心理学报 2013, Vol. 45, No.1, 114−124
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2013.00114

校正的 Bootstrap 方法对概化理论方差分量及其变异量估计的改善*

黎光明[1,2] 张敏强[1]

([1]华南师范大学心理应用研究中心，广州 510631) ([2]广州大学教育学院心理系，广州 510006)

摘 要 Bootstrap 方法是一种有放回的再抽样方法，可用于概化理论的方差分量及其变异量估计。用 Monte Carlo 技术模拟四种分布数据，分别是正态分布、二项分布、多项分布和偏态分布数据。基于 p×i 设计，探讨校正的 Bootstrap 方法相对于未校正的 Bootstrap 方法，是否改善了概化理论估计四种模拟分布数据的方差分量及其变异量。结果表明：跨越四种分布数据，从整体到局部，不论是“点估计”还是“变异量”估计，校正的 Bootstrap 方法都要优于未校正的 Bootstrap 方法，校正的 Bootstrap 方法改善了概化理论方差分量及其变异量估计。

关键词 概化理论; Bootstrap 方法；方差分量；方差分量变异量；蒙特卡洛模拟

分类号 B841

1 引言

概化理论，又称为方差分量模型，在心理与教育测量实践中有着广泛的应用。Bootstrap 方法，也称“自助法”，是一种有放回的再抽样方法，可用于概化理论的方差分量及其变异量估计(Brennan, 2001)。

Bootstrap 方法是美国斯坦福大学统计系 Efron (1979)提出的一种统计方法。这种方法是统计学上的新突破之一(Fan, 2003)，其实质是模拟了从原始数据(把它看作“总体”)中随机抽取大量样本的过程，既可以用于参数估计，也可用于非参数估计(Cui & Kolen, 2008)。使用 Bootstrap 方法的基本程序是：以原始数据(样本容量为 n) 为基础，在保证每个观察单位每次被抽到的概率相等(1/n)的情况下，作有放回的重复抽样，所得样本被称为 Bootstrap 样本。根据一个 Bootstrap 样本计算出相应的统计量，就得到参数的一个估计值。这样重复若干次(记为 B，常设 B = 1000)，就形成了一个参数的近似抽样分布。根据这个抽样分布，可以估计出平均值、标准差、相应的百分位数，进而获得参数的标准误和置信区间。在这个过程中，如果频数分布是正态的，可以用参数的标准误直接计算其 95%置信区间。如果频数分布不是正态的，可以用第 2.5 百分位数和第 97.5 百分位数来估计其 95%置信区间。

Bootstrap 对统计量的估计基本实现过程如下：首先，从服从 $D(\theta)$未知分布的原始数据中抽取 X_1, X_2, ……, X_n, 样本容量 n≤N (原样本容量)，每个 X_i 的几率为 1/N；然后，决定上一步中θ 的统计量 $\hat{\theta}$，如平均数、分位点、方差分量等等；再后，重复抽样 B 次(B 一般为 1000)，用第 1 样本估计 $\hat{\theta}_1^*$，用第 2 个样本估计 $\hat{\theta}_2^*$，……，用第 B 个样本估计 $\hat{\theta}_B^*$；最后，Bootstrap 估计：$\hat{\theta}_{Boot}=\sum_{b=1}^{B}\hat{\theta}_b^*/B$。根据 $\hat{\theta}_{Boot}=\sum_{b=1}^{B}\hat{\theta}_b^*/B$，可以求从与参数$\theta$ 的偏差(bias)和标准误：

$$Bais(\hat{\theta})=\hat{\theta}_{Boot}-\theta \quad (1)$$

收稿日期: 2011−11−25

* 教育部人文社会科学研究青年基金项目(12YJC190016)、全国教育科学“十二五”规划教育部重点课题(GFA111009)、广东省教育科学“十二五”规划 2011 年度研究项目(2011TJK161)。

通讯作者: 张敏强, E-mail: Zhangmq1117@yahoo.com.cn

$$se(\hat{\theta}_{Boot})=\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}_b^*-\hat{\theta}_{Boot}^*)^2}{B-1}} \quad (2)$$

如果 $\hat{\theta}_{Boot}$ 表示方差分量, 那么公式(1)表示方差分量(点)估计值与真值的偏差。相应地, 公式(2)则表示方差分量的标准误。对于方差分量标准误估计, B 取 50 至 200 次被认为是充分的(Efron & Tibshrani, 1986; Brennan, Harris, & Hanson, 1987)。Gao (1992)的研究表明, B 取 500 与取 2000 结果基本相同。本研究方差分量标准误估计 B (重复次数)取 1000, 其理由有三: 一是 1000 次更能保证估计值接近真值; 二是方差分量置信区间一般要求 1000, 为使两者统一, 所以选择 1000 次; 三是今天的计算机能够处理 1000 次的迭代。

虽然上述 Bootstrap 方法过程比较简单, 但是如何进行 Bootstrap 方法再抽样却是个问题。这是因为对于完全随机的 p×i 设计存在多种抽样, 不同的抽样方法抽取样本不同, 可能导致计算出的统计量(如标准误)有别。因此, 需要对 Bootstrap 方法再抽样进行考虑, 包括再抽样策略和未校正与校正策略。

第一, 再抽样策略。因为是完全随机的 p×i 设计, 进行再抽样既要考虑 p 和 i, 也要考虑残差项 r。根据 Bootstrap 再抽样原则, 可供考虑的 Bootstrap 策略有 10 种(Wiley, 2001), 分别是 boot-p、boot-i、boot-pi、boot-pr、boot-ir、boot-pir、boot-effects、boot-oneway、boot-individual 和 boot-main。前 7 种再抽样策略是“单一”的再抽样策略, 后 3 种再抽样策略是“综合”的再抽样策略。其中 boot-p 表示固定 i 和 r, 仅考虑对 p 抽样; boot-i 表示固定 p 和 r, 仅考虑对 i 抽样; Boot-pi 表示固定 r, 对 p 和 i 同时抽样; Boot-pr 表示固定 i, 考虑对 p 和 r 抽样; boot-ir 表示固定 p, 考虑对 i 和 r 抽样; boot-pir 表示考虑对 p、i 和 r 同时抽样; Boot-effects 表示用“概率机制”P 来抽取随机效应样本的一种方法(Efron & Tibshrani, 1993; Wiley, 2001; Othman, 1995); boot-oneway 是一种再抽样策略, 要求对比 p 和 i 的样本容量, 按照较大者进行“单一”的再抽样策略, 如 boot-p 或 boot-i, 但不可能是 boot-pi (Leucht & Simth, 1989)。boot-individual 表示对 p 使用 boot-p, 对 i 使用 boot-i, 对 pi 使用 boot-pi, 而 boot-main 则表示对 p 和 pi 使用 boot-p, 对 i 使用 boot-i (Othman, 1995)。Wiley (2001)认为 boot-effects 对于不平衡的数据不适合, 因此这种再抽样策略在实践中受到限制。在考虑残差抽样的几种策略(boot-pr、boot-ir、boot-pir)中, r 被作为一个随机因素。

第二, 未校正(non-adjusted)与校正(adjusted)策略。事实上, 对于上述各种再抽样策略, 因为是有放回抽样, 存在重复个案, 可能导致方差分量估计过大或过小。其中的一个理由是, 对于完全随机的 p×i 设计, $\hat{\sigma}_p^2$ 和 $\hat{\sigma}_i^2$ 的无偏估计公式如下(Brennan, Harris, & Hanson, 1987; Brennan, 2001; Brennan, 2007):

$$\hat{\sigma}_p^2=\frac{1}{n_i(n_i-1)}\sum_{i\neq}\sum_{i'}\left[\sum_{p}\frac{(X_{pi}-\bar{X}_i)(X_{pi'}-\bar{X}_{i'})}{n_i-1}\right] \quad (3)$$

$$\hat{\sigma}_i^2=\frac{1}{n_p(n_p-1)}\sum_{p\neq}\sum_{p'}\left[\sum_{i}\frac{(X_{pi}-\bar{X}_p)(X_{p'i}-\bar{X}_{p'})}{n_p-1}\right] \quad (4)$$

Brennan 等人认为公式(3)和公式(4)中的 $\hat{\sigma}_p^2$ 和 $\hat{\sigma}_i^2$ 容易被高估, 特别是当 n_i 过小时。Brennan 等人(1987)、Leucht 和 Smith (1989)都使用了 Bootstrap 的校正方法。Brennan 等人认为, 为了给出较好的概化理论方差分量估计值, 需要对 boot-main 中的 $\hat{\sigma}_p^2$ 和 $\hat{\sigma}_{pi}^2$ 分别乘以 n/(n−1), 对 $\hat{\sigma}_i^2$ 乘以 k/(k−1), 分别予以校正。Leucht 和 Smith (1989)给出了以下两个公式:

$$\hat{\sigma}_{A(boot)}^2=\frac{a-1}{a}\hat{\sigma}_A^2 \quad (5)$$

$$\hat{\sigma}_{AB,e(boot)}^2=\frac{a-1}{a}\hat{\sigma}_{AB,e}^2 \quad (6)$$

公式(5)和(6)其实与 Brennan 等人(1987)乘以校正系数估计方差分量等价。

Brennan, Leucht 和 Othman 等人均认为对概化理论方差分量的估计存在“最佳”的 Bootstrap 策略。但是, Wiley (2001)给出了完全随机的 p×i 设计下各种再抽样策略下的校正公式推导, 否定了这种说话, 即对于方差分量估计不存在“最佳”情况。也就是说, 如果对原来 Bootstrap 方法再抽样策略进行校正, 那么所有再抽样策略都能够准确估计方差分量。Wiley (2001)的研究具有极其重要的意义, Wiley 从理论上推导出如何对各种再抽样策略进行有效校正。Wiley (2001)认为在完全随机的 p×i 设计下, 虽然 Bootstrap 方法进行方差分量点估计不存在“最佳”策略, 但对于方差分量变异量的估计情况就不同了。因此, 需要探讨未校正的 Bootstrap 方法和校正的 Bootstrap 方法对于估计方差分量标准误和置信区间之间存在的差别。本研究借鉴和改进了 Wiley 的成果, 进一步采用了校正的 Bootstrap 方法来估计概化理论方差分量的变异量。

在 Bootstrap 方法中，概化理论方差分量置信区间估计使用的方法是分位数方法。这种方法用每个抽取样本的统计量表示百分位数，如 10%的分位数或 90%的分位数。如果能够用某个样本的方差分量平均数来表示最后所得方差分量的估计值。那么，同样的道理，我们也可以求取所有样本 10%或 90%方差分量的平均值来表示最后所得的 10%或 90%的方差分量。这样做的好处是，可以获得置信区间的上下限，以此表示置信区间。这个过程可以作如下表述(Efron, 1982)：对于构建某个统计量的 100(1−α)%置信区间，首先产生 $\hat{\theta}_b^*$ 的直方图，直方图能够表示出待估参数 θ 的实际分布；然后在这个实际分布中找到两个划界点：$\theta_{low(\alpha)}$ 和 $\theta_{up(\alpha)}$，并且 $P(\hat{\theta}_b^*) < \theta_{low(\alpha)} = \alpha/2$，$P(\hat{\theta}_b^*) > \theta_{up(\alpha)} = \alpha/2$。因此，对于某个统计量的 100(1−$\alpha$)% 置信区间，可以表示为：$\theta_{low(\alpha)} < \theta < \theta_{up(\alpha)}$，这种方法称之为百分位数方法(empirical percentile procedure)。对于方差分量置信区间估计，除了百分位数方法外，还可以使用 Bootstrap-t、Bootstrap-BCa (Bias-Corrected and Accelerated)和 Bootstrap-ABC (Approximate Bootstrap Confidence intervals)方法(Wiley, 2001, p. 30)。

根据 Bootstrap 方法的实现过程，一些学者(Leucht & Smith, 1989; Othman, 1995; Brennan, 2001; Wiley, 2001)已注意到，虽然 Bootstrap 方法乘以一定校正系数后对方差分量估计有所改善，但存在以下问题：

第一，未见学者对于校正的 Bootstrap 方法是否改善概化理论方差分量变异量估计，作出详细讨论。Wiley (2001)注意到校正的 Bootstrap 方法对于方差分量估计具有等价性(equivalence)，这种等价性表现在各种校正的 Bootstrap 策略估计的方差分量相当一致。但这却不能拓广至方差分量变异量。然而，Wiley 的研究仅是比较了各种 Bootstrap 策略，却没有对校正的与未校正的 Bootstrap 方法进行比较。

第二，鲜有学者真正意义上对非正态分布数据进行过比较，特别是对校正的和未校正的 Bootstrap 方法。Brennan、Leucht、Othman 和 Wiley 等人的研究，用模拟数据进行方法之间的比较，仅限于正态分布数据。但是，非正态分布数据具有常见性，一些考试数据，如选择题，就是 0 或 1 记分，是二项分布数据，一些心理测验数据，用 Likert 形式评分，如 0~4 分，就是多项分布数据，缺乏探讨非正态分布数据形式下方法之间的比较，显得不足。

第三，一些学者的研究仍不充分。黎光明和张敏强(2009a)认为，数据分布对概化理论方差分量变异量估计有影响，与 Traditional、Jackknife、MCMC 三种方法相比，Bootstrap 方法具有方法的“跨分布性”，但是这种性质是建立在 Bootstrap 方法“分而治之”策略之下，boot-p、boot-pi、boot-i 策略分别估计 p、i、pi 的方差分量变异量最佳。然而，boot-p、boot-pi、boot-i 策略是使用校正的方法好，还是使用未校正的方法好，未作出进一步的讨论和说明。

即使 Bootstrap 方法具有方法的“跨分布性”，且具有统一规则，也仅仅解释了如何使用 Bootstrap 方法的问题。对于 Bootstrap 的 boot-p、boot-pi、boot-i 策略如何选择方差分量标准误和置信区间估计方法，却是一个值得更为深入讨论的问题。Wiley (2001)认为，未校正的 Bootstrap 方法与校正的 Bootstrap 方法在变异量估计上相当。这个结论是粗糙的，从整体而言，并没有完全针对某一种 Bootstrap 策略进行深入研究，更忽略了两种方法在方差分量变异量估计上的精确比较。因此，需要更为深入地探讨，在多种数据分布下(如正态分布、二项分布、多项分布和偏态分布)，校正的 Bootstrap 方法估计概化理论方差分量及其变异量是否优于未校正的 Bootstrap 方法。或者说，跨越不同数据分布，相比于未校正的 Bootstrap 方法，校正的 Bootstrap 方法是否对概化理论方差分量及其变异量估计有所改善。

2 方法

2.1 数据产生

基于 p×i 设计概化理论模型，运用蒙特卡洛数据模拟技术产生各种分布数据。数据模拟所使用的软件为 R 软件，产生的模拟数据包括：正态分布、二项分布、多项分布和偏态分布数据。

2.2 比较标准

设定两种比较标准：一是方差分量及标准误估计的比较标准，为“偏差” (bias)，计算方法是 $bias = (\hat{\theta}_i - \theta)$，$\hat{\theta}_i$ 表示统计量的估计值，θ 为参数，偏差的绝对值(称为“绝对偏差”)越大，表明估计值与参数的差异越大，结果越不可靠；二是方差分量置信区间估计的比较标准，为“80%置信区间包含率”，包含率越接近 0.80，表明结果越可靠(黎光明，张敏强, 2009b)。

2.3 分析工具

分析工具为 R 软件和 HyperbolicDist 软件包。借助这些软件或软件包, 自编完成研究程序。通过自编的R程序, 产生1000批次的模拟数据, 直接实现 Bootstrap 方法对方差分量及其变异量的估计, HyperbolicDist 软件包的作用在于生成服从一定偏度的偏态分布数据。

3 结果

3.1 不同分布数据的方差分量偏差

对正态、二项、多项和偏态分布模拟数据, 分别计算校正的和未校正的 Bootstrap 方法的 boot-p、boot-i、boot-pi 策略估计的三个方差分量偏差(如表 1)。通过比较估计的方差分量偏差, 可以看出不同数据分布下不同方法估计方差分量的性能差异。

表 1 中 vc.p_bias、vc.i_bias、vc.pi_bias 分别表示人的方差分量偏差、题目的方差分量偏差、人与题目交互作用(包括残差)的方差分量偏差。adjust 和 non-adjust 分别表示校正的和未校正的 Bootstrap 方法。boot-p、boot-i、boot-pi 表示 Bootstrap 方法的三种策略。Normal、Dichotomous、Polytomous、Skewnessed 分别表示正态分布、二项分布、多项分布和偏态分布数据。表 1 中的数值表示使用 Bootstrap 方法所得的方差分量偏差, 这些偏差是将估计值与参数值相减后获得的, 例如, 0.0289 表示对正态分布数据使用校正的 Bootstrap 方法的 boot-p 策略估计的方差分量为 4.2089, 参数值为 4.000, 两者相减后即为 0.0289。又如, –0.0114 表示对正态分布数据使用未校正的 Bootstrap 方法的 boot-p 策略估计的方差分量为 3.9886, 参数值为 4.000, 两者相减后即为–0.0114, 其它类似解释。

3.2 不同分布数据的方差分量标准误偏差

对正态、二项、多项和偏态分布模拟数据, 分别计算校正的和未校正的 Bootstrap 方法的 boot-p、boot-i 、boot-pi 策略估计的三个方差分量标准误偏差(如表 2)。通过比较估计的方差分量标准误偏差, 可以看出不同数据分布下不同方法估计方差分量标准误的性能差异。

表 2 中 SE (vc.p)_bias、SE (vc.i)_bias、SE (vc.pi)_bias 分别表示人的方差分量标准误偏差、题目的方差分量标准误偏差、人与题目交互作用(包括残差)的方差分量标准误偏差。其它解释同表 1。例如, –0.0192 表示对正态分布数据使用未校正的 Bootstrap 方法的 boot-p 策略估计的方差分量标准误为 1.0095, 参数值为 1.0287, 两者相减后即为 –0.0192, 其它类似解释。

3.3 不同分布数据的方差分量置信区间包含率

对正态、二项、多项和偏态分布模拟数据, 分别计算校正的和未校正的 Bootstrap 方法的 boot-p、boot-i、boot-pi 策略估计的三个方差分量置信区间包含率(如表 3)。通过比较估计的方差分量置信区间包含率, 可以看出不同数据分布下不同方法估计方差分量置信区间的性能差异。

表 3 中 CI (vc.p)、CI (vc.i)、CI (vc.pi)分别表示人的方差分量置信区间包含率、题目的方差分量置信区间包含率、人与题目交互作用(包括残差)的方差分量置信区间包含率。其它解释同表 1。例如, 0.769 表示对正态分布数据使用未校正的 Bootstrap

表 1　不同分布数据 Bootstrap 方法估计的方差分量偏差

分布		vc.p_bias		vc.i_bias		vc.pi_bias	
		adjust	non-adjust	adjust	non-adjust	adjust	non-adjust
Normal	boot-p	0.0289	–0.0114	0.0855	0.7250	–0.0493	–0.6888
	boot-i	0.0291	3.2265	0.0788	–0.7252	–0.0518	–3.2493
	boot-pi	0.0310	3.1563	0.0830	–0.1134	–0.0490	–3.8545
Dichotomous	boot-p	0.0000	–0.0001	0.0000	0.0002	–0.0001	–0.0026
	boot-i	0.0001	0.0174	0.0000	–0.0001	–0.0001	–0.0126
	boot-pi	0.0001	0.0171	0.0000	0.0000	–0.0001	–0.0150
Polytomous	boot-p	–0.0032	–0.0082	–0.0051	0.0070	–0.0011	–0.0131
	boot-i	–0.0032	0.0571	–0.0053	–0.0162	–0.0010	–0.0613
	boot-pi	–0.0032	0.0516	–0.0053	–0.0047	–0.0010	–0.0728
Skewnessed	boot-p	–0.0068	–0.0234	0.0208	0.0374	–0.0033	–0.0199
	boot-i	–0.0069	0.0587	0.0212	0.0373	–0.0033	–0.1019
	boot-pi	–0.0068	0.0587	0.0212	–0.0477	–0.0033	–0.1020

表 2 不同分布数据 Bootstrap 方法估计的方差分量标准误偏差

分布		SE (vc.p)_bias		SE (vc.i)_bias		SE (vc.pi)_bias	
		adjust	non-adjust	adjust	non-adjust	adjust	non-adjust
Normal	boot-p	–0.0192	–0.0293	–3.9095	–3.9099	–0.0172	–0.0379
	boot-i	0.4687	0.4204	–0.4488	–0.6963	0.0935	0.2082
	boot-pi	1.0676	1.0024	–0.0023	–0.2720	1.6159	1.3956
Dichotomous	boot-p	–0.0002	–0.0002	–0.0004	–0.0004	–0.0055	–0.0056
	boot-i	0.0035	0.0033	0.0001	0.0001	0.0001	–0.0004
	boot-pi	0.0059	0.0055	0.0000	0.0000	0.0022	0.0015
Polytomous	boot-p	–0.0025	–0.0035	–0.1670	–0.1670	0.0000	–0.0003
	boot-i	–0.0660	–0.0666	–0.0423	–0.0497	0.0048	0.0028
	boot-pi	0.0115	0.0101	–0.0377	–0.0453	0.0289	0.0251
Skewnessed	boot-p	–0.0366	–0.0399	–0.7381	–0.7381	–0.0010	–0.0018
	boot-i	–0.2913	–0.2918	–0.1755	–0.2074	0.0000	–0.0041
	boot-pi	–0.0179	–0.0215	–0.1651	–0.1975	0.0561	0.0479

表 3 不同分布数据 Bootstrap 方法估计的方差分量置信区间包含率

分布		CI (vc.p)		CI (vc.i)		CI (vc.pi)	
		adjust	non-adjust	adjust	non-adjust	adjust	non-adjust
Normal	boot-p	0.769	0.768	0.289	0.299	0.803	0.782
	boot-i	0.918	0.056	0.763	0.728	0.847	0.400
	boot-pi	0.988	0.209	0.799	0.790	0.970	0.617
Dichotomous	boot-p	0.768	0.765	0.391	0.396	0.470	0.469
	boot-i	0.964	0.006	0.771	0.747	0.799	0.485
	boot-pi	0.997	0.049	0.817	0.846	0.888	0.485
Polytomous	boot-p	0.783	0.783	0.210	0.259	0.783	0.752
	boot-i	0.372	0.306	0.666	0.679	0.838	0.358
	boot-pi	0.836	0.795	0.671	0.671	0.980	0.539
Skewnessed	boot-p	0.737	0.742	0.088	0.087	0.808	0.783
	boot-i	0.219	0.242	0.663	0.641	0.812	0.535
	boot-pi	0.767	0.786	0.672	0.651	0.958	0.779

方法的 boot-p 策略估计的方差分量置信区间包含率，其它类似解释。计算 80%置信区间包含率的方法是：通过判断参数是否落入 10%到 90%两分位点对应的方差分量之间，如果某次成功，则包含次数加 1，最后计算落入的总次数，并除以 1000，即为最后的包含率。

4 分析与讨论

4.1 不同分布数据的方差分量偏差分析

根据表 1 中四种分布数据 Bootstrap 方法的 boot-p、boot-i 、boot-pi 策略估计的三个方差分量偏差，可以绘出它们对应的偏差图，如图 1(a)~(d)所示。

从图 1(a)可以看出，对于正态分布数据，在人的方差分量偏差上，未校正的 boot-pi 和 boot-i 策略偏差分别为 3.2265 和 3.1563，相对较大，远离横轴，未校正的 boot-p 策略偏差为-0.0114，相对较小，几乎接近横轴，校正的 boot-p、boot-i、boot-pi 策略偏差分别为 0.0289、0.0291、0.0310，相对较小。在题目的方差分量偏差上，未校正的 boot-pi 策略偏差相对较小，未校正的 boot-i 和 boot-p 策略偏差相对较大，校正的 boot-p、boot-i 、boot-pi 策略偏差相对较小。在人与题目交互作用(包括残差)的方差分量偏差上，未校正的 boot-p、boot-i、boot-pi 策略偏差相对较大，校正的 boot-p、boot-i、boot-pi 策略偏差相对较小。对图 1(a)的分析可知，对于正态分布数据，校正的 Bootstrap 方法各种策略估计方差分量较为接近，策略间具有“等价性”，这与 Wiley (2001)的结论一致。

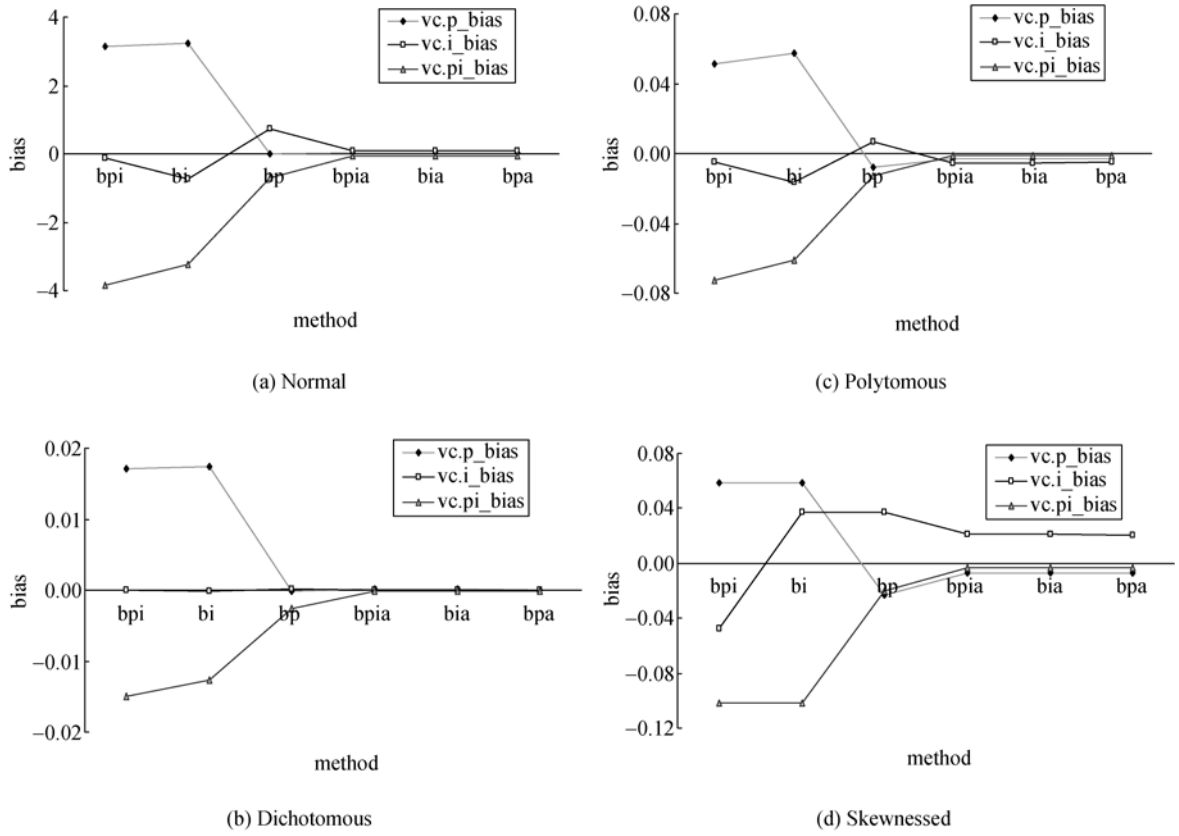

图 1　正态、二项、多项和偏态分布数据 Bootstrap 方法估计的方差分量偏差图

注：bpi 表示 boot-pi, bi 表示 boot-i, bp 表示 boot-p, bpia 表示 boot-piadj, bia 表示 boot-iadj, bpa 表示 boot-padj。

从图 1(b)可以看出，对于二项分布数据，在人的方差分量偏差上，未校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相对较大，校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相对较小。在题目的方差分量偏差上，校正的和未校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相当，相对较小。在人与题目交互作用(包括残差)的方差分量偏差上，校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相对较大，校正的 boot-p、boot-i、boot-pi 策略偏差相对较小。

从图 1(c)可以看出，对于多项分布数据，其解释可参考图 1(a)。

从图 1(d)可以看出，对于偏态分布数据，在人的方差分量偏差上，未校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相对较大，校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相对较小。在题目的方差分量偏差上，未校正的 boot-pi 策略方差分量偏差相对较大，未校正的 boot-p 和 boot-i 策略方差分量偏差相当，校正的 boot-p、boot-i、boot-pi 策略方差分量偏差相当，表现出 Wiley (2001)提出的“等价性”。但是，值得注意的是，偏差并没有接近横轴，这与图 1(a)~(c)结果不一致，这说明数据的“偏态性”会影响数据的方差分量估计精度，这是一些研究未注意到的地方(Othman, 1995; Wiley, 2001; Tong & Brennan, 2007)，对比校正的和未校正的 boot-p、boot-i、boot-pi 在题目上的方差分量偏差，前者仍然小于后者。在人与题目交互作用(包括残差)的方差分量偏差上，未校正的 boot-p、boot-i、boot-pi 策略偏差相对较大，校正的 boot-p、boot-i、boot-pi 策略方偏差相对较小。

4.2　不同分布数据的方差分量标准误偏差分析

根据表 2 中四种分布数据 Bootstrap 方法的 boot-p、boot-i、boot-pi 策略估计的三个方差分量标准误偏差进行配对样本的 t 检验，boot-p、boot-i、boot-pi 策略各有 12 对数据，这样总共有 36 对数据，对这 36 对数据进行 t 检验(张厚粲，徐建平，2004)，结果为 t=2.021，p=0.051，表明差异显著性检验相伴概率处于边缘显著水平($0.050 < p < 0.100$)，从总体趋势看，可以认为校正的和未校正的 boot-p、boot-i、boot-pi 策略在三个方差分量的标准误偏差存在差异，前者的偏差小于后者的偏差。

从表 2 可以看出，对于正态分布数据，校正的

和未校正的 boot-p 策略估计人的方差分量标准误偏差分别为-0.0192 和-0.0293, 绝对偏差相对较小, 而校正的和未校正的 boot-i 策略估计人的方差分量标准误偏差分别为 0.4687 和 0.4204, 校正的和未校正的 boot-pi 策略估计人的方差分量标准误偏差分别为 1.0676 和 1.0024, 绝对偏差相对较大。因此, 可以认为, 对于正态分布数据, boot-p 策略估计人的方差分量标准误最佳。同理, boot-pi 策略估计题目的方差分量标准误最佳, boot-p 策略估计人与题目交互作用(包括残差)的方差分量标准误最佳, boot-i 策略次佳。

在表 2 中, 依照正态分布数据的分析方法, 对于二项和多项分布数据, boot-p 策略估计人的方差分量标准误最佳, boot-pi 策略估计题目的方差分量标准误最佳, boot-i 策略估计人与题目交互作用(包括残差)的方差分量标准误最佳。对于偏态分布数据, boot-pi 策略估计人的方差分量标准误最佳, boot-p 策略次佳, boot-pi 策略估计题目的方差分量标准误最佳, boot-i 策略估计人与题目交互作用(包括残差)的方差分量标准误最佳。

根据以上分析, 整合四种分布数据, boot-p 策略估计 SE (vc.p)最佳, boot-pi 策略估计 SE (vc.i)最佳, boot-i 策略估计 SE (vc.pi)最佳, 这符合 Bootstrap 方法“分而治之”策略规则。根据这种“分而治之”策略规则, 对于概化理论方差分量标准误估计, 应该探索是否校正的方法优于未校正的方法, 如果前者更好, 应该采用校正的方法。接下来, 按照“分而治之”策略规则, 使用校正的和未校正的方法, 分别分析 boot-p 策略估计 SE (vc.p), boot-pi 策略估计 SE (vc.i), boot-i 策略估计 SE (vc.pi), 结果如图 2(a)~(c)所示。

从图 2(a)可以看出, 未校正的和校正的 boot-p 策略估计 p 的方差分量标准误偏差不一致, 表明两种方法估计 p 的方差分量标准误有差别, 校正的 boot-p 策略的标准误偏差更小。与未校正的方法相比, 校正的 boot-p 策略估计 p 的方差分量标准误相对更为准确。跨越不同分布, 发现偏态分布数据估计的标准误偏差最大, 然后是正态分布和多项分布, 标准误最小是二项分布。跨越不同分布估计的标准误偏差不同, 其原因与数据本身有关, 例如, 正态分布数据与二项分布数据, 其标准误是不同的, 二项分布数据是 1 或 0 的形式, 数据本身较小, 求出的方差分量较小, 其标准误偏差自然也较小。但是, 正态分布则不同, 它的数据可能较大(如 20), 这样的数据产生的方差分量较大, 其标准误偏差也较大, 其它分布情况类似。

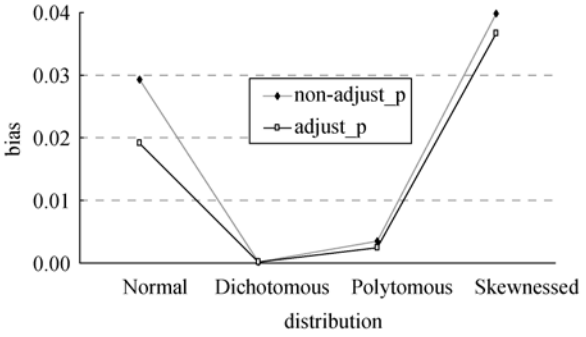

图2(a) boot-p策略估计SE (vc.p)

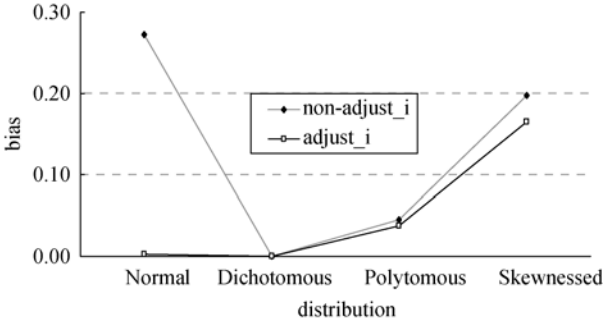

图2(b) boot-pi策略估计SE (vc.i)

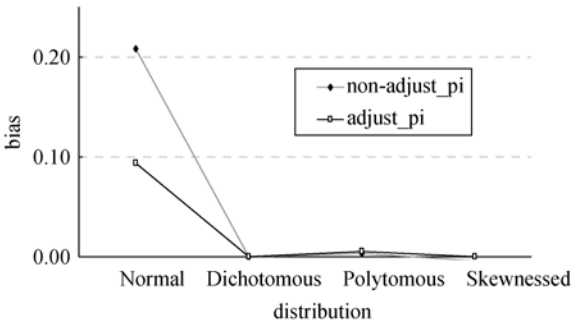

图2(c) boot-i策略估计SE (vc.pi)

从图 2(b)可以看出, 未校正的和校正的 boot-pi 策略估计 i 的方差分量标准误偏差存在差别, 特别是正态分布数据(Normal), 校正的和未校正的 boot-pi 策略估计 i 的方差分量标准误偏差分别是–0.0023 和–0.2720, 前者绝对偏差值明显小于后者, 其它结果与图 2(a)近似, 表明未校正的和校正的 boot-pi 策略估计 i 的方差分量标准误偏差表现出不一致, 从整体上看, 校正的方法要优于未校正的方法。在图 2(c)中, 不同数据分布下估计的方差分量标准误偏差也存在差别, 正态分布数据两种方法差异较大, 其它分布差异则不显著。

未校正的和校正的 boot-p、boot-i、boot-pi 策

略在估计 p、i、pi 的方差分量标准误时，结果存在差异，校正的方法相对较好，也考虑到校正的 boot-p、boot-i、boot-pi 策略在方差分量点估计上要优于未校正的 boot-p、boot-i、boot-pi 策略。因此，不同的数据分布下使用校正的 boot-p、boot-i、boot-pi 策略来估计 p、i、pi 的方差分量标准误，更为妥当些。

4.3　不同分布数据的方差分量置信区间包含率分析

本研究 Bootstrap 使用了 PC 方法来估计方差分量置信区间，并用包含率的大小来估价各种 Bootstrap 策略的准确性, boot-p、boot-i、boot-pi 策略隶属于 Bootstrap 策略，也是采用了 PC 方法来估计方差分量置信区间。如果包含率越接近 0.800，那么结果越准确。

根据表 3 的结果，为了对比校正的和未校正的 Bootstrap 方法在估计方差分量置信区间包含率时的差异，我们作了两种方法方差分量置信区间包含率散点图，如图 3(a)~(c)所示。

从图 3(a)和图 3(c)可以看出，校正的和未校正的 Bootstrap 方法散点无规律性，表明两种方法数据等级之间存在不一致性，两种方法存在差异。从图 3(b)可以看出，校正的和未校正的 Bootstrap 方法散点几乎在一条直线上，两种方法估计题目的方差分量置信区间包含率差异较小。跨越四种分布数据，分别计算 CI (vc.p)、CI (vc.i)、CI (vc.pi)校正的和未校正的 Bootstrap 方法的相关系数，其值分别为 −0.112 (p=0.729)、0.996 (p=0.000)、0.174 (p=0.589)，表明校正的和未校正的 Bootstrap 方法在 CI (vc.p)和 CI (vc.pi)存在差异，与上述散点图结果一致。

将校正的和未校正的 Bootstrap 方法在 CI (vc.p)、CI (vc.i)和 CI (vc.pi)的包含率与 0.800 相减，所得值表示偏差值，偏差越大，表示偏离参数越大，结果越不可靠。用配对样本 T 检验比较校正的和未校正的 Bootstrap 方法包含率的偏差差异，结果如表 4 所示。

从表 4 可知，校正的和未校正的 Bootstrap 方法在 CI (vc.p)、CI (vc.i)和 CI (vc.pi)的包含率偏差存在显著性差异且效果量相对较大(t=3.693，p≤0.001，d=0.764)，与未校正的方法相比，校正方法的包含率偏差更小，这表明从总体上看，校正的 Bootstrap 方法在估计方差分量置信区间时，优势更为明显。

类似于分析方差分量标准误，对于置信区间估计, Bootstrap 方法仍然遵循“分而治之”策略。分析这些策略目的是在分析总(整)体结果之后，对局部的考虑，探讨这些策略是否也服从总体所得到的结果。

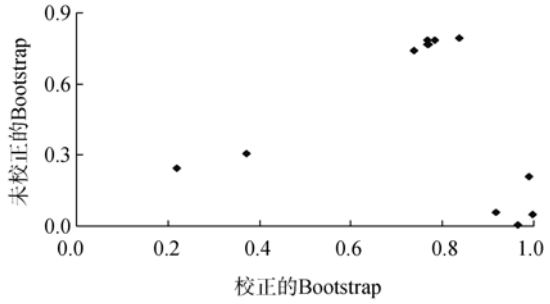

图3(a)　两种方法在CI (vc.p)上的散点图

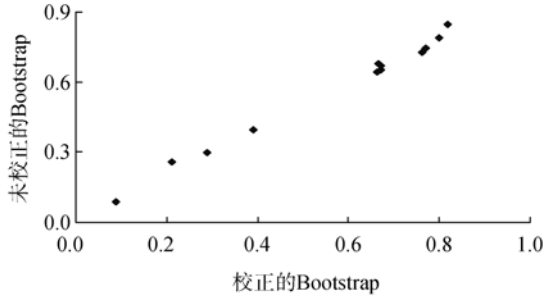

图3(b)　两种方法在CI (vc.i)上的散点图

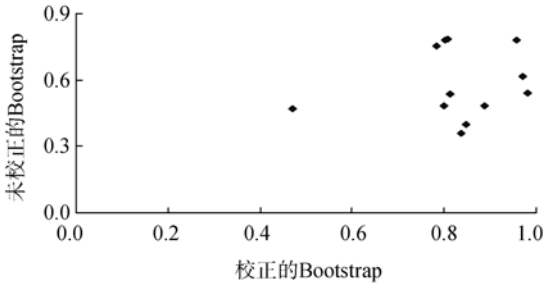

图3(c)　两种方法在CI (vc.pi)上的散点

表 4　校正的和未校正的 Bootstrap 方法包含率偏差的差异显著性检验

Bootstrap 方法	平均数	标准差	t	p
adjust	0.0813	0.2372	3.693	0.001
non-adjust	0.2643	0.2556		

与对表 2 的分析类似，对于表 3 中的正态和偏态分布数据, boot-p 估计 CI (vc.p)最佳, boot-pi 估计 CI (vc.i)最佳，boot-p 估计 CI (vc.pi)最佳，boot-i 次佳。对于表 3 中的二项和多项分布数据，boot-p 估计 CI (vc.p)最佳，boot-pi 估计 CI (vc.i)最佳，boot-i 估计 CI (vc.pi)最佳。综合考虑可认为，跨越四种分布数据，boot-p 估计 CI (vc.p)最佳，boot-pi 估计 CI

(vc.i)最佳, boot-i 估计 CI (vc.pi)最佳。

与上述标准误偏差分析类似, 在整体分析之后, 再考虑“局部”, 即根据这种“分而治之”策略规则, 对于概化理论方差分量置信区间估计, 应该探索是否校正的方法优于未校正的方法, 如果前者更好, 表明应该采用校正的方法。接下来, 按照“分而治之”策略规则, 使用校正的和未校正的方法, 分别分析 boot-p 策略估计 CI (vc.p), boot-pi 策略估计 CI (vc.i), boot-i 策略估计 CI (vc.pi), 结果如图 4(a)~(c)所示。

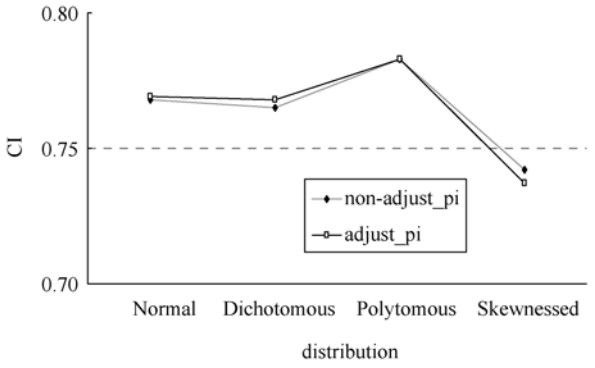

图4(a) boot-p策略估计 CI (vc.p)

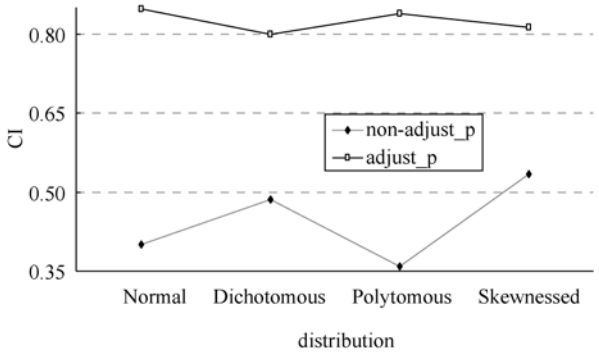

图4(b) boot-pi策略估计 CI (vc.i)

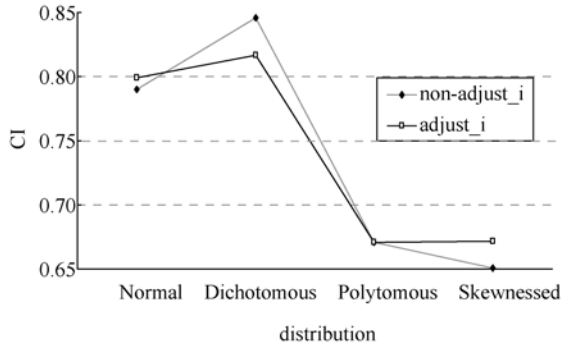

图4(c) boot-i策略估计 CI (vc.pi)

从图 4(a)可以看出, boot-p 策略估计 CI (vc.p), 在正态和二项分布数据下, 校正的方法与未校正的方法相比, 包含率更接近参数值 0.800, 表明校正的方法略优于未校正的方法; 在多项分布下, 两种方法相当; 在偏态分布下, 未校正的方法略优于校正的方法。

从图 4(b)可以看出, boot-pi 策略估计 CI (vc.i), 在正态、二项和偏态分布数据下, 校正的方法与未校正的方法相比, 包含率更接近参数值 0.800, 表明校正的方法略优于未校正的方法; 在多项分布下, 两种方法相当。

从图 4(c)可以看出, boot-i 策略估计 CI (vc.pi), 在正态、二项、多项和偏态分布数据下, 校正的方法与未校正的方法相比, 包含率更接近参数值 0.800, 而未校正的方法明显远离参数值 0.800, 表明校正的方法明显优于未校正的方法。

5 结论

(1)对于方差分量(点)估计, 跨越四种分布数据, 与未校正的 Bootstrap 方法相比, 校正的 Bootstrap 方法各种策略估计方差分量较为接近, 策略间具有“等价性”, 估计的方差分量偏差相对较小。

(2)对于方差分量标准误估计, 跨越四种分布数据, 从整体趋势看, 校正的 Bootstrap 方法与未校正的 Bootstrap 方法估计三个方差分量标准误偏差存在差异, 前者的偏差小于后者的偏差。从“分而治之”策略局部看, 与未校正的 boot-p、boot-pi 和 boot-i 策略分别估计 p、i、pi 的方差分量标准误相比, 校正的方法相对较好。

(3)对于方差分量置信区间估计, 跨越四种分布数据, 校正的 Bootstrap 方法和未校正的 Bootstrap 方法包含率偏差存在显著性差异, 校正的方法包含率偏差更小, 从整体上看, 校正的 Bootstrap 方法优势更为明显。从“分而治之”策略局部看, 校正的方法也较好。

(4)对上述结论进一步推论: 跨越四种分布数据, 从“整体”到“局部”, 不论是“点估计”还是“变异量估计”, 一致表明: 校正的 Bootstrap 方法要优于未校正的 Bootstrap 方法, 校正的 Bootstrap 方法改善了概化理论方差分量及其变异量估计。

参 考 文 献

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2007). Unbiased estimates of variance components with bootstrap procedures. *Educational and Psychological Measurement, 67*(5), 784–803.

Brennan, R. L., Harris, D. J., & Hanson, B. A. (1987). *The*

bootstrap and other procedures for examining the variability of estimated variance components in testing contexts (ACT Research Report Series87-7). Iowa City, IA: American College Testing Program.

Cui, Z. M., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32*(4), 334–347.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM CBMS-NSF Monograph 38.

Efron, B., & Tibshrani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54–57.

Efron, B., & Tibshrani, R. J. (1993). *An introduction to the Bootstrap*. New York. Chapman and Hall.

Fan, X. T. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63*(1), 24–50.

Gao, X. H. (1992). *Generalizability of a state-wide science performance assessment.* Unpublished doctorial dissertation. University of California.

Leucht, R. M., & Smith, P. L. (1989). *The effects of bootstrapping strategies on the estimation of variance components.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Li, G. M., & Zhang, M. Q. (2009a). A Cross-distribution study: Estimating the variability of estimated variance components for Generalizability Theory. In *The 1st psychology doctoral forum of china collection of summaries* (pp. 290–294). Beijing: Beijing Normal University.

[黎光明, 张敏强. (2009a). 一项跨分布研究: 基于概化理论的方差分量变异量估计. 见 首届全国心理学博士学术论坛论文集(pp. 290–294). 北京: 北京师范大学.]

Li, G. M., & Zhang, M. Q. (2009b). Estimating the variability of estimated variance components for Generalizability Theory. *Acta Psychologica Sinica, 41*(9), 889–901.

[黎光明, 张敏强. (2009b). 基于概化理论的方差分量变异量估计. *心理学报, 41*(9), 889–901.]

Othman, A. R. (1995). *Examining task sampling variability in science performance assessments.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67*(5), 804–817.

Wiley, E. W. (2001). *Bootstrap strategies for variance component estimation: Theoretical and empirical results.* Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Zhang, H. C., & Xu, J. P. (2004). *Modern psychological and educational statistics*. Beijing, China: Beijing Normal University Press.

[张厚粲, 徐建平. (2004). *现代心理与教育统计学*. 北京: 北京师范大学出版社.]

Using Adjusted Bootstrap to Improve the Estimation of Variance Components and Their Variability for Generalizability Theory

LI Guangming[1,2]; ZHANG Minqiang[1]

([1] *Center for Studies of Psychological Application, South China Normal University, Guangzhou* 510631, *China*)

([2] *Department of Psychology, School of Education, Guangzhou University, Guangzhou* 510006, *China*)

Abstract

Bootstrap is a returned re-sampling method used to estimate the variance component and their variability. Adjusted bootstrap method was used by Wiley in p×i design for normal data in 2001. However, Wiley did not compare the difference between adjusted method and unadjusted method when estimating the variability.

To expand Wiley's 2001 study, our study applied Monte Carlo method to simulate four distribution data. The aim of simulation is to explore the effects of four different estimation methods when estimating the variability of estimated variance components for generalizability theory. The four distribution data are normal distribution data, dichotomous distribution data, polytomous distribution data and skewed distribution data. It is common that researchers focus on normal distribution data and neglect non-normal distribution data, yet non-normal distribution data could always be seen in tests such as TOEFL and GRE. There are several methods to estimate the variability of variance components, including traditional, bootstrap, jackknife and Markov Chain Monte Carlo (MCMC). Former research by Li and Zhang (2009) shows that bootstrap method is significantly better than traditional, jackknife, and MCMC methods in estimating the variability for four distribution data.

Bootstrap method has superior cross-distribution quality when estimating the variability of estimated variance components. Li and Zhang (2009) also suggest that bootstrap method should be adopted with a “divide-and-conquer” strategy to obtain good estimated standard error and estimated confidence interval and the criteria of such strategy should be set to: boot-p for person, boot-pi for item, and boot-i for person and item. However, it is unclear that which of the bootstrap methods (adjusted and unadjusted) is better for boot-p, boot-pi, and boot-i. Therefore, our study intends to probe into this comparison as well.

This aim of the study is to explore whether adjusted bootstrap method is superior to unadjusted method in improving the estimation of variance components and their variability relative for generalizability theory. The simulation is implemented in *R* statistical programming environment. To simulate skewed data, *HyperbolicDist* package is used. Some criteria are set to compare the four methods. The bias is considered when variance components and their standard errors are estimated. The smaller the absolute bias is, the more reliable the result is. The criterion of confidence intervals is “80% interval coverage”. If the “80% interval coverage” is closer to 0.80, the confidence interval is more reliable.

The results indicate that for four distribution data, adjusted bootstrap method is superior to unadjusted bootstrap method whether in point estimation of variance components or in variability estimation of variance components. For its improvement of the estimation of variance components and their variability for generalizability theory, adjusted bootstrap should be adopted as soon as possible.

Key words Generalizability Theory; Bootstrap method; Variance component; variability of estimated variance components; Monte Carlo simulation