An introduction to the Beta-Binomial model

COMPSCI 3016: Computational Cognitive Science
Dan Navarro & Amy Perfors
University of Adelaide

Abstract

This note provides a detailed discussion of the "beta-binomial model" described in lecture 4, and pops up in several places in the lecture series. The goal is to (a) give you a more comprehensive resource to work from, and (b) to highlight some of the interesting technical issues that will arise when you try to build models of your own. Additionally, the notes aim to be self-contained – very little knowledge of statistics will be assumed. If anything the notes doesn't make sense please contact me (these notes were written by Dan: daniel.navarro@adelaide.edu.au) and I'll try to fix them!

Introduction

The beta-binomial model (along with the closely-related beta-Bernoulli model) is probably the simplest interesting Bayesian model. It is tractable, useful for simple situations (e.g., the coins problem from lecture 4), and is easy to extend to more interesting examples (e.g., the AI survey problem from lectures 11-13). As such, it forms one of the basic building blocks for Bayesian modelling in cognitive science as well as other disciplines. This note describes the model in some detail. Much of this material should be very familiar to you already, since it is in large part a rehash of basic statistics. However, because of the fact that it's such a basic model, it's not unusual to gloss over some of the details when talking about the model. This can be quite confusing to people if they happen to be missing one or two of the key facts. Given that the CCS class is open to students from quite diverse backgrounds (e.g., we have now discovered that some people know beta functions but not beta distributions, and others haven't come across the distinction between probabilities and probability densities before), we think it may be useful to describe the beta-binomial model from the beginning.

To introduce the basic model, consider the following situation, which is taken from Mark Schervish's excellent 1995 book *The Theory of Statistics*. Suppose we are flipping old-fashioned thumbtacks, like the ones shown in Figure 1. When tossed into the air, a thumbtack can either land on its head (outcome = `H`), with the point sticking directly upwards, or it can land on its side (outcome = `S`), resting on an angle with both the point and the head on the ground. When flipped, there is some probability $P(\texttt{H}) = \theta$ that a tack will land on its head, and some probability $P(\texttt{S}) = 1 - \theta$ that it will land on its side. What we would like is some simple statistical tools that we can use in this situation. In what

Figure 1. Nine old-fashioned thumbtacks. Five of them have landed on their heads (`H`), and four of them are on their sides (`S`).

follows, we'll discuss the problem from a Bayesian perspective (rather than an orthodox frequentist one), so we'll discuss everything in terms of priors and likelihoods.

Likelihood #1: The Bernoulli distribution

Suppose that we flip a tack n times, and in each case observe whether the tack lands on its head or on its side. Let x_i denote the outcome of the ith flip, such that

$$x_i = \begin{cases} 1 & \text{if the outcome is } \texttt{H}\text{; i.e., the tack lands on its head} \\ 0 & \text{if the outcome is } \texttt{S}\text{; i.e., the tack lands on its side} \end{cases} \tag{1}$$

Across the set of n flips, we have a length-n binary data vector $\mathbf{x} = (x_1, \ldots, x_n)$, where we may say that $\mathbf{x} \in \mathcal{X}^{(n)}$, where $\mathcal{X}^{(n)}$ is the set of all possible binary vectors of length n. In statistics, this set is called the *sample space* (in this course, we will sometimes also call it the *outcome space* or *data space*). In the Bernoulli distribution, we want to be able to assign probability to the *specific* set of observations $\mathbf{x}$ that we have made, $P(\mathbf{x}|\theta, n)$. If we treat each of the flips as independent events, we can say that:

$$P(\mathbf{x}|\theta, n) = \prod_{i=1}^{n} P(x_i|\theta) \tag{2}$$

Now, from our earlier discussion, we know that

$$P(x_i|\theta) = \begin{cases} \theta & \text{if the outcome is } \texttt{H}\text{; i.e., if } x_i = 1 \\ 1-\theta & \text{if the outcome is } \texttt{S}\text{; i.e., if } x_i = 0 \end{cases} \tag{3}$$

The distribution over the two possible x_i values is the Bernoulli distribution. The notation that we often use to describe this is:

$$x_i|\theta \sim \text{Bernoulli}(\theta) \tag{4}$$

If all n observations are independent and identically distributed (iid) Bernoulli variates, then we can assign probability to the complete data vector $\mathbf{x}$ in the obvious way. If we let k denote the number of times that the tack landed on its head, then

$$P(\mathbf{x}|\theta, n) = \theta^k (1-\theta)^{n-k} \tag{5}$$

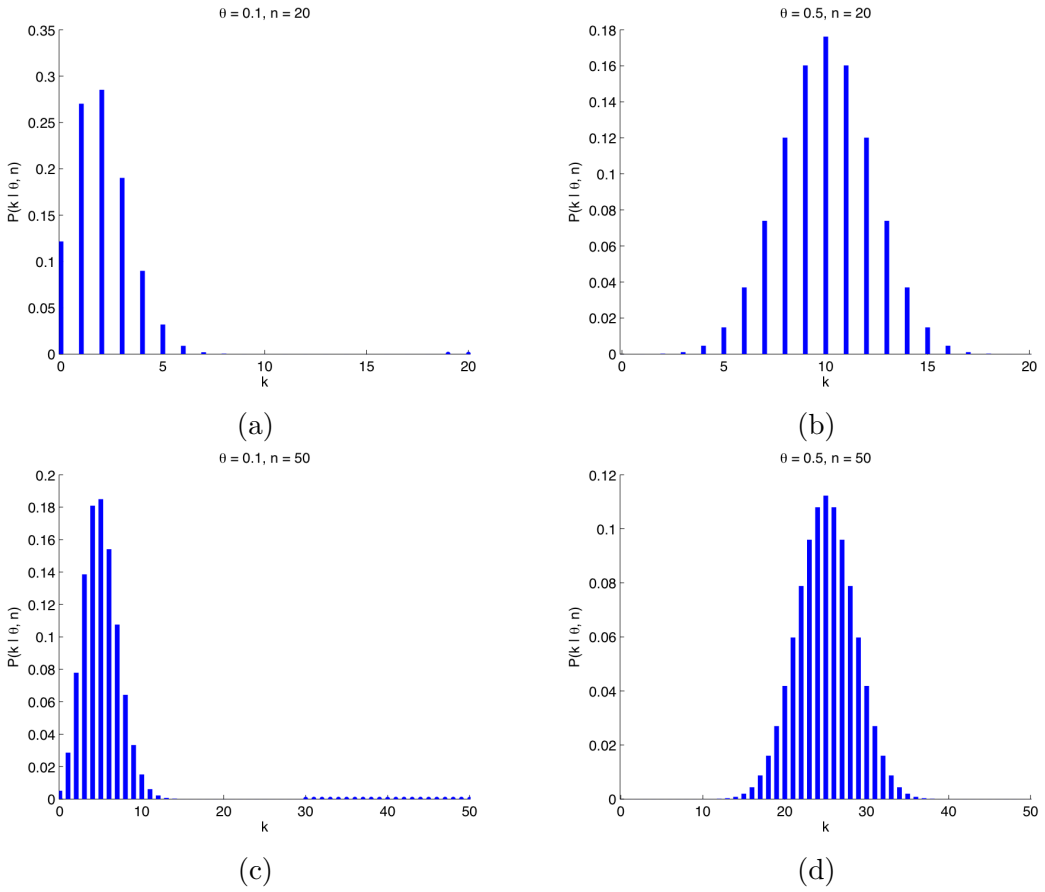

Figure 2. Binomial distributions.

This is the Bernoulli distribution, and it is a proper probability distribution since $P(\mathbf{x}|\theta, n)$ always lies between 0 and 1, and

$$\sum_{\mathbf{x}\in\mathcal{X}^{(n)}} P(\mathbf{x}|\theta, n) = 1 \tag{6}$$

Likelihood #2: The binomial distribution

In the Bernoulli distribution, our concern was with questions like "if I flip a tack 10 times, what is the probability that the result will be `HHSHSSHSHH`?" However, in many situations we might be more interested in questions like "if I flip a tack 10 times, what is the probability that 6 of the outcomes will be heads?" If the question is related to the *number* of heads observed (which is conventionally denoted k), rather than the *specific* set of outcomes $\mathbf{x}$, then we want to use the binomial distribution, not the Bernoulli distribution. The notation here is

$$k|n, \theta \sim \text{Binomial}(\theta, n) \tag{7}$$

Typically, the number of observations n is a fixed characteristic of the data, so in most models only θ is treated as a parameter to be learned. Not surprisingly, the binomial distribution is very similar to the Bernoulli. If we let $\mathcal{X}^{(k,n)}$ denote the set of all binary strings of length n that contain exactly k ones, the binomial likelihood can be derived from the Bernoulli likelihood as follows:

$$P(k|\theta, n) = \sum_{\mathbf{x}\in\mathcal{X}^{(k,n)}} P(\mathbf{x}|\theta, n) \tag{8}$$

$$= \sum_{\mathbf{x}\in\mathcal{X}^{(k,n)}} \theta^k(1-\theta)^{n-k} \tag{9}$$

$$= |\mathcal{X}^{(k,n)}|\ \theta^k(1-\theta)^{n-k} \tag{10}$$

where $|\mathcal{X}^{(k,n)}|$ is a count of the number of unique length-n binary vectors that contain exactly k ones. Elementary combinatorics tells us that:

$$|\mathcal{X}^{(k,n)}| = \tbinom{n}{k} = \frac{n!}{k!(n-k)!} \tag{11}$$

Thus, the binomial likelihood is

$$P(k|\theta, n) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} \tag{12}$$

Four different binomial distributions are plotted in Figure 2.

Prior: The beta distribution

Suppose we have a binomial or Bernoulli likelihood function. What should we use as the prior? A very common choice in this situation is the beta distribution. The beta distribution has two parameters, β_1 and β_2, and is designed to be very similar to Equations 5 and 12. The beta distribution is denoted in the following way,

$$\theta|\beta_1, \beta_2 \sim \text{Beta}(\beta_1, \beta_2) \tag{13}$$

and (if we ignore the constant of proportionality) for the moment,

$$P(\theta|\beta_1, \beta_2) \propto \theta^{\beta_1-1}(1-\theta)^{\beta_2-1}. \tag{14}$$

Now, notice that θ is a probability, so it can vary over the real interval $[0, 1]$. So the beta distribution actually describes a probability density function. Probability densities often confuse students the first time they encounter them. Whenever you have a distribution defined over a continuous sample space, you actually have a density, not a probability. The key thing with densities is that they *integrate* to 1, rather than sum to 1 (which is what probabilities do). On the surface, you'd think that this doesn't introduce any difficulties, but densities can be counterintuitive. The main problem is this: probabilities are required to sum to 1, so it is necessarily the case that the probability of event X lies in the range $[0, 1]$. However, densities don't behave this way: it is perfectly possible for a density to exceed

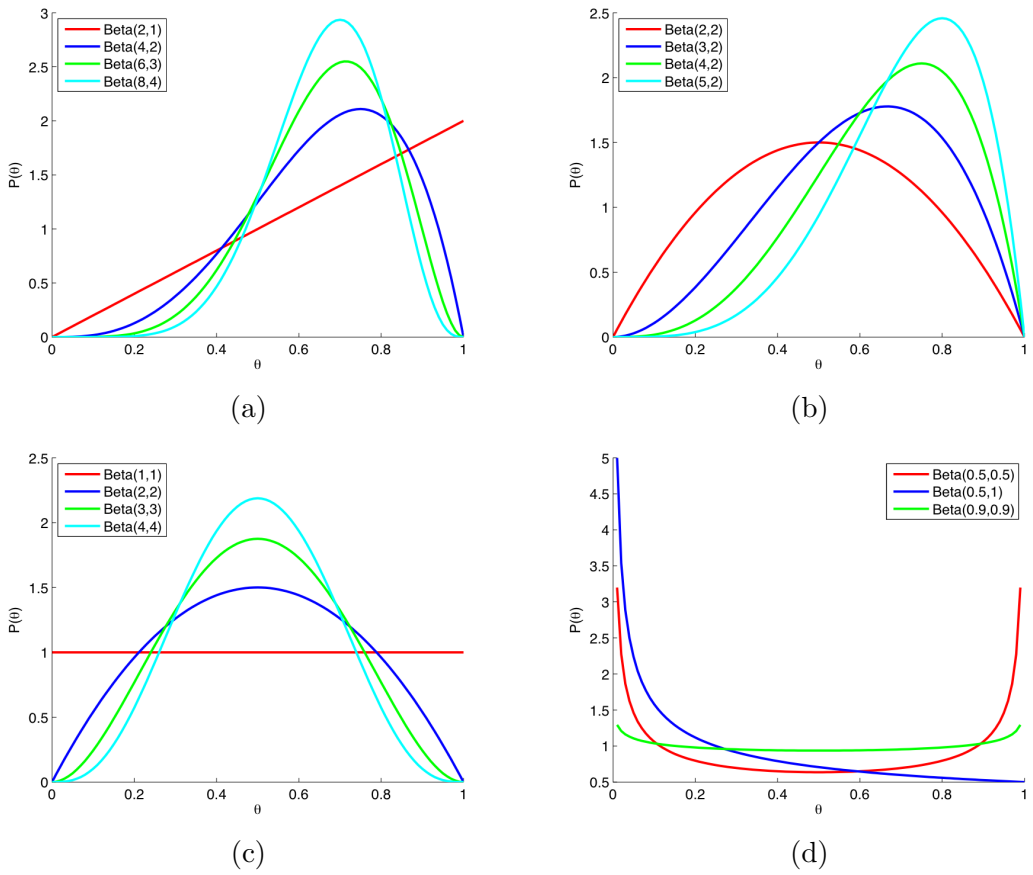

Figure 3. Beta distributions.

1, as long as the the integral over the (continuous) sample space is 1. As an example, try defining a uniform distribution over the range [.3, .4]. Every point in the interval has density 10. This isn't a mathematical statistics course, so we won't go into details: for now, the important thing to recognise is that we can only assign *probabilities* to a countable number of events. If you have continuous spaces, then you have an uncountably infinite set of events to work with, so you can only assign *probability density* to them.

In any case, since densities are required to integrate to 1 we may write:

$$P(\theta|\beta_1, \beta_2) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_2-1}}{\int_0^1 u^{\beta_1-1}(1-u)^{\beta_2-1}\, du} \tag{15}$$

The integral in the denominator is referred to as a beta function, denoted $B(\beta_1, \beta_2)$. Thus

$$P(\theta|\beta_1, \beta_2) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_2-1}}{\int_0^1 u^{\beta_1-1}(1-u)^{\beta_2-1}\, du} \tag{16}$$

$$= \frac{1}{B(\beta_1, \beta_2)}\theta^{\beta_1-1}(1-\theta)^{\beta_2-1} \tag{17}$$

$$= \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1} \tag{18}$$

where the last line exploits a well-known relationship between the beta function and the gamma function (see below), namely that

$$B(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)} \tag{19}$$

Equation 18 is probably the most common description of the beta distribution. Several beta distributions are plotted in Figure 3.

Refresher: The gamma function

Since the gamma function plays a key role in the beta distribution, and crops up in several places in the lectures, it may be useful to quickly remind you of what it is. The gamma function is given by:

$$\Gamma(c) = \int_0^\infty \exp(-u) u^{c-1} du \tag{20}$$

and, if c is an integer, :

$$\Gamma(c) = (c-1)! \tag{21}$$

Two of its key properties are

$$\Gamma(c) = (c-1)\Gamma(c-1) \tag{22}$$

and

$$\Gamma(c)\Gamma(1-c) = \frac{\pi}{\sin(\pi c)} \tag{23}$$

These properties hold even when c is not an integer. In general $\Gamma(c)$ does not have an analytic closed form, but it can be approximated to arbitrary precision by a number of different methods. Matlab code for one of these approximations (the Lanczos approximation) is given in Figure 4.

The conjugacy property: Posteriors are betas

At this stage we are in a position to see why the beta distribution is so frequently used as a prior whenever the data are Bernoulli (or binomial). Going back to our tack-flipping example, let's imagine that we set a Beta prior over θ. We've flipped n tacks, of which k landed on their heads. Our posterior distribution over θ is given by:

$$P(\theta|n, k, \beta_1, \beta_2) = \frac{P(k|n, \theta)P(\theta|n, \beta_1, \beta_2)}{P(k|n, \beta_1, \beta_2)} \tag{24}$$

$$\propto P(k|n, \theta)P(\theta|n, \beta_1, \beta_2) \tag{25}$$

$$= P(k|n, \theta)P(\theta|\beta_1, \beta_2) \tag{26}$$

```
function q = gamma_lanczos(c)

    % reflection formula
    if c < 0.5
        q = pi / ( sin(pi*c) * gmm_big(1-c));

    % normal version
    else
        q = gmm_big(c);
    end

function q = gmm_big(c)

    % coefficients
    g = 7;
    p = [0.99999999999980993, 676.5203681218851, -1259.1392167224028, ...
        771.32342877765313, -176.61502916214059, 12.507343278686905, ...
        -0.13857109526572012, 9.9843695780195716e-6, 1.5056327351493116e-7];

    % calculations
    c = c-1;
    x = p(1);
    for i = 1:g+1
        x = x + p(i+1) / (c+i);
    end
    t = c + g + 0.5;
    q = sqrt(2*pi) * t^(c+0.5) * exp(-t) * x;
```

Figure 4. MATLAB code implementing the Lanczos approximation to the Gamma function.

$$= \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}\theta^{\beta_1-1}(1-\theta)^{\beta_2-1} \tag{27}$$

$$\propto \theta^k(1-\theta)^{n-k} \times \theta^{\beta_1-1}(1-\theta)^{\beta_2-1} \tag{28}$$

$$= \theta^{k+\beta_1-1}(1-\theta)^{n-k+\beta_2-1} \tag{29}$$

This is exactly the same function that we saw in Equation 14. That is, our posterior distribution is also a beta distribution. Notationally, if

$$k|n,\theta \sim \text{Binomial}(\theta, n) \tag{30}$$

$$\theta|\beta_1,\beta_2 \sim \text{Beta}(\beta_1,\beta_2) \tag{31}$$

then

$$\theta|k,n,\beta_1,\beta_2 \sim \text{Beta}(\beta_1+k, \beta_2+n-k) \tag{32}$$

This property is called *conjugacy*, and the complete model is generally called a "beta-binomial model". Specifically, we say that "the beta prior is conjugate to the binomial likelihood". Whenever you have a conjugate prior, the posterior distribution belongs to the same family of distributions as the prior. As a consequence, conjugate priors are extremely useful tools in Bayesian statistics, since they make things a lot more analytically tractable.

Posterior predictions: the beta-binomial distribution

Suppose that you've constructed a beta-binomial model for some data (as per Equations 30 and 31), and you want to make a prediction about what you expect to happen next. The first thing that we might want to know is the *posterior mean*; that is, our best point estimate for θ. This is given by:

$$
\begin{aligned}
E[\theta|k,n,\beta_1,\beta_2] &= \int_0^1 \theta \; P(\theta|k,n,\beta_1,\beta_2) \; d\theta && (33)\\
&= \int_0^1 \theta \; \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \theta^{\beta_1+k-1}(1-\theta)^{\beta_2+n-k-1} \; d\theta && (34)\\
&= \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \int_0^1 \theta^{\beta_1+k}(1-\theta)^{\beta_2+n-k-1} \; d\theta && (35)\\
&= \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \times \frac{\Gamma(\beta_1+k+1)\Gamma(\beta_2+n-k)}{\Gamma(\beta_1+\beta_2+n+1)} && (36)\\
&= \frac{\Gamma(\beta_1+k+1)}{\Gamma(\beta_1+k)} \times \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+\beta_2+n+1)} && (37)\\
&= \frac{\beta_1+k}{\beta_1+\beta_2+n} && (38)
\end{aligned}
$$

What about the more general prediction about future data? That is, imagine that we started with a Beta(β_1, β_2) prior, and then we collected n "tack-flipping" observations, of which k turned out to be heads. Imagine that we then proposed to collect an additional m observations. What is the probability that exactly j of these are heads?

$$
\begin{aligned}
P(j|m,n,k,\beta_1,\beta_2) &= \int_0^1 P(j|m,\theta)P(\theta|n,k,\beta_1,\beta_2) \; d\theta && (39)\\
&= \int_0^1 \frac{m!}{j!(m-j)!}\theta^j(1-\theta)^{m-j} \\
&\qquad \times \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \theta^{\beta_1+k-1}(1-\theta)^{\beta_2+n-k-1} \; d\theta && (40)\\
&= \frac{m!}{j!(m-j)!} \times \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \\
&\qquad \times \int_0^1 \theta^{\beta_1+k+j-1}(1-\theta)^{\beta_2+n-k+m-j-1} \; d\theta && (41)\\
&= \frac{m!}{j!(m-j)!} \times \frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+k)\Gamma(\beta_2+n-k)} \\
&\qquad \times \frac{\Gamma(\beta_1+k+j)\Gamma(\beta_2+n+m-k-j)}{\Gamma(\beta_1+\beta_2+n+m)} && (42)
\end{aligned}
$$

The general name for the expected distribution over future observations is the *posterior predictive distribution*. The name for the probability distribution in Equation 42 is a "beta-binomial distribution". That is, if the prior distribution is beta and the likelihood is binomial, then the posterior predictive distribution is a beta-binomial distribution. Incidentally,

it is this beta-binomial distribution that is referred to on slide 148 of lectures 11-13 (i.e., the particle filtering demo using the AI survey data). Note that when implementing the beta-binomial distribution you don't need to calculate all of these gamma functions, because you can use the $\Gamma(c) = (c-1)\Gamma(c-1)$ property to simplify things a lot! To see this note that:

$$\begin{aligned} P(j|m,n,k,\beta_1,\beta_2) &= \frac{m!}{j!(m-j)!}\frac{\Gamma(\beta_1+\beta_2+n)}{\Gamma(\beta_1+\beta_2+n+m)} \\ &\quad \times \frac{\Gamma(\beta_1+k+j)}{\Gamma(\beta_1+k)}\frac{\Gamma(\beta_2+n+m-k-j)}{\Gamma(\beta_2+n-k)} \qquad (43) \\ &= \frac{m!}{j!(m-j)!}\frac{\left(\prod_{v=\beta_1+k}^{\beta_1+k+j-1} v\right)\left(\prod_{v=\beta_2+n-k}^{\beta_2+n+m-k-j-1} v\right)}{\left(\prod_{v=\beta_1+\beta_2+n}^{\beta_1+\beta_2+n+m-1} v\right)} \qquad (44) \end{aligned}$$

which is not a very elegant equation, but is very simple to write code for! Notice that in the special case where $m = 1$ and $j = 1$ we obtain:

$$P(j|m,n,k,\beta_1,\beta_2) = \frac{1!}{1!0!}\frac{(\beta_1+k)(0)}{(\beta_1+\beta_2+n)} = E[\theta|n,k,\beta_1,\beta_2] \qquad (45)$$