

小样本情况下参数区间估计的改进方法

孙慧玲, 胡伟文, 刘海涛

(海军工程大学 理学院应用数学系, 湖北 武汉 430033)

摘 要:小样本情况下实验数据的概率分布较难确定,传统小样本估计方法无法提供准确的参数估计;针对工程上常用的 Bayes Bootstrap 方法对小样本可靠性参数估计仅仅是原样本的重复,在参数区间估计上精度不够高的问题;在不改变原样本数据的基础上,依据时间序列将原样本分组并扩充,对扩充后的样本进行参数点估计和区间估计,提出针对小样本情况下参数区间估计的改进方法,给出了改进方法的算法。运用蒙特卡罗仿真方法进行建模仿真,结合具体算例分析,验证新方法对小样本情况下参数的区间估计精度有显著提高。

关键词:小样本; Bayes Bootstrap 方法; 区间估计

DOI: 10.15938/j.jhust.2017.01.019

中图分类号: O211 文献标志码: A 文章编号: 1007-2683(2017)01-0109-05

An Improvement to Interval Estimation for Small Samples

SUN Hui-ling, HU Wei-wen, LIU Hai-tao

(School of Sciences, Naval University of Engineering, Wuhan 430033, China)

Abstract: Because it is difficult and complex to determine the probability distribution of small samples, it is improper to use traditional probability theory to process parameter estimation for small samples. Bayes Bootstrap method is always used in the project. Although, the Bayes Bootstrap method has its own limitation, In this article an improvement is given to the Bayes Bootstrap method, This method extended the amount of samples by numerical simulation without changing the circumstances in a small sample of the original sample. And the new method can give the accurate interval estimation for the small samples. Finally, by using the Monte Carlo simulation to model simulation to the specific small sample problems. The effectiveness and practicability of the Improved-Bootstrap method was proved.

Keywords: small sample; Bayes Bootstrap method; interval estimation

0 引 言

样本容量 $n \leq 30$ 在工程上一般被认为是小样本。如果是正态分布,小样本的样本量界定可能更

小,甚至小于 10。随着高新技术在武器系统中的广泛应用,武器装备是否能保证每次成功完成任务与其可靠性直接相关,因此,可靠性是衡量装备性能的一个重要指标;导致在研制武器装备的过程中,其精度及可靠度要求越来越高;使得技术更复杂,造

收稿日期: 2015-09-19

基金项目: 国家自然科学基金(61074191); 海军工程大学青年基金(HGDQNJ15003); 海军工程大学理学院青年基金(HJGSK2014G125)。

作者简介: 孙慧玲(1982—),女,讲师,E-mail: moon0806sun@163.com;

胡伟文(1964—),男,教授,博士研究生导师;

刘海涛(1982—),男,讲师。

价更昂贵成为整个研制系统的大趋势。特别是某些破坏性试验,一次实验往往要付出巨大的代价^[1]。针对这类试验的传统鉴定方法已不再适用。因为传统鉴定方法是以经典统计理论为基础的^[2],也就意味着较大的样本量必不可少,而昂贵的武器装备从安全以及节约的方向考虑,显然不适合进行大量试验。

从统计学角度分析,武器装备的可靠性研究是参数估计的范畴,是参数估计的具体实例。目前,工程上已经积累了不少方法来处理小样本问题^[3-4],根据有无先验信息这点进行界定,它们大致可以分为两大类:一类是以 Bayes 方法为代表的传统估计方法。该方法仅利用原始积累实验数据也即历史信息来估计参数。另一类是以 Bootstrap 和 Bayes Bootstrap 方法为代表的方法。该方法仅仅利用当前实验数据,在样本量较小的情况下,可以对参数进行比较准确的估计。

本文先介绍 Bayes Bootstrap 方法的基本思想和基本步骤;随后分析该方法的不足之处,针对不足提出改进意见;最后通过具体算例验证改进方法的可行性。

1 小样本参数估计 Bayes Bootstrap 方法

1.1 Bayes Bootstrap 方法的基本步骤

定义1 观测样本 $X = (x_1, x_2, \dots, x_n)$ 为总体样本,其样本量是有限的,称该样本为原生样本,设 $x_i \sim F(x)$, $i = 1, 2, \dots, n$, $F(x)$ 未知,则这些原生样本构造的经验分布函数为

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ k/n & x_{(k)} \leq x < x_{(k+1)} \\ 1 & x \geq x_{(n)} \end{cases} \quad (1)$$

式中: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 是顺序统计量,是按 $x_1, x_2, \dots, x_n$ 从小到大的排序后得到的。

步骤1: 假设 $\theta = \theta(F)$ 是总体的某个参数(例如均值或方差), $\hat{\theta} = \hat{\theta}(F_n)$ 是总体参数 θ 的估计值,记:

$$\hat{\theta}(F_n) - \theta(F) = T_n = R(X, F) \quad (2)$$

T_n 即为 θ 的估计误差,易知, T_n 是随机变量是 X, F 的函数;

步骤2: 从 F_n 中抽取 N 组样本,方法如下:

1) 取自具有 Dirichlet 分布的随机变量 $V_i V_i = U_i - U_{i-1}$ ($i = 1, 2, \dots, n$), 其中, $U_1, U_2, \dots, U_{n-1}$ 为来自

$(0, 1)$ 均匀分布的随机数,令 $U_0 = 0, U_n = 1$ 。则 $V_1 + V_2 + \dots + V_n = 1$;

2) 这 N 组重抽样样本即为: $X^*(j) = (x_1^*, x_2^*, \dots, x_n^*)$, $j = 1, 2, \dots, N$, 此再生样本称为自助样本。

其中, $x_i^* = \sum_{j=1}^n V_j x_{(j)}$; 因此,每进行一次计算机抽样,就可以得到一组 V_i , 相应的就可以得到一个再抽样样本 x_i^* ;

步骤3: 计算下式中的 R_n ,

$$\hat{\theta}(F_n^*) - \hat{\theta}(F_n) = R_n = R^*(X^*, F_n) \quad (3)$$

式中 F_n^* 是自助样本的经验分布函数。易知, R_n 是随机变量是 X^*, F_n 的函数。借助计算机进行多次抽样,可以得到 R_n 的概率分布;

步骤4: 利用式(3)中 R_n 的分布近似式(2)中 T_n 的分布,可得 $\theta(F) \approx \hat{\theta}(F_n) - R_n$, 计算机模拟 N 次抽样(N 可以取 10000 或者更大), 就可以得到 N 个 $\theta(F)$, 未知参数的 θ 的分布、特征值就可用相应的统计方法求出。

重复步骤2~4, 逐步得到所求参数的平稳值。

1.2 Bayes Bootstrap 方法的算法实现

传统方法对未知参数均值 μ 的估计如下:

定理1^[5] 设总体 $X \sim N(\mu, \sigma^2)$, 样本为 $X_1, X_2, \dots, X_n, \bar{X}, (S^*)^2$ 分别为样本均值和样本修正方差, 设已给定置信水平为 $1 - \alpha$, 则 μ 的点估计和置信水平为 $1 - \alpha$ 的置信区间分别为:

$$\hat{\mu} = \bar{X}, (\bar{X} - t_{\frac{\alpha}{2}}(n-1) \frac{S^*}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}(n-1) \frac{S^*}{\sqrt{n}}) \quad (4)$$

Bayes Bootstrap 方法的算法实现:

1) 计算机模拟抽样 N 次, 得到 N 组数据 $R_n(i)$, $i = 1, 2, \dots, n$ 。

2) 算出 $R_n(i)$ 在置信水平为 $1 - \alpha$ 的情况下的最大值 $R_n(\max)$, 最小值 $R_n(\min)$, 均值 $\bar{R}_n$, 满足:

$$\bar{R}_n = \frac{1}{N} \sum_{i=1}^N R_n(i) \quad (5)$$

$$P(R_n(\max) < R_n < R_n(\min)) = 1 - \alpha \quad (6)$$

用 R_n 近似 T_n , 由式(4)可计算出 μ 的置信水平为 $1 - \alpha$ 的置信区间为:

$$(\bar{X} + R_n(\min), \bar{X} + R_n(\max))$$

参数 μ 的估计值为 $\hat{\mu} = \bar{X} + \bar{R}_n$ 。

1.3 Bayes Bootstrap 方法的分析

根据1.1的介绍可知 Bayes Bootstrap 方法没有添加任何样本以外的信息, 仅仅是在原样本基础上的重复抽样, 对样本点进行了一定的修正, 并且扩

大了样本容量对原有参数进行估计。据已有的成果^[4-7],小样本情况下(样本量为10),Bayes Bootstrap方法明显优于经典统计法,不仅在参数点估计更接近真实值,并且得到的估计置信区间更短^[8]。

研究中发现,Bayes Bootstrap方法对Dirichlet分布和原生样本依赖性较大^[9]。另外,Bayes Bootstrap方法的再生样本是取自Dirichlet分布随机数与原样本的加权平均,在(0,1)区间生成一序列的随机数结果有多种可能,一旦生成的随机数均匀性不好就会导致实验结果出现很大差别。鉴于以上局限性,有专家学者对Bayes Bootstrap方法提出了改进意见,一是对经验函数提出改进意见,重新构造更为合理的经验分布函数^[10];二是对小样本的Bootstrap抽样方法进行改进,目的在于调整抽样方法,增大样本容量^[11]。在具体工程问题中,这些改进方法都有较好的适应性^[12]。

为了更好地进行估计,本文进一步提出对Bayes Bootstrap方法的改进措施,即对原生样本按时间序列分组,每组利用插值法增大样本容量,从而对原小样本的数据进行了扩充,将小样本数据变成较小样本数据,这部分内容也是本文的创新点。最后用Bayes Bootstrap思想重抽样对未知参数进行估计,从而在样本量一定的情况下减少误差提高精度。

下面介绍改进方法的基本步骤和算法实现。

2 改进 Bayes Bootstrap 方法介绍

2.1 改进方法的基本步骤

定义2 假设 $X_1, \dots, X_n$ 是来自总体的按时间序列的简单随机小样本,按照样本来源的时间顺序将 n 个数据分为 K 组,每组数据长度为 h , (假设 n 可以整除 h) ,那么, $B_1 = (X_1, \dots, X_h)$, $\dots$, $B_K = (X_K, \dots, X_n)$, 其中 B_i 表示第 i 组数据。这 n/h 个数据组拼接在一起的样本量大小仍为 n 。

拆分之后对这 K 组数据进行重抽样,步骤如下:

步骤1: 将 $B_1 = (X_1, \dots, X_h)$ 中的数据按从小到大的顺序排列,排列好的数据为: $(X_{(1)}, X_{(2)}, \dots, X_{(h)})$, 对顺序统计量 $X_{(i)}$ 的观测值 $x_{(i)}$ 做如下邻域:

$$\begin{cases} U_1 = [x_{(1)} - (x_{(2)} - x_{(1)})/p, x_{(1)} + (x_{(2)} - x_{(1)})/p] \\ U_i = [x_{(i)} - (x_{(i)} - x_{(i-1)})/p, x_{(i)} + (x_{(i+1)} - x_{(i)})/p] \\ i = 2, \dots, n-1 \\ U_h = [x_{(h)} - (x_{(h)} - x_{(h-1)})/p, x_{(h)} + (x_{(h)} - x_{(h-1)})/p] \end{cases} \quad (7)$$

这里 $p \geq 2$ 。

步骤2: 在邻域 $U_1 = [x_{(1)} - (x_{(2)} - x_{(1)})/p, x_{(1)} + (x_{(2)} - x_{(1)})/p]$ 中取得 $x_{(0)}$, 在邻域 $U_h = [x_{(h)} - (x_{(h)} - x_{(h-1)})/p, x_{(h)} + (x_{(h)} - x_{(h-1)})/p]$ 中取得 $x_{(h+1)}$; 由此将第一组样本量扩充为 $h+2$ 个。

步骤3: 重复步骤1~2, 依次将 K 组数据进行样本量扩充, 扩充后样本容量即增加为 $n+2K$ 个。

步骤4: 将 K 组扩充后的样本合并作为再生样本, 运用Bayes Bootstrap方法针对该再生样本求参数的点估计和区间估计。

2.2 改进 Bayes Bootstrap 方法的算法实现

1) 将扩充后样本容量为 $n+2K$ 个样本进行计算机模拟抽样 N 次, 得到 N 组数据 $D_n(i)$, $i=1, 2, \dots, n$ 。

2) 算出 $D_n(i)$ 在置信水平为 $1-\alpha$ 的情况下的最大值 $D_n(\max)$, 最小值 $D_n(\min)$, 均值 $\bar{D}_n$, 满足:

$$\bar{D}_n = \frac{1}{N} \sum_{i=1}^N D_n(i) \quad (8)$$

$$P(D_n(\max) < D_n < D_n(\min)) = 1 - \alpha \quad (9)$$

用 D_n 近似2)中的 T_n , 由式(4)可计算出 μ 的置信水平为 $1-\alpha$ 的置信区间为

$(\bar{X} + D_n(\min), \bar{X} + D_n(\max))$ 。参数 μ 的估计值为 $\hat{\mu} = \bar{X} + \bar{D}_n$ 。

2.3 改进方法计算机建模仿真流程图

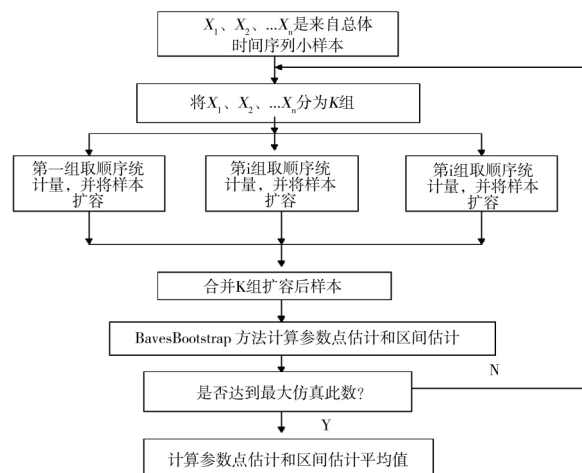


图1 改进方法仿真流程图

3 算例

前面介绍了小样本参数估计的传统方法和Bayes Bootstrap方法, 本文提出了基于Bayes Boot-

strap 方法的改进意见并给出了仿真流程,下面通过具体实例来比较 3 种方法在实际问题中的适应性,验证改进方法的优越性。

例 计算机生成服从正态分布 $N(2, 0.5)$ 的 10 个随机数 1.7837, 1.1672, 2.0627, 2.1438, 1.4268, 2.5955, 2.5946, 1.9812, 2.1636, 2.0873, 取置信度 $1 - \alpha = 0.95$, 分别用传统小样本估计方法、Bayes Bootstrap 方法以及改进 Bayes Bootstrap 方法对参数 μ 作点估计和区间估计。

解: 用传统方法计算,根据式(4)可得 μ 的点估计 $\hat{\mu} = 2.006$, μ 的置信度为 0.95 的置信区间为

$[1.7388, 2.2625]$ 。由于 $n = 10$ 是小样本数据,考虑运用 Bayes Bootstrap 方法和改进 Bayes Bootstrap 方法对 μ 进行估计,方法如下:

构造并产生 $N = 10\,000$ 组自助统计量(可以更大),根据式(5)、(6),运用 Bayes Bootstrap 方法得到参数 μ 的点估计值和区间估计^[10-11](见表 1), μ 的参数分布如图 2 所示。根据改进方法增大样本容量的思想,可将原样本数据分为 2 组,运用式(8)、(9),改进 Bayes Bootstrap 方法得到 μ 的估计值和区间估计(见表 1), μ 的参数分布如图 3 所示。

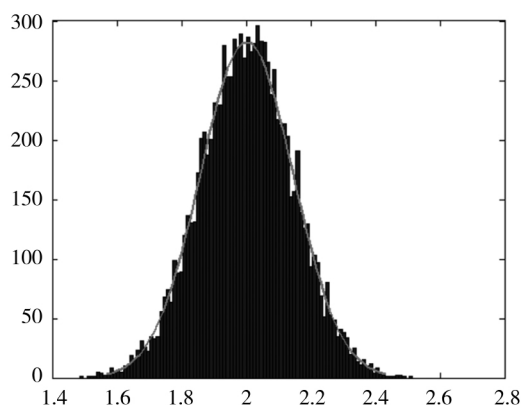


图 2 Bayes Bootstrap 方法的参数分布图

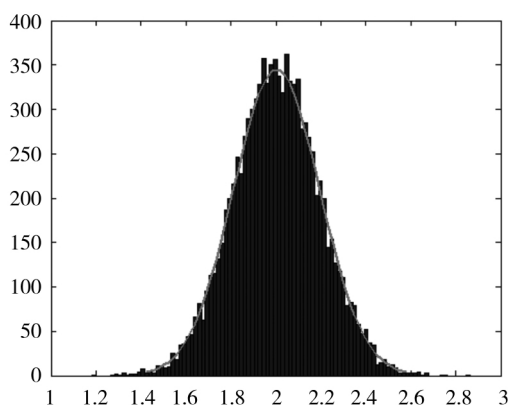


图 3 改进 Bayes Bootstrap 方法的参数分布图

表 1 评估结果比较表

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	2.006 0	2.0	0.006	[1.738 8, 2.262 5]	0.523 7
Bayes Bootstrap 方法	2.003 1	2.0	0.003 1	[1.999 9, 2.006 3]	0.006 4
改进 Bayes Bootstrap 方法	2.002 1	2.0	0.002 1	[1.999 7, 2.004 5]	0.004 8

4 改进方法的评价

鉴于原 Bayes Bootstrap 方法对原始数据及 Dirichlet 分布的依赖性较大,在样本量较小情况下很难得到满意的估计,改进方法在以下方面克服了原方法的不足:第一,先将样本按时间序列分组,在每一组中重构顺序统计量,克服了 Bayes Bootstrap 方法中再生样本数据向中间点集中的趋势;第二,调整了抽样方法从而扩展了样本容量,将每一组的样本容量都进行了扩充,并且将最大最小顺序统计量延拓至非观测点,极大地降低了再生样本与原样本的相似性。

5 结 论

表 1 的数据显示,改进方法对参数 μ 的点估计与原方法相差不大。而在相同置信度的情况下对参数 μ 的区间估计精度明显比 Bayes Bootstrap 方法更好,原因是改进方法对样本的延拓必然增大了样本信息,从理论上讲,在置信度一定的情况下,提高区间估计精度只能依靠增加样本容量^[12],所以,改进方法的实际建模效果与统计学原理也是一致的。

本文并未对参数 σ 进行估计,那么,改进方法对参数 σ 是否也具有良好的适应性还有待进一步研究。

参考文献:

- [1] 唐雪梅, 张金魁. 武器装备小子样实验分析与评估 [M]. 北京: 国防工业出版社, 2001: 6-73.
- [2] 陆璇. 数理统计基础 [M]. 北京: 清华大学出版社, 1998: 86-97.
- [3] 裴鹿成, 工仲奇. 蒙特卡罗方法及其应用 [M]. 海洋出版社, 1998: 5-63.
- [4] 孙锦, 李国林, 许城, 等. 小子样条件下导弹靶场实验综合评价方法 [J]. 火力与指挥控制, 2014(9): 23-27.
- [5] 吕鹏, 袁永生, 相荣霞, 等. 小样本下改进的 PWM 方法及其在导弹精度评定中的应用 [J]. 重庆理工大学学报(自然科学版), 2014(4): 139-142.
- [6] 马智博, 朱建士, 徐迺新, 等. 有关正态分布的小样本可靠性评估 [J]. 核科学与工程, 2003(4): 31-35.
- [7] 刘江, 姚安东, 王攀, 等. 新型弹药可靠性评估方法 [J]. 兵工自动化, 2014(5): 1-3.
- [8] 黄金龙, 汪立新. 基于改进 Bayes Bootstrap 陀螺仪寿命预测研究 [J]. 压电与声光, 2014(6): 384-387.
- [9] 孙慧玲, 胡伟文. 小样本条件下参数估计方法比较研究 [J]. 统计与决策, 2014(12): 1-5.
- [10] 戴邵武, 高华明, 肖支才, 等. 基于自助法的小样本数据分析方法研究 [J]. 海军航空工程学院学报, 2009(1): 26-28.
- [11] 张守玉, 封伟书. 基于 Bootstrap 方法的正态分布样本数据生成研究 [J]. 装备指挥技术学院学报, 2009(20): 97-100.
- [12] 邹艳, 罗文强. 改进 Bootstrap 方法对比及应用研究 [J]. 应用数学, 2008(21): 62-66.
- [13] 孙慧玲. 取定统计量下最优置信区间的估计 [J]. 统计与决策, 2009(7): 32-37.
- [14] 刘海涛, 张志华. 威布尔分布无失效数据的 Bayes 可靠性分析 [J]. 系统工程理论与实践, 2008(11): 103-108.
- [15] BERGER J. The Robust Bayesian Viewpoint in Robustness of Bayesian Analysis [J]. Orth-Holland, Amsterdam: J. Kadane (Ed), 1985: 375-390.
- [16] JAYNES ET. Prior Probability [J]. IEEE Tran. On Systems Science and Cybernetics. SSC4, 1968: 227-241.
- [17] VILLEGAS C. On the Representation of Ignorance. Jour. Amen. Statist. Assoc, 1977, 72: 651-654.
- [18] BRAD Efron, ROB Tibshirani. An Introduction to the Bootstrap [M]. NW: Chapman and Hall(FirstPublished), 1993: 153-165.
- [19] XU Di, ZHAO Wen-biao. Reliability Prediction Using Multivariate Degradation Data [C] //Proc of Annual Reliability and Maintainability Symposium, Institute of Electrical and Electronic Engineers, 2005: 337-341.
- [20] HISAH Tanizaki. Computational Methods in Statistics and Econometrics [M/OL]. Published in the Taylor Francis e-Library, 2005: 479-483.

(编辑: 温泽宇)