NIST Technical Note
NIST TN 2250

A Way of Estimating the Standard Errors of Bayes Factor and Weight of Evidence – A Case Study (Theoretical Framework)

Jin Chu Wu
John M. Libert

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2250

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Technical Note
NIST TN 2250

A Way of Estimating the Standard Errors of Bayes Factor and Weight of Evidence – A Case Study (Theoretical Framework)

Jin Chu Wu
John M. Libert

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2250

March 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST Technical Series Policies
Copyright, Fair Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2023-03-24

How to Cite this NIST Technical Series Publication
Wu, Jin Chu, Libert, John (2023) A Way of Estimating the Standard Errors of Bayes Factor and Weight of Evidence – A Case Study (Theoretical Framework. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Technical Note (TN) 2250. https://doi.org/10.6028/NIST.TN.2250

NIST Author ORCID iDs
Wu, Jin Chu: 0000-0002-6340-2467
Libert, John: 0000-0003-0796-6871

Contact Information
jinchu.wu@nist.gov

Abstract

The Weight of Evidence (WoE) is defined to be the logarithm of the Bayes factor (BF) with base 10, which is generally with single point hypothesis rather than diffuse hypothesis. They are used in applications such as forensic science, etc. To statistically estimate the standard error (SE) and the 95% confidence interval of BF and WoE, both parametric and nonparametric two-sample bootstrap algorithms are employed, respectively. Then, three challenging issues arise: 1. how to generate observed binomial variates; 2. how many variates are needed; 3. how to implement bootstrap algorithms. The observed binomial variates can be generated using either the stochastic function-call method (i.e., call rbinom in the R Stats Package) or the deterministic partition method via the expected binomial densities (i.e., call dbinom in the R Stats Package). To ensure the computational accuracies, the size of observed binomial variates is determined by the root-mean-square deviation between the observed and expected binomial distributions, as well as the bootstrap sampling variability study. Thereafter, the parametric two-sample bootstrap algorithm is implemented on observed binomial variates generated using the stochastic function-call method, whereas the nonparametric two-sample bootstrap algorithm is carried out on observed binomial variates created using the deterministic partition method. In this article, a case study is carried out.

Keywords

weight of evidence, Bayes factor, standard error, binomial distribution, two-sample bootstrap, forensic science.

Acknowledgment

The authors would like to thank Dr. Carina Hahn for reviewing our manuscript and her comments.

Table of Contents

1. Introduction

The Weight of Evidence (WoE) is defined to be the logarithm of the Bayes factor (BF) with base 10, which is generally with single point hypothesis rather than diffuse hypothesis [1-4]. They are used in applications such as forensic science, etc. It is imperative to statistically estimate the standard error (SE) and the 95% confidence interval (CI) of measures so that the conventional statistical evaluation and comparison of the performance accuracies of different classifiers can be implemented properly [5-8]. Moreover, for BF and WoE, to transfer information, the SEs of BF and WoE are also needed [9].

In our prior research regarding receiver operating characteristic (ROC) analysis on large datasets with or without data dependency, both observed genuine scores and impostor scores were all discrete and pre-generated by a classifier for decision making, and usually did not have well defined parametric distributions but nonparametric distributions. Scores might be integers, or real numbers, etc. Moreover, all statistics of interest were based on cumulative probabilities rather than probability densities. Thereafter, to estimate the SEs and 95% CIs of any statistics of interest, the nonparametric two-sample (two-layer if data dependency was involved) bootstrap algorithms were implemented in the light of our prior rigorous statistical research concerning the corresponding data structure and bootstrap algorithms [5-7, 10-16].

Specifically, the validation study was carried out to show that the nonparametric two-sample bootstrap algorithm could be applied to computing the SEs and 95% CIs of any measures in ROC analysis on large datasets in areas such as biometrics, speaker recognition, etc., when the analytical method cannot be used [14]. Therefore, the bootstrap algorithms could also be applied to estimating the SEs of the BF and WoE. The validation was conducted by computing the SEs of the area under ROC curve (AUC) using the well-established analytical Mann–Whitney statistic method and also using the bootstrap method. The analytical result is unique. The bootstrap results are expressed as a probability distribution due to its stochastic nature. The comparisons were carried out using relative errors and hypothesis testing. It was found that these two results matched very well. Such a validation study provides a sound foundation for the applications of the bootstrap algorithms in ROC analysis.

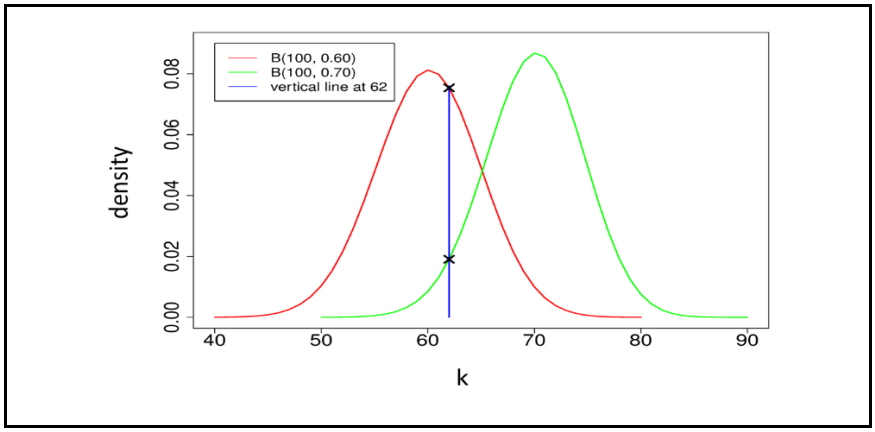

Figure 1 A case study: hypotheses H_1 and H_2 are the two expected binomial distributions B (100, 0.60) (red) and B (100, 0.70) (green), respectively, and the input data (i.e., the evidence) is at 62 (blue).

Certainly, the issue regarding the BF and WoE with single point hypothesis is quite different from those encountered in ROC analysis. If two competing hypotheses H_1 and H_2 are dealt with, the BF is defined to be the ratio of two probability densities $p(E \mid H_1) / p(E \mid H_2)$ (see Section 2), i.e., the ratio of the two conditional probabilities under these two hypotheses based on the input data, i.e., the evidence E. Here two binomial probability distributions may be involved (see Sections 2 and 3) [2-4].

All these can be illustrated by a case study depicted in Figure 1, in which the binomial distribution B (n, p) with two parameters is assumed, that is, n Bernoulli trials take place with success probability p for each trial [3-4]. In Figure 1, hypotheses H_1 and H_2 are the two expected binomial distributions B (100, 0.60) (red, i.e., the left curve) and B (100, 0.70) (green, i.e., the right curve), respectively; and the input data, i.e., the evidence E, is at 62 (blue). Therefore, the BF, i.e., $p(E \mid H_1) / p(E \mid H_2)$, is the ratio of the probability density of B (100, 0.60) at 62 to the probability density of B (100, 0.70) at 62, which is 3.9534. Thereafter the corresponding WoE is 0.5970.

Unlike ROC analysis stated above, here the two distributions are all well-defined discrete parametric probability distributions, namely, two binomial distributions. Binomial variates are only distributed at certain integers, and the probability densities of an expected binomial distribution can be computed exactly using the analytical formulas. However, the two sets of binomial variates (corresponding to the two sets of similarity scores in ROC analysis) are not provided before the bootstrap method is carried out.

Therefore, to statistically estimate the SEs and 95% CIs of BF and WoE using the bootstrap algorithm [10-11], three challenging issues arise. 1. How are the observed binomial variates generated to form a distribution given the total number of trials (n) and the success probability (p)? 2. How is the sample size, i.e., the number of variates, determined so that (1) the observed binomial distribution can match the corresponding expected binomial distribution and (2) it is appropriate for the applications of bootstrap algorithms as far as the sampling variability is concerned? 3. How are the bootstrap algorithms implemented?

To generate observed binomial variates, two methods are carried out: 1) the stochastic function-call method (see Section 3.2), and 2) the deterministic partition method (see Section 3.3). The function-call method is to generate a set of observed random binomial variates by calling function rbinom in the R Stats Package [17-18]. Different calls generate different sets of observed binomial variates. Thus, this method is a stochastic process. The partition method is to create observed binomial variates based on the expected binomial probability densities that can be computed analytically or obtained by calling function dbinom in the R Stats Package [17-18]. This method creates only one set of binomial variates that form an observed binomial distribution based on a set of given binomial parameters (see Section 3.1). Hence, this method is a deterministic process.

Certainly, to ensure the computation accuracy, any observed binomial distribution generated by either the function-call method or the partition method must match the corresponding expected binomial distribution in which the binomial probability densities are known. Thus, the total number of observed binomial variates, i.e., the sample size, is determined via the metric of the root-mean-square deviation (RMSD) between the observed and expected binomial distributions (see Section 4).

Moreover, while applying the bootstrap algorithms, like the bootstrap resampling variability, the sampling variability is also very important. As is known, the bootstrap resampling variability is because a limited number of bootstrap replications of a statistic of interest only constitute a subset of all possible bootstrap replications, which has been studied in Refs. [7, 11, 13-16, 19-20]; and the bootstrap sampling variability is because a finite number of samples usually form a subset of the entire population, which has not been studied yet. Therefore, in this article, the sample size is also determined by the coefficients of variation (CV) for some statistics of interest while implementing bootstrap algorithms and carrying out the bootstrap variability study, which is caused by sampling (see Section 6).

In this article, to statistically estimate the SE and 95% CI of BF and WoE, both the parametric and the nonparametric bootstrap algorithms are employed, respectively [10-11]. The parametric two-sample bootstrap algorithm is conducted on different sets of observed random binomial variates, each of which is stochastically generated using the function-call method and constitutes the corresponding observed binomial distribution with the same parameters n and p (see above). The nonparametric two-sample bootstrap algorithm is carried out on different sets of observed random binomial variates, and variates in each set are randomly selected with replacement (WR) from the same set of observed binomial variates, that is deterministically created using the partition method and forms only one observed binomial distribution with the same parameters n and p.

The BF and thus the WoE are defined in Section 2. In Section 3, to generate observed binomial variates which simulate the expected binomial distribution, two methods, i.e., the function-call method and the partition method, are explored. In Section 4, the sample size is determined by the metric RMSD between the observed and expected binomial distributions. In Section 5, explored are the parametric two-sample bootstrap algorithm carried out on observed binomial variates generated using the function-call method, and the nonparametric two-sample bootstrap algorithm conducted on those created using the partition method. In Section 6, the sample size is also investigated via bootstrap variability study. In this article, the computational part of a case study based on Refs. [3-4][1] as shown in Figure 1 is mentioned in Section 7; for other cases, the same procedure employed in this article can be applied.

2. The BF and the WoE

Let H_1 and H_2 denote two competing hypotheses. For instance, H_1 is the prosecutor's hypothesis and H_2 is the defense lawyer's hypothesis, etc. Each of them gives a prior marginal probability $p(H_1)$ and $p(H_2)$, respectively. Based on the input data, i.e., the evidence E, the prior conditional probabilities under each hypothesis are $p(E \mid H_1)$ and $p(E \mid H_2)$. With their posterior conditional probabilities $p(H_1 \mid E)$ and $p(H_2 \mid E)$, respectively, along with the posterior marginal probability $p(E)$, it holds true that $p(H_1 \mid E) \times p(E) = p(E \mid H_1) \times p(H_1)$ and $p(H_2 \mid E) \times p(E) = p(E \mid H_2) \times p(H_2)$ [3-4, 21]. Thus, it follows

[1] Specific hardware and commercial and non-commercial software products identified in this paper were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

$$\frac{p\,(H_1 \mid E)}{p\,(H_2 \mid E)} = \frac{p\,(E \mid H_1)}{p\,(E \mid H_2)} \times \frac{p\,(H_1)}{p\,(H_2)}\ . \tag{1}$$

In Eq. (1), $p\,(H_1) / p\,(H_2)$ is the prior odds, and $p\,(H_1 \mid E) / p\,(H_2 \mid E)$ is the posterior odds. Hence, the BF of H_1 and H_2 and the corresponding WoE are defined to be [1-4], respectively,

$$BF\,(H_1, H_2) = \frac{p\,(E \mid H_1)}{p\,(E \mid H_2)}\ ;\ WoE\,(H_1, H_2) = \log_{10}\,[BF\,(H_1, H_2)]. \tag{2}$$

The BF is the ratio of the two conditional probabilities under the two hypotheses H_1 and H_2 based on the input data, i.e., the evidence E, which transforms the prior odds to the posterior odds as shown in Eq. (1). The WoE is defined to be the logarithm of the BF with base 10 and is employed in applications such as forensic science, etc. The value of BF (H_1, H_2) greater than 1, i.e., WoE (H_1, H_2) greater than zero indicates that H_1 is more strongly supported by the evidence E under consideration than H_2 [1-4]. In this article, for illustration, two binomial distributions are assumed.

3. Generate observed binomial variates

3.1. Binomial distribution

In practice with single point hypothesis as illustrated in Section 1, the probabilities involved in BF and WoE are two binomial distributions [1-4]. As stated in Section 1, the first critical issue about investigating the SE of BF and thus WoE is how to generate an observed discrete binomial distribution that simulates the corresponding expected binomial distribution as accurately as possible. Let $n \in \{\text{positive integers}\}$ denote the total number of trials, $p \in [0, 1]$ represent the probability of success on each trial, and $k \in \{0, \ldots, n\}$ stand for the number of successes.

The probability densities $d_k\,(n, p)$ at any integer $0 \le k \le n$ of an expected binomial distribution B (n, p) with given n and p can be computed exactly using the analytical formula

$$d_k\,(n, p) = \binom{n}{k}\, p^k\,(1-p)^{n-k} \tag{3}$$

which is normalized to 1 with respect to integer k. Notice that the probability densities d_k (n, p) are only distributed at integers k between [0, n]. Thereafter, the expected binomial distribution with n and p is expressed as

$$B\,(n, p) = \{d_k\,(n, p) \mid k \in \{0, \ldots, n\} \text{ and } \textstyle\sum_{k=0}^{n} d_k\,(n, p) = 1\}. \tag{4}$$

On the other hand, an observed binomial distribution with n and p constituted by N binomial variates is expressed by

$$\widetilde{B}\,(N, n, p) = \{M_k\,(N, n, p) / N \mid k \in \{0, \ldots, n\} \text{ and } \textstyle\sum_{k=0}^{n} M_k\,(N, n, p) = N\}, \tag{5}$$

where M_k (N, n, p) denotes the frequency of such an observed binomial distribution at k. In other words, there are M_k (N, n, p) observed binomial variates at $k \in \{0, \ldots, n\}$, which are equal to k. So, M_k (N, n, p) / N is the probability of such an observed binomial distribution at a given k.

Indeed, it is impossible to create N random binomial variates to form an observed binomial distribution $\widetilde{B}$ (N, n, p), which can exactly match an expected binomial distribution B (n, p). In this article, two methods of generating binomial variates are explored, which are the stochastic function-call method and the deterministic partition method.

3.2. The stochastic function-call method

One way of generating N random binomial variates is to call the function rbinom (N, n, p) in the R Stats Package [17-18]. They form an observed binomial distribution $\widetilde{B}$ (N, n, p) in Eq. (5), where the frequency M_k (N, n, p) at k is the total number of ks in such a set of N random binomial variates. Each call rbinom (N, n, p) in the R Stats Package may produce different set of N random binomial variates. As a result, the function-call method is a stochastic process for a fixed set of N, n, and p.

3.3. The deterministic partition method

The alternative way is to partition a preset total number of binomial variates $\widetilde{N}$ to each number of success k as follows,

$$\widetilde{N} = \widetilde{N} \times \sum_{k=0}^{n} d_k\,(n, p) \tag{6}$$

thanks to the normalization of the binomial probability densities d_k (n, p) with respect to k for the expected binomial distribution B (n, p) (see Eq. (3)).

Such a probability density d_k (n, p) at k can be calculated exactly using Eq. (3) or estimated accurately by calling the function dbinom (k, n, p) in the R Stats Package [17-18]. In practice, the latter is adopted. As always $\widetilde{N} \times$ dbinom (k, n, p) ends up with a real number, but it can be uniquely converted to an integer by "rounding half away from zero." This unique integer is then assigned to the frequency of an observed binomial distribution at k, denoted by M_k ($\widetilde{N}$, n, p), which is certainly determined by the preset total number of binomial variates $\widetilde{N}$. Hence, the resultant total number of observed binomial variates N can be expressed by

$$N = \sum_{k=0}^{n} M_k\,(\widetilde{N}, n, p)\,. \tag{7}$$

Here, M_k ($\widetilde{N}$, n, p) and thus N can all be determined uniquely. In other words, if $\widetilde{N}$, n and p are assumed, then all N observed binomial variates distributed in [0, n] are determined uniquely. As a result, this partition method is a deterministic process for a fixed set of $\widetilde{N}$, n, and p.

Then, the question is how accurate such an observed binomial distribution formed by N binomial variates generated using the above deterministic partition method is with respect to the corresponding expected binomial distribution with the same n and p.

The resultant total number of observed binomial variants N is very close to the preset total number of observed binomial variants $\widetilde{N}$. After the observed binomial distribution is generated using the partition method, the resultant total number of observed binomial variates N rather than the preset total number $\widetilde{N}$ is used to characterize the observed binomial distribution. Thus, for the sake of convenience (see Section 5.2), the resultant total number N will be used in the notations of both the observed frequency M_k (N, n, p) at k and the observed binomial distribution $\widetilde{B}$ (N, n, p) created using the partition method.

From here on, in all computations wherever binomial variates are generated using the partition method, the resultant total number of observed binomial variants N will be employed, for instance, in the computation of RMSD (see Section 4), in the nonparametric two-sample bootstrap algorithm to estimate SEs (see Section 5.2), and in the bootstrap variability study to determine the total number of observed binomial variates (see Section 6).

4. Determine the total number of observed binomial variates using RMSD

As stated in Section 1, the second challenging issue is: How many observed binomial variates are needed to ensure the computation accuracies of the SE of BF and WoE? This issue is investigated in two respects. One is examining the discrepancy between the observed binomial distribution, which is generated using either the function-call method or the partition method, and the expected binomial distribution; and the other is investigating the related bootstrap variability. The former is explored in this section, and the latter is studied in Section 6.

To determine the statistical significance of the discrepancies between the observed and expected binomial distributions, the Kolmogorov-Smirnov test is too sensitive, because the binomial distribution is discrete and contains ties (i.e., repeated variates) [22]. On the other hand, in our case the chi-squared test for goodness of fit is not sensitive enough to differentiate between the observed and expected binomial distributions.

Hence, to this end, the following metric RMSD is employed,

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{k=0}^{n}[M_k\,(N, n, p) \,/\, N - \, d_k\,(n, p)]^2}{n+1}}\,, \tag{8}$$

where d_k (n, p) is the density of the expected binomial distribution B (n, p) at k, which can be in practice obtained by calling the function dbinom (k, n, p) in the R Stats Package [17-18]; and M_k (N, n, p) is the frequency at k of the N observed binomial variates, generated either by the function-call method (see Section 3.2) or by the partition method (see Section 3.3).

If binomial variates are created using the function-call method, because it is a stochastic process, the average of 1,000 RMSDs for a set of (N, n, p) is used for comparisons. If they are generated using the partition method, since it is a deterministic approach, the RMSD is determined uniquely by Eq. (8).

5. The two-sample bootstrap algorithm

The third critical issue as stated in Section 1 is how to implement bootstrap algorithms to statistically estimate the SE of BF and WoE. Any observed binomial distribution can also be expressed using its variates other than its frequencies shown in Eq. (5). Such an expression is suitable while describing bootstrap algorithms [5-7]. Hence two observed binomial distributions $\widetilde{B}_i\ (N_i, n_i, p_i)$, i = 1, 2 are denoted by

$$\widetilde{B}_i\ (N_i, n_i, p_i) = \{\alpha_{i\,j}\ (n_i, p_i) \mid j = 1, \ldots, N_i\},\ i = 1, 2, \tag{9}$$

where $\alpha_{i\,j}\ (n_i, p_i)$, j = 1, …, N_i, i = 1, 2, stand for two different sets of observed binomial variates corresponding to N_1, n_1, p_1, and N_2, n_2, p_2, respectively.

Further, regarding the two-sample bootstrap method, there are parametric and nonparametric [10-11]. Concerning how to generate binomial variates to constitute an observed binomial distribution, there are the function-call method (see Section 3.2) and the partition method (see Section 3.3). The parametric two-sample bootstrap algorithm can only be applied to observed distributions of binomial variates generated using the function-call method [10-11], whereas the nonparametric two-sample bootstrap algorithm is better to be carried out on observed distributions of binomial variates created using the partition method due to smaller RMSD.

5.1. The parametric two-sample bootstrap algorithm

To implement the parametric two-sample bootstrap algorithm, the two observed binomial distributions $\widetilde{B}_i\ (N_i, n_i, p_i)$, i = 1, 2, are generated by using the function-call method (see Section 3.2), i.e., calling function rbinom (N_i, n_i, p_i) in the R Stats Package [17-18] to generate N_i random binomial variates for preset n_i and p_i, respectively.

The parametric two-sample bootstrap algorithm is shown as follows [10-11].

Algorithm I (The parametric two-sample bootstrap)

1: **for** j = 1 **to** B **do**
2: call rbinom (N_1, n_1, p_1) in the R Stats Package
to form a new observed binomial distribution $\widetilde{B}_{1\,j}\ (N_1, n_1, p_1)$ with N_1 random variates
3: call rbinom (N_2, n_2, p_2) in the R Stats Package
to form a new observed binomial distribution $\widetilde{B}_{2\,j}\ (N_2, n_2, p_2)$ with N_2 random variates
4: $\widetilde{B}_{1\,j}\ (N_1, n_1, p_1)$ & $\widetilde{B}_{2\,j}\ (N_2, n_2, p_2) \Rightarrow$ statistics $\hat{S}_j^m$, m = 1, 2
5: **end for**
6: $\{\hat{S}_j^m \mid j = 1, \ldots, B\} \Rightarrow \widehat{SE}_B^m$ and $(\hat{Q}_B^m(\alpha/2),\ \hat{Q}_B^m(1 - \alpha/2))$, m = 1, 2
7: **end**

where B is the number of the two-sample bootstrap replications, m = 1 stands for BF and m = 2 represents WoE, and the two statistics of interest are $\hat{S}_j^1 = \widehat{BF}_j$ and $\hat{S}_j^2 = \widehat{WoE}_j$.

As shown from Step 1 to 5, Algorithm I runs B times. In the j-th iteration, N_1 (N_2) random binomial variates for n_1 and p_1 (n_2 and p_2) are generated by calling the function rbinom (N_1, n_1, p_1) (rbinom (N_2, n_2, p_2)) in the R Stats Package to form a new observed binomial distribution $\widetilde{B}_{1\,j}$ (N_1, n_1, p_1) ($\widetilde{B}_{2\,j}$ (N_2, n_2, p_2)), and then at Step 4 from these two new sets of binomial variates the j-th bootstrap replications of statistics of interest, i.e., $\hat{S}_j^1 = \widehat{BF}_j$ and $\hat{S}_j^2 = \widehat{WoE}_j$ are generated.

If the single point hypothesis is of interest as pointed out in Sections 2 and 3.1, the BF is a ratio of two conditional probabilities at the input data (i.e., the evidence) y based on Eq. (2). In the binomial case as illustrated in Section 1, such a conditional probability at a specific success number k can be estimated by dividing the frequency at k in the new observed binomial distribution by the total number of variates. And the WoE is obtained by Eq. (2) accordingly.

Finally, as indicated in Step 6, from the sets $\{\hat{S}_j^m \mid j = 1, \ldots, B\}$, m = 1, 2, the estimator of the SE, denoted by $\widehat{SE}_B^m$, i.e., the sample standard deviation of these B bootstrap replications, and the estimators of the $\alpha/2$ 100% and (1 - $\alpha/2$) 100% quantiles of the distribution of the bootstrap replications, denoted by $\hat{Q}_B^m(\alpha/2)$ and $\hat{Q}_B^m(1 - \alpha/2)$, at the significance level α can be calculated [11]. Definition 2 of quantile in Ref. [23] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus, $(\hat{Q}_B^m(\alpha/2),\ \hat{Q}_B^m(1 - \alpha/2))$ stands for the estimated bootstrap (1 - α) 100% CI. If 95% CI is of interest, then α is set to be 0.05.

Further, based on our extensive Monte Carlo studies of bootstrap variability in ROC analysis on large datasets with or without data dependency, the number of bootstrap replications B is determined to be 2,000 in order to reduce the bootstrap variance and ensure the accuracy of the computation (see Section 6) [7, 11, 13-16, 19-20].

5.2. The nonparametric two-sample bootstrap algorithm

For any set of N, n and p, the observed binomial distribution generated using the partition method has much less RMSD with respect to the corresponding expected binomial distribution than the one created using the function-call method. Moreover, the partition method is a deterministic process, meaning that it generates only one observed binomial distribution for a set of N, n and p as shown in Section 3.3.

Hence, the nonparametric two-sample bootstrap algorithm is carried out to the two observed binomial distributions $\widetilde{B}_1$ (N_1, n_1, p_1) and $\widetilde{B}_2$ (N_2, n_2, p_2) generated using the partition method with the resultant total numbers of binomial variates N_1 and N_2, respectively (see Section 3.3).

The nonparametric two-sample bootstrap algorithm is shown in the following [5-7, 10-11].

Algorithm II (The nonparametric two-sample bootstrap)

1: **for** j = 1 **to** B **do**
2: select N_1 binomial variates randomly WR from $\widetilde{B}_1$ (N_1, n_1, p_1)
to form a new observed binomial distribution $\widetilde{B}_{1\,j}$ (N_1, n_1, p_1)
3: select N_2 binomial variates randomly WR from $\widetilde{B}_2$ (N_2, n_2, p_2)
to form a new observed binomial distribution $\widetilde{B}_{2\,j}$ (N_2, n_2, p_2)
4: $\widetilde{B}_{1\,j}$ (N_1, n_1, p_1) & $\widetilde{B}_{2\,j}$ (N_2, n_2, p_2) ⇒ statistics $\hat{S}_j^m$, m = 1, 2
5: **end for**
6: $\{\hat{S}_j^m \mid j = 1, \ldots, B\} \Rightarrow \widehat{SE}_B^m$ and $(\hat{Q}_B^m(\alpha/2),\ \hat{Q}_B^m(1-\alpha/2))$, m = 1, 2
7: **end**

where WR stands for "with replacement," and m = 1 represents BF and m = 2 stands for WoE. By comparison with Algorithm I, only Steps 2 and 3 are different regarding how the bootstrap samples of binomial variates are generated. Here, in the j-th iteration, N_1 (N_2) binomial variates are randomly selected WR from the original observed binomial distribution $\widetilde{B}_1$ (N_1, n_1, p_1) ($\widetilde{B}_2$ (N_2, n_2, p_2)), that is generated by the partition method, to constitute a new observed binomial distribution $\widetilde{B}_{1\,j}$ (N_1, n_1, p_1) ($\widetilde{B}_{2\,j}$ (N_2, n_2, p_2)). Everything else stays the same.

6. Determine the sample size, i.e., the total number of observed binomial variates via bootstrap variability studies

To ensure the computational accuracies of the SE and 95% CI of BF and WoE, the sample size, i.e., the total number of observed binomial variates is determined not only by the discrepancy between the observed binomial distributions and the expected binomial distributions, but also by the bootstrap variability. The former is investigated using RMSD for both the function-call method and the partition method in Section 4. And the latter is studied in this section.

6.1. The bootstrap variability studies

While employing the parametric and nonparametric two-sample bootstrap algorithms to estimate the SE and the 95% CI of BF and WoE, to reduce the bootstrap variance and ensure the accuracy of computation, the bootstrap variability must be studied, which determines the sample size as well as the number of bootstrap replications.

As pointed out in the literature [10-11, 19-20], the substantial bootstrap variance is caused by the sampling variability and the bootstrap resampling variability. The former is because a finite number of samples usually form a subset of the entire population. The latter is because a limited number of bootstrap replications of a statistic of interest only constitute a subset of all possible bootstrap replications.

Further, the bootstrap variance produces the variance of the SE and the variance of the two bounds of the CI of the bootstrap distribution formed by the bootstrap replications of the statistic. Hence, these variances are functions of the sample size as well as the number of bootstrap replications. Inversely, the sample size and the number of bootstrap replications can be determined from these variances.

Concerning the number of bootstrap replications, based on our extensive Monte Carlo studies of bootstrap resampling variability in ROC analysis on large datasets with or without data dependency, the number of bootstrap replications is determined to be 2,000 (see Section 5) [7, 11, 13-16, 19-20].

Regarding the sample size, i.e., the total number of observed binomial variates, the study of sampling variability is carried out in terms of the six estimated CVs of SE, lower-bound (LB) and upper-bound (UB) of CI, i.e., CVSE, CVLB and CVUB, for BF and WoE, respectively. These six CVs are derived from 500 runs of SEs, LBs and UBs of CIs, using the parametric and nonparametric two-sample bootstrap algorithms on different specific total numbers of observed binomial variates generated by the stochastic function-call method and the deterministic partition method, respectively.

6.2. The sample size derived from the bootstrap variability studies

Concerning the sampling variability, it is explored using the following Algorithm III. As pointed out above, the sample size, i.e., the total number of observed binomial variates is determined not only by using RMSD in Section 4 while generating observed binomial distributions, but also by Algorithm III here while carrying out the bootstrap algorithm.

Algorithm III (Variability study concerning the number of binomial variates while bootstrap)

1: **for** i = 1 **to** L **do**
2: **for** j = 1 **to** B **do**
3: (Algorithm I or II: Step 2 through Step 3)$_{i\,j}$
4: $\widetilde{B}_{1\,i\,j}\,(N_1, n_1, p_1)$ & $\widetilde{B}_{2\,i\,j}\,(N_2, n_2, p_2) \Rightarrow$ statistics $\hat{S}^{m}_{i\,j}$, m = 1, 2
5: **end for**
6: $\{\hat{S}^{m}_{i\,j} \mid j = 1, \dots, B\} \Rightarrow \widehat{SE}^{m}_{B\,i}$ and $(\hat{Q}^{m}_{B\,i}(\alpha/2),\ \hat{Q}^{m}_{B\,i}(1-\alpha/2))$, m = 1, 2
7: **end for**
8: $\{\widehat{SE}^{m}_{B\,i}, \hat{Q}^{m}_{B\,i}(\alpha/2), \hat{Q}^{m}_{B\,i}(1-\alpha/2) \mid i = 1, \dots, L\} \Rightarrow$
$\widehat{CV}^{m}_{B,L}(\kappa), \kappa = \mathbf{SE}^{m}_{B,L}, \mathbf{Q}^{m}_{B,L}(\alpha/2), \mathbf{Q}^{m}_{B,L}(1-\alpha/2)$ m = 1, 2
9: **end**

where L is the number of Monte Carlo iterations, B is the number of bootstrap replications, and m = 1 stands for BF and m = 2 represents WoE. Step 3 in Algorithm III is equivalent to Step 2 through Step 3 in Algorithms I and II, respectively. Thus, Algorithm III can be applied to the bootstrap variability studies while either the parametric two-sample bootstrap Algorithm I or the nonparametric two-sample bootstrap Algorithm II is carried out.

Moreover, as shown in Sections 5.1 and 5.2, Step 3 in Algorithm III is related to the total numbers of observed binomial variates N_1 and N_2. As a result, all quantities in Steps 4, 6 and 8 of Algorithm III are the functions of N_1 and N_2. To make expressions simpler, N_1 and N_2 are not shown as independent variables here.

As indicated from Steps 1 to 7, Algorithm III runs L iterations for a specified B. The part from Steps 2 to 6 of Algorithm III is equivalent to Algorithms I and II, respectively, which generates the i-th $\widehat{\mathrm{SE}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}}$, $\widehat{\mathrm{Q}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}}(\alpha/2)$, and $\widehat{\mathrm{Q}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}}(1-\alpha/2)$ of BF (i.e., m = 1) and WoE (i.e., m = 2) in the i-th iteration for a given B.

As shown in Step 8, for a specified B, after L iterations of executing the two-sample bootstrap algorithm, the following six sets are generated,

$$\begin{aligned} \mathbf{SE}^{\mathrm{m}}_{\mathrm{B,L}} &= \{\widehat{\mathrm{SE}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}} \,|\, \mathrm{i}=1,\dots,\mathrm{L}\}, \\ \mathbf{Q}^{\mathrm{m}}_{\mathrm{B,L}}(\alpha/2) &= \{\widehat{\mathrm{Q}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}}(\alpha/2), |\, \mathrm{i}=1,\dots,\mathrm{L}\}, \qquad \mathrm{m}=1,2. \\ \mathbf{Q}^{\mathrm{m}}_{\mathrm{B,L}}(1-\alpha/2) &= \{\widehat{\mathrm{Q}}^{\mathrm{m}}_{\mathrm{B}\,\mathrm{i}}(1-\alpha/2), |\, \mathrm{i}=1,\dots,\mathrm{L}\}, \end{aligned} \tag{10}$$

Thereafter, from these six sets, the six estimated coefficients of variation $\widehat{\mathrm{CV}}$s of SE, LB and UB of CI, i.e., CVSE, CVLB and CVUB, for BF and WoE, respectively, can be obtained,

$$\widehat{\mathrm{CV}}^{\mathrm{m}}_{\mathrm{B,L}}(\kappa) = \frac{\sqrt{\widehat{\mathrm{VAR}}^{\mathrm{m}}_{\mathrm{B,L}}(\kappa)}}{\widehat{\mathrm{E}}^{\mathrm{m}}_{\mathrm{B,L}}}, \text{where } \kappa = \mathbf{SE}^{\mathrm{m}}_{\mathrm{B,L}}, \mathbf{Q}^{\mathrm{m}}_{\mathrm{B,L}}(\alpha/2), \mathbf{Q}^{\mathrm{m}}_{\mathrm{B,L}}(1-\alpha/2), \mathrm{m}=1,2. \tag{11}$$

It is clear that the estimated $\widehat{\mathrm{CV}}$s are functions of the number of bootstrap replications B, the number of Monte Carlo iterations L, the significance level α, and the total numbers of observed binomial variates N_1 and N_2.

In this study, L is set to be 500 and B is set to be 2,000 based on our prior investigation in Refs. [7, 13-16]; and α is set to be 0.05 if the 95% CI is of interest as stated in Section 5.1.

Hence, the total numbers of variates of the two binomial distributions N_1 and N_2 can be determined by the tolerable CVs. Indeed, N_1 and N_2 may very well be different. Because such a bootstrap variability study takes substantial CPU time, N_1 and N_2 are assumed to be equal in this article.

7. The computational part of a case study

The theoretical framework of estimating the SE of WoE was accomplished in this research paper. We are working on the computational part. After that, we can compute the SE and 95% CI of WoE in a case study using the parametric two-sample bootstrap algorithm as well as the nonparametric two-sample bootstrap algorithm, respectively. For other cases, the same procedure employed in this article can be applied.

References

[1] I.J. Good. *Probability and the Weighing of Evidence*. Charles Griffin & Co. Ltd, London, 1950.

[2] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, *Bayesian Data Analysis*. Third Edition. Chapman & Hall, New York, 2014.

[3] R.D. Morey, "What is a Bayes factor?" in R bloggers, 2014. [Online]. Available: https://www.r-bloggers.com/2014/02/what-is-a-bayes-factor/

[4] Technical Colloquium on the Weight of Evidence at the National Institute of Standards and Technology, USA, June 27-29, 2017. [Online]. Available: https://www.nist.gov/news-events/events/2017/06/technical-colloquium-weight-evidence
[5] J.C. Wu, M. Halter, R.N. Kacker, J.T. Elliott, and A.L. Plant, "A novel measure and significance testing in data analysis of cell image segmentation," *BMC Bioinformatics* 18: 168:1-13, 2017.
[6] J.C. Wu, A.F. Martin, C.S. Greenberg, and R.N. Kacker, "The impact of data dependence on speaker recognition evaluation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 5-18, 2017.
[7] J.C. Wu, A.F. Martin, and R.N. Kacker, "Measures, uncertainties, and significance test in operational ROC analysis," *J. Res. Natl. Inst. Stand. Technol*., vol. 116, no. 1, pp. 517–537, 2011.
[8] J.C. Wu, A.F. Martin, C.S. Greenberg, R.N. Kacker, and V.M. Stanford, "Significance test with data dependency in speaker recognition evaluation," in *Active and Passive Signatures IV*, Proc. SPIE 8734, 87340I, 2013.
[9] S.P. Lund and H. Iyer, "Likelihood Ratio as Weight of Forensic Evidence: A Closer Look," *J. Res. Natl. Inst. Stand. Technol*., vol. 122, no. 27, pp. 1–32, 2017.
[10] B. Efron, "Bootstrap methods: Another look at the Jackknife," *Ann. Statistics*, vol. 7, no. 1, pp. 1-26, 1979.
[11] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
[12] J.C. Wu and C.L. Wilson, "Nonparametric analysis of fingerprint data on large data sets," Pattern Recognition, vol. 40, no. 9, pp. 2574-2584, 2007.
[13] J.C. Wu, A.F. Martin, and R.N. Kacker, "Monte Carlo studies of bootstrap variability in ROC analysis with data dependency," *Communications in Statistics – Simulation and Computation*, vol. 48, no. 2, pp. 317-333, 2019.
[14] J.C. Wu, A.F. Martin, and R.N. Kacker, "Validation of nonparametric two-sample bootstrap in ROC analysis on large datasets," *Communications in Statistics – Simulation and Computation*, vol. 45, no. 5, pp. 1689-1703, 2016.
[15] J.C. Wu, A.F. Martin, and R.N. Kacker, "Bootstrap variability studies in ROC analysis on large datasets," *Communications in Statistics – Simulation and Computation*, vol. 43, no. 1, pp. 225–236, 2014.
[16] J.C. Wu, "Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap," *NISTIR 7449*, National Institute of Standards and Technology, Sep. 2007.
[17] V. Kachitvichyanukul and B. W. Schmeiser, "Binomial random variate generation," *Communications of the ACM*, vol. 31, no. 2, pp. 216–222, 1988.
[18] R: A Language and Environment for Statistical Computing, the R Development Core Team, the R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.r-project.org/
[19] P. Hall, "On the number of bootstrap simulations required to construct a confidence interval," *Ann. Statistics*, vol. 14, no. 4, pp. 1453-1462, 1986.
[20] B. Efron, "Better bootstrap confidence intervals," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 171-185, 1987.
[21] B. Ostle and L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, 4th ed. Iowa State University Press, Ames, 1988.

[22] J.D. Gibbons, *Nonparametric Statistical Inference*, CRC Press, Boca Raton, FL, 2020.
[23] R.J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *American Statistician*, vol. 50, pp. 361-365, 1996.