Accepted Manuscript

Semiparametric inference on the means of multiple nonnegative distributions with excess zero observations

Chunlin Wang, Paul Marriott, Pengfei Li

PII: S0047-259X(16)30233-0
DOI: https://doi.org/10.1016/j.jmva.2018.02.010
Reference: YJMVA 4332

To appear in: *Journal of Multivariate Analysis*

Received date : 12 December 2016

Please cite this article as: C. Wang, P. Marriott, P. Li, Semiparametric inference on the means of multiple nonnegative distributions with excess zero observations, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.02.010

Semiparametric inference on the means of 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, Canada N2L 3G1

Abstract

A non-standard, but not uncommon, situation is to observe multiple samples of nonnegative data which have a high proportion of zeros. This is the so-called excess of zeros situation and this paper looks at the problem of making inferences about the means of the underlying distributions. Under the semiparametric setup, proposed by Wang et al. [31], we develop a unified inference framework, based on an empirical likelihood ratio (ELR) statistic, for making inferences on the means of multiple such distributions. A chi-square-type limiting distribution of this statistic is established under a general linear null hypothesis about the means. This result allows us to construct a new test for mean equality. Simulation results show favorable performance of the proposed ELR when compared with other existing methods for testing mean equality, especially when the correctly specified basis function in the density ratio model is the logarithm function. A real data set is analyzed to illustrate the advantages of the proposed method.

Keywords: Density ratio model, Empirical likelihood, Estimating equation, Multinomial logistic regression, Non-standard mixture model, Semi-continuous data
2010 MSC: 62H15, 62H10, 62E20

1. Introduction

Making reliable inferences on the means of multiple distributions is an important, and fundamental, topic in statistics. In this paper, we investigate this topic in the case of multiple skewed nonnegative distributions with an excess of zero values. Specifically, suppose we have $m + 1$ independent samples modeled as

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

where n_i is the sample size of the ith group, $\mathbf{1}$ is an indicator function and the G_is are cumulative distribution functions with common support which may be continuous or discrete. Under the formulation (1), the mean of each mixture distribution F_i, with $i \in \{0, \ldots, m\}$, can be expressed as

$$\mu_i = \int_0^\infty x dF_i(x) = (1 - \nu_i) \int_0^\infty x dG_i(x).$$

Our interest is to make inferences about $\mu_0, \ldots, \mu_m$. These include testing the null hypothesis $\mathcal{H}_0^* : \mu_0 = \cdots = \mu_m$, and constructing confidence intervals for $\mu_i - \mu_j$ and μ_i/μ_j, for $i \neq j$.

Multiple samples, with a non-standard mixture structure as shown in (1), frequently arise from many research areas; see [31] and also a recent special issue of the *Biometrical Journal* [2] and references therein. The mean of a population with excess zeros has been considered an important summary quantity. For example, in fishery and health economics studies, the population total often has a crucial scientific meaning. It can provide information for

*Corresponding author.
Email address: pengfei.li@uwaterloo.ca (Pengfei Li)