Statistics and Probability Letters 121 (2017) 70–77

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Asymptotic tests for interval-valued means

Yan Sun

Department of Mathematics & Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322, United States

HIGHLIGHTS

- Asymptotic tests are developed for the means of interval-valued populations.
- The limiting null distributions are derived in analytical form.
- Large sample tests can be performed by numerical integration of Bessel function.

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form 26 August 2016
Accepted 11 October 2016
Available online 17 October 2016

Keywords:
Asymptotic test
Random interval
Limiting distribution

ABSTRACT

We develop asymptotic tests for the means of interval-valued population. The problem is formulated and studied in the framework of random compact convex sets. Under both one-sample and two-sample settings, we derive analytical forms of the probability density functions for the limiting null distributions. Large sample testing rules are given based on numerical integration of the Bessel function.

1. Introduction

Interval-valued data is gaining increasing popularity, arising from various circumstances such as lack of precision, grouping, and censoring. With the trend of big data recently, data aggregations are being extensively used to reduce the sample size, resulting in a large amount of interval-valued data. Precisely, interval-valued data refers to collections of observations in the format of intervals, as opposed to single numbers. Practical examples include measurement range, daily temperature range, [min, max] observation of a group of individuals, among many others. There has been a great deal of literature on statistical inferences with interval-valued data (e.g., Diamond, 1990; Gil et al., 2002; Jeon et al., 2014; Sun and Ralescu, 2015; Sun and Ralescu, 2015). This paper is concerned with hypothesis testing for the means of interval-valued data, in the formulation of random sets. That is, we view the observed intervals as realizations of random intervals, which are one-dimensional random sets. A solid probabilistic foundation for studying random sets has been provided by Kendall (1974) and Matheron (1975).

The problem was previously studied in Montenegro et al. (2008), where a weighted average of squared t statistics for the center and spread of the interval was proposed as the test statistic. However, having a complicated form, the null distribution had to be approximated by a bootstrap technique. In this paper, we propose test statistics for interval-valued means based on an L_2 distance in the space of intervals. The limiting null distributions are derived for the one-sample test, paired two-sample test, and two-sample test with unequal sample sizes. We give the corresponding probability density functions in closed form in terms of the Bessel function. Therefore, approximate large sample tests can be performed by numerical integration of the Bessel function.

E-mail address: yan.sun@usu.edu.

http://dx.doi.org/10.1016/j.spl.2016.10.013

The rest of the paper is organized as follows. Section 2 introduces some preliminaries of the random sets theory. The asymptotic results are presented in Section 3 for the one-sample test and in Section 4 for the two-sample test. We provide a numerical example in Section 5. Technical proofs are collected in the Appendix.

2. Preliminaries of random sets

Let $(\Omega, \mathscr{L}, P)$ be a probability space. Denote by $\mathcal{K}\left(\mathbb{R}^d\right)$ or $\mathcal{K}$ the collection of all non-empty compact subsets of $\mathbb{R}^d$. A random compact set is a Borel measurable function $A : (\Omega, \mathscr{L}) \to (\mathcal{K}, \mathscr{F})$, where $\mathscr{F}$ is the σ-algebra on $\mathcal{K}$ induced by the Hausdorff metric. If $A(\omega)$ is convex almost surely, then A is called a random compact convex set. The collection of all compact convex subsets of $\mathbb{R}^d$ is denoted by $\mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$ or $\mathcal{K}_{\mathcal{C}}$. The expected value of a random compact convex set A is defined by Aumann (1965) as

$$E(A) = \left\{Ef \,|\, f \in L^1(\Omega, \mathscr{L}, P), f(\omega) \in A(\omega)\ a.s.\right\}$$

where $f : \Omega \to \mathbb{R}^d$ as above is called a selection of A, and Ef denotes the classical expectation of a random vector. In particular, a measurable function $A : \Omega \to \mathcal{K}_{\mathcal{C}}(\mathbb{R})$ is called a random interval, and its expected value is given explicitly as

$$E(A) = [E \inf\{A\}, E \sup\{A\}].$$

For $A \in \mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$, define a function on the unit sphere S^{d-1} of $\mathbb{R}^d$ as

$$s_A(u) = \sup_{a \in A} \langle u, a \rangle, \quad \forall u \in S^{d-1}.$$

This function is called the support function of $A \in \mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$, and it plays a key role in establishing the theory for random compact convex sets. According to the embedding theorems (Rådström, 1952; Hörmander, 1954), $\mathcal{K}_{\mathcal{C}}$ can be embedded isometrically into the Banach space $C(S)$ of continuous functions on S^{d-1}. The embedding can be defined by the support function of $A \in \mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$. Therefore, letting $\mathcal{S}$ be the space of support functions of all non-empty compact convex subsets in $\mathbb{R}^d$, $\mathcal{S}$ is the image of $\mathcal{K}_{\mathcal{C}}$ into $C(S)$. In $\mathcal{S}$, an L_2 metric is given by the norm

$$\|s_A(u)\|_2 = \left[d \int_{S^{d-1}} |s_A(u)|^2 \mu(\mathrm{d}u)\right]^{\frac{1}{2}},$$

where μ is the normalized Lebesgue measure on S^{d-1}. Correspondingly, an L_2 distance in $\mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$ can be defined as

$$\begin{aligned}\delta(A, B) &:= \|s_A - s_B\|_2 \\ &= \left[d \int_{S^{d-1}} |s_A(u) - s_B(u)|^2 \mu(\mathrm{d}u)\right]^{\frac{1}{2}}, \quad \forall A, B \in \mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right).\end{aligned}$$

A more general L_2 distance – particularly for $\mathcal{K}_{\mathcal{C}}(\mathbb{R})$ – has been introduced in Körner and Näther (1998, 2001) via the support function as

$$D_K(A, B) = \left\{\sum_{S^0 \times S^0} [s_A(u) - s_B(u)]\,[s_A(v) - s_B(v)]\, K(u, v)\right\}^{\frac{1}{2}} \quad \forall A, B \in \mathcal{K}_{\mathcal{C}}(\mathbb{R}).$$

Here $K(\cdot, \cdot)$ is a symmetric positive definite kernel on S^0.

Denote a bounded closed interval by $X = [X^C - X^R, X^C + X^R]$ with center $X^C \in \mathbb{R}$ and radius $X^R \geq 0$. Alternatively, the interval is also denoted by $X = [X^L, X^U]$, where X^L and X^U are the lower and upper bounds satisfying $X^L \leq X^U$. The L_2 distance defined in (1) for two intervals $X, Y \in \mathcal{K}_{\mathcal{C}}(\mathbb{R})$ turns out to have the simple form

$$\delta(X, Y) = \left[\left(X^C - Y^C\right)^2 + \left(X^R - Y^R\right)^2\right]^{\frac{1}{2}}. \tag{1}$$

It is shown (Gil et al., 2002) that, with the constraints $K(1, 1) = K(-1, -1)$ and $K(1, -1) \geq 0$, D_K is equivalent to the so-called W-distance

$$d_W(X, Y) = \left\{\int_{[0,1]} [f_X(\lambda) - f_Y(\lambda)]^2\, dW(\lambda)\right\}^{\frac{1}{2}}, \tag{2}$$

where $f_X(\lambda) = \lambda X^U + (1 - \lambda)X^L, \forall \lambda \in [0, 1]$, and W is a non-degenerate measure on $[0, 1]$ symmetric about $1/2$. The apparent spirit of d_W lies in its distribution assumption for the points in the interval. On the other hand, it is also seen that

$$\begin{aligned}d_W(X, Y) &= \left\{\int_{[0,1]} \left[\left(X^C - Y^C\right) + (2\lambda - 1)\left(X^R - Y^R\right)\right]^2 dW(\lambda)\right\}^{\frac{1}{2}} \\ &= \left[\left(X^C - Y^C\right)^2 + \omega\left(X^R - Y^R\right)^2\right]^{\frac{1}{2}}, \end{aligned} \tag{3}$$