Part 7. Some Useful Functions

1. The Gamma Function $\Gamma(a)$ for $a \in \mathbb{N}$

Claim 1. The Gamma function, $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$, satisfies $\Gamma(a) = (a-1)!$ for $a \in \mathbb{N}$.

Proof. $\Gamma(1) = 0! = 1$ is true, because $\Gamma(1) = \int_0^\infty e^{-t}dt = 1$. For $a > 1$, proceed via integration by parts.

$$\begin{aligned} d(uv) &= udv + vdu \\ udv &= d(uv) - vdu \\ \int udv &= uv - \int vdu \end{aligned} \tag{1.1}$$

Let $\int udv = \Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$. Pick $u = e^{-t}$ and $dv = t^{a-1}dt$. Then $du = -e^{-t}dt$, and also, because $a - 1 \geq 1$, $v = \frac{1}{a}t^a$. Plug in to Equation 1.1 to find the following:

$$\Gamma(a) = \frac{1}{a}e^{-t}t^a\Big|_0^\infty + \frac{1}{a}\int_0^\infty e^{-t}t^a dt = \frac{1}{a}\Gamma(a+1)$$

Thus, it turns out that

$$\Gamma(a+1) = a\Gamma(a) \tag{1.2}$$

for $a > 1$. Because $\Gamma(1) = 1$, it can be shown by induction on a that $\Gamma(a) = (a-1)!$, for natural $a = 1...\infty$, i.e. all $a \in \mathbb{N}$, so long as $0! \equiv 1$. □

Claim 1 gives me some intuition for the meaning of the Gamma function, for real arguments at least. As an aside, it is easily shown that the integrand $e^{-t}t^{a-1}$ in Claim 1 is maximized for $t = a - 1$.

2. Brute-force Normalization of the Beta Distribution

Claim 2. The Beta distribution

$$Be(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.1}$$

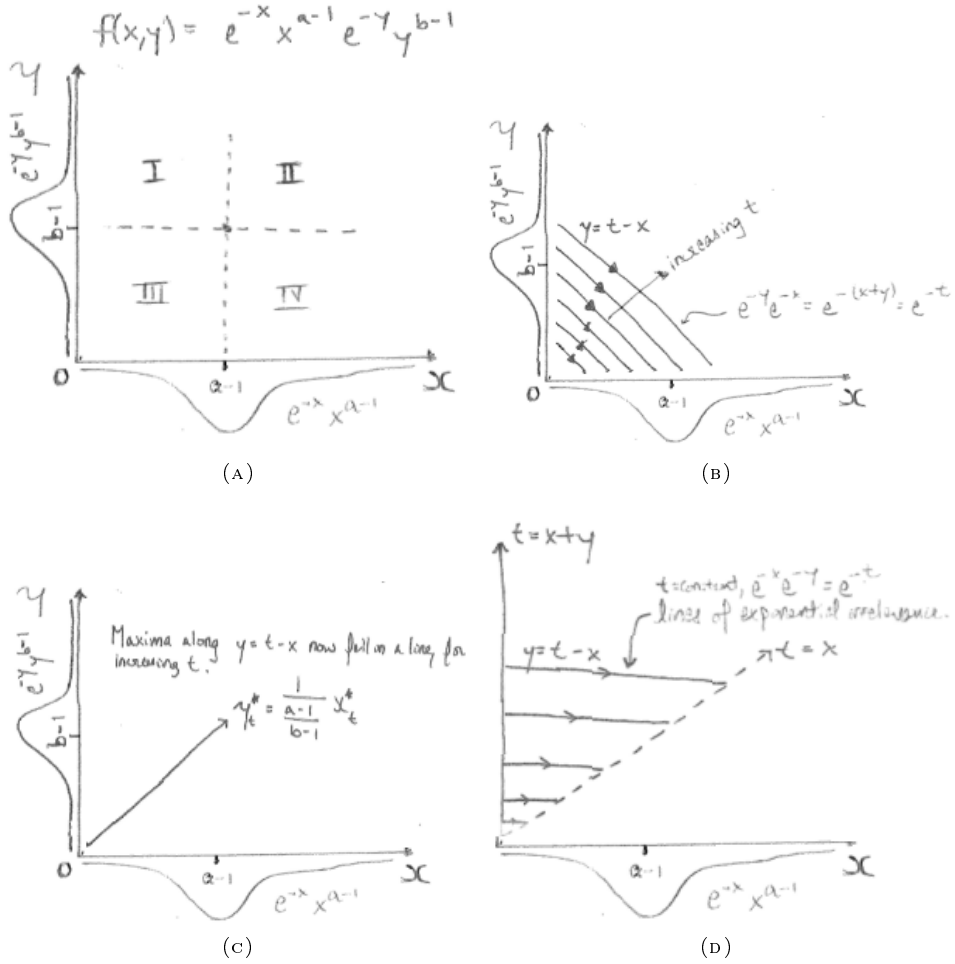
for $a, b \geq 1$, is normalized on $\mu \in [0,1]$. That is, $\int_0^1 Be(\mu|a,b)d\mu = 1$.

Proof. To prove this, I initially looked to problem 2.5 of Bishop, 2006 for hints, but it turns out to be more satisfying to intuit Bishop's hints.

The case $a, b = 1$ follows from easily from Claim 1. If $a, b > 1$, we must show that $\Gamma(a)\Gamma(b) = \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu$.

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \left[\int_0^\infty e^{-x}x^{a-1}dx\right]\left[\int_0^\infty e^{-y}y^{b-1}dy\right] \\ &= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}x^{a-1}y^{b-1}dy\right]dx \\ &= \int_0^\infty \left[\int_0^\infty f(x,y)dy\right]dx \end{aligned}$$

To see how to proceed, let us try and picture the function $f(x,y)$. Figure 2.1(a)



(A) (B) (C) (D)

FIGURE 2.1. Visualizing $f(x,y) = e^{-(x+y)}x^{a-1}y^{b-1}$. See main text for discussion.

sketches the functions $f(x,y)\Big|_y$ and $f(x,y)\Big|_x$ along the x and y-axes, respectively, to make the point that $(x,y) = (a-1, b-1)$ is a global maximum of $f(x,y)$. But what does $f(x,y)$ look like *around* its maximum?

At first glance, regions I-IV look different. Consider movement along the lines $y\Big|_t = t - x$ pictured in Figure 2.1(b). In region I, movement down such lines increases $f(x,y)$. In region IV, the same movement decreases $f(x,y)$. That makes sense, we're approaching and then leaving the global maximum.

In region III, as we walk down our lines, there's a tradeoff between the function $e^{-y}y^{b-1}$, which decreases, and the function $e^{-x}x^{a-1}$, which increases. A similar situation holds on region II. Because of this tradeoff, we expect $f(x,y)$ to achieve a local maximum along each line in both regions II and III.

So we have a simple picture of $f(x,y)$. There's a bulge in regions II and III around the global maximum at $(a-1, b-1)$.

Now, the lines $y\Big|_t = t - x$ pictured in Figure 2.1(b) seem special to the function $f(x,y)$. Suppose we travel in the positive x-direction along such a line, so $\delta x = -\delta y$, and keep track of the value of $f(x, y = t-x)\Big|_t$. For any given line (any fixed t), $e^{-(x+y)} = e^{-t}$ is a constant, so as we walk along our line, the exponentials play no role, other than to scale the polynomials $x^{a-1}y^{b-1}$ by a constant e^{-t}. Increases in x^{a-1} are still competing against decreases in y^{b-1}, of course, but this is much simpler to deal with. Analytically,

$$\begin{aligned}\frac{\partial f(x,y)}{\partial x}\Big|_{y_t = t-x} &= \partial_x[e^{-t=(x+y)}x^{a-1}(y=t-x)^{b-1}] \\ &= e^{-t}\partial_x[x^{a-1}(t-x)^{b-1}]\end{aligned}$$

Thus, the relevant competition, which determines where $f(x,y)$ is maximized along each line, is simply given by $\partial_x[x^{a-1}(t-x)^{b-1}]$. At the maximum for each line,

$$\begin{aligned}\partial_x[x^{a-1}(t-x)^{b-1}] &= (a-1)x^{a-2}(t-x)^{b-1} - (b-1)x^{a-1}(t-x)^{b-2} \\ &= 0\end{aligned}$$

We therefore have found that, for any positive x, t

$$x_{max} = \frac{\frac{a-1}{b-1}}{1+\frac{a-1}{b-1}}t$$

This result can be written another way, if we remember that our path is specified by $y_t = t - x$.

$$y_{max} = \frac{b-1}{a-1}x_{max}$$

Thus, the maxima of $f(x,y)$ along each line of the form $y_t = t - x$, for increasing t, themselves fall along a master line, as shown in Figure 2.1(c). We see that $f(x,y)$ has a natural directionality.

In Figure 2.1(d), we transform to t-x space, where $t = x + y$, and consider lines of constant t for $t \geq x$ (as $y \geq 0$). We are simply looking at the lines $y = t - x \iff t = x + y$ in a different way. In t-x space, these lines are simply horizontal lines of constant t. Along each of these horizontal lines, we have differential changes in $f(x, t-x)$, as a function of x, given by

$$\frac{\partial f(x,y)}{\partial x}\Big|_{y_t=t-x} = \frac{\partial f(x, y=t-x)}{\partial x}\Big|_t = e^{-t}\partial_x[x^{a-1}(t-x)^{b-1}]$$

as before. We've already shown that the point x_{max} at which $f(x,y)$ is maximized along these lines shifts out linearly with t. Now, because $y \geq 0$, each line of constant t is only of interest till $t = x$. What happens if we non-uniformly scrunch the x-axis as a function of t? Do our maxima collapse to the same place in this scrunched new axis? Let's define, for each fixed $t > 0$,

$$\mu\Big|_t = \frac{x}{t}$$

and consider the change in $f_1(\mu, t) = f(x = \mu t, y = t - x = t(1-\mu))$ as we move in μ at fixed t.

$$\begin{aligned}
\frac{\partial f_1(\mu,t)}{\partial \mu}\Big|_t &= \frac{\partial f(x(\mu,t), t - x(\mu,t))}{\partial \mu}\Big|_t \\
&= \frac{\partial f(x, t-x)}{\partial x}\Big|_{t, x=\mu t} \left[\frac{dx}{d\mu}\Big|_t\right] \\
&= \partial_x \left[e^{-t} x^{a-1}(t-x)^{b-1}\right]\Big|_{t,x=\mu t} [t] \\
&= te^{-t}\left[(a-1)x^{a-2}(t-x)^{b-1} - (b-1)x^{a-1}(t-x)^{b-2}\right]\Big|_{t,x=\mu t} \\
&= te^{-t}\left[(a-1)(\mu t)^{a-2}(t(1-\mu))^{b-1} - (b-1)(\mu t)^{a-1}(t(1-\mu))^{b-2}\right] \\
&= e^{-t} t^{a+b-2} \partial_\mu [\mu^{a-1}(1-\mu)^{b-1}]
\end{aligned}$$

This is wonderful! Our "special" lines in x-y space, $y_t = t - x$, are simply horizontal lines of constant t in μ-t space, and the function $f(x(\mu,t), y(\mu,t))$ along these lines achieves maxima in μ independent of t. We have found the natural axes of $f(x,y)$ in μ, t. Let's proceed to transform our original integrals to this new space.

$$\Gamma(a)\Gamma(b) = \int_0^\infty \left[\int_0^\infty e^{-(x+y)} x^{a-1} y^{b-1} dy\right] dx$$

Notice that x is fixed in the inner integral. Let $t = x + y$, so $dt\Big|_x = dy$. Then,

$$\Gamma(a)\Gamma(b) = \int_0^\infty \left[\int_x^\infty e^{-t} x^{a-1}(t-x)^{b-1} dt\right] dx$$

Picture this integral in the positive x-t plane, above the line $t = x$. It might help to refer to Figure 2.2. We can therefore change the order of integration as follows:

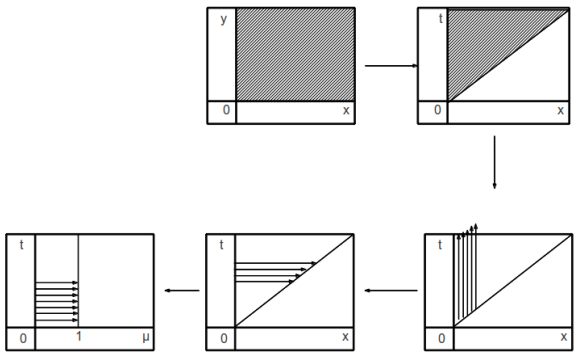


FIGURE 2.2. Visualizing the domains of integration over x-y, x-t, and t-μ space in Claim 2

$$\Gamma(a)\Gamma(b) = \int_0^\infty \left[\int_0^t e^{-t}x^{a-1}(t-x)^{b-1}dx\right] dt$$

Now, let $x = t\mu$. When $t = 0$, the inner integral is over $\int_0^0 dx$ and vanishes. If $t > 0$, $\mu = \frac{x}{t}$ is well defined for all x. Now, with t fixed in the inner integral, $dx\Big|_t = td\mu$. Refer once again to Figure 2.2 to visualize the transformation to t-μ space.

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \left[\int_0^t e^{-t}x^{a-1}(t-x)^{b-1}dx\right] dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}(t\mu)^{a-1}(t-t\mu)^{b-1}(td\mu)\right] dt \\
&= \left[\int_0^\infty e^{-t}t^{a+b-1}dt\right]\left[\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu\right] \\
&= \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu
\end{aligned}$$

or, reshuffled,

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \int_0^1 Be(\mu|a,b)d\mu = 1$$

Thus, the Beta distribution is normalized over $\mu \in [0,1]$. □

Remark. We placed no restrictions on a and b, so in principle the result above could be true for $a, b \in \mathbb{R}$ (or even $\mathbb{C}$?). However, we have not yet shown that the Gamma function is well defined for such inputs.

3. Combinatorial Normalization of the Beta Distribution

Claim. The Beta distribution $Be(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, for $a, b \geq 1$ and $a, b \in \mathbb{N}$, is normalized on $\mu \in [0,1]$. That is, $\int_0^1 Be(\mu|a,b)d\mu = 1$.

Proof. If we restrict ourselves to $a, b \in \mathbb{N}$, we can prove the normalization of the Beta distribution in an alternative way, following the clever combinatorial approach of Problem 3.30 fromBertsekas and Tsitsiklis, 2008.

Suppose we take $Y_1 ... Y_\alpha, Y, Y_{\alpha+1} ... Y_{\alpha+\beta}$ as i.i.d. uniform random variables on $[0,1]$. Consider the event

$$A = \{Y_1 \leq ... \leq Y_\alpha \leq Y \leq Y_{\alpha+1} \leq ... \leq Y_{\alpha+\beta}\}$$

Because our $Y_i \sim U(0,1)$, their probability density functions are simply $p(Y_i) = 1$ over $[0,1]$. The probability of the event A can be brute-force integrated, as follows, but let's try and be clever.

$$\begin{aligned}
P(A) &= \int_0^1 dY_1 p(Y_1)\int_{Y_1}^1 dY_2 p(Y_2) ... \int_{Y_\alpha}^1 dY p(Y) ... \int_{Y_{i-1}}^1 dY_i p(Y_i) ... \int_{Y_{\alpha+\beta-1}}^1 dY_{\alpha+\beta} p(Y_{\alpha+\beta}) \\
&= \int_0^1 dY_1 \int_{Y_1}^1 dY_2 ... \int_{Y_\alpha}^1 dY p(Y) ... \int_{Y_{i-1}}^1 dY_i ... \int_{Y_{\alpha+\beta-1}}^1 dY_{\alpha+\beta}
\end{aligned}$$

Try and picture that domain of integration for a simpler case, with only Y_1, Y, and Y_2. See Figure 3.1. In this figure, we cut the unit cube along the planes

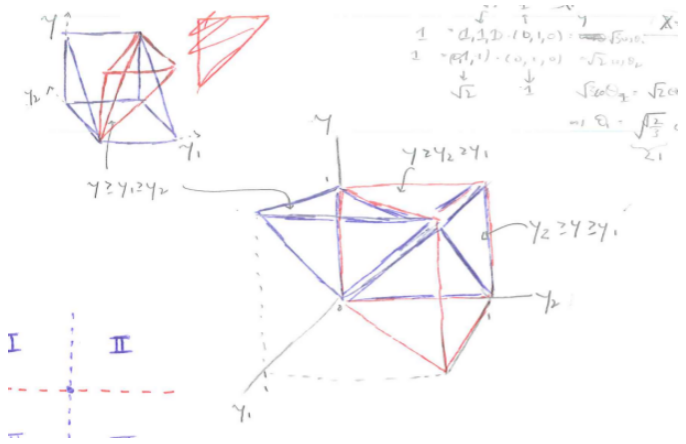


FIGURE 3.1. Visualizing all possible orderings of $\{Y_1, Y, Y_2\}$, which lie in a unit cube in $\mathbb{R}^3$. We can cut the cube into 6 equal volumes that each specify a different ordering of the three Y_i.

$Y \geq Y_1$, $Y \geq Y_2$, and $Y_2 \geq Y_1$, and find that this cutting procedure yields us 6 identical tetrahedrons, each of which encloses points that, except at equality boundaries, satisfy one and only one of the mutually exclusive orderings, $A_i \in \{\{Y_1 \leq Y \leq Y_2\}, \{Y_2 \leq Y_1 \leq Y\}, ...\}$.

That there are only 6 volumes (instead of $2^3 = 8$, for 3 cuts) might not be obvious. The three cuts we made were not independent. Consider the volume specified by knowing $Y > Y_1$ and $Y_2 > Y$. We therefore already know that $Y_2 > Y_1$, so the cut $Y_2 \geq Y_1$ does not do anything. The first two cuts yielded a set of points that already satisfied the third, so the volume after the third cut must be the same as the volume after the first two. On the other hand, cutting $Y > Y_1$ and $Y > Y_2$ does not tell us whether $Y_2 > Y_1$ so we need the third cut to fully specify the ordering of the three Y_i.

That these 6 volumes are equal is even less obvious. Figure 3.1 helps us visualize why, geometrically. Abstractly, consider the two orderings $A = \{Y_1 \leq Y \leq Y_2\}$ and $A' = (Y_2 \leq Y \leq Y_1)$. Let the set of all points (Y_1, Y, Y_2) satisfying A be called $\mathbb{A}$. This $\mathbb{A}$ must be associated with the volume of the tetrahedron corresponding to the ordering A. If we re-label Y_1 and Y_2 in the set $\mathbb{A}$, to make $\mathbb{A}' = (Y_2, Y, Y_1)$, the points in $\mathbb{A}'$ would satisfy A'. Moreover, with this re-labeling, $\mathbb{A}'$ must contain every possible point satisfying A', because if there were some other point such that $(Y_2 \leq Y \leq Y_1)$, we could just re-label again, so that the point satisfied A, and ask why this point was not in $\mathbb{A}$ to begin with.

Now, for any ordering A_i, if you swap any two elements of a point (Y_1, Y, Y_2), the point no longer satisfies A_i. This assumes that we are ignoring the equality planes $Y_i = Y_j$. With this caveat, when we do the re-labeling above, the points in $\mathbb{A}$ have no overlap with the points in $\mathbb{A}'$. That is, we have generated an entirely disjoint pair of equally sized sets, $\mathbb{A}$ and $\mathbb{A}'$, whose points can be associated with the volumes corresponding to A and A'.

Because our sets can be associated with the points in a volume of $\mathbb{R}^3$, they must be uncountable. The probability associated with the two sets must be identical, because we have Y_i uniform and i.i.d. over the unit cube, and we know we're talking

about the same "volume" in some sense. However, something is wrong. The set of points in a line has the same cardinality as the set of points in a cube (e.g. space filling curves). In our unit cube, we know that any tiny volume has some finite probability, but we wouldn't say that a line has any probability at all. We also wouldn't say a plane had any probability. This is why we ignored the equality planes, above. Also, the points in a tiny cube can be mapped to the points in the entire unit cube. So we were wrong; we cannot assign probability to points in a continuum.

Our logic still has a ring of truth to it. We can re-label the axes in Figure 3.1 and see that the tetrahedrons for all orderings are actually the same. So we know that all orderings have the same probability, but we can't yet prove it. Because every point in the unit cube satisfies one of the orderings, and the orderings are effectively disjoint, the following statement should hold:

$$\sum_{i=1}^{6} p(A_i) = 1$$

Because the volumes associated with each ordering are equal, $\forall i, p(A_i) = \frac{1}{6} = \frac{1}{3!}$, where 3! is exactly the number of ways we can order our three variables. Looking back at

$$A = \{Y_1 \leq ... \leq Y_\alpha \leq Y \leq Y_{\alpha+1} \leq ... \leq Y_{\alpha+\beta}\}$$

we can now say that the probability $P(A) = \frac{1}{(\alpha+\beta+1)!}$.

We define $B = \{\max\{Y_1, ..., Y_\alpha\} \leq Y\}$ and $C = \{\min\{Y_{\alpha+1}, ..., Y_{\alpha+\beta}\} \geq Y\}$. These statements are independent because the Y_i are independent. We can write the probability of these statements being true over our unit cube as follows:

$$\begin{aligned} P(B \cap C) &= \int_0^1 P(B \cap C | Y = y) P(Y = y) dy \\ &= \int_0^1 P(B | Y = y) P(C | Y = y) dy \\ &= \int_0^1 y^\alpha (1-y)^\beta dy \end{aligned}$$

which is very promising; that's the Beta distribution staring us in the face. We can also say that

$$P(A | B \cap C) = \frac{1}{\alpha! \beta!}$$

because there are, independently, $\alpha!$ ways of ordering $\{Y_1, ..., Y_\alpha\}$ and $\beta!$ ways of ordering $\{Y_{\alpha+1}, ..., Y_{\alpha+\beta}\}$, and therefore $\alpha!\beta!$ ways of ordering the two sets together. Only one of these orderings is specified in A, but we know that all are equally likely. Now,

$$P(A | B \cap C) P(B \cap C) = P(A \cap B \cap C) = P(A)$$

so we have found that

$$\frac{1}{\alpha! \beta!} \int_0^1 y^\alpha (1-y)^\beta dy = \frac{1}{(\alpha + \beta + 1)!}$$

or, equivalently, that

$$\frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = 1$$

which is exactly what we wanted to prove. □

4. Interpretation of the Beta Distribution

These two approaches to normalizing the Beta distribution were very different, and understanding the two took a lot of work. Yet, at the end of it all, have we learned anything about the Beta distribution itself? What *is* the Beta distribution?

Suppose we have a black box that emits particles of type A and B. We will model this process as a binomial process. We will call this model M. That is, M is the statement that we assume only A and B particles can be emitted, and they do so randomly, with probability $P(A) = \theta$ and $P(B) = 1 - \theta$, respectively. Unfortunately, we don't know θ. Let's say you don't know anything about θ initially, so you believe $P(\theta) = 1$ over $\theta \in [0, 1]$. Then, you see that out of $N = 50$ emissions, N_A are of type A. What can you now say about $P(\theta)$? Well,

$$P(N_A|N, M) = \int_0^1 P(N_A|N, M, \theta)P(\theta)d\theta = \int_0^1 \theta^{N_A}(1-\theta)^{N-N_A}d\theta = \frac{N_A!(N-N_A)!}{(N+1)!}$$

which we have recognized as the normalization integral for a Beta distribution, and

$$P(\theta|N_A, N, M) = \frac{P(N_A|\theta, N, M)P(\theta)}{P(N_A|N, M)}$$

so our belief about θ after we observe the N emissions takes the form of a normalized Beta distribution.

$$P(\theta|N_A, N, M) = \frac{(N+1)!}{N_A!(N-N_A)!}\theta^{N_A}(1-\theta)^{N-N_A}$$

4.1. **Perspective.** We feel like we understand what the Beta distribution means, in the context of belief about coin bias, and so on. I have as yet been unable to intuit the normalization of the distribution if I restrict myself to this perspective. If we look at the problem from another angle, we can use the combinatoric approach to get the normalization factor easily. Are the two perspectives related?

In some sense, yes. We are trying to evaluate

$$\int_0^1 y^a(1-y)^b dy$$

and we can either view y as the probability of "heads" or as the probability that an i.i.d. $U(0, 1)$ random variable falls between $\{0, y\}$. Thinking like a Frequentist (for shame!), in the first case, our experiments yield discrete results, heads or tails. In the second, our experiments yield a continuum of results, but we lump our observations into two bins, $\{0, y\}$ and $\{y, 1\}$. The difference is, in the second case, our sample space is now an $a + b + 1$ dimensional cube, whereas in the first we're stuck in a discrete heads-tails domain with no visual, geometric intuition.

5. The Multinomial & Dirichlet Distributions

Suppose we have a K-sided die, and we throw it N_K times, independently. The distribution of die-faces over the throws follows a multinomial distribution; this is intuitive if you are familiar with the binomial distribution.

We would like to update our belief about the probability that it falls on each of the K sides, $\underline{\mathrm{x}} = [x_1 ... x_K]^T$, where $\sum_{i=1}^K x_i = 1$ and $x_i \in [0,1]$, after seeing the N_K die thows. These throws yield $\{n_i\}$, with $\sum_{i=1}^K n_i = N_K$. We re-define our data N_K as $\{\alpha_i = n_i + 1\}$ for later convenience. This distribution, of our posterior belief of the die weighting, is one interpretation of the Dirichlet distribution.

Now, we can rewrite the normalization constraint on $\underline{\mathrm{x}}$ as $x_K = 1 - \sum_{i=1}^{K-1} x_i$, which means the possible values for $\underline{\mathrm{x}}$ lie on the $K-1$ dimensional simplex, $\mathbb{K}$, which is just the equivalent of a 2D (flat?) surface in 3D space. That is, in the simple example where $K = 3$, $x_3 = 1 - x_1 - x_2$, with all $x_i \in [0,1]$, which is just a section of a plane. Furthermore, we can write

$$P(\underline{\mathrm{x}}|N_K) = \frac{P(N_K|\underline{\mathrm{x}})P(\underline{\mathrm{x}})}{\int_{\mathbb{K}} P(N_K|\underline{\mathrm{x}})P(\underline{\mathrm{x}})d\underline{\mathrm{x}}}$$

and if we take our prior of $\underline{\mathrm{x}}$ over $\mathbb{K}$ to be uniform, we find

$$P(\underline{\mathrm{x}}|N_K) = \frac{P(N_K|\underline{\mathrm{x}})}{\int_{\mathbb{K}} P(N_K|\underline{\mathrm{x}})d\underline{\mathrm{x}}}$$

So, we just need to evaluate this integral, $\int_{\mathbb{K}} P(N_K|\underline{\mathrm{x}})d\underline{\mathrm{x}}$. It is very helpful to visualize this for the case $K = 3$.

We shall do this for all K by induction on $K \geq 2$. Now, when $K = 2$, $\mathbb{K}$ is just a line, $x_1 \in [0,1]$. Then, using $x_2 = 1 - x_1$, we find

$$\int_{\mathbb{K}} P(N_2|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \int_0^1 dx_1 x_1^{n_1} x_2^{n_2} = \int_0^1 dx_1 x_1^{\alpha_1 - 1}(1 - x_1)^{\alpha_2 - 1}$$

which we recognize as a Beta distribution. Thus, when $K = 2$, the integral equals

$$\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\sum_{i=1}^K \alpha_i)} = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$$

Note that this is true regardless of the α_i as long as $\alpha_i \geq 1$, so that our proof for the Beta distribution still holds.

We assume as our inductive hypothesis that

$$\int_{\mathbb{K}} P(N_K|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$$

holds for $K = 1 ... k-1$, for any $\alpha_i \geq 1$.

We then consider the case $K = k$. Thus, $\mathbb{K}$ is of dimension $k-1$, so we can only integrate over $k-1$ components.

$$\int_{\mathbb{K}} P(N_k|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \int_0^1 dx_1 \int_0^{1-x_1} dx_2 ... \int_0^{1-\sum_{i=1}^{k-3} x_i} dx_{k-2} \int_0^{1-\sum_{i=1}^{k-2} x_i} dx_{k-1} \left[\prod_{i=1}^{k-2} x_i^{\alpha_i - 1}\right] x_{k-1}^{\alpha_{k-1}-1}(x_k = 1 - \sum_{i=1}^{k-1} x_i)^{\alpha_k - 1}$$

In the last integral, we are integrating over x_{k-1} while holding $x_{1\ldots k-2}$ fixed. Let us define

$$\mu\Big|_{x_{1\ldots k-2}} = \frac{x_{k-1}}{1-\sum_{i=1}^{k-2} x_i}$$

so that we have

$$d\mu\Big|_{x_{i=1..k-2}} = \frac{dx_{k-1}}{1-\sum_{i=1}^{k-2} x_i}$$

Then, we can write

$$\int_{\mathbb{K}} P(N_k|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \int_0^1 dx_1 \int_0^{1-x_1} dx_2 ... \int_0^{1-\sum_{i=1}^{k-3} x_i} dx_{k-2} \int_0^1 d\mu (1-\sum_{i=1}^{k-2} x_i)$$
$$\left[\prod_{i=1}^{k-2} x_i^{\alpha_i - 1}\right] \left(\mu(1-\sum_{i=1}^{k-2} x_i)\right)^{\alpha_{k-1}-1} (1-\sum_{i=1}^{k-2} x_i - \mu(1-\sum_{i=1}^{k-2} x_i))^{\alpha_k - 1}$$

which can be simplified as follows:

$$\int_{\mathbb{K}} P(N_k|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \int_0^1 dx_1 \int_0^{1-x_1} dx_2 ... \int_0^{1-\sum_{i=1}^{k-3} x_i} dx_{k-2} \left[\prod_{i=1}^{k-2} x_i^{\alpha_i - 1}\right] (1-\sum_{i=1}^{k-2} x_i)^{\alpha_{k-1}+\alpha_k - 1}$$
$$\int_0^1 d\mu \mu^{\alpha_{k-1}-1}(1-\mu)^{\alpha_k - 1}$$

We see that the integral over μ is simply a Beta distribution with parameters α_{k-1} and α_k, and that the integrals over $x_{i=1\ldots k-2}$ are simply over a simplex of dimension $k-2$, with variables $x_{i=1..k-2}$ and $x'_{k-1} = 1-\sum_{i=1}^{k-2} x_i$, where we have "observed" x'_{k-1} effectively $\alpha_{k-1}+\alpha_k$ times. Thus, our inductive hypothesis tells us that

$$\int_{\mathbb{K}} P(N_k|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \frac{\left[\prod_{i=1}^{k-2}\Gamma(\alpha_i)\right]\Gamma(\alpha_{k-1}+\alpha_k)}{\Gamma(\sum_{i=1}^{k-2}\alpha_i + \alpha_{k-1}+\alpha_k)} \frac{\Gamma(\alpha_{k-1})\Gamma(\alpha_k)}{\Gamma(\alpha_{k-1}+\alpha_k)} = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$$

and the inductive proof is complete. Therefore, for all $K \geq 2$, and all observed $N_K = \{\alpha_i \geq 1\}$,

$$\int_{\mathbb{K}} P(N_K|\underline{\mathrm{x}})d\underline{\mathrm{x}} = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}$$

We can now write the normalized Dirichlet distribution as

$$P(\underline{\mathrm{x}}|N_K) \sim Dir(\underline{\mathrm{x}}|\alpha_{i=1..K}, \alpha_0 = \sum_{i=1}^{K}\alpha_i) = \Gamma(\alpha_0)\prod_{i=1}^{K}\left[\frac{x_i^{\alpha_i-1}}{\Gamma(\alpha_i)}\right] \tag{5.1}$$

It is not difficult to show how to update this distribution with new data. The updated posterior will remain $\sim Dir$. It is for this reason that the Dirichlet distribution is termed the conjugate prior of the multinomial distribution; as a special case, we have that the Beta distribution is the conjugate prior for the binomial distribution.