Bayesian Estimation

Jacob M. Montgomery

2018

Bayesian point estimation

Overview

- Last we talked about
  - “Simple” methods to make inferences using this approach
  - Some advanced approaches applicable both here and in MLE (the delta method and the parametric bootstrap)
- This time we are going to talk about Bayesian inference

Bayesian thinking

- Bayes' Theorem marks the beginnings of serious statistical inference.

Bayesian thinking

- Bayes' Theorem marks the beginnings of serious statistical inference.
- For many years Bayesian statistics was a backwater of statistics.
- However, as we have moved into the computer age, the popularity of Bayesian inference has waxed markedly.

The big picture

- Like the other approaches mentioned so far, we assume that the true DGP can be characterized by a parametric function

$$\mathcal{F} = \{f(x|\theta) : x \in \mathcal{X}, \theta \in \Theta\}.$$

The big picture

- Like the other approaches mentioned so far, we assume that the true DGP can be characterized by a parametric function

$$\mathcal{F} = \{f(x|\theta) : x \in \mathcal{X}, \theta \in \Theta\}.$$

- Here x is the observed data, $\mathcal{X}$ is the sample space.
- We think of θ as some point in the possible parameter space Θ.
- The basic idea is that we observe x generated by $f(x|\theta)$ and infer the value of θ.

Adding prior beliefs: The “drawback”

- The big difference is that we add to our model *prior beliefs* about the probable values of θ, which is characterized by a prior density

$$\pi(\theta).$$

Adding prior beliefs: The "drawback"

- The big difference is that we add to our model *prior beliefs* about the probable values of θ, which is characterized by a prior density
$$\pi(\theta).$$

- The basic idea is that "update" our prior beliefs about θ as we observe more data x.

Adding prior beliefs: The "drawback"

- The big difference is that we add to our model *prior beliefs* about the probable values of θ, which is characterized by a prior density
$$\pi(\theta).$$
- The basic idea is that "update" our prior beliefs about θ as we observe more data x.
- A formal statement of Bayes' Rule in this context is:
$$p(\theta|x) = \pi(\theta)\frac{f(x|\theta)}{f(x)}$$
where $f(x)$ is the marginal distribution of x
$$f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

Bayesian inference: The way you have always wanted to think about data

- In this setting x is fixed observed data.

Bayesian inference: The way you have always wanted to think about data

- In this setting x is fixed observed data. Inferences are *not* made based on asymptotic distributions of any statistic $t(x)$, the likelihood, or imagined repeated samples.

Bayesian inference: The way you have always wanted to think about data

- In this setting x is fixed observed data. Inferences are *not* made based on asymptotic distributions of any statistic $t(x)$, the likelihood, or imagined repeated samples.
- Instead, imagine that θ is some true parameter whose value we do not know.
- We use Bayes' formula to update our beliefs about θ.

Bayesian inference: The way you have always wanted to think about data

- In this setting x is fixed observed data. Inferences are *not* made based on asymptotic distributions of any statistic $t(x)$, the likelihood, or imagined repeated samples.
- Instead, imagine that θ is some true parameter whose value we do not know.
- We use Bayes' formula to update our beliefs about θ.
- Note that this is the **exact opposite of frequentist statistics** where we have assumed that θ is some fixed (but unknown) parameter and all inferences are generated by treating $t(x)$ as a random variable.

Bayesian inference in practice

- So we want to set up:

$$p(\theta|x) = \pi(\theta)\frac{f(x|\theta)}{f(x)}$$

- This can be re-written as:

$$p(\theta|x) = \pi(\theta)\frac{L(\theta)}{f(x)}$$

- Which can further be re-written as

$$p(\theta|x) = c_x \pi(\theta) L(\theta)$$

- How can we figure this out?:

$$p(\theta|x) = c_x \pi(\theta) L(\theta)$$

- How can we figure this out?:

$$p(\theta|x) = c_x \pi(\theta) L(\theta)$$

- The key here is that we can rarely directly carry out this operation

$$1/c_x = f(x) = \int_\Theta f(x|\theta)\pi(\theta) d\theta$$

- Instead, we use the knowledge that $p(\theta|x)$ must integrate to one.

- How can we figure this out?:

$$p(\theta|x) = c_x \pi(\theta) L(\theta)$$

- The key here is that we can rarely directly carry out this operation

$$1/c_x = f(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$$

- Instead, we use the knowledge that $p(\theta|x)$ must integrate to one.
- The common approach for today's class will be:
  - Write out the likelihood
  - Multiply it by a (carefully chosen) prior
  - Combine the two and think about θ as being the random variable.
  - *See* that resulting formula is the "kernel" of some known probability distribution.

Example: Bernoulli

Let $X_1, \ldots, X_n \sim Bern(p)$. Let $s = \sum X_i$. Suppose we take the uniform distribution as a prior $\pi(p) = 1$.

Example: Bernoulli

Let $X_1, \ldots, X_n \sim Bern(p)$. Let $s = \sum X_i$. Suppose we take the uniform distribution as a prior $\pi(p) = 1$.

$$\begin{aligned} p(p|\mathbf{x}) &\propto \pi(p)L(p) \\ &= p^s(1-p)^{n-s} \end{aligned}$$

Example: Bernoulli

Let $X_1, \ldots, X_n \sim Bern(p)$. Let $s = \sum X_i$. Suppose we take the uniform distribution as a prior $\pi(p) = 1$.

$$
\begin{aligned}
p(p|\mathbf{x}) &\propto \pi(p)L(p) \\
&= p^s(1-p)^{n-s}
\end{aligned}
$$

- The difficult part here is to adjust your mind to see that our random variable is no longer s but instead p.
- We need to "see" that this is the kernel of some known distribution.

- This is what we have

$$p^s(1-p)^{n-s}$$

- This is what we have

$$p^s(1-p)^{n-s}$$

- If some variable y is distributed according to a $Beta(\alpha, \beta)$ distribution, then the pdf is:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

.

$$\propto y^{\alpha-1}(1-y)^{\beta-1}$$

- This is what we have

$$p^s(1-p)^{n-s}$$

- If some variable y is distributed according to a $Beta(\alpha, \beta)$ distribution, then the pdf is:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

.

$$\propto y^{\alpha-1}(1-y)^{\beta-1}$$

- Keeping in mind that p in the top formula takes the place of y in the bottom formula, what is the posterior distribution of p?

What was the integration constant?

- We now showed that

$$p|\mathbf{x} \sim Beta(s+1, n-s+1).$$

What was the integration constant?

- We now showed that

$$p|\mathbf{x} \sim Beta(s+1, n-s+1).$$

- This means that the full posterior distribution was:

$$\frac{\Gamma(s+1+n-s+1)}{\Gamma(s+1)\Gamma(n-s+1)} p^{s+1-1}(1-p)^{n-s+1-1}$$

What was the integration constant?

- We now showed that

$$p|\mathbf{x} \sim Beta(s+1, n-s+1).$$

- This means that the full posterior distribution was:

$$\frac{\Gamma(s+1+n-s+1)}{\Gamma(s+1)\Gamma(n-s+1)} p^{s+1-1}(1-p)^{n-s+1-1}$$

- But the kernel we started out with was only

$$p^s(1-p)^{(n-s)}.$$

- So what was the integration constant $c_x = f(x)$?

Making a point estimate

- Now that we have found the posterior distribution of θ, making an estimate is easy.

Making a point estimate

- Now that we have found the posterior distribution of θ, making an estimate is easy.
- Let $p(\theta|\mathbf{x})$ be the posterior distribution of θ. The point estimate, $\hat{\theta}$ is just the first central moment of $p(\theta|\mathbf{x})$, $E(\theta)$.

Making a point estimate

- Now that we have found the posterior distribution of θ, making an estimate is easy.
- Let $p(\theta|\mathbf{x})$ be the posterior distribution of θ. The point estimate, $\hat{\theta}$ is just the first central moment of $p(\theta|\mathbf{x})$, $E(\theta)$.
- Alternatively, we might want to use some of our MLE methods to find the posterior mode. This would be the *modal a posteriori* (MAP) estimate.

Creating a an interval estimate

- To create a *credible interval* we need to find a and b such that

$$\int_{-\infty}^{a} p(\theta|\mathbf{x})d\theta = \int_{b}^{\infty} p(\theta|\mathbf{x})d\theta = \alpha/2$$

- If we can find this, then we can have an interval $C = (a, b)$ such that

$$P(\theta \in C|\mathbf{x}) = \int_{a}^{b} p(\theta|\mathbf{x})d\theta = 1 - \alpha$$

Creating a an interval estimate

- To create a *credible interval* we need to find a and b such that

$$\int_{-\infty}^{a} p(\theta|\mathbf{x})d\theta = \int_{b}^{\infty} p(\theta|\mathbf{x})d\theta = \alpha/2$$

- If we can find this, then we can have an interval $C = (a, b)$ such that

$$P(\theta \in C|\mathbf{x}) = \int_{a}^{b} p(\theta|\mathbf{x})d\theta = 1 - \alpha$$

- This will create a CI centered around the posterior mean.

Creating a an interval estimate

- To create a *credible interval* we need to find a and b such that

$$\int_{-\infty}^{a} p(\theta|\mathbf{x})d\theta = \int_{b}^{\infty} p(\theta|\mathbf{x})d\theta = \alpha/2$$

- If we can find this, then we can have an interval $C = (a, b)$ such that

$$P(\theta \in C|\mathbf{x}) = \int_{a}^{b} p(\theta|\mathbf{x})d\theta = 1 - \alpha$$

- This will create a CI centered around the posterior mean.
- An alternative is to create a Highest Posterior Density interval centered around the posterior mode(s).
- Both methods will typically be done numerically.

Example: Bernoulli

- We previously established that the posterior distrbution is

$$p(p|\mathbf{x}) \sim Beta(s+1, n-s+1)$$

Example: Bernoulli

- We previously established that the posterior distrbution is

$$p(p|\mathbf{x}) \sim Beta(s+1, n-s+1)$$

- The mean of the $Beta(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$ so our point estimate is

$$\hat{p} = \frac{s+1}{n+2}$$

Example: Bernoulli

- We previously established that the posterior distrbution is

$$p(p|\mathbf{x}) \sim Beta(s+1, n-s+1)$$

- The mean of the $Beta(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$ so our point estimate is

$$\hat{p} = \frac{s+1}{n+2}$$

- Notice that this is slighly off from the MLE we established.

Example: Bernoulli

- We previously established that the posterior distrbution is

$$p(p|\mathbf{x}) \sim Beta(s+1, n-s+1)$$

- The mean of the $Beta(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$ so our point estimate is

$$\hat{p} = \frac{s+1}{n+2}$$

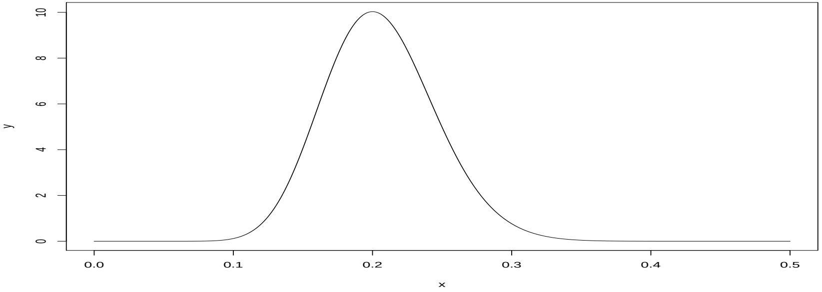
- Notice that this is slighly off from the MLE we established.
- However, notice also that this difference will diminish as $n \to \infty$
- Let $\lambda = n/(n+2)$, $\bar{x}$ be the MLE, and p^* be the prior mean $(1/2)$. Then $\hat{p} = \lambda\bar{x} + (1-\lambda)p^*$

- Now we need to figure out the credible interval
- Let's say that $n = 100$ and $s = 20$
- So

```
draws<-rbeta(n = 10000, shape1 = 20+1,
             shape2 = 100-20+1)
quantile(draws, probs = c(.025, .975))
```

```
##       2.5%     97.5%
## 0.1340576 0.2896866
```

```
x<-seq(0, .5, by=.001)
y<-dbeta(x, shape1 = 20+1,
          shape2 = 100-20+1)
plot(x, y, type="l")
```



```r
library(HDInterval)
draws<-rbeta(n = 10000, shape1 = 20+1,
             shape2 = 100-20+1)
hdi(draws, credMass=0.95)
```

```
##     lower     upper
## 0.1262027 0.2832161
## attr(,"credMass")
## [1] 0.95
```

Class Exercise

Let our data $X_1, \ldots, X_n$ be iid Poisson(λ). We assume that the prior distribution be a gamma distribution such that

$$\pi(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- Find the posterior distribution for λ
- Find the point estimate for λ
- Find the 95% credible interval for λ.

Posteriors for functions of parameters

- One of the nice features of the Bayesian approaches is that calculating posterior distributions for transformations of parameters is almost trivially easy.

Posteriors for functions of parameters

- One of the nice features of the Bayesian approaches is that calculating posterior distributions for transformations of parameters is almost trivially easy.
- Here is the process:
  1. Calculate the posterior distribution for the parameter.
  2. Simulate out of the posterior.
  3. Apply the transformation to the simulated parameters.
  4. Construct the credible interval (and even a point estimate) from this simulated sample.

Example: Log odds

- Imagine that in our previous example we are interested not in p but in the log odds:
$$log(\frac{p}{1-p})$$
- Take the code from our last example, and estimate the postior for this transformed parameter.

Example: Log odds

- Imagine that in our previous example we are interested not in p but in the log odds:
$$log(\frac{p}{1-p})$$
- Take the code from our last example, and estimate the postior for this transformed parameter.
- How would we make a posterior predictive interval?

Large sample properties of Bayesian statistics

- Let $\hat{\theta}$ be the MLE and let $\sigma_{\hat{\theta}}$ be the standard error,

$$\frac{1}{\sqrt{nI(\hat{\theta})}}.$$

Under appropriate regularity conditions.

Large sample properties of Bayesian statistics

- Let $\hat{\theta}$ be the MLE and let $\sigma_{\hat{\theta}}$ be the standard error,

$$\frac{1}{\sqrt{nI(\hat{\theta})}}.$$

  Under appropriate regularity conditions.
- The posterior will be approximately normal with mean $\hat{\theta}$ and standard deviation $\sigma_{\hat{\theta}}$.

Large sample properties of Bayesian statistics

- Let $\hat{\theta}$ be the MLE and let $\sigma_{\hat{\theta}}$ be the standard error,

$$\frac{1}{\sqrt{nI(\hat{\theta})}}.$$

  Under appropriate regularity conditions.
- The posterior will be approximately normal with mean $\hat{\theta}$ and standard deviation $\sigma_{\hat{\theta}}$.
- Thus, in asymptotic terms, the Bayesian posterior *will be exactly the same* as the asymptotic distribution of the MLE.

Large sample properties of Bayesian statistics

- Let $\hat{\theta}$ be the MLE and let $\sigma_{\hat{\theta}}$ be the standard error,

$$\frac{1}{\sqrt{nI(\hat{\theta})}}.$$

  Under appropriate regularity conditions.
- The posterior will be approximately normal with mean $\hat{\theta}$ and standard deviation $\sigma_{\hat{\theta}}$.
- Thus, in asymptotic terms, the Bayesian posterior *will be exactly the same* as the asymptotic distribution of the MLE.
- The differences between the approaches occur in finite samples.

Jargon Alert!: Types of priors

- Conjugate priors
- Informative priors
- Flat/noninformative priors
- Improper priors
- Jeffrey's priors

Bayesian statistics with multiple parameters

- Sometimes it's possible to divide the posterior so that we can *see* a distribution for one/both of the parameters

Bayesian statistics with multiple parameters

- Sometimes it's possible to divide the posterior so that we can *see* a distribution for one/both of the parameters
- In these cases the posteriors are conditionally independent.

Bayesian statistics with multiple parameters

- Sometimes it's possible to divide the posterior so that we can *see* a distribution for one/both of the parameters
- In these cases the posteriors are conditionally independent.
- Sometimes this calculation cannot be done, and we will have to give up on solving the problem analytically.
- Instead we will rely on more advanced algorithms we cover later in this class.
    - Gibbs sampler
    - Metropolis-hastings.
- We will return to these issues when we tackle the Bayesian t-test.

Multiple parameters

- Wasserman 11.7
- Overview of normal-gamma problem