Computational Statistics and Data Analysis 114 (2017) 146–157

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Testing homogeneity for multiple nonnegative distributions with excess zero observations

Chunlin Wang[a], Paul Marriott[b], Pengfei Li[b,*]

[a] Department of Statistics, School of Economics and Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, 361005, China
[b] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 27 April 2017
Accepted 28 April 2017
Available online 11 May 2017

Keywords:
Bootstrap
Density ratio model
Empirical likelihood
Multinomial logistic regression
Non-standard mixture model
Semicontinuous data

ABSTRACT

The question of testing the homogeneity of distributions is studied when there is an excess of zeros in the data. In this situation, the distribution of each sample is naturally characterized by a non-standard mixture of a singular distribution at zero and a positive component. To model the positive components, a semiparametric multiple-sample density ratio model is employed. Under this setup, a new empirical likelihood ratio (ELR) test for homogeneity is developed and a χ^2-type limiting distribution of the ELR is proved under the homogeneous null hypothesis. A nonparametric bootstrap procedure is proposed to calibrate the finite-sample distribution of the ELR. It is shown that this bootstrap procedure approximates the null distribution of the ELR test statistic under both the null and alternative hypotheses. Simulation studies show that the bootstrap ELR test has an accurate type I error, is robust to changes of underlying distributions, is competitive to, and sometimes more powerful than, several popular one- and two-part tests. A real data example is used to illustrate the advantage of the proposed test.

1. Introduction

Multiple groups of samples, with excess zero observations, are commonly encountered in many research fields, such as the life sciences (Taylor and Pollard, 2009; Wagner et al., 2011), epidemiology (Bascoul-Mollevi et al., 2005; Bedrick and Hossain, 2013), meteorology (Muralidharan and Kale, 2002), health economics (Tu and Zhou, 1999; Zhou and Tu, 1999), reliability (Lambert, 1992), and tobacco consumption (Johnson et al., 2015). For example, while monitoring rainfall distribution, Muralidharan and Kale (2002) presented a data set with daily rainfall measurements recorded over several years. There were often dry days which were recorded as having zero rainfall. Similarly, Zhou and Tu (1999) provided an example from the assessment of medical care expenditures, where the observations came from a control group and several intervention groups. In each group, a majority of inpatients had zero cost due to having no hospitalizations during the study.

Given multiple groups with excess zero observations, one fundamental problem is to test the homogeneity of their distributions (Lachenbruch, 1976, 2001, 2002; Tse et al., 2009; Bedrick and Hossain, 2013; Johnson et al., 2015). Specifically, suppose we have $m+1$ independent groups of samples distributed as follows:

$$x_{i1},\ldots,x_{in_i} \sim F_i(x) = \nu_i I(x=0) + (1-\nu_i) I(x>0) G_i(x), \quad i=0,\ldots,m, \tag{1}$$

[*] Corresponding author. Fax: +1 519 746 1875.
E-mail address: pengfei.li@uwaterloo.ca (P. Li).

http://dx.doi.org/10.1016/j.csda.2017.04.011
0167-9473/© 2017 Elsevier B.V. All rights reserved.

where n_i is the ith group's sample size, $I(\cdot)$ is an indicator function and the $G_i(\cdot)$'s are cumulative distribution functions with common support which may be continuous or discrete. In this paper, we concentrate on continuous distributions $G_i(\cdot)$'s whose support consists of all positive real numbers; but we propose, in Section 5, how the method can be applied to discrete distributions. For random samples with excess zero observations, the zero outcomes, in fact, contain valuable information and thus should not be simply discarded. The above formulation, (1), which is a non-standard mixture model of a point mass distribution at zero and a continuous positive component, is an intuitive way to account for the unique features of such data. Our interest is to test whether the $m+1$ mixture distributions F_i's are homogeneous, i.e., test whether $F_0 = \cdots = F_m$, or equivalently, $\nu_0 = \cdots = \nu_m$ and $G_0 = \cdots = G_m$.

In the literature, two-part tests have been widely used to compare groups of samples from the non-standard mixture structure (1). For example, Lachenbruch (2001, 2002) comprehensively studied two-part tests for two such populations. A two-part test is a two degrees of freedom test based on the sum of a test statistic for the equality of the proportions of zero counts and a conditional χ^2-test statistic for the positive part. The test for the latter part may be a nonparametric Wilcoxon–Mann–Whitney rank sum test or a two-sample t-test. If more than two populations are under consideration, we can replace these tests with a Kruskal–Wallis test or an ANOVA F-test, respectively. On the other hand, a parametric likelihood ratio test can also be used for the second part after assuming a parametric form on the nonzero data, such as a log-normal distribution or a gamma distribution. The two-part tests and their extensions have been successfully implemented in various applications; see for example, Bascoul-Mollevi et al. (2005), Taylor and Pollard (2009), and Wagner et al. (2011). Further ideas and comparisons of some existing one- and two-part procedures may be found in Delucchi and Bostrom (2004), Hallstrom (2010), and Zhang et al. (2010).

In numerical studies (see Section 3 and supplementary material, Appendix A), we show that the existing two-part tests are either inefficient when no parametric assumptions are made for the positive components or are not robust when the parametric models are assumed. In many applications, multiple populations may naturally share some common characteristics. It is therefore desirable to borrow efficiency across similar populations to improve testing power. At the same time, we also hope that a test is robust to deviations from the model assumptions. The semiparametric *density ratio model* (DRM) of Anderson (1979), which gained popularity after Qin and Zhang (1997), is a natural tool to use here. We propose to model the distributions of the positive components in (1), by the DRM to exploit information from all available samples. Let $dG_i(x)$ denote the density of $G_i(x)$ for $i = 0, \ldots, m$. The DRM postulates that

$$dG_i(x) = \exp\{\alpha_i + \boldsymbol{\beta}_i^\top \mathbf{q}(x)\} dG_0(x), \quad i = 0, \ldots, m, \tag{2}$$

for a non-trivial, pre-specified, basis function $\mathbf{q}(x)$ of dimension d, and unknown parameters α_i and $\boldsymbol{\beta}_i$. Clearly, $\alpha_0 = 0$ and $\boldsymbol{\beta}_0 = \mathbf{0}$ for an arbitrarily selected baseline group. Without specifying the baseline density $dG_0(x)$, we propose a test based on the DRM defined in (2) that does not depend on the form of $G_0(x)$ and hence is robust to the assumptions on $G_0(x)$.

The DRM is flexible and includes many parametric distribution families, such as the log-normal and gamma distributions, as special cases. In the literature, the DRM has been recognized as a powerful semiparametric tool in many statistical problems. For example, Qin and Zhang (1997) and Zhang (2002) showed the close relationship between the DRM and logistic regression models, and they further developed procedures to assess the goodness-of-fit of the logistic regression models based on case-control data. Qin (1999) and Zou et al. (2002) applied the DRM to a semiparametric mixture model. Fokianos et al. (2001) and Cai et al. (2017) considered hypothesis testing problems under the DRM without excess zero observations. Other recent publications include Chen and Liu (2013), who discussed quantile and quantile-function estimation under the DRM, and de Carvalho and Davison (2014), who considered modelling several multivariate extremal distributions by the DRM. To the best of our knowledge, the DRM has not been used in modelling multiple samples with excess zero observations.

Under this semiparametric setup, the empirical likelihood method (Owen, 2001) provides an effective platform for data analysis. We propose an empirical likelihood ratio (ELR) test for homogeneity under (1) and (2). We show that the proposed ELR test is also a two-part test: the first part tests the equality of zero proportions ν_i's and the second part tests homogeneity in the continuous components of the model. We show that the asymptotic null distribution of the ELR is of χ^2-type as the total sample size goes to infinity. We also explore using a nonparametric bootstrap procedure to calibrate the distributions of the proposed test statistic in finite-sample situations. This bootstrap procedure is shown to approximate the null distribution of the ELR test statistic for data generated from both the null and alternative hypotheses. Software implementing the bootstrap ELR test has been developed in the R language (R development core team, 2014) and is available in the online supplementary material (see Appendix A).

We note that developing the limiting distribution of the ELR is technically challenging. First, in the second part of the ELR, the number of observations, i.e., the number of positive observations in each group, is a random number, and thus this case differs from the work of Fokianos et al. (2001), Zhang (2002) and Cai et al. (2017). In particular, we have to deal with random sums of independent and identically distributed random variables. Second, the two parts of the ELR both have χ^2-type null limiting distributions. Hence we need to show their asymptotic independence so that their summation still has a χ^2-type null limiting distribution. We comment that investigating the asymptotic properties of the bootstrap procedure under both the null and alternative hypotheses is also technically challenging. Existing results may not be directly applied. We refer to Janssen and Pauls (2003) for more discussion.

We further note that under the null hypothesis of homogeneity, the DRM in (2) is automatically satisfied regardless of the choice of basis function $\mathbf{q}(x)$. This is because the null hypothesis corresponds to a reduced form of the DRM with all $\alpha_i = 0$ and $\boldsymbol{\beta}_i = \mathbf{0}$. This property is attractive because it ensures that the asymptotic size of the test can always be controlled at