The Dirichlet Family

Matt Beverlin and Paul Johnson

June 10, 2013

1 Introduction

The Dirichlet distribution is a multivariate distribution, meaning that a single outcome is actually a vector of numbers. The elements in this vector are are all between 0 and 1. A particular observation from a Dirichlet distribution would look like this

$$(0.2, 0.3, 0.5)$$

for a case describing a three dimensional outcome, or a $3-tuple$ for short.

The three numbers in this vector represent the probabilities of three different mutually exclusive events. Because the elements are probabilities, they sum to one.

The Dirichlet distribution gives a formula which tells how likely we are to observe a particular $3-tuple$.

Any logically meaningful combination of the elements in the vector can occur. Logically meaningful means that each element must be 0.0 or greater and that all of the numbers must sum to 1.

For example, suppose we are considering a situation in which a randomly drawn person has black, red/blonde, or brown hair. Speaking for myself as an individual, I would guess that the probabilities would be like this:

$$(0.15, 0.2, 0.65)$$

You, on the other hand, might have the opinion that the probabilities are like this:

$$(0.1, 0.15, 0.75)$$

And your friend Bob says the probabilities are like this:

$$(0.33, 0.33, 0.34)$$

If we wanted to go through and find out what everybody thinks, we would accumulate a lot

of these vectors, and the only thing they seem to have in common is that they all add up to 1.0.

We would like to have a probability model that tells us how likely each vector of beliefs is to appear in a sample. That is what Dirichlet is for. Dirichlet describes a wealth of possible distributions of opinion. It can be as simple as the statement that "all belief vectors are equally likely to occur". It need not place equal weight on all probability assignments, however. It has parameters which can lead you to expect that the most likely combination is mine and that other combinations with high weight on blondes are less likely.

Here is a point of caution. We are setting up a model that gives "probabilities about probabilities." That's confusing. There is inevitable confusion over various possible uses of the "letter p". Because the Dirichlet describes a vector of probabilities, the letter p is used to refer to the observed outcome. Possible values are labeled as $p = (p_1, p_2, ..., p_L)$. Then we are going to want to calculate the probability of observing that L-tuple. If you use the letter P for probabilities, then you end up with silly-looking notation like $P(p)$. Who can stand that? It might be better if we used some other letter, such as x_i or y_i, and think of the vectors in the same way we would think about the outcomes in any other kind of probability model. But we are not doing that, because doing so would cloud the fact that we really are discussing probabilities.

2 Mathematical Description

The Dirichlet Family generalizes the Beta family to a vector $p = (p_{0,}p_{1,...}p_L)$ in which $\sum_{i=0}^{L} p_i = 1$ and the $\{p_i\}$ are non-negative. Remember that p describes the outcome variable, the $L-tuple$, the one for which we want to calculate probability.

The shape of the probability model is determined by L shape parameters, $(\alpha_1, \alpha_2, ..., \alpha_L)$. These shape parameters are used to "tune" the distribution, to make certain $L-tuples$ more likely than others. As the figures presented below will illustrate, the large values of α_i correspond to actions which make outcome i "more likely".

Let the sum of the shape parameters be $\alpha = \sum_{l=0}^{L} \alpha_l$. The density function takes the form:

$$f_P(p) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_L)} p_0^{\alpha_{0-1}} \times ... \times p_\Gamma^{\alpha_{L-1}}$$

where

$$\{p_i\} \geq 0; \sum_{i=1}^{L} p_i = 1$$

and

$$\alpha \geq 0;$$
$$\sum_{i=0}^{L} \alpha_i = \alpha$$

Recall that the Gamma function $\Gamma(k)$ is a continuous variant of the factorial function. For integers, $\Gamma(k) = (k-1)!$ The Gamma function is described and illustrated in more depth in the discussion of the Gamma probability model.

3 Dirichlet is useful in Bayesian analysis

In Bayesian analysis, one needs probability models to summarize his/her beliefs about the world. Suppose you asked me the following. "We are going to survey people and ask them what fraction of the population has black, red/blonde, and brown hair. What do you expect will be the distribution of outcomes?" In response, I realize it is nonsense for me to simply give a univariate prediction, such as "the average proportion for brown will be 0.33." Not only must I specify probabilities for the other hair colors, I also have to be more modes in my view of the world. If I say the probabilities are

$$(0.2, 0.15, 0.65)$$

I should not act as though I think those probabilities are exactly right. This vector may represent my belief about the most likely combination, but it does not summarize the entirety of my view of the world. Instead, I should have some picture in my mind of how all other possible $3-tuples$ might will fit together into a mosaic. I have an idea of what the most likely 3-tuple is, and I also am pretty sure that

$$(0.05, 0.90, 0.05)$$

will almost never occur. But I don't think it is impossible.

The Beta distribution is a way that we summarize the distribution of one variable on the 0 to 1 continuum. If two or more variables on $[0,1]$ are being considered, then the Dirichlet is simply the multivariate generalization of the Beta.

4 Means, Variance, and Covariances

Consider a set of Dirichlet "shape" parameters $(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_L)$, the sum of which is $\alpha(\alpha = \sum \alpha_j)$. The expected value of any individual component is

$$E(p_j) = \frac{\alpha_j}{\alpha_1 + \alpha_2 + \cdots + \alpha_L} = \frac{\alpha_j}{\alpha}$$

The variance is

$$V(p_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}$$

and the covariance between two values is:

$$C(p_i p_j) = -\frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)}$$

Please observe that Covariance is very much in the nature of the beast with this distribution. Since $\sum p_j = 1$, any change in any one of the values must change at least one of the others.

5 Graphic Representation

Return again to the 3-tuple which gives the probability of black, red/blonde, and brown hair. If we specify the probability of black (p_1) and red/blonde (p_2), then the probability of brown is not open to question because the three probabilities must sum to one.

$$p_3 = 1 - p_1 - p_2$$

Furthermore, we know that the probability of observing $p_1 + p_2 > 1$ is equal to 0. It is impossible!

With that in mind, we have created some illustrations of a 3-tuple under the Dirichlet distribution in an effort to help the reader make a mental picture. The third element of the probability vector, p_3 , is implicit in these graphs because we show only the probability of observing (p_1, p_2).

In the package "gtools" the function ddirchlet gives the probability of observing any 3-tuple, given a set of shape parameters. The following R code will create the range of possible 3-tuples as inputs, and it calculates the probability of observing each one.

```
library (gtools)
N <- 100
y1 <-seq(0.001, 0.999, length.out=N)
y2 <-seq(0.001, 0.999, length.out=N)
z <- matrix (0,N,N)
myz <-function(a1,a2,a3){
    z <- matrix (NA,N,N)
    for (i in 1:N) {

        for (j in 1:N) {
          ddirchprob <- ddirichlet( c(y1[i],y2[j],1-y1[i]-y2[j]),
              c(a1,a2,a3))
          z[i,j] <- ifelse (y1[i]+y2[j]<1.02, ddirchprob, NA)
    }
 }
 z
 }
z1 <-myz(1,1,1)
persp(y1,y2,z1, theta = 100, phi = 40, xlab="p1", ylab="p2", zlab
  ="probability", ticktype="detailed")
```

If the shape parameters for all dimensions are identical, say $(\alpha_1, \alpha_2, \alpha_3) = (1,1,1)$ or $(2,2,2)$, then all feasible 3-tuples are equally likely. Any triplet of the form $(p_1, p_2, 1-p_1-p_2)$ will be equally possible. A figure which shows that all feasible outcomes are equally likely is presented in Figure 1.

On the other hand, suppose that there is decidedly less weight on the first element, and more is placed on the other two, as in $(\alpha_1, \alpha_2, \alpha_3) = (1,3,3)$. That probability density is shown in Figure 2

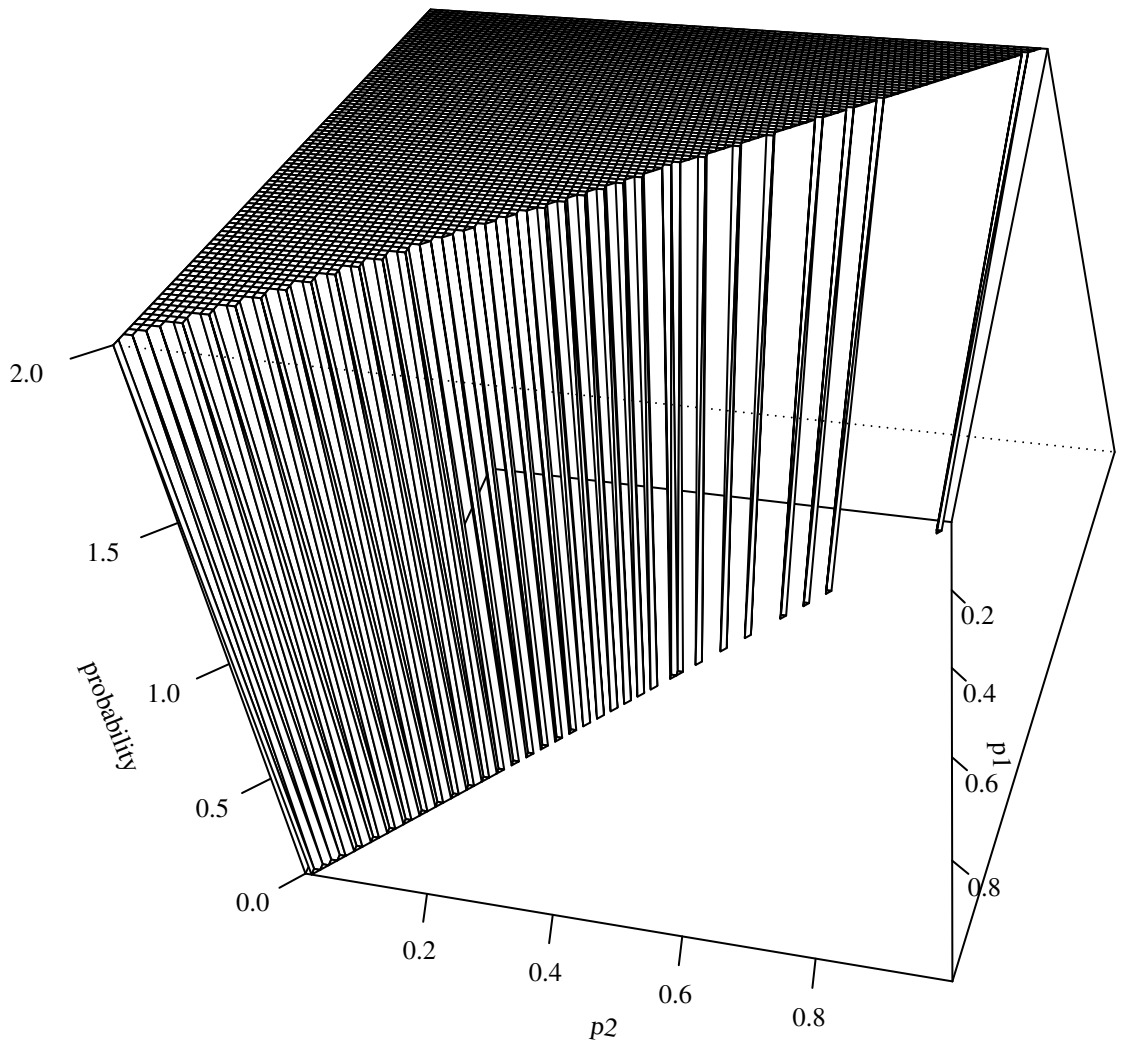

Figure 1: The Dirichlet Density (1,1,1)

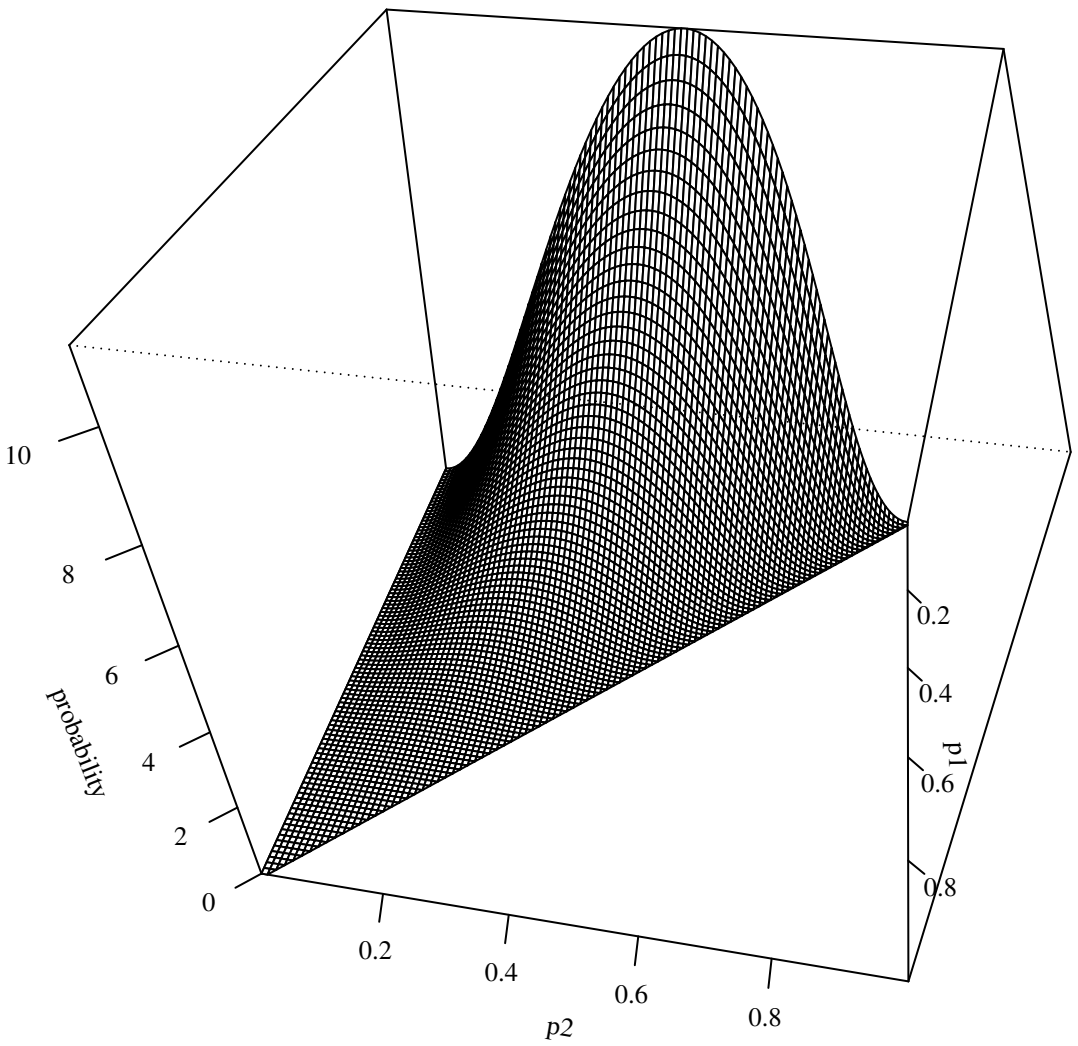

Figure 2: Dirichlet (1,3,3)

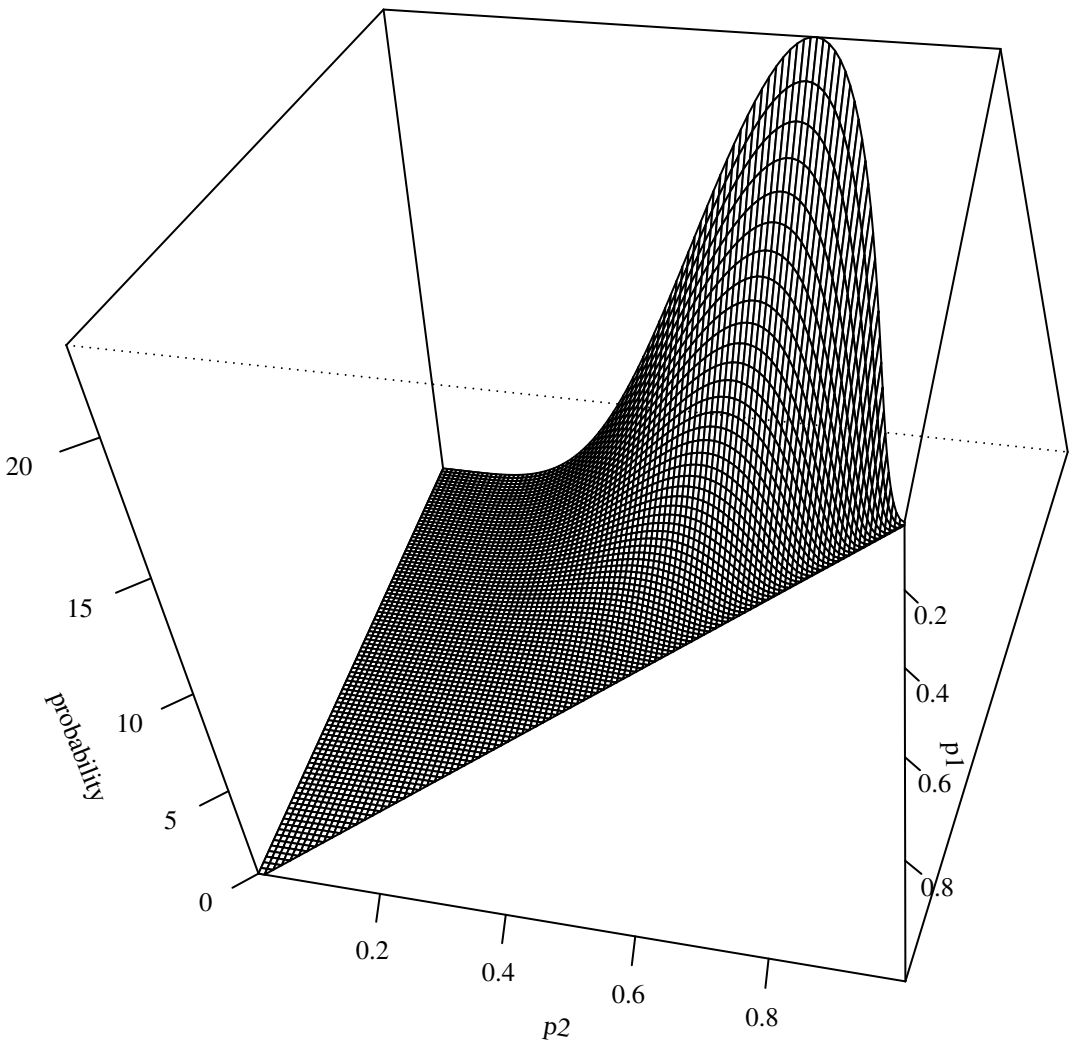

Figure 3: Dirichlet (1,6,3)

Suppose that we hike up the shape parameter on the second dimension, so that we have $(\alpha_1, \alpha_2, \alpha_3) = (1, 6, 3)$. That probability density is shown in Figure 3.

Once can continue in this vein forever, of course (and, please believe we have :)).