Bayesian Basketball

Charles Kankelborg

Revised April, 2024

Abstract

I introduce statistical inference using Bayes' theorem from minimal assumptions. As an example, I solve the following problem: A basketball player shoots N free throws, and all of them are good. What is the probability that she will sink the next shot?

1 Introduction to Bayes' Theorem

At this point, I will provide the background needed to apply Bayes' theorem not only to the free throw problem, but to larger and more difficult problems. In §1.1 I briefly introduce and justify Bayes' theorem. Then, in §1.2, I lay out the application of Bayes' theorem to parameter estimation.

1.1 Derivation

Two logical propositions, A and B, are capable of being either true or false. The expression $\Pr(A)$ means the probability that A is true. The expression $\Pr(A|B)$ denotes a conditional probability, which is read as the probability that A is true, given that B is true. Now, the probability that *both* A and B are true is given by a fundamental axiom of probability theory,

$$\Pr(A, B) = \Pr(A|B)\ \Pr(B).$$

In words, the probability that both A and B are true equals the probability of A under the assumption that B is true, multiplied by the probability that B is true. We take this as given, but I invite you to think about it carefully.

Equivalently, I could have written

$$\Pr(B, A) = \Pr(B|A)\ \Pr(A).$$

Notice, however, that $\Pr(A, B)$ and $\Pr(B, A)$ mean the same thing, the probability that both A and B are true. Consequently,

$$\Pr(A|B)\ \Pr(B) = \Pr(B|A)\ \Pr(A).$$

That is, the probability of A given B, multiplied by the probability of B, is equal to the probability of B given A, multiplied by the probability of A. The above equation is Bayes' theorem. It is frequently rearranged to solve for one of the two conditional probabilities,

$$\Pr(A|B) = \frac{\Pr(B|A)\ \Pr(A)}{\Pr(B)}.$$

1.2 Statistical Inference: Parameter Selection

While the above expression of Bayes' theorem is technically correct, we still need to develop the conceptual framework to use it for statistical inference. In particular, we will consider the problem of *parameter selection*.

Let us suppose that some data, $\mathbf{d}$, has been collected. We now assume that model $\mathcal{M}$ can be used to account for this data. The model has some number of tunable parameters, given by the parameter vector $\mathbf{x}$. We wish to use the data to asses the probability distribution for those parameters. We begin by using Bayes' theorem:

$$\boxed{\Pr(\mathbf{x}|\mathbf{d}, \mathcal{M}) = \frac{\Pr(\mathbf{d}|\mathbf{x}, \mathcal{M})\ \Pr(\mathbf{x}|\mathcal{M})}{\Pr(\mathbf{d}|\mathcal{M})}.} \tag{1}$$

This form of Bayes' theorem will motivate the rest of the paper. Often, $\mathcal{M}$ is omitted to simplify the notation, but it behooves us to remember that the model has been assumed to be true. The thing we wish to calculate, $\Pr(\mathbf{x}|\mathbf{d}, \mathcal{M})$, is called the *posterior distribution*. Can we evaluate all the quantities on the right hand side of the equation? The most straightforward part is probably $\Pr(\mathbf{d}|\mathbf{x}, \mathcal{M})$. After all, if assuming a model and choosing all its parameters does not give us enough information to assess the likelihood that our observations turn out a certain way, then the model is not very useful!

The factors $\Pr(\mathbf{x}|\mathcal{M})$ and $\Pr(\mathbf{d}|\mathcal{M})$ require careful thought. The notation seems to indicate that we are to calculate the probabily of some choice of model parameters without any knowledge of the data, and the probability of the data without having the slightest idea about the parameter values. That seems like an ill-defined problem, so I need to back up and think about $\Pr(\mathbf{x}|\mathcal{M})$ and $\Pr(\mathbf{d}|\mathcal{M})$ in a different way.

1.2.1 The Prior

We cannot calculate the probabilities $\Pr(\mathbf{x}|\mathcal{M})$ and $\Pr(\mathbf{d}|\mathcal{M})$ without starting from some assumption, so what is that starting point? The most important point is that $\Pr(\mathbf{x}|\mathcal{M})$ and $\Pr(\mathbf{d}|\mathcal{M})$ *must be compatible; that is, they must be derived from exactly the same set of assumptions.* We already know that both assume the same model, but it is usually necessary to add something. The way this is usually accomplished is to begin with an assumed parameter distribution $\Pr(\mathbf{x}|\mathcal{M})$. We will simply call this the *prior*, a too-short name that fails to express the weight on this starting assumption.[1]

If something is already known about the likelihoods of parameter choices before the data $\mathbf{d}$ are taken, then that information should be encoded in $\Pr(\mathbf{x}|\mathcal{M})$. If not, then usually the most honest choice is a *flat* prior, meaning that all parameter values are equally likely. If $\mathbf{x} \in \mathbb{R}^n$, the proper normalization of the prior would be

$$\int dx_1 \int dx_2 \cdots \int dx_n \, \Pr(x_1, x_2, ..., x_n|\mathcal{M}) \equiv \int \Pr(\mathbf{x}|\mathcal{M})) \, d^n x = 1. \quad (2)$$

The parameter integrals above are definite integrals over the appropriate domain for each parameter. Once we have defined a properly normalized multivariate distribution, the calculation of $\Pr(\mathbf{d}|\mathcal{M})$ will follow directly, as described in section 1.2.2.

[1] If it seems to you that assuming a probability distribution for $\mathbf{x}$ is begging the question, then I have done my job! That is exactly the danger we must keep in mind. One of the strengths of Bayesian inference is that it forces us to lay our cards on the table, making all assumptions explicit.

1.2.2 Marginalization

Once we have the prior, it becomes possible to work out $\Pr(\mathbf{d}|\mathcal{M})$. We do this by *marginalizing* over all possible values of the model parameters:

$$\Pr(\mathbf{d}|\mathcal{M}) = \int \Pr(\mathbf{d}|\mathbf{x}, \mathcal{M})\ \Pr(\mathbf{x}|\mathcal{M})\, d^n x. \tag{3}$$

Notice how the prior, $\Pr(\mathbf{x}|\mathcal{M})$, was required for marginalization. This establishes the required consistency between $\Pr(\mathbf{x}|\mathcal{M})$ and $\Pr(\mathbf{d}|\mathcal{M})$. Notice that if the prior not normalized, the normalization ultimately drops out in Bayes' theorem.

1.2.3 The flat prior

Let's think for a moment about a single model parameter, x. It is common practice to introduce a prior of the form

$$f(x) \equiv \Pr(x|\mathcal{M}) = c, \tag{4}$$

where c is a constant over the appropriate domain of x. This is called a "flat" prior. The idea behind a flat prior is to eliminate bias in parameter selection. Suppose, for example, that x is an astronomical distance whose uncertainty covers many orders of magnitude. It might then be more natural to work in terms of $y = \log x$:

$$g(y) \equiv \Pr(y|\mathcal{M}). \tag{5}$$

By conservation of probability,

$$\begin{aligned} f(x)\, dx &= g(y)\, dy \\ f(x) &= g(y)\, \frac{dy}{dx} = \frac{g(x)}{x} \\ \therefore\ g(\log x) &= cx \end{aligned}$$

Notice how the change of parameters affects the functional form of the prior. The distribution cannot be uniform in both x and $\log x$. So, which would be free of bias: flat $\Pr(x|\mathcal{M})$, flat $\Pr(\log x|\mathcal{M})$, or something else? Sometimes, the answer is obvious in context. But it is always important to spell out the prior clearly.

1.2.4 Posterior Distribution

We are now in a position to write down the probability distribution for the model parameters. Once again, this is called the posterior:

$$\Pr(\mathbf{x}|\mathbf{d}) = \frac{\Pr(\mathbf{d}|\mathbf{x})\ \Pr(\mathbf{x})}{\int \Pr(\mathbf{d}|\mathbf{x}')\ \Pr(\mathbf{x}')\, d^n x'}. \tag{6}$$

I have left the model $\mathcal{M}$ unstated to keep the notation compact. The primes are used to distinguish all the independent variables that disappear when the definite integral is calculated to marginalize over the parameter space. This notation helps us to see that the denominator does not introduce any dependence on the parameter vector, $\mathbf{x}$. If the prior is flat, then the posterior takes the same functional form as $\Pr(\mathbf{d}|\mathbf{x})$, and Bayes' theorem serves only to renormalize that distribution over the domain of the model parameter space. Nevertheless, correct normalization is important if we want to calculate a mean, a median, or confidence intervals from the distribution.

The posterior distribution, $\Pr(\mathbf{x}|\mathbf{d})$, is ostensibly the answer we have been looking for. Or is it? We have a range of possible models, described by the parameters $x_1, x_2, ..., x_n$. We also have a probability distribution for those parameters. We might prefer to have a well-defined best choice of parameters to describe the situation, $\mathbf{x}_{\text{best}}$. We will see from the basketball example in the next section that many possible choices of $\mathbf{x}_{\text{best}}$ could be made using $\Pr(\mathbf{x}|\mathbf{d})$.

2 The Free Throw Problem

We now pose the problem:

> A basketball player shoots N free throws, and all of them are good. What is the probability that she will sink the next shot?

This kind of statistical problem has many applications, including the analysis of risk in engineering.[2] Bayesian inference, as described in § 1, gives us a way to attack the problem.

[2]For example: A new kind of single-use solid rocket booster has had N flights, all successful. What is the likelihood that the next flight will be a success?

2.1 The Model and the Prior

In order to formulate a prior, I need to decide on a parameterized model. Let's characterize our basketball player by a single parameter, q, which is the probability of sinking a free throw. I have no other prior knowledge to incorporate, so I will use a flat prior:

$$\Pr(q) = 1, \quad 0 \leq q \leq 1.$$

Note the recursive use of the word probability: $\Pr(q)$ is the probability that the player's free throw probability has a certain value, q. This distribution is normalized:

$$\int_0^1 \Pr(q)\, dq = 1.$$

Exercise: Choosing a parameterized model restricts me from considering other models. What are some possible considerations that I have ignored by characterizing the player's free throw shooting ability with a single parameter, q?

2.2 Finding the Posterior Distribution

The next step is to calculate $\Pr(\mathbf{d}|q)$, the probability that the data would result, as a function of model parameter q. The data is that we have observed N good shots in a row; the sole model parameter is q. The chance of putting the ball through the hoop on any particular try is q, so the chance of two baskets in a row is q^2. Similarly, the chance of N baskets in a row is

$$\Pr(\mathbf{d}|\mathbf{x}) = \Pr(N|q) = q^N.$$

Now, I marginalize over all possible values of q to find $\Pr(\mathbf{d})$:

$$\Pr(\mathbf{d}) = \Pr(N) = \int \Pr(N|q)\ \Pr(q)\, dq = \int_0^1 q^N\, dq = \frac{1}{N+1}.$$

The posterior distribution is therefore

$$\Pr(q|N) = \frac{\Pr(N|q)\ \Pr(q)}{\Pr(N)} = (N+1)\, q^N. \tag{7}$$

2.3 Interpretation of the Posterior

The posterior distribution we have derived is zero for $q = 0$ and grows monotonically as q increases to 1. This makes a certain amount of sense. So long as $N > 0$, it has been demonstrated that she can hit the basket. If N is large, our player appears to be a very good shot. If we could decide on a single representative or "best" value for q, that would be our answer: the chance of making the next free throw is just $q_{\rm best}$. Some possible options for interpreting the posterior distribution would then include the usual measures of central tendency for a univariate (single parameter) probability distribution:

1. Mode: $q_{\rm best} = \arg\max \Pr(q|N)$.
2. Mean: $q_{\rm best} = \int_0^1 \Pr(q|N)\, q\, dq$.
3. Median: $\int_0^{q_{\rm best}} \Pr(q|N)\, dq = \frac{1}{2}$.

The mode is the *maximum likelihood* estimate. Since the distribution is monotonically increasing, the maximum likelihood estimate is $q = 1$. The larger N becomes, the more plausible this estimate becomes. However, it seems strange that our result would not depend on N. Should we think this player is a perfect shot after only a few successful tries? The other two estimates of $q_{\rm best}$ are more nuanced and will actually depend on N. How can we decide?

Let's look carefully at the original question: "What is the probability that she will sink the *next shot*?" I denote the answer as $\Pr(N+1|\mathbf{d}) = \Pr(N+1|N)$, by which I mean "the probability that the $N+1^{\rm st}$ free throw will be good, given the data that the first N free throws were good". I propose to calculate this probability by marginalizing over all possible values of q:

$$\Pr(N+1|N) = \int_0^1 \Pr(q|\mathbf{d})\ \Pr(N+1|q)\, dq.$$

In terms of q, the probability of sinking the next free throw is $\Pr(N+1|q) = q$. The answer is therefore:

$$\begin{aligned} \Pr(N+1|N) &= \int_0^1 \Pr(q|N)\, q\, dq \equiv \langle q \rangle \\ &= \int_0^1 (N+1)\, q^{N+1}\, dq = \boxed{\frac{N+1}{N+2}.} \end{aligned} \tag{8}$$

Notice that this corresponds to the mean value of the distribution, which is also called the *expectation value* of q, written $\langle q \rangle$. The reasonableness of this result becomes more apparent when we check extreme values:

1. In the limit $N \to \infty$, the player is a perfect shot: $\langle q \rangle \to 1$.
2. If $N = 0$, then we have no evidence, and $\langle q \rangle = \frac{1}{2}$.

Exercise: Calculate the median value of the posterior distribution, also called q_{50}, as in 50^{th} percentile. How does the median behave in the limits of extreme values of N?

Exercise: Perhaps you are not satisfied by a single "best" value for q. Using the posterior, calculate q_{90}, the smallest value of q to 90% confidence. In other words, there is only a 10% chance that $q < q_{90}$.

3 Further Exploration of the Problem

3.1 Non-flat prior

It might bother us to begin with a flat prior. Perhaps we have some minimal intuition about the likely results. For example, it seems unreasonable to expect that $q = 0$ because the player is presumed to have at least some skill. It is also unlikely that $q = 1$, because no athlete is perfect. A simple prior that goes to zero at the ends of the interval is:

$$\Pr(q) = 6q(1 - q).$$

Note that this is properly normalized. How does the new prior affect the result? We still have $\Pr(N|q) = q^N$. Marginalizing over q,

$$\Pr(N) = \int_0^1 6q^{N+1}(1 - q)\, dq = \frac{6}{(N + 2)(N + 3)}$$

The posterior is then:

$$\Pr(q|N) = (N + 2)(N + 3)(1 - q)q^{N+1}.$$

Exercise: Calculate $\langle q \rangle$ for the above posterior. Compare to the result for the flat prior.

3.2 Subsequent Trips to the Court

In §3.1, we modified the prior somewhat arbitrarily. A better use of a non-flat prior is to take account of results from previous experiments. For example:

> On Monday, a basketball player hits N free throws out of N attempts. She returns to the court on Tuesday, and hits K out of K. What is the probability that she will sink the next free throw?

Provided that the conditions have not changed, we can take both practice sessions into account in calculating the likelihood of hitting on the next attempt. The simplest approach would be to calculate it all at once. According to equation 8, $N + K$ baskets in a row implies that

$$\boxed{\langle q \rangle = \frac{N+K+1}{N+K+2}.} \tag{9}$$

It is more interesting (though more tedious) to use this example to demonstrate how Bayes' theorem takes account of the prior, $\Pr(\mathbf{x})$. We argue based on equation 7 that the posterior calculated from the first session should be the prior at the time we make the second trip to the court:

$$\Pr(\mathbf{x}) \to \Pr(q|N) = (N+1)\, q^N.$$

We must also marginalize over this distribution to get $\Pr(\mathbf{d})$ for the second session:

$$\begin{aligned} \Pr(\mathbf{d}) \to \Pr(K|N) &= \int_0^1 \Pr(K|q)\Pr(q|N)\, dq \\ &= (N+1)\int_0^1 q^{K+N}\, dq = \frac{N+1}{N+K+1}. \end{aligned}$$

Bayes' theorem for the posterior distribution, taking both datasets into account, is:

$$\Pr(q|K,N) = \frac{\Pr(K|q)\ \Pr(q|N)}{\Pr(K|N)}. \tag{10}$$

The posterior works out to be:

$$\Pr(q|K,N) = \frac{q^K(N+1)\, q^N}{(N+1)/(N+K+1)} = (N+K+1)\, q^{N+K},$$

which is exactly what we would have obtained by simply substituting $N \to N+K$ in equation 7. The expectation value (mean) of this posterior distribution is obviously what we predicted a moment ago, based on simpler reasoning (equation 9). This result should increase our confidence that *Bayes' theorem properly takes into account prior information.*

Often, prior information about a parameter or set of parameters comes in a very different form, and from a very different experiment, than the present experiment. If we are prepared to assume that the same model, with the same parameters, should hold on both occasions, we can use the method illustrated by equation 10 to build a posterior distribution that combines what has been learned from both experiments.

3.3 Example with Misses

So far, our star player has never missed. Let us generalize the problem as follows:

> A basketball player attempts N free throws, and sinks M of them.
> What is the probability that she will sink the next free throw?

Of course, $M \leq N$. We now have a more complicated data set, $\mathbf{d}$. Assuming underlying probability q, we can calculate $\Pr(\mathbf{d}|q)$ using the binomial distribution.

Exercise: Work out the rest of the example. Assuming a flat prior, what is $\Pr(q|M,N)$? What is $\langle q \rangle$? Explain the deviation from $\langle q \rangle = M/N$.

4 Summary

1. Bayes' theorem (equation 1) provides a rigorous way of calculating a posterior distribution, $\Pr(\mathbf{x}|\mathbf{d})$, that can be used for statistical inference about model parameters $\mathbf{x}$ based on data $\mathbf{d}$.

2. When we calculate the posterior distribution, it is essential to derive $\Pr(\mathbf{d})$ and $\Pr(\mathbf{x})$ using compatible assumptions. That means we marginalize over the model parameter space according to equation 3.

3. Bayes' theorem allows us to take prior information about the parameters, $\Pr(\mathbf{x})$, into account.

4. The interpretation of the posterior depends on the question being asked. In the case of parameter selection, for example, we might prefer the maximum likelihood, the expectation value, or some other descriptor of the posterior. If we are being more careful, we might integrate the posterior to obtain a 1- or 2-sided confidence interval.