Statistics and Probability Letters 129 (2017) 348–354

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Distribution free testing for conditional distributions given covariates

Estate Khmaladze

School of Mathematics and Statistics, Victoria University of Wellington, PO Box 600, Wellington, New Zealand

ARTICLE INFO

Article history:
Received 16 May 2017
Received in revised form 26 June 2017
Accepted 27 June 2017
Available online 8 July 2017

MSC 2010:
62E22
62F03
62F05
62F8600-01

Keywords:
Unitary operators
Empirical process
Kiefer process
Parametric models
Orthogonal projections

ABSTRACT

Given a sample $(\xi_i, X_i)_{i=1}^n$ of i.i.d. pairs, consider F_x and Q_x as two different models for conditional distribution of ξ_i given $X_i = x$. The paper shows how the empirical process for testing F_x can be transformed into a process with the same asymptotic behavior as the empirical process for testing Q_x thereby rendering the two testing problems equivalent.

1. Introduction

Below, by goodness of fit test we mean a statistical test with two properties: (a) under the null hypothesis its statistic is asymptotically distribution free (its limit distribution does not depend on the hypothetical family of distributions) and (b) it has some power against any contiguous (or "local") alternative, which approaches the hypothetical family from any given direction. People often express this latter property by saying that the test has "omnibus" character. We refer to Khmaladze (1993), Sec.3, for an earlier attempt to formulate the "goodness of fit problem".

Instead of finding separate goodness of fit statistics, it would be much more general and fruitful to construct asymptotically distribution free versions of empirical processes, the processes "themselves", if one may say so. If we are successful in this, then we will obtain the whole class of goodness of fit statistics — any appropriate functionals of an omnibus nature from these processes will lead to such statistics.

For empirical processes in finite-dimensional Euclidean spaces, based on continuous random variables, asymptotically distribution free transformations as suggested in Khmaladze (1981, 1993) are sufficiently known and we do not need to dwell upon it in this paper. However, what existed for random variables with discrete distributions was restricted to only one test — the K. Pearson's chi-square goodness of fit test, if we do not count various asymptotically equivalent forms of it.

The situation has changed only quite recently — in Khmaladze (2013) a method has been suggested to construct asymptotically distribution free empirical processes and, therefore, create a complete class of goodness of fit tests for discrete distributions. For readers' convenience, we recall the essence of this method in the remainder of this section.

E-mail address: Estate.Khmaladze@vuw.ac.nz.

http://dx.doi.org/10.1016/j.spl.2017.06.026
0167-7152/

Suppose $\xi_1, \dots, \xi_n$ are i.i.d. random variables with a discrete distribution $p = (p(k))_{k=1}^m$ with m finite. Denote by ν_{kn} the frequency of the outcome k in the sample of $(\xi_i)_{i=1}^n$ and consider the vector of "components" of the chi-square test statistics:

$$Y_n = (Y_{kn})_{k=1}^m, \quad Y_{kn} = \frac{\nu_{kn} - np(k)}{\sqrt{np(k)}},$$

so that the chi-square statistic itself is

$$\langle Y_n, Y_n \rangle = \sum_{k=1}^m Y_{kn}^2.$$

Although this statistic is asymptotically chi-square distributed no matter what the distribution p is, the limit distribution of the vector Y_n certainly does depend on p. Therefore, other statistics, such as, for example, an analogue of Kolmogorov–Smirnov statistic

$$\max_{1\le j\le m} \sum_{k=1}^j Y_{kn},$$

will have a limit distribution which very much depends on p. Let us append the index p to Y_n: $Y_n = Y_n^p$, and let $q = (q(k))_{k=1}^m$ be another m-dimensional distribution, so that we can consider now the pair Y_n^p and Y_n^q. What was suggested in Khmaladze (2013) is the unitary, and therefore one-to-one, transformation U, such that the transformed vector Z_n,

$$Z_n = UY_n^p,$$

and the vector Y_n^q will have the same limit distribution.

The transformation U here (see the definition in (1)) depends on p and q, or rather on the vectors $(\sqrt{p(k)})_{k=1}^m$ and $(\sqrt{q(k)})_{k=1}^m$, which we somewhat loosely will denote $\sqrt{p}$ and $\sqrt{q}$. Thus we have not one but a group of these unitary transformations and $U_{\sqrt{p},\sqrt{q}} U_{\sqrt{q},\sqrt{r}} = U_{\sqrt{p},\sqrt{r}}$. The consequence of this is that testing for any distribution p can be transformed into testing for a fixed distribution q. We will choose below $q = m^{-1}(1, \dots, 1)^T$. Any test statistic, based on Z_n, will have a limit distribution totally free from the hypothetical p.

The approach was extended in Khmaladze (2013) also to the case when p belonged to a parametric family of distributions with a finite-dimensional parameter. In Khmaladze (2016) it was carried from the components of chi-square statistic to the domain of empirical processes and testing parametric hypothesis in multidimensional spaces.

There are, however, many statistical problems when the distribution of ξ_i depends on covariates and this dependence has a parametric form. Moreover, each ξ_i should be allowed to have its own value of the covariates X_i. This makes ξ_i, $i = 1, \dots, n$, independent, but no longer identically distributed. In the present paper we want to extend the method, suggested in Khmaladze (2013), to this, much wider class of hypothetical distributions.

The problems with hypothetical distributions depending on covariates are technically close to the problems of regression. In the latter problems, or more precisely, in the problems of parametric regression, the expected value of "response" random variable ξ, given covariate X, $E(\xi|X = x) = m(x, \theta)$ is assumed to be a specific function of x depending on a finite-dimensional parameter θ.

Even closer in spirit are the problems of quantile regression, see Koenker (2005), when it is a quantile of the distribution of ξ of level α, which is supposed to be a specific parametric function of the covariate X.

In both of these situations, an asymptotically distribution free approach has been developed, see Koenker (2005) and Koenker and Xiao (2002), and Delgado and Stute (2008) and Khmaladze and Koul (2004), respectively. However, the general idea behind these papers is very different from what is suggested in the present note.

In the short Section 2 we show the main idea of the proposed method and stay with discrete distributions. In Sections 3 and 4 we present the general result — in particular, it incorporates families of hypothetical distributions, both discrete and continuous. In Section 5 we consider distributions depending on parameters and then, in Section 6 — the corresponding transformation of testing one parametric family into another.

2. Discrete distributions with covariates

From now on we assume that the random variables $\xi_1, \dots, \xi_n$ are, again, independent, and that there is a family of probability distributions $p_x = (p_x(k))_{k=1}^m$, where each probability is now a function of some variable x. Call this variable a *covariate*; it can be one- or multi-dimensional. Covariates are observable and for different ξ_i can take different values. In other words, the conditional distribution of ξ_i is one of p_x,

$$P(\xi_i = k|X_i = x) = p_x(k).$$

This latter statement we take as our null hypothesis. Alongside these distributions, consider another family of m-dimensional distributions $q_x(k)$. Denote, as in Section 1, by $\sqrt{p_x}$ the vector-function $(\sqrt{p_x(k)})_{k=1}^m$, and similarly use notation $\sqrt{q_x}$. Consider an operator in $\mathbb{R}^m$, defined as

$$U_{\sqrt{p_x},\sqrt{q_x}} = I - \frac{\langle \sqrt{p_x} - \sqrt{q_x}, \ \rangle}{1 - \langle \sqrt{p_x}, \sqrt{q_x} \rangle} \left(\sqrt{p_x} - \sqrt{q_x} \right), \tag{1}$$