Problem Set 3: Written Part

Your Name Goes Here

Details

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 15% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited.

Problem I: M&M's Example: Bias, Variance, Efficiency, and MSE of maximum likelihood estimator and Bayesian posterior mean

In lab 6, we estimated the proportion of M&M's that are blue, θ, based on a sample of n M&M's. We defined the random variable X, which was the count of how many M&M's were blue in the sample. Our model was $X \sim \text{Binomial}(n, \theta)$.

We have developed two approaches to inference for θ:

1. The maximum likelihood estimator $\hat{\theta}^{MLE} = \frac{X}{n}$. In lecture, we showed that $E(\hat{\theta}^{MLE}) = \theta$, $Var(\hat{\theta}^{MLE}) = \frac{\theta(1-\theta)}{n}$, and $MSE(\hat{\theta}^{MLE}) = \frac{\theta(1-\theta)}{n}$.
2. A Bayesian approach with conjugate prior distribution given by $\Theta \sim Beta(a, b)$. The posterior distribution is given by $\Theta | n, x \sim Beta(a + x, b + n - x)$. From this posterior distribution, we can obtain point estimates, with the most common choice being the posterior mean, which can be written as follows:

$$\hat{\theta}^{Bayes} = \frac{a + x}{(a + x) + (b + n - x)} = \cdots = (1 - w)\frac{a}{a + b} + w\frac{x}{n},$$

where $w = \frac{n}{n+a+b}$.

By considering what the posterior mean would be across different samples, we can view the posterior mean as specifying an estimator (replacing the lower case x above by a capital X),

$$\hat{\theta}^{Bayes} = (1-w)\frac{a}{a+b} + w\frac{X}{n}.$$

In the lab, we considered three different prior specifications for Θ with different level of informativeness, and we saw that for a large sample size it didn't matter which prior we used; the resulting estimates were essentially the same as each other and as the maximum likelihood estimate.

Let's now look more closely at how the Bayesian and Frequentist methods compare for smaller sample sizes. For concreteness, let's work with the intermediate of our three prior specifications, $\Theta \sim \text{Beta}(2, 8)$. Let's also imagine that our sample size is fixed at 10.

In this case, $w = \frac{n}{n+a+b} = \frac{10}{10+2+8} = 0.5$, and the Bayesian estimator reduces to

$$\hat{\theta}^{Bayes} = (1-w)\frac{a}{a+b} + w\frac{X}{n} = 0.5\frac{2}{10} + 0.5\frac{X}{10}$$

(1) Find the bias of $\hat{\theta}^{Bayes}$ (this will depend on the value of θ).

$$\begin{aligned}
Bias(\hat{\theta}_{Bayes}) &= E(\hat{\theta}_{Bayes}) - \theta \\
&= E((1-w)\frac{a}{a+b} + w\frac{X}{n}) - \theta \\
&= (1-w)\frac{a}{a+b} + w\frac{n\theta}{n} - \theta \\
&= (1-w)\frac{a}{a+b} + w\theta - \theta \\
&= (1-w)\frac{a}{a+b} - \theta(1-w) \\
&= 0.1 - 0.5\theta
\end{aligned}$$

(2) For what value of θ is $\hat{\theta}^{Bayes}$ unbiased? How does that relate to the prior distribution?

To find the value of θ for which $\hat{\theta}_{Bayes}$ is unbiased, we set the bias to 0 and solve for θ. This yields $\theta = \frac{a}{a+b}$. With our specific choice of prior, $\hat{\theta}_{Bayes}$ is unbiased if $\theta = 0.2$. This is the prior mean.

(3) Find the variance of $\hat{\theta}_{Bayes}$ in terms of θ.

$$\begin{aligned}
Var(\hat{\theta}_{Bayes}) &= Var((1-w)\frac{a}{a+b} + w\frac{X}{n}) \\
&= w^2 Var(\frac{X}{n}) \\
&= w^2\frac{\theta(1-\theta)}{n} \\
&= 0.5^2\frac{\theta(1-\theta)}{10}
\end{aligned}$$

(4) Find an expression for the MSE of $\hat{\theta}_{Bayes}$ in terms of θ.

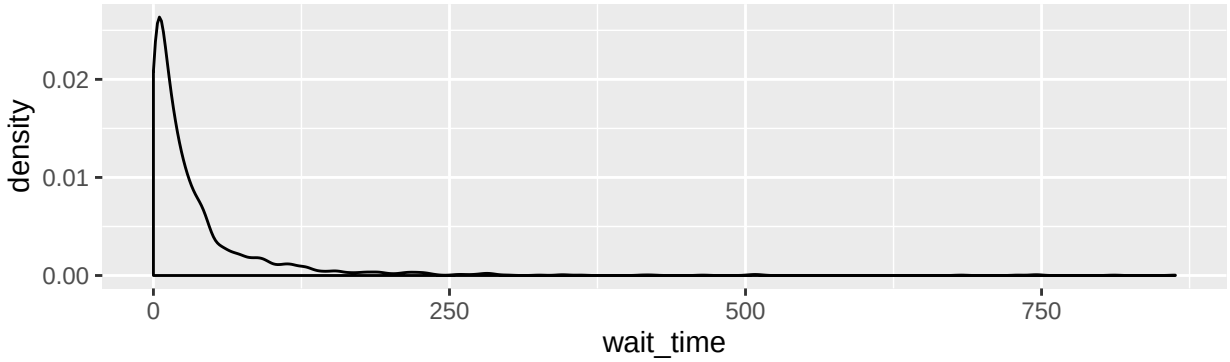
$$\begin{aligned}
MSE(\hat{\theta}_{Bayes}) &= \left\{Bias(\hat{\theta}_{Bayes})\right\}^2 + Var(\hat{\theta}_{Bayes}) \\
&= \left\{(1-w)\frac{a}{a+b} - \theta(1-w)\right\}^2 + w^2\frac{\theta(1-\theta)}{n} \\
&= (0.1 - 0.5\theta)^2 + 0.5^2\frac{\theta(1-\theta)}{10}
\end{aligned}$$

Problem II: Exponential model for hospital waiting times

Let's revisit an earlier example where we were waiting times for patients at hospital emergency departments. Here is the problem description from problem set 1:

The National Center for Health Statistics, a division within the U.S. Centers for Disease Control, conducts a nationally representative survey of hospitals each year to track the waiting times for emergency room visits (that is, how much time passed between when a patient arrived at the hospital and when they were seen by a doctor or registered nurse). In this problem, we will model the distribution of waiting times for 1874 emergency department visits from 2012. The output below shows a plot of the data.

```
ggplot(data = er_visits, mapping = aes(x = wait_time)) +
  geom_density()
```



An exponential distribution is often used to model waiting times. Suppose we adopt the data model

$X_i \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$, $i = 1, \ldots, n$,

where X_i is the waiting time for visit number i. In our data set we have observed values $x_1, \ldots, x_n$, where $n = 1874$.

We will use the parameterization of the Exponential distribution given on the common probability distributions handout: $f_{X_i|\Lambda}(x_i|\lambda) = \lambda e^{-\lambda x_i}$

(1) Show that the Gamma distribution is a conjugate prior for the parameter λ. State the posterior distribution and all of its parameters.

If we use the prior $\Lambda \sim \text{Gamma}(\alpha, \beta)$, then Λ has pdf $f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$

The posterior pdf can be found as follows:

$$
\begin{aligned}
f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots,x_n) &= \frac{f_\Lambda(\lambda)\prod_{i=1}^n f_{X_i|\Lambda}(x_i|\lambda)}{\int f_\Lambda(\lambda)\prod_{i=1}^n f_{X_i|\Lambda}(x_i|\lambda)d\lambda} \\
&\propto \lambda^{\alpha-1}e^{-\beta\lambda}\prod_{i=1}^n \lambda e^{-\lambda x_i} \\
&\propto \lambda^{\alpha+n-1}e^{-(\beta+\sum_{i=1}^n x_i)\lambda}
\end{aligned}
$$

Therefore, the posterior distribution is

$\Lambda|X_1, \ldots, X_n \sim \text{Gamma}(\alpha + n, \beta + \sum_{i=1}^n x_i)$

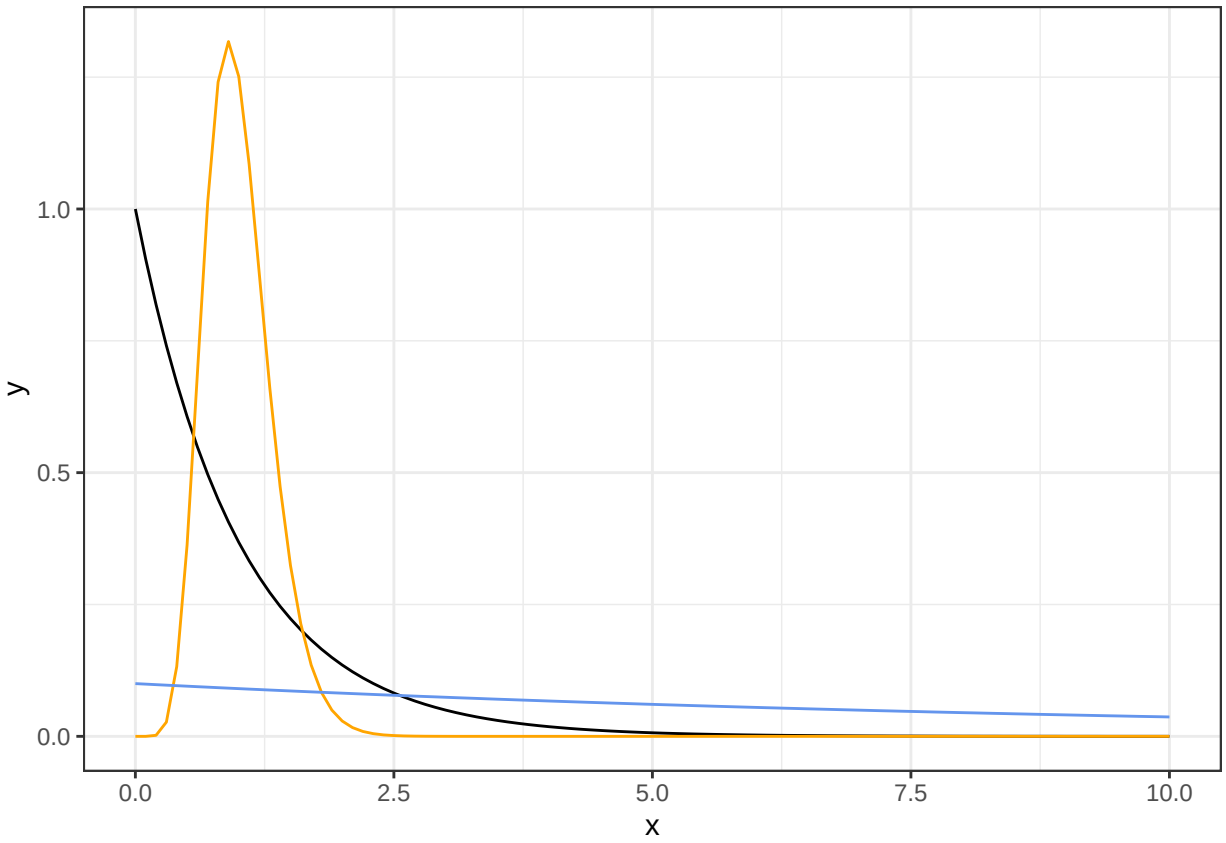
(2) Consider the three Gamma distribution probability density functions in the plot below. Which one would be most appropriate for use as a prior distribution by an analyst who did not know much about emergency room waiting times? Explain your answer in a sentence or two.

```
beta1 <- 1
alpha1 <- 1
```

```r
beta2 <- 10
alpha2 <- 10

beta3 <- 0.1
alpha3 <- 1

ggplot(data = data.frame(x = c(0, 10)), mapping = aes(x = x)) +
  stat_function(fun = dgamma, args = list(rate = beta1, shape = alpha1)) +
  stat_function(fun = dgamma, args = list(rate = beta2, shape = alpha2), color = "orange") +
  stat_function(fun = dgamma, args = list(rate = beta3, shape = alpha3), color = "cornflowerblue") +
  theme_bw()
```



The prior distribution plotted in blue, with parameters $\alpha = 1$ and $\beta = 0.1$, is least informative; it expresses that there is a wide range of values for λ that are plausible, whereas the distribution plotted in orange says there is a high probability that λ is between about 0.1 and 2, and the distribution plotted in black says there is a high probability that λ is between about 0 and 4. The distribution shown in blue would be most appropriate for an analyst with little prior knowledge about the value of the parameter.

Problem III: Multinomial model for Jane Austen word usage

Background on Multinomial distribution

The multinomial distribution is a distribution for a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ (a single observation from the multinomial distribution is a vector of length k).

Suppose $\mathbf{X} \sim \text{Multinomial}(n, \theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is a vector of counts for how many observations fell into each of k categories in n independent trials where the item sampled in each trial falls into category j with probability θ_j. More concretely, imagine rolling a weighted die with k sides n times, and on each roll, side j comes up with probability θ_j. The vector X records how many times each face of the die came up. (Note that since exactly one side of the die must come up on each roll, we must have $\sum_{j=1}^{k} \theta_j = 1$.)

The probability mass function of X is

$f_{\mathbf{X}}(x|\theta) = \frac{n!}{x_1!x_2!\cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}$

Background on Dirichlet distribution

The Dirichlet distribution is a distribution on a vector of probabilities that sum to 1, such as the probabilities for each side of the die coming up in the multinomial distribution. Its parameters are a vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$.

If $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_k)$ are jointly distributed as Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$, then the joint pdf of $\mathbf{\Theta}$ is given by:

$f_{\mathbf{\Theta}}(\theta) = \frac{1}{B(\alpha)}\prod_{j=1}^{k}\theta_j^{\alpha_j - 1}$

Here, $B(\alpha)$ is a complicated function of α that's just there to ensure that the joint pdf of $\mathbf{\Theta}$ integrates to 1.

Problem set up

This problem has been adapted (heavily) from problem 13.9 in Rice. Note that I think this is not actually the way I would analyze these data, but it's still an interesting example to think about. Here's a quote from Rice:

> When Jane Austin died, she left the novel *Sanditon* only partially completed, but she left a summary of the remainder. A highly literate admirer finshed the novel, attempting to emulate Austen's style, and the hybrid was published. Morton (1978) counted the occurrences of various words in several works: Chapters 1 and 3 of *Sense and Sensibility*, Chapters 1, 2, and 4 of *Emma*, Chapters 1 and 6 of *Sanditon* (written by Austen), and Chapters 12 and 24 of *Sanditon* (written by her admirer).

Problem 13.9 in the book says:

> This problem considers some more data on Jane Austen and her imitator (Morton 1978). The following table gives the relative frequency of the word *a* preceded by (PB) and not preceded by (NPB) the word *such*, the word *and* followed by (FB) and not followed by (NFB) the word *I*, and the word *the* preceded by and not preceded by *on*. [I have added column totals.]

To simplify this problem, we will examine only the uses of these phrases in the two sections of *Sanditon*.

Words	Austen	Imitator
a PB such	8	2
a NPB such	93	81
and FB I	12	1
and NFB I	139	153
the PB on	8	17
the NPB on	221	204
Total	481	458

The idea of this analysis is to treat the phrase counts in each part of the book as a realization of a multinomial random vector with unknown cell probabilities specific to that author and a known total count. For example, the vector of phrase counts for the part of *Sanditon* written by Jane Austen, (8, 93, 12, 139, 8, 221), is modeled as a realization of a multinomial random variable with size $n = 481$ and unknown probability vector $\theta^{Austen} = (\theta_1^{Austen}, \theta_2^{Austen}, \theta_3^{Austen}, \theta_4^{Austen}, \theta_5^{Austen}, \theta_6^{Austen})$. You might think of this vector as representing Austen's relative preference for each of the six word constructions in the table. Similarly, there is a second vector $\theta^{Imitator} = (\theta_1^{Imitator}, \theta_2^{Imitator}, \theta_3^{Imitator}, \theta_4^{Imitator}, \theta_5^{Imitator}, \theta_6^{Imitator})$ representing the imitator's relative preference for each of the word constructions in the table.

(1) Suppose that we have a single observation $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim$ Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$. That is, we will observe a vector of counts from each of the k possible categories in n trials. Show that the Dirichlet is a conjugate prior for the multinomial model. State the posterior distribution and all of its parameters.

The posterior pdf is obtained as follows (up to a constant of proportionality):

$$
\begin{aligned}
f_{\mathbf{\Theta}|\mathbf{X}} &\propto f_{\mathbf{\Theta}}(\theta) f_{\mathbf{X}|\Theta}(x|\theta) \\
&\propto \prod_{j=1}^{k} \theta_j^{\alpha_j} \prod_{j=1}^{k} \theta_j^{x_j} \\
&\propto \prod_{j=1}^{k} \theta_j^{\alpha_j + x_j}
\end{aligned}
$$

Therefore, $\mathbf{\Theta}|\mathbf{X} \sim \text{Dirichlet}(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_k + x_k)$

(2) Suppose a Dirichlet(1, 1, 1, 1, 1, 1) prior is adopted for each of the unknown parameter vectors $\theta^{Austen} = (\theta_1^{Austen}, \theta_2^{Austen}, \theta_3^{Austen}, \theta_4^{Austen}, \theta_5^{Austen}, \theta_6^{Austen})$ and $\theta^{Imitator} = (\theta_1^{Imitator}, \theta_2^{Imitator}, \theta_3^{Imitator}, \theta_4^{Imitator}, \theta_5^{Imitator}, \theta_6^{Imitator})$ described above. What are the posterior distributions? Give all parameters for the posterior distributions. There will be two separate posteriors, one for each of the parameter vectors.

We have:

$\mathbf{\Theta}^{Austen} \sim \text{Dirichlet}(9, 94, 13, 140, 9, 222)$

$\mathbf{\Theta}^{Imitator} \sim \text{Dirichlet}(3, 82, 2, 154, 18, 205)$

The parameters for the posterior distributions were obtained by adding 1 (α's from the prior) to the phrase counts x in the table from the problem statement above.

(3) Find the marginal posteriors for each of the 12 individual parameters θ_1^{Austen}, θ_2^{Austen}, θ_3^{Austen}, θ_4^{Austen}, θ_5^{Austen}, θ_6^{Austen}, $\theta_1^{Imitator}$, $\theta_2^{Imitator}$, $\theta_3^{Imitator}$, $\theta_4^{Imitator}$, $\theta_5^{Imitator}$, and $\theta_6^{Imitator}$.

To do this, use the fact that if $\mathbf{\Theta} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, then the marginal distribution of Θ_i is $\Theta_i \sim \text{Beta}(\alpha_i, \sum_{j=1}^{k} \alpha_j - \alpha_i)$. We are doing this because it's hard to visualize and think about a joint distribution for a vector of 6 parameters, but it's easier to think about the marginal distribution for each θ_j.

$\mathbf{\Theta}_1^{Austen}|X \sim \text{Beta}(9, 478)$

$\mathbf{\Theta}_2^{Austen}|X \sim \text{Beta}(94, 393)$

$\mathbf{\Theta}_3^{Austen}|X \sim \text{Beta}(13, 474)$

$\mathbf{\Theta}_4^{Austen}|X \sim \text{Beta}(140, 347)$

$\mathbf{\Theta}_5^{Austen}|X \sim \text{Beta}(9, 478)$

$\mathbf{\Theta}_6^{Austen}|X \sim \text{Beta}(222, 265)$

$\mathbf{\Theta}_1^{Imitator}|X \sim \text{Beta}(3, 461)$

$\mathbf{\Theta}_2^{Imitator}|X \sim \text{Beta}(82, 382)$

$\mathbf{\Theta}_3^{Imitator}|X \sim \text{Beta}(2, 462)$

$\mathbf{\Theta}_4^{Imitator}|X \sim \text{Beta}(154, 310)$

$\mathbf{\Theta}_5^{Imitator}|X \sim \text{Beta}(18, 446)$

$\mathbf{\Theta}_6^{Imitator}|X \sim \text{Beta}(205, 259)$

Problem IV: Bias

Is the following claim true or false? Justify your answer in a sentence or two.

If two estimators are unbiased, they are equally good and it does not matter which one you use.

This is false. There are other aspects of an estimator that should also be considered – in particular, two unbiased estimators may have different variances, and in that case we would generally prefer the estimator with smaller variance.

Problem V: Bayes Intuition

Write a couple of paragraphs explaining what a prior distribution is, what a posterior distribution is, and how Bayes' rule gets us from one to the other. You should include a formula for Bayes' Rule **and some written sentences describing what's happening in the formulas**. Your first paragraph should explain these ideas in a general setting, and your second paragraph should illustrate them in the context of a specific example such as estimating the proportion of M&Ms that are blue based on a sample of n M&Ms (you can pick a different example if you prefer). For the purpose of this assignment, a paragraph consists of at least three complete sentences.

This problem will be graded by Evan, and you will have an opportunity to submit a revision to your answer for up to full credit after receiving feedback.

I will post a solution for this problem after everyone has had a chance to submit a revision. In the meantime, I'm happy to discuss in person.