

# Bootstrap 估计标准化率及其可信区间

朱焱<sup>1</sup> 赵耐青<sup>2△</sup>

<sup>1</sup>贵阳医学院公共卫生学院卫生统计教研室 贵阳 550004;

<sup>2</sup>复旦大学公共卫生学院卫生统计与社会医学教研室 上海 200032)

**【摘要】** 目的 实现标化率的 95%可信区间计算。方法 以 2000 年全国人口构成为标准,对从已知总体中随机抽样产生的 100 个样本,用 Bootstrap 再抽样技术计算标准化率,并估计其 100 个 95%可信区间,考察有多少个可信区间包含了总体标准化率。结果 样本含量 100 000 和 500 000 时,100 个样本的标准化死亡率的 95% CI 包含总体标准化死亡率的覆盖率的 95%可信区间分别为(0.887 2,0.983 6)和(0.929 6,0.997 6)。结论 Bootstrap 再抽样技术对估计样本标准化率及其可信区间有较好的准确性。

**【关键词】** Bootstrap 法; 标准化率; 可信区间

**【中图分类号】** R 195.1 **【文献标志码】** A

## Bootstrap method estimates standardize rate and confidence intervals

ZHU Yan<sup>1</sup>, ZHAO Nai-qing<sup>2△</sup>

<sup>1</sup>Department of Health Statistic, School of Medical, Guiyan University, Guiyang 550004, China;

<sup>2</sup>Department of Health Statistic and Social Medicine, School of Public Health,

Fudan University, Shanghai 20032, China)

**【Abstract】** **Objective** This article introduces the bootstrap method to estimate standardize rate and confidence interval. **Methods** The national population of 2 000 was used as the standard population, and standardize rate and 95% CI of 100 random samples from a known population was estimated with bootstrap method. **Results** The 95% CI of one hundred standardize rates covered the standardize rate of total was (0.887 2,0.983 6) when sample size was 100 000, and the other was (0.929 6, 0.997 6) when sample size was 500 000. **Conclusions** The technique of bootstrap has good accuracy to estimate the sample standardize rate and confidence interval.

**【Key words】** bootstrap method; standardize rate; confidence interval

当在比较两个(或多个)地区死亡率(或患病率、发病率等)时,由专业知识可知某个地区死亡率与该地区年龄构成有着密切的联系,高年龄组的死亡率高于低年龄组的死亡率。由于各个地区的年龄构成存在差别,要比较各个地区总死亡率的差别,应当用标准化法。标准化法的基本思想就是统一各组的年龄构成,使资料具有可比性,即选择通用的、有代表性的、较稳定、数据量大的人群资料为标准人口构成,各组以标准人口构成为参照计算相应的标准化率,再进行标准化率的比较。这样计算出来的标准化率因为用了相同的人口年龄别构成比,从而消除

了年龄这个混杂因素的影响。但是,如果资料是一个抽样资料,此时计算所得尚是样本标准化率,若要估计其总体的标准化率,样本标准化率只是一个点估计,一般需要计算 95%可信区间。Swift<sup>[1]</sup>作了有关 Bootstrap 应用于标准化率可信区间估计的研究,认为计算机模拟 Bootstrap 法进行直接标准化率的可信区间估计是一种较准确的方法。近年来,Bootstrap 可信区间估计已广泛应用于医学统计学的各个领域<sup>[2,3]</sup>。国内尚未见有关 Bootstrap 应用于标准化率可信区间估计的相关报道<sup>[4,5]</sup>。本文将介绍 Bootstrap 法实现对总体标准化率的区间估计

<sup>△</sup>Corresponding author E-mail:nqzhao@gmail.com

并与曹素华老师主编《卫生统计学方法》<sup>[6]</sup>中对甲乙两地食管癌死亡率直接法标准化的方法进行比较。

## 资料和方法

Bootstrap 分为参数 Bootstrap 和非参数 Bootstrap,参数 Bootstrap 需要借助假定的模型进行估计然后借助估计模型进行重抽样统计,非参数 Bootstrap 借助经验分布理论进行重抽样统计<sup>[7,8]</sup>。在 Bootstrap 应用中,绝大多数研究采用非参数 Bootstrap 方法进行重抽样统计,故本研究将采用非参数 Bootstrap 计算标准化率的 95% 可信区间。Bootstrap 再抽样技术的具体方法为,假设有一个实际观测到的数据集(原始数据集),它含有  $n$  个观察单位,从这个数据集中有放回地随机抽取  $m$  个观察单位组成一个新的样本,称为 Bootstrap 样本。在有放回地随机抽样中,原始数据集中的每个观察单位每次被抽到的概率相等,为  $1/n$ ,这些观察单位有的只被抽到 1 次,有的超过 1 次,也有的没有被抽到。而 Bootstrap 样本含量  $m$  通常取实际观测随机样本含量  $n$ <sup>[9]</sup>。

本研究参照某人口普查资料的年龄组死亡率为背景,用计算机模拟得到已知总体及其各个年龄组的死亡率,以当年全国人口资料为标准人口,用直接标准化的方法得到该人群的标化死亡率。从该已知总体中随机抽取一定样本量的 100 个样本为研究对象(每一个随机样本为一个原始数据集)。以当年全国人口资料为标准人口,对每一个原始数据集应用 Bootstrap 法进行再抽样,Bootstrap 样本数为 200<sup>[9]</sup>,Bootstrap 样本量与原始数据的样本量相同,计算标准化率及估计标准化率的 95% 可信区间(CI),并验证用 Bootstrap 法估计标准化率 95% 可信区间的准确性。

### Bootstrap 计算标准化率及 95% 可信区间

2000 年某地某人口普查资料为已知总体,为了叙述和计算方便,不妨取 20~39 岁为研究人群,各年龄组死亡率见表 1,粗的总死亡率为 0.000 539 9,用全国人口数直接标化。记这 4 个年龄组的全国总人口数为  $N_1, N_2, N_3, N_4$ ,这 4 个年龄组的全国总人口数为  $N_{all}$ ,该地区 4 个年龄组的平均人口数分别为  $\bar{n}_1, \bar{n}_2, \bar{n}_3, \bar{n}_4$ ,该地区 4 个年龄组的死亡人数为  $d_1, d_2, d_3, d_4$ ,根据直接标化公式得到该地区 20~39 岁人群的标化死亡率(即总体标化的死亡率)为:

$$P_{\text{总体标化}} = \sum_{i=1}^4 \frac{N_i}{N_{all}} p_i = \sum_{i=1}^4 \frac{N_i}{N_{all}} \left( \frac{d_i}{\bar{n}_i} \right) \\ = 0.000 532 284$$

表 1 2000 年某地某人口年龄组死亡资料

Tab 1 The mortality rate of different ages somewhere in 2000

Age (year)	National population( $N_i$ )	The mortality status		
		Average population( $\bar{n}_i$ )	Death ( $d_i$ )	Mortality ( $p_i$ )
20-24	94 573 174	1 344 760	478	0.000 355 5
25-29	117 602 265	1 384 506	529	0.000 382 1
30-34	127 314 298	1 376 412	706	0.000 513 9
35-39	109 147 295	1 526 590	1 328	0.000 869 9
Total	448 637 032	5 632 268	3 041	0.000 539 9

参照表 1 所示总体,用计算机模拟简单随机抽样的方法,从中随机抽取样本量为  $n$  的 100 个样本,对 100 个随机样本采用 Bootstrap 再抽样。

对每个样本,运用 Bootstrap 再抽样方法,即在每个现有样本中有放回地再抽样 200 次,每次抽样的样本含量为原有样本含量(样本量为  $n$ )。得到 200 个 Bootstrap 的样本,对于每个 Bootstrap 样本,记抽到的这 4 个年龄组的 Bootstrap 样本人口数为  $m_{b1}, m_{b2}, m_{b3}, m_{b4}$ ,其中  $m_{b1} + m_{b2} + m_{b3} + m_{b4} = n$ ,记抽到的这 4 个年龄组的 Bootstrap 样本死亡人数为  $d_{b1}, d_{b2}, d_{b3}, d_{b4}$ ,以当年全国人口构成为标准人口构成比,用直接标化的方法计算每个 Bootstrap 样本的标准化死亡率,由此得到每个样本的 200 个 Bootstrap 样本的标准化死亡率,用百分位数计算 95% 可信区间<sup>[10,11]</sup>。

首先取随机抽样的样本量为  $n = 100 000$ ,从上述总体中随机抽取 100 个样本,对于每个样本进行 Bootstrap 再抽样,所以得到 100 个标准化死亡率的 95% 可信区间,考察这 100 个 95% 可信区间是否包含总体标准化死亡率,统计包含总体标准化死亡率的 95% 可信区间,由此推断用 Bootstrap 方法所估计的 95% 可信区间的可信度是否为 95%。

为了考查样本量增大对上述方法的影响问题,将随机抽样的样本量增加为  $n = 500 000$ ,从上述总体中再随机抽取 100 个样本,再次进行 Bootstrap 法估计参数及其 95% 可信区间。并验证 100 个可信区间中,包含标准化总率的可信区间的个数和推断用 Bootstrap 方法所估计的 95% 可信区间的可信度是否为 95%。

**不同方法的比较** 对甲乙两地死亡率的资料用自编程序计算甲乙两地死亡标准化率及其可信区间,并与教材上提供的方法进行比较<sup>[6]</sup>。

**Bootstrap 过程的计算机实现** 根据上述方法,运用 SPSS 11.5 及 Stata7.0 软件完成。相应程序如下:

```

program define pp
1.     version 7.0
2.     if “‘1’” == “?”{
3.         global S_1 “deadrate”
4.         exit
5.     }
6.     scalar  n1 = 94 573 174
7.     scalar  n2 = 117 602 265
8.     scalar  n3 = 127 314 298
9.     scalar  n4 = 109 147 295
10.    su dead if age == 1
11.    scalar m1 = r(mean) * n1
12.    su dead if age == 2
13.    scalar m2 = r(mean) * n2
14.    su dead if age == 3
15.    scalar m3 = r(mean) * n3
16.    su dead if age == 4
17.    scalar m4 = r(mean) * n4
18.    scalar ppd = (m1 + m2 + m3
                + m4)/(n1 + n2 + n3 + n4)
19.    post ‘1’ (ppd)
20.    end
.     set seed 10001
.     bstrap pp, reps(200)

```

## 结 果

与该人群的总体标准化死亡率:0.000 532 284 比较,结果有 95 个样本的标准化死亡率的 95%可信区间包含总体标准化死亡率,即样本标准化死亡率的 95%可信区间的包含总体标准化死亡率的样本覆盖率为 95%,对应覆盖率的 95%可信区间为(0.887 2,0.983 6),包含可信度理论值 0.95。未包含总体标准化率的可信区间的 5 个样本的标准化死亡率及其 95%可信区间由表 2 列出。

表 4 甲、乙两地各年龄食管癌死亡率(1/10 万)

Tab 4 The mortality rate of esophagoscope in different areas (Per 100 thousand)

Age(year)	Region A			Region B		
	Population	Death	Mortality	Population	Death	Mortality
0 -	420 664	1	0.2	313 865	0	0.0
30 -	71 048	14	19.7	44 176	7	15.8
40 -	62 048	73	117.7	45 751	49	107.1
50 -	47 216	179	379.1	37 472	135	360.3
60 -	28 067	196	698.3	25 947	176	678.3
70 -	11 966	85	710.4	16 148	112	693.6
Total	641 008	548	85.5	483 359	479	99.1

表 2 样本量为 100 000 时不包含总体标准化率的可信区间

Tab 2 The 95% confidence interval excluding SMR of population with sample size of 100 000

ID	Standardize rate	95% confidence interval	
		Lower	Upper
2	0.000 405 9	0.000 293 7	0.000 529 9
15	0.000 699 3	0.000 547 7	0.000 852 7
60	0.000 681 2	0.000 543 7	0.000 841 2
61	0.000 717 6	0.000 564 4	0.000 889 9
83	0.000 381 5	0.000 281 0	0.000 514 7

样本含量为 500 000,进行计算机模拟实验,方法同 1。结果 100 个标准化死亡率的 95%可信区间有 98 个包含总体标准化死亡率。即随机抽样样本量为 500 000 时的 100 个样本标准化死亡率的 95%可信区间的包含总体标准化死亡率的样本覆盖率为 98%,对应覆盖率的 95%可信区间为(0.929 6,0.997 6),包含可信度理论值 0.95。未包含总体标准化率的可信区间的 2 个样本的标准化死亡率及其 95%可信区间由表 3 列出。

表 3 样本量为 500 000 时不包含总体标准化率的可信区间

Tab 3 The 95% confidence interval excluding SMR of population with sample size of 500 000

ID	Standardize rate	95% confidence interval	
		Lower	Upper
20	0.000 457 9	0.000 387 3	0.000 521 4
55	0.000 455 9	0.000 382 0	0.000 524 3

**两种方法的比较及验证** 曹素华老师主编的《卫生统计学方法》第 27 页甲乙两地死亡率的比较<sup>[6]</sup>,是以甲乙两地人口合计作为标准人口,调用 Stata7.0 软件 dstdize 命令,对甲乙两地食管癌死亡率进行直接法标化。

利用该资料,仍以甲乙两地人口合计作为标准人口,调用自编程序,用 Bootstrap 再抽样方法分别对甲、乙两地资料计算标准化率及其 95%可信区间。结果见表 5。可见两方法结果相差甚微。

表 5 两种不同方法计算甲、乙两县标准化率的 95% 可信区间比较  
 Tab 5 Comparison of the 95% confidence intervals of SMR between different methods in two areas

Methods	Region A				Region B			
	Standardize rate	95% confidence interval			Standardize rate	95% confidence interval		
		Lower	Upper	Width		Lower	Upper	Width
Dstdize	0.000 933	0.000 855	0.001 012	0.000 157	0.000 890	0.000 810	0.000 969	0.000 159
Bootstrap	0.000 933	0.000 852	0.001 014	0.000 162	0.000 892	0.000 813	0.000 970	0.000 157

## 讨 论

Bootstrap 再抽样技术(自助法)是美国 Stanford 大学统计系教授 Efron 提出的一种统计推断方法,是以原始现有样本资料为基础,借助经验分布理论进行有放回地重抽样,从而估计和推断统计量的分布特征,特别适用于难以用常规方法进行参数估计和假设检验的问题,目前 Bootstrap 已经成为非常常用的统计分析方法。

随机模拟的特点是用于验证方法的结果是否符合已知总体,而实际资料是不知道已知总体参数的,所以无法用实际数据说明方法是否合适或正确的。随机模拟都是假定实际资料来自某个总体,模拟数据虽然为伪随机数,但基本满足随机模拟的要求,任何一个研究中的随机模拟都是采用计算机产生伪随机数进行随机模拟的。本研究采用 Bootstrap 法对计算机模拟总体实现标准化率及其可信区间的估计,其结果与曹素华老师主编《卫生统计学方法》中借助 Stata 软件进行两样本资料的标化统计量比较的结果有较好的一致性。但 dstdize 方法只能以两个样本合并作为标准人口构成比,不是真正的标准人口构成,当两个样本都比较小的时候,抽样误差就会比较大。用 Bootstrap 方法进行标化具有计算量较大的缺陷,不过目前计算机技术已经高度发展,计算量大的问题已经基本不能成为阻碍其应用的问题了。另外用 Bootstrap 方法可以引用一个独立的人群年龄构成比(如全国人群年龄构成比)同时对两个或多个样本进行标化和比较,因此具有推广应用

价值。

## 参 考 文 献

- [1] Swift MB. Simple confidence intervals for standardized rates based on the approximate bootstrap method[J]. *Stat Med*, 1995,14(17):1 875 - 1 888.
- [2] Hutson AD. Bootstrap-type confidence intervals for quantiles of the survival distribution[J]. *Stat Med*, 2001,20(11):1 693 - 1 702.
- [3] Jiang H, Zhou XH. Bootstrap confidence intervals for medical costs with censored observations[J]. *Stat Med*, 2004,23(21): 3 365 - 3 376.
- [4] 蔡雪亚,金丕焕,曹素华.用 Bootstrap 方法计算中位数的可信区间[J]. *中国卫生统计*, 2002,19(3):185 - 186.
- [5] 乔舰,景平,汪金芳.比估计的 Bootstrap 置信区间[J]. *河南科技大学学报:自然科学版*, 2004,25(6):85 - 89.
- [6] 曹素华. *卫生统计学方法*[M]. 上海:复旦大学出版社, 2003: 26 - 31.
- [7] 张文彤. Bootstrap 方法在 Cox 模型参数估计中的应用[J]. *中国卫生统计*, 2002,18(9):1 141 - 1 142.
- [8] 孔丹莉,丁元林.非参数 Bootstrap 方法及其应用[J]. *数理医药学杂志*, 2006,19(3):232 - 233.
- [9] Efron B, Tibshirani RJ. Bootstrap measures for standard errors, confidence intervals and other measures of statistical accuracy[J]. *Stat Sci*, 1986,1(1):54 - 77.
- [10] 陈峰,陆守曾,杨珉,等. Bootstrap 估计及其应用[J]. *中国卫生统计*, 1997,14(5):5 - 7.
- [11] 刘勤,金丕焕. Bootstrap 方法及其在医学统计中的应用[J]. *中华预防医学杂志*, 1998,32(1):52 - 53.

(收稿日期:2008 - 10 - 10;编辑:张秀峰)