More on Bayesian Methods: Part II

Rebecca C. Steorts
Bayesian Methods and Modern Statistics: STA 360/601

Lecture 5

Today's menu

- Confidence intervals
- Credible Intervals
- Example

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.
- It is also *not* a statement about individual subjects in the population.

Common Misconceptions in Statistical Inference

- Let a 95% confidence interval for the average amount of television watched by Americans be (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.

Confidence interval

- Conceptually, probability comes into play in a frequentist confidence interval *before* collecting the data.
- Ex: there is a 95% probability that we will collect data that produces an interval that contains the true parameter value.
- Awkard!
- We would like to make statements about the probability that the interval contains the true parameter value given the data that we actually observed.

Credible interval

- Meanwhile, probability comes into play in a Bayesian credible interval after collecting the data
- Ex: 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

- 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 UF, where 11 students recorded they slept at least eight hours each night.
- $X \sim \text{Binomial}(27, \theta)$.

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.