应用概率统计　第十八卷
第四期　2002 年 11 月

Chinese Journal of Applied Probability
and Statistics Vol. 18 No. 4 Nov. 2002

# 固定样组纵向调查“间歇式”期单元无回答的加权调整 *

杨宝慧　　孙山泽
(北京大学数学科学学院，北京，100871)

**摘　要**

期单元无回答误差是固定样组纵向调查中经常出现的一类非抽样误差．如果不对之进行调整，则往往造成估计量的偏差．已经提出的两种加权调整方法不易处理“间歇式”期单元无回答．在本文中，我们提出了纵横加权调整方法，这一方法克服了已有方法的不足．我们所作的模拟研究表明，纵横加权方法降低了估计量的偏差，并在作两调查期指标均值变化分析时，充分利用了两期回答状态的相关信息，提高了变化估计量的准确度．

**关 键 词**：固定样组纵向调查，期单元无回答，回答机制，加权，变化估计量．
**学 科 分 类 号**：O212.2.

## §1. 引　　言

固定样组纵向调查在调查期内对初始样本进行跟踪调查，使每一初始样本单元有一个纵向的记录，并对之进行纵向分析，如两调查期之间的变化的分析．例如，Lepkowski(1989)[1] 中提到的“Income Survey Development Program 1979 Research Panel(ISDP)”就是一个这样的调查．此调查中对约 7500 个住户中 16 岁及 16 岁以上成员每隔 3 个月进行一次调查，共进行了 6 期调查，以得到收入及参加政府提供的培训项目的情况，进而对政府项目进行评估．调查中不能从所有的样本单元及问卷中的所有问题获得有用的数据，称为“无回答”．对于一次性调查的无回答已有许多研究，文献中上百篇的文章讨论了预防、测定或消除无回答产生的影响．从范围上分，一次性调查的无回答分为单元无回答和项目无回答．单元无回答指样本单元没有接受调查，对样本单元没有收集到任何有价值的信息．而项目无回答是指样本单元接受了调查，但对调查中的部分项目没有回答．由于连续调查的复杂性，使得连续调查的无回答形式多样化．连续调查的无回答可分为如下几类：

单元无回答：合格的样本单元在调查时间内的各期调查均未提供任何数据信息，也可以说未参与调查，则称为单元无回答．

期单元无回答：同一样本单元在调查时间内合格期数为两期以上 (含两期) 时，样本单元在某些合格期未参与调查，而至少在一个合格期参与调查，造成部分合格期没有提供任何有价值的信息，称为期单元无回答．

项目无回答：样本单元在参与调查各期均未回答某些调查项目的信息，则称为项目无回答．

期项目无回答：样本单元在参与调查部分调查期未回答某一调查项目的信息，但至少在一个调查期回答了这一调查项目，则称为期项目无回答．

期单元无回答，可分为“间歇式”和“掉队式”期单元无回答．所谓“掉队式”期单元无回答是指合格样本单元一旦在某一期无回答，则在以后各期都无回答．否则，称为“间歇式”期单元无回答．

概括地说，数据缺失可能有两方面的影响．一方面，如果数据缺失是完全随机缺失 (MCAR)，则无回答数据与回答数据没有系统差别．直接利用回答数据分析、推断总体参数会产生有效的估计 (无偏或近似无偏估计)，只是各种类型的数据缺失 (单元缺失、期单元缺失、项目缺失、期项目缺失) 都使得分析时所用的有效样本量减少，这将降低估计量的精度．这一影响也许并不严重，因为可以通过在设计阶段设计一个更大的样本来加以预防．另一方面，当无回答造成的缺失数据与有回答数据有系统的差别时，不加调整地分析不完全

* 本论文受国家自然科学基金项目资助 (项目编号 10071091)．
本文 2000 年 7 月 3 日收到，2001 年 7 月 6 日收到修改稿．

数据往往会造成估计量的偏差．在样本量较大的大规模调查中这种偏差的平方是估计量的均方误差的主要成分，对估计量准确度影响极大 [2]. 因此，当出现无回答时，特别是非完全随机的数据缺失时，统计分析者都面临的一个任务是如何对无回答造成的不完全数据进行调整，以消除或减小估计量的偏差，最终提高估计量的准确度．本文首先介绍 Lepkowski(1989)[1] 和 Little 和 David(1983)[3] 对连续调查中期单元无回答的加权调整方法．然后提出我们建议的新方法．

## §2. 固定样组调查期单元无回答的已有的加权调整方法及存在的问题

### (一) Lepkowski(1989) 的方法[1]

Lepkowski(1989) 考查了期单元无回答的模式．以三期调查为例，用 X 代表回答， O 代表无回答．则回答模式可能为： XXX, XXO, XOO, OXX, OOX, OXO, XOX, OOO. 其中， OOO 代表单元无回答， XXX 代表单元各期都回答．Lepkowski 列举了 ISDP(1979)— 一个 6 期的连续调查和 SIPP(1984)— 一个 9 期的调查的前三期的不同回答模式所占的样本百分比．SIPP 对首期的不回答者不再进行调查，即一旦某样本单元在某一期不回答，则在后续调查中不再对其调查．因此，在 SIPP 中不出现 OXX, OOX, OXO 这些回答模式．为了对比，将 OXX, OOX, OXO 这些回答模式从 ISDP 的数据去除后，计算出 XXX, XXO, XOO, XOX 回答模式样本单元占第一期回答样本单元的百分比．见下表：

| 回答模式 [a] | | ISDP | | SIPP |
|---|---|---|---|---|
| | | 所有有回答样本 | 首期回答样本 | |
| 完全回答 | XXX | 80.2 | 83.3 | 90.0 |
| 掉队式回答 | XXO | 7.2 | 7.5 | 4.9 |
| | XOO | 6.7 | 7.0 | 4.2 |
| 间歇式回答 | XOX | 2.3 | 2.4 | 1.0 |
| | OXX | 2.2 | — | —[b] |
| | OXO | 0.6 | — | — |
| | OOX | 0.9 | — | — |

a: 回答 =X, 不回答 =O.

b: SIPP 对首期的不回答者不再进行调查．

表中 “所有有回答样本” 对应列是各种回答模式在 ISDP 中的所有有回答样本单元 (即在所有抽中样本中去除 OOO 型单元无回答样本) 中所占的百分比， “首期回答样本” 对应列是 ISDP 中 XXX, XXO, XOO, XOX 回答模式样本单元占第一期回答样本单元 (即在所有有回答样本中去除第一期无回答样本) 的百分比．完全回答者的比例最大， “掉队式” 无回答次之， “间歇式” 比例最小．随着调查期的增多，完全回答者的比例减少，回答模式增多，间歇式期单元无回答的比例可能增大．9 期的 SIPP 调查可能的回答模式有 $2^9 - 1 = 511$ 个．也许完全回答的比例最大，之后是有两期无回答的回答者，其它类型的期无回答者的比例可能很小．虽然单个期单元无回答类型所占的比例可能是可以忽略的，但总括起来却不可忽略．

Lepkowski 考虑了这些类型和所占的比例，并根据期单元无回答的具体情况提出了一些实用化的加权组调整方法．加权组的方法的关键是构造均值与回答概率同质度高的加权组，选择划分加权组的变量一般要与要分析的指标密切相关．Rubin(1976) 指出，如果加权组内回答概率同质度高，加权组间回答概率差异比较大，则加权组调整会降低估计量偏差 [4].

对于单元无回答作加权组调整时，用于划分加权组的变量可能只是设计阶段已知的一些信息，这些指标可能与要分析的指标联系不是很强．但对于在某些期作了回答的期无回答进行调整时，可能有更有效的划分加权组的指标．这些强联系的指标的使用可能提高加权组调整的准确度．另一个在作期无回答加权调整时所要考虑的问题是所作的分析．例如前面的三期调查的例子，可能要三期的横向参数的估计量：期均值、期总值，

还可能希望估计两期之间均值的变化等．当估计横向参数，如第一期的指标均值时，XXX, XXO, XOX, XOO 提供了第一期的数据，可对它们进行加权组调整以弥补第一期无回答的数据 OOO, OOX, OXX 和 OXO, 即利用多期调查数据中与要分析的指标及回答概率强联系的变量对样本划分加权组，在加权组内，将 OOO, OOX, OXX 和 OXO 样本的权分配给组内 XXX, XXO, XOX, XOO 样本．对其它期横向参数进行估计时，可对相应期无回答样本的权数作类似的调整．这种调整方法暗含假设组内期回答概率相等．当估计两期均值变化，如第一期到第二期均值的变化时，为了更多地利用数据，可用 XXX 和 XXO 型回答数据，Lepkowski(1989) 的方法是在加权组内作加权调整以弥补其它五种期无回答数据．这样，若要满足三期调查的所有横向估计与纵向比较，则需要 7 组权数．对于一个 9 期调查则需要 511 组权．在上述分析中暗含的假设是组内 XXX, XXO 与其它期无回答类型的回答概率相同．这种方法对不同的分析目的暗含了不同的假设，调整的权数也随之改变．随着调查期的增加，满足多种分析目的所需的权数组急剧增加，这在数据管理上是很大的困难．减小权数组数的一个方法是少用一些比例较小的回答类型的数据．例如，掉队式期无回答数据每一期只需要一组权．如果间歇式期单元无回答数据中间歇式无回答的比例较小，可以在分析时将这一小部分数据弃之不用，只用掉队式回答数据，每一期数据只需一组权．当间歇式回答数据比例比较大时，这样会浪费许多数据，可弃掉某些单元某些期的数据使之转成掉队式数据．总之，可对 Lepkowski 的加权调整方法概括如下：

(1) 利用各期都回答的数据，其它数据都弃之不用，然后将不利用的样本的权数分配给这些每一期都回答的样本单元．这样可能因为弃掉太多的数据而浪费大量信息．

(2) 只利用期期都回答的样本单元及掉队式回答单元的数据，将其它样本单元的权数在加权组内调整给利用的单元．在有较多的间歇式无回答时会浪费信息．有时可将一些间歇回答单元数据转成掉队式加以利用．

(3) 间歇式期单元无回答的加权调整：尽可能多地利用数据，根据不同分析目的构造不同的权数组以弥补未利用的样本．这种方法的缺点是权数组可能太多，且对不同的分析目的暗含了不同的假设，似乎不太合理．

### (二) Little 和 David(1983) 的方法[3]

Little 和 David(1983) 给出了一种对掉队型数据进行加权补救的方法．思想是将单元 $i$ 的第 1 期的回答标识变量 $r_{i1}$ 对 $z$ 进行线性回归，Probit 回归或 Logistic 回归，计算出第一期的回答概率 $p_{i1}$, 单元 $i$ 的第 1 期数据调整权数 $w_{1.0}$ 为回答概率 $p_{i1}$ 的倒数．对于第 1 期的回答者，第 2 期的回答标识变量 $r_{i2}$ 对辅助变量 $z$ 和第 1 期回答指标值 $y_{i1}$ 回归，从而得到从第 1 期到第 2 期的回答概率 $p_{i2.1}$, 从第 1 期到第 2 期的调整权数 $w_{2.1}$ 为 $p_{i2.1}$ 的倒数，第 2 期的最终的调整权数为 $w_{1.0}\cdot w_{2.1}$. 单元 $i$ 的第 $t$ 期的回答标识变量 $r_{it}$ 对辅助变量 $z$ 和前 $t-1$ 期的指标 $y_{i1}, y_{i2}, \cdots, y_{it-1}$ 回归，以估计从第 $t-1$ 期到第 $t$ 期的调整权数 $w_{t,t-1}$, 则第 $t$ 期的调整权数为

$$w_t = \prod_{i=1}^{t} w_{i,i-1}. \tag{2.1}$$

当用这种方法处理“间歇式”无回答数据时，需对各种回答模式分别建立回归方程，所需建立的回归方程数量随调查期的增多呈几何级数增长，这一方法变得复杂．详细论述请参看文献．Little 和 David(1983) 没有说明对于两期均值的比较分析时如何进行特别的加权调整．

## §3. 调整“间歇式”期单元无回答的一种新方法 - 纵横加权方法

从以上的介绍可以看到，Lepkowski 的方法以及 Little 和 David 的方法在处理间歇式期单元无回答时要么权数组太多，要么需要建立的回归方程太多，不利于数据管理和数据分析．为此，我们提出了一种新的加权调整方法 - 纵横加权方法，并通过计算机模拟实验，与 Lepkowski 的方法进行了对比．

### (一) 纵横加权调整方法介绍

Little and Su(1989) 提出了一种对期项目无回答调整的纵横复制方法 [5]. 受之启发，我们提出一种处理一般期单元无回答的纵横加权调整方法.

假设一个样本量为 $n$ 的 $T$ 期的连续调查. 根据已知辅助信息将总体分为 $H$ 类，样本也分为相应的 $H$ 类. 假设 $h$ 类内的单元 $i$ 在第 $t$ 期调查时回答的概率 $p_{iht}$ 服从下列模型:

$$p_{iht} = r_t \cdot c_h, \tag{3.1}$$

其中 $c_h$ 为类效应， $r_t$ 为调查期效应.

采用以下步骤对 $c_h$ 、 $r_t$ 进行估计:

(1) 计算第 $h$ 类第 $t$ 期的样本回答率 $p_{ht}$; 总样本第 $t$ 期回答率 $p_{tc}$, 具体算法依据抽样设计而定.

(2) 计算 $H$ 类 $T$ 期累计平均回答率 $p$.

(3) 估计调查期效应:

$$\widehat{r}_t = \min(p_{tc}/p,\ 1). \tag{3.2}$$

(4) 去除调查期效应，估计类效应. 令 $p'_{ht} = p_{ht}/r_t$, 类 $h$ 的效应采用估计量

$$\widehat{c}_h = \min\Big(\frac{1}{T}\sum_{t=1}^{T} p'_{ht},\ 1\Big). \tag{3.3}$$

(5) 估计 $h$ 类内的单元 $i$ 在第 $t$ 期调查时回答的概率 $p_{iht}$,

$$\widehat{p}_{iht} = \widehat{r}_t \cdot \widehat{c}_h. \tag{3.4}$$

(6) 若 $h$ 类内的样本单元 $i$ 在第 $t$ 期调查时回答，则其权数调整为

$$w_{iht} = w_i \cdot (\widehat{p}_{ith})^{-1}. \tag{3.5}$$

对于任意两期 $t_1$, $t_2$ 的均值或总值变化进行估计时，我们采用这两期都有回答的样本数据，假设单元 $i$ 就是这样一个样本. 此时，对单元 $i$ 的调整权数计算步骤如下:

$$w_{i,t_1t_2} = \frac{1}{p_{it_1}p_{it_2} + \rho_{irt_1t_2}\sqrt{p_{it_1}(1-p_{it_1})p_{it_2}(1-p_{it_2})}}. \tag{3.6}$$

这是因为

$$p(r_{it_1} = 1, r_{it_2} = 1) = \mathsf{E}(r_{it_1}r_{it_2}) = p_{it_1}p_{it_2} + \rho_{irt_1t_2}\sqrt{p_{it_1}(1-p_{it_1})p_{it_2}(1-p_{it_2})},$$

其中假设单元 $i$ 在 $t_1$, $t_2$ 两期回答标识变量 $r_{it_1}$ 、 $r_{it_2}$ 服从二点分布， $\rho_{irt_1t_2}$ 是 $r_{it_1}$ 、 $r_{it_2}$ 的相关系数. 在模拟计算时，我们假设同一类 $h$ 内所有单元在 $t_1$, $t_2$ 两期回答标识变量的相关系数都等于 $\rho_{rht_1t_2}$, 可用类 $h$ 内 $t_1$、$t_2$ 两期回答标识变量样本相关系数进行估计. 一个特殊的情况是，若假设 $t_1$, $t_2$ 两期回答标识变量相互独立， $\rho_{ihrt_1t_2} = 0$, 则

$$w_{i,t_1t_2} = \frac{1}{p_{it_1}p_{it_2}} = w_{i,t_1}w_{i,t_2}. \tag{3.7}$$

对于任意两期 $t_1$, $t_2$ 的均值或总值变化进行估计时，即可采用 $w_{i,t_1t_2}$ 对第 $t_1$, $t_2$ 两期期都有回答的样本单元 $i$ 的调查权 $w_i$ 进行调整，单元 $i$ 的最终权数为

$$w_i^* = w_i \cdot w_{it_1t_2}. \tag{3.8}$$

这种加权调整基于对回答机制的假定. 这里假设的回答机制既考虑了类效应又考虑了调查期效应，有合理性. 这种加权调整可以同时满足横向分析与纵向分析的需要. 与 Lepkowski 的加权组方法相比，只需保存一组参数就可以方便得到满足横向分析及任意两期比较分析的权数. 特别在进行两期比较研究时，利用了两

期回答标识变量的相关信息，而 Lepkowski 的方法没有利用这一信息．与 Little 和 David(1983) 的方法相比，则容易处理“间歇式”期单元无回答问题．

### (二) 纵横加权调整方法的模拟研究

我们通过模拟运算对纵横加权调整方法和 Lepkowski 的加权调整方法进行了对比研究．下面介绍模拟的过程与结果．

(1) 模拟数据的产生：模拟样本为两期固定样组调查，样本量为 2000. 用随机数发生器产生服从下列分布的随机数各 2000 个：

$$z_i \sim N(10,\ 25), \qquad dz_i \sim N(1,\ 1), \qquad z_i \sim N(0,\ 25), \qquad e_i \sim N(10,\ 25),$$
$$e_{i1} \sim N(0,\ 16), \qquad z_i \sim N(0,\ 16), \qquad w_{0i} \sim U(1,\ 10),$$

令

$$y_{i1} = z_i + e_i + e_{i1}, \qquad y_{i2} = z_i + dz_i + e_i + e_{i2}, \qquad dy_i = y_{i2} - y_{i1}.$$

设 $z_i$ 为设计阶段就已知的辅助信息，$y_{i1}$ 为计量指标第一期的观测值，$y_{i2}$ 为计量指标第二期的观测值，$dy_i$ 为两期指标变化值，$w_{0i}$ 抽选样本的初始权数．完整样本第一期指标的加权均值为 $\overline{y}_1$、第二期指标加权均值为 $\overline{y}_2$，两期变化的加权均值为 $\overline{y}_d$.

(2) 对样本分组

随机分组：将 2000 个样本随机分为大小同为 500 个样本的四个样本组，则一、二两期共有八个加权组．

Z 分组：以 $z_i$ 为辅助信息，按 $z_i$ 单调增的顺序将样本分为大小同为 500 个样本的四个样本组，则一、二两期共有相应的八个加权组．

(3) 不回答机制

不回答机制同为 (3.1)，并满足第一期回答的条件下，第二期的回答概率是 $rr$.

(4) 比较准则

设第 $k$ 次模拟运算，采用 Lepkowski 的方法得到的第一期指标均值的估计、第二期指标均值估计、两期变化的均值估计分别为 $\overline{y}_{1Lk}$，$\overline{y}_{2Lk}$，$\overline{y}_{dLk}$；采用纵横加权调整方法得到的第一期指标均值的估计、第二期指标均值估计、两期变化的均值估计分别为 $\overline{y}_{1Mk}$，$\overline{y}_{2Mk}$，$\overline{y}_{dMk}$. 经过 2000 次模拟运算，计算下列均值，进行比较：

$$\mathrm{bias}1L = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{1Lk} - \overline{y}_1), \qquad \mathrm{bias}1M = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{1Mk} - \overline{y}_1),$$
$$\mathrm{bias}2L = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{2Lk} - \overline{y}_2), \qquad \mathrm{bias}2M = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{2Mk} - \overline{y}_2),$$
$$\mathrm{bias}dL = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{dLk} - \overline{y}_d), \qquad \mathrm{bias}dM = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{dMk} - \overline{y}_d),$$
$$\mathrm{mse}1L = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{1Lk} - \overline{y}_1)^2, \qquad \mathrm{mse}1M = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{1Mk} - \overline{y}_1)^2,$$
$$\mathrm{mse}2L = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{2Lk} - \overline{y}_2)^2, \qquad \mathrm{mse}2M = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{2Mk} - \overline{y}_2)^2,$$
$$\mathrm{mse}dL = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{dLk} - \overline{y}_d)^2; \qquad \mathrm{mse}dM = \frac{1}{2000}\sum_{k=1}^{2000}(\overline{y}_{dMk} - \overline{y}_d)^2.$$

(5) 模拟结果

随机分组，$rr = 0.9$

$$r_1 = 0.8, \qquad r_2 = 0.7, \qquad c_1 = 0.9, \qquad c_2 = 0.8, \qquad c_3 = 0.7, \qquad c_4 = 0.6$$

$\text{bias}1L = 0.1756$ $\text{bias}1M = 0.1322$
$\text{bias}2L = 0.1116$ $\text{bias}2M = 0.0861$
$\text{bias}dL = 1.6897e-2$ $\text{bias}dM = 1.3745e-2$
$\text{mse}1L = 7.1753e-2$ $\text{mse}1M = 8.6601e-2$
$\text{mse}2L = 7.8163e-2$ $\text{mse}2M = 9.2652e-2$
$\text{mse}dL = 1.6361e-2$ $\text{mse}dM = 1.7036e-2$

随机分组， $rr = 0.8$

$r_1 = 0.8$, $r_2 = 0.7$, $c_1 = 0.8$, $c_2 = 0.7$, $c_3 = 0.6$, $c_4 = 0.5$

$\text{bias}1L = 0.1414$ $\text{bias}1M = 0.1141$
$\text{bias}2L = 0.1393$ $\text{bias}2M = 0.1241$
$\text{bias}dL = 1.8435e-2$ $\text{bias}dM = 1.2009e-2$
$\text{mse}1L = 5.3867e-2$ $\text{mse}1M = 6.6221e-2$
$\text{mse}2L = 6.3893e-2$ $\text{mse}2M = 7.4004e-2$
$\text{mse}dL = 3.7417e-2$ $\text{mse}dM = 3.1102e-2$

Z 分组， $rr = 0.9$

$r_1 = 0.8$, $r_2 = 0.7$, $c_1 = 0.9$, $c_2 = 0.8$, $c_3 = 0.7$, $c_4 = 0.6$

$\text{bias}1L = 0.0777$ $\text{bias}1M = 0.1345$
$\text{bias}2L = 3.8659e-2$ $\text{bias}2M = 0.1242$
$\text{bias}dL = 3.8508e-2$ $\text{bias}dM = 1.9502e-2$
$\text{mse}1L = 3.6127e-2$ $\text{mse}1M = 5.0506e-2$
$\text{mse}2L = 2.4052e-2$ $\text{mse}2M = 6.0211e-2$
$\text{mse}dL = 1.9507e-2$ $\text{mse}dM = 1.6541e-2$

Z 分组， $rr = 0.8$

$r_1 = 0.8$, $r_2 = 0.7$, $c_1 = 0.8$, $c_2 = 0.7$, $c_3 = 0.6$, $c_4 = 0.5$

$\text{bias}1L = 0.1060$ $\text{bias}1M = 0.1345$
$\text{bias}2L = 0.0340$ $\text{bias}2M = 0.1243$
$\text{bias}dL = -0.0657$ $\text{bias}dM = 0.0191$
$\text{mse}1L = 0.0302$ $\text{mse}1M = 0.0728$
$\text{mse}2L = 0.0589$ $\text{mse}2M = 0.1043$
$\text{mse}dL = 0.0531$ $\text{mse}dM = 0.0392$

将模拟结果概括如下：

a. 在以上的个组模拟中，对变化估计量而言，无论从“偏差”的角度，还是从“均方误差”的角度，纵横加权调整方法总是优于 Lepkowski 的方法．其原因可能在于纵横加权调整利用了两期回答标识变量的相关信息，而 Lpkowski 的方法没有利用这一信息．

b. 在随机分组下，从“偏差”角度看，对三个估计量，纵横加权方法都优于 Lepkowski 的方法，但从“均方误差”的角度，对两个横向估计量，纵横加权方法不如 Lepkowski 的方法．原因可能是纵横加权方法得到的估计量方差较大．

c. 在 Z 分组下，对于两个横向估计量， Lepkowski 的方法优于纵横加权方法．原因可能在于， Z 分组采用了与调查指标高度相关的辅助信息分组，加大了组内调查指标的同质性，使 Lepkowski 方法得到的估计量方差和偏差减小，相比之下，纵横加权方法的方差较大．

## §4. 结　　论

我们提出的处理固定样组纵向连续调查的“间歇式”期单元无回答的纵横加权调整方法是一种基于对回答机制的假定的调整方法．这里假设的回答机制既考虑了类效应又考虑了调查期效应，有合理性．这种加权调整可以同时满足横向分析与纵向分析的需要．与 Lepkowski 的加权组方法相比，只需保存一组参数就可以方便得到满足横向分析及任意两期比较分析的权数．所作的模拟研究表明在假设的回答机制下，特别在进行两期比较研究时，利用了两期回答标识变量的相关信息，提高了变化估计量的准确度，而 Lepkowski 的方法没有利用这一信息．与 Little 和 David(1983) 的方法相比，则容易处理“间歇式”期单元无回答问题，同时避免了 Little 和 David(1983) 的方法在作两期变化分析时可能的复杂性．同时，当所采用的分组方法不能保证有效提高组内调查指标的同质性性时，纵横加权方法可能会减小估计量偏差．

## 参 考 文 献


[1] Lepkowski, *Traetment of Wave Non-response in Panel Surveys* in Panel Surveys, Daniel Kasprzyk ect eds, John Wiley & Sons, 1989.

[2] Cochran, W.G., *Sampling Techniques*, 3rd ed., New York: Wiley, 1977.

[3] Little and David, Weighting adjustment for non-response in panel surveys (U.S. Bureau of the Census working paper), 1983.

[4] Rubin, D.B., Inference and missing data, *Biometrika*, **63**(1976), 581–592.

[5] Little and Su, *Item Non-response in Panel Surveys* in Panel Surveys, Daniel Kasprzyk ect eds, John Wiley & Sons, 1989.


## Weighting Adjustment of Wave Non-response in Panel Surveys


YANG BAOHUI　　SUN SHANZE

(*School of Mathematical Sciences, Peking University, Beijing, 100871*)



Wave non-response is a sort of non-sampling error that often occurs in panel surveys. It usually causes bias of the estimator without adjustment. The two existing weighting adjustments encounter complexities in dealing with non-monotone wave non-responses. In this paper, we propose a novel weighting method, column and row weighting method, to adjust for the non-monotone wave non-response in panel surveys. Our method avoids the complexity of the existing methods. The simulation study We did shows that our method reduce the bias of the estimators and increase the accuracy of the estimator of the net change between two waves.