Basketball, Beta, and Bayes

MATTHEW RICHEY
PAUL ZORN
St. Olaf College
Northfield, Minnesota 55057
richeym@stolaf.edu
zorn@stolaf.edu

Problem B1 on the 2002 Putnam competition [**1**] reads as follows:

> PUTNAM PROBLEM (PP): Shanille O'Keal shoots free throws on a basketball court. She hits the first, misses the second, and thereafter the probability that she hits the next shot is equal to the proportion of the shots she has hit so far. What is the probability that she hits exactly 50 out of her first 100 shots?

The answer is $1/99$, as the reader may enjoy showing. Indeed, the probability that Shanille hits k of her next 98 shots is $1/99$ for *all* k with $0 \le k \le 98$; we will prove this later as a corollary to a more general result.

The fact that the number of later hits is uniformly distributed on $\{0, 1, \dots, 98\}$ may seem unexpected, but it reflects the fact that the starting conditions—one hit and one miss—convey very little information. In Bayesian jargon, as we will see, these starting conditions would be called *noninformative*, not because they convey no information whatever but because they lead to a uniform distribution on the outcome space.

Let's generalize the problem slightly. Suppose Shanille begins with a hits and b misses, and then takes n additional shots. We will consider two associated random variables:

$$S_n := \text{the number of successes among } n \text{ attempts};$$

$$\theta_n := \frac{S_n + a}{a + b + n}.$$

We can think of θ_n as an "updated belief" in Shanille's shooting ability based on all accumulated data. In the Putnam problem (PP) we have $a = b = 1$ and we seek $\mathrm{P}(S_{98} = 49)$, the probability of exactly 49 hits among the next 98 shots. (We will write $\mathrm{P}(X = k)$ for the probability that the discrete random variable X has value k.)

> GENERALIZED PUTNAM PROBLEM (GPP): Shanille shoots free throws. To begin, she hits a and misses b shots; thereafter, she hits with probability equal to the proportion of hits so far. Determine the probability distribution of S_n. Equivalently, describe the distribution of θ_n.

As noted above, S_n is uniformly distributed on $\{0, 1, \dots, n\}$ for $a = b = 1$. We will soon see that S_n is *not* uniformly distributed for any other choice of a and b.

The GPP is a *probability* problem—at each stage the success probability θ is fixed—but its ingredients suggest the *Bayesian* approach to *statistics*:

- an initial belief (a *prior distribution*, in Bayesian jargon) about Shanille's shooting accuracy: $\theta = a/(a + b) = \text{hits}/(\text{hits} + \text{misses})$;
- data: the outcome of one or more shots;
- an updated (*posterior*) belief, based on the data, about Shanille's accuracy.

After solving the GPP we will propose and solve a Bayesian variant of the basketball problem, in which the GPP can be seen as "embedded." En route, we introduce Bayesian statistics in general and the *beta–binomial* model in particular. The Bayesian perspective can help explain some surprises in the GPP's solution.

Bayesian statistics: a primer

To put what follows in context, we note some differences between probability and statistics. Both disciplines deal with *parameters* (such as θ, the success probability in some experiment) and *data* or *random variables* (such as S_n in the GPP). A probabilist usually takes parameters as known, and studies properties of the data or random variables, such as their distribution and expected values. Statisticians study the inverse problem: Beginning from data, they try to describe parameters.

The difference between the Bayesian statistician and the (more common) *Frequentist* statistician centers mainly on how each views the roles of parameters and data. A Frequentist views parameters as *fixed but unknown* quantities, and the data as *random.* Inferences, such as those concerning confidence intervals for parameters, are obtained through thought experiments starting with "Imagine all possible data produced by the parameter" or "If we sampled arbitrarily often ...."

Consider, for instance, a Frequentist method, the one-sample t-test for estimating a population mean θ. Suppose that $X_1, X_2, \ldots, X_n$ is an independent, identically distributed sequence of random variables and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The distribution of the X_i may not be known exactly, but some assumptions are made, such as that the distribution is approximately normal and that the X_i have finite mean and variance.

Given specific data $x_1, x_2, \ldots, x_n$ and their sample mean $\bar{x}$, one obtains a specific 95% confidence interval $(\bar{x} - m, \bar{x} + m)$, with $\bar{x}$ a point estimate for θ. (The value m is the margin of error and is determined from the variability observed in the data.) Notice that, from the Frequentist perspective, 95% is *not* the probability that θ lies in the calculated interval. On the contrary, θ is fixed while the confidence interval is random. The 95% probability concerns "coverage": If one samples often, each time calculating a confidence interval, then about 95% of these intervals will contain θ.

A Bayesian, by contrast, sees the data as fixed, but expresses *belief* about a parameter θ as a probability distribution—subject to change as additional data arise. A Bayesian would concede that in some simple cases θ has a "true" value. If, say, θ is a coin's probability of landing "heads," then (according to the law of large numbers) one could approximate θ by flipping the coin many times. But in less repeatable cases, such as whether it will rain tomorrow, a Bayesian would rather model our *belief* (or uncertainty) about this likelihood.

Consider, for instance, how a Bayesian makes an inference about a population parameter, θ, contained in a parameter space Ω. The Bayesian starts with a *prior distribution* $\pi(\theta)$, which expresses an initial belief about the relative likelihood of possible values of θ. The data $\mathbf{X} = X_1, X_2, \ldots, X_n$ and their *joint distribution* $f(\mathbf{X} = \mathbf{x} \mid \theta)$ are known. To obtain the distribution of θ conditioned on $\mathbf{X}$ (called the *posterior distribution* and denoted by $f(\theta \mid \mathbf{X} = \mathbf{x})$), a Bayesian invokes *Bayes's rule*:

$$f(\theta \mid \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{X} = \mathbf{x} \mid \theta)\pi(\theta)}{\int_{\Omega} f(\mathbf{X} = \mathbf{x} \mid \hat{\theta})\pi(\hat{\theta})\, d\hat{\theta}}, \qquad (1)$$

where Ω is the parameter space for θ. (The integral is a sum if θ is a discrete random variable.) In simple cases (1) can be evaluated in closed form; other cases require numerical methods, such as Markov Chain Monte Carlo (MCMC) techniques. A variety of sources discuss Bayesian methods in general [**3, 4, 5, 6**] and MCMC techniques in particular [**2, 4, 5**].

Knowledge of $f(\theta \mid \mathbf{X} = \mathbf{x})$ amounts, from a statistical perspective, to knowing "everything" about θ. For example, a Bayesian might use the expectation $\mathrm{E}(\theta \mid \mathbf{X} = x)$ as a point estimate of θ. A 95% confidence interval for θ could be any interval (θ_1, θ_2) for which

$$\int_{\theta_1}^{\theta_2} f(\theta \mid \mathbf{X} = \mathbf{x})\, d\theta = 0.95. \tag{2}$$

Among all such intervals, one might choose the one symmetric about the expected value $\mathrm{E}(\theta \mid \mathbf{X} = \mathbf{x})$, or, alternatively, the narrowest interval for which (2) holds. In any event, (2) expresses the Bayesian's belief that θ lies in the confidence interval with 95% probability. Thus, Bayesians and Frequentists agree that a 95% confidence interval depends on the data, but a Bayesian expresses the dependence explicitly, using the posterior distribution.

Bayesians v. Frequentists Where do Bayesians and Frequentists disagree? Following is a somewhat caricatured discussion; in practice, many statisticians adopt aspects of both approaches.

Frequentists see the Bayesian notion of a prior distribution as too subjective. Bayesians counter that Frequentists make equally subjective assumptions, such as that a given distribution is normal. Frequentists claim that Bayesians can, by choosing a suitable prior, obtain any desired result. Not so, Bayesians reply: Sufficient data always "overwhelm the prior"; besides, ill-chosen priors are soon revealed as such. Frequentists appreciate the computational tractability of their methods, and see Bayesian posterior distributions as unduly complex. Not any more, say Bayesians—modern computers and algorithms make posterior distributions entirely manageable.

Bayesian inference the beta–binomial way

We illustrate Bayesian inference using the *beta–binomial* model (which we apply later to a Bayesian-flavored basketball problem). Suppose X is a *Bernoulli* random variable with values 0 or 1 and $\mathrm{P}(X = 1) = \theta$; we write this as $X \sim \mathrm{Bernoulli}(\theta)$. For fixed n, let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables and set $Y = X_1 + X_2 + \cdots + X_n$. Then Y is the number of successes (1s) in n Bernoulli trials, a *Binomial* random variable, and we write $Y \sim \mathrm{Binomial}(n, \theta)$.

As a prior distribution on θ Bayesians often choose a $\mathrm{Beta}(a, b)$ distribution on $[0, 1]$. This distribution is defined for arbitrary positive a and b, and has density

$$f_{a,b}(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1};$$

we write $\theta \sim \mathrm{Beta}(a, b)$. (The *gamma function* is defined for $x > 0$ by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt.$$

Among its familiar properties are $\Gamma(x) = (x-1)\Gamma(x-1)$ and $\Gamma(1) = 1$. In particular, $\Gamma(n) = (n-1)!$ for positive integers n. It is an interesting exercise to show directly that $\int_0^1 f_{a,b}(\theta)\, d\theta = 1$.)

FIGURE 1 shows Beta(a, b) densities $f_{a,b}(\theta)$ for several choices of a and b. Notice how the parameters control the shape of the distribution. Straightforward calculations give the expectation and variance for $\theta \sim$ Beta(a, b):

$$\mathrm{E}(\theta) = \frac{a}{a+b}; \qquad \mathrm{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3}$$

These formulas suggest that a and b can be thought of as imagined prior hits and misses among $n = a + b$ attempts. Indeed, suppose that a, b are actual hits and misses from a Binomial$(a + b, \theta)$ distribution. Then the Frequentist's (unbiased) estimate of θ,

$$\hat{\theta} = \frac{a}{a+b},$$

is precisely the Bayesian's expected value of θ. Similarly,

$$\mathrm{Var}\left(\hat{\theta}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{a+b} = \frac{ab}{(a+b)^3},$$

which is $(a + b + 1)/(a + b)$ times the Bayesian's Var(θ).

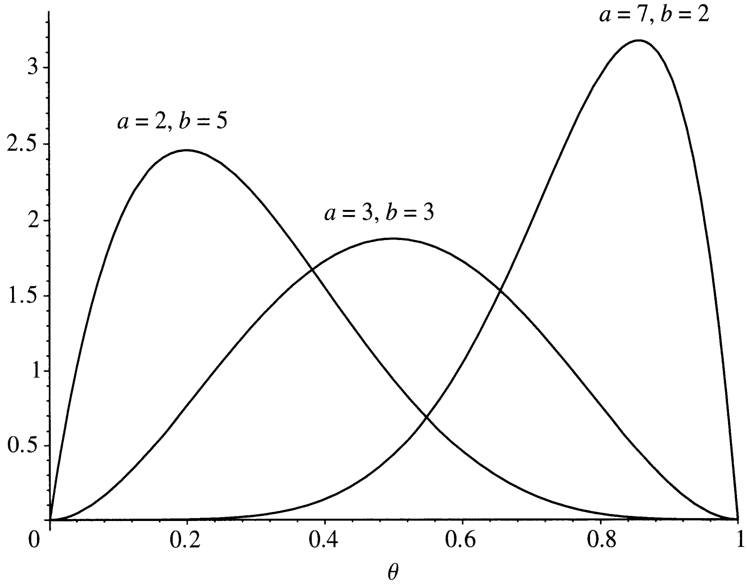

Figure 1 Plots of $f_{a,b}(\theta)$ for different values of a, b

Notice also that one can arrange *any* positive mean and variance by choosing a and b judiciously in (3). For example, setting $a = 7$ and $b = 2$ in (3) reflects Bayesian belief in a mean near 7/9; see FIGURE 1. Setting $a = b = 1$ produces the *uniform* distribution on [0, 1], known here as the *noninformative prior* because it reflects little or no prior belief about θ.

The following well-known result links the beta and the binomial distributions in the Bayesian setting.

PROPOSITION 1. *Suppose that* $\theta \sim$ Beta(a, b) *and* $X \sim$ Binomial(n, θ), *so the prior distribution has density function*

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

and

$$P(X = k \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Then θ has posterior distribution Beta$(a+k, b+n-k)$. *That is,*

$$f(\theta \mid X = k) = \frac{\Gamma(a+b+n)}{\Gamma(a+k)\Gamma(b+n-k)}\,\theta^{a+k-1}(1-\theta)^{b+n-k-1}.$$

Note Proposition 1 and equation (3) imply that the posterior expectation $E(\theta)$ satisfies

$$E(\theta) = \frac{a+k}{a+b+n} = \frac{a+b}{a+b+n}\cdot\frac{a}{a+b} + \frac{n}{a+b+n}\cdot\frac{k}{n}.$$

Thus, $E(\theta)$ is a convex combination of the prior mean $a/(a+b)$ and the sample mean k/n; the Bayesian method favors the prior for small samples but tends toward the Frequentist figure as the sample size increases.

Proof. Bayes's rule (1) gives

$$f(\theta \mid X = k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}\,\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \binom{n}{k}\hat{\theta}^k(1-\hat{\theta})^{n-k}\,\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\hat{\theta}^{a-1}(1-\hat{\theta})^{b-1}\,d\hat{\theta}}.$$

Canceling common factors and collecting powers of θ and $(1-\theta)$ gives

$$f(\theta \mid X = k) = C\,\theta^{a+k-1}(1-\theta)^{b+n-k-1}$$

for some constant C. Since $f(\theta \mid X = k)$ is a probability density, it must be Beta$(a+k, b+n-k)$. ■

The beta-binomial model's main ingredients resemble those of the GPP:

- a prior distribution: $\theta \sim$ Beta(a, b), with $E(\theta) = a/(a+b)$;
- data: $X \sim$ Binomial(n, θ);
- a posterior distribution: $\theta \sim$ Beta$(a+k, b+n-k)$ with

$$E_{\text{post}}(\theta) = \frac{a+k}{a+b+n}.$$

The analogy continues. For the GPP, the initial success probability has the fixed value $\theta_0 = a/(a+b)$; for the beta-binomial, θ_0 is random, but $E(\theta_0) = a/(a+b)$. After n shots with k successes, the GPP gives $\theta_n = (a+k)/(a+b+n)$, which is also the expected value of θ_n after a beta-binomial update. A similar theme will be seen in what follows.

Repeated updates Repeated beta–binomial updates turn out to be equivalent to a single beta–binomial update. More precisely, let n_1 and n_2 be positive integers and a and b fixed positive numbers, with $\theta \sim$ Beta(a, b) and $k_1 \sim$ Binomial(n_1, θ). Updating θ once gives $\theta \sim$ Beta$(a+k_1, b+n_1-k_1)$. If $k_2 \sim$ Binomial(n_2, θ_1), then updating θ again gives

$$\theta \sim \text{Beta}\left(a + (k_1+k_2),\ b + (n_1+n_2) - (k_1+k_2)\right),$$

which shows that θ can be obtained from *one* beta–binomial update with $k_1 + k_2$ successes in $n_1 + n_2$ trials. This result plays well with Bayesian philosophy: The data and the model—not an arbitrary subdivision into two parts—determine our posterior belief.

If $n = 1$, the binomial random variable reduces to the Bernoulli case; the result might be called a *beta-Bernoulli model.* The preceding observation implies that updating a Beta(a, b) prior distribution with a sequence of beta–Bernoulli updates is equivalent to a single beta–binomial update.

Bayesian prediction Our goal so far has been to combine a prior distribution with data to predict future values of a parameter θ. A natural next step is to use our new knowledge of θ to predict future values of the random variable X. If θ were fixed we would obtain the traditional binomial distribution:

$$\mathrm{P}(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Because our θ is random, we average over all values of θ, weighted by its posterior density. Not surprisingly, the randomness of θ leads to greater variability in X. With $\theta \sim$ Beta(a, b), we have

$$\begin{aligned}\mathrm{P}(X = k) &= \int_0^1 \mathrm{P}(X = k \mid \theta) f_{a,b}(\theta)\, d\theta \\ &= \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\, d\theta \\ &= \binom{n}{k}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{a+k-1}(1-\theta)^{b+n-k-1}\, d\theta \\ &= \binom{n}{k}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)}. \qquad (4)\end{aligned}$$

This result is called the *marginal distribution* of X, or, in Bayesian parlance, the *predictive posterior distribution.*

The distribution (4), known as the *diffuse binomial,* is close kin to the ordinary binomial distribution. FIGURE 2 suggests how: Both distributions have the same general

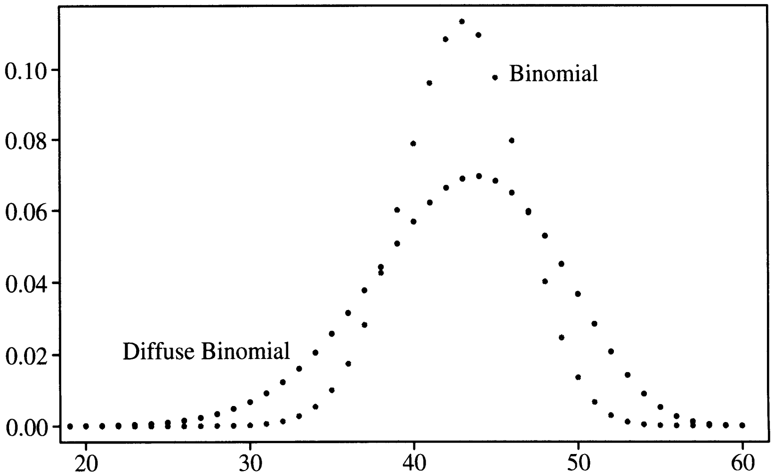

Figure 2 Distributions of $\mathrm{P}(X = k)$ for the diffuse binomial with $n = 60$, $a = 25$, $b = 10$; and for Binomial$(60, \theta)$ with $\theta = 25/35$

shape but the diffuse version has larger "tails." Equation (4) has more than passing interest—it will reappear when we solve the GPP.

Like the updating process, the predictive posterior distribution does not depend on a partitioning of n. To see this, suppose that $\theta_1 \sim \text{Beta}(a, b)$, $X_1 \sim \text{Binomial}(n_1, \theta_1)$, $\theta_2 \sim \text{Beta}(a + k_1, b + n_1 - X_1)$, and $X_2 \sim \text{Binomial}(n_2, \theta_2)$.

If $X = X_1 + X_2, n = n_1 + n_2$, and $k = k_1 + k_2$, then we have

$$\begin{aligned}\text{P}(X = k) &= \sum_{k_1=0}^{n_1} \sum_{k_2=k-k_1}^{n_2} P\,(X_1 = k_1 \text{ and } X_2 = k_2) \\ &= \sum_{k_1=0}^{n_1} \sum_{k_2=k-k_1}^{n_2} P(X_2 = k_2 \mid X_1 = k_1)\,\text{P}(X_1 = k_1).\end{aligned}$$

Substituting (4) in both factors above gives $\text{P}(X = k) =$

$$\begin{aligned}&\sum_{k_1=0}^{n_1} \sum_{k_2=k-k_1}^{n_2} \binom{n_2}{k_2} \frac{\Gamma(a+b+n_1)}{\Gamma(a+k_1)\Gamma(b+n_1-k_1)} \\ &\quad \times \frac{\Gamma(a+k_1+k_2)\Gamma(b+(n_1+n_2)-(k_1-k_2))}{\Gamma(a+b+(n_1+n_2))} \\ &\quad \times \binom{n_1}{k_1} \frac{\Gamma(a+b)}{\Gamma(b)\Gamma(a+b)} \frac{\Gamma(a+k_1)\Gamma(b+n_1-k_1)}{\Gamma(a+b+n_1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)} \sum_{k_1=0}^{n_1} \sum_{k_2=k-k_1}^{n_2} \binom{n_2}{k_2}\binom{n_1}{k_1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)} \binom{n}{k}.\end{aligned}$$

Thus, an imagined partitioning of the attempts does not—and to a Bayesian should not—change the predicted outcome.

No free lunch One might also ask about the marginal posterior distribution of θ, found by averaging over all possible outcomes of X. To a Bayesian this exercise is fruitless: Imagining *all* possible outcomes should not change the initial belief about θ. The following calculation explains this for θ a continuous and X a discrete random variable. (The same result holds for any combination of discrete and continuous random variables.)

Consider a continuous random variable θ, with prior density $\pi(\theta)$ and values in Ω, and suppose that a discrete random variable X, with values x in a set S, has conditional density $\text{P}(X = x \mid \theta)$. Then, for each x, the conditional posterior distribution for θ has density

$$f_{\text{post}}(\theta \mid X = x) = \frac{\text{P}(X = x \mid \theta)\,\pi(\theta)}{\int_\Omega \text{P}(X = x \mid \hat{\theta})\pi(\hat{\theta})\,d\hat{\theta}}.$$

Now we can find the marginal posterior distribution:

$$\begin{aligned}\hat{f}_{\text{post}}(\theta) &= \sum_{x\in S} f_{\text{post}}(\theta \mid X = x)\,\mathrm{P}(X = x)\\ &= \sum_{x\in S} \frac{\mathrm{P}(X = x \mid \theta)\pi(\theta)}{\int_\Omega \mathrm{P}(X = x \mid \hat{\theta})\pi(\hat{\theta})\,d\hat{\theta}}\,\mathrm{P}(X = x)\\ &= \sum_{x\in S} \frac{\mathrm{P}(X = x \mid \theta)\pi(\theta)}{\mathrm{P}(X = x)}\,\mathrm{P}(X = x) = \pi(\theta)\sum_{x\in S} P(X = x \mid \theta) = \pi(\theta).\end{aligned}$$

That the prior and the marginal posterior distributions for θ are identical illustrates a Bayesian "no free lunch" principle: Real data, not thought experiments, are needed to update a prior distribution.

Bayesian basketball Yet again the tireless Shanille shoots free throws and we seek to model our belief about her success probability θ. As cautious Bayesians we choose for θ the "noninformative prior"—the uniform distribution on [0, 1]. Lacking further information, we see *every* subinterval of [0, 1] of width $\Delta\theta$ as equally likely to contain the "true" θ. A more informed Bayesian fan might choose an informative prior, say, Beta(4, 4). FIGURE 3 shows both choices.

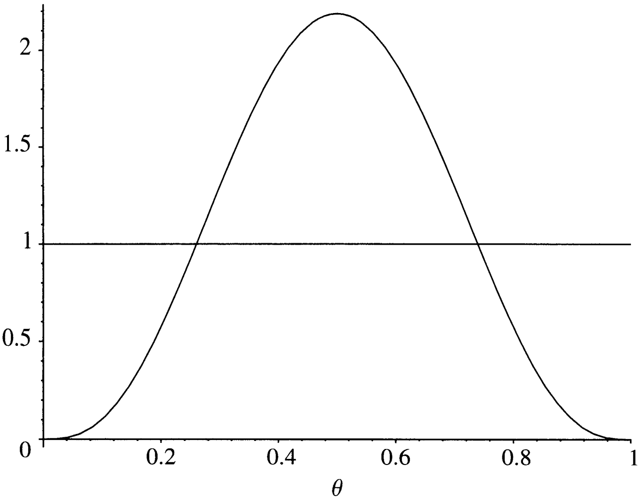

Figure 3 The Beta(1, 1) (uniform, noninformative) and Beta(4, 4) priors for θ, Shanille's success probability

Here, unlike in the GPP, we never know Shanille's exact success probability. (If we knew θ exactly we wouldn't bother to estimate it!) If Shanille shoots n free throws and hits k, then, using the beta–binomial model and invoking Proposition 1, we obtain the following posterior distributions for θ:

$$\text{for prior Beta}(1,1)\text{:}\quad \text{Beta}(1+k, 1+n-k),\ \text{with } \mathrm{E}_{\text{post}}(\theta) = \frac{k+1}{n+2}$$

$$\text{for prior Beta}(4,4)\text{:}\quad \text{Beta}(4+k, 4+n-k),\ \text{with } \mathrm{E}_{\text{post}}(\theta) = \frac{k+4}{n+8}.$$

In each case a 95% confidence interval (u, v) for θ can be obtained by finding (perhaps numerically) any numbers u and v for which

$$\frac{\Gamma(n+a+b)}{\Gamma(k+a)\Gamma(n-k+b)}\int_u^v \theta^{a-1+k}(1-\theta)^{b-1+n-k}\,d\theta = 0.95.$$

The following table compares the Frequentist (in this case binomial) and two different Bayesian point estimates and symmetric 95% confidence intervals for θ using different values for n and k.

TABLE 1: Frequentist and Bayesian estimates for θ

Model	(k, n)	Point estimate	95% confidence	Interval width
Frequentist	$(2, 3)$	$2/3 \approx .667$	$(.125, .982)$	.857
	$(30, 45)$	$30/45 \approx .667$	$(.509, .796)$	.287
	$(60, 90)$	$60/90 \approx .667$	$(.559, .760)$	.201
Bayesian, $a = 1, b = 1$	$(2, 3)$	$3/5 = .600$	$(.235, .964)$	.729
	$(30, 45)$	$31/47 \approx .660$	$(.527, .793)$	.266
	$(60, 90)$	$61/82 \approx .663$	$(.567, .759)$	.191
Bayesian, $a = 4, b = 4$	$(2, 3)$	$6/11 \approx .545$	$(.270, .820)$	.550
	$(30, 45)$	$34/54 \approx .642$	$(.514, .769)$	.254
	$(60, 90)$	$64/98 \approx .653$	$(.560, .747)$	.187

Observe:

- The Frequentist point estimate for θ is always 2/3, the sample proportion of successes. Bayesian point estimates, by contrast, are weighted averages of the sample proportions and the prior mean, $1/2$.
- For fixed k and n, widths of 95% confidence intervals decrease as we move from the Frequentist through the noninformative Bayesian to the informative Bayesian perspective.
- The Frequentist and Bayesian estimates come together as n increases. "Data overwhelm the prior," a Bayesian might say.
- For the informative prior Beta(4, 4), FIGURE 4 shows the densities for θ becoming narrower and moving toward the sample mean as n increases.

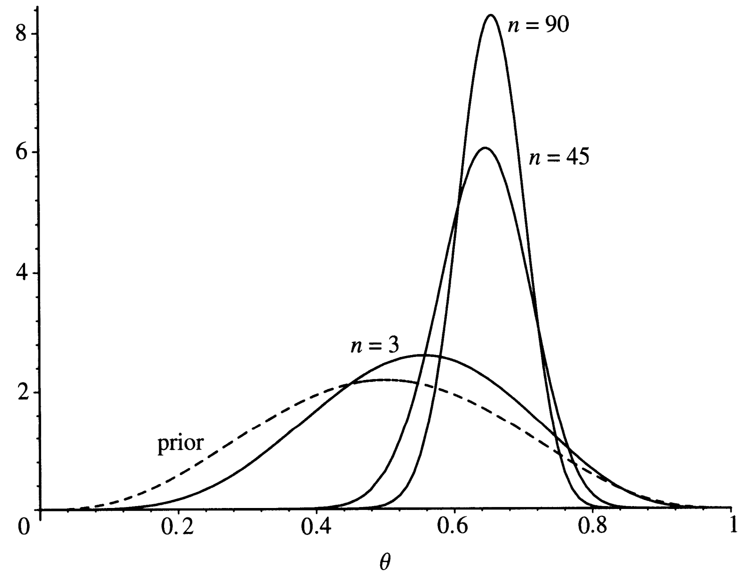

Figure 4 An informative prior (Beta(4, 4), dashed) and several posterior densities for θ

Solving the GPP

Leaving (for now) the realm of statistics, we now return to probability and the GPP. We asserted above without proof that in the original PP (with $a = b = 1$), the random variable S_n is uniformly distributed:

$$\mathrm{P}(S_n = k) = \frac{1}{n+1} \quad \text{for } k = 0, 1, \ldots, n.$$

Instead of proving this directly, we calculate $P(S_n = k)$ for arbitrary positive integers a and b. This takes a little work, but as a small reward we see the beta–binomial predictive posterior distribution (4) crop up again. We see, too, that (5) reduces to the uniform distribution if, but *only* if, $a = b = 1$.

PROPOSITION 2. *With notation as in the GPP, we have*

$$\mathrm{P}(S_n = k) = \binom{n}{k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)}. \tag{5}$$

Proof. There are $\binom{n}{k}$ ways to achieve k successes in n attempts. The key observation is that *all* sequences of k successes and $n - k$ failures are equally probable. To see why, consider any point at which s successes and f failures have occurred. The probability of a success followed by a failure is thus

$$\frac{s}{s+f} \frac{f}{s+f+1},$$

while the probability of a failure followed by a success is

$$\frac{f}{s+f} \frac{s}{s+f+1}.$$

Because these two quantities are equal we can rearrange *any* sequence of successes and failures so that (say) all k successes come first. The probability of this special arrangement is

$$\begin{aligned} &\frac{a(a+1)\cdots(a+k-1)\, b(b+1)\cdots(b+n-k-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)}; \end{aligned}$$

where the equality follows by rewriting the right side in terms of factorials. Finally, the probability of any k successes among n attempts is $\binom{n}{k}$ times the preceding quantity. ■

Back to the Beta Now we can start to explain why formulas (4) and (5) are identical. FIGURE 5 shows probability distributions of θ_{100} (Shanille's success rate on her 100th shot) for several choices of a and b.

The plots in FIGURE 5 closely resemble the corresponding beta density functions $f_{a,b}$ in FIGURE 1; they differ mainly by a vertical scale factor of $a + b + n$. We can describe this graphical similarity in probabilistic language. Let I be any interval contained in $[0, 1]$ and $\theta \sim \mathrm{Beta}(a, b)$. We will show that

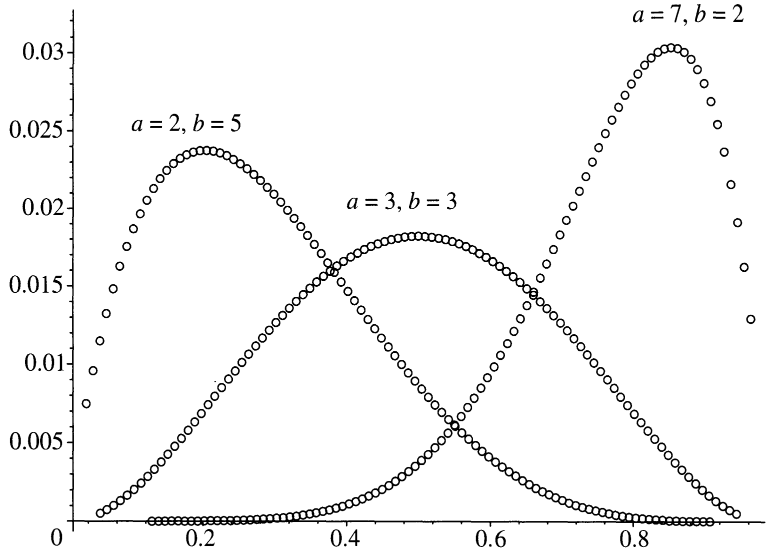

Figure 5 Density for plots of θ_{100} for different values of a and b; compare to FIGURE 1

$$\left|\mathrm{P}(\theta_n \in I) - \mathrm{P}(\theta \in I)\right| = O(1/n), \tag{6}$$

which means that if we start with a hits and b misses, then for large n the *discrete* probability distribution for θ_n is approximated by the *continuous* Beta(a, b) distribution.

To prove equation (6) is a routine but slightly messy exercise in O-arithmetic. First, let

$$t_{n,k} = \frac{a+k}{a+b+n} \quad \text{and} \quad P_{n,k} = \mathrm{P}(\theta_n = t_{n,k})$$

for $k = 0, 1, 2, \ldots, n$. Equation (6) will follow if we show that

$$f_{a,b}(t_{n,k}) = (a+b+n)\, P_{n,k} + O(1/n) \tag{7}$$

uniformly in k. Because $f_{a,b}(t_{n,k})$ and $P_{n,k}$ have the common factor

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},$$

to prove (7) it will suffice to show that

$$D_{n,k} := (a+b+n)\binom{n}{k}\frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)} = t_{n,k}^{a-1}\,(1-t_{n,k})^{b-1} + O(1/n),$$

uniformly in k. Writing out $D_{n,k}$ in terms of factorials gives

$$D_{n,k} = (a+b+n) \times \frac{(a+k-1)(a+k-2)\cdots(k+1)\,(b+n-k-1)(b+n-k-2)\cdots(n-k+1)}{(a+b+n-1)(a+b+n-2)\cdots(n+1)}.$$

Note that both numerator and denominator have $a+b-1$ factors. Dividing *all* factors by $N = a+b+n$ and substituting $t_{n,k} = (a+k)/N$ and $1 - t_{n,k} = (b+n-k)/N$ now gives

$$\frac{\left(t_{n,k}-\tfrac{1}{N}\right)\left(t_{n,k}-\tfrac{2}{N}\right)\cdots\left(t_{n,k}-\tfrac{a-1}{N}\right)\left(1-t_{n,k}-\tfrac{1}{N}\right)\left(1-t_{n,k}-\tfrac{2}{N}\right)\cdots\left(1-t_{n,k}-\tfrac{b-1}{N}\right)}{\left(1-\tfrac{1}{N}\right)\left(1-\tfrac{2}{N}\right)\cdots\left(1-\tfrac{a+b-1}{N}\right)\cdots}$$

$$=\frac{\left(t_{n,k}+O(1/n)\right)^{a-1}\left(1-t_{n,k}+O(1/n)\right)^{b-1}}{(1+O(1/n))^{a+b-1}}$$

$$=t_{n,k}^{a-1}\left(1-t_{n,k}\right)^{b-1}+O(1/n)=D_{n,k},$$

as desired. (The assiduous reader is invited to verify the O-arithmetic in the last step.)

Next we use (7) to prove (6). To begin, write $I=[\alpha,\beta]$, fix n, and suppose that I contains the $t_{n,k}$ ranging from t_{n,k_0} to t_{n,k_M}, where $0\le k_0<k_M\le n$. Now the continuity of $f_{a,b}$ on $[0,1]$ implies that

$$\mathrm{P}(\theta\in I)=\int_\alpha^\beta f_{a,b}(t)\,dt=\int_{t_{n,k_0}}^{t_{n,k_M}} f_{a,b}(t)\,dt+O(1/n).$$

The last integral can, in turn, be approximated by a convenient approximating sum with step size $1/(a+b+n)$:

$$\int_{t_{n,k_0}}^{t_{n,k_M}} f_{a,b}(t)\,dt=\frac{f_{a,b}(t_{n,k_0})+f_{a,b}(t_{n,k_0+1})+\cdots+f_{a,b}(t_{n,k_M})}{a+b+n}+O(1/n).$$

(The $O(1/n)$ assertion holds because the integrand $f_{a,b}(t)$ has bounded first derivative.) Now (7) implies that

$$\frac{f_{a,b}(t_{n,k})}{a+b+n}=P_{n,k}+O\left(1/n^2\right)$$

for all k, which in turn gives

$$\begin{aligned}\frac{f_{a,b}(t_{n,k_0})+\cdots+f_{a,b}(t_{n,k_M})}{a+b+n}&\\ &=P_{n,k_0}+P_{n,k_0+1}+\cdots+P_{n,k_M}+M\cdot O\left(1/n^2\right)+O(1/n)\\ &=P_{n,k_0}+P_{n,k_0+1}+\cdots+P_{n,k_M}+O(1/n),\end{aligned}$$

and (6) follows.

Great expectations Next we find expected values of S_n and θ_n; perhaps surprisingly, we can do so without recourse to their explicit distributions, given in (2). First we write $S_n=S_{n-1}+X_n$, where X_n is 1 if the nth shot hits and 0 otherwise. Then we have

$$\begin{aligned}\mathrm{P}(X_n=1)&=\sum_{k=0}^{n-1}\mathrm{P}(X_n=1\mid S_{n-1}=k)\mathrm{P}(S_{n-1}=k)\\ &=\sum_{k=0}^{n-1}\frac{a+k}{a+b+n-1}\mathrm{P}(S_{n-1}=k)\\ &=\frac{a\sum\mathrm{P}(S_{n-1}=k)+\sum k\,\mathrm{P}(S_{n-1}=k)}{a+b+n-1}=\frac{a+\mathrm{P}(S_{n-1})}{a+b+n-1}.\end{aligned}\tag{8}$$

Thus,

$$P(S_n) = P(S_{n-1}) + P(X_n) = P(S_{n-1}) + P(X_n = 1)$$
$$= P(S_{n-1}) + \frac{a + P(S_{n-1})}{a+b+n-1} = \frac{a}{a+b+n-1} + P(S_{n-1})\frac{a+b+n}{a+b+n-1}.$$

Now clearly $E(S_0) = 0$, and it follows by induction that

$$P(S_n) = \frac{na}{a+b}$$

for $n \geq 0$. The expected value of θ_n is readily found—and seen to be constant:

$$P(\theta_n) = \frac{a + P(S_n)}{a+b+n} = \frac{a + \frac{na}{a+b}}{a+b+n} = \frac{a}{a+b}.$$

In particular, (8) shows that

$$P(X_n = 1) = P(\theta_{n-1}) = \frac{a}{a+b}.$$

That $P(X_n = 1)$ is constant recalls the earlier Bayesian "no free lunch" result, in which averaging over all possible outcomes did not produce new knowledge. Here, too, Shanille's initial success probability—$a/(a+b)$—remains unaltered as we imagine future outcomes.

Another way to think of this is to imagine a large number, M, of Shanille-clones, each starting with a hits and b misses and hence an initial $a/(a+b)$ success probability. Thereafter, each Shanille updates her own probability by the GPP rule. At each stage the distribution of all M Shanilles' success probabilities is as described in Proposition 2, but the group's *average* success probability remains at $a/(a+b)$, and its expected number of hits is $Ma/(a+b)$. After many stages, the distribution of the individual success probabilities strongly resembles a Beta(a, b) distribution.

Bayesian basketball Finally, we consider the *Bayesian Beta–Binomial Basketball Putnam Problem* (BBBPP):

> Shanille O'Keal, now a converted Bayesian, shoots free throws. Starting with a Beta distribution, at each stage she draws a value of θ from the distribution and, with success probability, shoots a basket and then updates her distribution by the beta-binomial method. Describe the marginal distributions of θ_n and of S_n, her success probability and total number of successes after n shots.

The BBBPP extends the GPP in several senses. First, the GPP starting points, a hits and b misses, mirror the initial Beta parameters a and b. Second, while the GPP Shanille is certain of her success probability at each stage, the BBBPP Shanille has less certainty—but she could always return to her GPP ways by replacing the random draw of θ_n with its expected value at each stage.

In the BBBPP setting S_n and θ_n have a new relationship: S_n is still a discrete random variable, but θ_n is now a continuous random variable that describes Shanille's skill. Now S_n is conditioned on θ_k, for $0 \leq k < n$, and θ_n is conditioned on S_n. Each θ_k is obtained by a beta-Bernoulli update of θ_{k-1}.

We might seem to have traded the relatively simple discrete GPP, with its single discrete random variable S_n, for a more complex BBBPP, with two intertwined random variables, one discrete and the other continuous. The bargain is better than it might seem—properties of the beta–binomial and of the Bayesian scheme turn out to simplify our solution. For instance, we can replace the sequence $\theta_0, \theta_1, \ldots, \theta_n$ of beta-Bernoulli updates with a *single* beta–binomial, starting with θ_0 and producing θ_n.

The marginal distribution of θ_n follows from the "no free lunch" principle: The marginal posterior and the prior distribution are identical. Thus, for all n, the success probability θ_n has marginal posterior distribution Beta(a, b), and $\mathrm{E}(\theta_n) = a/(a+b)$. In the GPP, by comparison, the distributions of θ_n *approximate* the Beta(a, b) distribution for large n, and $\mathrm{E}(\theta_n) = a/(a+b)$ for all n.

Consider, in particular, the noninformative (uniform) Beta(1, 1) prior distribution. Averaging over all possible outcomes assures that, at the nth stage, all values of θ_n *remain* equally likely. This brings the original PP to mind: All values θ_n for the updated success probability are equally likely at each stage.

Now we consider S_n. For the GPP the distribution of S_n, found in Proposition 2, is identical to the predictive posterior of a Beta(a, b) prior, shown in (4). The same result holds for the BBBPP, but here the connection is more natural. A sequence of n Bernoulli updates is equivalent, as shown earlier, to a single binomial update with n trials. The probability we seek, $\mathrm{P}(S_n = k)$, was found in (4):

$$\mathrm{P}(S_n = k) = \binom{n}{k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)}.$$

It follows that $\mathrm{E}(S_n) = na/(a+b)$ for all n, as in the GPP.

Conclusion

Here we end our tour of Shanille's basketball adventures and our detour through Bayesian statistics. We explored the GPP, a problem in probability, by "embedding" it in a Bayesian context. The embedding amounts, essentially, to replacing a fixed θ with a random variable with *expectation* θ. Almost every property of the fixed-θ case is reflected in a property of the variable-θ setting. The embedding helped reveal some interesting properties of the GPP and links to Bayesian principles, such as the "no free lunch" property.

Finally, a confession: Even the BBBPP smacks more of probability than of Bayesian statistics at its purest. To a fully committed Bayesian, what changes over successive free throw attempts is not really Shanille's success probability, as the BBBPP implies. Shanille's skill remains unchanged throughout—only our belief is knowable, and subject to updating. For Bayesians, it's all about belief.

REFERENCES

1. 63rd Annual William Lowell Putnam Mathematical Competition, this MAGAZINE **76** (2003), 76–80.
2. Siddhartha Chib, Edward Greenberg, Understanding the Metropolis-Hastings algorithm, *The American Statistician* **49** (1995), 327–335.
3. Peter Congdon, *Bayesian Statistical Modeling*, John Wiley and Sons, New York, NY, 2001.
4. Bradley Carlin and Thomas Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, CRC Press, Boca Raton, FL, 2000.
5. Andrew Gelman, John Carlin, Hal Stern, and Donald Rubin, *Bayesian Data Analysis*, CRC Press, Boca Raton, FL, 1995.
6. Robert V. Hogg, Allen T. Craig, *Introduction to Mathematical Statistics*, Prentice Hall, Upper Saddle River, NJ, 1995.
7. Sheldon Ross, *A First Course in Probability*, Prentice Hall, Upper Saddle River, NJ, 2002.