Computational Statistics and Data Analysis 104 (2016) 1–9

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A distribution-free *m*-out-of-*n* bootstrap approach to testing symmetry about an unknown median

Vyacheslav Lyubchich [a], Xingyu Wang [b], Andrew Heyes [a], Yulia R. Gel [c,*]

[a] *University of Maryland Center for Environmental Science, P.O. Box 38, 146 Williams St., Solomons, MD, 20688, USA*
[b] *Morneau Shepell Ltd., 411 Dunsmuir St., Vancouver, BC, Canada, V6B 1X4*
[c] *University of Texas at Dallas, 800 West Campbell Road, Richardson, TX, 75080, USA*

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 2 May 2016
Accepted 5 May 2016
Available online 16 June 2016

Keywords:
Testing for symmetry
Bootstrap
Resampling
Environmental monitoring
Mercury bioaccumulation
Skewness

ABSTRACT

Testing for symmetry about an unknown median is a ubiquitous problem in mathematical statistics, particularly, for nonparametric rank-based methods, and in a broad range of applied studies, from economics and business to biology, ecology, and medicine. However, the challenge still remains on how to derive a symmetry test with a good power performance and at the same time delivering a reliable Type I Error estimate. To overcome this problem, a new data-driven *m*-out-of-*n* bootstrap method is introduced for testing symmetry about an unknown median. The asymptotic properties of the developed *m*-out-of-*n* bootstrap tests are investigated along with their empirical finite-sample performance. The new tests are illustrated by applications to legal studies and wildlife monitoring.

1. Introduction

Testing for symmetry is an essential and ubiquitous problem in statistics, with applications ranging from ecology studies and environmental monitoring to implementation of economic policies and analysis of business development (for some most recent overview see Ng et al., 2013; Laska et al., 2014; Mukherjee et al., 2014; Henderson and Parmeter, 2015, and references therein). Our primary interest is in developing a robust and reliable test for evaluating whether the shape of unknown distribution of observed data is symmetric about the unknown median, i.e., a critical question in many nonparametric rank-based procedures (Lehmann and Romano, 2005; Hollander, 2005). In particular, consider a sample $X_1, \dots, X_n$ of independent and identically distributed observations from an absolutely continuous distribution F_x with unknown mean μ, median ν, and standard deviation σ. Let $f(x)$, $x \in R$ be the corresponding density function. Then, we are interested in testing that for any x in the range of $f(x)$

$$
\begin{aligned}
&H_0 : f(\nu - x) = f(\nu + x), \\
&H_a : f(\nu - x) \neq f(\nu + x).
\end{aligned} \tag{1}
$$

While tests for (1) have been considered by numerous authors (Gupta, 1967; Gastwirth, 1971; Doksum et al., 1977; Antille et al., 1982; Bhattacharya et al., 1982; Boos, 1982; Cabilio and Masaro, 1996; Mira, 1999, amongst others), the problem of deriving robust and computationally efficient methods for assessing symmetry about an unknown median

* Corresponding author.
E-mail addresses: lyubchic@umces.edu (V. Lyubchich), xwang@morneaushepell.com (X. Wang), heyes@umces.edu (A. Heyes), ygl@utdallas.edu (Y.R. Gel).

http://dx.doi.org/10.1016/j.csda.2016.05.004
0167-9473/

remains an active area of research and attracts a growing attention in statistics, econometrics, and other disciplines (Miao et al., 2006; Ekström and Jammalamadaka, 2007; Zheng and Gastwirth, 2010; Ghosh, 2011; Ekström and Jammalamadaka, 2012; Bacci and Bartolucci, 2014; Mukherjee et al., 2014). As noted by Ekström and Jammalamadaka (2007), the challenge is to derive a symmetry test with good power performance and at the same time being either asymptotically distribution-free or at least yielding a Type I Error that remains close to the desired level for probability distributions under the null hypothesis of symmetry (1). The latter issue is intrinsically related to the fact that many symmetry test statistics follow asymptotic distributions that depend on the unknown underlying distribution F_x (for example, on the density estimated at the median, $f(\nu)$); thus, distribution of the test statistic also remains unknown. A usual remedy is to circumvent the problem by approximating the unknown distribution of the test statistic under unknown F_x by a distribution of the same test statistic under assumption that $X_1, \ldots, X_n$ follow a standard normal distribution Φ. However, such approximation yields unsatisfactory performance when F_x further deviates from Φ (Cabilio and Masaro, 1996; Mira, 1999; Miao et al., 2006; Ghosh, 2011). Unfortunately, the classical bootstrap modifications (Zheng and Gastwirth, 2010) do not yield noticeable improvement either: while classical bootstrap correctly estimates the limiting distribution of a sample median, it fails to estimate the second term in the Edgeworth expansion.

As a remedy, in this paper we propose to employ a nonparametric m-out-of-n bootstrap and develop a new data-driven and distribution-free tests for symmetry about an unknown median. The key idea is to take advantage of the fact that m-out-of-n bootstrap of median gives the correct Edgeworth expansion up to the second term, and the m-bootstrap expansion agrees with the expansion of median in the non-bootstrap setting (Sakov and Bickel, 2000; Arcones, 2003; Bickel and Sakov, 2008). The new bootstrap tests are shown to deliver reliable estimates of Type I Error and competitive power across a variety of distributions, including both short- and heavy-tailed shapes. While we primarily focus on the Cabilio–Masaro statistic, the Mira test, and the robustified test of Miao et al. (2006), a similar m-out-of-n bootstrap approach can be readily applied to other symmetry tests such as the Ghosh test (Ghosh, 2011) and the Ekström–Jammalamadaka test (Ekström and Jammalamadaka, 2007). In addition, we show that measures of symmetry can be used as an environmental indicator of water pollution levels and shed light on association of various pollutants and fish sizes.

The paper is organized as follows. In Section 1 we present the nonparametric m-out-of-n bootstrap-based tests and discuss their asymptotic properties. In Section 2 we evaluate finite sample performance of the new bootstrap tests with Monte Carlo simulations. In Section 3 we illustrate applications of the new tests to legal studies and wildlife monitoring. The paper is concluded by discussion in Section 4.

2. The *m*-out-of-*n* bootstrap-based approach and its asymptotic properties

Denote the sample mean, median, standard deviation and mean absolute deviation from a sample median by $\overline{X}$, M, s_n and rs_n. To test hypotheses (1), we consider the following three variants of the symmetry tests about an unknown median, namely, the Cabilio–Masaro statistic (Cabilio and Masaro, 1996), based on the Yule coefficient of skewness (Yule, 1911; Cabilio and Masaro, 1996):

$$S_n = \frac{\overline{X} - M}{s_n},$$

the Mira test based on the Bonferroni measure of skewness (Mira, 1999):

$$\gamma_n = 2(\overline{X} - M),$$

and the robustified test of Miao et al. (2006) based on the Groeneveld–Meeden measure of skewness (Groeneveld and Meeden, 1984):

$$T_n = \frac{\overline{X} - M}{rs_n}.$$

Under H_0 of symmetry, i.e., $f(\nu - x) = f(\nu + x)$, all three test statistics, S_n, γ_n, and T_n, are shown to be asymptotically normally distributed (Cabilio and Masaro, 1996; Mira, 1999; Miao et al., 2006)

$$\begin{aligned} \sqrt{n}S_n &\xrightarrow{d} N\left(0, \frac{1}{\sigma^2}\Gamma\right), \\ \sqrt{n}\gamma_n &\xrightarrow{d} N\left(0, 4\Gamma\right), \\ \sqrt{n}T_n &\xrightarrow{d} N\left(0, \frac{2}{\pi\tau^2}\Gamma\right), \end{aligned} \tag{2}$$

where

$$\Gamma = \sigma^2 + \frac{1}{4f^2(\nu)} - \frac{\tau}{f(\nu)} \quad \text{and} \quad \tau = \mu - 2\int_{-\infty}^{\nu} xf(x)dx.$$