

# On inference of partially correlated data

## Editorial

Statistical inferential methods in the fields of the social, behavioral, economic, biological, medical, epidemiologic, health, public health, and drug developmental sciences need has grown exponentially in the last few decades. Study designs in the aforementioned applied sciences give rise to correlated and partially correlated data due to missing responses. For instances correlated data arise when subjects are matched to controls because of confounding factors and there are missing values in either or both groups. Other situations arise when subjects are repeatedly measured over time as in repeated measures designs. One assumption to consider is that observations are missing completely at random (MCAR).<sup>1,2</sup> However, Akritas et al.<sup>3</sup> consider another missing value mechanism, missing at random (MAR). For quantitative responses, statistical methods, including linear and nonlinear models, are established for correlated data. However, for partially correlated data there are concerns which to be addressed due to the complexity of the analysis. In particular, for small sample sizes and when a normality assumption of the underlying populations is not valid.

As an example of partially correlated data for the MCAR design, consider the case where the researcher compares two different treatment regimens for eye redness or allergy and randomly assigns one treatment to each eye for each experimental subject. Some patients may drop out after the first treatment, while other patients may drop out before the first treatment and came back for the second treatment. In this situation, we may have two groups of patients: the first group of patients who received both treatments in each eye, and are considered as paired matched data; and the second group who received only one of the treatments in one of the eyes, and are considered as unmatched data. Moreover, additional examples for partially correlated data can be found in the literature.<sup>4-6</sup> Several authors have presented various tests considering the problem of estimating the difference of means of a bivariate normal distribution when some observations corresponding to both variables are missing. Under the assumption of bivariate normality and MCAR, Ekbohm<sup>7</sup> summarized five procedures for testing the equality of two means. Using Monte Carlo results Ekbohm<sup>7</sup> indicated that the two tests based on a modified maximum likelihood estimator are preferred: one due to Lin and Stivers<sup>8</sup> when the number of complete pairs is large, and the other proposed in Ekbohm's paper otherwise, provided the variances of the two responses do not differ substantially. When the correlation coefficient between the two responses is small, two other tests may be used: a test proposed by Ekbohm when the homoscedasticity assumption is not strongly violated, and otherwise a Welch-type statistic suggested by Lin and Stivers<sup>8</sup> (for further discussion, see Ekbohm<sup>7</sup>).

Alternatively, researchers tend to ignore some of the data – either the correlated or the uncorrelated data depending on the size of each subset. However, in case the missingness not completely at random (MCAR), Looney and Jones<sup>9</sup> argued that ignoring some of the correlated observations would bias the estimation of the variance of the difference in treatment means and would dramatically affect the performance of the statistical test in terms of controlling type I error rates and statistical power.<sup>10</sup> They propose a corrected z-test method to overcome the challenges created by ignoring some of the correlated observations. However, our preliminary investigation shows that the

Volume 2 Issue 1 - 2015

**Hani Samawi, Robert Vogel**

Department of Biostatistics, Georgia Southern University, USA

**Correspondence:** Hani Samawi, Department of Biostatistics, JPHCOPH, Georgia Southern University, Statesboro, GA 30460, USA, Tel 912-478-1345, Fax 912-478-5811, Email [hsamawi@georgiasouthern.edu](mailto:hsamawi@georgiasouthern.edu)

**Received:** January 22, 2015 | **Published:** January 26, 2015

method of Looney and Jones<sup>9</sup> pertains to large samples and is not the most powerful test procedure. Furthermore, Rempala & Looney<sup>11</sup> studied asymptotic properties of a two-sample randomized test for partially dependent data. They indicated that a linear combination of randomized t-tests is asymptotically valid and can be used for non-normal data. However, the large sample permutation tests are difficult to perform and only have some optimal asymptotic properties in the Gaussian family of distributions when the correlation between the paired observations is positive. Other researchers, such as Xu & Harra<sup>12</sup> and Konietzschke et al.<sup>13</sup> also discuss the problem for continuous variables including the normal distribution by using weighted statistics. However, the procedure suggested by Xu & Harra<sup>12</sup> is a functional smoothing to the Looney & Jones<sup>9</sup> procedure. As such, the Xu and Hara procedure is not a practical alternative for the non-statistician researcher. The procedure suggested by Konietzschke et al.<sup>13</sup> is a nonparametric procedure based on ranking.

Samawi & Vogel<sup>14</sup> presented weighted test procedure to combined the correlated and non-correlated data. The aforementioned methods cannot be used for non-normal and moderate, small sample size data and categorical data. Samawi & Vogel<sup>15</sup> introduced several weighted tests when the variables of interest are categorical. They showed that their test procedures compete with other tests in the literature. Moreover, there are several attempts to provide nonparametric test procedures under MCAR and MAR designs.<sup>1-3,16,17</sup> However, there is still a need for intensive investigation to develop more powerful nonparametric testing procedures for MCAR and MAR designs. Samawi et al.,<sup>18</sup> discussed and proposed some nonparametric testing procedures to handle data when partially correlated data is available without ignoring the cases with missing responses. They introduced more powerful testing procedure which combined all cases in the study. All the above suggested procedures will be of special importance in meta-analysis where partially correlated data is a concern when combining results of various studies.

## Acknowledgments

None.

## Conflicts of interest

The authors declare that there are no conflicts of interest.

## Funding

None.

## References

1. Brunner E, Puri ML. Non parametric methods in design and analysis of experiments. In: Ghosh S, Rao CR, editors. *Handbook of Statistics*. Elsevier, Amsterdam, North-Holland: Netherlands; 1996. 631–703 p.
2. Brunner E, Domhof S, Langer F. *Non parametric analysis of longitudinal data in factorial designs*. John Wiley & Sons, New York: USA; 2002.
3. Akritas MG, Kuha J, Osgood DW. A nonparametric approach to matched pairs with missing data. *Sociological Methods & Research*. 2002;30(3):425–454.
4. Dimery IW, Nishioka K, Grossie B. Polyamine metabolism in carcinoma of oral cavity compared with adjacent and normal oral mucosa. *Am J of Surg*. 1987;154(4):429–433.
5. Nurnberger J, Jimerson D, Allen JR, et al. Red cellouabain-sensitive Na<sup>+</sup>-K<sup>+</sup>-adenosine triphosphatase: a state marker in affective disorder inversely related to plasma cortisol. *Bol Psychiatry*. 1982;17(9):981–992.
6. Steere AC, Green J, Schoen RT, et al. Successful parenteral penicillin therapy of established Lyme arthritis. *New England Journal of Medicine*. 1985;312(14):869–874.
7. Ekbohm G. Comparing means in the paired case with missing data on one response. *Biometrika*. 1976;63(1):169–172.
8. Lin P, Stivers LE. On difference of means with incomplete data. *Biometrika*. 1974;61(2):325–334.
9. Looney SW, Jones PW. A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Stat Med*. 2003;22(9):1601–1610.
10. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. IA: Iowa State University Press, Ames: USA; 1980.
11. Rempala G, Looney S. Asymptotic properties of a two-sample randomized test for partially dependent data. *Journal of Statistical Planning and Inference*. 2006;136(1):68–89.
12. Xu J, Harrar SW. Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical Journal*. 2012;54(2):281–295.
13. Konietschke F, Harrar SW, Lange K, et al. Ranking procedures for matched pairs with missing values-Asymptotic theory and a small sample approximation. *Computational Statistics & Data Analysis*. 2012;56(5):1090–1102.
14. Samawi HM, Vogel RL. Notes on Two Sample Tests for Partially Correlated (Paired) Data. *Journal of Applied Statistics*. 2014;41(1):109–117.
15. Samawi HM, Vogel RL. Tests of Homogeneity for Partially Matched-Pairs Data. *Statistical Methodology*. 2011;8(3):304–313.
16. KyungAh IM. A modified signed rank test to account for missing in small samples with paired data. MS Thesis, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania: USA; 2002.
17. Tang X. New test statistic for comparing medians with incomplete paired data. MS Thesis, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania: USA; 2007.
18. Samawi HM, Yu L, Vogel RL. On Some Nonparametric Tests for Partially Correlated Data: Proposing a New Test. 2014.