Computational Statistics and Data Analysis 56 (2012) 1090–1102

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Ranking procedures for matched pairs with missing data — Asymptotic theory and a small sample approximation

F. Konietschke [a,*], S.W. Harrar [b], K. Lange [a], E. Brunner [a]

[a] Department of Medical Statistics, University of Göttingen, Humboldtallee 32, D-37073 Göttingen, Germany

[b] Department of Mathematical Sciences, University of Montana, Missoula, MT 59812-0864, USA

ARTICLE INFO

Article history:
Available online 15 April 2011

Keywords:
Behrens–Fisher problem
Rank test
Nonparametric hypothesis
Ordered categorical data
Ties
Repeated measures design
Missing values

ABSTRACT

Nonparametric methods for matched pairs with data missing completely at random are considered. It is not assumed that the observations are coming from distribution functions belonging to a certain parametric or semi-parametric family. In particular, the distributions can have different shapes under the null hypothesis. Hence, the so-called nonparametric Behrens–Fisher problem for matched pairs with missing data is considered. Moreover, a new approach for confidence intervals for nonparametric effects is presented. In particular, no restriction on the ratio of the number of complete and incomplete cases is required to derive the asymptotic results. Simulations show that for arbitrary settings of complete data and missing values, the resulting confidence intervals maintain the pre-assigned coverage probability quite accurately. Regarding the power, none of the proposed tests is uniformly superior to the other. A real data set illustrates the application.

1. Introduction

In many sociological, psychological or medical studies, the subjects are observed repeatedly under different conditions, which are called treatments in the terminology of experimental designs. Hence, the repeated measurements on each subject may be dependent and therefore, statistics must be derived which take the correlation of the data into account. The simplest repeated measures design occurs if the subjects are observed twice, which is called a matched-pairs design. Moreover, the special problem of *missing values* is an additional difficulty arising in studies with repeated measures. Missing values occur frequently in practice and therefore appropriate statistical procedures must be derived. The key question for analyses with missing data is that of under what circumstances, if any, do the analyses that we would perform lead to valid answers. The answers depend on the missing value mechanism, which is the probability that a set of values are missing given the values taken by the observed and missing observations. Data are said to be *missing completely at random* (MCAR) if the probability of an observation being missing does not depend on observed or unobserved measurements. If the missingness mechanism does not depend on the unobserved data, the observations are said to be *missing at random* (MAR). In practice, however, the missing value mechanism is rarely known. For a detailed overview of missing value mechanisms, we refer the reader to the textbook by Little and Rubin (1987). With respect to the missing value mechanisms, different test procedures including *complete pairs*, *all available cases* and *(multiple-)imputation* methods are extensively discussed in the literature (see, e.g., Akritas et al., 2002 and Akritas et al., 2006). A complete-pairs statistic can only be applied when data are MCAR but it does not use all the information from the data. All-available test procedures and imputation methods use all the observed data. Hereby, all-available approaches must be able to deal with strongly imbalanced sample sizes of the complete and incomplete

* Corresponding author. Tel.: +49 0551 394989; fax: +49 0551 39 4995.
E-mail address: fkoniet@gwdg.de (F. Konietschke).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.022

data, whereas multiple-imputation methods estimate the unobserved data, and variance estimation of the imputed data becomes an issue. Therefore, multiple-imputation methods can only be applied when sample sizes are very large (see, e.g., Akritas et al., 2002, Section 5). In this paper, we only consider test procedures based on all available data and we will derive inference methods for small sample sizes.

For the case of MCAR or that of MAR, the analysis of matched pairs with missing values using parametric or semi-parametric mixed models is well developed and is well described in several textbooks (Diggle et al., 1994; Lindsey, 1993; Verbeke and Molenberghs, 2000). All of these procedures are based on specific model assumptions, e.g. the existence of an expectation or homogeneous variances. In practice, model assumptions can rarely be verified. If the data do not reflect the assumptions, e.g. skewed distributions, outliers or small sample sizes, the statistical procedures may result in obtaining wrong conclusions. For the special case of dichotomous data, Lin et al. (2009) propose a Bayesian approach under the assumption of MCAR. For the analysis of general non-normal data, in particular discrete data or even ordered categorical data, rank-based nonparametric methods are preferred. Let $\mathbf{X}_k = (\Delta_{1k}X_{1k}, \Delta_{2k}X_{2k})$, $k = 1, \ldots, n$, denote paired observations with marginal distributions $X_{ik} \sim F_i$, $i = 1, 2$. Here, Δ_{ik} denotes an indicator with $\Delta_{ik} = 1, 0$ if X_{ik} is observed or non-observed, respectively. Further let $n_c = \sum_{k=1}^{n} \Delta_{1k}\Delta_{2k}$ denote the number of the complete cases and let $n_g = \sum_{k=1}^{n} \Delta_{gk}(1 - \Delta_{sk})$, $s \neq g$, denote the number of incomplete cases in sample g, $g = 1, 2$. Brunner and Puri (1996) discuss nonparametric ranking methods for the hypothesis $H_0^F : F_1 = F_2$ formulated in terms of the marginal distribution functions under the assumption of MCAR. Akritas et al. (2002) propose a new ranking approach for $H_0^F : F_1 = F_2$ under the assumption of MAR. They overcome the perception that ranking methods can only be applied when data are MCAR. Gao (2007) extends the nonparametric imputation approach by Akritas et al. (2002) to the two-factor mixed model with MCAR data. All of these procedures are based on the asymptotic normality of estimates of so-called relative marginal effects $p = P(X_{11} \leq X_{22})$ (continuous distributions) under the hypothesis H_0^F. For independent samples, p is estimated by the numerator of the Wilcoxon–Mann–Whitney test (Mann and Whitney, 1947). Browne (2010) emphasizes that in the case of (asymptotic) normality, p can help to express t-test results in terms of differences between individuals randomly chosen from the two populations, rather than in terms of differences in population means. In the case of ordinal data, p is also called the *ordinal effect size measure* (Ryu, 2008; Ryu and Agresti, 2008). All of these ranking procedures for testing the hypothesis H_0^F are not consistent with respect to alternatives of the form $H_1^F : F_1 \neq F_2$. Additionally, they are limited to testing problems and cannot be used to construct confidence intervals for the effect size measure. Therefore, test procedures which test the hypothesis $H_0^p : p = \frac{1}{2}$ in terms of the treatment effect are more appropriate. We focus on the interpretation of an effect size instead of a probability (p-value), following the ICH-E9 (ICH, 1998) recommendation for randomized clinical trials: *Estimates of treatment effects should be accompanied by confidence intervals, whenever possible* [...].

Brunner and Neumann (1984) propose a nonparametric approach for testing H_0^p for continuous MCAR data, which was generalized to arbitrary distributions by Brunner and Puri (1996). Simulation studies show that the test statistics tend to not maintain the pre-assigned type-I error level for small sample sizes or in unbalanced sample size allocations. In this work, we extend the idea of Konietschke and Brunner (2009), who propose weighted rank estimators for relative effects in factorial diagnostic trials with clustered data, to MCAR data and derive test procedures for H_0^F, H_0^p and confidence intervals for p under the assumption that the sums $n_c + n_g$ of the complete and the incomplete cases, for $g = 1, 2$, tend to infinity. This includes the particular cases in

Assumption 1.1.

1. $n_c \to \infty, n_1, n_2 \leq M < \infty$, or
2. $n_c \to \infty, n_1 \to \infty, n_2 \leq M < \infty$, or
3. $n_c \to \infty, n_1 \to \infty, n_2 \to \infty$, or
4. $n_c \leq N_c, n_1 \to \infty, n_2 \to \infty$,

which are the most practice-oriented patterns of sample sizes. All previous procedures are not valid for all the four cases listed in Assumption 1.1. Simulation studies show that the type-I error level is controlled quite accurately even for small sample sizes. The approach is purely nonparametric and allows one to formulate hypotheses in terms of the effect size measure p. Hence, the so-called *nonparametric Behrens–Fisher problem* (see, e.g., Brunner and Munzel, 2000 and Munzel, 1999b) for matched pairs with missing values will be considered. To the best of our knowledge, there is no other unified approach in the literature which can lead to confidence intervals for p in this general setup. A real data set illustrates the application of the new methods to a clinical trial involving ordered categorical data.

The paper is organized as follows. To motivate the ideas, a real data example is discussed in Section 2. In Section 3, the statistical model and the nonparametric effect size measure are introduced. Section 4 presents the new estimation approach for p in matched pairs with missing data, and the asymptotic normality of the estimator is shown. The estimation of the variance is explained in Section 5. New test procedures and confidence intervals under the assumption of MCAR are provided in Section 6. Simulation results will be presented in Section 7. The paper closes with the statistical analysis of the example in Section 8 and with a discussion of the methods proposed in Section 9. Technical proofs are given in the Appendix.

2. A motivating example: migraine sufferers

As an example, we consider the migraine trial published by Kostecki-Dillon et al. (1999), which was investigated in detail by Gao (2007).