Why Beta Priors: Invariance-Based Explanation

Olga Kosheleva[1], Vladik Kreinovich[1], and
Kittawit Autchariyapanitkul[2]
[1]University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu
[2]Faculty of Economics, Maejo University
Chiang Mai, Thailand, kittar3@hotmail.com

Abstract

In the Bayesian approach, to describe a prior distribution on the set $[0, 1]$ of all possible probability values, typically, a Beta distribution is used. The fact that there have been many successful applications of this idea seems to indicate that there must be a fundamental reason for selecting this particular family of distributions. In this paper, we show that the selection of this family can indeed be explained if we make reasonable invariance requirements.

1 Formulation of the Problem

In the Bayesian approach (see, e.g., [2, 4]), when we do not know the probability $p \in [0, 1]$ of some event, it is usually recommended to use a Beta prior distribution for this probability, i.e., a distribution for which the probability density function $\rho(x)$ has the form

$$\rho(x) = c \cdot x^{\alpha-1} \cdot (1 - x)^{\beta-1},$$

where α and β are appropriate constants and c is a normalizing constant – guaranteeing that

$$\int_0^1 \rho(x)\, dx = 1.$$

There have been numerous successful application of the use of the Beta distribution in the Bayesian approach. How can we explain this success? Why not use some other family of distributions located on the interval $[0, 1]$?

In this paper, we provide a natural explanation for these empirical successes.

Comment. The need for such an explanation is especially important now, when the statistician community is replacing the traditional p-value techniques with more reliable hypothesis testing methods (see, e.g., [3, 7]), methods such as the Minimum Bayesian Factor (MBF) method which is based on using a specific class of Beta priors

$$\rho(x) = c \cdot x^{\alpha-1}$$

that correspond to $\beta = 1$; see, e.g., [5].

2 Analysis of the Problem and the Main Result

Main idea. We want to find a natural prior distribution on the interval $[0, 1]$, a distribution that describes, crudely speaking, how frequently different probability values p appear. In determining this distribution, a natural idea to take into account is that, in practice, all probabilities are, in effect, conditional probabilities: we start with some class, and in this class, we find the corresponding frequencies.

From this viewpoint, we can start with the original probabilities and with their prior distribution, or we can impose additional conditions and consider the resulting conditional probabilities. For example, in medical data processing, we may consider the probability that a patient with a certain disease recovers after taking the corresponding medicine. We can consider this original probability – or, alternatively, we can consider the conditional probability that a patient will recover – e.g., under the condition that the patient is at least 18 years old.

We can impose many such conditions, and, since we are looking for a universal prior, a prior that would describe all possible situations, it makes sense to consider priors for which, after such a restriction, we will get the exact same prior for the corresponding conditional probability.

Let us describe this main idea in precise terms. In general, the conditional probability $P(A \,|\, B)$ has the form

$$P(A \,|\, B) = \frac{P(A \,\&\, B)}{P(B)}.$$

Crudely speaking, this means that when we transition from the original probabilities to the new conditional ones, we limit ourselves to the original probabilities which do not exceed some value $p_0 = P(B)$, and we divide each original probability by p_0.

In these terms, the above requirement takes the following form: for each $p_0 \in (0, 1)$, if we limit ourselves to the interval $[0, p_0]$, then the ratios p/p_0 should have the same distribution as the original one.

Definition 1. *We say that a probability distribution with probability density $\rho(x)$ on the interval $[0, 1]$ is* invariant *if for each $p_0 \in (0, 1)$, the ratio x/p_0 (restricted to the values $x \le p_0$) has the same distribution, i.e., if*

$$\rho(x/p_0 : x \le p_0) = \rho(x).$$

Proposition 1. *A probability distribution is invariant if and only if it has a form*

$$\rho(x) = c \cdot x^a$$

for some c and a.

Proof. The conditional probability density has the form

$$\rho(x/p_0 : x \le p_0) = C(p_0) \cdot \rho(x/p_0),$$

for an appropriate constant C depending on p_0. Thus, the invariance condition has the form

$$C(p_0) \cdot \rho(x/p_0) = \rho(x).$$

By moving the term $C(p_0)$ to the right-hand side and denoting $\lambda \stackrel{\text{def}}{=} 1/p_0$ (so that $p_0 = 1/\lambda$), we conclude that

$$\rho(\lambda \cdot x) = c(\lambda) \cdot \rho(x), \tag{1}$$

where we denoted $c(\lambda) \stackrel{\text{def}}{=} 1/C(1/\lambda)$.

The probability density function is an integrable function – its integral is equal to 1. It is known (see, e.g., [1]) that every integrable solution of the functional equation (1) has the form

$$\rho(x) = c \cdot x^a$$

for some c and a. The proposition is thus proven.

Comment. It is worth mentioning that namely these distributions – corresponding to $\beta = 1$ – are used in the Bayesian approach to hypothesis testing [5, 6].

How to get a general prior distribution. The above proposition describes the case when we have a single distribution corresponding to a single piece of prior information. In practice, we may have many different pieces of information. Some of these pieces are about the probability p of the corresponding event E, some may be about the probability $p' = 1 - p$ of the opposite event $\neg E$.

According to Proposition 1, each piece of information about p can be described by the probability density

$$c_i \cdot x^{a_i},$$

for some c_i and a_i. Similarly, each piece of information about $p' = 1 - p$ can be described by the probability density

$$c'_j \cdot x^{a'_j}$$

for some c'_j and a'_j. In terms of the original probability $p = 1 - p'$, this probability density has the form

$$c'_j \cdot (1 - x)^{a'_j}.$$

Since all these piece of information are independent, a reasonable idea is to multiply these probability density functions. After multiplication, we get a distribution of the type

$$c \cdot x^a \cdot (a - x)^{a'},$$

where $a = \sum_i a_i$ and $a' = \sum_j a'_j$. This is exactly the Beta distribution – for $\alpha = a + 1$ and $\beta = a' + 1$.

Thus, we have indeed justified the use of Beta priors.

Acknowledgments

This work was supported by the Institute of Geodesy, Leibniz University of Hannover. It was also supported in part by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

This paper was written when V. Kreinovich was visiting Leibniz University of Hannover.

References

[1] J. Aczel and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 1989.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, Florida, 2013.

[3] A. Gelman and C. P. Robert, "The statistical crises in science", *American Scientist*, 2014, Vol. 102, No. 6, pp. 460–465.

[4] K. R. Kock, *Introduction to Bayesian Statistics*, Springer, 2007.

[5] H. T. Nguyen, "How to test without p-values", *Thailand Statistician*, 2019, Vol. 17, No. 2, pp. i-x.

[6] R. Page and E. Satake, "Beyond p-values and hypothesis testing: using the Minimum Bayes Factor to teach statistical inference in undergraduate introductory statistics courses", *Journal of Education and Learning*, 2017, Vol. 6, No. 4, pp. 254—266.

[7] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: context, process, and purpose", *American Statistician*, 2016, Vol. 70, No. 2, pp. 129–133.