Statistics and Probability Letters 99 (2015) 215–222

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Distribution free testing of goodness of fit in a one dimensional parameter space

Leigh A. Roberts

School of Economics and Finance, Victoria University, Wellington, New Zealand

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form 29 December 2014
Accepted 1 January 2015
Available online 9 January 2015

Keywords:
Brownian bridge
q-projected Brownian motion
Distribution free
Goodness of fit testing
Pareto
Weibull

ABSTRACT

We propose two versions of asymptotically distribution free empirical processes. When a composite null hypothesis contains a family of distributions indexed by a one dimensional parameter space, and when that single parameter is estimated by maximum likelihood, the resulting distribution free goodness of fit tests are simpler than tests applying the Khmaladze transformation. For the Pareto distribution, the process we advocate is especially simple. The theory is illustrated by fitting the Pareto distribution to threshold exceedances of stock returns, and the Weibull distribution to fibre strength data.

1. Introduction

The distribution free testing of goodness of fit of empirical data against a simple null hypothesis, under which the data is assumed to arise from a distribution function of the form $F_\theta(x)$, and with the parameter θ known, is a reasonably familiar process. The extension of the test to composite null hypotheses, however, for which $\theta \in \Theta$ for some set Θ, has proved elusive: replicating the simpler test using the estimated $\widehat{\theta}$ in place of the true θ is not satisfactory, in that the standard statistics, such as the Kolmogorov–Smirnov, Anderson–Darling and chi-squared statistics, have distributions which depend on the family $F_\theta(x)$, and generally on the specific values of $\widehat{\theta}$ and θ.

A standard way to effect distribution free goodness of fit tests for a composite null hypothesis is to apply the Khmaladze transformation (Khmaladze, 1981). While this transformation is finding increasing use in practice, the computations depend on the particular family $F_\theta(x)$ concerned, and may be complicated; and bootstrap procedures for more general estimators of θ may not be any simpler.

When there is but a single parameter to be estimated, and that parameter is estimated by maximum likelihood, then we provide two relatively simple distribution free processes for goodness of fit testing of a distribution. Two Brownian bridges are found, in different times, for which the standard statistics listed above may be applied as if one were testing a simple null hypothesis.

After some definitions and preliminaries, in Section 3 we mimic an approach taken for vector valued θ in Khmaladze (1981, p. 250). We consider an empirical process using a slightly unconventional centring (see (2)), and show that a certain function of this process is asymptotically a Brownian bridge in a particular time when the composite null hypothesis is valid.

E-mail address: leigh.roberts@vuw.ac.nz.

http://dx.doi.org/10.1016/j.spl.2015.01.002
0167-7152/

The final theoretical section is Section 4, in which we make use of Khmaladze's recent theory of the q-projected F-Brownian motion to 'rotate' the first Brownian bridge found to obtain a Brownian bridge in the simpler time $F_\theta(x)$.

The penultimate Section 5 considers applications, the first testing the validity of the customary usage of the Pareto distribution to model threshold exceedances of daily stock returns, from which the particularly simple form for the Pareto of the processes introduced will be apparent. The second application tests the goodness of fit of a Weibull distribution to fibre strength, utilising data in Smith and Naylor (1987), which is followed by a short conclusion.

2. The empirical process and a variant

For testing goodness of fit of data against a simple null hypothesis, viz. a distribution function of the form $F_\theta(x)$ with specified θ, one conventionally defines

$$v_n(x,\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbb{1}_{\{X_i\le x\}} - F_\theta(x)\right] = \sqrt{n}\left[F_n(x) - F_\theta(x)\right] \tag{1}$$

in which $X_1, X_2, \ldots, X_n$ is an independent and identically distributed sample from the distribution function $F_\theta(x)$; $\mathbb{1}$ is the indicator function; and $F_n(x)$ is the empirical distribution function. As $n \to \infty$, $v_n(x,\theta)$ tends to a Brownian bridge in time $F_\theta(x)$; and thence leading, by a transformation of the time, to the standard distribution free tests of goodness of fit of the simple null hypothesis.

In like vein, using a different centring we define

$$w_n(x,\theta) = \sqrt{n}\left[F_n(x) - \int_{-\infty}^{x}\frac{1-F_n(y)}{1-F_\theta(y)}dF_\theta(y)\right]. \tag{2}$$

The process $w_n(x,\theta)$ tends to a Brownian motion in time $F_\theta(x)$ as $n\to\infty$, and again leads to standard distribution free tests of goodness of fit of a simple null hypothesis. Although less common than the conventional centring in (1), the centring used in (2) may be more appealing for studies in survival analysis, in which there is a primary interest in the hazard rate or force of mortality.

2.1. Mathematical preliminaries

We assume that all distributions are absolutely continuous, with density $f(x,\theta) = F'(x,\theta)$. We set $\dot{F}(x,\theta) = (\partial/\partial\theta)F(x,\theta)$, and write interchangeably $F(x,\theta)$ and $F_\theta(x)$, with analogous notational conventions being used throughout the paper. In addition to standard regularity conditions, allowing us to differentiate under integral signs, etc., we assume that $\dot{F}_\theta(x)/\sqrt{1-F_\theta(x)} \to 0$ as $x\to\infty$, for each $\theta\in\Theta$, a condition easily verified for the Pareto and Weibull distributions.

Denoting the score function by $\psi_\theta(x)$, the Fisher information, denoted by $\Gamma(\theta)$, is given by

$$\Gamma(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right] = -E\dot{\psi}_\theta(x) = E\psi_\theta^2(x).$$

It is convenient to define $h_\theta(x) = \psi_\theta(x)/\sqrt{\Gamma(\theta)}$; and we further define

$$H_\theta(x) = \int_{-\infty}^{x} h_\theta(y)\,dF_\theta(y) = \frac{1}{\sqrt{\Gamma(\theta)}}\int_{-\infty}^{x}\dot{f}_\theta(y)\,dy = \frac{\dot{F}_\theta(x)}{\sqrt{\Gamma(\theta)}} \tag{3}$$

noting that $H_\theta(\infty) = 0$ and $\int_{-\infty}^{\infty} h_\theta^2(x)\,dF_\theta(x) = 1$. Finally (Cramer, 1946, Chapter 3)

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{1}{\sqrt{\Gamma(\theta)}}\int_{-\infty}^{\infty} h(x,\theta)v_n(dx,\theta) + o_P(1). \tag{4}$$

2.2. From the first centring to the second

A standard result linking a Brownian bridge $v_n(x,\theta)$ and a Brownian motion $w_n(x,\theta)$ in time $F_\theta(x)$ states that (Shorack and Wellner, 1986; Khmaladze, 1981)

$$dv_n(x,\theta) = -f_\theta(x)dx\int_{-\infty}^{x}\frac{w_n(dy,\theta)}{1-F_\theta(y)} + w_n(dx,\theta).$$

Substituting this expression into the integral in (4) yields

$$\int_{-\infty}^{\infty} h_\theta(x)\,dv_n(x,\theta) = -\int_{-\infty}^{\infty} dH_\theta(x)\int_{-\infty}^{x}\frac{w_n(dy,\theta)}{1-F_\theta(y)} + \int_{-\infty}^{\infty} h_\theta(x)\,w_n(dx,\theta). \tag{5}$$