The multinomial distribution

The binomial distribution can be generalized to allow for more than two mutually exclusive outcomes. We consider a random variable taking one of of k possible outcomes and count the number of occurrences of each type of outcome. The vector of such counts, y has the density

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}, \;\; \sum_{j=1}^{k} \theta_j = 1$$

We assume that $n = \sum_{j=1}^{k} y_j$ is known.

The Dirichlet distribution

A generalization of the beta distribution to k components is given by the Dirichlet distribution

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}, \sum_{j=1}^{k} \theta_j = 1, \;\; \alpha_j > 0, \; j = 1, \ldots, k$$

This distribution is a conjugate prior for the multinomial likelihood. The corresponding posterior distribution is

$$p(\theta|y) \propto \prod_{j=1}^{k} \theta_j^{y_j + \alpha_j - 1}$$

This implies that a posterior the distribution is a Dirichlet with parameters $\alpha_j + y_j, j = 1, \ldots, k$.

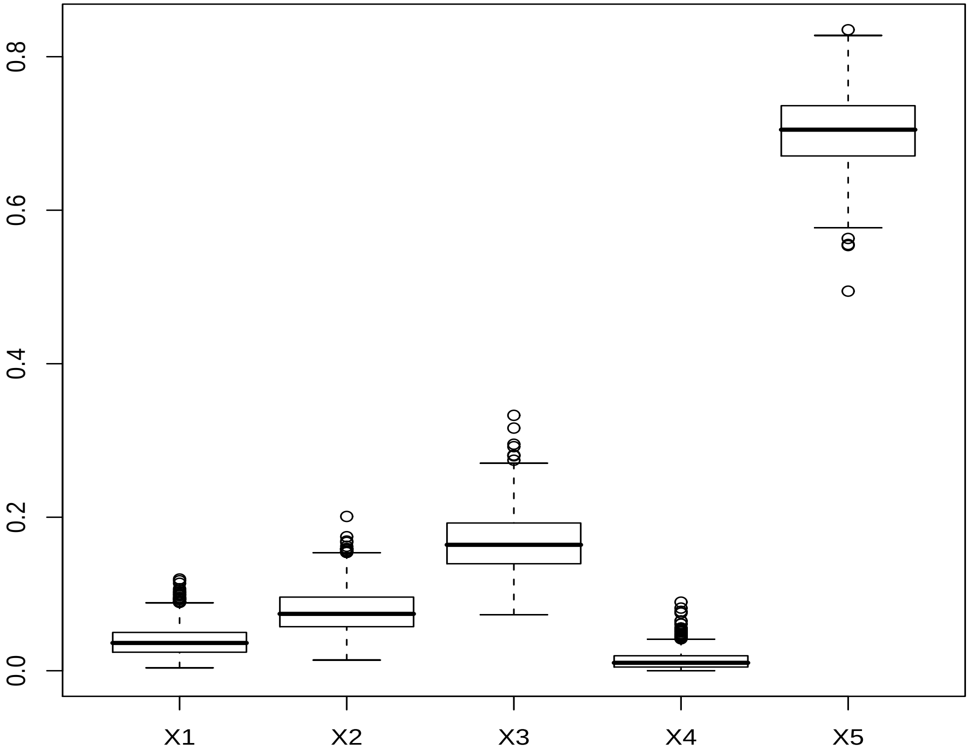
Sampling from a Dirichlet

To obtain a sample form a Dirichlet $\alpha = (\alpha_1, \ldots, \alpha_k)$ generate k independent gamma random variates $x_1, \ldots, x_k$, with shape parameters $\alpha_1, \ldots, \alpha_k$. Let

$$\theta_j = \frac{x_j}{\sum_i x_i}$$

We notice that the scale that is used to generate the gamma variates is irrelevant, as it cancels in the ratio. Incidentally, only $k-1$ variates need to be generated as the k-th is obtained from the fact that $\sum_j \theta_j = 1$.

This sample was generate with with the shape parameters $\alpha = (3.5, 7, 15, 1.2, 62)'$. There are 1,000 samples in the figure.



Properties of the Dirichlet

The marginal distribution of θ_j is beta$(\alpha_j, \alpha_0 - \alpha_j)$, where $\alpha_0 = \sum_j \alpha_j$. The marginal distribution of any sub-vector of a Dirichlet is also a Dirichlet, so

$$(\theta_i, \theta_j, 1 - \theta_i - \theta_j) \sim Dir(\alpha_i, \alpha_j, \alpha_0 - \alpha_i - \alpha_j)$$

The conditional distribution of a sub-vector of θ given the remaining elements of θ is also a Dirichlet.

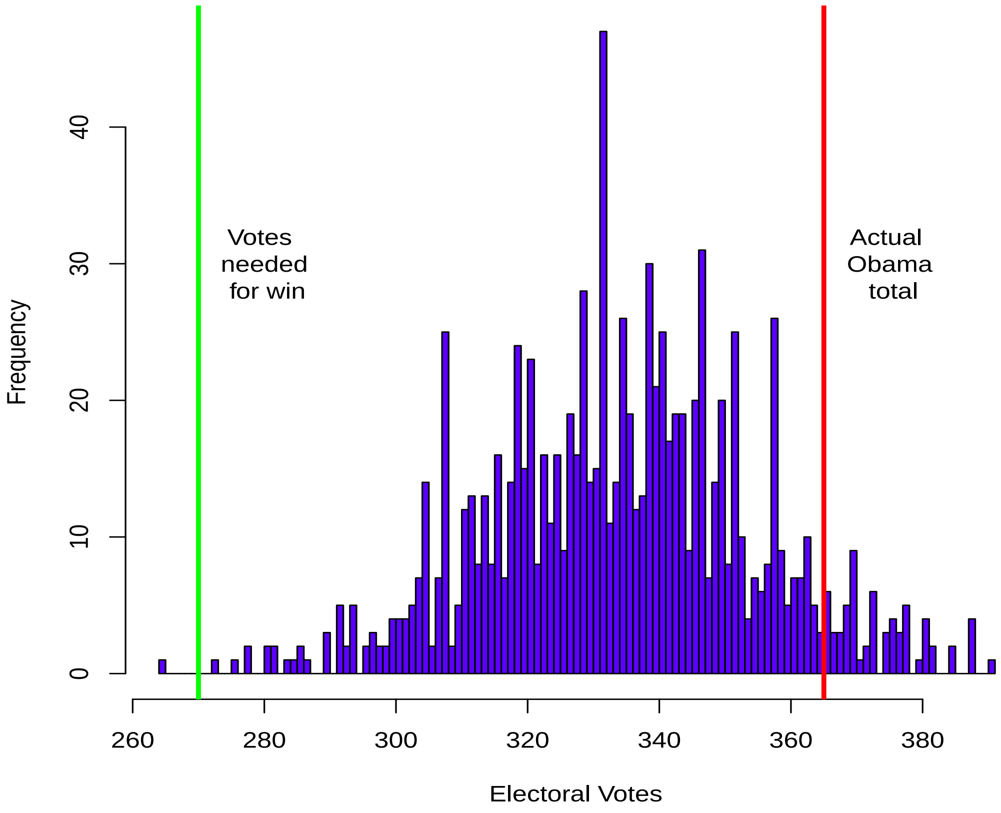
Multinomial Example

The Sunday before Election Day `www.cnn.com` gives the results of a nationwide poll by state. Let θ_{Oj} and θ_{Mj} denote the probabilities for State j, corresponding to Obama and McCain, respectively. The data available as `election.2008` in `LearnBayes`.

Assuming a uniform on $(\theta_{Oj}, \theta_{Mj}, 1 - \theta_{Oj} - \theta_{Mj})$ we have that, for a sample of 500 voters, the posterior for State j is a Dirichlet$(500q_{Oj} + 1, 500q_{Mj} + 1, 500(1 - 500q_{Oj} - 500q_{Mj} + 1))$, where q denote the reported proportions.

We use these distributions to sample from the predictive posterior distributions of the number of electoral votes carried by Obama out of 538.

This histogram was obtained from 1,000 replicates from the predictive posterior distribution of the number of electoral votes obtained by Obama.



Multivariate Normal

An k-dimensional vector y follows a multivariate normal distribution with mean vector μ and $k \times k$ variance-covariance matrix V if the density is given by

$$p(y|\mu, V) \propto |V|^{-1/2} \exp\left\{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)\right\}$$

Here μ is an arbitrary vector and V is a symmetric and *positive definite* matrix. If V is not of full rank, the distribution is defined, but it does not have a density.

A positive definitive matrix V satisfies

1. $x'Vx > 0, \forall x, \;\; x'Vx = 0 \iff x \equiv 0$
2. If λ is an eigenvalue of V, i.e., $\exists\, x \;\; s.t. Vx = \lambda x$, then $\lambda > 0$.
3. $|V| = \prod_{i=1}^{k} \lambda_i > 0$, where λ_i is an eigenvalue of V.

Properties

1. Any subset of y has a normal distribution with the corresponding sub-mean and sub-covariance matrix.

2. Any linear combination of y is also normal. Thus, for any A (compatible with y)

$$Ay \sim N(A\mu, AVA')$$

3. The conditional distributions are also normal $p(y^{(1)}|y^{(2)}) = N(m, W)$ where

$$m = \mu^{(1)} + V^{(12)}(V^{(22)})^{-1}(y^{(2)} - \mu^{(2)})$$

$$W = V^{(11)} - V^{(12)}(V^{(22)})^{-1}V^{(21)}$$

Likelihood

Given a random sample $y_1, \ldots, y_n$ from a multivariate normal we have the likelihood

$$p(y_1, \ldots, y_n | \mu, V) \propto |V|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' V^{-1} (y_i - \mu) \right\}$$

Notice that, for a collection of matrices $A_1, \ldots, A_n$

$$\text{tr}\left(\sum_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} \text{tr}(A_i)$$

Also, tr(AB) =tr(BA). Thus

$$\sum_{i=1}^{n} (y_i - \mu)' V^{-1} (y_i - \mu) = \sum_{i=1}^{n} \text{tr}\left((y_i - \mu)' V^{-1} (y_i - \mu) \right) =$$

$$\sum_{i=1}^{n} \text{tr}\left(V^{-1} (y_i - \mu)(y_i - \mu)' \right) = \text{tr}\left(V^{-1} \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)' \right)$$

So the likelihood can be written as

$$|V|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(V^{-1}S_1\right)\right\} \quad S_1 = \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)'$$

Also, by adding and subtracting $\overline{y}$ we obtain

$$|V|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[V^{-1}(S + n(\overline{y} - \mu)(\overline{y} - \mu)')\right]\right\}$$

with

$$S = \sum_{i=1}^{n}(y_i - \overline{y})(y_i - \overline{y})'$$

which shows that $\overline{y}$ and S are sufficient statistics for μ and V.

Unknown μ Vector

We place a Multivariate normal prior on $\mu \sim N_k(m_0, C_0)$. The posterior $p(\mu|Y, m_0, C_0, V)$ is a multivariate normal such that

$$\begin{aligned} p(\mu|Y, m_0, m_0) &= p(Y|\mu, V) \times p(\mu) \\ &\propto \exp[-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^{'} V^{-1}(y_i - \mu)] \\ &\quad \times \exp[-\tfrac{1}{2}(\mu - m_0)^{'} C_0^{-1}(\mu - m_0)] \\ &\sim N_k(m_1, C_1) \end{aligned}$$

where $m_1 = C_1(C_0^{-1} m_0 + nV^{-1}\bar{Y})$ and $C_1 = (C_0^{-1} + nV^{-1})^{-1}$.

Inverse Wishart Distribution

A matrix V of dimension $k \times k$ is said to follow an Inverse Wishart distribution, denoted as $V \sim W^{-1}(r_0, S_0)$, if the density is

$$p(V|r_0, S_0) \propto |V|^{-\frac{r_0+k+1}{2}} \exp[-\frac{1}{2}\mathrm{tr}(S_0 V^{-1})].$$

Unknown Covariance Matrix

Let $Y = (y_1, \ldots, y_n)$ be a sample from a multivariate normal with known mean and unknown covariance matrix V. The posterior distribution for V is

$$\begin{aligned} p(V|Y, r_0, S_0) &= p(Y|\mu, V) \times p(V) \\ &\propto |V|^{-\frac{n}{2}} \exp[-\tfrac{1}{2}\mathrm{tr}(S_1 V^{-1})] \\ &\quad \times |V|^{-\frac{r_0+k+1}{2}} \exp[-\tfrac{1}{2}\mathrm{tr}(S_0 V^{-1})] \\ &\propto |V|^{-\frac{r_1+k+1}{2}} \exp[-\tfrac{1}{2}\mathrm{tr}(S_2 V^{-1})] \\ &\sim W^{-1}(r_1, S_1) \end{aligned}$$

where $r_1 = r_0 + n$ and $S_2 = S_0 + S_1$.

Both Parameters Unknown

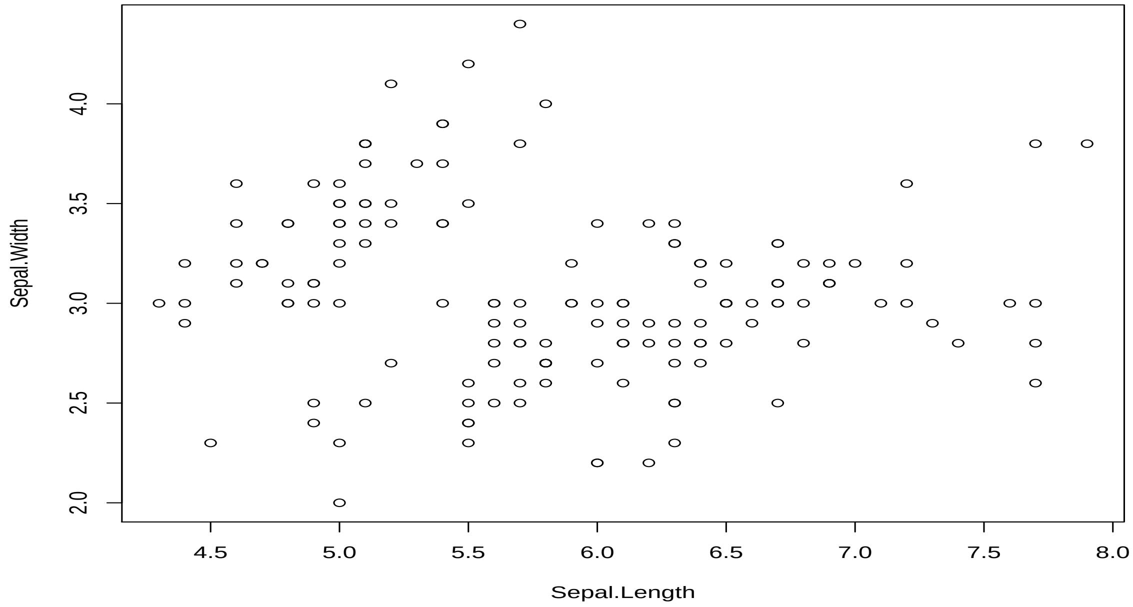
If both μ and V are unknown, we place a joint **Multivariate Normal Inverse Wishart** (NIW) prior distribution. We denote this as $(\mu, V) \sim NIW(m_0, \beta, r_0, S_0)$ with the density,

$$p(\mu, V|m_0, \beta, r_0, S_0) \quad = \quad N_k(\mu|m_0, \beta^{-1}V) \times W^{-1}(V|r_0, S_0)$$

For $Y = (y_1, \ldots, y_n)$, a sample from a multivariate normal with an unknown mean vector μ and unknown covariance matrix V, we have that the prior distribution for (μ, V) is conjugate.

Iris Sepal Length/Width

We consider Sepal Length and Sepal Width for 150 flowers. Let $Y = (y_1, \ldots, y_{150})$ y_i follows a bivariate normal distribution with unknown mean μ and covariance matrix V.

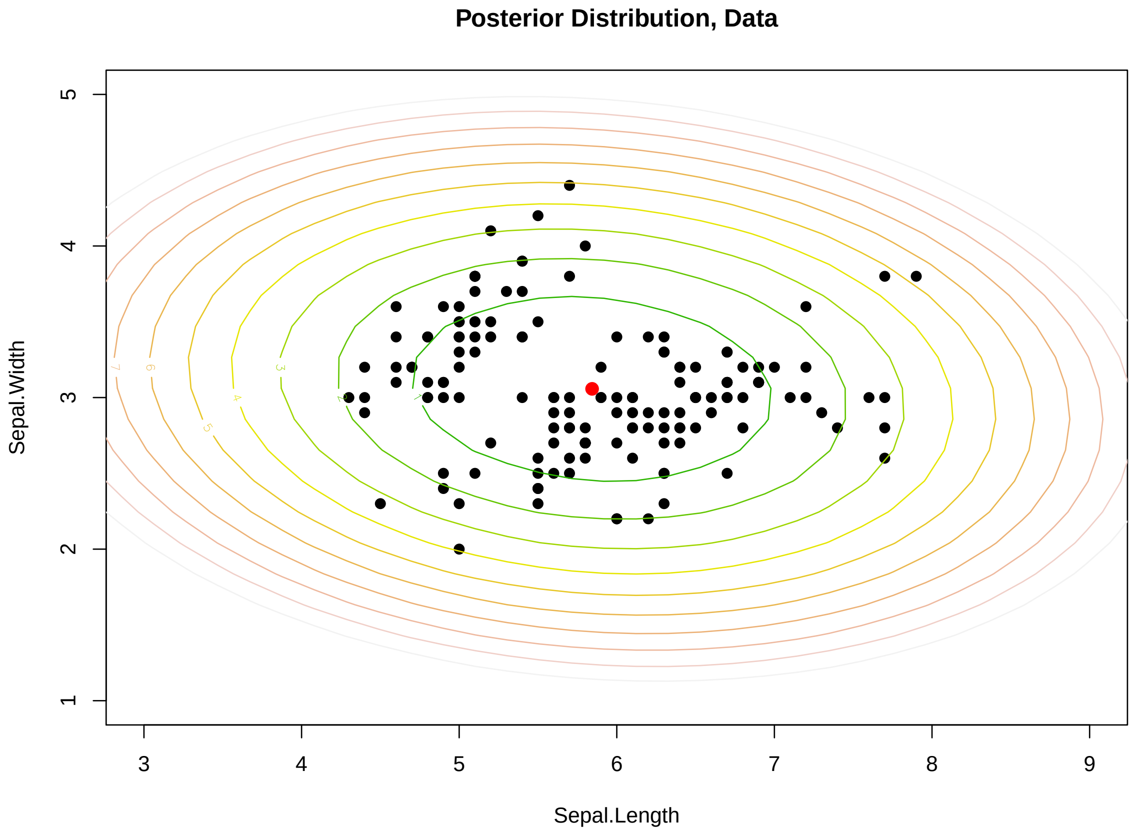


Iris Sepal Length/Width

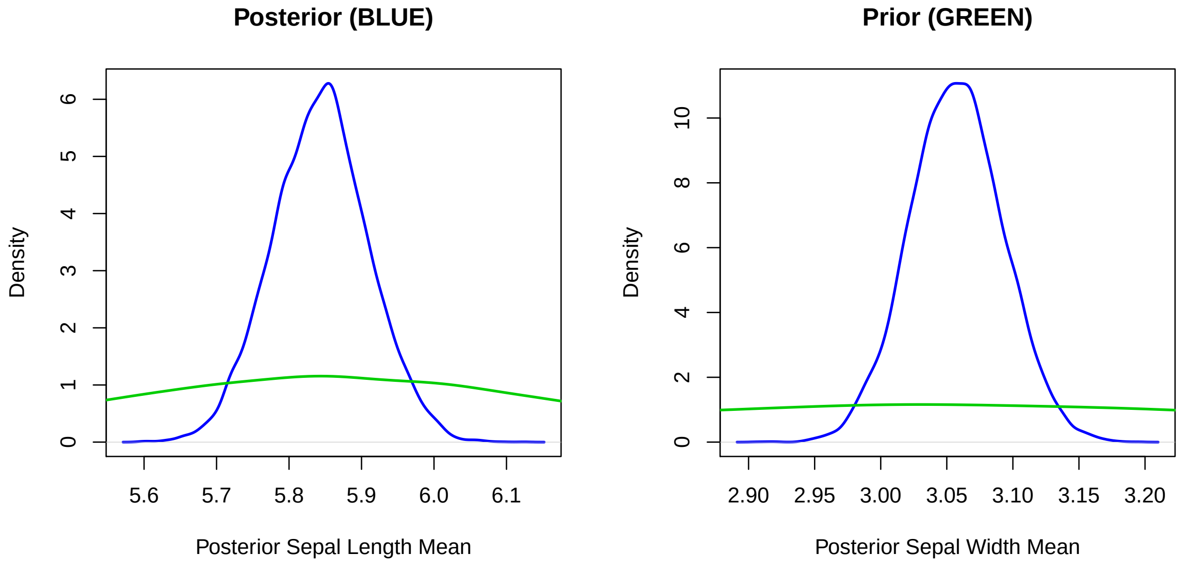
We use $(\mu, V) \sim NIW(m_0, \beta, r_0, S_0)$ with $(m_0 = Y_{MLE}, \beta = 1, r_0 = 10, S_0 = \mathbf{I_2})$. Using direct sampling we estimate the posterior distributions for μ and V.

	2.5%	50%	97.5%
Sepal Length	5.716	5.844	5.9741
Sepal Width	2.989	3.057	3.127
SL Var	0.526	0.652	0.818
SW Var	0.150	0.185	0.233
SLW CoVar	-0.099	-0.041	0.014

Iris Sepal Length/Width



Iris Sepal Length/Width



Simulating from a multivariate normal

V is positive definite, so it can be decomposed in the Cholesky factors $V = LL'$ where L is a lower triangular matrix.

If $z \sim N_k(0, I)$ then define $w = Lz$, then $Ew = 0$ and $\mathrm{var}(LZ) = LL' = V$, so, $w \sim N(0, V)$. Thus, to obtain a sample from a $N(\mu, V)$:

1. Obtain the Cholesky decomposition of $V = LL'$
2. Generate k i.i.d. $N(0, 1)$ variates
3. Pre-multiply the vector z by L
4. Add μ

Simulating from a Wishart

The Wishart is the generalization of the χ^2. A χ^2 is the sum of squares of normals. A k-dimensional Wishart with ν degrees of freedom is given as $\sum_{i=1}^{\nu} z_i z_i'$, where $z_i \sim N_k(0, I)$ and they are independent.

The density of a Wishart $V \sim Wi_k(\nu, S)$ is given by

$$p(V) \propto |V|^{(\nu-k-1)/2} \exp\left\{\frac{1}{2} tr(S^{-1}V)\right\}$$

One way to obtain a sample from a $Wi_k(\nu, S)$, when $\nu \geq k$ is to generate ν independent normal vectors $x_i \sim N_k(0, S)$ and calculate $\sum_{i=1}^{\nu} x_i x_i'$.

Note that this method implies generating $\nu \times k$ random variables. For a sample of an inverse Wishart, we invert the results of the previous procedure.

An alternative (much faster procedure) is to use Bartlet's decomposition:

1. Generate a lower triangular $k \times k$ matrix A, s.t.:
   (a) $a_{ii} \sim \sqrt{\chi^2_{\nu-i+1}}$
   (b) $a_{ij} \sim N(0,1), \quad j < i$
2. Compute the Cholesky decomposition $S = LL'$.
3. Calculate $LAA'L'$

This method requires only $k \times (k+1)/2$ random variable generations (no dependence on the degrees of freedom). Furthermore, the method directly produces the Cholesky factor of the sample: LA.