Session 1: The Basics of Bayesian Inference

Bayesian Inference for the Social Sciences
November 3, 2017

Kevin Quinn

Overview of Workshop

Goal of this workshop is to introduce you to key Bayesian ideas and hopefully encourage you to learn more on your own.

Overview of Workshop

Goal of this workshop is to introduce you to key Bayesian ideas and hopefully encourage you to learn more on your own.

Four sessions:

- Session 1: Introduction to Bayesian Inference
- Session 2: Introduction to Simulation-Based Inference
- Session 3: Gibbs Sampling
- Session 4: Auxiliary Variables / Data Augmentation

Overview of Workshop

Goal of this workshop is to introduce you to key Bayesian ideas and hopefully encourage you to learn more on your own.

Four sessions:

- Session 1: Introduction to Bayesian Inference
- Session 2: Introduction to Simulation-Based Inference
- Session 3: Gibbs Sampling
- Session 4: Auxiliary Variables / Data Augmentation

Especially in the later sessions there will be a number of hands-on examples.

Please note that we won't have time to go into as much depth with these examples as would be ideal.

Also please note that I've tried to make the code as easy to read and understand as possible at the cost of some computational efficiency and programming elegance.

Bayesian Inference

Bayesian inference is a means of making rational probability statements about quantities of interest (observables, model parameters, functions of model parameters).

It answers the questions that researchers are really interested in, e.g. “What is the probability that...”

Natural way to combine information from multiple studies.

Provides a formal method for combining prior qualitative information with observed quantitative information.

Natural way to deal with missing data.

Bayesian Inference

Probability is thought of as subjective degrees of belief.

Bayesian Inference

Probability is thought of as subjective degrees of belief.

Implies that one should either:

- elicit and defend real subjective beliefs
- perform sensitivity analyses to demonstrate that one's results don't depend heavily on idiosyncratic subjective beliefs

The Process of Bayesian Inference

1. setting up a probability model
2. calculating and interpreting the posterior distribution
3. evaluating model adequacy

Probability Models

Consider an observed sample of data y.

Probability Models

Consider an observed sample of data y.

A probability model for y consists of two things:

Probability Models

Consider an observed sample of data y.

A probability model for y consists of two things:

- An assumption about the probability distribution with density $p(y|\theta)$ that generated y

Probability Models

Consider an observed sample of data y.

A probability model for y consists of two things:

- An assumption about the probability distribution with density $p(y|\theta)$ that generated y
- The set Θ of possible values of the model parameters θ

Probability Models

$p(y|\theta)$ is sometimes called the sampling density. It is the joint density of all the observed data points.

Probability Models

$p(y|\theta)$ is sometimes called the sampling density. It is the joint density of all the observed data points.

When $p(y|\theta)$ is viewed as a function of θ for fixed y it is referred to as the likelihood function and is written $L(\theta|y)$.

Probability Models

$p(y|\theta)$ is sometimes called the sampling density. It is the joint density of all the observed data points.

When $p(y|\theta)$ is viewed as a function of θ for fixed y it is referred to as the likelihood function and is written $L(\theta|y)$.

The likelihood function can actually be any function of y and θ that is proportional to $p(y|\theta)$.

Probability Models: Example

Consider a random sample of size *n* from the population of registered US voters.

Probability Models: Example

Consider a random sample of size n from the population of registered US voters.

We observe that y of the n citizens voted in the 2008 presidential election. The remainder abstained.

Probability Models: Example

Consider a random sample of size n from the population of registered US voters.

We observe that y of the n citizens voted in the 2008 presidential election. The remainder abstained.

The setup for the problem suggests that a reasonable distributional assumption for y is that it follows a Binomial distribution with sample size n and probability of success θ.

Probability Models: Example

Consider a random sample of size n from the population of registered US voters.

We observe that y of the n citizens voted in the 2008 presidential election. The remainder abstained.

The setup for the problem suggests that a reasonable distributional assumption for y is that it follows a Binomial distribution with sample size n and probability of success θ.

n is fixed and it is reasonable to assume that $\theta \in (0, 1)$. This determines the parameter space.

Probability Models: Example

Consider a random sample of size n from the population of registered US voters.

We observe that y of the n citizens voted in the 2008 presidential election. The remainder abstained.

The setup for the problem suggests that a reasonable distributional assumption for y is that it follows a Binomial distribution with sample size n and probability of success θ.

n is fixed and it is reasonable to assume that $\theta \in (0, 1)$. This determines the parameter space.

Thus our probability model is

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}, \qquad \theta \in (0, 1)$$

Calculating and Summarizing the Posterior

The goal of Bayesian inference is to make probability statements about model parameters θ and/or functions of model parameters $g(\theta)$ given a probability model and observed data.

In other words, we want to know $p(\theta|y)$ (the posterior density).

Calculating and Summarizing the Posterior

The goal of Bayesian inference is to make probability statements about model parameters θ and/or functions of model parameters $g(\theta)$ given a probability model and observed data.

In other words, we want to know $p(\theta|y)$ (the posterior density).

Note that our probability model is defined in terms of $p(y|\theta)$ which is not quite what we want.

Calculating and Summarizing the Posterior

The goal of Bayesian inference is to make probability statements about model parameters θ and/or functions of model parameters $g(\theta)$ given a probability model and observed data.

In other words, we want to know $p(\theta|y)$ (the posterior density).

Note that our probability model is defined in terms of $p(y|\theta)$ which is not quite what we want.

How do we get from $p(y|\theta)$ to $p(\theta|y)$?

Calculating the Posterior

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

This identity is known as Bayes' rule

Calculating the Posterior

Bayes' rule gives us a formula for calculating the posterior density of θ given y (denoted $p(\theta|y)$ from knowledge of the sampling density (denoted $p(y|\theta)$) and the prior density of θ (denoted $p(\theta)$).

The function $p(\theta)$ plays a crucial role here.

Since $p(\theta)$ doesn't depend on the observed data it represents the researcher's subjective a priori beliefs about the likely values of θ.

The fact that $p(\theta)$ is a subjective probability implies that $p(\theta|y)$ is a subjective probability.

Calculating the Posterior

Since $p(y)$ is a constant for fixed y we can write:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In words, the posterior density is proportional to the sampling density times the prior density

Summarizing the Posterior

Once a probability model is formed and a prior is specified Bayesian inference proceeds by summarizing $p(\theta|y)$

Interesting quantities include (but are not limited to):

- The posterior mean of θ

$$\mathbb{E}[\theta|y] = \int_{\Theta} \theta p(\theta|y) d\theta$$

- The posterior variance of θ

$$\mathbb{V}[\theta|y] = \int_{\Theta} (\theta - \mathbb{E}[\theta|y])^2 p(\theta|y) d\theta$$

- a $100 \times (1 - \alpha)\%$ credible set $C \subset \Theta$ where C is chosen to satisfy

$$1 - \alpha = \int_{C} p(\theta|y) d\theta$$

Prior and Posterior Predictive Distributions

In some cases the parameters of our probability model are of direct substantive interest, but in many other cases the model parameters are of only indirect interest—we are really interested in making claims about the likelihood that future values of y will take particular values.

The Bayesian approach provides a straightforward method of making these sorts of inferences.

The Prior Predictive Distribution

Suppose Y is our outcome variable.

Question: What does our prior $p(\theta)$ imply about the distribution of Y?

Answer: Calculate the prior predictive distribution of Y:

$$p(y^{\text{rep}}) = \int_{\Theta} p(y^{\text{rep}}|\theta)p(\theta)d\theta$$

(Note: the y^{rep} notation is designed to make it clear that this is a hypothetical replication of y and not an observed y.)

The Prior Predictive Distribution

We can generate a sample of M y^{rep} values from $p(y^{\text{rep}})$ by the following simple procedure:

- for $i = 1, \ldots, M$ do:
  1. sample $\theta_{(i)}$ from $p(\theta)$
  2. sample $y^{\text{rep}}_{(i)}$ from $p(y^{\text{rep}} | \theta_{(i)})$
  3. store $y^{\text{rep}}_{(i)}$

The Posterior Predictive Distribution

Suppose Y is our outcome variable.

Question: What does our posterior $p(\theta|y)$ imply about the distribution of future replicates of Y?

Answer: Calculate the posterior predictive distribution of Y:

$$p(y^{\text{rep}}|y) = \int_{\Theta} p(y^{\text{rep}}|\theta)p(\theta|y)d\theta$$

(Note: the need for the y^{rep} notation becomes more clear here where we need to distinguish between the observed y and hypothetical replications y^{rep}.)

The Posterior Predictive Distribution

We can generate a sample of M y^{rep} values from $p(y^{\text{rep}}|y)$ by the following simple procedure:

- for $i = 1, \ldots, M$ do:
  1. sample $\theta_{(i)}$ from $p(\theta|y)$
  2. sample $y^{\text{rep}}_{(i)}$ from $p(y^{\text{rep}}|\theta_{(i)})$
  3. store $y^{\text{rep}}_{(i)}$

Assessing Model Adequacy

Inferences depend on assumptions

Assessing Model Adequacy

Inferences depend on assumptions

In a Bayesian analysis the modeling assumptions are encoded in two objects:

- $p(y|\theta)$
- $p(\theta)$

Assessing Model Adequacy

Inferences depend on assumptions

In a Bayesian analysis the modeling assumptions are encoded in two objects:

- $p(y|\theta)$
- $p(\theta)$

(How) do inferences change across reasonable choices of $p(y|\theta)$ and $p(\theta)$?

Assessing Model Adequacy

“Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration. ... Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.”
— George E. P. Box, 1976

“All models are wrong. Some are useful.”
— George E. P. Box

Bayesian Inference for the Binomial θ Parameter

Bayesian Inference for the Binomial θ Parameter

Suppose we have n independent Bernoulli trials:

$$Y_i \overset{ind.}{\sim} \mathcal{B}ernoulli(\theta) \quad i = 1, \ldots, n$$

Bayesian Inference for the Binomial θ Parameter

Suppose we have n independent Bernoulli trials:

$$Y_i \overset{ind.}{\sim} \mathcal{B}ernoulli(\theta) \quad i = 1, \ldots, n$$

What are some substantive examples that fit this data generating process?

Bayesian Inference for the Binomial θ Parameter

Suppose we have n independent Bernoulli trials:

$$Y_i \overset{ind.}{\sim} \mathcal{B}ernoulli(\theta) \quad i = 1, \ldots, n$$

What are some substantive examples that fit this data generating process?

Define $Y = \sum_{i=1}^{n} Y_i$.

It follows that

$$Y \sim \mathcal{B}inom(n, \theta)$$

If we observe y and n, what should we infer about the value of θ?

The Probability Model

From the setup above, we know that the probability model is a binomial model:

$$p(y|\theta) = \binom{n}{y} \theta^{y} (1-\theta)^{(n-y)}$$

The Prior Distribution

Since θ is a probability it must take values between 0 and 1.

We need a prior distribution that has support on $(0, 1)$

The beta distribution is one possibility:

$$\theta \sim \mathcal{B}eta(a, b)$$

with density

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{(a-1)} (1-\theta)^{(b-1)}$$

The Posterior Distribution

Let's use the proportional form of Bayes' rule:

$$
\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&\propto \binom{n}{y}\theta^{y}(1-\theta)^{(n-y)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{(b-1)}
\end{aligned}
$$

The Posterior Distribution

Let's use the proportional form of Bayes' rule:

$$
\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&\propto \binom{n}{y}\theta^{y}(1-\theta)^{(n-y)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{(b-1)}
\end{aligned}
$$

Ignoring terms that are constant for fixed n, y, a, and b we can write

$$
\begin{aligned}
p(\theta|y) &\propto \theta^{y}(1-\theta)^{(n-y)}\theta^{(a-1)}(1-\theta)^{(b-1)} \\
&\propto \theta^{(y+a-1)}(1-\theta)^{(n-y+b-1)}
\end{aligned}
$$

Which is proportional to a $\mathcal{B}eta(y+a,\ n-y+b)$ density.

The Posterior Distribution

Interesting: binomial likelihood $\times$ beta prior = beta posterior

The Posterior Distribution

Interesting: binomial likelihood $\times$ beta prior = beta posterior

This is an example of conjugacy

The Posterior Distribution

Interesting: binomial likelihood $\times$ beta prior = beta posterior

This is an example of conjugacy

Interpreting prior parameters as additional data

Summarizing the Posterior Distribution

Summarizing the posterior in this case is quite straightforward because we know that the posterior distribution is a member of a known family of distributions whose properties are well understood.

Recall the analytic formulas for the mean and variance of a beta random variable.

We can also make use of `pbeta()` and `qbeta()`.

We can also use simulation to summarize the posterior.

Examples to follow.

Example 1: Voter ID Requests in Boston

Cobb, Greiner, and Quinn (2012) conducted an exit poll in Boston in 2008 and calculated the rates at which voters report being asked for ID at the polling location.

The exit poll covered 39 polling locations, but to make life simple we'll look at data from a single polling location. (We'll also ignore some complications caused by survey non-response, etc.)

In this polling location 314 voters were sampled. (The selection probability was 1/16 = 0.0625. 86 reported being asked for ID. What should we infer about the fraction of all (roughly 5000) voters in this polling location who were asked for ID?

Example 1: Voter ID Requests in Boston

`BostonVoters1.R`

Example 2: *Castaneda v. Partida* 430 U.S. 482 (1977)

79.1% of a local population is Mexican-American.

Over 2.5 years 220 people were called from this local population to serve on grand juries.

100 were Mexican-Americans.

Is this evidence of discrimination?

Example 2: *Castaneda v. Partida* 430 U.S. 482 (1977)

`Castaneda.R`

The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take only two distinct values:

The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take only two distinct values:

- A toss of a coin comes up either heads or tails

The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take only two distinct values:

- A toss of a coin comes up either heads or tails
- The roll of a six-sided die is either even or odd

The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take only two distinct values:

- A toss of a coin comes up either heads or tails
- The roll of a six-sided die is either even or odd
- A randomly chosen citizen either turned out to vote in a particular election or did not

The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take only two distinct values:

- A toss of a coin comes up either heads or tails
- The roll of a six-sided die is either even or odd
- A randomly chosen citizen either turned out to vote in a particular election or did not
- A particular Supreme Court justice either joined the majority opinion in a particular case or did not.

The Bernoulli Distribution

Without loss of generality we specify the values of the two possible outcomes to be 0 and 1.

We will call 1s “successes” and 0s “failures”.

This is purely a convention and doesn’t affect any results that follow.

The Bernoulli Distribution

The Bernoulli distribution is governed by a single parameter θ. θ is the probability of a "success".

If X follows a Bernoulli distribution with probability of success θ we write:

$$X \sim \mathcal{B}ernoulli(\theta)$$

The Bernoulli Distribution

The Bernoulli probability mass function is given by

$$f_{\mathcal{B}ern}(x|\theta) = \begin{cases} (1-\theta) & \text{if } x = 0 \\ \theta & \text{if } x = 1 \end{cases}$$

or equivalently as

$$f_{\mathcal{B}ern}(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

The Bernoulli distribution function is given by

$$F_{\mathcal{B}ern}(x|\theta) = \begin{cases} (1-\theta) & \text{if } x = 0 \\ 1 & \text{if } x = 1 \end{cases}$$

The Bernoulli Distribution

If $X \sim \mathcal{B}ernoulli(\theta)$ then

$$\mathbb{E}[X] = \theta$$

and

$$\mathbb{V}[X] = \theta(1 - \theta)$$

The Binomial Distribution

The binomial distribution arises as the distribution of a random variable X that is the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ that all have the same probability of success θ.

The Binomial Distribution

The binomial distribution arises as the distribution of a random variable X that is the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ that all have the same probability of success θ.

If X follows a binomial distribution with probability of success θ and sample size n we write

$$X \sim \mathcal{B}inomial(n, \theta)$$

The Binomial Distribution

The binomial probability mass function is

$$f_{\mathcal{B}in}(x|n,\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{(n-x)}$$

where $\binom{n}{x}$ is read "n choose x" and is equal to $\frac{n!}{x!(n-x)!}$.

$\binom{n}{x}$ is the number of combinations of size x that can be taken from a set of n elements.

The binomial probability mass function is implemented as `dbinom()` in `R`

The Binomial Distribution

The binomial distribution function is

$$F_{\mathcal{B}in}(x|n, \theta) = \sum_{z=0}^{\lfloor x \rfloor} \binom{n}{z} \theta^z (1 - \theta)^{(n-z)}$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x.

The binomial distribution function is implemented as `pbinom()` in `R`

Sampling from a binomial distribution can be accomplished in `R` with `rbinom()`

The Binomial Distribution

If $X \sim \mathcal{B}inomial(n, \theta)$ then

$$\mathbb{E}[X] = n\theta$$

and

$$\mathbb{V}[X] = n\theta(1 - \theta)$$

The Binomial Distribution

How do we get to the binomial distribution from iid Bernoulli trials?

The Binomial Distribution

How do we get to the binomial distribution from iid Bernoulli trials?

Let $X_i \overset{iid}{\sim} \mathcal{B}ernoulli(\theta) \quad i = 1, \ldots, n$

The Binomial Distribution

How do we get to the binomial distribution from iid Bernoulli trials?

Let $X_i \overset{iid}{\sim} \mathcal{B}ernoulli(\theta) \quad i = 1, \ldots, n$

and $X = \sum_{i=1}^{n} X_i$

The Binomial Distribution

Because of the assumption of independence we can calculate the probability of one particular realized sequence $(x_1, \ldots, x_n)$ as:

$$
\begin{aligned}
\Pr(X_1 = x_1, \ldots, X_n = x_n | \theta) &= \theta^{x_1}(1-\theta)^{(1-x_1)} \times \ldots \times \theta^{x_n}(1-\theta)^{(1-x_n)} \\
&= \theta^{(\sum_{i=1}^{n} x_i)}(1-\theta)^{(n-\sum_{i=1}^{n} x_i)} \\
&= \theta^{x}(1-\theta)^{(n-x)}
\end{aligned}
$$

The Binomial Distribution

Now, note that by definition there are $\binom{n}{x}$ ways that x successes could occur in n Bernoulli trials.

The Binomial Distribution

Now, note that by definition there are $\binom{n}{x}$ ways that x successes could occur in n Bernoulli trials.

Since the Bernoulli draws are independent and identically distributed the probability of each of these $\binom{n}{x}$ sequences has equal probability.

The Binomial Distribution

Now, note that by definition there are $\binom{n}{x}$ ways that x successes could occur in n Bernoulli trials.

Since the Bernoulli draws are independent and identically distributed the probability of each of these $\binom{n}{x}$ sequences has equal probability.

To get the probability of x successes in n Bernoulli trials we need to calculate the probability of any single sequence of x successes and then multiply by the total number of sequences that could give rise to x successes in n trials.

The Binomial Distribution

Now, note that by definition there are $\binom{n}{x}$ ways that x successes could occur in n Bernoulli trials.

Since the Bernoulli draws are independent and identically distributed the probability of each of these $\binom{n}{x}$ sequences has equal probability.

To get the probability of x successes in n Bernoulli trials we need to calculate the probability of any single sequence of x successes and then multiply by the total number of sequences that could give rise to x successes in n trials.

The first part is $\theta^x(1-\theta)^{(n-x)}$ and the second part is $\binom{n}{x}$ so the probability of x successes in n trials is

$$\binom{n}{x}\theta^x(1-\theta)^{(n-x)}$$

which is the rhs of the binomial probability mass function.

The Beta Distribution

The beta distribution is a common choice for modeling a continuous random variable that takes values in the $(0, 1)$ interval.

The beta distribution is governed by two positive shape parameters a and b.

If the random variable X follows a beta distribution with parameters a and b we write

$$X \sim \mathcal{B}eta(a, b)$$

The Beta Distribution

The probability density function of the beta distribution is

$$f_{\mathcal{B}eta}(x|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{(a-1)}(1 - x)^{(b-1)}$$

where Γ is the gamma function

In `R` :

- density function: `dbeta()`
- distribution function: `pbeta()`
- pseudo-random variate generation: `rbeta()`

The Beta Distribution

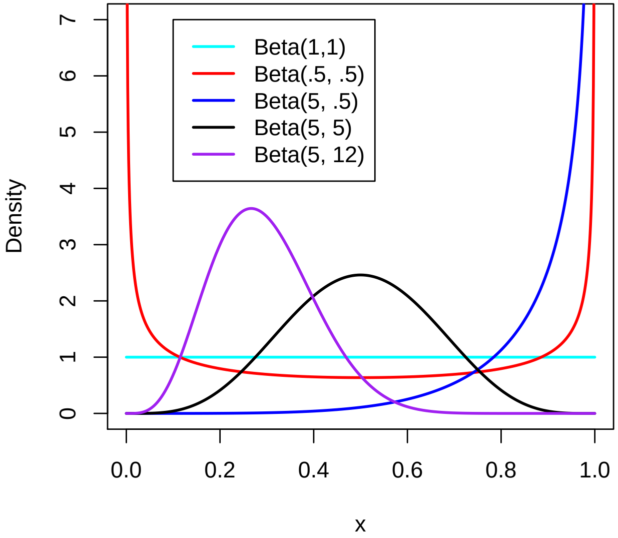
If $X \sim \mathcal{B}eta(a, b)$ then

$$\mathbb{E}[X] = \frac{a}{a+b}$$

and

$$\mathbb{V}[X] = \frac{ab}{(a+b)^2(a+b+1)}$$

Examples of Some Beta Density Functions



The Beta-Binomial Distribution

We can also consider situations where the probability of success θ is also a random variable.

Exactly what we are willing to assume about the distribution of θ determines the distribution for X.

The Beta-Binomial Distribution

Consider the following two-step data generating process:

1. draw $\theta \sim \mathcal{B}eta(\alpha, \beta)$
2. draw $X \sim \mathcal{B}inomial(n, \theta)$

Here X follows a beta-binomial distribution with sample size n and parameters α and β which we can also write as:

$$X \sim \mathcal{B}eta\text{-}\mathcal{B}inom(n, \alpha, \beta)$$

The Beta-Binomial Distribution

The beta-binomial probability density function is

$$f_{\mathcal{B}eta\text{-}\mathcal{B}inom}(x|n, \alpha, \beta) = \binom{n}{x} \frac{B(x + \alpha, n - x + \beta)}{B(\alpha, \beta)}$$

where B is the beta function.

The Beta-Binomial Distribution

An implementation of the beta-binomial density function in `R` :

```
dbetabinom <- function(x, n, alpha, beta,
  log=FALSE){
    logfun <- lchoose(n, x) +
                  lbeta(x + alpha, n - x + beta) -
                    lbeta(alpha, beta)
    if (log==TRUE){
      return(logfun)
    }
    else{
      return(exp(logfun))
    }
  }
```

The Beta-Binomial Distribution

If $X \sim \mathcal{B}eta\text{-}\mathcal{B}inom(n, \alpha, \beta)$ then

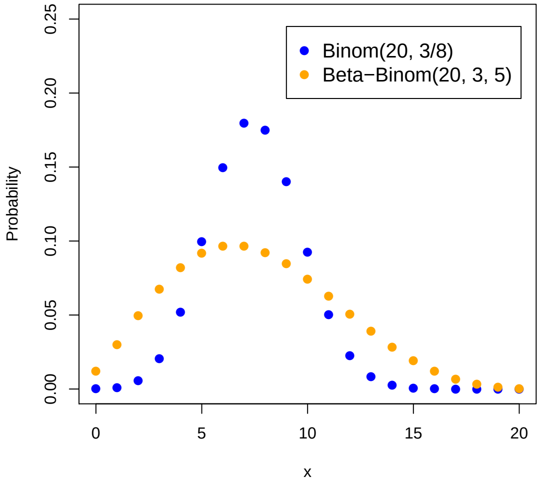
$$\mathbb{E}[X] = \frac{n\alpha}{\alpha + \beta}$$

and

$$\mathbb{V}[X] = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

The Beta-Binomial Distribution

The beta-binomial is over-dispersed relative to the binomial distribution:



The Multinomial Distribution

If a k-vector Y follows a multinomial distribution, we write

$$Y \sim \mathcal{M}ultinom(n, \boldsymbol{\theta})$$

where $\sum_{j=1}^{k} Y_j = n$ and $\sum_{j=1}^{k} \theta_j = 1$ with each $\theta_j \in [0, 1]$.

The multinomial probability mass function is:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{n!}{y_1!y_2!\cdots y_k!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k}$$

The multinomial distribution is a generalization of the binomial distribution to more than 2 categories.

The Dirichlet Distribution

If a k-vector Y follows a Dirichlet distribution, we write

$$Y \sim \mathcal{D}irichlet(\boldsymbol{\alpha})$$

where $\sum_{j=1}^{k} Y_j = 1$ and $\alpha_j > 0 \quad j = 1, \ldots, k$.

The Dirichlet probability density function is:

$$p(\mathbf{y}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} y_1^{(\alpha_1 - 1)} \cdots y_k^{(\alpha_k - 1)}$$

The Dirichlet distribution is a generalization of the beta distribution to more than 2 probabilities.