Prior distributions

Alexandre Tchourbanov *
PKI 357, UNOmaha, 1110 South 67th St.
Omaha, NE 68182-0694, USA
Phone: (402)554-6004
E-mail: tchourba@cse.unl.edu

July 29, 2002

Abstract

This document introduces prior distributions for the purposes of Bayesian statistics. Using prior beliefs we can significantly improve our statistical inferences based on observations. Most of the books on statistics do not cover the material presented. Here we try to collect the information available on conjugate priors to certain distributions.

1 Introduction

According to the Bayesian rule [1], we can express posterior probability of certain event H given some *data* with the formula

$$P(H|data) = \frac{P(data|H)P(H)}{P(data)}$$

The probability of H given the *data* is called the posterior probability of H. The posterior equals to the likelihood time the prior divided by marginal probability of *data*.

The paper shows what priors we can have and how they affect posterior distributions given likelihood.

2 Prior and posterior distributions

Sometimes a prior distribution can be approximated by one that is in a convenient family of distributions, which combines with the likelihood to produce a posterior that is manageable.

We see that an objective way of building priors for the binomial parameter was to use the conjugate family distribution that has the property that the updated distribution is in the same family. In general, if the prior distribution belongs to a family G, the data have a distribution belonging to a family H, and the posterior distribution also belongs to G, then we say that G is a family of *conjugate priors* to H. Thus, the beta distribution is a conjugate prior to the binomial, and the normal is self conjugate. Conjugate priors may not exist; when they do, selecting a member of the conjugate family as a prior is done mostly for mathematical convenience, since the posterior can be evaluated very simply. More generally, numerical methods of integration would have to be used to evaluate the posterior.

*I would like to thank professors Hesham Ali and Jitender Deogun for the opportunity to work on this project

Observations	Prior	Posterior
Bernoulli	Beta	Beta
Poisson	Gamma	Gamma
Binomial	Beta	Beta
Normal	Normal	Normal
Normal	Gamma	Gamma

Table 1: Conjugate priors

3 Beta priors

From Bayes 1763: A white billiard ball W is rolled along a line and we look at where it stops, scale the table from 0 to 1. We suppose that it has a uniform probability of falling anywhere on the line. It stops at a point p. A red billiard ball R is then rolled n times under the same uniform assumption. X then denotes the number of times R goes no further than W went.

Given X, what inference can we make about p? Here we are looking for the posterior distribution of p given X. The prior distribution of p is uniform $g(p) = Uniform(0,1) = Beta(1,1) = 1$. Given p, X has binomial distribution

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

The overall distribution of the number of successes is the sum of probabilities for all possible p's

$$P(a < p < b, X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$$

Suppose we throw all $n+1$ balls on the table, and choose the red one. Then the probability that the red one has x whites to the left of it is $\frac{1}{n+1}$. So we have

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp = \binom{n}{x} \int_0^1 p^x (1-p)^{n-x} dp = \frac{1}{n+1}$$

$$\int_0^1 p^x (1-p)^{n-x} dp = \frac{x!(n-x)!}{(n+1)!}$$

according to definition formula for beta function

$$B(r,s) = \int_0^1 p^{r-1}(1-p)^{s-1} dp = \frac{(r-1)!(s-1)!}{(r+s-1)!} = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

we have

$$X \sim B(n,p)$$

$$P(X = x) = \binom{n}{x} \int_0^1 p^x (1-p)^{n-x} dp = \binom{n}{x} B(x+1, n-x+1)$$

$$P(a < p < b | X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp}{\binom{n}{x} B(x+1, n-x+1)}$$

$$P(a < p < b | X = x) = \frac{\int_a^b p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

which is a beta distribution of p with parameters $x+1$ and $n-x+1$.
The density function f of the beta distribution is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \qquad 0 \leq p \leq 1$$

Example
Suppose that the prior distribution of p is $Beta(a, b)$, i.e.

$$g(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}$$

Likelihood has a binomial distribution

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

The posterior distribution of p given x is

$$\begin{aligned} h(p|x) &= \frac{f(x|p)g(p)}{f(x)} \\ &= \frac{B(a,b)\binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1}}{B(a,b)\int_0^1 \binom{n}{x} p^{x+a-1}(1-p)^{n-x+b-1} dp} \\ &= \frac{p^{a+x-1}(1-p)^{n+b-x-1}}{B(a+x, n+b-x)} \\ &= Beta(a+x, n+b-x) \end{aligned}$$

This distribution is thus beta as well with parameters $a' = a + x$ and $b' = b + n - x$.

4 Normal prior

Here we follow example on page 589 [2], which proves the Normal conjugate prior for Normal distribution.

The conjugate for a Normal likelihood is the Normal distribution.

Example
We consider inference concerning an unknown mean with known variance.

First, suppose that the prior distribution of μ is $N(\mu_0, \sigma_0)$. A single observation $X \sim N(\mu, \sigma^2)$ is taken. The posterior distribution of μ is

$$h(\mu|x) = \frac{f(x|\mu)g(\mu)}{\int_{-\infty}^{\infty} f(x|\mu)g(\mu)d\mu} \propto f(x|\mu)g(\mu)$$

$$\begin{aligned} f(x|\mu)g(\mu) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ &\propto \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right)\right) \end{aligned} \quad (1)$$

Let a, b and c be the coefficients in the quadratic polynomial in μ that is the last expression. Then (1) may then be written

$$h(\mu|x) = \exp\left(-\frac{a}{2}\left(\mu^2 - \frac{2b}{a}\mu + \frac{c}{a}\right)\right) \quad (2)$$

To simplify this further, we use the technique of completing the square and rewrite the expression (2) as

$$\begin{aligned} h(\mu|x) &= \exp\left(-\frac{a}{2}\left(\mu-\frac{b}{a}\right)^2\right) \exp\left(-\frac{a^2}{2}\left(\frac{c}{a} - \left(\frac{b}{a}\right)^2\right)\right) \\ &\propto \exp\left(-\frac{a}{2}\left(\mu-\frac{b}{a}\right)^2\right) \end{aligned}$$

We see that posterior distribution of μ is normal with mean

$$\mu_1 = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

For practical reasons, we define the precision as the inverse of the variance: we denote by $\xi = \frac{1}{\sigma^2}$ and and $\xi_0 = \frac{1}{\sigma_0^2}$

Theorem 1 *Suppose that $\mu \sim N(\mu_0, \sigma_0^2)$. Then the posterior distribution of μ is normal with mean*

$$\mu_1 = \frac{\sigma_0\mu_0 + \xi x}{\xi_0 + \xi}$$

and precision

$$\xi_1 = \xi_0 + \xi$$

The posterior mean is a weighted average of the prior mean and the data, weights being proportional to the respective precisions. With a very gentle prior we would have a very low precision ξ_0, a very at prior and mostly the posterior is Normal with x as its mean. Of course what we are usually interested in is the posterior given an iid sample of size n, what you could expect happens it is equivalent to adding one observation $\bar{x}$ from a distribution that has variance $\frac{\sigma^2}{n}$.

5 Multinomial Dirichlet priors

Dirichlet prior Dirichlet prior is conjugate to multinomial distribution. This is a probability distribution on on the n simplex.

$\Delta_n = \{\tilde{p} = (p_1, p_2, \ldots, p_n), p_1 + \ldots + p_n = 1, p_i \geq 0\}$

The Dirichlet distribution can be written as

$$D(\Theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \Theta_i^{\alpha_i - 1}$$

where

$\alpha = \alpha_1, \ldots, \alpha_K$, with $\alpha_i > 0$ are constants specifying the Dirichlet distribution

Θ_i satisfy $0 \leq \Theta_i \leq 1$ and $\sum_{i=1}^{K} \Theta_i = 1$

The multinomial distribution corresponding to k balls dropped into n boxes with fixed probability $(p_1, \ldots, p_n)$, (with ith box containing k_i balls) is

$$\binom{k}{k_1, \ldots, k_n} p_1^{k_1} \cdots p_n^{k_n}$$

For two variables $K = 2$ the Dirichlet distribution reduces to Beta distribution, and normalizing constant becomes Beta function.

The Dirichlet is a convenient prior because the posterior $\tilde{\text{p}}$ having observed $(k_1, \ldots, k_n)$ is Dirichlet with probability $(\alpha_1 + k_1, \ldots, \alpha_n + k_n)$. An important characterization of the Dirichlet: it is the only prior that predicts outcomes linearly in the past. One frequently used special case is the symmetric Dirichlet when all $\alpha_i = c > 0$. We denote this prior as D_c.

Dirichlet priors are important because

- They are natural conjugate priors for multinomial distributions, i.e. posterior parameter distribution, after having observed some data from a multinomial distribution with Dirichlet prior, also have form of Dirichlet distribution
- The Dirichlet distribution can be seen as multivariate generalization of the beta distribution, over the space of distributions P, with a constant on the average distance (*relative entropy*) to a reference distribution determined by Θ and α.

References

[1] Rev. Thomas Bayes, *An essay towards solving a problem in the doctrine of chances*, Philosophical Transactions of the Royal Society of London (1763).

[2] John. A. Rice, *Mathematical statistics and data analysis*, Duxbury Press, 1995.