STA 250 Lecture Notes

10/9/2013

Recorded by Jeremy Werner

Announcements

- Approximately when Homework 1 is assigned, a tutorial on correctly syncing to github will be provided.

Bayesian Inference

Interpretation of confidence interval: Under repeated sampling, $100(1-\alpha)\%$ of confidence intervals would contain θ. We would prefer to say things like "there is a 95% chance that θ is between .1 and .9," for example.

Idea:

$$\begin{array}{rll} \underline{\text{Likelihood}} & P(y|\theta) & [P(\text{data}|\text{parameters})] \\ \underline{\text{Goal}} & P(\theta|y) & [P(\text{parameters}|\text{data})] \end{array}$$

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

To be able to get $P(\theta|y)$ we need:

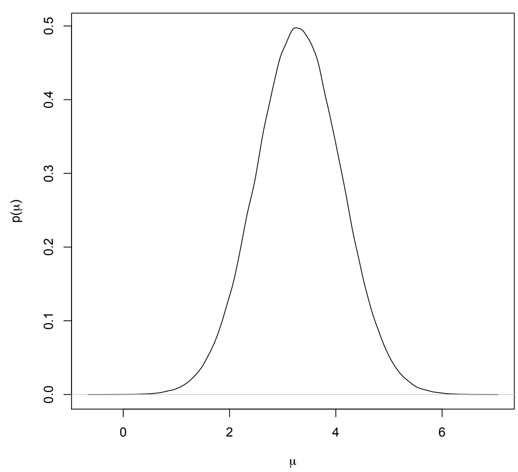
(1) Likelihood $P(y|\theta)$

(2) Prior $P(\theta)$

(3) Posterior $P(\theta|y)$ which follows from using basic probability rules once we have (1) and (2)

A prior distribution $P(\theta)$ encodes an analyst's beliefs about what values of θ are plausible before seeing any data. In other words, the prior is *"How you expect θ to be distributed before you see any data"*

Example: Undergraduate GPA of STA 250 students
Model $y_i|\theta \sim$ Truncated Normal$(\mu, \sigma^2, [0, 4])$
Assume σ^2 is known, a possible prior distribution could look like the graph below.



How can a prior be specified? First, it is necessary to note that no prior is unique or correct! With that being said, there are two (three) major *camps of belief/schools of thought.*

① *Subjective Bayes* ⇒ Prior encodes the belief of the analyst

② *Objective Bayes* ⇒ Priors are determined by formal rules/criteria

③ (*Pragmatic Bayes* ⇒ Do whatever works!)

Formal Rules

① Reference Priors (Bernardo, Berger ∼ 1979)

- *Idea:* Maximize the "distance" (e.g. KL divergence) between the prior and posterior.
  - Excellent properties, can be tricky to derive for complex models

② Probability Matching Prior (Welch, Peers ∼ 1963)

- *Idea:* Select a prior such that the posterior distribution allows the construction of intervals with frequentist coverage (i.e. confidence intervals)
  - Nice theory, not practical!

③ Invariance

- *Idea:* Construct a rule such that the prior distributions constructed in different parametrizations are consistent (e.g. prior μ : $\mu \sim N(0,1)$ Reparametrize to $\theta = e^{\mu}$ ⇒ Prior on θ is transformed accordingly)
  - The most famous invariant prior is Jeffreys Prior $P(\theta) \propto ||I(\theta)||^{1/2}$ where $I(\theta)$ is the Fisher Information and $||\cdot||$ is the determinant of the matrix for multivariate forms.

Recipe 1: Derive Jeffreys Prior for θ
Recipe 2: Derive Jeffreys Prior μ, then transform via $\theta = e^{\mu}$ and find the induced prior on θ using Jeffreys Prior. Both recipes give the same answer.

Example: $y_i|\theta \sim Bin(n_i, \theta)$ $i = 1, ..., m$

Need prior on θ... $P(\vec{y}|\theta) = \prod_{i=1}^{m} \binom{n_i}{y_i} \theta^{y_i}(1-\theta)^{n_i - y_i} \propto \theta^{\sum y_i}(1-\theta)^{\sum(n_i - y_i)}$

(Aside: Recall that $\frac{P(\theta)P(y|\theta)}{P(y)} \propto P(\theta)P(y|\theta)$)

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

(Aside: Get the fisher with a calculator or taking the long way with derivatives, etc...)

It turns out that $I(\theta) = \frac{\sum n_i}{\theta(1-\theta)}$ ⇒ Jeffreys Prior = $P(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2}$

⇒ 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: if $x \sim Beta(a,b)$ then $P(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} I_{\{0 \le x \le 1\}}$)

We have $P(\theta|y) \propto \theta^{(\sum y_i + 1/2)-1}(1-\theta)^{(\sum(n_i - y_i)+1/2)-1} \Rightarrow \theta \sim Beta(1/2 + \sum y_i, 1/2 + \sum(n_i - y_i))$

Note: Jeffreys Prior is actually a $Beta(1/2, 1/2)$ in this case. It turns out that if we use a $Beta(a, b)$ prior, we get a posterior of the form $\theta|y \sim Beta(a + \sum y_i, b + \sum(n_i - y_i))$

We call a prior a conjugate prior if the posterior distribution remains in the same family as the prior. (e.g. the prior for θ was $Beta$, posterior was also $Beta$)

Note: Jeffreys priors are not always conjugate priors

Suppose we have $\theta|y \sim Beta(1/2+\sum y_i, 1/2+\sum(n_i-y_i))$ and we have data $\sum y_i = 10 \quad \sum(n_i - y_i) = 20$ which are the total successes and total failures, respectively.
$\Rightarrow$ Posterior is $\theta|y \sim Beta(10.5, 20.5)$ which is shown in Figure 1.

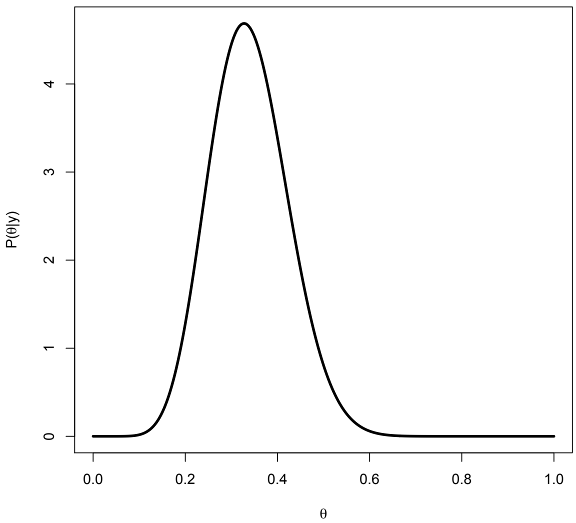


Figure 1: $Beta(10.5, 20.5)$

To get a point estimate for θ we can use either the

(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 the 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

$$S = \left\{ \theta : P(\theta) > P(\theta') \ \forall \theta \in S, \theta' \notin S, \int_S P(\theta|y)d\theta = 1 - \alpha \right\}$$

Code Example

View `binom_coverage_sim.R` in the GitHub repo for some illustrations from this lecture. Also, the function `binom.test()` can be helpful in comparison. The results should be similar. In the end, under certain conditions, Jeffreys Prior returns essentially the same results as Maximum Likelihood will.