Multinomial inference and the Dirichlet distribution

14–21 Feb 2018; 7 Feb 2020

1 The Multinomial Distribution

The multinomial distribution is a model for situations where there are K possible categories ($K \geq 2$) for observations, each with independent probability α_i ($i = 1$ to K), and one reports the number of times n_i that each outcome was observed after some total number of observations $N = \sum_i n_i$. We will call the result of a single observation (adding a single count to one of the n_i) a *trial*; when the observations are done, the collection of N observations, summarized by the K numbers n_i, we will call a *sample*. Under these conditions, the probability for the data given known values of the α_i is,

$$p(n|\alpha) = N! \prod_{i=1}^{M} \frac{\alpha_i^{n_i}}{n_i!}, \tag{1}$$

where I use α to stand for the whole set, $\{\alpha_i\}$, and n to stand for $\{n_i\}$. I've used the symbol α_i for the value of the probability given to outcome i instead of something like p_i so that the notation doesn't get confusing when we talk about probabilities for parameters whose values happen to be the value of a probability! I chose α because the probabilities for the outcomes are often thought of as proportions in an infinite ensemble (the symbol for proportionality is an elongated alpha, $\propto$), although this interpretation is neither necessary nor desirable in all applications.

When $K = 2$, equation (1) becomes the better-known *binomial distribution*,

$$p(n_1, n_2|\alpha) = \frac{(n_1 + n_2)!}{n_1! n_2!} \alpha^{n_1} (1 - \alpha)^{N_2}, \tag{2}$$

where I've changed notation a bit for simplicity: $\alpha_1 \to \alpha$, and $\alpha_2 \to (1 - \alpha)$. This is often used to model coin flips, for example. With $K = 6$, the mulitnomial distribution is often used to model tosses of a die.

This establishes the basic notation and terminology. You can skip to the next section for a summary of basic multinomial inferences; the remainder of this section merely collects some standard properties of the mulitnomial that may come in handy.

Normalization and moments. Since it's a probability for n, it had better be normalized over all possible sets n. We can readily verify this with a trick you may have previously encountered in a basic probability or StatMech course. Do a multinomial expansion of $(\alpha_1 + \alpha_2 + \cdots + \alpha_K)^N$ to find that

$$(\alpha_1 + \alpha_2 + \cdots + \alpha_K)^N = \sum_{\mathcal{N}} N! \prod_{i=1}^{K} \frac{\alpha_i^{n_i}}{n_i!}, \tag{3}$$

where $\mathcal{N}$ denotes the set of all K-tuples of whole numbers n_i such that $\sum_i n_i = N$. Note that each term in this sum is of the same form as the multinomial probability in equation (1). Now note that since the α_i themselves are values of probabilities, $(\alpha_1 + \alpha_2 + \cdots + \alpha_K) = 1$. Thus the left hand side of equation (3) is unity, and equation (3) thus proves normalization.

This multinomial expansion trick comes in handy for other calculations, too. Take the derivative of equation (3) with respect to α_j; a couple lines of algebra yields

$$N(\alpha_1 + \cdots)^{N-1} = \frac{1}{\alpha_j} \sum_{\mathcal{N}} n_j p(n|\alpha). \tag{4}$$

Normalization of the α_i implies that the left hand side equals N. The sum on the right hand side is just the expectation value for n_j. Rearranging, we thus have,

$$\langle n_j \rangle = \alpha_j N, \tag{5}$$

that is, the expected value for the number of trials with outcome j is just α_j times the total number of trials. Take another derivative, and it's a bit messier, but you can get the second moment of the distribution, $\langle n_j^2 \rangle$. Combine it with the first moment to get the standard deviation,

$$\sigma_{n_j} \equiv \langle (n_j - \langle n_j \rangle)^2 \rangle^{1/2} = \sqrt{N\alpha_j(1-\alpha_j)}. \tag{6}$$

Thus the standard deviation grows like "root-N."

Marginal distributions. An important property of the multinomial distribution is that the marginal distribution for the number of trials in any particular category (say, category j) is a binomial distribution (in this case, with probability α_j). To see this, let $\mathcal{N}_j$ denote the set of all possible K-tuples of whole numbers $\{n_i\}$ with the j'th integer fixed at n_j, and the rest constrained so that the total is N. (This is just the set of $K-1$ whole numbers that sum to $N - n_j$.) Then the marginal distribution for n_j (the probability for n_j, regardless of the values of the other n_i) can be found as follows:

$$p(n_j|\alpha) = \sum_{\mathcal{N}_j} p(n|\alpha) \tag{7}$$

$$= \frac{N!\alpha_j^{n_j}}{n_j!} \sum_{\mathcal{N}_j} \prod_{i \neq j} \frac{\alpha_i^{n_i}}{n_i!} \tag{8}$$

$$= \frac{N!\alpha_j^{n_j}}{n_j!(N-n_j)!} \times \left[(N-n_j)! \sum_{\mathcal{N}_j} \prod_{i \neq j} \frac{\alpha_i^{n_i}}{n_i!} \right] \tag{9}$$

$$= \frac{N!}{n_j!(N-n_j)!} \alpha_j^{n_j} (1-\alpha_j)^{N-n_j}, \tag{10}$$

where to get the last line we identified the multinomial expansion of $(\sum_{i \neq j} \alpha_i)^{N-n_j}$, and used the fact that normalization implies $\sum_{i \neq j} \alpha_i = (1 - \alpha_j)$. This is convenient because when we are interested in a particular outcome, this result tells us we can ignore details about the other outcomes.

Finding modes. One last trick that is good to know for handling distributions for discrete (as opposed to continuous) quantities is a trick for finding the integer n that maximizes some function of interest $f(n)$ (in our case we seek the value of n_j that maximizes the probability $p(n_j|\alpha)$). We can locate the maximum by treating n as a continuous variable and taking the integer part of our solution. We could try the obvious thing and set a derivative equal to zero, but this gets to be messy real quick in our problem because n_j appears in factorials.

Instead, imagine a plot of $f(n)$ vs. n, which we can represent as a discrete set of points on a f vs. n plot. Draw a smooth curve through these points, representing $f(n)$ when we let n be continuous. Now imagine sliding around a little horizontal line segment of length 1 with its left endpoint on the curve. At some point near the peak of the curve, its right endpoint will hit the curve; the mode (the integer value of n that maximizes the probability) will lie within the interval spanned by the segment. Thus if we can find the real number x such that $f(x) = f(x-1)$, the integer that maximizes $f(n)$ is the greatest integer less than or equal to x (the "floor" of x, denoted $\lfloor x \rfloor$). Cute! I wish I knew who invented this trick; I learned of it from Ed Jaynes's writing.

For the problem at hand, we thus set

$$\frac{N!}{x!(N-x)!}\alpha_j^x(1-\alpha_j)^{N-x} = \frac{N!}{(x-1)!(N-x+1)!}\alpha_j^{x-1}(1-\alpha_j)^{N-x+1}. \tag{11}$$

Thanks to the fact that $n! = n(n-1)!$, there is a lot of cancelation and we're left with $x = \alpha_j(N+1)$, so that the mode of the marginal distribution is

$$\hat{n}_j = \lfloor \alpha_j(N+1) \rfloor. \tag{12}$$

So, for example, if $\alpha_j = \frac{1}{2}$ and $N = 100$, the most probable value of n_j is 50. Note that this is the mode of the *marginal* distribution for n_j. If we want to find the mode of the joint distribution for all the n_i, we need to solve a bunch of simultaneous equations (which may have multiple solutions). Let's not bother with this right now! But it is perhaps worth noting in closing that the mode of a marginal distribution for some quantity need not equal the value of that quantity that maximizes a higher dimensional joint distribution.

2 Bayesian Inference

In the Bayesian approach to statistical inference, we assess hypotheses H_i for data D by calculating probabilities for the hypotheses given the data and any other contextual information, $\mathcal{C}$, we may have (including such things as a model or models connecting the data and the hypotheses). A main tool for doing such calculations is Bayes's theorem,

$$p(H_i|D,\mathcal{C}) = p(H_i|\mathcal{C})\frac{p(D|H_i,\mathcal{C})}{p(D|\mathcal{C})}. \tag{13}$$

For standard terminology and a basic introduction to Bayesian parameter estimation and model comparison in just a few pages, see § 2 of Gregory and Loredo (1992) (GL92 hereafter). All we need to know here is that the *posterior probability* for a hypothesis, $p(H_i|D,\mathcal{C})$, is proportional to the product of a *prior probability*, $p(H_i|\mathcal{C})$ ("prior" meaning prior to

consideration of the data), and $p(D|H_i, \mathcal{C})$, which is called the *sampling distribution* for the data, or the *likelihood* for the hypothesis. The likelihood terminology emphasizes that what is relevant about $p(D|H_i, \mathcal{C})$ is its dependence on H_i; the likelihood is often written more simply as $\mathcal{L}(H_i)$.

We are typically concerned with hypotheses about parameterized models. Calculations proceed differently depending on the types of hypotheses of interest. If we presume a particular parameterized model $\mathcal{M}$ is true (taking $I = \mathcal{M}$, though I could include information in addition to specification of the model), and we are interested in hypotheses about the values of the parameters, the problem is one of *parameter estimation*. If we have rival parameterized models and we want to compare them, we would take $I = \mathcal{M}_1 + \mathcal{M}_2 + \cdots$ (denoting the proposition, "model $\mathcal{M}_1$ is true, or model $\mathcal{M}_2$ is true, or..."), and the problem is one of *model comparison*. The resulting calculations differ, but are related; we consider them in turn.

2.1 Parameter Estimation

Parameter estimation with a multinomial model refers to estimation of the trial probabilities, α. Bayes's theorem for the joint posterior distribution for all K parameters, α_i, is,

$$p(\alpha|n, \mathcal{M}) = \frac{p(\alpha|\mathcal{M})\,\mathcal{L}(\alpha)}{Z}, \tag{14}$$

where the likelihood for α is the multinomial distribution for n,

$$\mathcal{L}(\alpha) \equiv p(n|\alpha, \mathcal{M}) \tag{15}$$

$$= N! \prod_{i=1}^{M} \frac{\alpha_i^{n_i}}{n_i!}, \tag{16}$$

and Z is a normalization constant,

$$Z \equiv p(n|\mathcal{M}) \tag{17}$$

$$= \int d^K\alpha\; p(\alpha|M)\mathcal{L}(\alpha). \tag{18}$$

This quantity is called the prior predictive probability for n, or the average likelihood for α, the marginal likelihood, or the global likelihood for $\mathcal{M}$; reasons for the latter names will become apparent below.

To go further, we must specify the prior distribution for α. Any such prior must account for the fact that the α_i sum to unity, so it must be proportional to a δ-function, $\delta(1 - \sum \alpha_i)$. If we are starting in a state of initial ignorance, an intuitively appealing choice is a flat distribution,

$$p(\alpha|\mathcal{M}) = C_K \delta\left(1 - \sum_{i=1}^{K} \alpha_i\right), \tag{19}$$

where C_K is a normalization constant that depends only on the number of outcomes, K. [*Note:* This prior is not as uninformative as intuition may suggest and is probably a poor

choice when K is large; we'll discuss this elsewhere.] To find C_K we will use the *generalized beta integral* (GBI),

$$\int_0^\infty dx_1 \ldots \int_0^\infty dx_m x_1^{k_1-1} \ldots x_m^{k_m-1}\, \delta\left(a - \sum_{j=1}^m x_j\right) = \frac{\Gamma(k_1)\ldots\Gamma(k_m)}{\Gamma(k)}\, a^{k-1}, \tag{20}$$

where $k = \sum_j k_j$, and $\Gamma(x)$ is the Gamma function, with $\Gamma(n) = (n-1)!$ when n is a positive integer. (Note that the GBI is used in GL92, but there's a typo where it's defined there.) We are interested in the case with $a = 1$ and all $k_j = 1$, so $k = K$. Using this to integrate equation (19) gives

$$C_K = (K-1)!. \tag{21}$$

With the prior now specified, we can calculate Z according to equation (18). This requires another (less trivial) application of the GBI, giving

$$Z = \frac{(K-1)!N!}{(N+K-1)!}. \tag{22}$$

We now have all the ingredients needed to calculate the joint posterior for α;

$$p(\alpha|n,\mathcal{M}) = \frac{(N+K-1)!}{\prod_i n_i!}\, \alpha_1^{n_1}\cdots\alpha_K^{n_K}\, \delta\left(1 - \sum \alpha_i\right). \tag{23}$$

This looks a lot like the multinomial distribution for n (the likelihood), but it has a δ-function factor and a different factor in the numerator reflecting the fact that it is now a distribution for α rather than n, and must be appropriately normalized over α. A distribution of this form (a product of powers of the α_i and a δ-function enforcing normalization of $\sum \alpha_i$) is called a *Dirichlet distribution*.

The Bayesian solution to the parameter estimation problem is the entire joint posterior distribution, $p(\alpha|n,\mathcal{M})$. But usually we will want to summarize its salient features. We can use the GBI to find interesting properties of it. The posterior expectation value for α_j is,

$$\langle\alpha_j\rangle \equiv \int d\alpha\; \alpha_j\, p(\alpha|n,\mathcal{M}) \tag{24}$$

$$= \frac{n_j+1}{N+K}. \tag{25}$$

For large N, this is very nearly equal to n_j/N, the fraction of trials with outcome j, as one would expect. A measure of our uncertainty in the value of α_j is given by the posterior standard deviation σ_j, which satisfies

$$\sigma_j^2 = \frac{(n_j+1)(N+K-n_j)}{(N+K)^2(N+K+1)}. \tag{26}$$

For large N and $n_j \ll N$, this gives $\sigma_j \approx \sqrt{n_j}/N$; another familiar "root-n" law. If one is going to use these moments of the distribution as summaries of the posterior, it is good to keep in mind that since they are integral quantities referring to a single parameter, you have to be careful using estimates of different $\alpha - j$ jointly. In particular, the sum of the posterior

means of a set of parameters will not in general obey constraints that the parameters must obey jointly (e.g., normalization), although in this case it ends up being true that $\sum_i \langle \alpha_i \rangle = 1$.

We can also use the GBI to get the marginal distribution for a particular α_i; for example,

$$p(\alpha_1|n, \mathcal{M}) \equiv \int d\alpha_2 \cdots \int d\alpha_K \, p(\alpha|n, \mathcal{M}) \tag{27}$$

$$= \frac{(N+K-1)!}{n_1!(N+K-n_1-2)!} \alpha_1^{N_1} (1-\alpha_1)^{N+K-n_1-2}. \tag{28}$$

Such a single-parameter case of a Dirichlet distribution is called a *beta distribution*. The mean and standard deviation must be equal to the values found using the joint distribution above; using the beta integral you can easily verify this.

Now let's find the joint mode, $\{\hat{\alpha}_i\}$—the values of α_i that maximize the joint posterior. We can't just differentiate because of the δ-function. To get around this, we remove the δ-function and maximize the rest of the posterior, inserting the normalization constraint via the *method of Lagrange multipliers*. So we thus seek to maximize

$$Q = \sum_i n_i \ln \alpha_i + \lambda \left(1 - \sum_i \alpha_i \right), \tag{29}$$

where I have taken the logarithm of the posterior and dropped terms that don't affect the α dependence; we must maximize with respect to both α and the Lagrange multiplier, λ. Taking the derivative with respect to λ and requiring that it vanish simply recovers the normalization constraint. Taking the derivative with respect to α_j and requiring that it vanish gives,

$$\frac{n_j}{\hat{\alpha}_j} = \lambda, \tag{30}$$

so that $\hat{\alpha}_j = n_j/\lambda$. Inserting this into the constraint reveals that $\lambda = N$, so that

$$\hat{\alpha}_j = \frac{n_j}{N}, \tag{31}$$

as one might have guessed. Note that the mode and mean are different (though the difference becomes negligible for large samples); the posterior distribution is a bit lopsided and asymmetrical.

Point estimates like the posterior mean or mode aren't incredibly useful summaries of the posterior by themselves, since they don't tell you how uncertain you are of the values. The posterior standard deviation, calculated above, is a common summary of the uncertainty. Another common (and more easily interpreted) summary of uncertainty is a *credible region*—a region containing some fraction of the posterior probability, found by integrating the distribution over α. Marginal credible regions can easily be found using the incomplete beta integral to integrate the marginal distribution for a particular α_i of interest; the incomplete beta integral is defined as,

$$B_x(a, b) = \int_0^x t^{a-1}(1-t)^{b-1} dt. \tag{32}$$

SciPy's `scipy.special.betainc` function can compute this.

You can use quadrature (numerical integration) to find a joint credible region for all the parameters, plotted as a contour in the α space. Although there are K parameters, α_i, this space is really $K-1$ dimensional because they must all sum to unity. The parameter space is thus really an $K-1$ dimensional simplex. For $K=2$ (coin flipping), the simplex is just 1-dimensional, so it's easy to display results. For $K=3$, the simplex is a 2-dimensional triangle. It is possible to plot results in a 2-dimensional square plot parameterized by any two of the α_i. But in this special case it is also possible to make a 2-dimensional plot that displays the posterior dependence on all three parameters. I discuss this at the end of these notes. In higher dimensions, it is difficult to usefully visualize the full joint posterior, so one would usually focus attention on lower dimensional marginals.

2.2 Model Comparison

Suppose we have two parameterized models for some data, $\mathcal{M}_1$ and $\mathcal{M}_2$. We can use Bayes's theorem to calculate their posterior probabilities. Since there are only two alternatives, it is convenient to report the ratio of their probabilities rather than the two probabilities separately; this is called the *odds*. The odds favoring model 1 over model 2 is,

$$O_{12} = \frac{p(\mathcal{M}_1|\mathcal{C})}{p(\mathcal{M}_2|\mathcal{C})} \times \frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_1)}. \tag{33}$$

In the second factor, I've made the identification $p(D|\mathcal{M}_1) = p(D|\mathcal{M}_1)$, etc.; that is, I've used the fact that the proposition $(\mathcal{M}_1, \mathcal{C})$ is equivalent to the proposition $\mathcal{M}_1$ by itself (saying "model 1 is true AND one of the two models is true" is the same as saying "model 1 is true"). Also, note that the normalization constant, $p(D|\mathcal{C})$, cancels in the odds, which adds to its convenience.

The first factor in (22) is a ratio of prior probabilities for the competing models, the *prior odds*. If there is quantitative information available discriminating between the models, we can use it here; otherwise, one typically gives the models an equal shot a priori, so the prior odds is set to unity. The odds is thus equal to the second factor, which is called the *Bayes factor*,

$$B_{12} \equiv \frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_1)}. \tag{34}$$

This is a ratio of the likelihoods for the models (the probability for the data given some hypothesis is the likelihood for the hypothesis; here the hypotheses are the models).

Now note that the likelihood for a model, $p(D|\mathcal{M}_i)$, is just the quantity that we needed to normalize Bayes's theorem for parameter estimation (denoted C^{-1} above). So we know how to calculate it; if model $\mathcal{M}_i$ has parameters θ_i and likelihood function $\mathcal{L}_i(\theta_i) \equiv p(D|\theta_i, \mathcal{M}_i)$, then

$$p(D|\mathcal{M}_i) = \int d\theta_i \; p(\theta_i|\mathcal{M}_i)\mathcal{L}(\theta_i). \tag{35}$$

The likelihood for $\mathcal{M}_i$ as a whole is just the average likelihood for its parameters, with the prior $p(\theta_i|\mathcal{M}_i)$ being the averaging weight.

Now lets specialize to models with a multinomial likelihood. Suppose $\mathcal{M}_1$ is a model with *no* free parameters; from some information other than the data (e.g., theory), we are

given a set of values for each of the α_i; call these β_i (collectively, β). Then

$$p(D|\mathcal{M}_1) = \mathcal{L}(\beta), \tag{36}$$

where $\mathcal{L}(\beta)$ is just the multinomial likelihood of equation (16), evaluated with $\alpha = \beta$.

For model $\mathcal{M}_2$, we just declare the data to come from *some* multinomial distribution, but with unknown α. This is just the model $\mathcal{M}$ we considered for parameter estimation, above. We thus already know the model likelihood; it is given by (16). The Bayes factor can now be calculated;

$$B_{12} = \frac{(N+K-1)!\mathcal{L}(\beta)}{(K-1)!N!} \tag{37}$$

$$= \frac{(N+K-1)!}{(K-1)!n_1!\cdots n_K!}\,\beta_1^{n_1}\cdots\beta_K^{n_K}. \tag{38}$$

It's a simple enough equation, but not incredibly illuminating intuitively. So let's examine some special cases.

First let's consider the problem of assessing whether the results of flipping a coin N times supports the hypothesis that the flipping is fair. This is the $K = 2$ (binomial) case. Let outcome 1 be heads (h) and 2 be tails (t). The fair-flip hypothesis, $\mathcal{M}_1$, corresponds to taking $\beta_h = \beta_t = \frac{1}{2}$. Thus (26) becomes,

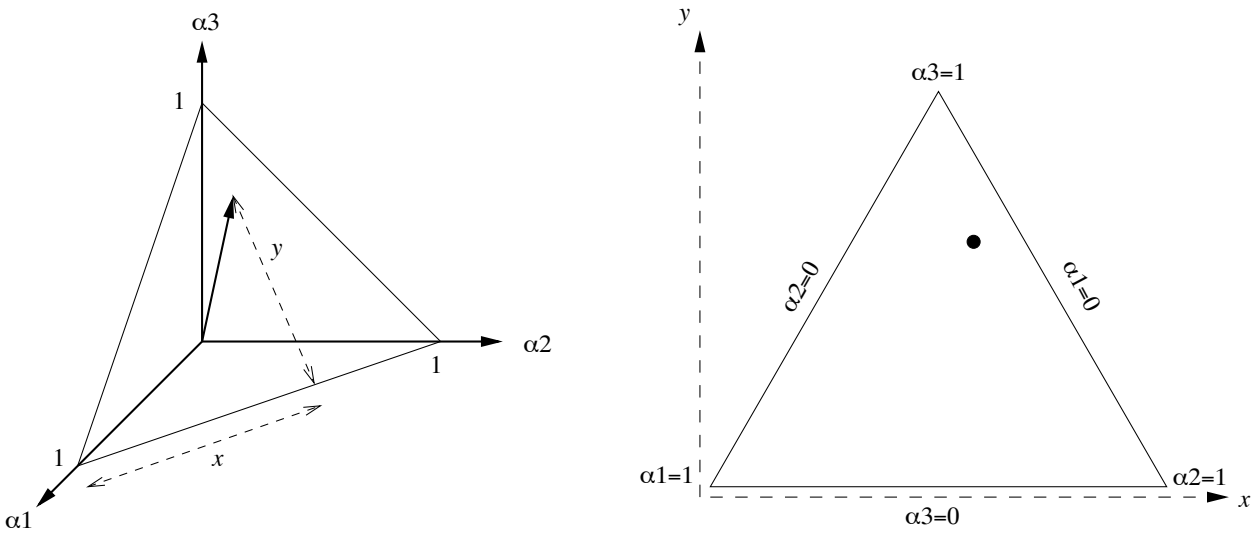
$$B_{12} = \frac{(N+1)!}{n_h!n_t!} \times \frac{1}{2^N}. \tag{39}$$

Now let's consider the case where we see $n_h = 8$ heads in $N = 12$ flips. Plugging in, we find $B_{12} = 1.57$, a Bayes factor that slightly favors the fair flip hypothesis. Yet we saw somewhat more heads than expected with fair flipping (6 are expected). If we were to estimate α_h with these data, the posterior would have a mean $\pm$ standard deviation of $\alpha_h = 0.64 \pm 0.12$. This is about 1.2σ away from the fair-flip value of 0.5. The Bayes factor tells us this isn't far enough away to lead us to prefer the hypothesis that the flipping was unfair.

This is an interesting result, because you can readily verify that by taking $\alpha_h = 0.64$ (and thus $\alpha_t = 0.36$), the data are more probable than taking $\alpha_h = 0.5$; that is, $\mathcal{L}(\alpha) > \mathcal{L}(\beta)$ if we use the best-fit value of α. But the Bayes factor doesn't use the *maximum* likelihood under $\mathcal{M}_2$, it uses the *average* likelihood, which penalizes $\mathcal{M}_2$ for its prior uncertainty in α (and thus weaker ability to predict the data). This leads to an "Occam's razor" effect that favors the simpler model unless the data significantly prefer the alternative. For more discussion of how this comes about, see Gregory and Loredo (1992).

Let's look a little more at this binomial case. Suppose we had seen $n_h = 6$ heads, just what we expect for fair flipping. In this case, we find $B = 2.93$. This favors fair flipping more strongly than $n_h = 8$, but it's by no means decisive—you wouldn't bet too much for $3:1$ odds! Thus with a dozen flips, we can never be really sure that the flipping is fair.

Now look at the opposite extreme: suppose we saw $n_h = 0$ heads. Now $B \approx 1/315$, *very* strong odds against fair flipping. So even though a dozen flips can never convince us the coin is fair, there are data such that a dozen flips can convince us it's *not* fair. This kind of asymmetry is common in Bayesian model comparison calculations.



It is perhaps worth noting that it is easy to generalize this calculation to deal with models $\mathcal{M}_1$ that are more complicated than the "simple hypothesis" (a technical term for a hypothesis with no free parameters) we considered here. For example, in a case with $K = 3$, you might have a theoretical calculation that specifies one of the α_i values, but doesn't specify how the rest of the probability is split between the other two. You would put that in via a prior for α that had a δ-function factor for the known probability. Or you might have a calculation or observation that gave you estimates for the α_i, but with some uncertainty. In this case you would use a prior that peaked at the estimates but had some width; it is easy to build such a prior with a Dirichlet distribution.

3 Plotting the $K = 3$ Case

I mentioned above that there is a way to plot contours of the $K = 3$ posterior that displays its dependence on all three α_i parameters, even though there are only two degrees of freedom. The figure shows how to do this.

On the left is shown the 3-dimensional α space. Constraining the α_i to be positive and to sum to one means the allowed parameter space is the part of the plane given by the equation $\sum \alpha_i = 1$ that lies in the positive octant. This is just the 2-dimensional tilted isosceles triangle shown on the left and duplicated by itself on the right. Since it's just 2-dimensional, we can parameterize the points in it with two coordinates, (x, y), identified in the figures. The points where each $\alpha_i = 1$ correspond to the vertices of the triangle. Each side corresponds to setting one of the $\alpha_i = 0$; only the other two parameters vary along each side. Lines with one α_i equal to a constant are parallel to the sides.

A page or two of straightforward algebra gives the following mappings between the α

coordinates and the (x, y) coordinates. To get (x, y) corresponding to $(\alpha_1, \alpha_2, \alpha_3)$, use:

$$x = \frac{1}{\sqrt{2}}(1 - \alpha_1 + \alpha_2), \tag{40}$$

$$y = \alpha_3\sqrt{\frac{3}{2}} \tag{41}$$

$$= (1 - \alpha_1 - \alpha_2)\sqrt{\frac{3}{2}}. \tag{42}$$

The x coordinate varies over $[0, \sqrt{2}]$, and the y coordinate varies over $[0, \sqrt{3/2}]$. To go in the reverse direction, pick an (x, y) pair and first check to see that it lies in the triangle. To check this, first make sure x and y individually lie in the intervals just specified. Then, if $x \leq 1/\sqrt{2}$, the point is in or on the triangle if $y \leq x\sqrt{3}$. If $x > 1/\sqrt{2}$, the point is in or on the triangle if $y \leq \sqrt{6} - x\sqrt{3}$. Inside the triangle, you can find the α_i as follows:

$$\alpha_1 = 1 - \frac{y}{\sqrt{6}} - \frac{x}{\sqrt{2}}, \tag{43}$$

$$\alpha_2 = \frac{x}{\sqrt{2}} - \frac{y}{\sqrt{6}}, \tag{44}$$

$$\alpha_3 = y\sqrt{\frac{2}{3}}. \tag{45}$$

Note that the transformation is linear, so the Jacobian is a constant; thus a probability density for (x, y) is directly proportional to one for α. You can thus plot contours of the posterior for $(\alpha_1, \alpha_2, \alpha_3)$ by stepping through a rectangular grid of (x, y) values, transforming to α inside the triangle, and recording the value of $p(\alpha|n, \mathcal{M})$; draw the resulting contours from this array with your favorite contour algorithm.

— *Tom Loredo*