How to Make a Decision Based on the Minimum Bayes Factor (MBF): Explanation of the Jeffreys Scale

Olga Kosheleva, Vladik Kreinovich, Nguyen Duc Trung, and Kittawit Autchariyapanitkul

Abstract In many practical situations, we need to select a model based on the data. It is, at present, practically a consensus that the traditional p-value-based techniques for such selection often do not lead to adequate results. One of the most widely used alternative model selection techniques is the Minimum Bayes Factor (MBF) approach, in which a model is preferred if the corresponding Bayes factor – the ratio of likelihoods corresponding to this model and to the competing model – is sufficiently large for all possible prior distributions. Based on the MBF values, we can decide how strong is the evidence in support of the selected model: weak, strong, very strong, or decisive. The corresponding strength levels are based on a heuristic scale proposed by Harold Jeffreys, one of the pioneers of the Bayes approach to statistics. In this paper, we propose a justification for this scale.

1 Formulation of the Problem

Why Minimum Bayes Factor. In many practical situations, we have several possible models M_i of the corresponding phenomena, and we would like to decide, based on the data D, which of these models is more adequate. To select the most appropriate model, statistics textbooks used to recommend techniques based on p-values. However, at present, it is practically a consensus in the statistics community that the use of p-values often results in misleading conclusions; see, e.g., [3, 4, 7].

Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: olgak@utep.edu, vladik@utep.edu

Nguyen Duc Trung
Banking University HCMC, Ho Chi Minh City, Vietnam, e-mail: trungnd@buh.edu.vn

Kittawit Autchariyapanitkul
Maejo University, Maejo, Thailand, kittar3@hotmail.com

To make a more adequate selection, it is important to take prior information into account, i.e., to use Bayesian methods. It is reasonable to say that the model M_1 is more probable than the model M_2 if the likelihood $P(D|M_1)$ of getting the data D under the model M_1 is larger than the likelihood $P(D|M_2)$ of getting the data D under the model M_2, i.e., if the *Bayes factor*

$$K \stackrel{\text{def}}{=} \frac{P(D|M_1)}{P(D|M_2)}$$

exceeds 1. Of course, if the value is only slightly larger than 1, this difference may be caused by the randomness of the corresponding data sample. So, in reality, each of the two models can be more adequate. To make a definite conclusion, we need to make sure that the Bayes factor is sufficiently large – and the larger the factor K, the more confident we are that the model M_1 is indeed more adequate.

The numerical value of the Bayes factor K depends on the prior distribution π: $K = K(\pi)$. In practice, we often do not have enough information to select a single prior distribution. A more realistic description of the expert's prior knowledge is that we have a *family* F of possible prior distributions π. In such a situation, we can conclude that the model M_1 is more adequate than the model M_2 if/ the corresponding Bayes factor is sufficiently large for all possible prior distributions $\pi \in F$, i.e., equivalently, that the *Minimum Bayes Factor*

$$\text{MBF} \stackrel{\text{def}}{=} \min_{\pi \in F} K(\pi)$$

is sufficiently large; see, e.g., [4, 5].

Jeffreys scale. In practical applications of Minimum Bayes Factor, the following scale is usually used; this scale was originally proposed in [2]:

- when the value of MBF is between 1 and 3, we say that the evidence for the model M_1 is barely worth mentioning;
- when the value of MBF is between 3 and 10, we say that the evidence for the model M_1 is substantial;
- when the value of MBF is between 10 and 30, we say that the evidence for the model M_1 is strong;
- when the value of MBF is between 30 and 100, we say that the evidence for the model M_1 is very strong;
- finally, when the value of MBF is larger than 100, we say that the evidence for the model M_1 is decisive.

Remaining problem and what we do in this paper. Jeffreys scale has been effectively used, so it seems to be adequate, but why? Why do we select, e.g., 1 to 3 and not 1 to 2 and 1 to 5?

In this paper, we provide a possible explanation for the success of Jeffreys scale. This explanation is based on a general explanation of the half-order-of-magnitude scales provided in [1].

2 Our Explanation

Towards the precise formulation of the problem. A scale means, crudely speaking, that instead of considering all possible values of the MBF, we consider discretely many values

$$\ldots < x_0 < x_1 < x_2 < \ldots$$

corresponding to different levels of strength. Every actual value x is then approximated by one of these values $x_i \approx x$.

What is the probability distribution of the resulting approximation error $\Delta x \stackrel{\text{def}}{=} x_i - x$? This error is caused by many different factors. It is known that under certain reasonable conditions, an error caused by many different factors is distributed according to Gaussian (normal) distribution (see, e.g., [6]; this result – called the *Central Limit Theorem* – is one of the reasons why Gaussian distributions are ubiquitous). It is therefore reasonable to assume that Δx is normally distributed.

It is known that a normal distribution is uniquely determined by its two parameters: its average μ and its standard deviation σ. For situations in which the approximating value is x_i, let us denote:

- the mean value of the approximation error Δx by Δ_i, and
- the standard deviation of the approximation error by σ_i.

Thus, when the approximate value is x_i, the actual value $x = x_i - \Delta x$ is distributed according to the Gaussian distribution, with the mean $x_i - \Delta_i$ (which we will denote by $\widetilde{x_i}$), and the standard deviation σ_i.

For a Gaussian distribution with mean μ and standard deviation σ, the probability density is everywhere positive, so theoretically, we can have values which are as far away from the mean value μ as possible. In practice, however, the probabilities of large deviations from μ are so small that the possibility of such deviations can be safely ignored. For example, it is known that the probability of having the value outside the "three sigma" interval $[\mu - 3\sigma, \mu + 3\sigma]$ is $\approx 0.1\%$ and therefore, in most applications in science and engineering, it is assumed that values outside this interval are impossible.

There are some applications where we cannot make this assumption. For example, in designing computer chips, when we have millions of elements on the chip, allowing 0.1% of these elements to malfunction would mean that at any given time, thousands of elements malfunction and thus, the chip would malfunction as well. For such critical applications, we want the probability of deviation to be much smaller than 0.1%, e.g., $\leq 10^{-8}$. Such small probabilities (which practically exclude any possibility of an error) can be guaranteed if we use a "six sigma" interval $[\mu - 6\sigma, \mu + 6\sigma]$. For this interval, the probability for a normally distributed variable to be outside it is indeed $\approx 10^{-8}$.

In accordance with the above idea, for each x_i, if the actual value x is within the "three sigma" range $I_i = [\widetilde{x_i} - 3\sigma_i, \widetilde{x_i} + 3\sigma_i]$, then it is reasonable to take x_i as the corresponding approximation.

What should be the standard deviation σ_i of the approximation error? We are talking about a very crude approximation, when, e.g., all the values from 1 to 3 are assigned the same level. Thus, the approximation error has to be reasonably large. The only limitation on the approximation error is that we want to make sure that all values that we are covering are indeed non-negative, i.e., that for every i, even the extended "six sigma" interval $[\widetilde{x}_i - 6\sigma_i, \widetilde{x}_i + 6\sigma_i]$ only contains non-negative values. Other than that, there should not be any other limitations on the approximation error – i.e., the value σ_i should be the largest for which the above property holds.

We want to cover all possible values x, so that each positive real number x be covered by one of the intervals I_i. In other words, we want the union of all these intervals to coincide with the set of all positive real numbers. We also want to make sure that to each value x, we assign exactly one strength level, i.e., that the intervals I_i corresponding to different strength levels do not intersect – except maybe at the borderline point.

Thus, we arrive at the following definitions.

Definition.

- *We say that an interval $I = [\mu - 3\sigma, \mu + 3\sigma]$ is* reliably non-negative *if every real number from the interval $[\mu - 6\sigma, \mu + 6\sigma]$ is non-negative.*
- *We say that an interval $I = [\mu - 3\sigma, \mu + 3\sigma]$ is* realistic *if for the given μ, the corresponding value σ is the largest for which the corresponding interval is reliably non-negative.*
- *We say that a set of realistic intervals $\{I_i = [\underline{x}_i, \overline{x}_i]\}$ with*

$$\ldots \le \underline{x}_1 \le \underline{x}_2 \le \ldots$$

describes strength levels *if these intervals form a partition of the set $\mathbb{R}^+$ of all positive real numbers: $\bigcup\limits_i I_i = \mathbb{R}^+$ and for each $i \ne j$, the intersection $I_i \cap I_j$ is either an empty set or a single point.*

Proposition. *A set of realistic intervals $I_i = [\underline{x}_i, \overline{x}_i]$ describes strength levels if and only if these intervals have the form $[\underline{x}_i, \overline{x}_i] = [3^i \cdot x_0, 3^{i+1} \cdot x_0]$.*

Discussion. In other words, we have intervals

$$[x_0, 3 \cdot x_0], \ \ [3 \cdot x_0, 9 \cdot x_0], \ \ [9 \cdot x_0, 27 \cdot x_0], \ldots$$

This is (almost) what the Jeffreys scale recommends, with $x_0 = 1$ – the only difference is that in the Jeffreys scale, we have 10 instead of 9. Modulo this minor issue, we indeed have an explanation for the empirical success of the Jeffreys scale.

Proof of the Proposition. Each interval

$$I_i = [\underline{x}_i, \overline{x}_i] = [\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$$

is realistic. This means that when the value μ_i is fixed, the corresponding value σ_i is the largest for which all the numbers from the interval $[\mu_i - 6\sigma_i, \mu_i + 6\sigma_i]$ are non-negative. One can easily see that this largest value corresponds to the case when

$\mu_i - 6\sigma_i = 0$, i.e., when $\sigma_i = \frac{1}{6} \cdot \mu_i$. For this value σ_i, we have $\underline{x}_i = \mu_i - 3\sigma_i = \frac{1}{2} \cdot \mu_i$ and $\overline{x}_i = \mu_i + 3\sigma_i = \frac{3}{2} \cdot \mu_i$. Thus, for each realistic interval $I_i = [\underline{x}_i, \overline{x}_i]$, we have

$$\overline{x}_i = 3 \cdot \underline{x}_i.$$

In particular, this is true for $i = 0$, so we have $\overline{x}_0 = 3x_0$, where we denoted $x_0 \stackrel{\text{def}}{=} \underline{x}_0$. Let us prove, by induction, that for every i, we have $\underline{x}_i = 3^i \cdot x_0$ and $\overline{x}_i = 3^{i+1} \cdot x_0$. Indeed, we have just proved these equalities for $i = 0$, i.e., we have the induction base.

Let us now prove the induction step. Suppose that

$$I_i = [\underline{x}_i, \overline{x}_i] = [3^i \cdot x_0, 3^{i+1} \cdot x_0].$$

The intervals I_i form a partition, so the next interval I_{i+1} intersects with I_i at exactly one point: $\underline{x}_{i+1} = \overline{x}_i = 3^{i+1} \cdot x_0$. Since the interval I_{i+1} is realistic, we have

$$\overline{x}_{i+1} = 3 \cdot \underline{x}_{i+1} = 3^{(i+1)+1} \cdot x_0.$$

The induction step is thus proven, and so is the proposition.

Acknowledgments

This work was supported in part by the 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).

References

1. J. Hobbs and V. Kreinovich, "Optimal choice of granularity in commonsense estimation: why half-orders of magnitude", *International Journal of Intelligent Systems*, 2006, Vol. 21, No. 8, pp. 843–855.
2. H. Jeffreys, *Theory of Probability*, Claredon Press, Oxford, 1989.
3. H. T. Nguyen, "Why p-values are banned?", *Thailand Statistician*, 2016, Vol. 24, No. 2, pp. i-iv.
4. H. T. Nguyen, "How to test without p-values?", *Thailand Statistician*, 2019, Vol. 17, No. 2, pp. i-x.
5. 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.
6. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
7. R. L. Wasserstein and N. A. Lazar, "The American Statistical Association's statement on p-values: context, process, and purpose", *American Statistician*, 2016, Vol. 70, No. 2, pp. 129–133.