STA 250 Lecture 4

Qi GAO

October 12, 2013

Bayesian Inference

- Based on Bayes' Theorem.
- Different from classical inference by regarding parameters as random variables.

Introduction

Interpretation of a confidence interval: under repeated sampling, $100(1-\alpha)\%$ of confidence intervals would contain θ.

$\rightsquigarrow$We would prefer to say things like "there is a 95% chance that θ is between 0.1 and 0.9".

Idea: The likelihood function $p(y|\theta)$ can be interpreted as $p(data|parameters)$.

Goal: We are interested to know $p(\theta|y)$ i.e., $p(parameter|data)$.

Here θ becomes a random variable!

Using Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

. Therefore, to be able to get $p(\theta|y)$, we need

(1) likelihood $p(y|\theta)$

(2) "prior" $p(\theta)$

Once we have these, basic probability rules allow us to get

(3) "posterior" $p(\theta|y)$.

Prior Distribution

A prior distribution $p(\theta)$ encodes an analyst's beliefs about what values of θ are plausible before seeing any data.

Example: Undergraduate GPA of STA 250 students. Since most students enrolled in STA

250 are graduate students, it is reasonable to believe that they have higher undergraduate GPA on average. Assuming that σ^2 is known, we can use a truncated normal distribution centered around say 3.6 to model μ.

How to specify priors?

Note: There is no unique/correct prior!

Two (or Three) Camps

(1) Subjective Bayes. Prior encodes the belief of the analyst.

(2) Objective Bayes. Priors are determined by formal rule/criteria.

(3) Pragmatic Bayes. In real world, we just do whatever works!

Formal Rules

(1) Reference priors (Bernardo, Berger. $\sim$ 1970)
Idea: Maximize the "distance" (e.g. KullbackCLeibler divergence) between the prior and the posterior. These priors have excellent properties, but it can be tricky to derive them for complex models.

(2) Probability matching priors (Welch&Peers. $\sim$ 1956)
Idea: Select a prior such that posterior distribution allows the construction of intervals with frequentist coverage (i.e., confidence intervals). Nice theory exists, but not practical (computation-intensive)!

(3) Invariance
Idea: Construct a rule such that the prior distributions constructed in different parametrizations are consistent.
E.g., For prior μ where $\mu \sim N(0,1)$, if we reparametrize it to $\theta = e^{\mu}$, prior on *theta* is transformed accordingly.
The most famous invariance prior is *Jefferey's Prior* where $p(\theta) \propto \|I(\theta)\|^{1/2}$. $I(\theta)$ is the Fisher information and in multivariate case, $\|I(\theta)\|^{1/2}$ is the square root of the determinant of $I(\theta)$. We are able to obtain equivalent priors after transformation. Considering the following two recipes:

Recipe 1 Derive Jeffrey's prior for θ.

Recipe 2 Derive Jeffrey's prior for μ, then transform via $\theta = e^{\mu}$ and find the induced prior in θ.

Using Jeffrey's prior, both recipes give the same answer.

Example: $y_i|\theta \overset{ind.}{\sim} Bin(n_i, \theta), i = 1, ..., m$. Need prior on θ.

(Aside: Recall $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$ since $p(y)$ does not involve θ.)

$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \theta^{y_i}(1-\theta)^{n_i - y_i} \propto \theta^{\sum y_i}(1-\theta)^{\sum(n_i - y_i)}$.

Posterior: $p(\theta|y) \propto p(\theta)\theta^{\sum y_i}(1-\theta)^{\sum(n_i - y_i)}$.

It turns out that $I(\theta) = \frac{\sum n_i}{\theta(1-\theta)}$.

$\Rightarrow$ Jeffrey's prior: $p(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2}$.

$\Rightarrow$ Posterior: $p(\theta|y) \propto \theta^{(\sum y_i) - 1/2}(1-\theta)^{\sum(n_i - y_i) - 1/2}$.

We need $\int p(\theta|y)d\theta = 1$.

Recall that if $X \sim Beta(a, b)$, then $p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} I_{\{0<x<1\}}$.

We have $p(\theta|y) \propto \theta^{(\sum y_i + 1/2) - 1}(1-\theta)^{[\sum(n_i - y_i) + 1/2] - 1}$.

$\Rightarrow \theta|y \sim Beta(\sum y_i + 1/2, \sum(n_i - y_i) + 1/2)$.

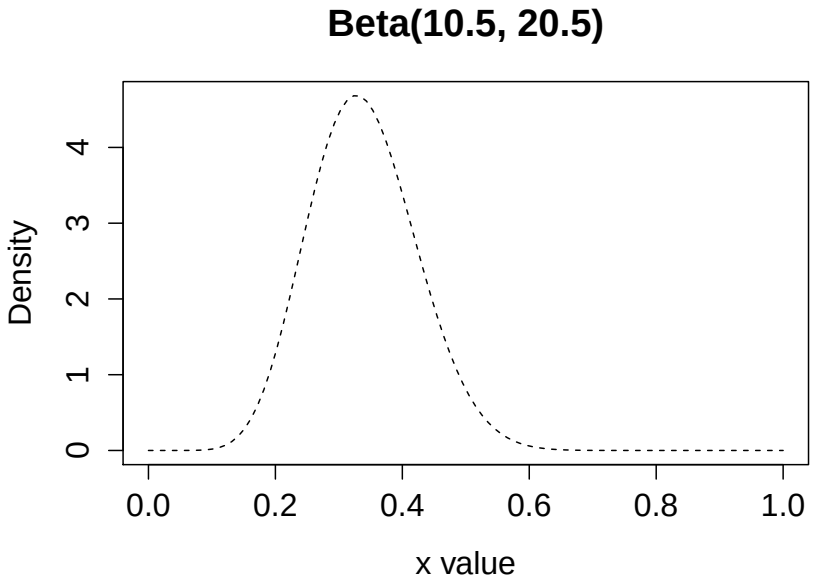
Jeffrey's prior is actually a $Beta(1/2, 1/2)$. It turns out that is we use a $Beta(a, b)$ as the prior, we get a posterior $\theta|y \sim Beta(a + \sum y_i, b + \sum(n_i - y_i))$.

We call a prior a **conjugate prior** if the posterior distribution remains in the same family as the prior.

E.g., Prior for θ was Beta, posterior was also Beta.

We have $\theta|y \sim Beta(\sum y_i + 1/2, \sum(n_i - y_i) + 1/2)$. Suppose I have data $\sum y_i = 10$, $\sum(n_i - y_i) = 20$.

$\Rightarrow$ Posterior is $\theta|y \sim Beta(10.5, 20.5)$. The plot of $p(\theta|y)$ is shown below.



Credibel Interval

To get a point estimate for θ, we can use:

(1) Posterior mean.

(2) Posterior median.

(3) posterior mode (can be hard to compute in practice).

We also need an uncertainty quantification/interval. In the Bayesian context, a posterior interval is known as a credible interval.
$S^{1-\alpha}(y)$ is defined to be a $100(1-\alpha)\%$ **credible interval** for θ if $\int_{S^{1-\alpha}(y)} p(\theta|y)d\theta = 1-\alpha$.
A central credible interval takes the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the posterior.
A highest posterior density (HPD) interval is an interval S such that

$$S = \{\theta : p(\theta) > p(\theta') \quad \forall \theta \in S, \theta' \notin S, \int_S p(\theta|y)d\theta = 1-\alpha\}.$$

An illustration of these two intervals can be found at `http://www.bayesian-inference.com/credible`.

Program Demonstration

The R code for binomial coverage simulation can be found at `https://github.com/STA250/Stuff/blob/master/Lecture_Code/Bayes/binom_coverage_sim.R`.
As we can see from the output of the simulation, if we use Jeffrey's prior, the frequentist coverage of Bayesian interval is reasonably good.