STA 601: Lecture 2

Loss Functions, Bayes Risk and Posterior Summaries

Merlise Clyde

Last Time ...

- Introduction to "ingredients" of Bayesian analysis
- Illustrated a simple Beta-Binomial conjugate example
- Posterior $\pi(\theta \mid y)$ is a $\textsf{Beta}(a+y, b+n-y)$

Today ...

- an introduction to loss functions
- Bayes Risk
- optimal decisions and estimators

Bayes estimate

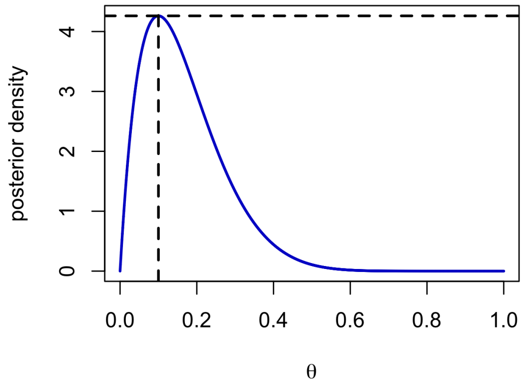
- As we've seen by now, having posterior distributions instead of one-number summaries is great for capturing uncertainty.
- That said, it is still very appealing to have simple summaries, especially when dealing with clients or collaborators from other fields, who desire one.

1) What if we want to produce a single "best" estimate of θ?

2) What if we want to produce an interval estimate (θ_L, θ_U)?

These would provide alternatives to the frequentist MLEs and confidence intervals

Heuristically



1) "best" estimate of θ is the maximum *a posteriori estimate* (**MAP**) or posterior mode

- what do we really mean by "best"?

2) find an interval such that $P(\theta \in (\theta_L, \theta_U) \mid y) = 1 - \alpha$

- lots of intervals that satisfy this! which one is "best"?

Loss Functions for Estimators

Introduce loss functions for decision making about what to report!

- a loss function provides a summary for how bad an estimator $\hat{\theta}$ is relative to the "true" value of θ
- Squared error loss $(L2)$

$$l(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

- Absolute error loss $(L1)$

$$l(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

But how do we deal with the fact that we do not know θ?

Bayes Risk

- **Bayes risk** is defined as the expected loss of using $\hat{\theta}$ averaging over the posterior distribution.

$$R(\hat{\theta}) = \mathsf{E}_{\pi(\theta|y)}[l(\theta, \hat{\theta})]$$

- the **Bayes optimal estimate** $\hat{\theta}$ is the estimator that has the lowest posterior expected loss or Bayes Risk

- Depends on choice of loss function

- **Frequentist risk** also exists for evaluating a given estimator under true value of θ

$$\mathsf{E}_{p(y|\theta_{\text{true}})}[l(\theta_{\text{true}}, \hat{\theta}))]$$

Squared Error Loss

A common choice for point estimation is squared error loss:

$$R(\hat{\theta}) = \mathsf{E}_{\pi(\theta|y)}[l(\theta, \hat{\theta})] = \int_{\Theta} (\hat{\theta} - \theta)^2 \pi(\theta \mid y) \, d\theta$$

- Expand and take derivative of $R(\hat{\theta}$ with respect to $\hat{\theta}$)

Let's work it out!

Steps

$$R(\hat{\theta}) = \int_{\Theta} (\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2)\pi(\theta \mid y)\, d\theta$$

$$R(\hat{\theta}) = \hat{\theta}^2 - 2\hat{\theta} \int_{\Theta} \theta\pi(\theta \mid y)\, d\theta + \int_{\Theta} \theta^2\pi(\theta \mid y)\, d\theta$$

$$R(\hat{\theta}) = \hat{\theta}^2 - 2\hat{\theta}\mathsf{E}[\theta \mid y] + \mathsf{E}[\theta^2 \mid y]$$

- Quadratic in $\hat{\theta}$ minimized when $\hat{\theta} = \mathsf{E}[\theta \mid y]$
- Posterior mean is the **Bayes optimal estimator** for θ under squared error loss
- In the beta-binomial case for example, the optimal Bayes estimate under squared error loss is

$$\hat{\theta} = \frac{a + y}{a + b + n},$$

What about other loss functions?

- Clearly, squared error is only one possible loss function. An alternative is **absolute loss**, which has

$$l(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

- Absolute loss places less of a penalty on large deviations & the resulting Bayes estimate is the **posterior median**.

- Median is actually relatively easy to estimate.

- Recall that for a continuous random variable Y with cdf F, the median of the distribution is the value z, which satisfies

$$F(z) = \Pr(Y \leq z) = \frac{1}{2} = \Pr(Y \geq z) = 1 - F(z).$$

- As long as we know how to evaluate the CDF of the distribution we have, we can solve for z.

Beta-Binomial

- For the beta-binomial model, the CDF of the beta posterior can be written as

$$F(z) = \Pr(\theta \leq z|y) = \int_0^z \text{Beta}(\theta|a+y, b+n-y)d\theta.$$

- Then, if $\hat{\theta}$ is the median, we have that $F(\hat{\theta}) = 0.5$.
- To solve for $\hat{\theta}$, apply the inverse CDF $\hat{\theta} = F^{-1}(0.5)$.
- In R, that's simply

```
qbeta(0.5,a+y,b+n-y)
```

- For other popular distributions, switch out the beta.

Loss Functions in General

- A **loss function** $l(\theta, \delta(y))$ is a function of the parameter θ and $\delta(y)$ based on just the data y
- For example, $\delta(y) = \bar{y}$ can be the decision to use the sample mean to estimate θ, the true population mean.
- $l(\theta, \delta(y))$ determines the penalty for making the decision $\delta(y)$, if θ is the true parameter or state of nature; the loss function characterizes the price paid for errors.
- Bayes optimal estimator or action is the estimator/action that minimizes the expected posterior loss marginalizing out any unknowns over posterior/predictive distribution.

MAP Estimator

- What about the MAP estimator? Is it an optimal Bayes estimator & under what choice of loss function?

- L_∞ loss:

$$R_\infty(\hat{\theta}) = \lim_{p \to \infty} \int_\Theta (\theta - \hat{\theta})^p \pi(\theta \mid y)\, d\theta$$

- Essentially saying that we need the estimator to be right on the truth or the error blows up!

- Is this a reasonable loss function?

Interval Estimates

Recall that a frequentist confidence interval $[l(y),\ u(y)]$ has 95% frequentist coverage for a population parameter θ if, before we collect the data,

$$\Pr[l(y) < \theta < u(y)|\theta] = 0.95.$$

- This means that 95% of the time, our constructed interval will cover the true parameter, and 5% of the time it won't.

- There is NOT a 95% chance your interval covers the true parameter once you have collected the data.

- In any given sample, you don't know whether you're in the lucky 95% or the unlucky 5%. You just know that either the interval covers the parameter, or it doesn't (useful, but not too helpful clearly).

- Often based on aysmptotics i.e use a Wald or other type of frequentist asymptotic interval $\hat{\theta} \pm 1.96\,\mathrm{se}(\hat{\theta})$

Bayesian Intervals

- We want a Bayesian alternative to confidence intervals

for some pre-specified value of α

- An interval $[l(y),\ u(y)]$ has $1-\alpha$ 100% Bayesian coverage for θ if

$$\Pr(\theta \in [l(y),\ u(y)] \mid y) = 1 - \alpha$$

- This describes our information about where θ lies *after* we observe the data.

- Fantastic! This is actually the interpretation people want to give to the frequentist confidence interval.

- Bayesian interval estimates are often generally called **credible intervals** or **credible sets**.

How to choose $[l(y),\ u(y)]$?

Bayesian Equal Tail Interval

- The easiest way to obtain a Bayesian interval estimate is to use posterior quantiles with **equal tail areas**. Often when researchers refer to a credible interval, this is what they mean.

- To make a $100 \times (1 - \alpha)$ equi-tail quantile-based credible interval, find numbers (quantiles) $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

  1. $\Pr(\theta < \theta_{\alpha/2} \mid y) = \frac{\alpha}{2}$; and

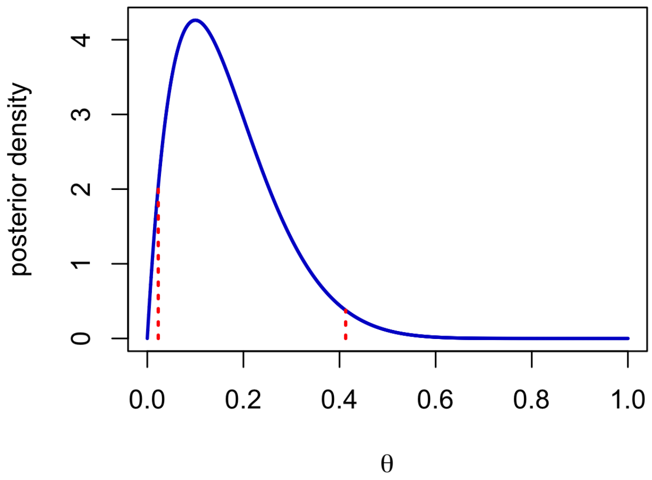
  2. $\Pr(\theta > \theta_{1-\alpha/2} \mid y) = \frac{\alpha}{2}$.

Convenient conceptually and easy as we just take the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\pi(\theta \mid y)$ as $l(y)$ and $u(y)$, respectively.

Beta-Binomial Equal-tailed Interval

```
a = 1; b= 1; y = 1; n = 10
ly = qbeta(0.025, a + y, b + n - y)
uy = qbeta(0.975, a + y, b + n - y)
c(ly, uy)
```

```
## [1] 0.0228312 0.4127799
```



Monte Carlo Version

- Suppose we don't have $\pi(\theta \mid y)$ is a simple form, but we do have samples $\theta_1, \ldots, \theta_T$ from $\pi(\theta \mid y)$

- We can use these samples to obtain Monte Carlo (MC) estimates of posterior summaries

$$\hat{\theta} = \mathsf{E}[\theta \mid y] \approx \frac{1}{T} \sum_{t=1}^{T} \theta_t$$

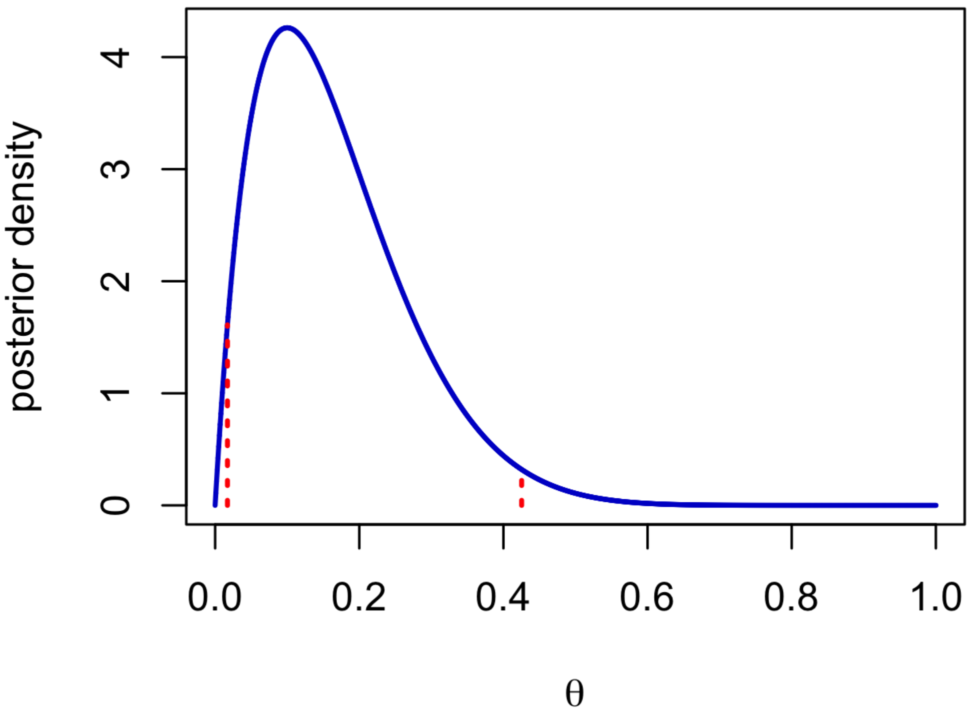
- what about MC quantile estimates?

- Find the 2.5th and 97.5th percentile from the empirical distribution

```
theta = rbeta(1000, a + y, b + n - y)
quantile(theta, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.01710151 0.42527760
```

Equal-Tail Interval



Note there are values of θ outside the quantile-based credible interval, with higher density than some values inside the interval.

HPD region

- A $100 \times (1-\alpha)$ highest posterior density (HPD) region is a subset $s(y)$ of the parameter space Θ such that

    1. $\Pr(\theta \in s(y) \mid y) = 1-\alpha$; and

    2. If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$, then $p(\theta_a \mid y) > p(\theta_b \mid y)$ (highest density set)

- $\Rightarrow$ **All** points in a HPD region have higher posterior density than points outside the region.

- The basic idea is to gradually move a horizontal line down across the density, including in the HPD region all values of θ with a density above the horizontal line.

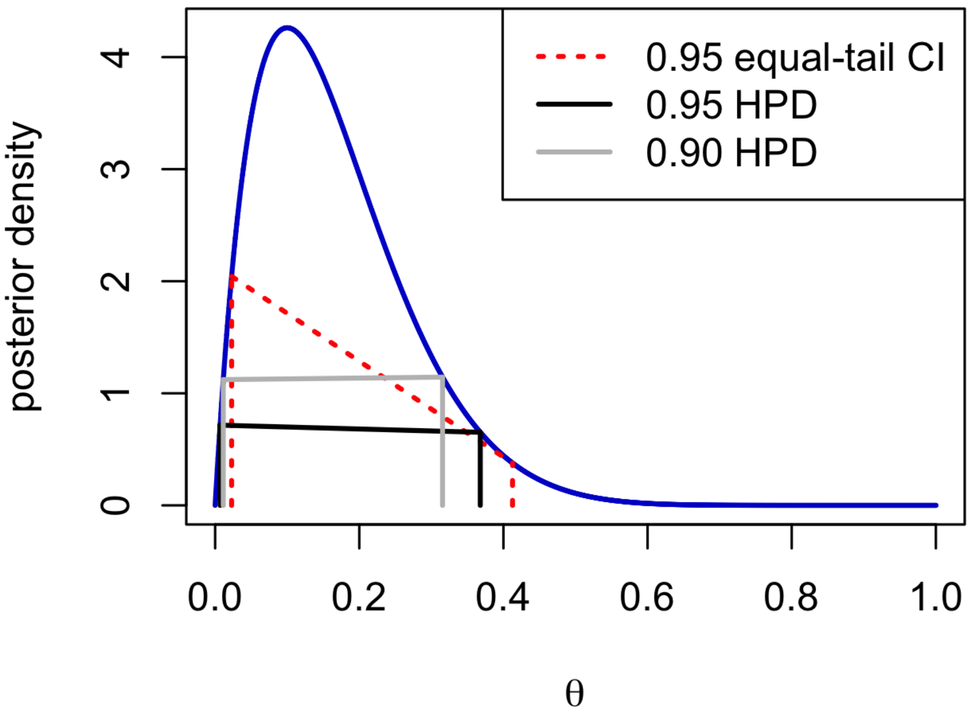
- Stop moving the line down when the posterior probability of the values of θ in the region reaches $1-\alpha$.

Simulation Based

```
suppressMessages(library(rjags))
HPDinterval(as.mcmc(theta))
```

```
##           lower     upper
## var1 0.01010097 0.3697792
## attr(,"Probability")
## [1] 0.95
```

HPD Intervals



Properties of HPD Sets

- Shortest length interval (or volume) for the given coverage
- Equivalent to Equal-Tail Intervals if the posterior is unimodal and symmetric
- May not be an interval if the posterior distribution is multi-modal
- In general, not invariant under monotonic transformations of θ
- More computationally intensive to solve!

Loss Functions for Interval Estimation

See "The Bayesian Choice" by Christian Robert Section 5.5.5

Connections between Bayes and MLE Based Frequentist Inference

Berstein von Mises (BvM) Theorems aka Bayesian Central Limit Theorems

- examine limiting form of the posterior distribution $\pi(\theta \mid y)$ as $n \to \infty$
- $\pi(\theta \mid y)$ goes to a Gaussian under regularity conditions
  - centered at the MLE
  - variance given by the inverse of the Expected Fisher Information (var of MLE)
- The most important implication of the BvM is that Bayesian inference is asymptotically correct from a frequentist point of view
- Used to justify Normal Approximations to the posterior distribution (eg Laplace approximations)

Model Misspecification ?

- We might have chosen a bad sampling model/likelihood
- posterior still converges to a Gaussian centered at the MLE under the misspecified model, but wrong variance
- 95% Bayesian credible sets do not have correct frequentist coverage
- See Klein & van der Vaart for more rigorous treatment if interested
- parametric model is "close" to the true data-generating process
- model diagnostics & changing the model can reduce the gap between model we are using and the true data generating process