More on Bayesian Methods

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

November 2015

- When are Bayesian and frequentist methods the same?
- Example: Normal-Normal
- Posterior predictive inference
- Credible intervals

Notation

$$p(x|\theta) \qquad \text{likelihood}$$

$$\pi(\theta) \qquad \text{prior}$$

$$p(x) = \int p(x|\theta)\pi(\theta)\, d\theta \qquad \text{marginal likelihood}$$

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \qquad \text{posterior probability}$$

$$p(x_{new}|x) = \int p(x_{new}|\theta)\pi(\theta|x)\, d\theta \qquad \text{predictive probability}$$

Normal-Normal

$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathsf{N}(\theta, \sigma^2)$$
$$\theta \sim \mathsf{N}(\mu, \tau^2),$$

where σ^2 is known. Calculate the distribution of $\theta | x_1, \ldots, x_n$. Using a ton of math and algebra, you can show that

$$\begin{aligned} \theta | x_1, \ldots, x_n &\sim N\left(\frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right) \\ &= N\left(\frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2} \right). \end{aligned}$$

Two Useful Things to Know

Definition

The reciprocal of the variance is referred to as the *precision*. Then

$$\text{Precision} = \frac{1}{\text{Variance}}.$$

Suppose the loss we assume is squared error. Let $\delta(x)$ be an estimator of true parameter θ. Then

$$MSE(\delta(x)) = Bias^2 + Variance \tag{1}$$

$$= \left\{\theta - E_\theta[\delta(x)]\right\}^2 + E_\theta[\left\{\delta(x) - E_\theta[\delta(x)]\right\}^2 \tag{2}$$

Theorem

Let δ_n be a sequence of estimators of $g(\theta)$ with mean squared error $E(\delta_n - g(\theta))^2$. Let $b_n(\theta)$ be the bias.

(i) *If $E[\delta_n - g(\theta)]^2 \to 0$ then δ_n is consistent for $g(\theta)$.*

(ii) *Equivalent to the above, δ_n is consistent if $b_n(\theta) \to 0$ and $Var(\delta_n) \to 0$ for all θ.*

(iii) *In particular (and most useful), δ_n is consistent if it is unbiased for each n and if $Var(\delta_n) \to 0$ for all θ.*

We omit the proof since it requires Chebychev's Inequality along with a bit of probability theory. See Problem 1.8.1 in TPE for the exercise of proving this.

Normal-Normal Revisited

We write the posterior mean and posterior variance out.

$$E(\theta|x) = \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

$$= \frac{\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \frac{\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

$$V(\theta|x) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

Can someone given an explanation of what's happening here? How does this contrast frequentist inference?

What happens as $n \to \infty$?

Divide the posterior mean (numerator and denominator) by n. Now take $n \to \infty$. Then

$$E(\theta|x) = \frac{\frac{1}{n}\frac{n\bar{x}}{\sigma^2} + \frac{1}{n}\frac{\mu}{\tau^2}}{\frac{1}{n}\frac{n}{\sigma^2} + \frac{1}{n}\frac{1}{\tau^2}} \to \frac{\frac{\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2}} = \bar{x} \quad \text{as } n \to \infty.$$

In the case of the posterior variance, divide the denominator and numerator by n. Then

$$V(\theta|x) = \frac{\frac{1}{n}}{\frac{1}{n}\frac{n}{\sigma^2} + \frac{1}{n}\frac{1}{\tau^2}} \approx \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty.$$

Since the posterior mean is unbiased and the posterior variance goes to 0, the posterior mean is consistent by our Theorem.

Posterior Predictive Distributions

- We have just seen how estimation can be done in Bayesian analysis.
- Another goal might be prediction.
- That is given some data y and a new observation $\tilde{y}$, we may wish to find the conditional distribution of $\tilde{y}$ given y.
- This distribution is referred to as the *posterior predictive distribution*.
- That is, our goal is to find $p(\tilde{y}|y)$.

Posterior Predictive Distributions

Consider

$$\begin{aligned} p(\tilde{y}|y) &= \frac{p(\tilde{y}, y)}{p(y)} \\ &= \frac{\int_\theta p(\tilde{y}, y, \theta)\, d\theta}{p(y)} \\ &= \frac{\int_\theta p(\tilde{y}|y, \theta) p(y, \theta)\, d\theta}{p(y)} \\ &= \int_\theta p(\tilde{y}|y, \theta) p(\theta|y)\, d\theta. \end{aligned}$$

In most contexts, if θ is given, then $\tilde{y}|\theta$ is independent of y, i.e., the value of θ determines the distribution of $\tilde{y}$, without needing to also know y. When this is the case, we say that $\tilde{y}$ and y are *conditionally independent* given θ. Then the above becomes

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

Theorem

If θ is discrete and $\tilde{y}$ and y are conditionally independent given θ, then the posterior predictive distribution is

$$p(\tilde{y}|y) = \sum_{\theta} p(\tilde{y}|\theta)p(\theta|y).$$

If θ is continuous and $\tilde{y}$ and y are conditionally independent given θ, then the posterior predictive distribution is

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

Negative Binomial Distribution

- We reintroduce the Negative Binomial distribution.
- The binomial distribution counts the numbers of successes in a fixed number of iid Bernoulli trials.
- Recall, a Bernoulli trial has a fixed success probability p.
- Suppose instead that we count the number of Bernoulli trials required to get a fixed number of successes. This formulation leads to the *Negative Binomial distribution*.
- In a sequence of independent Bernoulli(p) trials, let X denote the trial at which the rth success occurs, where r is a fixed integer.
  Then

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \; x = r, r+1, \ldots$$

  and we say $X \sim$ Negative Binom(r, p).

Negative Binomial Distribution

- There is another useful formulation of the Negative Binomial distribution.
- In many cases, it is defined as $Y =$ number of failures before the rth success. This formulation is statistically equivalent to the one given above in term of $X =$ trial at which the rth success occurs, since $Y = X - r$. Then

$$f(y) = \binom{r+y-1}{y} p^r (1-p)^y, \; y = 0, 1, 2, \ldots$$

  and we say $Y \sim$ Negative Binom(r, p).
- When we refer to the Negative Binomial distribution in this class, we will refer to the second one defined unless we indicate otherwise.

$$X|\lambda \sim \text{Poisson}(\lambda)$$
$$\lambda \sim \text{Gamma}(a, b)$$

Assume that $\tilde{X}|\lambda \sim \text{Poisson}(\lambda)$ is independent of X. Assume we have a new observation $\tilde{x}$. Find the posterior predictive distribution, $p(\tilde{x}|x)$. Assume that a is an integer. First, we must find $p(\lambda|x)$.

Recall

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)(p(\lambda) \\ &\propto e^{-\lambda}\lambda^{x}\lambda^{a-1}e^{-\lambda/b} \\ &= \lambda^{x+a-1}e^{-\lambda(1+1/b)}. \end{aligned}$$

Thus, $\lambda|x \sim \text{Gamma}(x+a, \frac{1}{1+1/b})$, i.e.,
$\lambda|x \sim \text{Gamma}(x+a, \frac{b}{b+1})$. Finish the problem for homework.

- Suppose that X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month.
- The discrete count nature of the data plus its natural interpretation as an arrival rate suggest modeling it with a Poisson likelihood.
- To use a Bayesian analysis, we require a prior distribution for θ having support on the positive real line. A convenient choice is given by the Gamma distribution, since it's conjugate for the Poisson likelihood.

The model is given by

$$X|\lambda \sim \text{Poisson}(\lambda)$$
$$\lambda \sim \text{Gamma}(a, b).$$

- We are also told 42 moms are observed arriving at the particular hospital during December 2007. Using prior study information given, we are told $a = 5$ and $b = 6$.
- (We found a, b by working backwards from a prior mean of 30 and prior variance of 180).

We would like to find several things in this example:

1. Plot the likelihood, prior, and posterior distributions as functions of λ in $\mathtt{R}$.
2. Plot the posterior predictive distribution where the number of pregnant women arriving falls between [0,100], integer valued.
3. Find the posterior predictive probability that the number of pregnant women arrive is between 40 and 45 (inclusive). Do this for homework.

Confidence intervals vs credible intervals

A confidence interval for an unknown (fixed) parameter θ is an interval of numbers that we believe is likely to contain the true value of θ. Intervals are important because they provide us with an idea of how well we can estimate θ.

Confidence intervals vs credible intervals

- A *confidence interval* is constructed to contain θ a percentage of the time, say 95%.
- Suppose our confidence level is 95% and our interval is (L, U). Then we are 95% confident that the true value of θ is contained in (L, U) in *the long run*.
- In the long run means that this would occur nearly 95% of the time if we repeated our study millions and millions of times.

Common Misconceptions in Statistical Inference

- A confidence interval is a statement about θ (a population parameter). It is not a statement about the sample.
- Remember that a confidence interval is *not* a statement about individual subjects in the population.
- As an example, suppose that I tell you that a 95% confidence interval for the average amount of television watched by Americans is (2.69, 6.04) hours.
- This *doesn't* mean we can say that 95% of all Americans watch between 2.69 and 6.04 hours of television. We also *cannot* say that 95% of Americans in the sample watch between 2.69 and 6.04 hours of television.
- Beware that statements such as these are false.
- However, we can say that we are 95 percent confident that the *average* amount of televison watched by Americans is between 2.69 and 6.04 hours.

Credible intervals

Let θ be a random variable (parameter). A confidence (credible region) on θ is to determine $C(X_n)$ such that

$$\pi(\theta \in C(X_n) \mid X_n) = 1 - \alpha,$$

where α is predetermined such as 0.05.

Simple definition of credible interval

A Bayesian credible interval of size $1-\alpha$ is an interval (a, b) such that

$$P(a \leq \theta \leq b|x) = 1 - \alpha.$$

$$\int_a^b p(\theta|x)\, d\theta = 1 - \alpha.$$

Remark: When you're calculating credible intervals, you'll find the values of a and b by several means. You could be asked do the following:

- *Find the a, b using means of calculus to determine the credible interval or set.*
- *Use a Z-table when appropriate.*
- *Use `R` to approximate the values of a and b.*

Important Point

Our definition for the credible interval could lead to many choices of (a, b) for particular problems.
Suppose that we required our credible interval to have equal probability $\alpha/2$ in each tail. That is, we will assume

$$P(\theta < a|x) = \alpha/2$$

and

$$P(\theta > b|x) = \alpha/2.$$

Important Point

Is the credible interval still unique? No. Consider

$$\pi(\theta|x) = I(0 < \theta < 0.025) + I(1 < \theta < 1.95) + I(3 < \theta < 3.025)$$

so that the density has three separate plateaus. Now notice that any (a, b) such that $0.025 < a < 1$ and $1.95 < b < 3$ satisfies the proposed definition of a ostensibly "unique" credible interval. To fix this, we can simply require that

$$\{\theta : \pi(\theta|x) \text{ is positive}\}$$

(i.e., the support of the posterior) must be an interval.

Comparisons

- Conceptually, probability comes into play in a frequentist confidence interval *before* collecting the data, i.e., there is a 95% probability that we will collect data that produces an interval that contains the true parameter value. However, this is awkward, because we would like to make statements about the probability that the interval contains the true parameter value given the data that we actually observed.
- Meanwhile, probability comes into play in a Bayesian credible interval *after* collecting the data, i.e., based on the data, we now think there is a 95% probability that the true parameter value is in the interval. This is more natural because we want to make a probability statement regarding that data after we have observed it.

Sleep Example

- Consider that we were interested in the proportion of the population of American college students that sleep at least eight hours each night (θ).
- Suppose a random sample of 27 students from Duke, where 11 students recorded they slept at least eight hours each night.
- So, we assume the data is distributed as Binomial$(27, \theta)$.

Suppose that the prior on θ was Beta(3.3,7.2). Thus, the posterior distribution is

$$\theta|11 \sim \text{Beta}(11 + 3.3, 27 - 11 + 7.2), \text{ i.e.,}$$
$$\theta|11 \sim \text{Beta}(14.3, 23.2).$$

Sleep Example

- Suppose now we would like to find a 90% credible interval for θ.
- We cannot compute this in closed form since computing probabilities for Beta distributions involves messy integrals that we do not know how to compute.
- However, we can use `R` to find the interval.

We need to solve

$$P(\theta < c|x) = 0.05$$

and

$$P(\theta > d|x) = 0.05 \text{ for } c \text{ and } d.$$

Sleep Example

We cannot compute this in closed form because we need to compute

$$\int_0^c \text{Beta}(14.3, 23.2)\ d\theta = 0.05$$

and

$$\int_d^1 \text{Beta}(14.3, 23.2)\ d\theta = 0.05.$$

Note that Beta(14.3,23.2) represents

$$f(\theta) = \frac{\Gamma(37.5)}{\Gamma(14.3)\Gamma(23.2)}\theta^{14.3-1}(1-\theta)^{23.2-1}.$$

Sleep Example

The R code for this is very straightforward:

```
a = 3.3
b = 7.2
n = 27
x = 11
a.star = x+a
b.star = n-x+b

c = qbeta(0.05,a.star,b.star)
d = qbeta(1-0.05,a.star,b.star)
```

Running the code in R, we find that a 90% credible interval for θ is (0.256, 0.514), meaning that there is a 90% probability that the proportion of Duke students who sleep eight or more hours per night is between 0.256 and 0.514 given the data.