心理学探新 2020,Vol.40,No.5,431 -437
PSYCHOLOGICAL EXPLORATION

# 概化理论方差分量及其变异量估计:跨分布的模拟研究

甄锋泉 张敏强 刘 颖
(华南师范大学心理学院,广州 510631)

**摘 要**:为考察概化理论中方差分量及其变异量估计的准确性,采用模拟研究的方法,探究Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法在 $p\times i\times h$ 和 $p\times(i:h)$2 种双侧面设计和正态、二项、多项、偏态分布 4 种数据类型下的表现。结果显示:(1)4 种方法均能准确估计方差分量;(2)估计方差分量的标准误时,若数据正态分布,Traditional 法最优,非正态分布时 Bootstrap 法最优;(3)估计方差分量的 90% 置信区间时,Bootstrap 法在不同分布的数据下表现稳定,但容易受到侧面水平数的影响。综合来说,若数据呈正态分布,建议选用 Traditional 法;若数据呈非正态分布,建议选用 Bootstrap 法。

**关键词**:概化理论;方差分量;模拟研究;数据分布
**中图分类号**:B841.2 **文献标识码**:A **文章编号**:1003 -5184(2020)05 -0431 -07

## 1 引言

概化理论(Generalizability Theory,GT)是一种测量行为可靠性的统计理论。作为经典测量理论的拓展,它引入了实验法和方差分析(Analysis of Variance,ANOVA)的思想,将误差分解到各个侧面及其交互作用,然后通过调整侧面的水平数以及侧面之间的关系,观察概化系数和可靠性指数的变化,为决策提供依据。由于在分解和控制误差方面的优势,概化理论在实践中被应用到考试研究(Karami,2012;Kellermargulis,Mercer,& Thomas,2015;Lin,2017)、心理测验分析(Jiang & Raymond,2018;Mantzicopoulos,French,& Patrick,2018)、教学评价(Casabianca,Lockwood,& Mccaffrey,2015;Meyer,Liu,& Mashburn,2014)及人才测评(Nalbantoglu Yilmaz,2017;李向阳,2015;姚若松,刘泽,赵葆楠,苗群鹰,2015)等方面。

分解和控制测量误差的关键在于方差分量的准确估计。近年来,研究者发展出一系列方法估计方差分量及其变异量。应用最多的是 Traditional 法,在此基础上又发展出 TBGJL 法估计方差分量的置信区间(Burdick & Graybill,1992;Ting,Burdick,Graybill,Jeyaratnam,& Lu,1990),但使用 Traditional 法需要分数效应服从多元正态分布(Brennan,2001;Searle,Casella,& Mcculloch,1992)。故研究者又引入了其他非参数的方法,如 Jackknife 法、Bootstrap 法和 MCMC 法,它们的共同之处在于不需要对分数效应的分布进行前提假设。

为了比较估计方法的性能,研究者进行了一系列模拟研究。Brennan、Harris 和 Hanson(1987)基于概化理论 $p\times i$ 设计,比较了 Traditional 法、Jackknife 法和 Bootstrap 法的表现,结果发现,Jackknife 法在正态分布和二项分布上均表现较优,Traditional 法仅在正态分布上表现良好,Bootstrap 法比较复杂,需要分类讨论。该研究仅考虑了正态分布和二项分布两种数据类型,其中二项分布数据是从实证数据中抽取的,而非模拟产生。Othman(1995)基于概化理论 $p\times i$ 设计,比较了 Traditional 方法和 Bootstrap 方法的表现。结果显示,在正态分布的情况下,Traditional 方法的表现优于 Bootstrap 方法,但在二项分布的情况下,没有找到准确的估计方法。然而,该研究中二项分布数据的方差分量需要进行校正,由于校正系数不能提前获取,这种校正的方法没有被后续的研究者采纳。Wiley(2001)对比了矫正的 Bootstrap 法与 Traditional 法的表现,发现正态分布情况下,在方差分量估计上,Traditional 法与矫正的 Bootstrap 法表现接近,但在对应的标准误和置信区间估计上,Traditional 法更优。该研究基于概化理论 $p\times i$ 设计,考虑了正态分布和二项分布数据,但由于二项分布数据是实证数据,无法与真值进行比较。Mao,Shine 和 Brennan(2005)基于概化理论 $p\times i$ 设计比较了 Traditional 法和 MCMC 法的表现,发现在二项分布情况下,Traditional 法和 MCMC 法均容易出现较大偏差。该研究仅考虑了正态分布和二项分布两种数据类型,其中二项分布数据是从考试数据中直接抽取的。黎光明(2010)糅合前人研究,对比了 Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法的表现,认为 Bootstrap 法表现最佳,具有跨分布的稳定性。该研究仅考虑了概化理论 $p\times i$ 设计,并且

Jackknife 方法分成 Jack - p、Jack - i 和 Jack - pi 策略来报告估计结果，没有对伪值进行合成，无法与前人研究对接（Brennan et al.，1987；Feng，2002）。

回顾以往文献，目前方差分量估计的研究还存在以下局限性：

（1）实证研究中，为了在充分发挥概化理论优势的同时平衡研究成本的限制，研究者多采用两侧面完全交叉设计或者两侧面嵌套设计（Clayson & Miller，2016；Medvedev，Krägeloh，Narayanan，& Sieg-ert，2017），而大多数的模拟研究则集中探讨单侧面设计。双侧面设计包含的侧面更多，侧面之间的关系更加复杂，参数估计的难度也相应增大，所以基于单侧面设计的模拟研究，其结论推广到双侧面设计时可能存在偏差。因此，需要进行模拟研究，进一步探讨概化理论双侧面设计的方差分量及其变异量估计。

（2）大部分模拟研究仅涉及一种或两种估计方法（Luecht & Smith，1989；Mao et al.，2005；Moore，2010；Othman，1995；Tong & Brennan，2007；Wiley，2001），缺少多种估计方法的系统比较，因此，有必要创设不同的模拟条件，同时探讨概化理论多种估计方法的表现。

（3）对方差分量变异量的探讨大多停留在标准误上，缺乏对置信区间的探讨（Feng，2002；Mao et al.，2005；Tong & Brennan，2007）。以往研究指出，方差分量服从何种分布无法确定（Searle，1971；Searle et al.，1992），很有可能呈一个非对称分布（Moore，2010），因此，仅报告方差分量的标准误不足以说明方差分量的变异性，需要进一步探讨方差分量的置信区间估计。

综上，本文将在两种双侧面研究设计（$p \times i \times h$ 和 $p \times [i:h]$ 设计）、四种数据分布（正态、二项、多项和偏态分布）条件下，比较 Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法对方差分量及其标准误和置信区间估计的准确性。

## 2 工具和方法

首先基于 $p \times i \times h$ 和 $p \times (i:h)$ 设计，使用蒙特卡洛模拟生成正态分布、二项分布、多项分布和偏态分布数据各 1000 批数据。参考前人文献（Tong & Brennan，2007；黎光明，2010），$p$、$i$、$h$ 三个侧面的水平数分别设定为 100、20、4。然后，自编程序实现 Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法，分别估计每一批数据的方差分量，标准误和 90% 置信区间。分析工具采用 JAGS 软件、R 软件以及其中的 R2jags、coda 和 HyperbolicDist 软件包。其中，JAGS 软件用于实现 MCMC 法，HyperbolicDist 软件包用于产生偏态分布数据。

### 2.1 数据生成

正态分布数据。假设分数效应服从多元正态分布，$p \times i \times h$ 和 $p \times (i:h)$ 设计下的观察分数可以分别用以下公式来表示：

$$X_{pih} = \sigma(p)z_p + \sigma(i)z_i + \sigma(h)z_h + \sigma(pi)z_{pi} + \sigma(ph)z_{ph} + \sigma(ih)z_{ih} + \sigma(pih)z_{pih} \quad (1)$$

$$X_{pi:h} = \sigma(p)z_p + \sigma(h)z_h + \sigma(ph)z_{ph} + \sigma(i:h)z_{i:h} + \sigma(pi:h)z_{pi:h} \quad (2)$$

其中，$z_p$、$z_i$、$z_h$、$z_{pi}$、$z_{ph}$、$z_{ih}$ 和 $z_{pih}$ 表示独立随机的标准正态分布，$\sigma(p)$、$\sigma(i)$、$\sigma(h)$、$\sigma(pi)$、$\sigma(ph)$、$\sigma(ih)$ 和 $\sigma(pih)$ 表示方差分量的参数值。

二项分布数据。对于 $p \times i \times h$ 和 $p \times (i:h)$ 设计，分别使用公式（1）和公式（2）模拟正态分布数据，然后令其中小于等于 1 的值为 0，其余为 1，这样就得到了两种设计下的二项分布数据。

多项分布数据。类似于正态分布，多项分布的数据可以通过若干个二项分布的线性累加获得。$p \times i \times h$ 设计和 $p \times (i:h)$ 设计的观测分数可以分别用下面两条公式来获得：

$$X_{pih} = B(2,0.7966) + B(1,0.8570) + B(1,0.98785) + B(2,0.7313) + B(1,0.98579) + B(1,0.9975) + B(2,0.8025) \quad (3)$$

$$X_{pi:h} = B(1,0.713) + B(1,0.843) + B(1,0.93) + B(2,0.713) + B(5,0.63) \quad (4)$$

偏态分布数据。$p \times i \times h$ 设计和 $p \times (i:h)$ 设计的观测分数可以分别用下面两条公式来获得：

$$X_{pih} = GH(p) + GH(i) + GH(h) + GH(pi) + GH(ph) + GH(ih) + GH(pih) \quad (5)$$

$$X_{pi:h} = GH(p) + GH(h) + GH(ph) + GH(i:h) + GH(pi:h) \quad (6)$$

其中，GH 为广义双曲线分布（Generalized Hyperbolic Distribution，GH），参考黎光明（2010）设定 $\lambda = 1$，$\mu = 0$，$\delta = 1$，$\alpha = 3$，$\beta = -2$。

参数值的设置要考虑模拟研究的便利性，因此有一定的主观性，但更重要的是要尽可能与客观实际相符，不可胡乱设置。参考 Tong 和 Brennan（2007），方差分量及标准误的参数值设定如表 1、表 2 所示。

表1　$p \times i \times h$ 设计的方差分量和标准误设定

| | 侧面 | $p$ | $i$ | $h$ | $pi$ | $ph$ | $ih$ | $pih$ |
|---|---|---|---|---|---|---|---|---|
| 正态分布 | 方差分量 | 16.0000 | 4.0000 | 1.0000 | 64.0000 | 2.0000 | 3.0000 | 144.0000 |
| | 标准误 | 3.0662 | 1.8774 | 0.9970 | 3.3305 | 0.7670 | 0.8321 | 2.7110 |
| 二项分布 | 方差分量 | 0.0108 | 0.0027 | 0.0007 | 0.0448 | 0.0014 | 0.0020 | 0.1870 |
| | 标准误 | 0.0021 | 0.0013 | 0.0007 | 0.0030 | 0.0009 | 0.0007 | 0.0034 |
| 多项分布 | 方差分量 | 0.3244 | 0.1225 | 0.0119 | 0.3929 | 0.0141 | 0.0025 | 0.3170 |
| | 标准误 | 0.0509 | 0.0577 | 0.0536 | 0.0142 | 0.0062 | 0.0057 | 0.0062 |
| 偏态分布 | 方差分量 | 1.6605 | 1.6669 | 1.6491 | 1.6601 | 1.6614 | 1.6559 | 1.6609 |
| | 标准误 | 0.4210 | 0.9189 | 1.9915 | 0.0899 | 0.1985 | 0.4316 | 0.0443 |

表2　$p \times (i:h)$ 设计的方差分量和标准误设定

| | 侧面 | $p$ | $h$ | $ph$ | $i:h$ | $pi:h$ |
|---|---|---|---|---|---|---|
| 正态分布 | 方差分量 | 16.0000 | 1.0000 | 2.0000 | 7.0000 | 208.0000 |
| | 标准误 | 2.7268 | 1.2058 | 1.0367 | 1.4734 | 3.3912 |
| 二项分布 | 方差分量 | 0.0109 | 0.0007 | 0.0013 | 0.0047 | 0.2317 |
| | 标准误 | 0.0019 | 0.0009 | 0.0011 | 0.0011 | 0.0025 |
| 多项分布 | 方差分量 | 0.2050 | 0.1308 | 0.0651 | 0.4099 | 1.1656 |
| | 标准误 | 0.0252 | 0.1502 | 0.0144 | 0.0592 | 0.0174 |
| 偏态分布 | 方差分量 | 1.6628 | 1.6640 | 1.6609 | 1.6524 | 1.6599 |
| | 标准误 | 0.4059 | 1.9610 | 0.1951 | 0.4001 | 0.0401 |

### 2.2　Bootstrap 抽样策略

Bootstrap 法可以对不同的侧面或者侧面的组合进行重抽样。就本文的两种双侧面设计,可能的抽样策略共有 7 种:Boot－p、Boot－i、Boot－h、Boot－pi、Boot－ph、Boot－ih 和 Boot－pih。经过矫正后,每种抽样策略估计的方差分量是一致的,但标准误和置信区间的结果有较大差别。本文参考 Tong 和 Brennan(2007)的结论,选择最优策略进行结果的报告,具体如表3所示。

表3　方差分量变异量的 Bootstrap 最优抽样策略

| $p \times i \times h$ 设计 | | | | | | | |
|---|---|---|---|---|---|---|---|
| 方差分量 | $\sigma^2(p)$ | $\sigma^2(i)$ | $\sigma^2(h)$ | $\sigma^2(pi)$ | $\sigma^2(ph)$ | $\sigma^2(ih)$ | $\sigma^2(pih)$ |
| 最优策略 | Boot－p | Boot－i | Boot－h | Boot－p | Boot－p | Boot－i | Boot－p |
| $p \times (i:h)$ 设计 | | | | | | | |
| 方差分量 | $\sigma^2(p)$ | $\sigma^2(h)$ | $\sigma^2(ph)$ | $\sigma^2(i:h)$ | $\sigma^2(pi:h)$ | | |
| 最优策略 | Boot－p | Boot－h | Boot－p | Boot－i | Boot－p | | |

### 2.3　MCMC 先验分布

使用 MCMC 方法需要对待估计的方差分量 $\sigma^2$ 设置先验分布。依据共轭先验分布的性质,假设 $\tau=1/\sigma^2$ 服从 Gamma 分布(Mao et al.,2005),其中,形状参数固定为2,因为形状参数大于1时分布为正。尺度参数设定为 Traditional 法估计的方差分量值。

### 2.4　比较标准

不同条件下的参数值之间的差异很大,相对偏差百分比(Relative Percentage Bias,RPB,计算公式为$(\hat{\theta}-\theta)/\theta$,其中 $\hat{\theta}$ 和 $\theta$ 分别为参数的平均估计值和真值)用于衡量参数估计的准确性,$|RPB|$ 越接近0,表明估计越准确。由于置信区间的参数值无法得知,采用覆盖率来衡量置信区间估计的准确性,覆盖率越接近0.9表明估计越准确。

## 3　结果与分析

### 3.1　方差分量的估计

各个条件下方差分量的 RPB 如图1所示。绝大多数情况 RPB 都比较小,仅有少数极端情况|RPB|偏大,并且这些情况通常参数值都很小,数值上偏差较小。为了进一步检验各种方法的优劣,对方差分量的$|RPB|$进行4(数据类型:正态、二项、多项、偏态分布)×4(估计方法:Traditional 法、Jackknife 法、Bootstrap 法、MCMC 法)的两因素方差分析。结果显示,数据类型的主效应不显著($F(3,176)=2.528$,$p>0.05$),估计方法的主效应不显著($F(3,176)=2.312$,$p>0.05$),两者间的交互作用亦不显著($F(9,176)=0.539$,$p>0.05$),显示方差

分量的估计在不同的估计方法以及数据类型上都没有差异。

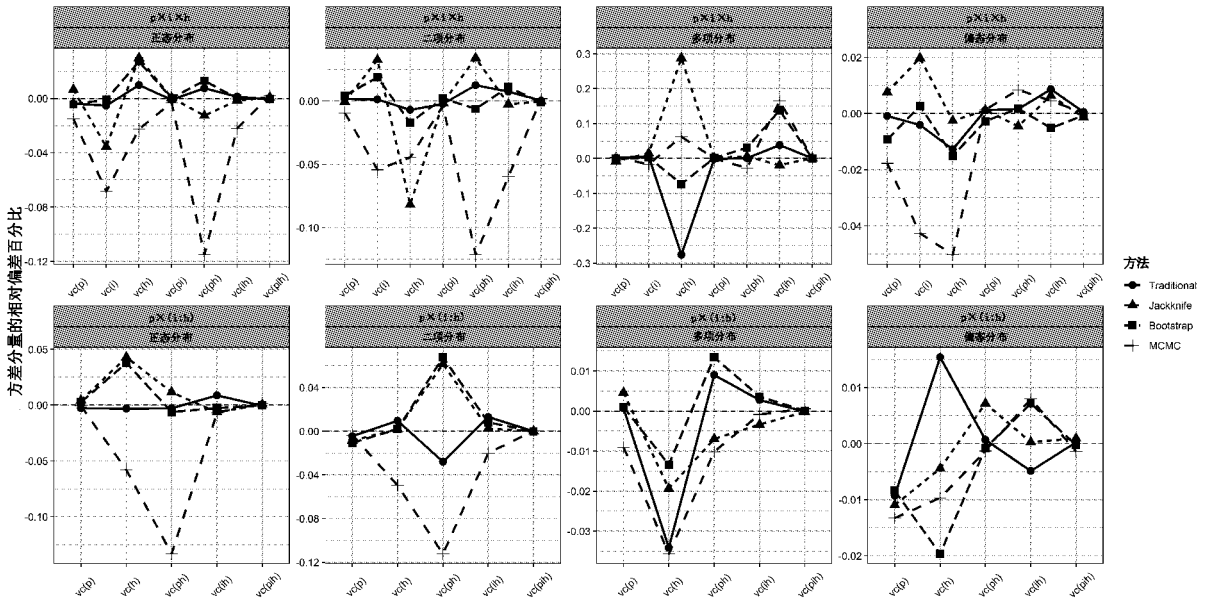


**图 1 方差分量的相对偏差百分比**

### 3.2 方差分量标准误的估计

各个条件下方差分量标准误的 RPB 如图 2 所示。Traditional 法在数据正态分布时表现较好，在二项分布时亦尚可，但在多项分布和偏态分布时|*RPB*|明显增大。而 Bootstrap 法表现稳健，在正态分布和二项分布时表现与 Traditional 法相当，在多项分布和偏态分布时|*RPB*|也相对其他方法较小，显示出跨分布的稳定性。Jackknife 法和 Bootstrap 法的折线趋势有一致性，但是|*RPB*|的大小绝大部分情况都比 Bootstrap 法大。MCMC 法表现稍逊，|RPB|容易发生波动。

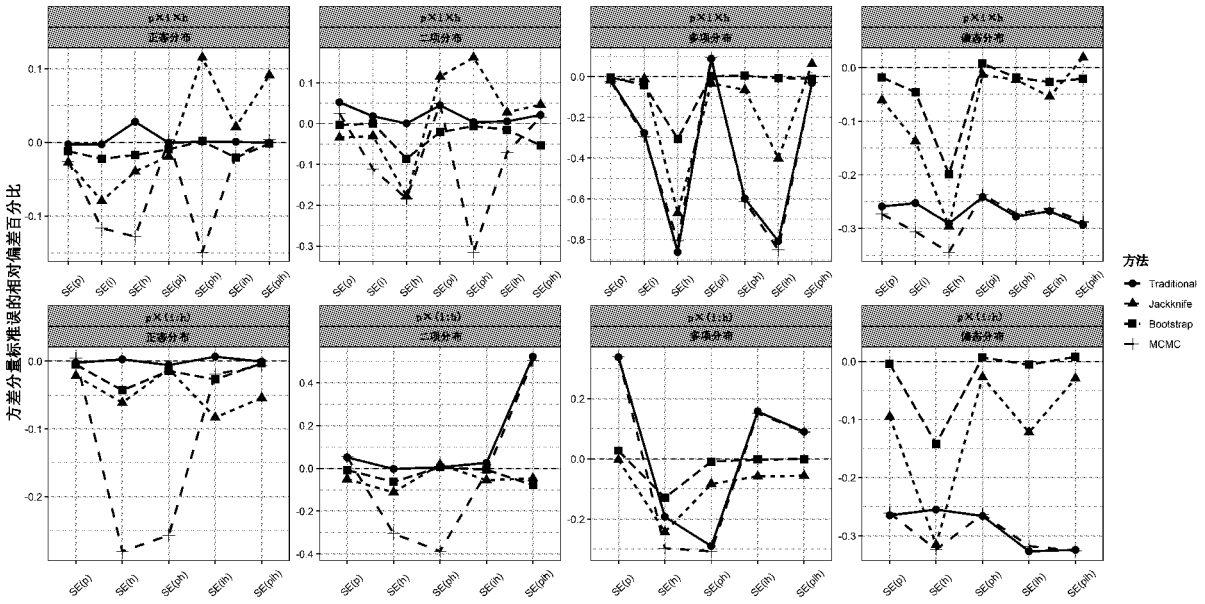


**图 2 方差分量标准误的相对偏差百分比**

为了进一步检验各种方法的优劣，对方差分量标准误的|*RPB*|进行 4（数据类型：正态、二项、多项、偏态分布）×4（估计方法：Traditional 法、Jackknife 法、Bootstrap 法、MCMC 法）的两因素方差分析。结果显示，数据类型的主效应显著（$F(3,176)=15.744, p<0.05, \eta_p^2=0.212$），估计方法的主效应显著（$F(3,176)=16.824, p<0.05, \eta_p^2=0.223$），两者间的交互作用亦显著（$F(9,176)=2.761, p<0.05, \eta_p^2=0.124$）。于是进一步进行简单效应分析，如图 3 所示。数据呈正态分布时，Tra-

ditional 法表现最优,其他分布下 Bootstrap 法最优。

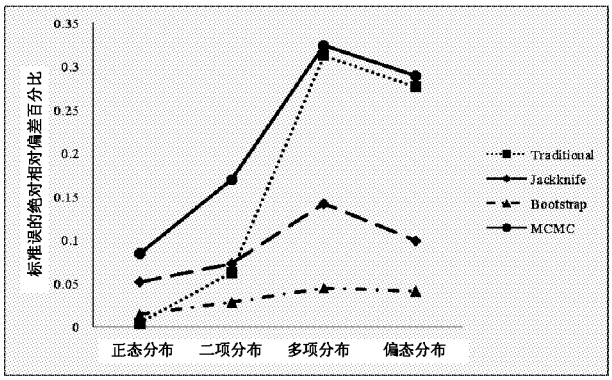


**图 3　不同估计方法和数据类型下方差分量标准误 |*RPB*| 的简单效应分析图**

### 3.3　方差分量的 90% 置信区间估计

相较于方差分量以及标准误的估计,置信区间的估计更加不易,容易出现较大的偏差。各个条件下方差分量 90% 置信区间的覆盖率如图 4 所示。总体来说,正态分布和二项分布的置信区间估计要优于多项分布和偏态分布。Traditional 法在正态分布和二项分布时表现最好,在多项分布和偏态分布时表现变差。Jackknife 法和 Bootstrap 法对分布形态不敏感,而又以 Bootstrap 法略优。但这两种方法容易受到样本量的影响,对 $\sigma^2(h)$ 和 $\sigma^2(ih)$ 的置信区间覆盖率低估。MCMC 法较特殊,相对其他方法更容易高估覆盖率。

为了进一步检验各种方法的优劣,对方差分量 90% 置信区间覆盖率的 |*RPB*| 进行 4(数据类型:正态、二项、多项、偏态分布)×4(估计方法:Traditional 法、Jackknife 法、Bootstrap 法、MCMC 法)的两因素方差分析。结果显示,数据类型的主效应显著($F(3,176)=11.724$, $p<0.05$, $\eta_p^2=0.167$),估计方法的主效应不显著($F(3,176)=0.357$, $p>0.05$),两者间的交互作用不显著($F(9,176)=0.373$, $p>0.05$),显示影响置信区间估计的主要因素是数据类型,四种数据类型的置信区间覆盖率从优到劣依次为:正态分布,二项分布,偏态分布,多项分布。

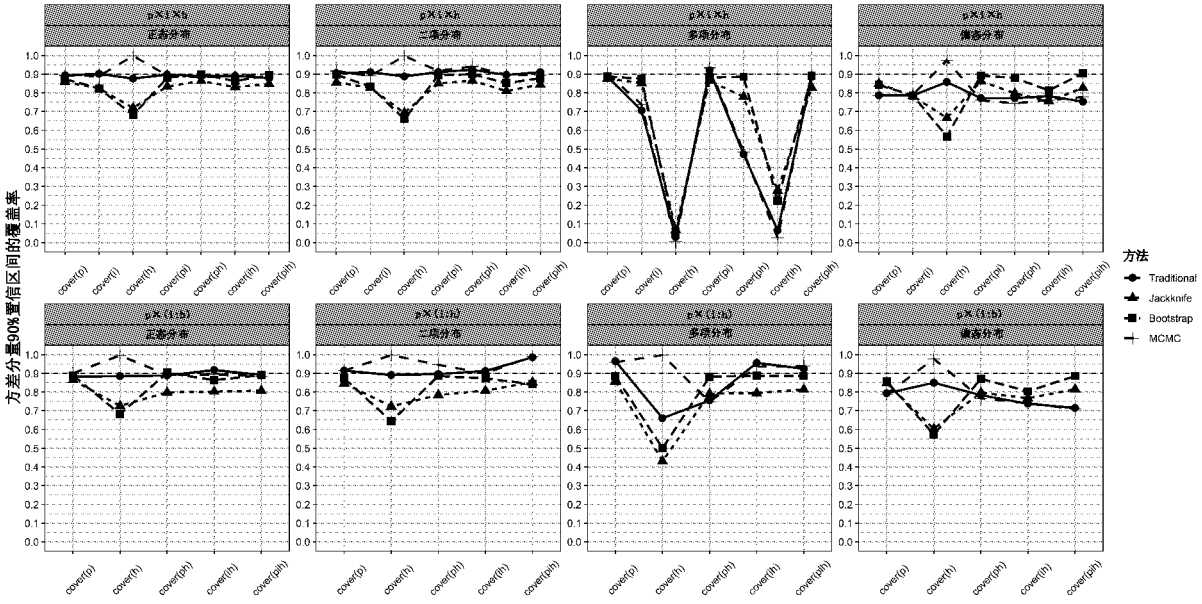


**图 4　方差分量 90% 置信区间的覆盖率**

## 4　讨论

本文采用模拟研究的方法,探讨在概化理论双侧面设计中采用 Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法估计四种类型的数据时,方差分量及其变异量的准确性。结果显示四种方法各有优劣,应该根据实际情况选择合适的估计方法。

### 4.1　方差分量估计

与前人研究一致(Brennan et al.,1987;Mao et al.,2005;Moore,2010;Tong & Brennan,2007),不论数据的分布形态,四种方法都能准确估计方差分量,为方差分量变异量的估计提供基础。特别要指出的是,Bootstrap 法的全部七种抽样策略均能够准确估计所有的方差分量,再一次证明了 Wiley(2001)以及 Tong 和 Brennan(2004)提出的校正公式的有效性。

### 4.2　方差分量的标准误估计

不同方法估计标准误的表现存在差异,当数据为正态分布时,Traditional 法表现最优,Bootstrap 法略次之,在其他三种分布下均以 Bootstrap 法表现最优。

与预期结果一致,Traditional 法在数据为正态分

布时表现最好，在另外三种分布下估计精度下降，这是因为数据呈非正态分布时，分数效应无法满足多元正态分布（Brennan，2001）。二项分布的标准误估计结果与前人研究不完全一致。本研究中，Traditional 法在二项分布时表现尚可，多数情况 $|RPB|$ 较小，而 Brennan 等（1987）和 Othman（1995）的研究均显示，Traditional 法在二项分布上表现不佳。这可能是因为，Brennan 等（1987）的研究二项分布用的不是模拟数据，Othman（1995）的研究中二项分布虽然是模拟数据，但是方差分量经过了矫正，这种做法在后来的研究中鲜有人采用。此外，这些研究都是基于 $p \times i$ 设计的，并没有双侧面的研究。因此，出现不一致的具体原因还有待进一步研究。MCMC 法与 Traditional 法表现类似，容易受到分布形态的影响。Mao 等（2005）的研究也有类似的结论，一个可能的原因是数据非正态分布时，分数效应通常不服从多元正态分布，但在模型中均使用正态分布作为先验分布。

Bootstrap 法在估计 SE（h）时容易出现波动，Jackknife 法也有类似的表现，并且波动更大。Tong 和 Brennan（2007）在研究 Bootstrap 法的时候同样发现估计 SE（h）的偏差更大，并且这种偏差在 $n_h = 2$ 增加到 $n_h = 4$ 的时候会减少。可能侧面水平数太少是这两种重抽样技术产生较大偏差的原因。

### 4.3 方差分量的 90% 置信区间估计

总体上置信区间的估计结果不太理想，但是置信区间的估计与标准误有一致性。正态分布、二项分布下的结果均优于多项分布和偏态分布，并且 Traditional 法和 MCMC 法的估计易受到数据分布形态的影响。Jackknife 法的曲线形状与 Bootstrap 法相似，大部分情况下 Bootstrap 法更优，但 Bootstrap 法易受到侧面水平数的影响，在侧面水平数较小时，例如 $\sigma^2(h)$ 和 $\sigma^2(ih)$ 的置信区间，估计结果会产生较大的低估。Moore（2010）的研究也有类似的结果。这可能是由于百分位置信区间的性质决定的，当侧面水平数较少时，例如，h 侧面仅有 4 个水平，进行 Bootstrap 抽样时只有 $C_7^4$ 种可能的情况，相对于 999 次重抽样来说数量很少。因此，在排序的时候会出现非常多重复的方差分量，这可能是侧面水平数影响 Bootstrap 法置信区间估计的原因，具体需要水平数达到多少，还有待未来进一步探讨。

## 5 结论

基于 $p \times i \times h$ 和 $p \times (i:h)$ 2 种双侧面设计，以及正态、二项、多项和偏态分布 4 种数据类型，采用模拟研究的方法，探讨 Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法在估计概化理论方法分量及其变异量时候的表现，得出结论如下。

（1）估计方差分量时，Traditional 法、Jackknife 法、Bootstrap 法和 MCMC 法在各个条件下均较为准确，在此基础上探讨方差分量变异量的估计是有意义的。

（2）估计方差分量标准误时，Traditional 法在数据呈正态分布时较为准确，二项分布次之，多项分布和偏态分布偏差较大。Bootstrap 法在数据呈正态分布和二项分布时表现与 Traditional 法相当，多项分布和偏态分布时偏差也比其他方法小，显示出跨分布的稳定性。Jackknife 法类似 Bootstrap 法，但偏差相对更大。MCMC 法表现欠佳，波动较大。

（3）估计置信区间时，四种方法互有利弊。Traditional 法易受数据分布形态影响，在数据呈正态分布和二项分布时估计较为准确，多项分布和偏态分布偏差较大。Jackknife 法和 Bootstrap 法类似，在样本量比较小的时候会低估覆盖率，而 MCMC 法则比较容易出现高估的情况。

（4）综合来说，在估计方差分量及其变异量时，若数据呈正态分布，建议选择 Traditional 法；若数据非正态分布，建议选择 Bootstrap 法。

## 参考文献


黎光明.（2010）.*基于概化理论的方差分量变异量估计*（博士学位论文）.华南师范大学.

李向阳.（2015）.高校辅导员半结构化面试的概化理论研究.*内蒙古工业大学学报（社会科学版）*,*24*（1）,33－36.

姚若松，刘泽，赵葆楠，苗群鹰.（2015）.结构化面试中多源变异的概化分析.*心理学探新*,*35*（4）,344－349.

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

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 Series 87－7）. Iowa City, IA: American College Testing Program.

Burdick, R. K., & Graybill, F. A.（1992）. *Confidence intervals on variance components*. New York: Dekker.

Casabianca, J. M., Lockwood, J. R., & Mccaffrey, D. F.（2015）. Trends in classroom observation scores. *Educational & Psychological Measurement*,*75*（2）,311－337.

Clayson, P. E., & Miller, G. A.（2016）. ERP reliability analysis（ERA）toolbox: An open－source toolbox for analyzing the reliability of event－related brain potentials. *International Journal of Psychophysiology*,*111*,68－79.

Feng, W.（2002）. *Applicability of the jackknife procedure for estimating standard errors of variance component estimates in selected random effects G study designs*. Unpublished Doctorial Dissertation, University of Iowa.

Jiang, Z., & Raymond, M.（2018）. The use of multivariate gen-

eralizability theory to evaluate the quality of subscores. *Applied Psychological Measurement*,*42*(8),595 – 612.

Karami,H. (2012). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. *Psychological Test & Assessment Modeling*,*54*(3),211 – 226.

Kellermargulis,M. A. ,Mercer,S. H. ,& Thomas,E. L. (2015). Generalizability theory reliability of written expression curriculum – based measurement in universal screening. *School Psychology Quarterly*,*31*(3),383 – 392.

Lin,C. K. (2017). Working with sparse data in rated language tests:Generalizability theory applications. *Language Testing*,*34*(2),271 – 289.

Luecht,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.

Mantzicopoulos,P. ,French,B. F. ,& Patrick,H. (2018). The mathematical quality of instruction(MQI) in kindergarten:An evaluation of the stability of the MQI using generalizability theory. *Early Education & Development*,*29*(6),893 – 908.

Mao,X. ,Shin,D. ,& Brennan,R. L. (2005). *Estimating the variability of estimated variance components and related statistics using the MCMC procedure:An exploratory study*. Paper presented at the annual meeting of the National Council on Measurement in Education,Montreal.

Medvedev,O. N. ,Krägeloh,C. U. ,Narayanan,A. ,& Siegert,R. J. (2017). Measuring mindfulness:Applying generalizability theory to distinguish between state and trait. *Mindfulness*,*8*(4),1 – 11.

Meyer,J. P. ,Liu,X. ,& Mashburn,A. J. (2014). A practical solution to optimizing the reliability of teaching observation measures under budget constraints. *Educational & Psychological Measurement*,*74*(2),280 – 291.

Moore,J. L. (2010). *Estimating standard errors of estimated variance components in generalizability theory using bootstrap procedures*. Unpublished Doctorial Dissertation,University of Iowa.

Nalbantoglu Yilmaz,F. (2017). Reliability of scores obtained from self – ,peer – ,and teacher – assessments on teaching materials prepared by teacher candidates. *Educational Sciences Theory & Practice*,*17*(2),395 – 409.

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

Searle,S. R. (1971). *Linear models*. New York:Wiley.

Searle,S. R. ,Casella,G. ,& Mcculloch,C. E. (1992). *Variance components*. New York:Wiley.

Ting,N. ,Burdick,R. K. ,Graybill,F. A. ,Jeyaratnam,S. ,& Lu,T. F. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation & Simulation*,*35*(3 – 4),135 – 143.

Tong,Y. ,& Brennan,R. L. (2004). *Bootstrap procedures for estimating standard errors of estimated variance components for two – facet designs*(CASMA Research Report No. 5). Iowa City,IA:Center for Advanced Studies in Measurement and Assessment,University of Iowa.

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

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

# Estimating Variance Components and Their Variabilities in Generalizability Theory:A Cross – distribution Simulation Study

Zhen Fengquan　Zhang Minqiang　Liu Ying

(School of Psychology,South China Normal University,Guangzhou 510631)

**Abstract**:In terms ofeight simulation conditions(data with normal,dichotomous,polytomous and skew distribution with the$p \times i \times h$and$p \times [i:h]$designs),this article examined the applicability of Traditional method,Jackknife method,Bootstrap method and MCMC method for estimating variance components and their variabilities,in order to find out the best method to estimate variance components in various occasions. The result showed that:(1) All these methods could accurately estimate variance components without significant difference. (2) For standard errors of estimated variance components,Traditional method performed well in normally – distributed data,and Bootstrap method behaved best in nonnormally – distributed data. (3) In terms of 90% confidence intervals of estimated variance components,Bootstrap method was robust across different data types,but sensitive to sample sizes. Above all,we recommend traditional method for normally – distributed data,and Bootstrap method for nonnormally – distributed data.

**Key words**:Generalizability Theory;variance component;simulation study;data distribution