Statistics and Probability Letters 139 (2018) 61–66

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Testing equivalence to families of multinomial distributions with application to the independence model

Vladimir Ostrovski

Talanx Asset Management, Charles-de-Gaulle-Platz 1, 50679 Cologne, Germany

ARTICLE INFO

Article history:
Received 22 September 2017
Received in revised form 21 January 2018
Accepted 20 March 2018
Available online 30 March 2018

Keywords:
Equivalence
Testing
Multinomial
Collapsible
Independence
Contingency

ABSTRACT

We introduce tests for equivalence to families of multinomial distributions. The finite sample performance of the tests is improved by bootstrapping. We apply the tests to the independence model of two-way contingency tables and study finite sample performance by simulation. We apply tests to real data sets.

1. Introduction

Equivalence testing has received an increasing attention in applied statistics over the last decade. A comprehensive overview of equivalence testing can be found in Wellek (2010). The equivalence testing has also a general value in a broader context to test the agreement between theory and observations. Two probability distributions P and Q are considered equivalent with respect to distance d if $d(P, Q) < \varepsilon$ for some sufficiently small positive ε, such that the distances smaller than ε are of little practical significance.

There are few tests for equivalence to a fully specified multinomial distribution, see Wellek (2010), Frey (2009) and Ostrovski (2017). We consider a more general case of showing equivalence to a family of multinomial distributions. This testing problem arises in many applications if some specific model for the data generating process is assumed. The testing for approximate collapsibility of two-way contingency tables is a common example where the equivalence testing can be applied. Multinomial distributions with k categories correspond to the probability vectors from the simplex $\mathcal{S}_k \subset \mathbb{R}^k$. Let $p \in S_k$ denote a probability vector, where p_i is the probability of ith category. Let $\mathcal{M} \subset S_k$ be a family of multinomial distributions. The distance between p and $\mathcal{M}$ is defined by

$$d(p, \mathcal{M}) = \inf_{q \in \mathcal{M}} d(p, q), \tag{1}$$

where the infimum becomes a minimum if $\mathcal{M}$ is compact. The distance $d_{\min}$ can be calculated numerically using conventional optimization techniques. The generalized equivalence test problem is given by

$$H_0 = \{d(p, \mathcal{M}) \geq \varepsilon\} \text{ against } H_1 = \{d(p, \mathcal{M}) < \varepsilon\}, \tag{2}$$

where $\varepsilon > 0$ is a tolerance parameter and p ranges over the simplex S_k.

E-mail address: vladimir_ostrovski@web.de.

https://doi.org/10.1016/j.spl.2018.03.014
0167-7152/

To our best knowledge no published work investigates equivalence testing of (2). However there are few articles on the goodness-of-fit tests with a tolerance region. Hodges and Lehmann (1954) consider the weighted Euclidean distance to build a tolerance zone around the null hypothesis in a number of testing problems. Then they apply the usual chi-square test for composite hypothesis. Rudas et al. (1994) present a framework based on distribution mixtures for evaluating goodness-of-fit of contingency tables. This approach is extended to the Kullback–Leibler distance in Liu and Lindsay (2009), where the multinomial likelihood ratio test for composite hypothesis is applied. Unfortunately the chi-square tests as well as likelihood ratio tests are not suitable for the equivalence test problem (2) because the asymptotic distributions of the test statistics have a singular point at zero with probability 0.5, see Hodges and Lehmann (1954) for the chi-square statistic and Rudas et al. (1994) for the likelihood ratio statistic.

We take a different approach using a normalized estimator of the distance $d(p, \mathcal{M})$ as the tests statistic. We observe the vector $p_n = (p_{n1}, \ldots, p_{nk})$ of relative frequencies from n independent realizations of a random variable which is distributed according to p. The vector p_n can be used as a plug-in estimator of the probability vector p. Thus we put p_n in expression (1) and obtain the test statistic $T_n = \sqrt{n}\,(d(p_n, \mathcal{M}) - \varepsilon)$ for the generalized test problem (2). An equivalence test rejects H_0 if T_n is smaller than a critical value, which can be calculated asymptotically or by means of the parametric bootstrap. In the next section we derive the asymptotic distribution of T_n and show the local asymptotic optimality (LAN) of the proposed tests.

2. Asymptotic distribution and local asymptotic optimality

First we show that the function $p \mapsto d(p, \mathcal{M})$ is differentiable under mild regularity conditions.

Theorem 1. *Let p_0 be a fixed probability vector. Assume that there exists a continuous function $q_{\min} : \mathbb{R}^k \to \mathcal{M}$ on an open neighborhood U of p_0 with $d(p, \mathcal{M}) = d(p, q_{\min}(p))$. Let d be continuously differentiable on an open neighborhood of $(p_0, q_{\min}(p_0))$. Let $\mathring{d}(p, q)$ denote the partial derivative $\frac{\partial}{\partial p} d(p, q)$. Then the function $p \mapsto d(p, \mathcal{M})$ is differentiable at p_0 with the derivative $\mathring{d}(p_0, q_{\min}(p_0))$.*

The proof of Theorem 1 is given in the supplementary material. The derivative of $d(p, \mathcal{M})$ can be computed in two steps:
1. Calculate the value of $q_{\min}(p) \in \mathcal{M}$ numerically.
2. Evaluate the known partial derivative function $\mathring{d}(p_0, q_{\min}(p_0))$.

Remark 2. Euclidean distance is everywhere differentiable and therefore meets the requirements of Theorem 1. The total variation distance is not differentiable at some points because the absolute value is not differentiable at zero. However, there is a smooth version of the total variation distance, which is everywhere differentiable, see Ostrovski (2017) for details. Thus the smooth version of the total variation distance fulfills the requirements of Theorem 1.

Remark 3. The existence of a continuous minimizer $q_{\min}(p)$ on an open neighborhood of p_0 is a basic requirement for the numerical calculation of $d(p, \mathcal{M})$. However the existence of a global continuous minimizer is usually very difficult to show. Therefore, in the most practical cases we assume the existence of a continuous minimizer on an open neighborhood of the true distribution density. The assumption can be validated numerically using different starting points for the optimization.

It is a well-known fact (see Bishop et al. (1975), Theorem 14.3-4) that the normalized vector $\sqrt{n}\,(p_n - p)$ of relative frequencies converges weakly to a random variable Z, which is Gaussian with mean zero and covariance matrix $\Sigma(p) = D_p - pp^t$, where D_p is the square diagonal matrix, whose diagonal entries are $p_1, \ldots, p_k$.

Corollary 4. *Let p_0 be a boundary point of H_0. Under the assumptions of Theorem 1 the asymptotic distribution of the test statistic T_n under p_0 is Gaussian with mean zero and variance $\sigma(p_0) = \mathring{d}(p_0)\,\Sigma(p_0)\,\mathring{d}(p_0)^t$, where $\mathring{d}(p_0)$ is a shorthand notation for $\mathring{d}(p_0, q_{\min}(p_0))$.*

Proof. The assertion follows from Theorem 1 by application of the delta method, see van der Vaart (1998), p. 26, Theorem 3.1. □

Let l_α denote the lower α-quantile of the normal distribution and let p_0 be a boundary point of H_0. The equivalence tests based on T_n are asymptotically optimal as the following assertion states.

Corollary 5. *Let the assumptions of Theorem 1 hold. Let c_n be a critical value such that $c_n \to l_\alpha$ in probability. Let σ_n be a consistent estimator of $\sigma(p_0)$ such that $\sigma_n \to \sigma(p_0)$ in probability. Then the test that rejects H_0 if $T_n \le c_n\sigma_n$ is locally asymptotically most powerful.*

Proof. It follows from Theorem 1 and Ostrovski (2017), Proposition 3. □

The last component of the asymptotic equivalence test is an estimator of $\sigma(p_0)$, which is given in the next proposition.

Proposition 6. *Under the assumptions of Theorem 1 the estimator $\sigma(p_n)$ converges to $\sigma(p_0)$ a.s.*