综述

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.

Keywords: missing data　missingness mechanism　survey and research

通讯作者: 谭红专，Email: tanhz99@qq.com

作者简介: 沈琳，硕士研究生，主要从事流行病与卫生统计学研究。
作者Email: tanhz99@qq.com

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): I-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.

本刊中的类似文章