Bootstrap 方法及其在生物学研究中的应用

赵亮[1,3]，程锦秀[2]，许木启[3]，李明[3*]

（1. 安徽宿州学院化学与生命科学系，安徽宿州 234000；2. 安徽宿州学院继续教育学院，安徽宿州 234000；
3. 中国科学院动物研究所，动物生态与保护遗传学重点实验室，北京 100101）

摘要： Bootstrap 方法是以原始数据为基础的模拟抽样统计推断法，特别适用于那些难以用常规方法导出的参数的区间估计、假设检验等问题。本文介绍了该方法的基本思想及具体步骤，并附有生物学研究中应用的实例。生物科学中有许多数据总体分布信息往往很难确定，难以用常规的方法进行统计分析，因此 Bootstrap 方法在生物科学研究中具有很大的应用价值。

关键词： Bootstrap 方法；区间估计；假设检验

中图分类号： O212　　**文献标识码：** B　　**文章编号：** 1000－7083(2010)04－0638－04

Bootstrap Method and its Application in Biology

ZHAO Liang[1,3], CHENG Jin-xiu[2], XU Mu-qi[3], LI Ming[3*]

(1. Department of Biology, Suzhou College, Suzhou, Anhui Province 234000, China; 2. Department of Continuing Education, Suzhou College, Suzhou, Anhui Province 234000, China; 3. Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: The bootstrap method is a data-based simulation to carry out familiar statistical calculations, such as confidence intervals estimated, statistic inference, *et al*. By purely computation means rather than using of statistical formulas, it is useful especially when the statistical formulas are hard to be got. This article introduced the bootstrap method, including bootstrap basic ideas and procedures and illustrated its application in biology with some examples. Along with the quick development of computer techniques, this method is now surging into widely practical use in biological studies.

Key words: bootstrap; confidence intervals estimated; statistic inference

统计推断是从样本资料推断相应的总体特征，包括参数估计和假设检验。早期的统计推断是以大样本为基础的，在处理问题时，通常假定是正态分布；WS. Gosset 在 1908 年发现了 t 分布后，开创了小样本的研究。半个多世纪以来，这种思维一直占有主导地位，统计学家研究的主流就是如何将这种思维付诸实践（Rao，1973；陈峰等，1997）。实际上，对总体分布做出假定，一般出于计算上的考虑，真实分布通常是无法精确了解的，同时在很多情况下关于总体参数的某些推断（如均值的方差、均值的分位数、置信区间等）几乎不可能推导出明确的解析式来。当今计算机技术的高度发展，使统计研究及其应用跃上了一个新台阶。特别是个人电脑的发展及普及，把统计学家从求解数学难题中解放出来，并逐渐形成一种面向应用的、基于大量计算的统计思维——模拟统计推断（Roff，2006；谢益辉等，2008）。模拟统计推断在解决具体问题时，只依据观测信息作分析和判断。模拟计算方法很多，主要包括 Bootstrap 方法、Jackknife 方法、Permutation 方法、Monte Carlo 模拟等（Roff，2006）。本文介绍其中的 Bootstrap 法，并附有生物学研究中应用的实例，若与传统的统计思维相比较，确有新意。本研究实例分析使用 R-软件编制完成（Ihaka *et al*.，1996）。

1　Bootstrap 方法简介

Bootstrap 方法是 Efron（1979）提出来的一种统计方法，它根据给定的原始样本复制观测信息，不需要进行分布假设或增加新的样本信息，可对总体的分布特性进行统计推断，属于非参数统计方法（Money *et al*.，1993）。Efron 和 Tibshirani（1993）、Davison 和 Hinkley（1997）较为系统地介绍了 Bootstrap 方法的理论成果。在过去 20 多年的时间里，该方法在理论和应用上都得到发展，尤其在金融、经济、生物、医学等领域应用广泛（刘勤等，1998；Davidson *et al*.，2002；龙志和等，2008；钱俊等，2008）。Bootstrap 方法的核心是利用自助样本（或称为 Bootstrap 样本、自举样本）来估计未知概率测度的某种统计量的统计特性。

Bootstrap 方法中心思想为：假设我们希望估计某一分布 $F(\theta;x)$ 的某一统计量为 $\theta(x)$：$\theta(x)=\int g(x)dF(\theta;x)$。由于总体分布经常是未知的，Bootstrap 方法通过由样本获得的经验分布 $\hat{F}(\theta;x)$ 来对总体分布 $F(\theta;x)$ 进行估计得到：$\hat{\theta}(x)=\int g(x)dF(\theta;x)$。根据极限定理，经验分布 $\hat{F}(\theta;x)$ 是总体理

收稿日期：2009-10-12　　接受日期：2009-12-06　　基金项目：安徽省教育厅重点项目（KJ2009A052Z）；宿州学院人才基金（2007YSS10）

作者简介：赵亮，男，博士，副教授，研究方向：保护遗传学，E-mail：zhaoliang@ioz.ac.cn

* 通讯作者 Corresponding author，E-mail：lim@ioz.ac.cn

论分布 F(θ; x) 的一致性估计。

Bootstrap 方法的具体步骤如下: 首先有一个实际观测到的数据集(称之为原始数据集) ,它含有 n 个观测,然后根据分析的需要确定计算某个统计量的公式。从这个数据集中有放回地随机抽取 m 个观测组成一个样本,称之为 *Bootstrap* 样本。在一次随机抽样中,原始数据集中的观测有的只被抽到 1 次,有的超过 1 次,也有的没有被抽到。利用这个被抽到的样本,按照事先确定的公式,计算出所需要的统计量。如此反复抽样和估计,称之为复制。最后由复制出的统计量的值组成一个数据集(θ_i; i = 1,2,……m) ,并利用这个数据集来反映该统计量的抽样分布,即产生经验分布,这样,即使我们对总体分布不确定,也可以近似估计出一些统计量及其置信区间(例如均数、中位数等) 。尤其值得说明的是,利用 *Bootstrap* 方法所获得的经验分布的特征,可用于解决那些难以用常规方法导出的对参数的区间估计、假设检验等问题。

2 Bootstrap 方法应用实例

2.1 参数的区间估计

【例 1】从一个 $\lambda = 1$ 的 Poisson 分布总体中随机抽取一个样本含量为 100 的样本。对该样本用 Bootstrap 方法来估计平均数和 95 可信区间。

本例数据来源于 Poisson 分布总体,为非对称分布,不能用正态分布或 t 分布的原理来估计其均值的置信区间。Bootstrap 方法不需要进行分布假设,步骤如下:

(1) 产生一个样本容量为 100,符合 Poisson 分布的随机样本; 计算其平均值$\bar{x}_0$。

(2) 以上面的 100 个随机数作为原始数据集,从中有放回地随机抽取 1 个含量仍为 100 的 Bootstrap 样本,计算其平均值$\bar{x}_0^*$ 。

(3) 重复步骤(2) m 次(本研究 m = 1000) ,共获得 m 个值。

(4) 检验 m 个平均值是否符合正态分布,平均值频数分布符合正态分布时,以其平均数$\bar{x}_0$ 作为点估计,用正态原理估计可信区间; 频数分布为偏态时,以其中位数作为点估计,以自举统计数分布的上、下 2.5% 分位数作为其 95% 可信限。

(5) 分别计算平均数(或中位数) 的估计值和 95% 的置信区间。

本例 Bootstrap 方法所获得 1000 个平均值的分布见图 1。

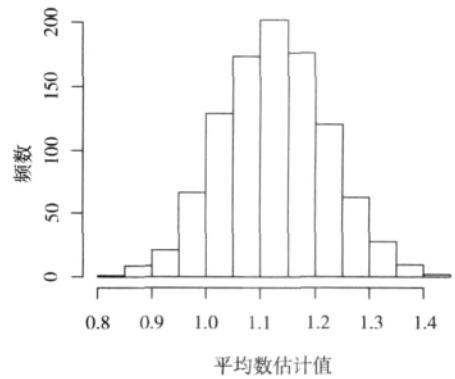

图 1 Bootstrap 样本平均数的分布

Fig. 1 Distribution of the Bootstrap sample average

Shapiro-Wilk 检验表明符合正态分布(W = 0.9993, P = 0.972) ,其平均数的估计值和 95% 的置信区间为: $\bar{x}_0 \pm z_{a/2} S(x) = 1.130 \pm 0.192$。

2.2 假设检验

2.2.1 两均数比较

【例 2】冬小麦“东方红 3 号”和“农大 193”的蛋白质含量(%) 测定结果如表 1,两种冬小麦蛋白质含量是否有显著差异?

本例原假设为两种冬小麦蛋白质含量没有显著差异,即 $H_0: \mu_1 - \mu_2 = 0$。上述两样本经测验为方差异质($F = 6.35$, $P < 0.05$) ,此时若用常规 t 测验,得出 t 值的抽样分布并不服从 t 分布,因此难以用常规 t 测验来检测。用 Bootstrap 方法进行同样的假设测验,步骤如下:

表 1 两种冬小麦蛋白质含量测定结果

Table 1 Protein content of two winter wheats

品种 Species	蛋白质含量(%) Protein content(%)	平均值 Average	标准差 Standard Deviation
东方红 3 号 Dongfanghong number 3	11.8, 15.1, 16.4, 14.8, 13.5, 14.2, 14.5, 14.8, 13.8, 15.0	14.39	1.209
农大 193 号 Nongda number 193	11.7, 10.8, 11.6, 11.9, 12.0	11.60	0.474

(1) 利用原始数据计算两种冬小麦蛋白质含量平均数之差,记为 $W_0 = \bar{x}_1 - \bar{x}_2$。

(2) 从东方红 3 号中有放回的抽取一个容量为 10 的自举样本,记为 $x_{11}, x_{12} \cdots, x_{110}$,计算出$\bar{x}_1^* = \frac{1}{10}\sum_{i=1}^{10} x_{1i}$; 从农大 193 中有放回的抽取一个容量为 5 的自举样本,记为 $x_{21}, x_{22} \cdots, x_{25}$,计算出$\bar{x}_2^* = \frac{1}{5}\sum_{i=1}^{5} x_{2i}$,然后计算 $W = \bar{x}_1^* - \bar{x}_2^*$ 。

(3) 重复步骤(2) m - 1 次(本研究 m = 1000) ,共获得 m 个 W 值(包括 W_0) ,按由小到大的顺序排列,由此得出 W 值分布图(图 2) ,得到 $W_{0.025}$ 和 $W_{0.975}$,若区间($W_{0.025}, W_{0.975}$) 包含 0,则不拒绝原假设,否则拒绝原假设。

本例两种冬小麦蛋白质含量自举样本平均数之差的第 2.5% 分位数到第 97.5% 分位数所包括的区间是(1.95, 4.98) ,不包含 0,拒绝原假设,即两种冬小麦蛋白质含量存在显著差异。

2.2.2 多个样本平均数的比较

【例 3】采用三种黑光灯诱捕昆虫,捕获数见表 2,其差异有无统计学意义?

对该多样本进行正态性检验, Shapiro-Wilk 检验证明随机误差不服从方差分析中的正态性假定(W = 0.8332, P = 0.01012) ; 对该多样本进行方差同质性测验,证明该资料也不

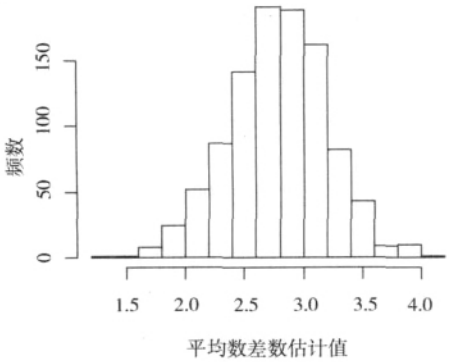

图 2　Bootstrap 样本平均数差数的分布

Fig. 2　Distribution of the difference between two Bootstrap sample averages

表 2　3 种黑光灯诱捕昆虫数量

Table 2　Amount of insect captured by three different black-light lamps

处理 Treatment	黑光灯 Black-light lamps		
	Ⅰ	Ⅱ	Ⅲ
1	19	50	123
2	23	166	407
3	39	224	398
4	23	59	229
5	17	65	251
平均数 Arverage	24.2	112.8	281.6
标准差 Standard Deviation	8.67	77.97	120.55

满足方差分析中的方差同质性假定(Bartlett's K-squared = 14.0365，$P = 0.00089$)，因此不能按常规的方差分析方法对该资料进行检验。

使用 Bootstrap 方法则不需这些假定，步骤如下:

(1) 在 H_0:$\mu_i = \mu_0$，即不同黑光灯昆虫捕获数间没有统计学意义。根据 F 值计算公式 $F = MS_t / MS_e$(MS_t 为处理均方，MS_e 为误差均方) 算出样本统计量 F_0 为 12.398。

(2) 把 15 个样本观察值混合成新的样本，通过 Bootstrap 抽样，得到具有 15 个观察值的 Bootstrap 自举样本，将 15 个观察值随机地归为Ⅰ、Ⅱ、Ⅲ三类，且每类均有 5 个观察值; 计算各自举样本相应的 F 值。

(3) 重复步骤(2) m - 1 次，共得到 m 个 F 值(包括 F_0)，由此得出 F 值分布图(图 3)，由之得到临界值 $F_{0.95} = 3.884$。本例实际样本统计量 $F_0 = 12.398$ 落在接受区之外($F_0 > F_{0.95}$)，因此拒绝 H_0，即判断三种黑光灯诱捕昆虫有明显差异。

2.2.3　基因功能约束序列检验

【例 4】假设在一个长度为 10 000 bp 的基因的 DNA 序列中有一段序列为 ATATAT，它在整个基因序列中出现了 10 次，我们认为这不是一个随机现象，可能为功能约束片段，试分析之。

本例需进行如下检验:

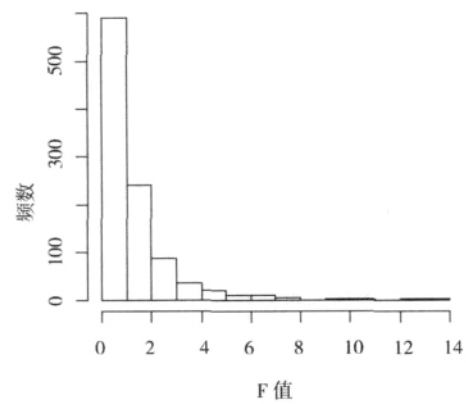

图 3　Bootstrap 样本统计量 F 值的分布

Fig. 3　Distribution of F-value for Bootstrap samples

H_0: ATATAT 序列出现 n = 10 次为随机现象;

H_A: ATATAT 序列出现 n = 10 次为功能约束的结果。

检验本例的问题首先要知道在 DNA 序列随机排列的假设下这段序列(ATATAT) 重复出现次数的概率分布，显然这很难用常规的统计学知识去推导，而用 Bootstrap 方法可以方便地解决该问题，步骤如下:

(1) 该基因的长度为 10 000 个碱基，我们可以从 4 种碱基(A，T，C，G) 中进行 10 000 次独立的 Bootstrap 抽样，按抽到的顺序组成一个新的序列。

(2) 新得到的序列中碱基的排列顺序是完全随机的，然后数一数其中 ATATAT 这段序列重复出现了多少次(本例记为 Ti) 。

(3) 重复上述过程(本例是 i = 1，2，3…，1500)，每次都计算这段序列重复出现的次数，可以得到 1500 个数，它们就构成了在原假设下该段序列重复出现次数的经验分布。

(4) 求重复出现次数 Ti 经验分布的 95% 的置信区间，若 n = 10 落在该置信区间内，说明在整个基因序列中出现了 10 次 ATATAT 是一种随机现象; 否则可能为一功能约束序列。

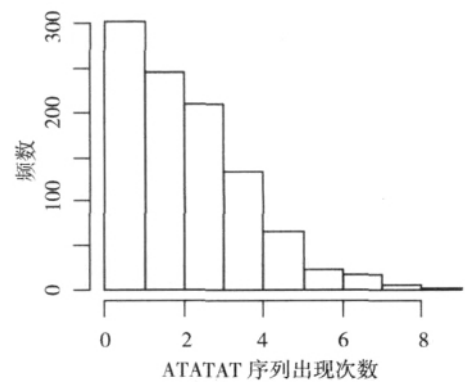

图 4　ATATAT 序列出现次数的频率分布

Fig. 4　Distribution of ATATAT sequences for Bootstrap samples

使用 Bootstrap 方法得到本例 ATATAT 重复出现次数 T_i 直方图如图 4 所示，其 95% 的置信区间为(0，6) 。ATATAT 重复出现次数 n = 10 落在该置信区间之外，拒绝原假设，说明在整个基因序列中出现了 10 次 ATATAT 不是一种随机现象，可能该序列为一功能约束序列。

3 讨论

参数的点估计和区间估计要求已知或可以导出数据的分布特征，生物科学研究中的许多问题非常复杂，对其总体分布信息往往很难确定，因而难以用常规的方法进行区间估计。在处理两个平均数的假设测验中，要求统计数的理论分布是已知或可以导出，同时要求样本的方差是同质的，在样本方差异质的情况下，统计量的分布不再服从正态分布、t 分布或 F 分布(陈红等，1997；敖雁等，2006)。多个平均数的假设测验，即方差分析也是建立在一些假定的基础上：① 处理效应和误差(环境)效应是可加的；②试验误差是独立的随机变量，并作正态分布的；③ 所有处理的误差方差都是同质的。但实际上生物科学研究中的许多试验资料并不能满足三个假定，这种资料不能直接进行方差分析。此时，一种方法是使用符号检验、秩和检验、等级相关检验等非参数检验方法，非参数统计由于采用编秩的方法进行处理，信息比较粗糙，效率上损失较多，一般来说检验效率低于参数方法(宋昕等，2004)；另一种方法是将变量作适当的数据转换，再对转换后的资料进行方差分析，但该方法缺点是较烦琐，且经转换过的数据缺乏描述问题的直接性(敖雁等，2006)。

Bootstrap 方法根据给定的原始样本复制观测信息对总体的分布特性进行统计推断，不需要额外的信息，Efron(1979)认为该方法也属于非参数统计方法。Bootstrap 方法从观察数据出发，不需任何分布假定，针对统计学中的参数估计及假设检验问题，利用 Bootstrap 方法产生的自举样本计算的某统计量的数据集可以用来反映该统计量的抽样分布，即产生经验分布，这样，即使我们对总体分布不确定，也可以近似估计出该统计量及其置信区间，由此分布可得到不同置信水平相应的分位数——即为通常所谓的临界值，可进一步用于假设测验。因而，Bootstrap 方法能够解决许多传统统计分析方法不能解决的问题。在 Bootstrap 的实现过程中，计算机的地位不容忽视(Diaconis *et al.*，1983)，因为 Bootstrap 涉及到大量的模拟计算。可以说如果没有计算机，Bootstrap 理论只可能是一纸空谈。随着计算机的快速发展，计算速度的提高，计算费时大大降低。在数据的分布假设太牵强或者解析式太难推导时，Bootstrap 为我们提供了解决问题的另一种有效的思路。因此，该方法在生物科学研究中有一定的利用价值和实际意义(敖雁等，2006；谢益辉等，2008)。

值得说明的是 Bootstrap 的应用在很大程度上取决于经验分布的选取和样本数的大小(Roff，2006)。而且 Bootstrap 是在原始样本及其经验分布的基础上作有放回的再抽样，其结果是针对现有资料做出统计推断，所得结论不具一般性，只能靠大量复制样本从而优化经验分布，得到较准确的检验及推理结果。在利用 Bootstrap 方法进行区间估计和假设测验时，需要确定合适的抽样次数，即抽样次数 m 的取值不能太小，一般要求在 500 次以上(张勤，2007)。应该指出 Bootstrap 方法并不是在一切场合下都适用的，当样本量小不足以提供总体分布信息时更是如此(Beran，1982；Hardle *et al.*，1991)。作为一种新统计方法，Bootstrap 方法尚在发展中，许多问题有待于进一步研究和解决(陈红等，1997；谢益辉等，2008)。

4 参考文献

敖雁，王学枫，汤在祥，等. 2006. Bootstrap 方法在平均数假设测验中的应用[J]. 中国卫生统计，23(6)：542～544.

陈峰，陆守曾，杨珉. 1997. Bootstrap 估计及其应用[J]. 中国卫生统计，14(5)：5～7.

陈红，吴汇川. 1997. Bootstrap 方法及其应用[J]. 青岛大学学报，(3)：78～83.

刘勤，金丕焕. 1998. Bootstrap 方法及其在医学统计中的应用[J]. 中国预防医学杂志，(1)：52～53.

龙志和，欧变玲. 2008. Bootstrap 方法在经济计量领域的应用[J]. 工业技术经济，(7)：231～234.

钱俊，陈平雁. 2008. Bootstrap 和 Permutation 方法在样本率多重比较中的应用[J]. 中国医学统计，(1)：34～36.

宋昕，蔡泳，徐刚，等. 2004. 非参数检验方法概述[J]. 上海口腔医学，(6)：561～563.

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

张勤. 2007. 动物遗传育种中的计算方法[M]. 北京：科学出版社.

Beran R. 1982. Estimated Sampling Distributions，The Bootstrap and Competitors[J]. Ann Statist，(10)：212～215.

Davision AC，Hinkley DV. 1997. Bootstrap Methods and Their Application[M]. Cambridge：Cambridge University Press.

Davidson AC，MacKinnon G. 2002. Bootstrap Inference In Econometric[J]. The Canadian Journal of Economics，35(4)：615～645.

Diaconis P，Efron B. 1983. Computer-Intensive Methods In Statistics[J]. Scientific American，(5)：116～130.

Efron B. 1979. Bootstrap Methods：Another Look at the Jackknife[J]. The Annals of Statistics，7(1)：1～26.

Efron B，Tibshirani R. 1993. An Introduction to The Bootstrap[M]. New York：Chapman & Hall Ltd.

Hardle W，Marron JS. 1991. Bootstrap Simultaneous Error Bars for Nonparametric Regression[J]. Ann statist，(2)：778～796.

Ihaka R，Gentleman RR. 1996. A Language for Data Analysis and Graphics[J]. Journal of computational and Graphical Statistics，(5)：299～314.

Money CZ，Duval RD. 1993. Bootstrapping. A Nonparametric Approach to Statistical Inference[M]. Sage University Paper series on Quantitative Applications in the Social Sciences，Sage：Newbury Park.

Rao CR. 1973. Linear Statistics Inference and Its Application (2nd)[M]. New York. John Wiley & sons Inc.

Roff DA. 2006. Introduction to Computer-Intensive Methods of Data Analysis In Biology[M]. Cambridge University Press.