \sim \text{Beta}(\alpha, \beta)$ and $S \mid \Theta = \theta \sim \text{Binomial}(n, \theta)$. That is, S is binomially distributed with n trials and success probability Θ drawn from a Beta distribution with parameters α and β.

If we sample first Θ, then sample S given Θ, the number of successes S is drawn from a beta-binomial distribution (see lab 1). In this problem we will learn how to compute its moments without direct integration. Rather we will chain together moments of familiar distributions (the Beta and the Binomial). This is a good strategy that is worth remembering. It is included as a warm up for the rest of this homework.

(a) Recall that the n^{th} raw moment of S is defined, $\mathbb{E}_S[S^n]$. Write down (don't solve) the explicit integral form for the n^{th} raw moment. Comment on the feasibility of direct integration.

(b) In this case, it is easier to use the rules of conditional expectation, and the law of total variance, to compute the moments. Expand $\mathbb{E}[S]$ and $\text{Var}[S]$ by first conditioning on Θ. Show that:

$$\mathbb{E}[S] = n\mathbb{E}[\Theta], \quad \text{Var}[S] = n(n + \alpha + \beta)\text{Var}[\Theta]. \tag{1}$$

(c) Use this result to compute the marginal variance in the MLE estimator for Θ. That is, for $\hat{\theta}_{\text{MLE}}(S; n, \alpha, \beta) = S/n$. We will use this answer later.

Problem 6*: Regularization - Bias in Posterior Point Inference

Consider the same joint Θ, S model established in problem 5. Under this model:

(a) Write down the posterior distribution for $\Theta|S = s$ as a function of s, the number of trials, n, and the prior parameters α, β. Then, write down the formulas for the posterior expectation, $\hat{\Theta}_{\text{mean}}(s; n, \alpha, \beta) = \mathbb{E}[\Theta|S = s]$, and MAP estimator, $\hat{\Theta}_{\text{MAP}}(s; n, \alpha, \beta)$, as functions of s, n, α, and β. You do not need to rederive the equations for the mean and mode of the beta distribution.

(b) Compute the bias in $\hat{\Theta}_{\text{MLE}}(S)$, $\hat{\Theta}_{\text{mean}}(S)$, $\hat{\Theta}_{\text{MAP}}(S)$ conditional on $\Theta = \theta$. Recall that the bias in an estimator is the expected error, $\mathbb{E}[\hat{\Theta} - \theta]$. In this case, we want $\mathbb{E}_{S|\Theta=\theta}[\hat{\theta}(S) - \Theta]$.

(c) Which estimators are conditionally unbiased given θ? Which estimator is most biased?

(d) In what direction are the posterior estimators biased relative to the MLE estimate? Rationalize this effect in terms of the prior.

(e) Repeat the same calculation, but now marginalize over Θ as well. That is, compute the joint bias, $\mathbb{E}_{\Theta,S}[\hat{\theta}(S) - \Theta]$. Which estimators are unbiased with respect to the full data generating process (joint model)?

(f) Interpret your conclusions: (i) suggest a scenario where you would prefer the conditional notion of bias, and a situation where you would prefer the joint notion of bias. (ii) give a reasonable standard for the largest bias you would accept (i.e. the bias should be ... relative to the posterior standard deviation, or the standard deviation in the sampling distribution of the estimators).

Problem 7*: Principles for Selecting an Estimator

Given an observation, say $S = s$, we usually don't return the full posterior for an unknown, $\Theta|S = s$. Instead, we return inferential summaries, at minimum, the posterior mean, $\hat{\theta}_{\text{mean}}(s) = \mathbb{E}[\Theta|S = s]$, or posterior mode, $\hat{\theta}_{\text{MAP}}(s)$. Why adopt these points as summaries?

One way to justify an estimation procedure is to select the optimizer to control properties of the conditional error. Prove each statement below regarding a Bayesian inference problem where S denotes a generic observable that is drawn jointly with an unknown Θ taking values on the real line.

(a) The estimator $\hat{\theta}(s)$ that minimizes the expected square error $L_2(\hat{\theta}(s), \Theta) = (\hat{\theta}(s) - \Theta)^2$, for each possible observed s, is the posterior mean:

$$\hat{\theta}_{\text{mean}}(s) = \text{argmin}_{\theta}\{\mathbb{E}_{\Theta|S=s}[L_2(\theta, \Theta)]\} \tag{2}$$

(b) Let $t(s)$ denote an arbitrary real-valued, function of the observable. We will consider t a "test" function. Let $\text{Corr}_{\Theta,S}[t(S), \hat{\theta}(S) - \Theta]$ denote the correlation between the test function value and the estimation error $\hat{\theta}(S) - \Theta$. Here the correlation is evaluated jointly over the unknown and the observable. Note that, if the correlation is nonzero for any test function, then that test function carries information about the estimation error. So, it could be used to try and improve the estimator.

The only estimator whose estimation error has zero correlation with all test functions is, up to a constant, the posterior mean:

$$\text{Corr}_{\Theta,S}[t(S), \hat{\theta}(S) - \Theta] = 0 \text{ for all } t \text{ if and only if } \hat{\theta}(\cdot) = \hat{\theta}_{\text{mean}}(\cdot). \tag{3}$$

(c) The estimator $\hat{\theta}(s)$ that minimizes the expected absolute error $L_1(\hat{\theta}(s), \Theta) = |\hat{\theta}(s) - \Theta|$, for each possible observed s, is the posterior median:

$$\hat{\theta}_{\text{median}}(s) = \text{Median}(\Theta|S = s) = \text{argmin}_{\theta}\{\mathbb{E}_{\Theta|S=s}[L_1(\theta, \Theta)]\} \tag{4}$$

Problem 8*: Regularization - Reducing Variance

Consider the same joint Θ, S model established in problem 5. A common reason to introduce a prior is to regularize inference. Regularization typically aims to reduce the sampling variance in an estimator by biasing it towards answers that are more likely, a priori.

We've already studied the biasing effects of the prior. In this problem, we will study its role in reducing sample variance relative to Frequentist estimators.

(a) Compute the variance in the estimation error $\hat{\theta}_{\text{MLE}}(S) - \theta$ given a fixed value for θ. That is $\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{MLE}}(S) - \theta]$. This variance measures the expected (squared) magnitude of the error due to sampling variability.

(b) Compute the expectation over Θ of the variance in the estimation error $\hat{\theta}_{\text{MLE}}(S) - \Theta$ over S. That is, $\mathbb{E}_{\Theta}[\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{MLE}}(S) - \theta]]$. Interpret this quantity as a measure of the accuracy of the maximum likelihood estimator.

(c) Repeat this same pair of calculations for the posterior mean estimator, $\hat{\theta}_{\text{mean}}(S)$. Compute the ratios:

$$\frac{\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{mean}}(S) - \theta]}{\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{MLE}}(S) - \theta]} \text{ and } \frac{\mathbb{E}_{\Theta}[\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{mean}}(S) - \Theta]]}{\mathbb{E}_{\Theta}[\text{Var}_{S|\Theta=\theta}[\hat{\theta}_{\text{MLE}}(S) - \Theta]]} \tag{5}$$

Then, use your answer to compare the sampling variability in the estimators. Is the variance/expected variance for the posterior mean estimator always less than the variance/expected variance in the maximum likelihood estimator?

(d) Use the law of total variance to show that the *expected sampling* variance in the maximum likelihood estimator's error equals the *joint* variance in the maximum likelihood estimator error. That is, that your answer to (b) equals $\text{Var}_{\Theta,S}[\hat{\theta}(S) - \Theta]$.

(e) Use the law of total variance to show that the joint variance in the posterior mean estimation error is:

$$\text{Var}_{\Theta,S}[\hat{\theta}_{\text{mean}}(S) - \Theta] = \frac{(\alpha+\beta)}{(n+\alpha+\beta)}\text{Var}_{\Theta}[\Theta]. \tag{6}$$

(f) Show that the joint variance in the posterior mean estimator's error is strictly less than the joint variance in the maximum likelihood estimator's error, and that the ratio of variances is $\frac{n}{n+\alpha+\beta}$.

Problem 9*: Information

In the Bayesian paradigm, we learn from observations by conditioning on them. Before observing $S = s$, we believe $\Theta \sim p_{\Theta}$. After observing $S = s$, we believe that $\Theta \sim p_{\Theta|S=s}$. Observations provide information through the likelihood, which pushes the prior to the posterior. In this problem, we will compare the prior and posterior variance after observing $S = s$. As usual, we restrict our attention to the (beta, binomial) model established in question 5.

(a) Write down the expressions for the prior and posterior variance as functions of $\mathbb{E}[\Theta]$, $\mathbb{E}[\Theta|S = s]$, and the parameters α, β, n.

(b) Use the law of total variance to show that, in expectation, observation always reduces the variance in the unknown. That is:

$$\mathbb{E}_S[\text{Var}[\Theta] - \text{Var}[\Theta|S]] > 0 \tag{7}$$

(c) Show that the expected percent reduction in variance due to observation approaches 1 as n diverges. In particular, show that:

$$\mathbb{E}_S\left[\frac{\text{Var}[\Theta] - \text{Var}[\Theta|S]}{\text{Var}[\Theta]}\right] = \frac{n}{n+\alpha+\beta}. \tag{8}$$

(d) While observation reduces the variance in expectation, some observations may increase the variance in the unknown. Let's show these are atypical. First, use Markov's inequality to show that:

$$\Pr(\text{Var}[\Theta|S] > \text{Var}[\Theta]) \leq \frac{\alpha+\beta}{n}. \tag{9}$$

Interpret this inequality. What does it imply as n diverges?

(e) Solve for the region of s/n given n, α and β such that the posterior variance is *greater* than the prior variance. Don't worry about simplifying (your answer will be expressed in terms of an ugly quadratic equation). Write a code to illustrate this region for $\alpha = \beta = 0.1$, $\alpha = \beta = 1$, $\alpha = \beta = 3$, and $\alpha = 8$, $\beta = 2$ as a function of n (sweep n from 1 to 30). I'd suggest plotting n on the vertical axis, and s/n on the horizontal. Attach your plots.

(f) What do you notice about the observations that increase our uncertainty about the unknown? What prerequisites do we need to place on the prior, and the number of data points collected, such that an observation can increase our uncertainty?

Problem 0: Spot Grading

Select two problems from the problems marked with an asterisk for spot grading.