The beta density, Bayes, Laplace, and Pólya

Saad Mneimneh

1 The beta density as a conjugate form

Suppose that \( k \) is a binomial random variable with index \( n \) and parameter \( p \), i.e.

\[
P(k|p) = \binom{n}{k} p^k (1-p)^{n-k}
\]

Applying Bayes’s rule, we have:

\[
f(p|k) \propto p^k (1-p)^{n-k} f(p)
\]

Therefore, a prior of the form

\[
f(p) \propto p^{\alpha-1} (1-p)^{\beta-1}
\]

is a conjugate prior since the posterior will have the form:

\[
f(p|k) \propto p^{k+\alpha-1} (1-p)^{n-k+\beta-1}
\]

It is not hard to show that

\[
\int_0^1 p^{\alpha-1} (1-p)^{\beta-1} dp = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}
\]

Let’s denote the above by \( B(\alpha, \beta) \). Therefore,

\[
f(p) = Be(\alpha, \beta)
\]

where \( Be(\alpha, \beta) \) is called the beta density with parameters \( \alpha > 0 \) and \( \beta > 0 \), and is given by:

\[
\frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}
\]

Note that the beta density can also be viewed as the posterior for \( p \) after observing \( \alpha - 1 \) successes and \( \beta - 1 \) failures, given a uniform prior on \( p \) (here both \( \alpha \) and \( \beta \) are integers).

\[
f(p|\alpha, \beta) \propto p^{\alpha-1} (1-p)^{\beta-1}
\]
Example: Consider an urn containing red and black balls. The probability of a red ball is $p$, but $p$ is unknown. The prior on $p$ is uniform between 0 and 1 (no specific knowledge). We repeatedly draw balls with replacement. What is the posterior density for $p$ after observing $\alpha - 1$ red balls and $\beta - 1$ black balls?

$$ f(p|\alpha - 1 \text{ red}, \beta - 1 \text{ black}) \propto \left( \frac{\alpha + \beta - 2}{\alpha - 1} \right) p^{\alpha - 1}(1 - p)^{\beta - 1} $$

Therefore, $f(p) = Be(\alpha, \beta)$. Note that both $\alpha$ and $\beta$ need to be equal to at least 1. For instance, after drawing one red ball only ($\alpha = 2$, $\beta = 1$), the posterior will be $f(p) = 2p$. Here’s a table listing some possible observations:

<table>
<thead>
<tr>
<th>observation</th>
<th>posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1$, $\beta = 1$</td>
<td>$f(p) = 1$</td>
</tr>
<tr>
<td>$\alpha = 2$, $\beta = 1$</td>
<td>$f(p) = 2p$</td>
</tr>
<tr>
<td>$\alpha = 2$, $\beta = 2$</td>
<td>$f(p) = 6p(1 - p)$</td>
</tr>
<tr>
<td>$\alpha = 3$, $\beta = 1$</td>
<td>$f(p) = 3p^2$</td>
</tr>
<tr>
<td>$\alpha = 3$, $\beta = 2$</td>
<td>$f(p) = 12p^2(1 - p)$</td>
</tr>
<tr>
<td>$\alpha = 3$, $\beta = 3$</td>
<td>$f(p) = 30p^2(1 - p)^2$</td>
</tr>
</tbody>
</table>

2 Laplace’s rule of succession

In 1774, Laplace claimed that an event which has occurred $n$ times, and has not failed thus far, will occur again with probability $(n + 1)/(n + 2)$. This is known as Laplace’s rule of succession. Laplace applied this result to the sunrise problem: What is the probability that the sun will rise tomorrow?

Let $X_1, X_2, \ldots$ be a sequence of independent Bernoulli trials with parameter $p$. Note that this notion of dependence is conditional on $p$. More precisely:

$$ P(X_1 = b_1, X_2 = b_2, \ldots, X_n = b_n|p) = \prod_{i=1}^{n} P(X_i = b_i) $$

In fact, $X_i$ and $X_j$ are not independent because by observing $X_i$, one could say something about $p$, and hence about $X_j$. This is a consequence of the Bayesian approach which treats $p$ itself as a random variable (unknown). Let $S_n = \sum_{i=1}^{n} X_i$. We would like to find the following probability:

$$ P(X_{n+1} = 1|S_n = k) $$
Observe that:

\[ P(X_{n+1} = 1|S_n = k) \]

\[ = \int_0^1 P(X_{n+1} = 1|p, S_n = k)f(p|S_n = k)dp \]

\[ = \int_0^1 P(X_{n+1} = 1|p)f(p|S_n = k)dp = \int_0^1 pf(p|S_n = k)dp \]

Therefore, we need to find the posterior density of \( p \). Assuming we know nothing about \( p \) initially, we will adopt the uniform prior \( f(p) = 1 \) between 0 and 1. Applying Bayes’ rule:

\[ f(p|S_n = k) \propto P(S_n = k|p)f(p) \propto p^k(1-p)^{n-k} \]

We conclude that:

\[ f(p|S_n = k) = \frac{1}{B(k+1, n-k+1)}p^{(k+1)-1}(1-p)^{(n-k+1)-1} \]

Finally,

\[ P(X_{n+1} = 1|S_n = k) = \int_0^1 pf(p|S_n = k)dp = \frac{k + 1}{n + 2} \]

We obtain Laplace’s result by setting \( k = n \).

3 Generalization

Consider a coin toss that can result in head, tail, or edge. We denote by \( p \) the probability of head, and by \( q \) the probability of tail, thus the probability of edge is \( 1 - p - q \). Observe that \( p, q \in [0, 1] \) and \( p + q \leq 1 \). In \( n \) coin tosses, the probability of observing \( k_1 \) heads and \( k_2 \) tails (and thus \( n - k_1 - k_2 \) edges) is given by the multinomial probability mass function (this generalizes the binomial):

\[ P(k_1, k_2) = \binom{n}{k_1} \binom{n-k_1}{k_2} p^{k_1} q^{k_2} (1-p-q)^{n-k_1-k_2} \]

The Dirichlet density is a generalization of beta and is conjugate to multinomial. For instance:

\[ f(p, q) = \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}p^{\alpha-1}q^{\beta-1}(1-p-q)^{\gamma-1} \]
4 Pólya’s urn

Pólya’s urn represents a generalization of a Binomial random variable. Consider the following scheme: An urn contains $b$ black and $r$ red balls. The ball drawn is always replaced, and, in addition, $c$ balls of the color drawn are added to the urn. When $c = 0$, drawings are equivalent to independent Bernoulli processes with $p = \frac{b}{b+r}$. However, with $c \neq 0$, the Bernoulli processes are dependent, each with a parameter that depends on the sequence of previous drawings.

For instance, if the first ball is black, the (conditional) probability of a black ball at the second drawing is $\frac{b+c}{b+r+c}$. The probability of the sequence black, black is, therefore, $\frac{b}{b+c} \frac{b}{b+r+c}$.

Let $X_n$ be a random variable denoting the number of black balls drawn in $n$ trials. What is $P(X_n = k)$? It is easy to show that all sequences with $k$ black balls have the same probability $p_n$ and, therefore,

$$P(X_n = k) = \binom{n}{k} p_n$$

We now compute $p_n$ as:

$$p_n = \frac{\prod_{i=1}^{k} \left[ b + (i-1)c \right] \prod_{i=1}^{n-k} \left[ r + (i-1)c \right]}{\prod_{i=1}^{n} \left[ b + r + (i-1)c \right]}$$

Rewriting in terms of the Gamma function (assuming $c > 0$), we have:

$$p_n = \frac{\Gamma\left( \frac{b}{c} + k \right) \Gamma\left( \frac{c}{c} + n - k \right)}{\Gamma\left( \frac{b}{c} + \frac{c}{c} + n \right)}$$

Therefore, the important parameters are $b/c$ and $r/c$. Note that we can rewrite the above as (verify it):

$$p_n = \int_0^1 p^k (1-p)^{n-k} Be\left( \frac{b}{c}, \frac{r}{c} \right) dp$$

So,

$$P(X_n = k) = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} Be\left( \frac{b}{c}, \frac{r}{c} \right) dp$$
5 Pólya’s urn generates beta

We now show that Pólya’s urn generates a beta distribution at the limit. For this, we will consider \( \lim_{n \to \infty} X_n / n \).

First note that we can write \( P(X_n = k) \) as follows:

\[
P(X_n = k) = \frac{\Gamma\left(\frac{b}{c} + \frac{r}{c} + \frac{n}{c}\right)}{\Gamma\left(\frac{b}{c}\right) \Gamma\left(\frac{r}{c}\right) \Gamma\left(\frac{n}{c} + 1\right) \Gamma\left(\frac{n}{c} + \frac{b}{c} + \frac{r}{c} + 1\right)}
\]

Using Stirling’s approximation \( \Gamma(x) \approx \sqrt{2\pi x} x^x e^{-x} \) as \( x \) goes to infinity, we can conclude that when \( x \) goes to infinity,

\[
\frac{\Gamma(x + a)}{\Gamma(x + b)} \approx x^{a-b}
\]

Therefore, when \( k \to \infty \) (but \( k \leq x \) for some \( 0 < x < 1 \)),

\[
P(X_n = k) = \frac{1}{B\left(\frac{b}{c}, \frac{r}{c}\right)} k^{\frac{b}{c} - 1} (n - k)^{\frac{r}{c} - 1} n^{1 - \frac{b}{c} - \frac{r}{c}}
\]

Now,

\[
P\left(\frac{X_n}{n} \leq x\right) = P\left(\frac{X_n}{n} = 0\right) + P\left(\frac{X_n}{n} = \frac{1}{n}\right) + \ldots + P\left(\frac{X_n}{n} = \frac{\lfloor nx \rfloor}{n}\right)
\]

As \( n \) goes to infinity, \( 1/n \) goes to zero; therefore:

\[
\int_0^x P\left(\frac{X_n}{n} = u\right) du = \lim_{n \to \infty} \frac{1}{n} \left[ P\left(\frac{X_n}{n} = 0\right) + P\left(\frac{X_n}{n} = \frac{1}{n}\right) + \ldots + P\left(\frac{X_n}{n} = \frac{\lfloor nx \rfloor}{n}\right) \right]
\]

\[
P\left(\frac{X_n}{n} \leq x\right) = n \int_0^x P\left(\frac{X_n}{n} = u\right) du = n \int_0^x P(X_n = nu) du
\]

And since \( nu \to \infty \), we can replace \( k \) by \( nu \) in the limiting expression we obtained for \( P(X_n = k) \) to get:

\[
P\left(\frac{X_n}{n} \leq x\right) = \int_0^x \frac{1}{B\left(\frac{b}{c}, \frac{r}{c}\right)} u^{\frac{b}{c} - 1} (1 - u)^{\frac{r}{c} - 1} du
\]

It is rather interesting that this limiting property of Pólya’s urn depends on the initial condition. Even more interesting is that if \( Y = \lim_{n \to \infty} X_n/n \), then conditioned on \( Y = p \) we have independent Bernoulli trials with parameter \( p \) (stated without proof).

\[
P(X_n = k|Y = p) = \binom{n}{k} p^k (1-p)^{n-k}
\]