

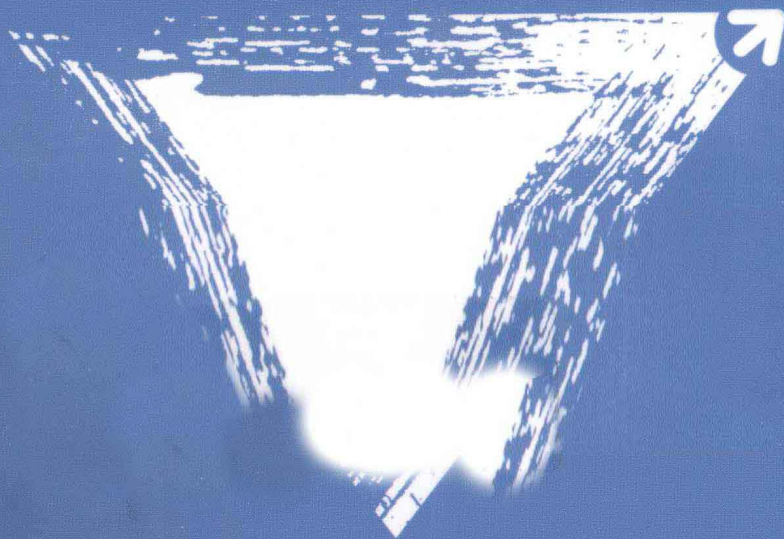
北京市人才强教深化计划资助项目 (项目编号: PHR20090505)

DIAOCHA YU FENXI

调查与分析

李朝鲜 主编

中国商业统计学会市场调查与教学研究分会
北京工商大学经济学院统计学系



经济科学出版社
Economic Science Press

北京市人才强教深化计划资助项目（项目编号：PHR20090505）

调查与分析

中国商业统计学会市场调查与教学研究分会
北京工商大学经济学院统计学系

李朝鲜 主编

经济科学出版社

图书在版编目 (CIP) 数据

调查与分析/李朝鲜主编. —北京: 经济科学出版社, 2011. 5

ISBN 978 - 7 - 5141 - 0519 - 3

I. ①调… II. ①李… III. ①市场调查②市场分析
IV. ①F713.52

中国版本图书馆 CIP 数据核字 (2011) 第 043598 号

责任编辑: 刘怡斐

责任校对: 王肖楠

版式设计: 代小卫

技术编辑: 王世伟

调查与分析

中国商业统计学会市场调查与教学研究分会

北京工商大学经济学院统计学系

李朝鲜 主编

经济科学出版社出版、发行 新华书店经销

社址: 北京市海淀区阜成路甲 28 号 邮编: 100142

总编部电话: 88191217 发行部电话: 88191540

网址: [www. esp. com. cn](http://www.esp.com.cn)

电子邮件: [esp@ esp. com. cn](mailto:esp@esp.com.cn)

北京欣舒印务有限公司印刷

三佳装订厂装订

787 × 1092 16 开 17 印张 330000 字

2011 年 5 月第 1 版 2011 年 5 月第 1 次印刷

ISBN 978 - 7 - 5141 - 0519 - 3 定价: 35.00 元

(图书出现印装问题, 本社负责调换)

(版权所有 翻印必究)

目 录

调查方法探讨

缺失数据处理方法的比较	庞新生	陈建成	1
第三种平衡的三层次轮换模式		侯志强	9
编制指数基础数据质量控制方法研究			
——以“义乌指数”为例		蒋剑辉	16
浅谈 CATI 技术在统计调查中的应用	安 颖	安瑞娟	26
网络调查应用问题分析	赵 然	安瑞娟	33
数据挖掘在营销调研中的应用		程开明	42

调研实践及报告

北京居民生活指数相关分析及差异分析	马立平	刘沛沛	49
北京市公共服务满意度指数研究	刘 娟	王 莉	54
居民收入对其安全感高低的影响分析	李 锋	张玉春	68
北京市城镇居民幸福感调查分析	彭 展	石 燕	74
导演评价研究			
——基于网络调查数据	艾小青	张净雨	84

对媒体廉政监督作用及现状的调查研究	李 宇	91
北京市最低工资标准调研分析	刘黎明 庞洪涛 李 锋 吴启富	100
北京市民新能源汽车购买意向调查分析	王 虹 杨 洋	110
北京市高校教师幸福感调研分析报告	周广军	122
北京市高校离退休职工的生活状况探析	吴海健 刘海燕	136
关于北京市民国际意识调查统计分析	唐 静 马立平	143
高校大学生满意度模型的构建及实证研究	刘 娟 赵倩倩	154
有效的时间管理是提高统计学成绩的必要条件	姚丽芳	166

统 计 分 析

北京市空气污染的空间统计分析	胡芳芳 董寒青	173
基于灰色模糊综合评价的建筑业竞争力研究	李 勇	179
试算内蒙古绿色国内生产总值	程士富 刘 倩	184
交通运输业对浙江省经济发展的影响分析	陈忠伟 张香云	195
房屋属性对价格的影响	罗玉波	204
房地产价格及其影响因素分析	李玉梅	211
中国股价指数与宏观经济变量之间影响关系 的实证分析	许玉娟 柴林涛	220
我国区域经济发展状况的多元统计分析	孔璐曼	229
我国科技发展水平影响因素及地区差异因子分析	杨双慧	236
ARMA 模型在我国居民消费价格指数预测中的应用	杨立圆	244
北京餐饮业影响因素的实证分析	周 媛	251
对我国居民消费价格指数的研究	宋红凤 陶林花	256

国 际 观 察

新西兰劳动力调查的现状与历史	侯志强	263
----------------------	-----	-----

调查方法探讨

缺失数据处理方法的比较

庞新生 陈建成

摘要：本文简要介绍了常用的缺失数据处理方法，讨论了缺失数据的处理方法评价标准，并对各种缺失数据的处理方法的特点及适用情况进行了比较。

关键词：缺失数据；处理方法；比较评价

设 $Y = (Y_1, Y_2, \dots, Y_N)^T$ 表示有缺失数据的所有变量的集合，为方便论述，以下称为研究变量。 Y 可以是一维，也可以是多维，通常， Y 指包括那些调查者感兴趣的目标变量。当然，若辅助变量数据发生缺失，也是可以纳入 Y 中。 Y_{obs} 为 Y 中的观测值； Y_{mis} 为 Y 中的缺失值。如果 Y 是一维的，在缺失数据处理中首先要识别缺失数据的产生机制；如果 Y 是多维的，对缺失数据进行处理时，要识别缺失数据的缺失模式。缺失数据的产生机制、缺失模式不同，所采用的方法也不同。目前常用的缺失数据处理方法大体上可以概括为列表删除、成对删除、加权调整、缺失数据插补方法等（如图 1 所示），其中，有些处理方法是基于完全随机缺失假设（Missing Completely at Random, MCAR）。一般来说，当数据不是 MCAR 而是随机缺失（Missing at Random, MAR）时，这些方法是不适用的；而有些方法（如似然估计法）在 MAR 的假设下是适用的。因此，在进行缺失数据处理时，首先需要认真分析缺失数据产生的原因，然后采取有针对性的补救措施，这样才能够获得无偏或弱偏估计。

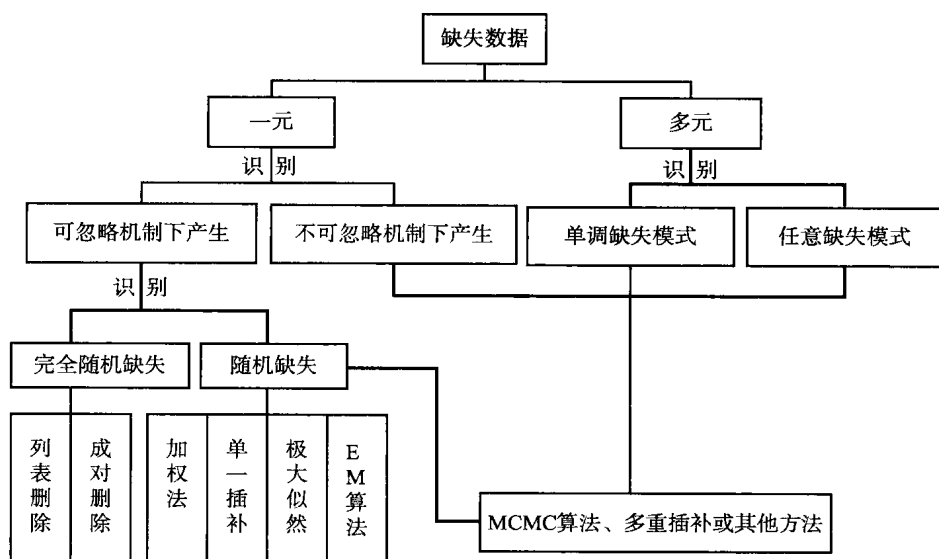


图1 缺失数据处理技术

一、常用缺失数据的处理方法

（一）忽略不完全观测值，直接估计

不考虑缺失数据的影响，直接在目前获取的数据基础之上进行分析，包括列表删除和成对删除。

1. 列表删除（或称为完全单位分析法）

列表删除是处理缺失数据最简单的方法，即删除观测不完全的变量，针对所有回答项目采用完全数据进行分析。这种方法简便，易于实施，不做任何修正，在 MCAR 的假设下，完全单位是原单位的一个有效子样本，丢弃不完全单位数据不会造成有偏估计。当观测变量较多时，样本规模的损失会相当大，当缺失数据多的时候，采用列表删除会放弃相当数量的信息，特别是当样本量较小的时候，采用这种方法会使数据量变得更少，估计效果较差；特别是当缺失数据为非随机缺失时，估计效果会更差。对表1设定的例子，如果采用列表删除法，推断分析时仅考虑单位2和单位4，其余都删除不考虑。

2. 成对删除（或称为有效单位分析法）

多元数据处理的另一种选择就是把目标变量进行过回答的所有单位都包括进来，这种方法就称为成对删除法。以表1数据为例，采用这种方法处理缺失数据，每个目标变量使用的样本构成如表2所示。它的缺点是根据缺失数据形式不

同，各个变量的样本基础总是不断变化。使用有效单位分析过程中，在 MCAR 下，均值和方差的估计可以直接计算，但要估计协方差或相关系数需要进行修整，具体方法见相关文献。

表 1 列表删除统计表

变量 单位编号	y_1	y_2	y_3	y_4	y_5
1	1	0	1	0	0
2	1	1	1	1	1
3	1	0	0	0	1
4	1	1	1	1	1

注：1 表示完全数据，0 表示缺失数据。

表 2 成对删除样本单位构成

目标变量	分析采用的样本构成
y_1	单位 1、单位 2、单位 3、单位 4
y_2	单位 2、单位 4
y_3	单位 1、单位 2、单位 4
y_4	单位 2、单位 4
y_5	单位 2、单位 3、单位 4

(二) 基于插补的技术

通过上面的分析可以看出，简单地将数据删除并不是好的方法。替换缺失数据技术的意义在于比列表删除浪费更少的信息，而且当缺失数据为非随机缺失时，替换缺失数据技术比列表删除更稳健。基于插补的缺失数据处理技术是用适当的估计补全缺失数据，这样就允许将标准完全数据分析方法用于分析插补后的数据集。插补法主要用于项目无回答情况下的调整，根据每个缺失值替代值的个数，可以分为单一插补和多重插补。

1. 单一插补

单一插补是指对每一个由于无回答造成的缺失值只构造一个替代值。单一插补是以估算为基础的方法，是在替代缺失数据后对新合成的数据进行相应的统计分析。

(1) 根据插补中辅助信息来源的不同，单一插补方法可以划分为热卡插补和冷卡插补。热卡插补是从每一个缺失数据的估计分布抽取插补值替代缺失值，使

用回答单元的抽样分布作为抽取分布是最常见的方法。与冷卡插补相比,热卡插补使用当前调查的回答值,而冷卡插补则使用其他资料中的回答数据或先前同样的调查或普查中的历史数据。

(2) 根据插补值与缺失值关系,单一插补可以划分为相似回答单一插补与其他类型的单一插补。相似回答插补的基本思想是找寻与含有缺失数据的样本单位最相近的不存在缺失值的样本单位,用相近样本单位的数据去补全缺失数据。其他类型的单一插补又划分为均值插补、随机插补、回归插补和随机回归插补,具体方法见相关文献。

单一插补改变了传统方法将缺失值忽略不考虑的习惯,使得各种统计分析均可以在插补后的完整数据集上展开;但单一插补的缺点也是显而易见的:无论采用何种方法,都存在扭曲样本分布的问题。

2. 多重插补

多重插补是在单一插补的基础上衍生来的,由鲁宾(Rubin)在1977年首先提出。它是指给每个缺失值都构造一个以上的替代值,这样就产生了若干个完全数据集,对每个完全数据集分别使用相同的方法处理,得到若干个处理结果,最后再综合这些处理结果,最终得到目标变量的估计。多重插补法的出现弥补了单一插补法的缺陷:第一,多重插补过程产生多个中间插补值,可以利用插补值之间的变异反映无回答的不确定性,包括无回答原因已知情况下抽样的变异性和无回答原因不确定造成的变异性。第二,多重插补通过模拟缺失数据的分布,较好地保持变量之间的关系。第三,多重插补能给出衡量估计结果不确定性的大量信息,单一插补给出的估计结果则较为简单。与单一插补相比,多重插补唯一的缺点是需要做大量的工作来创建插补集并进行结果分析,因为它主要是执行若干次相同的任务而非一次,然而数据分析中大量工作在今天的计算环境下是比较容易实现的。

(三) 基于模型的方法

既不是删除缺失值也不是采用插补方法去补全缺失值,而是首先要考虑缺失数据的缺失机制(MCAR、MAR和NMAR),在此基础上为部分缺失数据定义模型,模型的参数可以通过极大似然估计或极大后验估计,常用的方法有极大似然估计、EM算法、MCMC方法(其中包含DA算法)。这里主要讨论完全信息极大似然估计,这是基于模型的方法,可直接用于不完全数据的分析,其最大特点在于即使缺失数据不是完全随机缺失,估计的结果也是无偏的。

完全信息极大似然估计是建立在对数极大似然估计基础上的,假定数据来源于多元正态分布,对于不完全服从多元正态的数据还是稳健的。极大似然估计的不足之处在于需要相对比较大的数据集,而且可供推断的信息是有限的。当样本

量太小时,不宜采用完全信息极大似然估计。

二、缺失数据处理方法评价

通过上面的分析,我们可以看出解决缺失数据的方法各有特点,不存在既简便易行,估计效果又理想的方法,而且大部分方法在使用之前要求满足如下假定:一是多元正态分布;二是所有变量独立同分布;三是数据随机缺失。因此在进行缺失数据处理之前,应该对于上述假定逐一判断。应该指出的是,一个好的缺失数据处理方法应该合理反映缺失数据的不确定性,能够保持数据分布特征,保证变量间重要关系不会因调整后发生变化。

(一) 缺失数据的处理方法评价标准

对缺失数据的处理并不是为了寻找最理想的点预测,而是要获得有效的统计推断。通常情况下,在讨论处理缺失数据技术时,应该注意以下几个问题:第一,是不是每一种缺失数据处理方法对于由非随机缺失引起的偏差都是稳健的;第二,是不是每一种缺失数据处理方法都可能引入偏差;第三,缺失数据处理方法利用信息的程度如何;第四,缺失数据处理方法所适用的数据缺失类型。

如果缺失数据不是随机缺失的,数据分析可能导致偏差,除非分析方法能够纠正由于非随机缺失所造成的偏差。以抽样为基础的方法如多重插补和相似回答插补假定数据是 MCAR,而模型为基础的方法如完全信息极大似然估计仅假定数据是 MAR,利特尔(Little)和鲁宾指出,即使 MAR 的条件并不能严格满足,使用完全信息极大似然估计将减少偏差。也就是说,即使 MAR 的条件并不能严格满足,完全信息极大似然估计是一致的和有效的,不幸的是,完全根据经验弄清楚一个样本分布是 MCAR、MAR 或非随机缺失是很困难的,在决定这个问题上,先验知识是必要的。

如果缺失数据处理方法导致偏差,我们可以比较原始数据集和经缺失数据处理方法处理后获得的数据集的数字特征,如中位数、均值和标准差,通过统计检验来判断两者的异同,即使这些检验看起来非常简单,但在大部分关于缺失数据处理方法的统计研究中是必须的。假如采用列表删除的数据集分布与原始数据集分布相似时,我们说缺失数据处理方法在列表删除数据集中没有引入偏差,而该方法不适宜于完全信息极大似然估计。信息损失阻止程度可通过比较每个数据集完全单位的数目与原始数据集所有单位数目进行评价。

由于插补技术是一种非常重要的缺失数据处理方法,因此,在对各种插补方法进行比较时,需要注意以下几个原则:第一,插补必须是建立在缺失数据的预测分布基础之上;第二,在考虑插补时,完全回答变量必须考虑在内;第三,插

补必须基于需要插补变量的辅助信息；第四，超越数据取值过分的外推是要避免的；第五，为保持完全数据集的分布，插补值必须从预测分布中抽取；第六，必须提供一种把插补值考虑在内的抽样估计误差计算方法。均值插补中所有缺失数据采用同一插补值，是唯一不满足任何原则的方法。回归插补和基于 EM 算法的多重插补满足其中的两个原则；随机回归插补和基于 DA 算法的多重插补满足四条原则，在四条原则的基础上，随机回归插补和基于 DA 算法看起来最有发展前景，其次是回归插补、基于 EM 算法的多重插补，最差的是均值插补，具体比较见表 3。

表 3 插补方法比较

缺失数据处理方法	预测分布	是否包含完全回答变量	从预测分布抽取	考虑不确定性
均值插补	-	+	-	-
回归插补	+	+	-	-
随机回归插补	+	+	+	+
极大似然估计	+	+	-	-
基于 EM 算法的 MI	+	+	-	-
基于 DA 算法的 MI	+	+	+	+

注：① + 表示满足此原则；
② - 表示不满足此原则。

（二）缺失数据的处理方法比较

上述不同缺失数据的处理方法的选择主要取决于研究数据的性质和质量、数据的使用者和用途，以及缺失数据的广度和内在机制。如前面提到的，忽略不完全观测值仅仅在缺失数据比例很小，而且是可以忽略的情况下才不至于引入显著的偏差。

根据上述标准，对各种缺失数据处理方法进行比较分析，结果见表 4。从该表可以看出，在大样本情况下，完全信息极大似然估计是最好的选择，当存在缺失数据时，由于完全信息极大似然估计具有比其他方法更好的稳健性。如果缺失数据是非完全随机缺失，在使用列表删除时，应首先讨论是否是 MCAR，如果观测值与缺失值之间存在系统差异，采用列表删除是非常危险的。事实上，没有哪一种缺失数据处理方法能够非常满意地纠正由于缺失数据造成的偏差，更危险的是，没有缺失数据处理方法和检验可以发现这种偏差。因此，必须依靠专业知识去判断是否为随机缺失，假如缺失数据很多，而且可能为非随机缺失，不提倡使用任何缺失数据处理方法，包括列表删除，在这种环境中唯一的解决办法就是找寻真值，例如对无回答者的再次访问。但是，如果缺失数据非常少，数据缺失在 5% 之内，可以采用列表删除或其他缺失数据处理方法，而不至于引入大的误差。对于以模型为基础的方法如极大似然估计、EM 算法、DA 算法，不容易操

作, 仅适合于非常熟悉缺失数据机制并且具有处理缺失数据的专长和工具的专业人士。

表 4 各种缺失数据处理方法的比较

缺失数据处理方法	前提	信息利用程度	难易	适用范围	稳健性	偏差
列表删除	MCAR	删除很多有用信息	易	缺失 5% 内	很差	大
成对删除	MCAR	局限于回答信息	易	几乎不用	很差	大
均值插补	MCAR	局限于回答信息	易	少用	很差	低估方差及抽样误差
回归插补	MAR	局限于回答信息	较难	广泛	差	低估方差及抽样误差
冷卡插补	MAR	利用了过去信息	易	少用	差	低估方差及抽样误差
热卡插补	MAR	局限于回答信息	易	广泛	差	方差估计较好及抽样误差不易控制
随机回归插补	MCAR	局限于回答信息	难	少用	差	偏差较小
极大似然估计	多元正态分布	信息利用充分	难	大样本	很好	弱偏或无偏
多重插补	MAR	利用充分, 考虑了缺失值的不确定性	难	较广	好	弱偏
基于 EM 算法多重插补	MAR	利用充分, 考虑了缺失值的不确定性	很易	较广	好	弱偏
基于 DA 算法多重插补	MAR	利用充分, 考虑了缺失值的不确定性	很难	较广	好	弱偏

注: ①MCAR 为完全随机缺失;

②MAR 为随机缺失。

综上所述, 没有哪一种处理缺失数据的方法是普遍适用的, 每种方法都存在不尽如人意之处, 但这毕竟是对不完全数据集分析的一种尝试。对于这一问题的研究越来越引起人们的重视, 对于现有的方法, 应该持一种科学态度谨慎对待, 根据每一种方法的特点结合实际问题加以分析、选择和应用。

参考文献

- [1] Donald. B. Rubin, Multiple Imputation For Nonresponse In Surveys [M], New York: John Wiley & Sons Inc. 1987: 75 - 112.
- [2] Little, R. J. A. , Discussion Proceedings Of The Survey Research Methods Section Of The American Statistical Association [J], Alexandria, VA, 1995: 60 - 62.
- [3] Little, R. J. A. and Rubin, D. B. , The Analysis Of Social Science Data With Missing Value [J]. Sociological Methods And Research, Vol. 18. nos. 2/3, 1990: 292 - 326.
- [4] Roderick J. A. Little, Donald B. Rubin, Statistical Analysis With Missing Data [M], New York: John Wiley & Sons Inc. 2002: 218 - 265.

作者简介

庞新生（1970 年 ~ ），男，山西榆次人，博士、北京林业大学副教授

研究方向：抽样技术

电子邮箱：pxinsheng@sina.com

通信地址：北京林业大学经管学院，邮编 100083

陈建成（1963 年 ~ ），男，山西芮城人，博士、北京林业大学教授

研究方向：应用统计

第三种平衡的三层次轮换模式*

侯志强

摘要：三阶段抽样连续调查需要对每级单元实施样本轮换，如何使得三个级别单元的每次轮换相互协调是关键问题所在。本文针对此提出了一种新的轮换模式，并将此模式与之前提出的两种模式做了三个方面的对比。

关键词：三阶段抽样；连续调查；样本轮换；轮换模式

一、引言

根据国务院办公厅《关于建立劳动力调查制度的通知》^[1]，我国于2005年11月开始进行第一次全国劳动力调查，2006年每半年进行一次，2007年每个季度进行一次。根据世界大多数国家或地区劳动力调查^[2]的历史经验，随着我国经济的进一步发展，劳动力调查必将发展为月度调查。我国劳动力调查抽样设计的主体部分是三阶段抽样。^[3]一级抽样单元为乡级单元，即乡、镇或街道。二级抽样单元为村级单元，即村委会或居委会。三级抽样单元为小区，每个小区由约30户最相邻的住址组成，是村级单元的一个划分。由于劳动力调查采用了连续调查的形式，所以每级抽样单元均存在样本老化^[4]问题。当劳动力调查发展为月度调查的时候，各级单元样本老化的问题将更加突出。关于三个级别单元如何实施轮换才能保证每次轮换的协调统一问题，世界上还没有文献论述。对于月度三阶段抽样连续调查，笔者曾经设计了两种轮换模式。在此基础上，本文又提出了一种新的轮换模式。

二、第三种平衡三层次轮换模式的构造

所谓平衡三层次轮换模式，是指对于任何一级单元的任何一次轮换不仅各级单元的样本量不变，而且对于同一级单元而言，任何轮换掉的单元均接受过相同次数的调查。为此，每次调查时三个级别的单元均应同时进行轮换，并且各级单元的轮换模式可以确定如下：一级单元24 in，二级单元4-8-4（8），三级单元

* 全国统计科学研究（计划）项目（2010LC16）；北京市教育委员会科技计划面上项目；北京市大学生科学研究与创业行动计划项目。

2-10-2 (4)。按照这个模式,三个级别单元的每次轮换均相互衔接,构成一个有机整体。为了详细说明具体的轮换过程,特制作表1如下。

表1

样本轮换表

		第 m 子样本	第 $m+1$ 子样本
第 t 年	1 月	ABCDEF GHIJ KLMNOP QRSTUV WX	
		333322221111333322221111	
	2 月	BCDEF GHIJ KLMNOP QRSTUV WX	A
		333322221111333322221111	1
	3 月	CDEF GHIJ KLMNOP QRSTUV WX	AB
		333322221111333322221111	11
	4 月	DEF GHIJ KLMNOP QRSTUV WX	ABC
		333322221111333322221111	111
	5 月	EFGHIJ KLMNOP QRSTUV WX	ABCD
		333322221111333322221111	1111
	6 月	FGHIJ KLMNOP QRSTUV WX	ABCDE
		333322221111333322221111	21111
第 $t+1$ 年	1 月	GHIJ KLMNOP QRSTUV WX	ABCDEF
		333322221111333322221111	2211111
	2 月	HIJ KLMNOP QRSTUV WX	ABCDEF G
		333322221111333322221111	22211111
	3 月	IJKLMNOP QRSTUV WX	ABCDEF GH
		333322221111333322221111	222211111
	4 月	JKLMNOP QRSTUV WX	ABCDEF GHI
		333322221111333322221111	3222111111
	5 月	KLMNOP QRSTUV WX	ABCDEF GHIJ
		333322221111333322221111	3322211111
	6 月	LMNOP QRSTUV WX	ABCDEF GHIJ K
		333322221111333322221111	33322211111
第 $t+2$ 年	1 月	MNOP QRSTUV WX	ABCDEF GHIJ KL
		333322221111333322221111	333322211111
	2 月	NOP QRSTUV WX	ABCDEF GHIJ KLM
		333322221111333322221111	133332221111
	3 月	OP QRSTUV WX	ABCDEF GHIJ KLMN
		333322221111333322221111	1133332221111
	4 月	PQRSTUV WX	ABCDEF GHIJ KLMNO
		333322221111333322221111	11133332221111
	5 月	QRSTUV WX	ABCDEF GHIJ KLMNOP
		333322221111333322221111	111133332221111
	6 月	RSTUV WX	ABCDEF GHIJ KLMNOPQ
		333322221111333322221111	2111133332221111
第 $t+3$ 年	1 月	STUV WX	ABCDEF GHIJ KLMNOPQR
		333322221111333322221111	22111133332221111
	2 月	TUV WX	ABCDEF GHIJ KLMNOPQRS
		333322221111333322221111	222111133332221111
	3 月	UV WX	ABCDEF GHIJ KLMNOPQRST
		333322221111333322221111	2222111133332221111
	4 月	V WX	ABCDEF GHIJ KLMNOPQRSTU
		333322221111333322221111	3222111133332221111
	5 月	W WX	ABCDEF GHIJ KLMNOPQRSTUV
		333322221111333322221111	33222111133332221111
	6 月	X WX	ABCDEF GHIJ KLMNOPQRSTUVW
		333322221111333322221111	333222111133332221111

三、轮换模式的解释

(一) 表1说明

在表1中,每个大写英文字母表示一个一级轮换组,每个一级轮换组由若干个一级单元组成;每个阿拉伯数字表示一个二级子样本,每个二级子样本由若干个二级单元组成;每个算术运算符表示一个三级子样本,每个三级子样本由若干个三级单元组成。

每个子样本包含24个一级轮换组,分别用“A”~“X”表示;每个一级轮换组包含3个二级子样本,分别用“1”、“2”和“3”表示;每个二级子样本又包含2个三级子样本,分别用“+”和“-”表示。

每一个月的样本由3行24列共72个符号表示。每一行表示一个某一级别的样本。例如,第 t 年1月的第一行表示一个由24个一级轮换组所组成的一级样本;第二行表示一个由24个二级子样本所组成的二级样本;第三行表示一个由24个三级子样本所组成的三级样本。每一列的三个符号依次表示一级轮换组、一级轮换组内的二级子样本及二级子样本内的三级子样本。例如,第1列的3个符号分别表示一级轮换组“A”、“A”内的二级子样本“3”及“3”内的三级子样本“-”。

每次调查的样本来自于一个或两个子样本。例如,第 t 年1月的样本来自于第 m 子样本,第 t 年2月的样本来自于第 m 子样本和第 $m+1$ 子样本。

(二) 轮换解释

1. 一级轮换组的轮换

每个一级轮换组从首次进入样本到退出样本,连续接受24次调查,历时两年。以第 m 子样本内一级轮换组“X”为例。从第 t 年1月到第 $t+1$ 年12月,一级轮换组“X”连续接受24次调查,之后永久退出样本。

每个月,有1个一级轮换组接受第1次调查,有1个一级轮换组接受第2次调查,……,有1个一级轮换组接受第24次调查。如第 $t+1$ 年12月,“W”接受第1次调查,“V”接受第2次调查,……,“X”接受第24次调查。

2. 一级轮换组内二级子样本的轮换

每个二级子样本连续接受4次调查,然后暂时退出样本8个月,接着再次进入样本连续接受4次调查,之后永久退出样本。以第 m 子样本内一级轮换组“X”内的二级子样本“1”为例,从第 t 年1月到4月,“1”连续接受4次调查,之后暂时休息8个月,第 $t+1$ 年1月“1”再次进入样本,并从1月到4月

连续接受4次调查,之后永久退出样本。每个二级子样本总共接受8次调查,历时1年零4个月。

下面以一级轮换组“X”内二级子样本的轮换为例,说明一级轮换组内二级子样本的轮换过程。第 t 年1月到4月,一级轮换组“X”内的“1”连续接受4次调查,5月到8月,“2”连续接受4次调查,9月到12月,“3”连续接受4次调查,第 $t+1$ 年1月到4月,一级轮换组“X”内的“1”再次接受4次调查,5月到8月,“2”再次接受四次调查,9月到12月,“3”再次接受4次调查。

每个月,有3个二级子样本接受第1次调查,有3个二级子样本接受第2次调查,……,有3个二级子样本接受第8次调查。如第 $t+1$ 年12月,“W”内的“1”、“S”内的“2”和“O”内的“3”接受第1次调查;“V”内的“1”、“R”内的“2”和“N”内的“3”接受第2次调查,……,“H”内的“1”、“D”内的“2”和“X”内的“3”接受第8次调查。

每个月有3个二级子样本在接受完第4次调查后进入8个月的暂时休息期,有3个二级子样本在经过8个月的暂时休息后重新返回样本接受第5次调查。如第 $t+1$ 年12月,“T”内的“1”、“P”内的“2”和“L”内的“3”在接受第4次调查后进入8个月的暂时休息期,而“K”内的“1”、“G”内的“2”和“C”内的“3”在经过8个月的休息后重新返回样本接受第5次调查。

3. 二级子样本内三级子样本的轮换

每个三级子样本连续接受2次调查,然后进入10个月的暂时休息期,接着再次进入样本接受2次调查,之后永久退出样本。以第 m 子样本内一级轮换组“X”内二级子样本“1”内的三级子样本“+”为例。第 t 年1月到2月,“+”连续接受2次调查,之后暂时休息10个月,第 $t+1$ 年1月到2月,“+”再次连续接受2次调查,之后永久退出样本。每个三级子样本总共接受4次调查,历时1年零2个月。

下面以一级轮换组“X”内二级子样本“1”内的三级子样本的轮换为例,说明二级子样本内三级子样本的轮换过程。第 t 年1月到4月,“X”内“1”内的“+”接受前两个月的调查,“-”接受后两个月的调查,第 $t+1$ 年1月到4月,“X”内“1”内的“+”再次接受前两个月的调查,“-”也再次接受后两个月的调查。

每个月有6个三级子样本接受第1次调查,有6个三级子样本接受第2次调查,……,有6个三级子样本接受第4次调查。例如,第 $t+1$ 年12月,“W”内“1”内的“+”、“U”内“1”内的“-”、“S”内“2”内的“+”、“Q”内“2”内的“-”、“O”内“3”内的“+”和“M”内“3”内的“-”接受第1次调查;“V”内“1”内的“+”、“T”内“1”内的“-”、“R”内“2”内的“+”、“P”内“2”内的“-”、“N”内“3”内的“+”和“L”内“3”内的“-”接受第2次调查之后,这些三级子样本进入10个月的暂时休息期;