DOI:10.3969/j.issn.1672-7347.2013.12.014
http://xbyx.xysm.net/xbwk/fileup/PDF/2013121289.pdf

缺失数据的识别与处理

沈琳[1]，陈千红[2]，谭红专[1]

(1. 中南大学公共卫生学院流行病与卫生统计学系，长沙 410078；2. 湖南邮电职业技术学院，长沙 410015)

[摘要] 目的：数据缺失在调查研究中是一个非常普遍的现象，它的出现造成部分原始样本信息的损失，在一定程度上危害研究结果的有效性，需要引起研究者的重视。缺失数据产生的3类机制为完全随机缺失、随机缺失、非随机缺失。目前常见的缺失数据处理方法包括删除法、加权调整法、插补法、参数似然法，其各有优缺点，应针对缺失数据产生的机制选择相应的处理方法。

[关键词] 缺失数据；缺失机制；调查研究

Identification and treatment of missing data

SHEN Lin[1], CHEN Qianhong[2], TAN Hongzhuan[1]

*(1. Department of Epidemiology and Health Statistics, Central South University, Changsha 410078;
2. Hunan Post and Telecommunication College, Changsha 410015, China)*

ABSTRACT Missing data plagues almost all surveys and researches. The occurrence of missing data will cause losses of original sample information and undermine the validity of the research results to some extents, so researchers should attach great importance to this problem. In this article, we introduced 3 kinds of missingness mechanism, namely missing completely at random, missing at random, and not missing at random. We summarized some common approaches to deal with missing data, including deletion, weighting approach, imputation and parameter likelihood method. Since these methods had its pros and cons , we should carefully select the proper way to handle missing data according to the missingness mechanism.

KEY WORDS missing data; missingness mechanism; survey and research

数据缺失(missing data)在调查研究中是一个非常普遍的现象，它的出现造成部分原始样本信息的损失，在一定程度上危害研究结果的有效性[1]。在以往的流行病学调查研究中，人们对于处理缺失数据的重视程度不够或存在方法误用的情况，往往只是简单地将有缺失值的对象剔除，仅对完全记录对象进行分析，甚至不予处理[2]。在科学技术快速发展的今天，传统的缺失数据处理方式

收稿日期(Date of reception)：2013–06–06

作者简介(Biography)：沈琳，硕士研究生，主要从事流行病与卫生统计学研究。

通信作者(Corresponding author)：谭红专，Email: tanhz99@qq.com

显然已无法满足现实研究的需求，合理有效的缺失数据处理方法亟待应用与规范。2007年出版的关于流行病学研究中的观察性研究的报告要点(strengthening the reporting of observational studies in epidemiology，STROBE)声明中就明确指出，规范化的观察性研究文献应将缺失数据的处理方法表述清楚，并报道个体或单元无响应的原因[3]。可见，如何正确有效地处理缺失数据应引起流行病学研究者的重视。

缺失数据给研究结果带来的危害程度取决于数据缺失的机制、缺失数据的数量和造成缺失的原因，其中最为重要的是数据缺失的机制[4]。各种处理缺失数据的方法均建立在缺失数据机制的某种假定之上，因此，明确数据缺失的机制是正确选择缺失数据分析方法的前提。本文系统地介绍了缺失数据产生的机制以及如何识别这些机制的方法，在此基础上总结比较了几种最新的处理缺失数据的方法及其优缺点，并对目前认可度较高的多重插补方法予以举例说明，以期为医疗卫生工作者正确处理缺失数据提供实用参考。

1 数据缺失的机制

缺失数据机制表述了缺失数据与数据集中变量值之间的关系，目前公认的分类方法是采用Little和Rubin于1976年提出的理论框架[5]，分为3类。

1.1 完全随机缺失

完全随机缺失(missing completely at random，MCAR)指目标变量是否缺失与自身或其他观察变量的取值无关，此类缺失会导致信息的缺失。例如，在高血压危险因素研究的问卷调查中，调查人员不慎遗失了几份问卷，没有理由表明丢失问卷(数据缺失)这一事件与被调查者的高血压值或其他变量有任何关系，即缺失的发生完全随机，此时，可以把观察到的单元看作是从样本单元中简单随机抽取的子样本。在这种缺失机制下，对含有缺失值的数据集采用通常的统计分析方法是可行的，估计量无偏，但不同方法的估计效率存在差别，基于模型的方法比简单的估计方法有更高的效率。现实中存在MCAR的情况，但并不普遍。

1.2 随机缺失

随机缺失(missing at random，MAR)指有缺失值的变量，其缺失情况的发生只与已观察到的变量值有关。此类数据缺失较为常见，不仅导致信息的缺失，更可能导致分析结论发生偏差。例如，某中老年女性骨密度的调查资料中骨密度值有缺失，缺失情况主要发生在高龄组，是由于高龄组受访者行动不便未能到现场接受访谈和检查造成数据缺失。此时若直接删除有缺失值的观察单位，可能造成骨密度值的错误高估[6]。大多数缺失数据的分析方法是基于MAR的假设。

1.3 非随机缺失

非随机缺失(not missing at random，NMAR)，也称为不可忽略的缺失(nonignoarble missing，NIM)，指缺失数据不仅与其他变量有关，也与自身取值有关。例如，对某人群进行收入调查，高收入者不愿填写家庭人均年收入值。这种缺失比较难于处理，进行处理时需要基于目标变量和协变量模型比较强的假定条件。解决非随机缺失的一种思路是将其有条件地转化为随机缺失模式。例如，利用辅助变量将样本单元类别细化，使同类别中样本单元的目标变量值接近。

2 数据缺失机制的识别

在资料的分析处理过程中，研究者需要了解哪些变量有缺失值、缺失数据的数量和范围以及数据缺失发生的原因，这些信息可以借助一些统计软件来获得。例如，SPSS软件中的“Missing Value Analysis”模块，就提供了针对缺失值问题全面而强大的分析能力，包括对缺失值的描述和快速诊断等。

带缺失值的数据，其缺失的机制是MCAR或MAR，还是NMAR？一般来说，研究者缺少有关数据缺失原因的详细信息，对数据缺失机制的推测主要依靠其在数据收集阶段对数据缺失原因的了解和在这个研究领域丰富的知识背景。研究者也许可以根据某些研究对象经常无应答，特别是年老者或患重病者易缺失，排除数据MCAR的可能。另一种识别MCAR的方法是为每一个有缺失值的变量建立一个指示变量(当观测变量没有缺失值时，指示变量赋值0，有缺失值时指示变量赋值1)，并把这个缺失值指示变量作为一个协变量或预测因子放入回归模型中分析，这样所有观察对象都能利用分析。当指示变量的回归系数有统计学意义时，研究者可以推断该观察变量的无应答者情况不同于应答者的。因此，该观测变量的缺失机制不是MCAR。然而这种方法会高估回归的残差。Little等[5]提出了用似然比法检验MCAR的假设，当卡方值超过临界值时，该数据缺失的机制不是MCAR；也可以假设数据缺失的机制分别是MCAR，MAR或NMAR，用不同的分析方法去分

析，然后检查不同分析方法结果的灵敏性，这样也能为了解数据缺失的机制提供信息。另外还可对无应答者进行追踪调查，调查无应答的原因以及应答者与无应答者之间有无本质差别，进而识别数据缺失的机制[7]。

3 缺失数据处理方法的选择

要减少调查中的缺失数据，主要应从事前预防和事后补救两方面入手。事前预防是处理缺失数据最简便有效的方法，但现实中由于条件限制，往往不能完全解决问题。一般而言，在事后补救上，缺失数据处理方法大体上可以概括为以下几种方法(图1)。

3.1 删除法

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(listwise deletion)和成对删除(pairwise deletion)。

3.1.1 列表删除

也称为完全记录分析(complete case analysis)，这是一种最简单的缺失数据处理方法，即只对要分析的变量都有观察值的对象进行分析，放弃对有缺失值的对象进行分析的方法。此法优点在于简便易行，不存在编造的数据，当数据缺失机制是MACR时，完全记录分析的结果一般来说是有效的，这是因为此时有完全记录的对象是原人群的一个随机样本。不足是当有大量数据缺失时，进入分析的对象会很少，这样会降低统计检验的效能，浪费许多信息[8]。因此，此法仅适用于缺失数据少而样本量较大的情况，且为完全随机缺失的情况。

3.1.2 成对删除

也称有效单位分析法(available case analysis)，指用数据集中所有能利用的数据进行参数估计的方法。这种方法相较于列表删除法而言所利用到的信息更多，使用了所有有效的变量值，它的缺点是根据缺失数据形式不同，分析各个变量时的样本总在不断变化，一般用于回归分析、区组设计资料分析和因子分析中，当关键分析变量缺失较少且为完全随机缺失时适用。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就需要进行修整[9]。

3.2 加权调整法

加权调整就是当出现缺失单元时，通过某种方式把缺失单元的权数分解到非缺失单元(即观测值)身上。它通过增大调查中有观测数据的权数，以减小由于缺失数据可能对估计量带来的偏差。加权调整主要针对单元无应答，尽管有些场合下也可应用于项目无应答。在用该方法处理缺失值时，缺失值的数量很重要。假设不存在无应答偏倚的情况下，该方法要求有足够的未应答个体的信息。当单元无应答和项目无应答都存在的情况下，该方法变得无计可施。该方法的优点是偏倚较低，但它使得标准误计算变的更难[10]。几种主要的加权调整方法包括加权组调整法、再抽样调整法、事后分层调整法、迭代调整法、校准法、双重稳健加权法[11-12]等。

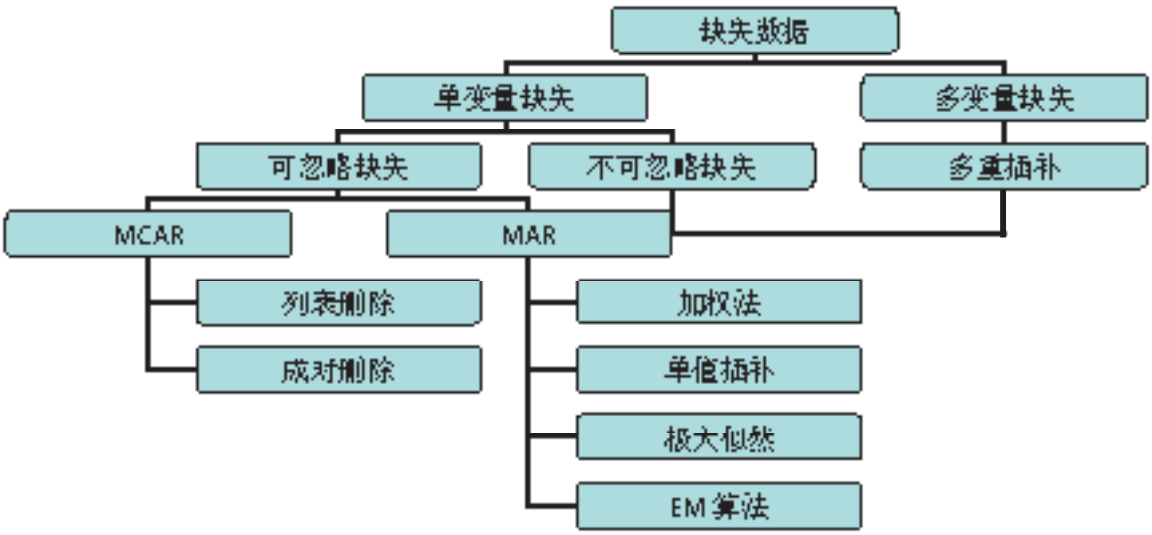

图1 缺失数据常用处理方法。

Figure 1 Common approaches to dealing with missing data.

3.3 插补方法

所谓“插补”(imputation)是指给每一个缺失数据一些替代值，这样可以得到“完整的数据集”，然后用标准的完全数据统计方法进行数据分析和推断，这些替代值称为插补值。如果说加权调整法主要用于单元无应答造成的缺失数据，那么插补法则主要处理项目无应答造成的缺失数据。一般说来，插补值不会提高估计的精度，因为插补值毕竟是“假信息”，但它的优点是可以采用完全数据分析的方法，并减少估计偏差。根据对每个缺失值的插补值的个数，插补法可分为单一插补法(single imputation)和多重插补法(multiple imputation)。

3.3.1 单一插补

单一插补是对每个缺失值，用预测分布的平均值或从中抽取一个值充缺失值。它以观测数据为基础，为插补创建一个预测分布。如何预测这个分布可以有两种途径：其一，明确建模，即预测模型基于一个常用的统计模型，这类方法有均值插补、比率插补、回归插补等；其二，模糊预测，即采用某个算法，该算法蕴含一个基本模型，但假定是模糊的，使用时需判断假定是否合理，这类方法主要有最近距离插补、热卡插补、冷卡插补等[13-15]。

3.3.2 多重插补

多重插补由Rubin最早提出，它是要求在数据随机缺失情况下，用两个或更多能反映数据本身概率分布的值来插补缺失或不完善数据的一种方法。在多重插补中，数据填补是关键环节，对每一个缺失数据填补m(m>1)次。这样，第一次填补就产生第一个完全数据集，以此类推，将产生m个完全数据集。对每一个完全数据集都采用标准的完全数据分析的方法进行分析，并将所得结果进行综合，最终得到对目标变量的估计[16](图2)。多重插补并不是试图通过模似值去估计每个缺失值，而是去代表缺失值的一个随机样本。其优点是根据被观测数据的后验预测概率分布，并充分考虑缺失值的不确定性，对缺失值进行多次填充，其结果估计是比较可靠和有效的。多重插补作为统计工具，能充分有效地利用被观测到的信息。多重插补方法需要根据资料类型和缺失模式来选择相应的填补方法。对于单调缺失模式(即若矩阵中元素X_{ij}缺失，通过对数据矩阵进行适当行列变换后，对任意的$k \geqslant i$ 和$n \geqslant j$，X_{kn}亦缺失)，有多种方法可以选用，如针对连续型变量可选用预测均数匹配(predictive mean matching，PMM)法、趋势得分(propensity score，PS)法；针对离散型变量可选用判别分析和logistic回归；对于任意缺失模式(对数据矩阵进行任意的变换都无法呈现上述现象)，可采用马尔科夫链蒙特卡罗(Markov chain Monte Carlo，MCMC)方法[16-18]。

3.4 参数似然方法

EM算法的最大似然法(maximum likelihood methods using the EM algorithm)可以用于估计一个模型的未知参数。对于完全数据，参数(如均数和线性回归系数)很容易用最大似然法进行估计。当有缺失数据发生时，也可使用相同原理，根据观测数据的似然函数进行参数的估计。但是当有缺失值发生时，观测数据的似然函数很难最大化。例如，当观测数据是完全时，某变量均数的最大似然估计是该变量值的总和与样本量之比。当该变量的一些值缺失时，该变量值的总和是未知的，这样就不能对该变量均数进行估计[19]。这时可使用EM算法进行迭代运算，对缺失值进行填充和参数估计，其原理和方法是EM算法分二步迭代估计。1)预测步：给定未知参数的某个估计值，预测充分统计量中有关缺失数据的部分。2)估计步：利用预测步得到的充分统计量计算参数最大似然估计的校正值，重复以上两步，直到前后两次计算结果达到规定的收敛标准。这种方法应用的条件是数据为多元正态分布和数据缺失的机制是可忽略的。最大似然法的主要优点是根据观测数据的分布对缺失值进行填充，其结果的估计比较精确和有效。但这种方法的有效性依赖于明确的模型假定，如纵向或结构方程模型，并且通常只能借助特定的软件平台才能实现，如AMOS(analysis of moment structure)和LISREL(linear structural relations)等，由于该方法不能计算标准误而限制了其使用[20]。

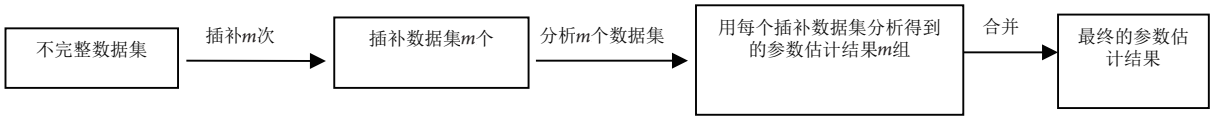

图2 多重插补的步骤及其统计推断原理。

Figure 2 Phases of multiple imputation and its theory of statistical inference.

4 实例分析

在上述的几种方法中，前面几种传统的缺失值处理方法较为简单，易于理解，并且在软件(如SPSS)中可通过勾选相应的条目来轻松实现。而多重插补法的原理相对复杂，作为近年来备受瞩目的一种缺失值处理方法，它的实用性和有效性相对较好[21]，与参数似然方法相比，其可操作性更强，因此，这里着重针对多重插补法的应用方法予以举例说明。

4.1 资料来源

本研究资料节选自1999年11月至2000年5月对湖南省洞庭湖洪灾区7~15 岁儿童的创伤性应激障碍(posttraumatic stress disorder，PTSD)发生情况及其影响因素的流行病学调查，随机选取500例资料完整的研究对象，在R软件(2.15.2)中模拟出随机缺失20%的不完整数据。选取4个对儿童发生PTSD有关的变量作为自变量，以PTSD(二分类变量) 为因变量，进行logistic 回归分析，这些变量包括年龄(定量变量)、性别(二分类变量)、受灾程度(等级变量)、受灾经历(即曾被水围困等待救援与否，二分类变量)。

4.2 软件平台

多重插补可通过常规的统计软件实现[22-23]，如SPSS软件中“analyze”菜单下的“multiple imputation”模块、SAS软件中的MI和MIANALYZE两个过程、R软件中的MICE软件包等。本例采用SPSS17.0软件进行统计分析。

4.3 分析步骤

4.3.1 缺失机制判断

本例中通过计算机模拟缺失机制为MAR的情况下缺失量为20%的情况。而在实际运用中，可从专业角度出发，根据前文中的缺失机制的定义来进行判断。

4.3.2 缺失模式判断

对数据集进行任意的行列对换后，仍无法达到该矩阵中X_{ij}缺失，则对任意的$k \geqslant i$ 和$l \geqslant j$，X_{kl}亦缺失的情形，故该数据为任意缺失模式。因此宜选择MCMC方法的多重插补。

4.3.3 数据分析过程

分析的过程首先是用多重插补法对缺失值进行多重插补，插补次数一般可以选择5~10次(有研究[24]表明插补次数在这个范围内的插补效率较高)，本例中设定为默认值即插补5次。多重插补以后将产生5个插补数据集，然后可用logistic回归来对插补后的数据进行常规分析，软件会自动对每个插补后的数据集进行相同的分析，最后将这些分析结果进行合并从而产生最终的统计推断。

4.3.4 结果比较

分别用列表删除和多重插补法处理本例中的不完整数据集后，其统计分析结果见表1。令用不同缺失值处理方法计算出的回归系数为b_1，用完整数据集计算出的回归系数为b_2，则相对误差为$(b_1-b_2)/b_2\times100\%$。从表1中可见，与列表删除法相比，多重插补的处理效果更好，各回归系数更接近于完整数据集下的真值，其相对误差的绝对值几乎都比列表删除法要小，且相对误差的范围不超过50%。

表1 不同方法处理缺失数据后的回归系数及其相对误差

Table 1 Regression coefficient and its relative error with the use of different methods dealing with missing data

变量	不同处理方法下的回归系数 (相对误差 /%)		
	完整数据集 (n=500)*	列表删除 (n=163)*	多重插补 (n=2500)*
受灾程度	–0.27(—)	–0.21(–22.2)	–0.25(–7.4)
性别	0.31(—)	0.43(38.7)	0.29(–6.5)
年龄	–0.10(—)	–0.17(70.0)	–0.15(50.0)
受灾经历	2.76(—)	21.62(683.3)	2.50(–9.4)

*n表示纳入分析的有效例数。

5 研究前沿

20世纪90年代初至今属于缺失数据研究的方法完善期，其理论发展主要体现在对已有方法的改进和扩展、方法比较研究以及应用研究方面。例如，Tang等[25]分别提出了两种不同的运用似然函数的半参数方法来处理非随机缺失数据问题。Qin等[26]从充分利用辅助信息的角度出发，提出了一些借助辅助信息的双重稳健方法，使估计值更加稳定有效。Stekhoven等[27]将数据挖掘技术运用到缺失值插补中，提出了针对复杂类型数据的缺失森林算法。这些方法为缺失值处理提供了新的

思路，其理论价值和可推广性有待更多的应用研究和相关软件模块技术的发展来验证和支持。

综上所述，目前尚没有哪一种处理缺失数据的方法是绝对普遍适用的，正如上面所分析的，每种方法都有利有弊，且处在不断的发展与改进之中。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，必要时可以将两种或多种方法结合使用，以最合理、有效的方式来处理相应的缺失数据。

参考文献

1. Abraham WT, Russell DW, et al. Missing data：a review of current methods and applications in epidemiological research [J]. Curr Opin Psychiatry, 2004, 17(4): 315-321.
2. Karahalios A, Baglietto L, Carlin JB, et al. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures [J]. BMC Med Res Methodol, 2012, 12: 96.
3. Vandenbroucke JP, Von EE, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration [J]. PLoS Med, 2007, 4(10): 1628-1654.
4. Potthofff RF, Tudor GE, Pieper KS, et al. Can one assess whether missing data are missing at random in medical studies [J]. Stat Methods Med Res, 2006, 15(3): 213-234.
5. Little RJA, Rubin DB. Statistical analysis with missing data [M]. 2nd ed. New York: John Wiley & Sons, 2002: 1-10.
6. Collins LM , Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modem missing data procedures[J]. Psychol Methods, 2001, 6(4): 330-351.
7. 周艺彪, 姜庆五, 赵根明. 调查研究中数据缺失的机制及处理方法[J]. 中国卫生统计, 2005, 22(5): 318-321.
   ZHOU Yibiao, JIANG Qingwu, ZHAO Genming. Mechanisms and treatment of missing data in survey research [J]. Chinese Journal of Health Statistics, 2005, 22(5): 318-321.
8. Fay RE. Alternative paradigms for the analysis of imputed survey data [J]. J Am Stat Assoc, 1996, 91(434): 490-498.
9. Pigott TD. A review of methods for missing data [J]. Educ Res Eval, 7(4): 353-383.
10. Streiner DL, Finkle WD. The case of the missing data: Methods of dealing with dropouts and other research vagaries [J]. Can J Psychiatry, 2002, 47(1): 68-75.
11. Fielding S, Fayers PM, Loge JH. Methods for handling missing data in palliative care research [J]. Palliat Med, 2006, 20(8): 791-798.
12. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses [J]. Am J Epidemiology, 1995, 142(12): 1255-1264.
13. 金勇进, 朱琳. 不同差补方法的比较[J]. 数理统计与管理, 2000, 19(2): 50-54.
   JIN Yongjin, ZHU Lin. Comparison of imputation methods [J]. Application of Statistics and Management, 2000, 19(2): 50-54.
14. Plaia A, Bondi AL. Single imputation method of missing values in environmental pollution data sets[J]. Atmos Environ, 2006, 40(38): 7316-7330.
15. Harel O. Inferences on missing information under multiple imputation and two-stage multiple imputation[J]. Stat Methodol, 2007, 4(1): 75-89.
16. 曹阳, 谢万军, 张罗漫. 多重填补的方法及其统计推断原理[J]. 中国医院统计, 2003,10(2): 77-81.
   CAO Yang, XIE Wanjun, ZHANG Luoman. Methods of multiple imputation and related inference theory [J]. Chinese Journal of Hospital Statistics, 2003,10(2): 77-81.
17. Buuren SV, Oudshoom KG. MICE: Multivariate imputation by chained equations in R [J]. J Stat Softw, 2010, 45(3): l-67.
18. Raghunathan TE, Lepkowski JM, Hoewyk JV, et al. A Multivariate technique for multiply imputing missing values using a sequence of regression models[J]. Surv Methodol, 2001, 27(1): 85-95.
19. Fulufhdo VN, Shakir M, Tshilidzi M. Missing data: a comparison of neural network and expectation maximization techniques[J]. Curr Sci, 2007, 93(11): 1514-1521.
20. Enders CK. Applied Missing Data Analysis [M]. New York: The Guilford Press, 2010: 103-112.
21. Cummings P. Missing Data and Multiple Imputation [J]. JAMA Pediatr, 2013, 167(7): 656-661.
22. Azur MJ, Stuart EA, Frangakis C, et al. Multiple Imputation by Chained Equations: What is it and how does it work [J]. Int J Methods Psychiatr Res, 2011, 20(1): 40-49.
23. Yucel RM. State of the multiple imputation software [J]. J Stat Softw, 2011, 45(1): 1-7.
24. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice [J]. Stat Med, 2011, 30(4): 377-399.
25. Tang G, Little RJA, Raghunathan TE. Analysis of multivariate missing data with nonignorable nonresponse [J]. Biometrika, 2003, 90(4): 747-764.
26. Qin J, Shao J, Zhang B. Efficient and doubly robust imputation for covariate-dependent missing responses [J]. J Am Stat Assoc, 2008, 103(482): 797-810.
27. Stekhoven DJ, Buhlmann P. MissForest-non-parametric missing value imputation for mixed-type data [J]. Bioinformatics, 2012, 28(1): 112-118.

(本文编辑　彭敏宁)

本文引用： 沈琳, 陈千红, 谭红专. 缺失数据的识别与处理[J]. 中南大学学报:医学版, 2013, 38(12): 1289-1294. DOI:10.3969/j.issn.1672-7347.2013.12.014

Cite this article as: SHEN Lin, CHEN Qianhong, TAN Hongzhuan. Identification and treatment of missing data[J]. Journal of Central South University. Medical Science, 2013, 38(12): 1289-1294. DOI:10.3969/j.issn.1672-7347.2013.12.014