

A Worst-Case Bound Analysis for Nonignorable Missing Data

Masayuki Henmi

The Institute of Statistical Mathematics, Tokyo
henmi@ism.ac.jp

Abstract

In order to make statistical inference with missing data, we usually need to make some assumption on the missing-data mechanism. If the data can be assumed to be missing completely at random (MCAR), then we can estimate a parameter of interest in a statistical model specified for a targeted population by ignoring missing data. If not, however, the parameter of interest cannot be identified unless some other assumption is made on the missing-data mechanism. A crucial difficulty in this problem is that such an assumption is untestable from observed data and often strong for the parameter to be identifiable. We may have a misleading result unless the assumption is strongly supported from background information other than data. In most applications, however, such information (if exists) is not necessarily sufficient to make the parameter of interest identifiable.

The purpose of this research is to propose a method of calculating the bounds for statistical quantities such as bias of estimates, confidence intervals and P-values under some (weak) assumption on the missing-data mechanism for statistical inference with missing data. Our starting point is our recent paper [1] on publication bias in meta-analysis. Meta-analysis is a statistical analysis to strengthen some statistical evidence by combining results from several studies. However, study selection in meta-analysis is often not random and biased. This is called the problem of publication bias and can be thought of as a missing-data problem in the sense that some studies are missing (unpublished). In [1], we first considered a class of all possible selection functions (conditional selection probabilities given the data) satisfying the assumption that smaller studies are more likely to be missing than larger studies, which is often the case in meta-analysis and is much weaker than usual assumptions. Then we derived the bounds for confidence intervals and P-values within the above class of selection functions, and proposed the use of these bounds for a sensitivity analysis. (Note that once a selection function is given, the parameter of interest is identified and these statistical quantities can be calculated.) We aim to develop our method by extending this work to more general missing-data problems.

Manski and his colleagues, in a series of their papers ([2] and its references for example), developed a method of calculating the bound for a (nonidentifiable) parameter of interest itself under no or few assumptions on the missing-data mechanism. Although the target for which the bound is calculated is different from ours, it seems that their work is closely related to our work. Another difference is that their approach is considered in a nonparametric setting, whereas we mainly consider parametric or semiparametric models for complete data.

In the poster presentation, we first introduce our previous work [1] on publication bias, and then propose some (tentative) ideas for more general missing-data problems. As well as Manski's bound, we would like to discuss possible relationships with some other concepts of imprecise probability.

Keywords. nonignorable missing data, worst-case bounds, confidence intervals, P-values, selection functions

References

- [1] M. Henmi, J. B. Copas and S. Eguchi. Confidence Intervals and P-values in Meta-Analysis with Publication Bias. *Biometrics*, 63:475–482, 2007.
- [2] J. L. Horowitz and C. F. Manski. Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics*, 132:445–459, 2006.