Stats 102C - Lecture 1-3: Bayesian Inference

Miles Chen, PhD

Week 1 Friday

Clarification

The likelihood function of a binomial probability is NOT the beta distribution.

The likelihood function of a binomial probability for N observations with z successes is:

$$\binom{N}{z}\theta^z(1-\theta)^{N-z}$$

This can be found in R with `dbinom(z, N, theta)` where `z` and `N` are fixed values from the data and `theta` is a variable.

The likelihood function of a binomial probability for N observations with z successes is **proportional to, but not equal to** a beta distribution with $\alpha = z + 1$ and $\beta = N - z + 1$.

The likelihood function does not integrate to 1. The density function of a beta distribution will integrate to 1.

In Wednesday's lecture, I plotted a beta distribution instead of the likelihood function to amplify the shape of the likelihood function. If I plotted the true likelihood function on the same scale as the beta distribution, it would be "too flat" to be seen.

Section 1

Lecture 1-3: Basic Bayesian Inference

The Beta-Binomial Model

The beta distribution is the conjugate prior for data with a binomial likelihood.

Let the observed data come from a binomial distribution. There are a total of N trials and z successes. The probability of success for each trial is θ.

$$L = \Pr(z|\theta, N) \propto \theta^z(1-\theta)^{N-z}$$

The prior distribution of the parameter θ is a beta distribution with parameters α and β.

$$\Pr(\theta) = Beta(\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The resulting posterior distribution will also be a Beta distribution with parameters $z + \alpha$ and $N - z + \beta$.

$$\Pr(\theta|z) = Beta(z + \alpha, N - z + \beta) = \frac{1}{B(z + \alpha, N - z + \beta)}\theta^{z+\alpha-1}(1-\theta)^{N-z+\beta-1}$$

Mean of the Beta distribution

The mean of the beta distribution with shape parameters α and β is $\frac{\alpha}{\alpha+\beta}$. Proof:

$$
\begin{aligned}
E[\theta] &= \int_0^1 \theta \times \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 \theta \times \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+1-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}B(\alpha+1,\beta) \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

Note: $\Gamma(x+1) = x!$, and $\Gamma(x+1) = x\Gamma(x)$

Mode of the Beta distribution

The mode (maximum) of the beta distribution with shape parameters α and β is $\frac{\alpha-1}{\alpha+\beta-2}$. Proof:

The maximum occurs where the derivative is 0.

$$
\begin{aligned}
0 &= \frac{d}{d\theta}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
0 &= \frac{1}{B(\alpha,\beta)}\left((\alpha-1)\theta^{\alpha-2}(1-\theta)^{\beta-1} + \theta^{\alpha-1}(\beta-1)(1-\theta)^{\beta-2}(-1)\right) \\
0 &= \theta^{\alpha-2}(1-\theta)^{\beta-2}((\alpha-1)(1-\theta) - \theta(\beta-1)) \text{ the derivative is also 0 when } \theta = 0 \text{ and } \theta = 1 \\
0 &= (\alpha-1)(1-\theta) - \theta(\beta-1) \\
0 &= \alpha - 1 - \alpha\theta + \theta - \theta\beta + \theta \\
0 &= (\alpha-1) - \theta(\alpha+\beta-2) \\
\theta(\alpha+\beta-2) &= (\alpha-1) \\
\theta &= \frac{\alpha-1}{\alpha+\beta-2}
\end{aligned}
$$

Probability intervals

With Bayesian statistics, we can make probability intervals. These are the Bayesian alternative to a confidence interval.

In frequentist statistics, the unknown parameter is fixed and so we cannot talk about the probability that the parameter falls within some chosen bounds.

In Bayesian statistics, the unknown parameter is a random variable, so we can say that there is a 95% probability that the parameter falls within the chosen bounds.

There are a couple popular methods for determining the bounds of the interval:

- Equal tail credibility interval
- Highest posterior density interval

Equal tail credibility interval

An equal tail credibility interval ensures that the area above the credibility interval is equal to the area below the credibility interval.

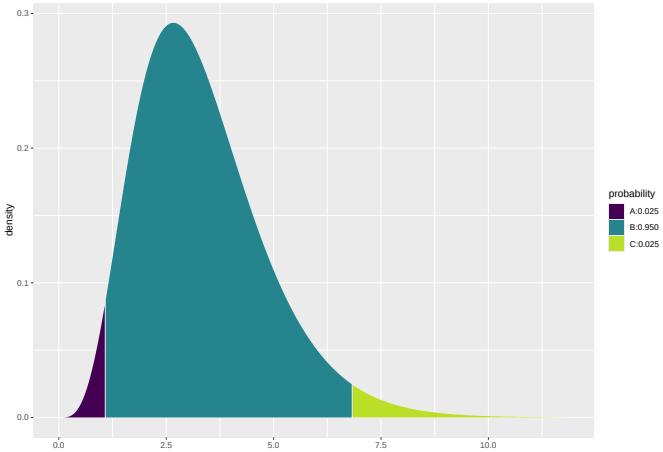
If the posterior distribution is known, the bounds can be found easily using the quantile function.

```
# posterior distribution is known to be gamma dist with the following parameters
shape = 5
rate = 1.5
bounds <- qgamma(c(.025, .975), shape, rate) # bounds of 95% credibility interval
bounds
```

```
## [1] 1.082324 6.827726
```

Equal tail credibility interval

```
library(mosaic)
xpgamma(bounds, shape, rate) # useful function in mosaic that plots and shades densities
```




```
## [1] 0.025 0.975
```

Highest Density interval

The highest (posterior) density interval finds the narrowest range of values that cover 95% of the posterior distribution.

The value of the density function at the lower bound will be equal to the value of the density function at the upper bound.

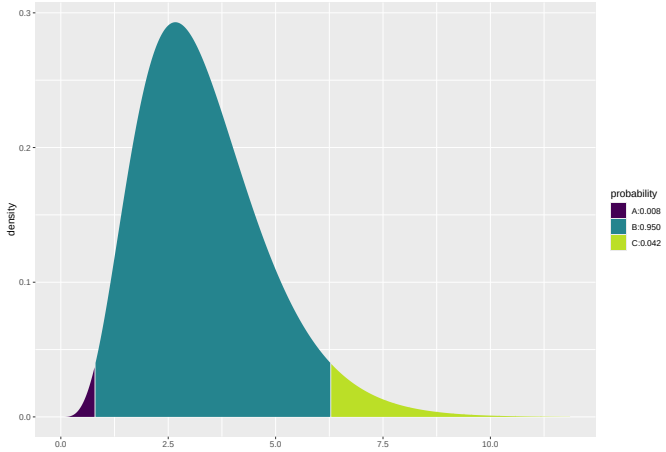
The highest (posterior) density interval can be found using the HDInterval package.

```
library(HDInterval)
# posterior distribution is known to be gamma dist with the following parameters
shape = 5
rate = 1.5
bounds <- hdi(qgamma, credMass = 0.95, shape = shape, rate = rate) # 95%
bounds
```

```
##     lower     upper
## 0.8046401 6.2868113
## attr(,"credMass")
## [1] 0.95
```

Highest Density interval

```
xpgamma(bounds, shape, rate)
```




```
##       lower    upper
```

Credibility Interval interpretations

Whether you select the equal-tailed interval or the highest posterior density interval, it is correct to say that there is a 95% probability that the unknown parameter falls within the bounds of the interval.

Keep in mind this probability reflects our subjective belief about the random variable.

The credibility interval incorporates our prior knowledge in the form of the prior distribution used in the calculation of the posterior distribution. In cases where there is little observed data, the credibility interval will be greatly influenced by the prior beliefs. In cases where lots of observed data is available, the interval will be influenced more by the data.

A frequentist confidence interval on the other hand is determined only by the observed data. Intervals based on little data have large margins of error and intervals based on more data will have smaller margins of error.

Conjugate priors

Wikipedia: In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x|\theta)$.

The beta and binomial distributions are a pair of conjugate distributions.

Wikipedia's table of conjugate distribution provides a list of important pairs

https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

Section 2

Making predictions with the posterior distribution

Making predictions with the posterior distribution

Let's revisit the baseball example.

Our prior distribution for θ is a beta distribution with $\alpha = 81$ and $\beta = 219$.

We observed a new player with 10 at bats and 5 hits.

Our posterior distribution for θ is now a beta distribution with $\alpha = 86$ and $\beta = 224$.

Let's say we are faced with the question:

"If this player has three at bats in the next game, what is the probability he gets exactly two hits?"

Strategies

"If this player has three at bats in the next game, what is the probability he gets exactly two hits?"

The answer depends on the value of θ: $\binom{3}{2}\theta^2(1-\theta)^1$

Unfortunately, we do not know the value of θ because it is a random variable.

There are a few strategies we can use for dealing with unknown θ

- Use a single value: the mean of the posterior distribution
- Use a single value: the mode of the posterior distribution
- Find the expected value of the probability analytically
- Estimate the expected value of the probability via Monte Carlo

Use a single value: the mean of the posterior distribution

The posterior distribution for θ is a beta distribution with $\alpha = 86$ and $\beta = 224$.

The mean of the posterior is $\frac{\alpha}{\alpha+\beta} = \frac{86}{86+224} \approx 0.2774194$

"If this player has three at bats in the next game, what is the probability he gets exactly two hits?"

Answer:

$$\binom{3}{2} 0.2774194^2 (1 - 0.2774194)^1 \approx 0.1668327$$

Use a single value: the mode of the posterior distribution

The posterior distribution for θ is a beta distribution with $\alpha = 86$ and $\beta = 224$.

The mode of the posterior is $\frac{\alpha-1}{\alpha+\beta-2} = \frac{86-1}{86+224-2} \approx 0.275974$

"If this player has three at bats in the next game, what is the probability he gets exactly two hits?"

Answer:

$$\binom{3}{2} 0.275974^2 (1 - 0.275974)^1 = 0.165429$$

Find the expected value of the probability analytically

Use "Law of the unconscious statistician": used to calculate the expected value of a function $g(X)$ of a random variable X when one knows the probability distribution of X but one does not know the distribution of $g(X)$

$$E[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx$$

In our case, $f(x)$ is the PDF of a beta distribution with $\alpha = 86$ and $\beta = 224$.

The function $g(x)$ is the probability of getting 2 hits in 3 at bats: $\binom{3}{2}\theta^2(1-\theta)^1$

Thus, the expected probability of getting 2 hits in 3 at bats for a player whose batting average is a random variable modeled by a beta distribution with $\alpha = 86$ and $\beta = 224$.

$$E[g(\Theta)] = \int_0^1 \binom{3}{2}\theta^2(1-\theta)^1 \cdot \frac{1}{B(86, 224)}\theta^{86-1}(1-\theta)^{224-1} d\theta$$

Find the expected value of the probability analytically

$$
\begin{aligned}
E[g(\Theta)] &= \int_0^1 \binom{3}{2}\theta^2(1-\theta)^1 \cdot \frac{1}{B(86,224)}\theta^{86-1}(1-\theta)^{224-1}d\theta \\
&= 3 \cdot \frac{1}{B(86,224)} \int_0^1 \theta^2(1-\theta)^1\theta^{86-1}(1-\theta)^{224-1}d\theta \\
&= 3 \cdot \frac{1}{B(86,224)} \int_0^1 \theta^{88-1}(1-\theta)^{225-1}d\theta \\
&= 3 \cdot \frac{1}{B(86,224)} B(88,225) \\
&= 3 \cdot \frac{\Gamma(86+224)}{\Gamma(86)\Gamma(224)} \frac{\Gamma(88)\Gamma(225)}{\Gamma(88+225)} \\
&= 3 \cdot \frac{\Gamma(310)}{\Gamma(313)} \frac{\Gamma(88)}{\Gamma(86)} \frac{\Gamma(225)}{\Gamma(224)} \\
&= 3 \cdot \frac{309!}{312!} \frac{87!}{85!} \frac{224!}{223!} \\
&= 3 \cdot \frac{1}{310 \cdot 311 \cdot 312} \frac{87 \cdot 86}{1} \frac{224}{1} = 0.1671515
\end{aligned}
$$

Estimate the expected value of the probability via Monte Carlo

Rather than work through the math analytically, we can use Monte Carlo methods to estimate the expected value of $g(x)$

$$E[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx \approx \frac{1}{N} \sum_{j=1}^{N} g(x_j)$$

Where x_j are values randomly drawn from a distribution with PDF $= f(x)$.

$f(x)$ is the PDF of a beta distribution with $\alpha = 86$ and $\beta = 224$. We are able to use R's `rbeta()` function to draw random values from this distribution.

The function $g(x)$ is the probability of getting 2 hits in 3 at bats: $\binom{3}{2}\theta^2(1-\theta)^1$

```
set.seed(1)
samp <- rbeta(10^6, 86, 224)
g <- function(theta){ 3 * theta ^ 2 * (1 - theta) }
mean(g(samp))
```

```
## [1] 0.1671577
```

Comments

When we compare the Monte Carlo estimate of the expected value with the true expected value based on analytical math, we see that the monte carlo estimate is a very good approximation.

For many problems, the analytic math will be intractable or simply too hard to solve.

For these types of problems, we will rely on the using Monte Carlo methods to estimate the desired quantity.

Monte Carlo methods are frequently used in the context of Bayesian statistics because the parameter is a random variable and solving problems often involve some complex integrals.