Computational Statistics and Data Analysis 106 (2017) 65–76

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Simultaneous confidence intervals for comparisons of several multinomial samples☆

Frank Schaarschmidt[a,*], Daniel Gerhard[b], Charlotte Vogel[a]

[a] Leibniz Universitaet Hannover, Institute of Biostatistics, Herrenhaeuserstr. 2, 30419, Hannover, Germany

[b] University of Canterbury, School of Mathematics & Statistics, Private Bag 4800, Christchurch 8041, New Zealand

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 1 September 2016
Accepted 6 September 2016
Available online 10 September 2016

Keywords:
Multiple comparisons
Polytomous data
Dirichlet
Baseline logit
Coverage probability

ABSTRACT

Multinomial data occur if the major outcome of an experiment is the classification of experimental units into more than two mutually exclusive categories. In experiments with several treatment groups, one may then be interested in multiple comparisons between the treatments w.r.t several definitions of odds between the multinomial proportions. Asymptotic methods are described for constructing simultaneous confidence intervals for this inferential problem. Further, alternative methods based on sampling from Dirichlet posterior distributions with vague Dirichlet priors are described. Monte Carlo simulations are performed to compare these methods w.r.t. their frequentist simultaneous coverage probabilities for a wide range of sample sizes and multinomial proportions: The methods have comparable properties for large samples and no rare events involved. In small sample situations or when rare events are involved in the sense that the expected values in some cells of the contingency table are as low as 5 or 10, the method based on sampling from the Dirichlet posterior yields simultaneous coverage probabilities closest to the nominal confidence level. The methods are provided in an R-package and their application is illustrated for examples from developmental toxicology and differential blood counts.

1. Introduction

In a number of toxicological assays, the major outcome is the classification of each experimental unit into one of several categories. For example cells may be classified by visual assessment into several categories, where categories distinguish undamaged cells from different types of unusual characteristics or malformation. In clinical trials, the primary outcome may be the classification of individual patients into one of several categories reflecting disease severity, or clinical subtypes of a certain disease. Often, such categories are ordinal. In some applications, however, the order of categories can be ambiguous, that is, there is no clear order of severity among categories, or there may be no order at all, such that the categories are best described as a nominal variable.

In such trials, multiple treatments can be of interest, for example, multiple dose groups compared to a control group in toxicological assays or different therapeutic interventions in a clinical trial. Counting the number of individuals in each category and each treatment group gives rise to a 2-dimensional contingency table with several rows and columns. In the following, we will assume that the individual experimental units are assigned to treatment groups in a completely

☆ R-code to reproduce the examples and tables containing the simulation settings are available as supplementary material (see Appendix A).

* Correspondence to: Institut fuer Biostatistik, Herrenhaeuser Strasse 2, D-30419 Hannover, Germany. Fax: +49 511 762 4966.
E-mail address: schaarschmidt@biostat.uni-hannover.de (F. Schaarschmidt).

http://dx.doi.org/10.1016/j.csda.2016.09.004
0167-9473/

randomized design and that the sample size per treatment group is fixed by the experimental design (i.e., it is not the result of a random process as, for example, in an epidemiological exposure study). Under these conditions, we may assume that the counts of the different categories follow a multinomial distribution, independently in each treatment group.

Such contingency tables may be analyzed by applying the χ^2 tests for independence. A significant result of such a test will only produce the rather general interpretation: The probability to fall into some of the categories does significantly differ between some of the treatment groups. In practice, this will rarely be an exhaustive interpretation of the data. On the contrary, interest will be in a more detailed interpretation: Which categories increase or decrease in probability between which of the treatment groups, and if so, by what extent? If multiple comparisons between treatments with respect to several categories contribute to an overall hypothesis in the sense of a union intersection test (e.g. Casella and Berger, 2002), simultaneous confidence intervals are necessary for such interpretations. But, depending on the application, not all possible comparisons between categories are of interest and not all comparisons between treatments may play a role for the overall hypothesis. Rather, particular categories and treatments in a given assay or trial will give rise to a special set of comparisons which are of interest.

Methods for simultaneous confidence intervals (SCI) in multiple comparisons in contingency tables have been proposed by Gold (1963) and Goodman (1964). Gold (1963) describes an asymptotic Scheffe-type-approach for SCI suitable for all possible linear combinations of the proportions of several multinomial vectors by using a χ^2-quantile with degrees of freedom as in the corresponding global test. Such approaches are inherently two-sided, and the resulting intervals will be unnecessarily large if only a small subset of comparisons (out of all possible comparisons) is of interest. Goodman (1964) considers asymptotic methods for all possible comparisons as well as a selected subset of comparisons of multinomial proportions on the log-scale, assuming a single multinomial distribution for a contingency table with multiple rows and columns (as suitable, e.g. for epidemiological studies). He shows that Bonferroni-adjusted standard normal quantiles may yield narrower intervals than the Scheffe-type approach, when only few comparisons are of interest. Still this approach can be improved because the Bonferroni-adjustment ignores the correlation between the estimators (or the related test statistics).

Since then, numerous authors have considered simultaneous confidence intervals for proportions or pairwise comparisons of proportions in a single multinomial sample (e.g. Glaz and Sison, 1999; Piegorsch and Richwine, 2001; Hou et al., 2003; Wang, 2000; Chafai and Concordet, 2009). To our knowledge, simultaneous confidence intervals for the comparison of multiple odds between multiple multinomial samples have not been considered any further, although there is room for improvement compared to the seminal methods of Gold (1963) and Goodman (1964): The test statistics related to comparisons of multiple logits of multinomial proportions asymptotically follow a multivariate normal distribution (e.g., Agresti, 2013) and multiple multinomial samples can be considered as a special case for the application of multivariate generalized linear models (e.g. McCullagh and Nelder, 1989; Agresti, 2013). One can thus use quantiles of the multivariate normal distribution (Bretz et al., 2001) based on a sample estimate of the correlation structure to construct asymptotic simultaneous confidence intervals according to Hothorn et al. (2008). Such intervals will be narrower than Bonferroni-adjusted intervals in cases where only a limited subset of parameters with correlated estimators is of interest, because their quantiles account for the correlation structure that is ignored by Bonferroni or Scheffe-type approaches. Although all necessary computational methods are available, these methods have so far not been investigated with respect to their properties when applied with small sample sizes. Also, they suffer from infinite interval bounds, when single cells of the contingency table happen to contain zeros. Further improvements compared to these asymptotic methods might be achievable by sampling from the joint distribution of interest, for example from the posterior of a Bayesian model with a vague prior. Simultaneous confidence intervals can then be computed from such samples by percentile methods as described in Besag et al. (1995), or Mandel and Betensky (2008).

In the remaining part of the paper, we will first describe asymptotic simultaneous confidence intervals for user-defined sets of logits compared between several multinomial samples. Additionally, we will consider simultaneous percentile intervals applied on samples of Dirichlet posteriors with vague Dirichlet priors. The small sample performance of these methods will be compared in frequentist simulation studies. Finally, the methods are applied to two data sets.

2. Material and methods

2.1. Data structure and notation

We consider $g = 1, \ldots, G$ treatment groups in a randomized design, where n_g is the sample size in group g that has been fixed by the experimental design. As the experimental outcome, each individual or experimental unit in group g is categorized into exactly one of C possible categories, with index $c = 1, \ldots, C$. Furthermore we assume that due to the randomized assignment of treatments to individuals or experimental units, there is no further subgrouping of individuals or heterogeneity among individuals and also, that there are no secondary factors or covariates that affect the outcome. Thus we assume that the counted number of individuals of categories $c = 1, \ldots, C$ in group g, $\mathbf{x}_g = \left(x_{g1}, x_{g2}, \ldots, x_{gC}\right)$, follows a multinomial distribution

$$\left(x_{g1}, x_{g2}, \ldots, x_{gC}\right) \sim \mathit{multinomial}\left(n_g, \left(\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC}\right)\right),$$