Statistical Science
2008, Vol. 23, No. 3, 321–324
DOI: 10.1214/08-STS244A
Main article DOI: 10.1214/07-STS244
© Institute of Mathematical Statistics, 2008

Comment: Quantifying the Fraction of Missing Information for Hypothesis Testing in Statistical and Genetic Studies

Tian Zheng and Shaw-Hwa Lo

INTRODUCTION

The authors suggest an interesting way to measure the fraction of missing information in the context of hypothesis testing. The measure seeks to quantify the impact of missing observations on the test between two hypotheses. The amount of impact can be useful information for applied research. An example is, in genetics, where multiple tests of the same sort are performed on different variables with different missing rates, and follow-up studies may be designed to resolve missing values in selected variables.

In this discussion, we offer our prospective views on the use of relative information in a follow-up study. For studies where the impact of missing observations varies greatly across different variables and where the investigators have the flexibility of designing studies that can have different efforts on variables, an optimal design may be derived using relative information measures to improve the cost-effectiveness of the follow-up.

Using the simple motivation example in their paper, we examine the estimation of relative information by $\mathcal{R}I_1$ and $\mathcal{R}I_0$ in terms of unbiasedness and variability, and discuss issues that require further research. Although the relative information measure developed in their paper estimates the mean impact of the missing data, the actual impact may be highly variable when the amount of information in the observed data is moderate or small, which makes the estimated mean relative information a less reliable prediction of the actual impact of missing observations. For this reason, we suggest a simple way to estimate the variability of relative information between complete data and observed data in the simple motivation example. Further investigation is required in incorporating these variability estimates into the optimal design of follow-up studies.

RELATIVE INFORMATION AND FOLLOW-UP STUDY DESIGNS

Missing values can occur for many reasons and can have different effects on a given test. Nicolae, Meng and Kong pointed out that the impact of missing values (in terms of *relative information*) on a test may not be as simple as the "face value" of n_0/n, where n_0 is the number of observed values and n is the number of individuals ($n - n_0$ is then the number of missing values). Therefore, a more accurate estimation of the information gain due to the resolution of missing values is important for the design of follow-up studies.

Given an existing data with n individuals (with missing values), if n_1 additional independent samples are collected (possibly with the same missing rate) to expand this data set, it is intuitive to assume that the ratio of information in the original data and the expanded data is approximately $n/(n + n_1)$. Now consider a test on the existing data with n individuals that has some missing values (say, n_0 observed values). The *relative information* is estimated to be 80%, meaning that if the data used for this test is "resolved" to become complete, the expected log likelihood ratio is about $1/80\% = 125\%$ of the observed log likelihood ratio. To achieve the same level of information by adding new independent observations, one would need to collect a sample of additional $n_1 = n \times 25\%$ individuals. In many situations, resolving missing values, if possible, turns out to be much cheaper than collecting data on additional samples. In Section 2 of the NMK paper, an example was given on genotyping ambiguity in genetic linkage analysis (meaning that the exact inheritance vectors needed for the lod score computation cannot always be derived given the genotypes observed on the individuals). Here, let Y_{ob} be current data with unambiguous genotypes. For a follow-up study, a researcher can decide between (1) increasing the density of genetic markers on the observed individuals to *resolve* the ambiguities and (2) increasing the sample size by genotyping more independent individuals on the same set of markers for the previously observed

Department of Statistics, Columbia University, New York, New York, USA (e-mail: tzheng@stat.columbia.edu; slo@stat.columbia.edu).