2007 年 6 月 数值计算与计算机应用 第 28 卷第 2 期
Jun., 2007 Journal On Numerical Methods and Computer Applications Vol.28, No.2

双重两步迭代估计及其应用[*1)]

丁树良 罗 芬
(江西师范大学计算机信息工程学院, 南昌 330027)

摘 要

两参数 Logisitic 模型 (2PLM) 是一个使用较广的项目反应模型. 用专业软件 BILOG 估计 2PLM 中未知项目参数时, 随着样本容量上升其估计精度改善很小. 为了解决这一问题, 本文改变 BILOG 的估计方案中迭代算法的初值, 即建议使用本文给出的双重两步迭代估计 (DTIE) 作为迭代初值, 所得新的估计方案记为 MMLE/EM(DTIE). 大量 Monte Carlo 模拟结果显示, 在样本容量不小于 2000 时, 使用 MMLE/EM(DTIE) 进行项目参数估计, 其估计精度超过 BILOG.

关键词: 两参数 Logistic 模型, 估计精度, BILOG, 迭代初值

MR (2000) 主题分类: 62F12, 65C05

DOUBLE TWO-STAGE ITERATIVE ESTIMATION AND ITS APPLICATIONS

Ding Shuliang Luo Fen
(*College of Computer information engineering, Jiangxi Normal University, Nanchang 330027, Jiangxi, China*)

Abstract

For solving the unknown parameters in 2-parameter Logistic model (2PLM) under item response theory (IRT), a new estimation scheme, named double two-stage iterative estimation (DTIE) is proposed. DTIE could be employed to provide the iterative initial value implementing M-step in some EM algorithms. It makes the accuracy of estimating the unknown parameters in 2PLM being higher than that of professional software BILOG when sample size is over 2000.

Key words: 2PLM, estimate accuracy, BILOG, initial value

2000 Mathematics Subject Classification: 62F12, 65C05

∗ 2005 年 9 月 19 日收到.

1) 国家自然科学基金项目 (60263005), 江西省科技攻关项目, 卫生部项目 (JM20060070) 及江西师范大学青年基金项目资助.

1. 引 言

项目反应理论 (Item Response Theory, IRT) 是计算机化自适应测验 (Computerized Adaptive Testing, CAT) 的理论基础 [1,2]. IRT 用数学模型 (称之为项目反应模型 Item Response Model, IRM) 来表达考生反应与项目特征之间的关系. IRM 中未知参数估计是应用 IRT 的基础.

目前国内使用最多的是 0-1 评分方式的两参数 Logistic 模型 (2PLM). 该模型如下:

$$P_{ij} = P(U_{ij} = 1|\theta_i) = \{1 + \exp[-Da_j(\theta_i - b_j)]\}^{-1} \tag{1}$$

其中 $U_{ij} = 1$ 表示考生 i 答对第 j 题, 否则 $U_{ij} = 0$, 而考生 i 能力值为 θ_i, 第 j 题的区分度为 a_j, 难度为 b_j, 这里 θ_i, a_j, b_j 便是模型中未知参数, 且 θ_i 是不能直接观察的潜变量 (Latent variable)[3,4]. 下文中, 我们还用 λ_j 表示 $-a_jb_j$.

2PLM 中未知参数估计引起了广泛关注且国际上已有专业软件包 BILOG 处理相应数据 [5], 似乎这一问题已经圆满解决, 然而我们使用 BILOG 时, 通过大量 Monte Carlo 模拟发现, 当考生人数 N 从 1000 上升到 2000, 3000, 4000, 5000 时, 若项目参数估计精度用估计值与真值绝对偏离的平均值来衡量, 估计精度得不到明显改进. 而对于国内许多测验来讲, 考生人数常以万为计算单位, 如果样本容量上升可以提高项目参数估计精度, 为什么不寻找一种较好的估计方法呢? 这个问题促使我们对 BILOG 的计算方法进行研究.

BILOG 使用 EM 算法实现边际似然估计 (MMLE/EM) 以估计项目参数. 悉知, MMLE 是强相合估计 [6]. 设 $x_1, x_2, \cdots, x_n$ 是容量为 n 的样本, 称估计量 $\hat{\beta}(x_1, ..., x_n)$ 是未知参数 β 的强相合估计, 如果 [7]

$$P(\lim_{n\to\infty} \hat{\beta}(x_1, x_2, \cdots, x_n) = \beta) = 1 \tag{2}$$

这里包含一个极限过程, 可见对项目参数估计时 BILOG 程序在某种意义上收敛速度较慢.

在 IRT 中各考生反应相互独立, 同一考生对不同项目反应条件独立的假设下, 项目参数是可以逐个进行估计的. 用 ξ_j=(a$_j$,b$_j$)' 表示第 j 个项目参数向量, 则 ξ_j 的迭代估计式为:

$$\xi_j^{(t+1)} = \xi_j^{(t)} - \Delta\xi_j^{(t)}, t = 0, 1, 2, \cdots, \tag{3}$$

BILOG 中迭代初值 $\xi_j^{(0)}$ 是根据经典测量理论给出 [2,3], 且在程序编制时也注意到为了使迭代收敛, 初值选择应是相合估计 [3].

本文给出一种新的初值估计方法 —— 双重两步迭代估计 (double two-stage iterative estimation, DTIE) 方法, 以改进 (3) 中初值, 再通过 EM 算法来实现项目参数的边际似然估计 (MMLE), 故将该方法记成为 MMLE/EM(DTIE).

本文第 2 节介绍双重两步迭代估计 (DTIE) 原理及算法, 第 3 节介绍 DTIE 在 2PLM 参数估计中的 MonteCarlo 模拟结果及应用, 在第 4 节中引出对 2PLM 中的未知参数的组合估计方式.

2. 双重两步迭代估计及算法

为了对问题有一个整体印象而不陷入枝节, 我们先给出双重两步迭代估计的轮廓以及为什么要采用这种估计的理由. 双重两步迭代估计建立在经验 Logistic 回归基础之上, 它适合于求解 (1) 中未知参数, 且有的未知参数还是潜在变量的情形. 所谓经验 Logistic 回归 [8], 是对观察比例 $p_{\alpha j}$ 用 logit 变换, 得到一个线性模型 (见下文式 7), 通常这个线性模型用迭代重加权最小二乘估计方法 (iteratively reweighted least squares method, IRLSM)[4,9,10] 求解, 但在我们的问题中, 这个线性模型的设计阵也是未知的, 故我们再通过使残差平方和最小的手段求出设计阵中未知参数 (能力参数) 后, 重新进行迭代, 由于这里必须对权重阵和设计阵分别进行修正和迭代, 故称之为双重迭代估计. 我们通过正交试验设计, 做了大量 Monte Carlo 模拟试验, 发现对设计阵进行修改可明显改善项目参数估计的精度.

2.1 双重两步迭代估计算法的理论依据

我们先引入一些记号, 根据得分矩阵 U, 我们记 $Z_i = \sum\limits_{j=1}^{m} u_{ij}$, 即 Z_i 是考生 i 在这次测验所得的总分; 根据总分 Z_i 将 N 个人进行划分, 设有 n_i 个人得 Z_i 分, $i = 1, 2. \cdots s$, $1 \le s \le \min\{N, m+1\}$; 且 $\sum\limits_{i=1}^{s} n_i = N$. 又分别用 r_{ij} 和 $n_i - r_{ij}$ 记总分为 Z_i 的 n_i 个考生答对和答错第 j 题的人数, 令 $p_{ij} = r_{ij}/n_i$, $q_{ij} = 1 - p_{ij}$, 若 P_{ij} 由 (1) 定义; $Q_{ij} = 1 - P_{ij}$, 则有如下引理

引理 1[8]**.** $Ep_{ij} = P_{ij}$, $Eq_{ij} = 1 - P_{ij} = Q_{ij}$, P_{ij} 是总分为 Z_i 的被试答对第 j 题的概率; 若 $p_{ij} \ne 0$ 且 $p_{ij} \ne 1$, 记 $y_{ij} = \ln(p_{ij}/q_{ij}) = \ln(r_{ij}/(n_i - r_{ij}))$, 则当 $n_i \to \infty$ 时, y_{ij} 的渐近分布为 $N(\ln(P_{ij}/Q_{ij}), 1/(n_i P_{ij} Q_{ij}))$; 如果 r_{ij}=0 或 $r_{ij} = n_i$ 时, $E\ln((p_{ij}+t)/(q_{ij}+t))$ 与 $\ln(P_{ij}/Q_{ij})$ 最接近的 t 是 0.5.

记

$$y_j = (y_{1j}, \cdots, y_{sj})^T, \quad \sigma_{ij}^2 = Var(y_{ij}) = 1/(n_i P_{ij} Q_{ij}),$$

$$\sum\nolimits_j = diag(\sigma_{1j}^2, \cdots, \sigma_{sj}^2) \quad X_j^T = \begin{pmatrix} 1 & \cdots & 1 \\ \theta_1 & \cdots & \theta_s \end{pmatrix} \tag{4}$$

注意这里 X_j(通常称为设计阵) 只与 $\theta_1, \cdots, \theta_s$ 有关而与 j 无关, 由于 a_j, λ_j 及 θ_i 未知, 通常用 $(n_i p_{ij} q_{ij})^{-1}$ 作为 σ_{ij}^2 的估计量, 如果 $p_{ij} = 0$ 或 $p_{ij} = 1$, 则可以用下式对 σ_{ij}^2 的估计量进行修正 [8],

$$[(n_i+1)(n_i+2)]/[n_i(r_{ij}+1)(n_i - r_{ij}+1)] \tag{5}$$

而 θ_i 可以用下式估计 (仍记为 θ_i)

$$\theta_i = \begin{cases} \ln[Z_i/(m - Z_i)], Z_i \ne 0 \text{ 且 } Z_i \ne m \\ \ln[(Z_i + 0.5)/(m - Z_i + 0.5)], \text{ 否则} \end{cases} \tag{6}$$

根据经验 Logistic 回归, 用上述 (4) 中记号可以写成一个线性模型

$$y_j = X_j \beta_j + \varepsilon_j, \quad Var(\varepsilon_j) = \sum\nolimits_j, \tag{7}$$

若 X_j 与 $\sum_j$ 为已知时, β_j 的最佳线性无偏估计 (BLUE) 为:

$$\hat{\beta}_j=(X_j'\sum\nolimits_j^{-1}X_j)^{-1}X_j'\sum\nolimits_j^{-1}y_j \tag{8}$$

实际上 (7) 是能力值为已知的前提下, Berkson 给出的最小变换 χ^2 方法 (例如, 可参见 [3]) 导出估计 λ_j 与 a_j 的式子. 可见最小变换 χ^2 方法就是统计中常用的经验 Logistic 回归方法.

在设计阵 X_j 为已知时, 即能力为已知时, 统计中对经验 Logistic 回归的通常处理办法是迭代重加权最小二乘方法 [4,9,10], 然而实际上能力未知, 故设计阵 X_j 也是估计出来的, 我们认为对 X_j 也应作相应的修正.

2.2 双重两步迭代估计算法

本小节中省去表示项目 j 的足码, 而 $\hat{\beta}_t$, X_t, Σ_t, P_{kt}, Q_{kt} 分别表示 β, X, Σ 第 t 次迭代值, 特别地, P_{kt} 是将 θ_k, a,b 的第 t 次迭代值代入 2PLM 而得到的, Q_{kt}=1-P_{kt}.

求 β 的估计的双重两步迭代估计算法为

$$\hat{\beta}_t=\left(X_t'\sum\nolimits_t^{-1}X_t\right)^{-1}X_t'\sum\nolimits_t^{-1}y_t \quad t=0,1,2,\cdots.$$

(i) $\hat{\beta}_0$ 中 X_0 中的 θ 值由 (6) 式给出, 对角阵 Σ_0 的对角元由 (5) 式给出.

(ii) 对 $t=0,1,2,\cdots$, $\hat{\beta}_{t+1}$ 中 X_{t+1}, Σ_{t+1} 如下定义:

(a) X_{t+1} 中 θ_k 值由极小化 $(y-x\hat{\beta}_t)'\Sigma_t^{-1}(y-x\hat{\beta}_t)$ 导出.

(b) 对角阵 Σ_{t+1} 中对角元由 $1/n_kP_{kt}Q_{kt}$ 给出 $(k=1,2,\cdots,s)$.

迭代直到 $\left\|\hat{\beta}_{t+1}-\hat{\beta}_t\right\|<\varepsilon$ 为止.

最后, 将估出的项目参数看成已知, 求能力参数条件似然估计 (如可参见 [2][3]).

2.3 双重两步迭代中为什么要修正设计阵

考察如下二水平的四个因素 A, B, C, D 及相应的交互效应:

因素 A 是设计阵 X 变化与不变 (水平 1: X 变, 水平 2: X 不变)

因素 B 是方差阵 $\sum$ 变化与不变 (水平 1: $\sum$ 变, 水平 2: $\sum$ 不变)

因素 C 是被试人数的变化 (水平 1: 2000 人, 水平 2: 3000 人)

因素 D 是项目数的变化 (水平 1: 20 题, 水平 2: 30 题)

讨论这些因素对 β_j 的估计的影响, 我们在计算机上进行正交试验, 采用 $L_{16}(2^{15})$ 安排试验.

试验重复进行 3 次, 并且我们分别观察了相关系数和 $RMSD$ 两类指标, 包括区分度、难度及能力值的估计值与相应真值的相关系数, 区分度和难度及能力值的估计值对应的 $RMSD$, (参见下文 (10) 式). 计算结果进行方差分析 ($ANOVA$) 后, 可得到如下结论:

(i) 交互效应均不显著且迭代比不迭代效果好;

(ii) 概括地说, a) X 的变化可显著地使 β_j 的估计值的 $RMSD$ 下降 (显著性水平 $\alpha<0.01$); b) $\sum$ 的变化对提高 a 的估计值与真值的相关系数有显著影响 ($\alpha=0.005$); 人数增多使 a 的相关系数有显著提高 ($\alpha=0.05$);

c) 题数越多, 对提高能力相关系数, 降低 $RMSD$ 越有显著效果;

d) 在显著性水平 $\alpha = 0.1$ 上, 人数越多, b 的估计值与真值的相关越高.

我们做了大量的计算机模拟试验, 这些结果支持上述计算机模拟所做的正交试验的结论.

3. 双重两步迭代估计的应用 —— 作为迭代初值

BILOG 求项目参数的方法为 MMLE/EM, 对能力的估计采用后验期望估计 (EAP) 或贝叶斯众数估计 (Bayesian modal estimation, BME), 并且如果被试样本容量超过 1000, 且估计时对样本容量采用默认值, 则 BILOG 样本容量的默认值为 1000[5].

为估计 2PLM 中未知参数, 我们将 DTIE 作为两种不同估计方法 (MMLE/EM 和 SQRT/EM) 的初值, 取得较好的效果. 在 Monte Carlo 模拟中, 我们采用以下两个指标, 以评价参数估计程序对参数真值的修复程度 (recovery), 一个为 ABS, 另一个是 RMSD:

$$ABS(x) = \sum_{r=1}^{R}\sum_{j=1}^{J} |\hat{x}_j^{(r)} - x_j|/(JR) \tag{9}$$

$$RMSD(x) = \sqrt{\sum_{r=1}^{R}\sum_{j=1}^{J} (\hat{x}_j^{(r)} - x_j)^2/(JR)} \tag{10}$$

(9), (10) 式中 x_j 表示参数向量 X 第 j 个分量的真值 (即 Monte Carlo 模拟值), 而 $\hat{x}_j^{(r)}$ 表示在第 r 次重复中对 x_j 的估计值, 同时用这两个指标的原因是 ABS 对越轨值 (outlier) 较稳健 (robust), 而在一定条件下 $RMSD$ 可以给出其分布, 由定义式 (9)(10), 我们知道, $ABS(RMSD)$ 越小, 则估计精度越高. 以下我们都用 $ABS(RMSD)$ 来描述估计精度.

我们通过 Monte Carlo 模拟来考查新算法的估计效果. 试验条件是项目数 m 保持不变 ($m = 60$), 对于每一种试验条件. 我们重复 $R = 30$ 次试验, 然后分别代入 (9) 和 (10) 式计算相应参数的 ABS, $RMSD$. 在 2PLM 的项目参数估计中我们采用 MMLE/EM(DTIE), 估计结果见表 1.

表 1 三种程序的估计结果

		程序	1000	2000	3000	4000	5000
ABS	a	BILOG	0.0906	0.0898	0.0907	0.0894	0.0884
		DTIE	0.1350	0.1119	0.1018	0.0936	0.0886
		MMLE/EM(DTIE)	0.1071	0.0792	0.0649	0.0559	0.0519
	b	BILOG	0.1438	0.1420	0.1449	0.1485	0.1399
		DTIE	0.1430	0.1174	0.1116	0.1056	0.1029
		MMLE/EM(DTIE)	0.1421	0.1052	0.0930	0.0816	0.0746
$RMSD$	a	BILOG	0.1270	0.1251	0.1243	0.1248	0.1215
		DTIE	0.2244	0.1717	0.1385	0.1189	0.1090
		MMLE/EM(DTIE)	0.2270	0.1919	0.1314	0.0968	0.0822
	b	BILOG	0.2588	0.2530	0.2359	0.2580	0.2257
		DTIE	0.2198	0.1870	0.1838	0.1759	0.1744
		MMLE/EM(DTIE)	0.2217	0.1662	0.1477	0.1349	0.1234

从表 1 可以看出, 使用 BILOG 做参数估计, 估计精度基本上不随被试样本容量的变大而提高, 但 DTIE 及 MMLE/EM(DTIE) 的项目参数估计精度却随样本容量的变化而大大提高, 以 ABS(a) 及 ABS(b) 为例, 如图 1, 2 所示, 图中横坐标为人数, 纵坐标为相应的 (ABS) 值.

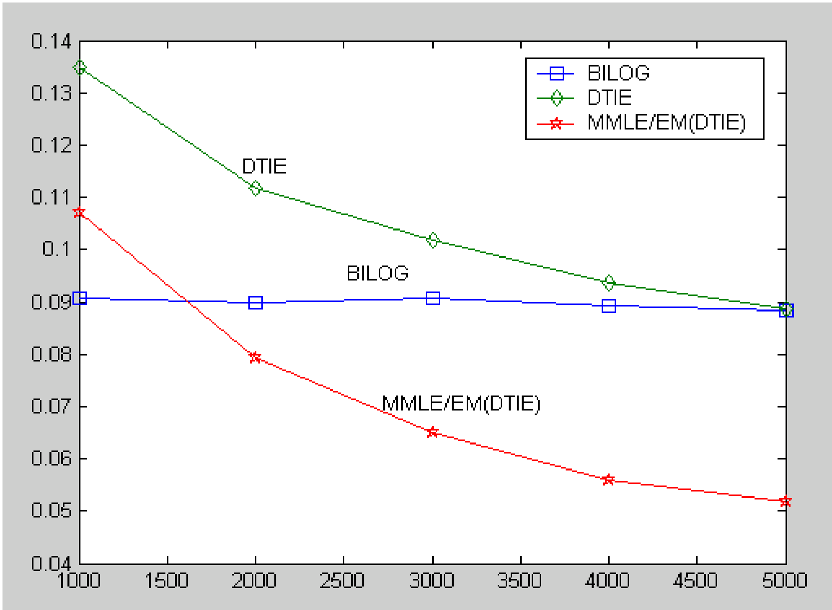

图 1　用三种估计方法估计区分度参数时的 ABS(a) 图

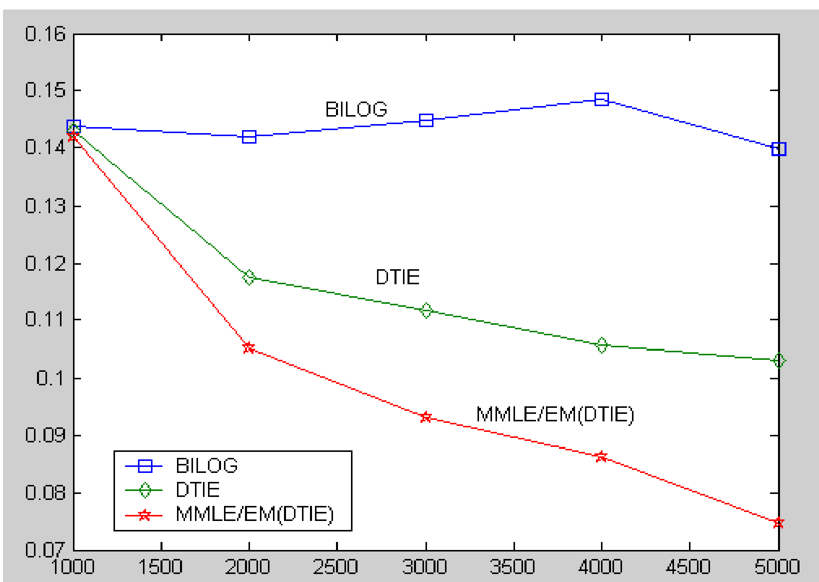

图 2　用三种估计方法估计难度参数时的 ABS(b) 图

当样本容量 $N \geq 500$ 的时候, DTIE 作为另一种估计 2PLM 中未知参数的方法 (SQRT/EM)[11] 的迭代初值, 效果也很好, 所得项目参数估计值的 ABS 也比相同条件下 BILOG 给出的估计值的 ABS 要小. 所谓 SQRT/EM 估计方法, 是用 EM 算法, 求取项目参数估计值 $\hat{\zeta}_j$, 记 $P_{\alpha j}(\hat{\zeta}_j)$ 为 $\hat{P}_{\alpha j}$, 使

$$\sum_{\alpha=1}^{N}\sum_{j=1}^{m}(\sqrt{p_{\alpha j}}-\sqrt{P_{\alpha j}})^2 \geq \sum_{\alpha=1}^{N}\sum_{j=1}^{m}(\sqrt{p_{\alpha j}}-\sqrt{\hat{P}_{\alpha j}})^2$$

式中 $p_{\alpha j}$ 和 $P_{\alpha j}$ 分别为 2PLM 的观察值与理论值 [11].

4. 结论与讨论

MMLE/EM(DTIE) 在样本容量大时 (比如 $N \geq 2000$) 比 BILOG 的估计值要精确一些, 如人数较多时, 使用 MMLE/EM(DTIE) 较好. 但对于 $N \leq 1000$ 时, 及项目数较小时, 其估计值不如 BILOG, 故目前对人数及题数较少时, 仍应使用 BILOG 进行估计. 这种估计的方法, 我们称之为组合估计的方法, 即针对所给条件的变化, 选用给定条件下的较好的估计方法, 以期得到较好的估计结果.

对于 SQRT/EM 方法, 如 [11] 中所报告的在 $N \geq 500$ 时, 使用 DTIE 作初值, 误差较小; 但 $N < 500$ 时, 则宜采用由经典测量理论给出的初值.

在一般的统计抽样中, 样本容量 1000 完全可以看成是一个大样本, 然而我们的模拟试验中, 新的参数估计方法都只能在样本容量不小于 2000 时表现出其优越性来, 这是一个很奇怪的现象. 然而这种奇怪的现象的产生却有两个深刻的统计背景.

第一, 本文 (2) 式所给出的模型 (2PLM) 中含有两类未知参数 [3], 一类是项目参数 a_j, b_j(或 a_j 及 $\lambda_j = -a_j b_j$). 这类参数的个数不随考生人数 (样本容量) 的变化而变化, 统计上称之为伴随参数 (incidental parameter), 又称为讨厌参数 (nuisance parameter). 一般统计书上很少讨论伴随参数. 由于伴随参数的出现, 使得估计变得比较困难. [3] 中所引的结论是用联合极大似然估计 (JMLE) 估计 2PLM 中未知参数, 只有当项目数大于 30, 考生人数大于 500 时才能估得较准, 而对 3PLM 则需项目数大于 60, 考生人数大于 1000. 所以对 IRT 模型中未知参数作估计时样本容量通常都比较大. 对于样本容量仅有 50, 60(甚至 100, 200), 我们要得到较好的估计, 只好寻求软计算方法 (如神经网络, 遗传算法等). 这种现象的另一个解释是通常统计抽样是从同一个母体抽取, 但对于 2PLM 来讲, 有 N 个考生, m 个项目, 则有 $N \times m$ 个总体, 每个总体分布列为 $\{P_{\alpha j}, Q_{\alpha j}\}$, $\alpha = 1, 2, \cdots, N$; $j = 1, 2, \cdots, m$, 于是只有考生人数 N 相当大时, 某些 θ_α 才会相差不大, 所以对于 IRT 模型, 欲估计模型中未知参数, 样本容量都要相当大.

第二, 我们给出的估计初值的方法 DTIE, 是建立在引理 1 的基础之上, 只有在引理 1 的条件得到满足时才能指望 DTIE 的行为表现较好, 否则, 我们建立的模型 (8) 就有相当大的误差.

而引理 1 强调在对考生按能力 (考分) 进行分组时, 每组中考生人数 n_i 应趋于无穷 ($n_i \to \infty$), 如果考生人数太少, 则相近能力 (考分) 的考生太少满足不了 $\mathrm{n}_i \to \infty$ 的要求; 如进行组

合并, 则可能同一组的考生能力 (考分) 相差甚远, 这又违背了分组应尽量使同一组考生得分 (能力) 相近的原则.

由于 IRT 模型中未知参数的特殊性, 故通常用统计方法进行参数估计时, 样本容量都很大, 而由引理 1 的要求, DTIE 算法中样本容量还要进一步扩大. 幸好我国的许多考试考生都很多. 故我们介绍的方法还可以用于实际.

参 考 文 献

[1] Embretson S, Reise S, Item Response Theory for psychologists, New jersey, Lawrence Erlbaum Associates, Publishers, 2000.

[2] 漆书青, 戴海琦, 丁树良. 现代教育与心理测量学原理 [M], 北京: 高等教育出版社, 2002.

[3] Frank B. Baker S. Item Response Theory: Parameter Estimation Techniques[M]. New York: Marcel Dekker, Inc., 1992.

[4] Bartholomew D J, Knott M. Latent variable models and factor analysis (2^{nd}, ed.), London: Arnold, 1999.

[5] Mislevy R J, Bock R D. BILOG-3; Item analysis and test scoring with binary logistic models[CP]. Mooresville, IN: Scientific Software, 1990.

[6] Kiefer J, Wolfowitg J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters[J]. Annals of Mathematical Statistics, 1956, 27: 887-903.

[7] 陈希孺. 概率论与数理统计 [M]. 北京: 科学出版社, 2000.

[8] 张尧庭. 定性资料的统计与分析 [M]. 桂林: 广西师范大学出版社, 1991.

[9] Stone C J. A Course in Probability and Statistics[M]. 北京: 机械工业出版社, 2003.

[10] Alan Agresti. Categorical Data Analysis[M]. New York: John Wiley & Sons, 1990.

[11] 黄华彩, 丁树良, 罗芬. IRT 模型下的 SQRT/EM 参数估计方法及应用研究 [J]. 江西师范大学学报, 2005, 29(3): 231-234.