Mathematics Bootcamp

Part V: Bayesian Analysis

Jordan Bryan[1] Brian Cozzi[1] Michael Valancius[1] Graham Tierney[1] Becky Tang[1]

[1]Department of Statistical Science
Duke University

Graduate Orientation, August 2020

Outline

Introduction to Bayesian Analysis

Different Interpretations of Probability

In general, there are two main interpretations of probability, both of which are consistent with the axioms of probability discussed to this point.

- **Frequentists** posit that the probability of an event is its relative frequency over time, i.e., its relative frequency of occurrence after repeating a process a large number of times under similar conditions.
- The **Bayesian** interpretation, gives the notion of probability a subjective status by regarding it as a measure of the 'degree of belief' of the individual assessing the uncertainty of a particular situation.

Who doesn't love coin flips?

Take the example of tossing a fair coin! How would Frequentists and Bayesians interpret the event?

- Frequentists: if the coin is tossed infinitely (or a large enough number of) times, 50% of the time it will land on heads and 50% on tails.
- Bayesians: when the coin is tossed, there is a 50% chance it will land on heads, and a 50% chance for tails

Importance of the Distinction

There are some key consequences to the Bayesian interpretation of probability:

- No assumption about the randomness of a particular event. Instead, probability is a measure of our own uncertainty.
- The importance of this distinction is that probability statements can be made about a much larger class of objects.

Namely, in parametric models, the parameters are often assumed to have "true" (albeit unknown) values, Bayesian methods can use probability to describe the uncertainty about parameters. In this general framework, anything unknown can be described by a probability distribution.

Bayes Rule

Bayesian inference uses Bayes' theorem to update probabilities after more evidence or data is obtained:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Two quantities of interest:

1. $y \in \mathcal{Y}$: the data is a member of $\mathcal{Y}$, the *sample space* or the set of all possible datasets;
2. $\theta \in \Theta$: the parameter (Θ: the parameter space), expressing the population characteristics.

Bayes Rule

The posterior distribution is obtained from the prior distribution and sampling model via **Bayes' rule**:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

That is,

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{normalizing constant}}$$

In practice, evaluating the normalizing constant is often intractable, so we instead obtain

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

and the form of the right hand side can help us determine $p(\theta|y)$.

Main components of Bayesian analysis

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Three Distributions:

1. For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, the **sampling model** or **likelihood** $p(y|\theta)$ is the probability of observing our data, given unknown parameters;
2. For each numerical value $\theta \in \Theta$, the **prior distribution** $p(\theta)$ describes our belief that θ represents the true population characteristics;
3. For each numerical value of $\theta \in \Theta$, the **posterior distribution** $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y.

Prior Interpretation

Generally, the prior distribution for a parameter θ is a probability distribution that reflects our uncertainty about θ before data (or, if updating, new data) is taken into account.

The prior distribution is the choice of the person conducting the analysis and ideally provides useful information that might be known about θ a priori. For example, if we are interested in describing the probability that the US Women's National team defeating Thailand in a soccer match, we might a priori have more belief that the probability θ is closer to 1 than to 0.

Posterior Interpretation

The posterior updates our prior, conditioned on the data that we have observed.

As $p(\theta|y) \propto p(y|\theta)p(\theta)$, a high prior probability for a given $\tilde{\theta}$ or a high likelihood value under that $\tilde{\theta}$ can result in a high posterior probability.

Brief aside

- Duke is a Bayesian department, but that does not mean you need to be Bayesian yourself! Some of our professors do non-Bayesian work!
- Would be interested to see how many students become converts...
- Areas where Bayesian statistics is useful

Analytical Methods

Conjugacy

Definition

A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

- For example, as seen in the following exercise, the beta distribution is conjugate for the binomial sampling model. Thus, if $p(\theta) \sim Beta(a, b)$ and $p(y|\theta) \sim Binomial(\theta)$, then $p(\theta|y) \sim Beta(c, d)$.
- Conjugate priors have a practical advantage: they provide computational convenience and interpretability since the posterior will follow a known parametric form.
- However, they might not always be flexible enough, and for more complicated or higher dimensional problems they quickly become impossible to use.

Binomial Model

Let $Y \sim Binomial(n, \theta)$, where $Y \in \{0, 1, ...n\}$. Having observed $Y = y$, we conduct posterior inference: $p(\theta|y)$. Using Bayes rule:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)} \\ &\propto c(y)\theta^y(1-\theta)^{n-y}p(\theta) \end{aligned}$$

If we choose a conjugate prior $p(\theta)$ to our likelihood, then we will have a closed form expression for the posterior.

Finding Conjugate Prior

For a binomial sampling model:

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)$$

Thus, a conjugate prior is a distribution of θ such that $p(\theta) \propto \theta^{c_1}(1-\theta)^{c_2}$ as a function of θ. The Beta distribution satisfies this requirement, as is illustrated in the following example.

Binomial Model Continued

Recall that if $\theta \sim Beta(a, b)$, then $p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$.

$$
\begin{aligned}
p(\theta|y) &\propto c(y)\theta^{y}(1-\theta)^{n-y}p(\theta) \\
&= c(y)\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&= \left(c(y)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\theta^{y+a-1}(1-\theta)^{n+b-y-1} \\
&= c_2(y)\theta^{y+a-1}(1-\theta)^{n+b-y-1}
\end{aligned}
$$

How do we confirm that the posterior distribution is also beta? Since the posterior distribution is a proper probability distribution, it integrates to 1. This fact, combined with some algebra, reveals the posterior distribution to be Beta(y + a, n + b - y).

Binomial Model Continued

$$
\begin{aligned}
1 &= \int_0^1 c_2(y)\theta^{y+a-1}(1-\theta)^{n+b-y-1}d\theta \\
&\implies 1 = c_2(y)\int_0^1 \theta^{y+a-1}(1-\theta)^{n+b-y-1}d\theta \\
&\implies 1 = c_2(y)\frac{\Gamma(y+a)\Gamma(n+b-y)}{\Gamma(n+a+b)} \\
&\implies c_2(y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n+b-y)} \\
&\implies p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n+b-y)}\theta^{y+a-1}(1-\theta)^{n+b-y-1} \\
&\implies p(\theta|y) = Beta(y+a, n+b-y)
\end{aligned}
$$

A Binomial Example

A survey is carried out to study the support rate θ $(0 < \theta < 1)$ of a policy. 100 people are surveyed, and a binary response Y_i is obtained from each person i $(i = 1, 2, \ldots, 100)$, $Y_i \sim \text{Bernoulli}(\theta)$ (that is, $Y = \sum_{i=1}^{100} Y_i \sim \text{Binomial}(100, \theta)$).

Before the survey, we believe that $\theta \sim \text{Beta}(5, 5)$, while the result of the survey is $Y = 60$. We'd like to obtain the posterior distribution of θ given the survey outcome.

A Binomial Example (Cont'd)

The prior distribution is $\theta \sim \text{Beta}(5, 5)$, that is

$$p(\theta) = \frac{\theta^{5-1}(1-\theta)^{5-1}}{B(5,5)} \propto \theta^{5-1}(1-\theta)^{5-1}$$

The sampling distribution is $Y \sim \text{Binomial}(100, \theta)$, that is, for each $\theta \in (0, 1)$ and $y = 0, 1, \ldots, 100$,

$$P(Y = y|\theta) = \binom{100}{y} \theta^y (1-\theta)^{100-y}.$$

Using Bayes' rule, the posterior distribution of θ given that $Y = 60$ is

$$\begin{aligned} p(\theta|Y = 60) &\propto p(Y = 60|\theta)p(\theta) \\ &= \theta^{60}(1-\theta)^{100-60}\theta^{5-1}(1-\theta)^{5-1} \\ &= \theta^{65-1}(1-\theta)^{45-1}, \end{aligned}$$

which has the form of the p.d.f. of a $\text{Beta}(65, 45)$ distribution. Thus, we have $\theta|Y = 60 \sim \text{Beta}(65, 45)$.

Bayesian Updating

Bayesian inference provides a framework for updating beliefs upon observing data. There is an initial belief, described by the prior distribution. Data is observed. Following Bayes Rule, the beliefs are updated into what is called the posterior distribution.

Question: If we observe data $D_1 = (x_1 ... x_n)$ and find $p(\theta|x_1 ... x_n)$ and then later observe more data $D_2 = (x_{n+1} ... x_{n+m})$, is $p(\theta|D_1, D_2) \propto p(D_2|\theta)p(\theta|D_1)$?

In other words, can the first posterior we derived after observing D_1 be used as the prior for conducting posterior inference when new data D_2 is observed?

Bayesian Updating Continued

So long as D_1 and D_2 are treated as conditionally independent given θ, the answer is yes. Thus, Bayesian inference gives us a powerful tool for repeatably updating a model every time more data is observed. The former posterior distribution becomes the new prior once more data is observed.

Example: In our previous example, we found that for a Binomial(100,θ) model with a prior of $\theta \sim$Beta(5,5) and observed data of $Y = 60$, the posterior distribution is $\theta|Y = 60 \sim$Beta(65,45). Now suppose we observe 100 more surveys, this time with $Y_2 = 55$. How do our beliefs change?

$$\begin{aligned} p(\theta|Y_2 = 55) &\propto p(Y_2 = 55|\theta)p(\theta|Y_1 = 60) \\ &= \theta^{55}(1-\theta)^{100-55}\theta^{65-1}(1-\theta)^{45-1} \\ &= \theta^{120-1}(1-\theta)^{90-1} \end{aligned}$$

Thus, our updated posterior is Beta(120,90).

Bayesian Updating Continued

What if, instead of updating, we restarted the analysis with the original prior of $Beta(5,5)$ and now had $n = 100 + 100 = 200$ and $Y = Y_1 + Y_2 = 115$?

$$\begin{aligned} p(\theta|Y) &\propto p(Y = 115|\theta)p(\theta) \\ &= \theta^1 15(1-\theta)^{200-115}\theta^{5-1}(1-\theta)^{5-1} \\ &= \theta^{120-1}(1-\theta)^{90-1} \end{aligned}$$

This, as before, has the form of Beta(120,90).

Bayesian Estimates: More than Point Estimation

- In maximum likelihood estimation, a random sample $X_1 ... X_n$ is drawn from a population with a probability distribution that is indexed by an unknown, fixed θ.
- The maximum likelihood estimate is a "best" guess at the true value of θ based on the sample.
- In Bayesian approaches, uncertainty about θ is itself described by a probability distribution. This distribution, $p(\theta)$, is called the prior distribution.
- After a sample is obtained from the sampling model $p(y|\theta)$, beliefs about θ are **updated** through Bayes rule: $p(\theta|y_1 ... y_n)$.

Comparison via Simple Example

Let $X_1 ... X_{10}$ be 10 coin tosses where $X_i = 1$ if the coin lands heads up and 0 if it lands tails up.

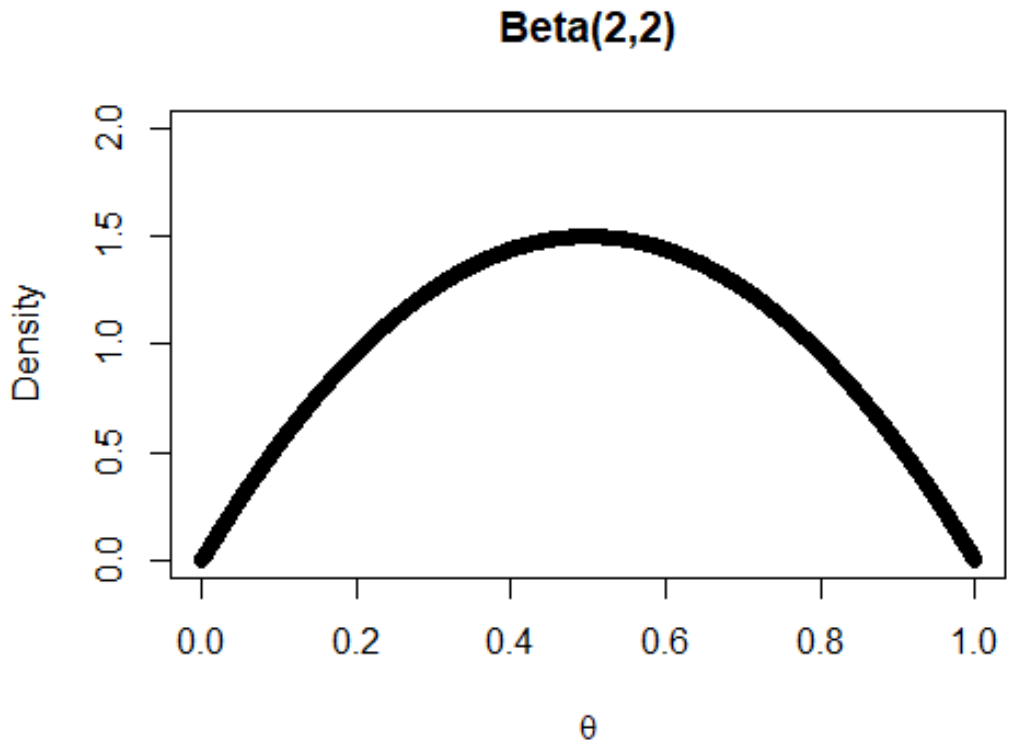
In both cases, we are assuming that the data are independent and identically distributed, with $p(X|\theta) \sim$ Bernoulli(θ). The goal: conducting inference on the unknown θ.

The maximum likelihood estimator is $\bar{X}$. Once data are observed, estimate is $\bar{x}$.

In Bayesian estimation, a prior distribution for θ is chosen representing beliefs of what values θ might be **before** observing the data. Given no other information, one might suppose that there is a better chance that θ is somewhere around 0.5 as opposed to 0.1 or 0.9.

Example Continued

A prior distribution should reflect this belief and be consistent with the structure of the problem (that is, $p(\theta) > 0$ only in $[0, 1]$). The calculations are simplified if the prior is conjugate, as discussed before. For this example, a beta(2,2) satisfies these beliefs.



Example Continued

Tossing the coin 10 times gives the sequence

$$H, H, T, T, T, T, T.H, H, T,$$

and we are interested in the $\theta = P(\text{heads})$. The MLE estimate is $\hat{\theta}_{MLE} = \bar{Y} = 0.40$.

In the Bayesian context, we know that with the observed data and Beta(2,2) prior, the posterior distribution is:

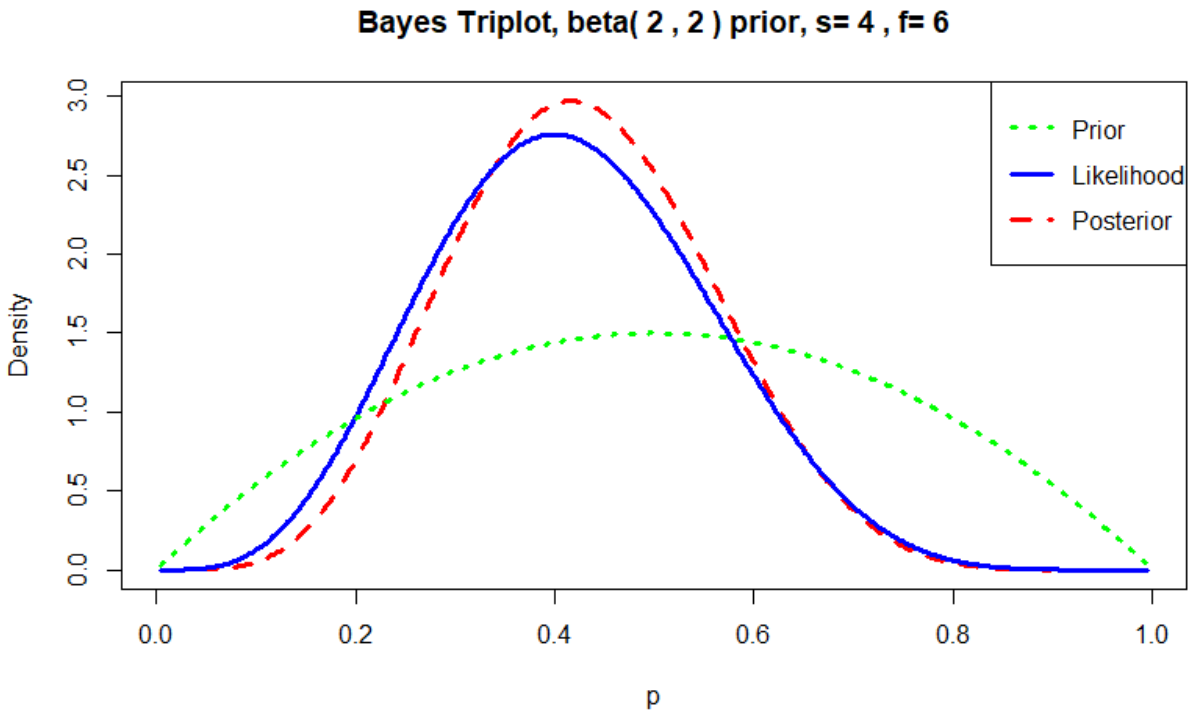
$$\theta|Y \sim Beta(6, 8)$$

Another reasonable estimator of θ might be the posterior mean:

$$\hat{\theta}_B = E[\theta|y] = \frac{a + y}{a + b + n} = \frac{6}{14} \approx 0.43$$

.

Example Continued



Revisiting MSE Example from Day 4

How do we know which estimator, $\hat{\theta}_{MLE}$ or $\hat{\theta}_B$ is "better"? Might consider calculating MSE:

$$\begin{aligned} MSE(\hat{\theta}_B) &= [Bias\hat{\theta}_B]^2 + Var(\hat{\theta}_B) \\ &= \left(\frac{a-(a+b)\theta}{n+a+b}\right)^2 + \left(\frac{n}{n+a+b}\right)^2 \frac{\theta(1-\theta)}{n} \end{aligned}$$

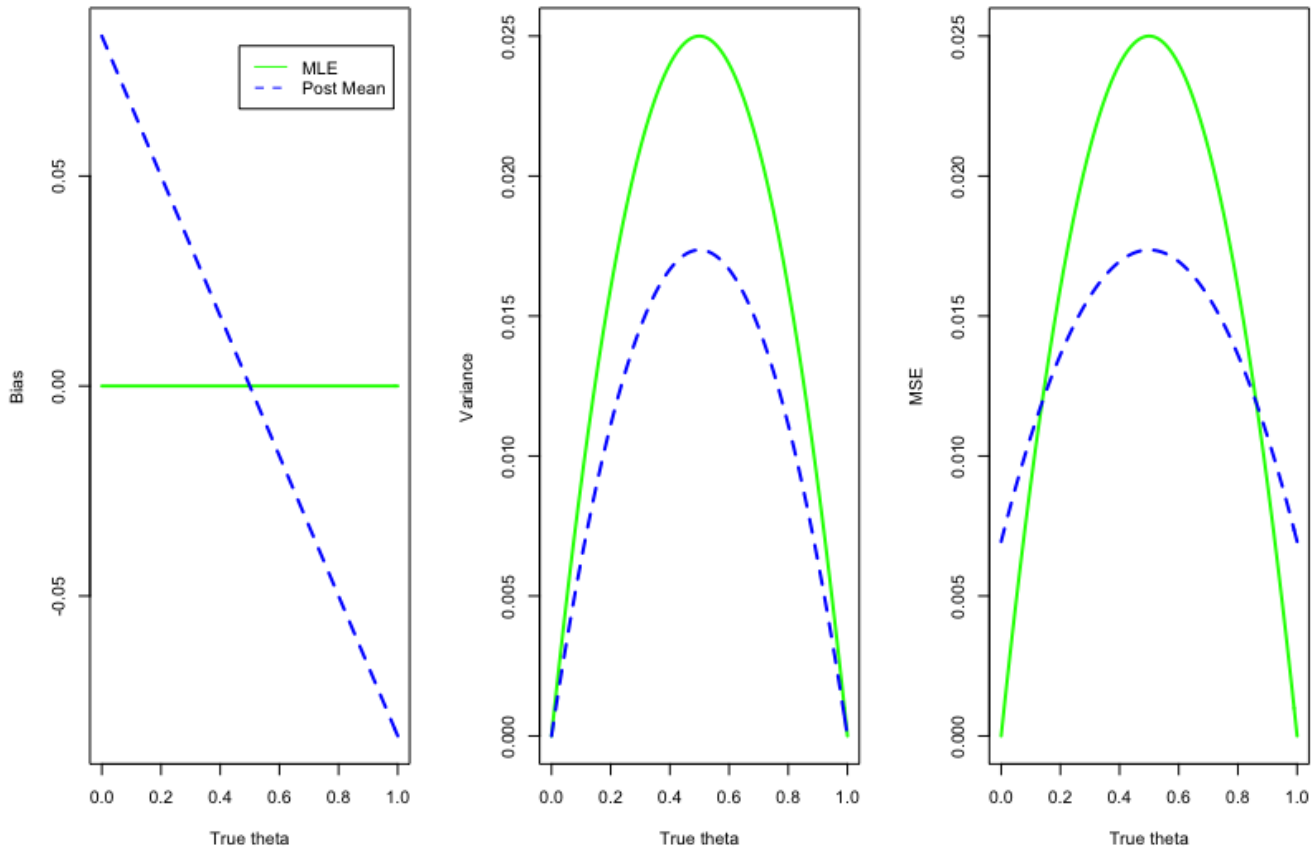
Recall that for this same data, we found the MSE of $\hat{\theta}_{MLE}$ to be

$$MSE(\hat{\theta}_{MLE}) = \frac{\theta(1-\theta)}{n}$$

.

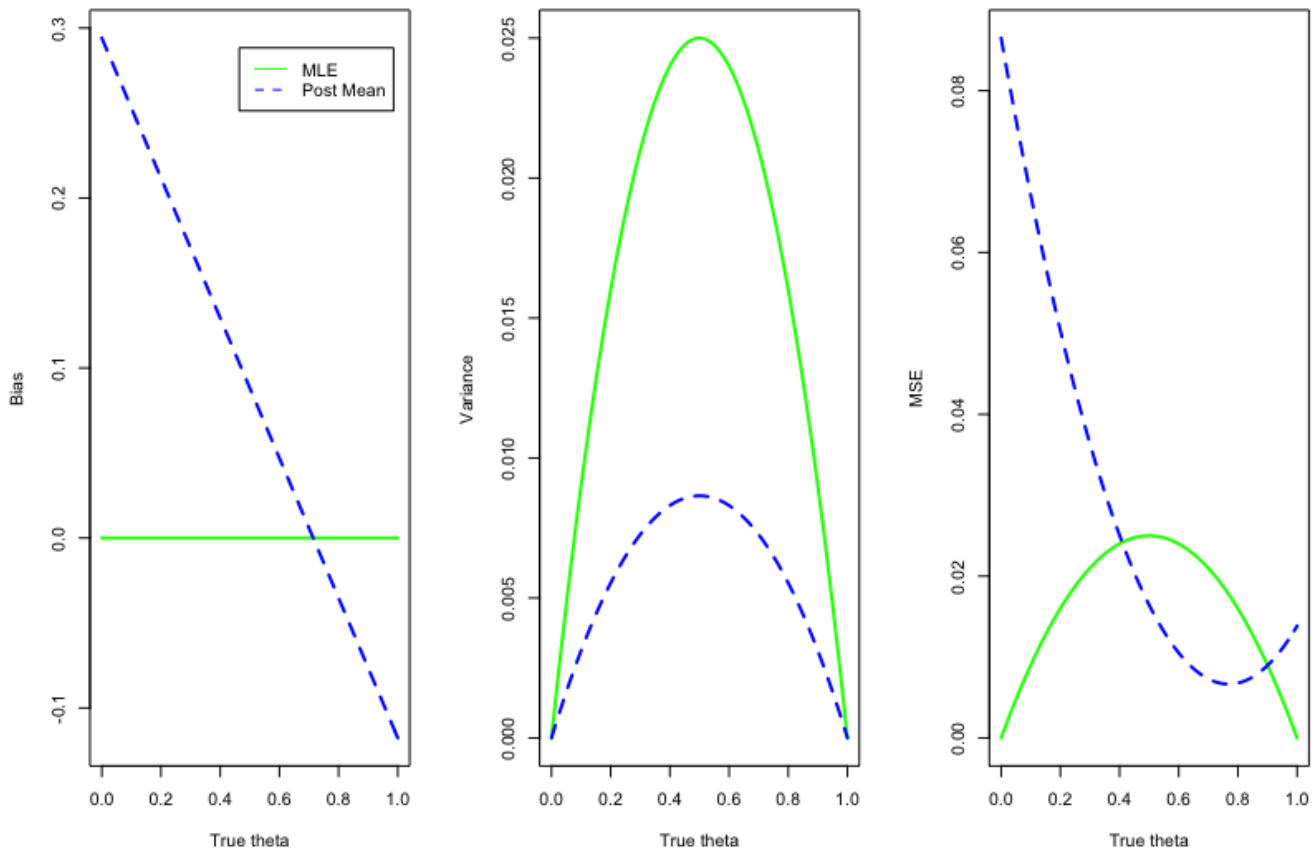
Comparing estimators

In terms of MSE, does one estimator always outperform the other? Not necessarily! With these two estimators, let us first consider a Beta(1,1) prior on θ and $n = 10$.



MSE Example Continued

Here, with a Beta(5,2) prior on θ and still $n = 10$.



Summary of Differences

- In the example, we decided to model the data as Binomial.
- In Maximum Likelihood Estimation, we find the value $\hat{\theta}$ that maximizes the likelihood of the data. This single point has the property that it is consistent and asymptotically normal.
- In Bayesian estimation, the likelihood of the data is combined with prior knowledge to produce a posterior distribution for θ. This probability distribution can be used to describe many features of $\theta|y_1...y_n$.

Bayesian Confidence Region

- As discussed before, in the frequentist context, a **random interval** has 95% coverage for θ if, before data are gathered,

$$Pr(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})|\theta) \geq 0.95$$

- An **interval**, based on observed data, has 95% Bayesian coverage for the **random variable** θ if

$$Pr(L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})|\mathbf{x}) \geq 0.95$$

- The two main types of confidence regions in Bayesian Analysis are (1) quantile-based regions and (2) highest posterior density regions, both of which will be discussed in detail in STA 601.