Reviewer's report

Title: How does correlation structure differ between real and fabricated data-sets?

Authors:
Dr. Noori Akhtar-Danesh (daneshn@mcmaster.ca)
Mahshid Dehghan-Kooshkghazi (mahshid@ccc.mcmaster.ca)

Version: 1 Date: 12 Jul 2003

Reviewer: Mohammad Hossein H Rahbar

Level of interest: A paper of limited interest

Advice on publication: Unable to decide on acceptance or rejection until the authors have responded to the compulsory revisions

Reviewer's Comments
on

How does correlation structure differ between real and fabricated data-sets?

The authors attempt to demonstrate, through two examples, a higher correlation between the variables in the fabricated data sets as compared to their respective real data sets. The third example is an attempt to investigate whether the same theory holds for association rather than correlation when one of the two variables is categorical. The problem might be of potential interest to situations where this type of data fabrications and fraud is common. The data is selected from two universities in different COUNTRIES. The participants WERE asked to make up data based on a summary statistics such as mean, standard deviation and range of the real data. It is not however clear what was the main purpose of having data sets from two different settings.

The method section could be strengthened. I suggest that the authors provide details about the simulation procedures. In Example 1 where the correlation between the two variables is positive it is not clear how these 2500 samples were drawn from the real data sets. Similarly, in Example 2 further explanations could be provided about the simulations samples. I also suggest that some of the materials should be reorganized between the method and the result sections. It is hard to assess the quality of the data. However, lack of association between gestational age and birth weight of the children from Shiraz Hospitals is surprising. Most literature that I am familiar with indicates association between these two variables.

A relatively high percentage of participants did not provide fabricated data sets. This is an important issue, which needs to be addressed in the manuscript in light of the difficulty in generalizing these results. I also believe that the results of this study are dependent on the setting and the mind set of the participants who have made these fabricated data sets. This raises a very important issue whether these results could be generalized to settings in North America or these findings only pertain to developing countries. The authors do not provide justification for their sample size in all three examples. I believe there should be some explanation about the sample size and its adequacy for achieving the objectives of this study.

The discussion section is very weak. The authors should provide a strong argument in finding the unexpected relationship they have found with the data in Example 3 from Canada. Does this mean this theory cannot be extended to the case of association between a continuous and a categorical
variable? Could this be attributed to differences in the type of data or the difference between the two countries where these two data sets are collected? Does this indicate potential differences between the mindset of people making fabricated data sets between the participants in the developing countries and developed countries?

The abstract and the main text could be further improved both in terms of language and the quality of the information presented. In general the manuscripts need some editing with specific attention in the usage of scientific language.

I hope the authors find these comments useful in revising their manuscript. I believe the authors must respond to the aforementioned comments and issues raised before I can make my final recommendation on the suitability of this manuscript.

**Competing interests:**

None declared.