文章编号: 0254 - 5357(2007)01 - 0029 - 04

自举法对我国进口铁矿二氧化硅含量的代表值估计

纪 雷, 孙 健, 林雨霏, 杜恒清, 王 岩, 刘心同
(山东出入境检验检疫局技术中心, 山东 青岛 266002)

摘要: 对我国近两年490批次、23个进口国的进口铁矿中SiO_2含量进行了总体统计分析,在数据统计分布特征研究基础上,使用内核密度估计对进口铁矿SiO_2含量进行数据多态性分析,使用自举法对原始数据样本值重复取样以获得稳健的SiO_2含量代表值估计及标准偏差,并证明以自举法重新取样样本分布的均值与标准偏差作为有限单次样本代表值是合理、有效的。

关键词: 铁矿; 二氧化硅; 代表值; 统计学描述; 内核密度估计; 自举法
中图分类号: O212:O613.72 **文献标识码:** A

Estimation for Representative Values of SiO_2 Content in Imported Iron Ores by Bootstrap Method

JI Lei, SUN Jian, LIN Yu-fei, DU Heng-qing, WANG Yan, LIU Xin-tong
(Technical Centre, Shandong Entry/Exit Inspection & Quarantine Bureau, Qingdao 266002, China)

Abstract: Content characteristics of silica dioxide in iron ores imported from 23 countries, totally 490 batches, in recent two years are investigated by statistical methods. Based on the study on data distribution characteristics, a new robust statistics method, kernel density estimation, coupled with bootstrap method is introduced to acquire the representative values of silica dioxide content and the standard deviations for imported iron ores. It is clearly demonstrated that this method shows a prior advantage to give a robust description in explanation for central tendency and variation of data profiles. It has also proved that the conclusion of substitution of bootstrap mean value and its standard deviation for the representative value of silica dioxide content from single specimen analysis is reasonable and effective.

Key words: ion ore; silica dioxide; representative value; statistical description; kernel density estimation; bootstrap

实验结果数据集的代表值估计是实验结果的重要表述特征,是工农业生产及科学实验的数据分析基础。这项工作在冶金、采矿领域及国际矿产贸易中意义尤显重要,在这些领域中代表值的准确估计对冶金工艺、采矿可行性及国际贸易结算都有重要影响。实验结果的代表值估计属基本统计学参数描述,分为参数法和非参数法,通常使用的参数法描述是以数据符合正态分布为前提,依据实验结果的数据分布特征,有多种方法可以对实验结果进行参数描述。一般情况下,对于符合正态分布的实验结果,采用平均值±标准偏差(Means ± SD)的描述体系,对不符合正态分布的实验结果,多采用稳健统计描述,较常见的如四分位稳健统计描述(IQR),这可以较好地克服异常值对结果的影响[1-2]。

我国铁矿资源缺乏,铁矿进口量逐年大幅递增,了解我国进口铁矿中各项质量指标的含量水平,对准确掌握进口铁矿整体质量、及时跟踪铁矿质量变化有着重要意义。本文旨在数据统计分布

收稿日期: 2006-02-27; **修订日期:** 2006-06-10
作者简介: 纪雷(1970 -),男,山东青岛市人,高级工程师,从事分析化学研究。E-mail: jilei70@ yahoo. com. cn。

特征研究基础上,对我国近两年 490 批次、23 个进口国的进口铁矿中的 SiO_2 含量进行了总体统计分析,采用正态概率分布函数对实验数据的正态性进行验证,使用内核密度分布估计对数据的多态性进行分析,并用自举法对多态性分布数据进行多次模拟重复取样,获得稳健的 SiO_2 含量代表值估计及标准误差,同时探讨了这一方法在代表值估计方面的特点。

自举法(Bootstrap)是一种基于对原始样本采用有回放的重新取样的模拟统计方法[3-5],使用自举法对样本总体参数进行估计的最大优势在于不需对被考察样本的总体分布形态做出假设,因此,可以克服通常对样本分布呈非正态所采用的异常值取舍、稳健统计描述,并且可以对目前较难处理的多态性分布的样本的总体分布参数做出较好解释。自举法已在一些关键代表值的估计工作中得到重要应用,主要有实验室能力水平测试中指定值的获得,毒性试验中重要毒理数据的稳健性估计,实验结果重要参数的估计;自举法在这些领域中克服了传统取样理论中的瓶颈,即大规模实验样本获得的可能性,取样成本因素及取样的权威性、严肃性。自举法已逐渐成为在对样本取样有较大难度、甚至不可能重新取样的情况下的标准处理方法。

1　采样与数据分析

1.1　样品来源与测量方法

本次调查内容为我国近两年进口的 490 批次,共计 23 个国家的铁矿样品,代表我国进口铁矿的整体质量水平。取制样标准按 ISO 3082:2000[6],SiO_2 含量测定采用 ISO 2598:1992[7]。

1.2　数据分布特征及多态性分析

采用正态概率分布函数对进口铁矿 SiO_2 含量的实验数据的正态性进行验证,使用内核密度分布估计对数据的多态性进行分析,选择合适的内核密度参数对原始数据样本集的数据多态性进行合理描述,依据数据的多态性分布特征决定是否采用自举法对铁矿 SiO_2 含量代表值进行估计。

1.3　以自举法估计二氧化硅含量代表值

使用自举法对多态性分布数据进行多次模拟重复取样,获得进口铁矿 SiO_2 含量的稳健统计描述,在模拟原始数据样本集大量取样的基础上,获得稳健的 SiO_2 含量代表值估计及标准偏差。

数据基础统计学分析、内核密度分布及相关数据处理均在 Matlab ® 6.1(Release11)下编制、处理,自举法算法程序在 R 语言[8]下编程,在装有 Windows ® 2000 台式 PC 机上调试、运行。

2　结果与讨论

2.1　二氧化硅含量分布特征及多态性分析

对于实验结果数值分布通常采用点分布图或直方图,其存在明显的缺点:对离散分布的点分布图的多态性特性难以确定,无法准确给出数据的多形态分布;直方图更具主观色彩,但得到的直方图分布受计算机给出的坐标比例、分布宽度等因素干扰严重,甚至会出现截然不同的结果。为克服观察方法对数据多态性解释的粗略、不准确缺陷,一种更深入分析数据分布的方法是采用正态概率图方法(normal probability plot)。一个服从正态分布的结果应在正态概率分布图的直线上下均匀分布。对所采集样本 SiO_2 含量(以质量分数 w 表示)的正态性分布进行验证,结果见图 1,可以看出,SiO_2 含量不服从正态分布特征,但是正态概率图方法不能给出 SiO_2 含量的数据多态性分布情况。

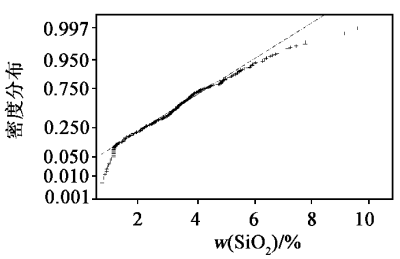

图 1　进口铁矿 SiO_2 含量正态性分布

Fig. 1　Normal probability plot of silica dioxide content in imported iron ore

Lowthian 等[9]提出采用内核密度估计(kernel density estimation)对实验数据的多态性进行分析以获得数据的整体估计值。该方法基于以下原理[10]:样本点 x_i($i = 1, 2, \cdots, n$)表达为一个标准偏差 s,面积为 $1/n$ 的正态分布,n 表示正态分布被加和的个数,任意一点 x 的内核密度 y 表示为:

$$y = \frac{1}{ns}\sum_{i=1}^{n}\phi\left(\frac{x - x_i}{s}\right) \quad (1)$$

式(1)中,$\phi(z)=\exp(-z^2/2)/\sqrt{2\pi}$为标准正态密度分布;$z$ 为概率密度值。

通过对实验数据的内核密度处理,可将原先离散的样本点构造为连续、平滑的内核密度分布图,更方便于数据分布的多态性观察及整体估计值获取。由式(1)可以看出,内核密度参数 s 的选择对平滑结果影响显著。s 的选择有多种方法,本文使用较简单、通用的直观观察方法选择 s[10-11],通过内核密度函数获得缺省平滑窗口宽度,再通过调节缺省平滑窗口宽度系数得到内核密度分布曲线(图 2)。从图中可看出,2 倍平滑单位得到的内核密度分布对表现进口铁矿中 SiO_2 含量的分布特征的分辨力较差,基本呈现一近似正态分布;1/2 倍平滑单位得到的内核密度分布对表现进口铁矿中 SiO_2 含量的分布特征的分辨力过强,对数据分布得细节表现太多,而忽略了数据的总体分布趋势描述;缺省平滑单位得到的内核密度分布对表现进口铁矿中 SiO_2 含量的分布特征的分辨力适中,表现为省略了数据分布的细节,突出数据分布的总体趋势。通过内核密度函数构造的进口铁矿 SiO_2 含量的分布特征总体表现为一典型的双态型(bimodality)多态分布,通过已有工作[9],对于双态型多态分布样本的代表值估计常规方法难以奏效,各种方法得到的代表值(真值)估计差别较大,这主要是由于双态型分布中每一单态数据的分布中心及分布权重无法确定。目前,国际上对处理这一类分布的代表值估计开始倾向于使用自举法重新取样模拟来获得对样本真值的估计[12-14]。

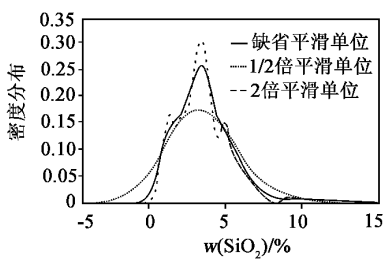

图 2　进口铁矿 SiO_2 含量的内核密度估计描述

Fig. 2　Description of silica dioxide content in imported iron ore by kernel density estimation

2.2　二氧化硅含量代表值自举法估计结果

实验样本数据的自举法重新取样模拟是通过对原始数据集的大量重新取样,再对大量取样后新构成的样本集的代表值作出估计,同时可给出置信区间及不确定度等统计参数。自举法重新取样模拟的效果可通过对多次重新取样中每一次取样指定代表值的正态分布图来考察,如果多次取样样本点在多次重新取样指定代表值的正态分布图中基本满足正态分布,则可认为得到的代表值估计是稳健的;但对于不满足正态分布的自举法重新取样模拟,也不能否定该方法的有效性。对进口铁矿 SiO_2 含量代表值进行自举法重新取样估计,取样次数采用缺省值 1000 次[15-16],研究表明,这一取样次数对绝大多数自举法应用都有足够精度[9,12]。自举法过程的分析结果见图 3,图中的条形图代表单次自举重新取样值的密度分布,图中同时给出了自举重新取样值的内核密度估计,可以看出,自举重新取样值的密度分布呈现明显的单态分布特征。

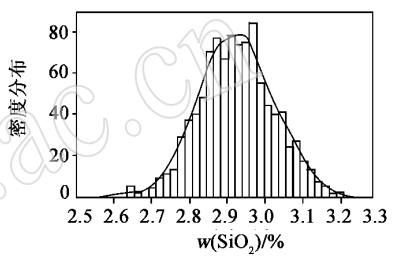

图 3　进口铁矿 SiO_2 含量的自举法参数估计

Fig. 3　Estimation of standing value of silica dioxide content in imported iron ore by bootstrap method

自举法重新取样理论认为,自举法重新取样样本的数值分布符合样本总体分布形态,自举法样本数值分布的均值和标准误差与样本总体一致,因此,可以用符合正态分布的自举法重新取样样本分布的均值和标准偏差作为有限单次样本代表值的稳健估计。自举标准误差(bootstrap standard error)是表征自举法样本分布标准偏差的统计量,自举标准误差 $SE_{\mathrm{boot},\bar{x}}$ 的定义[4]由式(2)给出:

$$SE_{\mathrm{boot},\bar{x}}=\sqrt{\frac{1}{B-1}\sum\left(\bar{x}^*-\frac{1}{B}\sum\bar{x}^*\right)^2} \qquad (2)$$

式(2)中,B 为自举重新取样次数;$\bar{x}^*$ 代表单次自举重新取样的平均值。计算结果 $w(SiO_2)$ 的代表值为 2.92,自举标准误差为 0.0982。

为进一步表明以自举法重新取样样本分布的均值与标准偏差作为有限单次样本代表值的稳健估计的有效性,考察单次自举重新取样平均值在正态概率分布图中的分布特征(图 4)。

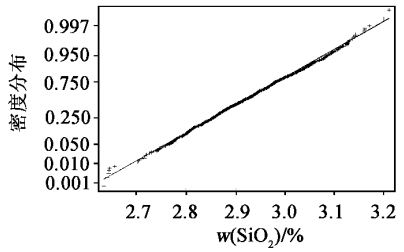

图 4　自举法模拟结果的正态分布

Fig. 4　Normal probability plot of standing value of silica dioxide content by bootstrap method

可以看出,进口铁矿 SiO_2 含量的自举法模拟结果满足正态分布特征,证明以自举法重新取样样本分布的均值与标准误差作为有限单次样本代表值是合理、有效的。

3　结语

对我国近两年 490 批次、23 个进口国的进口铁矿中 SiO_2 含量进行了总体统计分析,在数据统计分布特征研究基础上,用内核密度估计对进口铁矿 SiO_2 含量进行数据多态性分析;使用自举法对原始数据样本值重复取样获得了稳健的 SiO_2 含量代表值估计及标准误差,并证明以自举法重新取样样本分布的均值与标准误差作为有限单次样本代表值是合理、有效的。

4　参考文献

[1] 谢玉龙,王继红,梁逸曾,等. 化学计量学中的稳健估计方法[J]. 分析化学,1994,22(3):294 – 300.

[2] Vankeerberghen P, Vandenbosch C, Smeyers-Verbeke J, et al. Some Robust Statistical Procedures Applied to the Analysis of Chemical Data[J]. *Chemometrics and Intelligent Laboratory Systems*,1991,12:3 – 13.

[3] Zoubir A M, Boashash B. The Bootstrap and Its Application in Signal Processing[J]. *IEEE Signal Processing Magazine*,1998,15(1):55 – 76.

[4] Ron W, Hein P, Lutgrade M C. The Bootstrap: A Tutorial[J]. *Chemometrics and Intelligent Laboratory System*,2000,54:35 – 52.

[5] 黄玮,冯蕴雯,吕震宙. 极小子样试验的虚拟增广样本评估方法[J]. 西北工业大学学报,2005,23(3):384 – 387.

[6] ISO 3082:2000, Iron Ores—Sampling and Sample Preparation Procedures[S].

[7] ISO 2598-2:1992, Iron Ores—Determination of Silicon Content—Part 2: Reduced Molybdosilicate Spectrophotometric Method[S].

[8] R Development Core Team. URL http://www.R-project.org. R: A Language and Environment for Statistical Computing[CP/OL]. R Foundation for Statistical Computing, Vienna, Austria, 2003.

[9] Lowthian P J, Thompson M. Bump-hunting for the Proficiency Tester: Searching for Multimodality[J]. *Analyst*,2002,127:1359 – 1364.

[10] Bowman A W. Applied Smoothing Techniques for Data Analysis[M]. Cambridge: Oxford University Press, 1997:98 – 102.

[11] Jones M C, Marron J S, Sheather S J. A Brief Survey of Bandwidth Selection for Density Estimation[J]. *Journal of American Statistical Association*, 1996, 91: 401 – 407.

[12] Iwi G, Millard R K, Palmer A M, et al. Bootstrap Resampling: A Powerful Method of Assessing Confidence Intervals for Doses from Experimental Data[J]. *Physics in Medicine and Biology*,1999,44(4):55 – 62.

[13] Mokhlis Nahed A, Ibrahim Sahar. Efficient Bootstrap Resampling for Dependent Data[J]. *Communications in Statistics Part B: Simulation and Computation*, 2002, 31(3):345 – 355.

[14] Malzahn Dorthe, Opper Manfred. An Approximate Analytical Approach to Resampling Averages[J]. *Journal of Machine Learning Research*, 2003, 4(6): 1151 – 1173.

[15] 刘润幸,黄渭铭. 标准误差和可信区间的 Bootstrap 法及其计算机实现[J]. 数理医药学杂志,1996,9(1):67 – 70.

[16] Cole Stephen R. Simple Bootstrap Statistical Inference using the SAS System[J]. *Computer Methods and Programs in Biomedicine*,1999,60(1):79 – 82.