Journal of Multivariate Analysis 160 (2017) 117–133

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A new test of independence for bivariate observations

D. Bagkavos [a,*], P.N. Patil [b]

[a] Department of Mathematics and Applied Mathematics, University of Crete, Heraklion, 71500, Greece
[b] Department of Mathematics and Statistics, Mississippi State University, 410 Allen Hall, 175 President's Circle, MS, USA

ARTICLE INFO

Article history:
Received 11 September 2016
Available online 23 June 2017

AMS 2010 subject classifications:
62G10
62G20
60G42
60G44

Keywords:
Independence
Hypothesis test
Power
Quantiles

ABSTRACT

This research contributes a new methodological advance on bivariate independence hypothesis testing. It is based on the property that under independence, every quantile of Y given $X = x$ is constant. Apart from the asymptotic distributions of the test statistic under the null and alternative hypotheses, this work establishes their first order Edgeworth expansion. This is used to construct a bandwidth selection rule, designed to maximize power while the size is controlled by a given significance level. Finally, numerical evidence is given on the test's benefits against standard independence tests, frequently encountered in the literature.

1. Introduction

Independence of bivariate random vectors is a concept encountered quite frequently in statistics either as an assumption or for inferential purposes. Naturally many hypothesis tests have been proposed for its investigation in a wide range of contexts, see Tjøstheim [22] for an account of important contributions. Consistent independence tests in the general setting of mathematical statistics have been developed in the past by Hoeffding [13], Blum et al. [2], Feuerverger [5], Einmahl and McKeague [4], and Genest and Rémillard [9]. They all use as basis the necessary and sufficient condition that for a bivariate random vector with independent components, the Cramér–von Mises distance between the joint and the product of the marginal distributions is zero. The proposals of Rosenblatt [20] and Rosenblatt and Wahlen [21] were based on a density version of the same condition, weighted by an appropriate function.

The applicability of most of the above tests is still discussed in the literature. In Mudholkar and Wilding [16] Section 3, it is acknowledged that although the tests of Hoeffding [13] and Blum et al. [2] perform very well, use of their asymptotic distribution for finite sample sizes can lead to large discrepancies. Rosenblatt [20] and Feuerverger [5] report that density-based tests are less powerful compared to tests based on sample distribution functions. In contrast, the test of Genest and Rémillard [9], available in the R package `copula`, works perfectly well in finite samples. As all of the above proposals, however, it does not offer an inherent way to control the power and size of the test simultaneously.

The present research develops a test of independence which is feasible to implement even for finite samples and avoids the above difficulties. As a point of origin it uses the works of Chan [3] and Patil and Sengupta [19] who instead of the Cramér–von Mises distance used the idea that for a bivariate population (X, Y) independence is implied if every regression quantile of Y versus $X = x$ is constant. The test statistic results by repeated, for each $p \in (0, 1)$, sign tests for the constancy of

* Corresponding author.
E-mail address: dimitrios.bagkavos@gmail.com (D. Bagkavos).

http://dx.doi.org/10.1016/j.jmva.2017.06.004
0047-259X/

the regression quantile p in the neighborhood of every design point, adding the regression results all together and integrating over p. By construction, small test statistic values indicate independence between X and Y; on the contrary large positive values favor the alternative hypothesis of dependence. Patil and Bagkavos [18] compared its power against the tests of Blum et al. [2] and Hoeffding [13]. This preliminary research indicated that the test is more sensitive in detecting dependence in low correlation levels and through the bandwidth selection mechanism it can be configured to control power and size.

What is missing from the literature is a quantification of the test's asymptotic distribution as well as a bandwidth selection rule so as to use the procedure optimally in practice. The present research addresses both issues. It starts by providing the asymptotic distribution of the proposed test statistic under both the null and alternative hypotheses. These are interesting on their own right however their first order Edgeworth expansion with leading terms as functions of the bandwidth parameter are better suited for developing a bandwidth selection rule for finite samples. For this reason, these are established following the work of Gao and Gijbels [6], and used to develop a closed form bandwidth expression which maximizes power while at the same time keeps the size constant based on a given significance level. This last feature is also missing from the literature as existing independence tests generally do not keep size constant as power grows. Further, the obtained smoothing parameter is optimal irrespectively of whether we are under the null or alternative hypotheses.

The reasoning from which the test arises is given in Section 2, together with the specific form of the null and alternative hypotheses, an ideal (infeasible in practice) version of the test statistic and its data driven counterpart. Its asymptotic properties are addressed in Section 3. Bandwidth selection is treated in Section 5. Finally, performance of the proposed test in practice is given in Section 6.

It has to be noted that the test developed in Zheng [26], independently of Patil and Sengupta [19], can be viewed as related to the present test even though its construction is based on cumulative distribution functions. However Zheng [26] did not discuss bandwidth selection or practical performance and implementation of his test and thus the corresponding part of the present article together with its background theory is novel. Still, the present work can be viewed as a continuation of the work in Zheng [26] under the present construction scheme. Even though investigation of the equivalence of the two tests is beyond the scope of this article, Section 3 provides insights on the connection between the two tests.

2. Motivation and test statistic

Let $(X, Y) \in \mathbb{R} \times \mathbb{R}$ be a random pair, assumed to have a joint probability density function (p.d.f.) f with cumulative distribution function (c.d.f.) F. Denote with $F_Y(y \mid x)$ the marginal c.d.f. of Y conditional on $X = x$ and with $F_Y^{-1}(y \mid x)$ its inverse. The marginal (unconditional) c.d.f. of Y is denoted by F_Y and by F_Y^{-1} its inverse. Obviously under independence between X and Y, $F_Y(y \mid x) = F_Y(y)$ and $F_Y^{-1}(y \mid x) = F_Y^{-1}(y)$ for every $(x, y) \in \mathbb{R}_X \times \mathbb{R}_Y$ where $\mathbb{R}_X$, $\mathbb{R}_Y$ denote the support of X and Y, respectively.

The basis of the proposed test is that under independence, for every quantile $p \in (0, 1)$, we have $F_Y^{-1}(p \mid x) = c_p$, where c_p does not vary with x. That is

$$F_Y^{-1}(p \mid x) = F_Y^{-1}(p) = c_p. \tag{1}$$

For a fixed p, if one were to express this model in terms of observations Y_i it would mean $Y_i = m_p(x_i) + \varepsilon_i$ for all $i \in \{1, \ldots, n\}$, where $m_p(x_i)$ is the pth quantile of Y given $X = x_i$ and error ε_i is such that its pth quantile is zero; see also Basset and Koenker [1]. That is, for a fixed p,

$$F_Y^{-1}(p \mid x_i) = m_p(x_i) + F_\varepsilon^{-1}(p) = c(p, x_i), \tag{2}$$

and when Y and X are independent, $c(p, x_i) = c_p = c_p + F_\varepsilon^{-1}(p)$ which does not vary with x. For brevity henceforth we will refer to pth regression quantile simply as $m(x)$ instead of $m_p(x)$.

Now, let $\psi_p(u) = \text{sign}(u) + 2p - 1$ and let (X_1, Y_1) and (X_2, Y_2) be two independent random vectors with common p.d.f. f. Then observe that for every fixed $p \in (0, 1)$, under independence,

$$g_p(x) = \mathrm{E}\big[\psi_p\big\{Y_1 - F_Y^{-1}(p)\big\} \mid X_1 = x\big] = 0 \tag{3}$$

and thus for every $p \in (0, 1)$ with $K_h(X_1, X_2) = h^{-1}K\{(X_1 - X_2)/h\}$, where K is a kernel function, we have

$$J(p) = \mathrm{E}\big[K_h(X_1, X_2)\psi_p\big\{Y_2 - F_Y^{-1}(p)\big\}g_p(X_1)\big] = 0$$

and consequently

$$J = \int_0^1 J(p)dp = 0. \tag{4}$$

On the contrary, under dependence between X and Y we have

$$g_p(X_1) = \mathrm{E}\big[\psi_p\big\{Y_1 - F_Y^{-1}(p)\big\} \mid X_1\big] \neq 0 \text{ a.s.}$$