Statistics and Application 统计学与应用, 2022, 11(5), 1264-1269
Published Online October 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.115131

Hans汉斯

Bayes Bootstrap改进方法

赵　燕

长安大学理学院，陕西　西安

收稿日期：2022年9月26日；录用日期：2022年10月24日；发布日期：2022年10月31日

摘　要

小样本情况下实验数据的分布较难确定，工程上常采用Bootstrap和Bayes Bootstrap方法。在现有的文献中，该方法对小样本可靠性参数估计仅仅是重复利用原样本信息，通过扩大样本容量进行参数估计。在样本量较小的情况下，再生样本极易淹没原生样本信息导致估计偏差。本文在原方法的基础上，提出对Bayes Bootstrap方法的改进意见，在抽样过程中增加样本容量并通过对最大(最小)次序统计量领域进行扩充而达到对原始样本扩充的目的，最后用指数分布修正经验分布函数，以提高估计的精度。实验结果表明，改进后的Bayes Bootstrap方法对精度的估计有所提高，比原方法效果更好。

关键词

小样本，Bayes Bootstrap方法，点估计

A Improvement Method of Bayes Bootstrap

Yan Zhao

College of Science, Chang'an University, Xi'an Shaanxi

Received: Sep. 26th, 2022; accepted: Oct. 24th, 2022; published: Oct. 31st, 2022

Abstract

It is difficult to determine the distribution of experimental data under the condition of small samples. Bootstrap and Bayes Bootstrap methods are often used in engineering. In the existing literature, this method only reuses the original sample information to estimate the reliability parameters of small samples by expanding the sample size. When the sample size is small, it is easy for the regenerated sample to drown the original sample information, resulting in the estimation deviation. On the basis of the original method, this paper proposes suggestions on improving Bayes Bootstrap. In the sampling process, the sample size is increased, and the purpose of expanding the original sample is achieved by expanding the maximum (minimum) order statistics. Finally, the empirical distribution function is modified by exponential distribution to improve the estimation

文章引用: 赵燕. Bayes Bootstrap 改进方法[J]. 统计学与应用, 2022, 11(5): 1264-1269.
DOI: 10.12677/sa.2022.115131

accuracy. The experimental results show that the improved Bayes Bootstrap method has better accuracy estimation than the original method.

Keywords

Small Sample, Bayes Bootstrap Method, Point Estimation

1. 引言

1977 年由美国统计学教授 Efron [1]在前人的基础上提出了增广样本的新统计方法——Bootstrap 方法。Bootstrap 方法是仅利用验前数据确立验前分布，通过增广样本来对总体分布做出推断。近年来，国内外在 Bootstrap 方法用于小样本试验评估方面开展了许多工作，针对传统 Bootstrap 方法在产生随机样本方面的不足，刘建，吴翊等[2]提出了改进的 Bootstrap 方法，对自助样本的生成范围做了拓展，提高了估计精度。孙慧玲，胡伟文[3]将插值法应用到 Bayes Bootstrap 中，相对于较小样本量来说估计精度更高。曹军海，杜海东等[4]用指数分布函数拟合经验分布函数右尾部样本，增加了随机样本的随机性，扩大了其取值范围。根据研究前人文献，发现目前已经有了不少处理小样本问题的方法。根据有无先验信息，大致可以分成两类。一类是传统的 Bayes 方法，利用大量的先验信息就可获得小样本信息比较正确的估计。另一类是以 Bootstrap 和 Bayes Bootstrap 为代表的方法[5] [6]。该方法不依赖于先验信息，在样本量较小的情况下，利用验前样本实验数据确定验前分布，仅利用当前样本信息就可对总体参数做出估计。Bootstrap 和 Bayes Bootstrap 均是重复利用原始样本信息。由此，本文对 Bayes Bootstrap 做出改进，在抽样过程中增加样本容量并通过对最大(最小)次序统计量领域扩充达到对原始样本量的有效扩充，最后用指数分布修正经验分布函数来提高估计的精度。

2. Bayes Bootstrap 方法基本思想

Bayes Bootstrap 方法是对 Bootstrap 方法生成样本数据做了改进，极大地降低了生成样本数据与原样本数据的重复率[7]。

设 $X=(x_1,x_2,\cdots,x_n)$ 为来自未知总体的观测样本，$x_i \sim F(x), i=1,2,\cdots,n$，$F(x)$ 未知，样本容量是有限的，称其为原生样本。将 $x_1,x_2,\cdots,x_n$ 按照由大到小排序后仍记为 $x_1,x_2,\cdots,x_n$，此时 $x_1,x_2,\cdots,x_n$ 是顺序统计量，则由 $X=(x_1,x_2,\cdots,x_n)$ 构造的经验累积分布函数为：

$$F_n(x)=\begin{cases}0, & x\le x_1\\ \dfrac{k}{n}, & x_k\le x\le x_{k+1}\\ 1, & x\ge x_n\end{cases} \tag{1.1}$$

Bayes Bootstrap 方法的步骤为：

从 $F_n(x)$ 中抽取 N 个样本，基于仿真生成服从经验累积分布函数的随机抽样原理如下：

1) 在区间 $(0,1)$ 上产生分布均匀的随机数 $\eta_i, i=1,2,\cdots,n$ 构成向量 $\eta=(\eta_1,\eta_2,\cdots,\eta_n)$；

2) 令 $\beta=(n-1)\eta, i=[\beta]+1$，$[\beta]$ 为向下取整；

3) 令 $x_i^*=x_i+(\beta-i+1)(x_{i+1}-x_i)$，$i=1,2,\cdots,n-1$，则 $X^*=(x_1^*,x_2^*,\cdots,x_n^*)$ 为抽取得到的再生样本。

4) 计算再生样本均值 $\hat{\theta}_\mu$ 和方差 $\hat{\theta}_s$，用 $\hat{\theta}_\mu = \frac{1}{N}\sum_{i=1}^{N} X_i^*$ 用来估计原生样本均值 θ_μ，用 $\hat{\theta}_s = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^* - \hat{\theta}_\mu\right)^2$ 估计原生样本方差 θ_s。

5) 重复上述步骤，直达得到 4)的稳定值，得到的结果作为原生样本均值 θ_μ 和方差 θ_s 的最终估计值。

根据 Bayes Bootstrap 方法的步骤不难看出，这种方法对原生样本和 Dirichlet 分布的依赖性较大，且这种方法并没有增加任何样本以外的信息。所得样本其实是对原样本的加权平均，权值来源于(0, 1)之间的随机数。若生成随机数在(0, 1)之间均匀性不好，就会导致实验结果出现较大偏差[8]。鉴于以上不足，相关专家提出的改进意见大可分为两类：一类是对抽样方法进行改进[9]，目的在于扩大样本容量，提高估计精度。另一类是对经验分布函数进行修正[10]，构造更为合理的经验分布函数。笔者通过结合这两种方法，对 Bayes Bootstrap 方法进行改进。Bayes Bootstrap 方法的改进在样本量较小的情况下，通过 Bayes Bootstrap 方法确实可以扩大样本容量，但由于生成的再生样本数据过于集中，会造成有效信息的丢失。经验累积分布函数在拟合头部和尾部时效果不好。因此，笔者对 Bayes Bootstrap 方法提出了改进意见，首先对原生样本进行分组扩充，将扩充后的样本合并作为再生样本，构造经验分布函数并用指数分布修正尾部，运用 Bayes Bootstrap 方法对再生样本求点估计和区间估计。

3. 改进的 Bayes Bootstrap 方法

假设 $X = (x_1, x_2, \cdots, x_n)$ 是来自总体的按照时间顺序的简单随机小样本，按照时间先后顺序将 n 个数据分为 K 组，每组数据长度为 h。记 $B_1 = (x_1, x_2, \cdots, x_h)$，…，$B_k = (x_K, \cdots, x_n)$，其中 $K = n - h + 1$；如果 n 可以整除 h，将 n/h 个数据重新合在一起样本容量大小仍为 n。对每一组样本重组扩充步骤如下：

1) 将 $B_1 = (x_1, x_2, \cdots, x_h)$ 中的数据按从小到大排列，若排列好的数据仍记为 $(x_1, x_2, \cdots, x_h)$，对顺序统计量 $X_i, i = 1, h$ 做如下领域：

$$U_1 = \left[x_1 - (x_2 - x_1)/p, x_1 + (x_2 - x_1)/p\right] \tag{2.1}$$

$$U_h = \left[x_h - (x_h - x_{h-1})/p, x_h + (x_h - x_{h-1})/p\right], \quad p \geq 2 \tag{2.2}$$

2) 在 U_1 的左领域取得 x_0，在 U_h 的右领域取得 x_{h+1}；由此将第一组样本扩充为 $h+2$ 个。

3) 重复步骤 1)，2)，依次将 K 组数据进行扩充，扩充后数据总个数为 $n+2K$ 个。

4) 将 K 组样本合并在一起作为再生样本，按照从小到大重新排序，得到顺序统计量 $X = (x_0, x_1, x_2 \cdots, x_N), N = n + 2K$。

为避免随机样本的最大值受到再生样本范围限制，为弥补该方法的不足，我们用指数分布函数拟合右尾部样本。

1) 在 $N - m$ 个样本之前逐段构造经验分布函数，m 为尾部样本数，通常取 5 以下整数。

2) 尾部样本用指数分布拟合并修正经验分布函数，使得修正后的经验分布均值与原样本一致。修正的经验分布函数为：

$$F_n(x) = \begin{cases} 0 & x < x_1 \\ \dfrac{i}{n} + \dfrac{x - x_i}{n(x_{i+1} - x_i)} & x_i \leq x < x_{i+1}, i = 1, 2, \cdots, N - m \\ 1 - \dfrac{m}{n}\exp\left(-\dfrac{x - x_{n-m}}{\theta}\right) & x_{N-m} \leq x \end{cases} \tag{2.3}$$

式中：$\theta = \frac{1}{m}\left[\frac{x_{N-m}}{2} + \sum_{i=N-m+1}^{n}(x_i - x_{N-m})\right]$。

仿真服从修正的经验分布函数的随机样本的方法如下：

1) 产生$[0,1]$区间均匀分布的随机数η。

2) 若$\eta > 1-m/N$，则$\lambda = x_{N-m} - \theta\ln\left(1-\eta\dfrac{N}{m}\right)$为所求随机数；否则转到 3)；

3) 令$\beta=(n-1)\eta, i=[\beta]+1$。则$\lambda = x_i + (\beta - i + 1)(x_{i+1} - x_i)$为所求随机数。

仿真得到所需自主样本，用样本参数估计总体参数。具体流程图如图 1 所示。

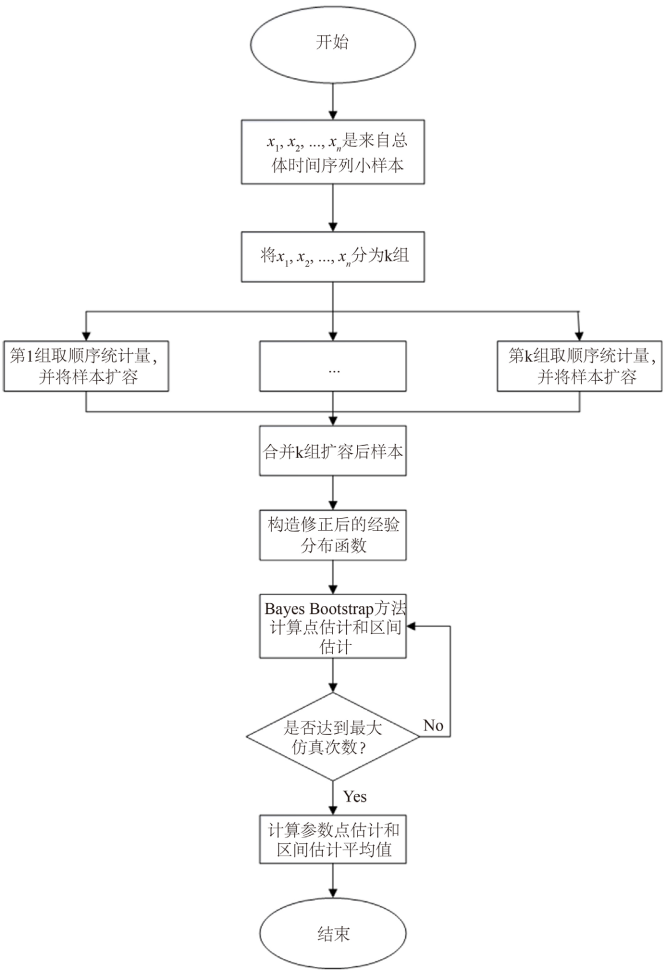

Figure 1. Flow chart
图 1. 流程图

4. 算例

计算机生成服从正态分布$N(100,16)$的 15 个随机数 96.7901，99.8024，98.0155，95.9829，99.7404，98.9578，100.4433，103.2569，101.7745，106.0590，105.5157，93.0443，92.9420，103.6162，102.5938，$n=15$，分别用 Bayes Bootstrap 和改进的 Bayes Bootstrap 方法对参数进行点估计和区间估计。

用经典传统方法计算μ的点估计$\hat{\mu}=99.8357$$\mu$的置信度为 0.95 的置信区间为[97.6023, 102.0690]。用 Bayes Bootstrap 和改进的 Bayes Bootstrap 方法对μ进行点估计和区间估计，分别构造 10,000 组自助样本，μ的参数分布如图 2 和图 3 所示，参数μ的分布形式都近似于正态分布，对多组μ的分布假设检验，在 0.05 的显著性水平上接受μ服从正态分布的假设。三种方法评估结果比较如表 1 所示。

Table 1. There methods were used to compare the results
表 1. 三种方法结果比教

方法	参数点估计			置信区间	
	估计值	期望值	误差	估计值	区间长度
经典统计方法	99.8357	100	0.1643	[97.6023, 102.0690]	4.4667
Bayes Bootstrap 方法	99.9229	100	0.0771	[99.9219, 99.9577]	0.0358
改进 Bayes Bootstrap 方法	99.9903	100	0.0097	[99.9473, 100.0332]	0.0859

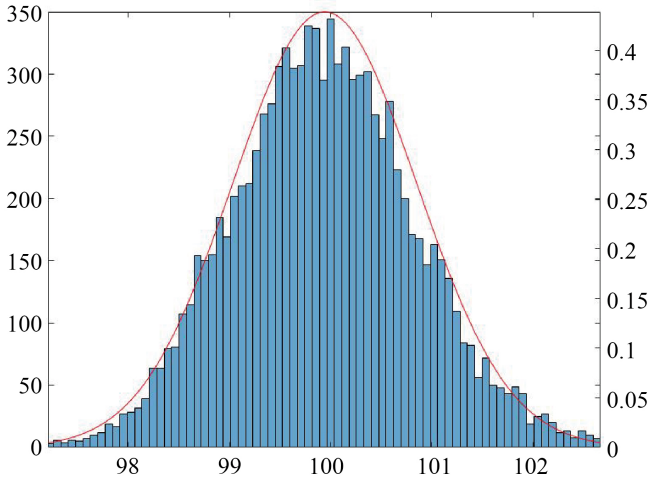

Figure 2. Parameter Distribution Cart
图 2. Bayes Bootstrap 方法的参数分布

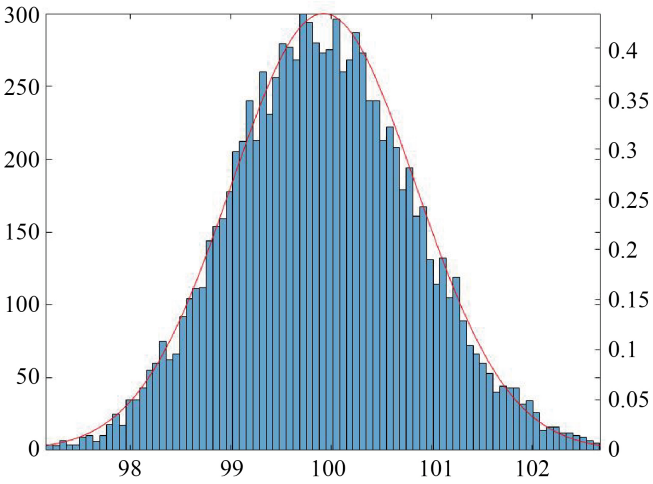

Figure 3. Parameter Distribution Cart of the improved method
图 3. 改进的 Bayes Bootstrap 方法的参数分布

5. 结论

5.1. 改进方法的评价

Bayes Bootstrap 方法对原始数据以及迪利克雷分布的依赖性较大，改进的方法克服了原方法的不足：

第一，将原样本按时间序列分组，在每一组重构顺序统计量并增加了样本量，克服了 Bayes Bootstrap 方法生成的样本数据向中间点集中的趋势[11]。第二，用指数分布函数修正经验分布函数右尾部，将最大顺序统计量延拓到非观测点，降低了再生样本与原样本的相似度。

5.2. 结论

表 1 数据显示，在相同置信度的前提下，改进的 Bayes Bootstrap 方法对 μ 的区间估计与原方法相差不大，但对 μ 的点估计精度明显比原方法更好。三种方法的估计值分别为 99.8357，99.9229，99.9903，误差分别为 0.1643，0.0771，0.0097，置信区间长度分别为 4.4667，0.0358，0.0859。总体来看，传统方法的估计值效果最差，Bayes Bootstrap 方法比传统方法好，而改进 Bayes Bootstrap 方法的估计效果是最好的。这表明本文所提的改进 Bayes Bootstrap 方法对估计值的精度有所提升，为 Bootstrap 方法提供了一种新的改进方法。对于经验分布函数的修正，指数分布函数只拟合了经验分布函数的右尾部，延拓了样本的最大顺序统计量，对于经验分布函数的左尾部并没有延拓。是否存在更合适的分布函数使其对经验分布函数整体拟合效果更好有待进一步研究。

参考文献

[1] Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York, 1-15. https://doi.org/10.1007/978-1-4899-4541-9_1

[2] 刘建, 吴翊, 谭璐. 对 Bootstrap 方法的自助抽样的改进[J]. 数学理论与应用, 2006(1): 69-72.

[3] 孙慧玲, 胡伟文, 刘海涛. 小样本情况下参数区间估计的改进方法[J]. 哈尔滨理工大学学报, 2017, 22(1): 109-113.

[4] 曹军海, 杜海东, 申莹. 基于改进 Bayes-Bootstrap 方法的系统可靠性仿真评估[J]. 装甲兵工程学院学报, 2016, 30(1): 95-98.

[5] 邹艳, 罗文强. 改进的 Bootstrap 方法对比及应用研究[J]. 应用数学, 2008, 21(S1): 62-66.

[6] Grimshaw, G.D. (2012) An Introduction to the Bootstrap. *Technometrics*, **37**, 49-54. https://doi.org/10.2307/1269918

[7] 黄玮, 冯蕴雯, 吕震宙. 基于 Bootstrap 方法的小子样试验评估方法研究[J]. 机械科学与技术, 2006(1): 31-35.

[8] 冯计才, 刘力维, 常宝娴, 张敏. Bootstrap 方法的仿真实现及其在系统偏差估计中的应用[J]. 南京理工大学学报(自然科学版), 2007(3): 399-402.

[9] 谢益辉, 朱钰. Bootstrap 方法的历史发展和前沿研究[J]. 统计与信息论坛, 2008(2): 90-96.

[10] 李磊, 叶友皓, 袁永生. 基于改进 Bayesian Bootstrap 方法的产品性能参数评估[J]. 电子设计工程, 2018, 26(2): 14-17+21.

[11] 孙慧玲, 胡伟文, 刘海涛. 基于插值法的 Bayes Bootstrap 方法的改进[J]. 统计与决策, 2017(9): 74-77.