Goodness-of-fit tests for the Weibull and Pareto distributions

Gennadi Martynov
Institute for information transmission problems
of the Russian Academy of Sciences (Kharkevich Institute)
Moscow, Russia
martynov@iitp.ru

Abstract

We will consider the goodness-of-fit tests for testing a form of the distribution function of the observed random variable. Let a distribution function belongs under hypothesis to a parametric family. Generally, the limit distributions of statistics, based on the empirical process, depend on the unknown parameters. It was stated in 1955 (see [8]) that this dependence is absent for the distribution family $\{G((x-\mu)/\sigma),\ \sigma>0\}$. This class includes the normal distribution. We present here the second class of the parametric distribution families with such a property. This is the family $\{R((x/\beta)^{\alpha})),\ \alpha>0,\ \beta>0,\ x\in\mathcal{X}\subset[0,\infty)\}$, where α and β are unknown parameters. This class includes the Pareto and Weibull distribution families. The exponential distribution family belongs to both classes. [1]

1 Introduction

Let $X^n=\{X_1,X_2,...,X_n\}$ be the sample from the r.v. with the distribution function $F(x),\ x\in R_1$. We will test the hypothesis

$$H_0:\ \ F(x)\in\mathcal{G}=\{G(x,\theta),\ \ \theta=(\theta_1,\theta_2,...\theta_k)^{\top}\in\Theta\subset R_k\},$$

where θ is an unknown vector of parameters. We will consider the Cramér-von Mises statistic

$$\omega_n^2=n\int_{-\infty}^{\infty}\psi^2(G(x,\theta_n))(F_n(x)-G(x,\theta_n))^2\,dG(x,\theta_n),$$

θ_n is an estimator of θ, $\psi(t)$ is the weight function, $F_n(x)$ is the empirical distribution function. The results below are applicable also to the Kolmogorov-Smirnov statistic

$$D_n=\sqrt{n}\sup_{-\infty<x<\infty}|\psi(G(x,\theta_n))(F_n(x)-G(x,\theta_n))|.$$

The exact methods for calculating the limit distribution are developed mostly for the Cramér-von Mises statistic (see [4], [8], [10], [11], [12], [14]).
Let θ_n be the likelihood maximum estimator of θ. Under the certain number of the regularity conditions and under H_0 limit distribution of the statistic ω_n^2 coincides (see [8]) with the distribution of the functional

$$\omega^2=\int_0^1\psi^2(t)\xi^2(t,\theta_0)dt$$

of the Gauss process $\psi(t)\xi(t,\theta_0)$ with $E\psi^2(t)\xi(t,\theta_0)=0$, and with the covariance function

$$K(t,\tau)=E(\psi(t)\xi(t,\theta_0)\psi(\tau)\xi(\tau,\theta_0))=\psi(t)\psi(\tau)(K_0(t,\tau)-q^{\top}(t,\theta_0)I^{-1}(\theta_0)q(\tau,\theta_0)),$$

where $K_0(t,\tau)=\min(t,\tau)-t\tau,\ \ t,\tau\in(0,\ 1)$, θ_0 is an unknown value of the parameter θ,

$$q^{\top}(t,\theta)=\left(\partial G(x,\theta)/\partial\theta_1,...,\partial G(x,\theta)/\partial\theta_k\right)|_{t=G(x,\theta)}\,,$$

[1] This research was partially supported by Russian foundation for fundamental research: 09-01-00740-a.

$I(\theta)$ is the Fisher information matrix,

$$I(\theta) = (E((\partial/\partial\theta_i)\log g(X,\theta)(\partial/\partial\theta_j)\log g(X,\theta)))_{1\le i,j\le k}, \quad g(x,\theta) = (\partial G(x,\theta)/\partial x).$$

The follow condition must be fulfilled:

$$\int_0^1 \psi^2(t)K(t,t)dt < \infty.$$

The limit distribution for D_n coincides with the distribution of

$$D = \sup_{0<t<1} |\psi(t)\xi(t,\theta_0)|,$$

but the conditions on $\psi(t)$ and another conditions are different from the conditions for ω^2. They was studied in [3], [13]. The distribution of ω^2 depends generally from θ_0 and the distribution family $\mathcal{G}$. Khmaladze [9] has proposed the method of empirical process transformation for eliminate such dependance. Khmaladze and Haywood [7] has applied this method to exponentiality testing by the Cramér-von Mises statistic.
We will use here the traditional approach consistsing in using of the statistic ω_n^2. It is well known (see for example [8], [10]]) that the empirical process does not depend on unknown parameter θ_0 for the family of the form

$$\mathcal{G} = \{G((x-m)/\sigma),\ -\infty < x < \infty,\ \sigma > 0\}.$$

Most known example of such family is the normal distribution family (see [5], [8)].
We will propose here another class of the distribution family with such property:

$$\mathcal{R} = \{R((x/\beta)^\alpha),\ \alpha > 0,\ \beta > 0, x \in \mathcal{X} \subset [0,\infty)\},$$

where $\mathcal{X}$ is the support of the distribution $R((x/\beta)^\alpha)$. Here $R(z)$ is a distribution function with the support $\mathcal{Z} \subset [0,\infty)$. Particular cases of such families are Weibull and Pareto distributions. The limit distributions of Cramér-von Mises and Kolmogorov-Smirnov statistics do not depend on the unknown parameters in both families. Additionally, the limit distribution for Pareto family coincide with analogous distribution for exponential family. The goodness-of-fit tests was discussed for the general Pareto distribution in many articles, particularly, in [1], [2], [6].

The ω^2-distribution can be calculated exactly with using the method of calculation the eigenvalues of the covariance operator. It was presented in [12]. This method is applicable for the power function $\psi(t) = t^\alpha\ \alpha > -1$. The method for the corresponding quadratic forms calculation was particularly presented in [10].

2 General result

Let $X^n = \{X_1, X_2, ..., X_n\}$ be the sample from the r.v. with a distribution function $F(x),\ x \in R_1$. We will test the hypothesis

$$H_0 :\ \ F(x) \in \mathcal{R} = \{R((x/\beta)^\alpha)),\ \alpha > 0,\ \beta > 0,\ x \in \mathcal{X} \subset [0,\infty)\},$$

where α and β are unknown parameters. The set of the alternative distributions contains all another distributions. Here $R(z)$ is the distribution function with the support $\mathcal{Z} \subset [0,\infty)$. We note the corresponding density function by $r(z)$. $\mathcal{R}$ is the family of Pareto distributions with $R(z) = 1 - 1/z,\ z > 1$ and $x > \beta$. The family $\mathcal{R}$ consists of Weibull distributions when $R(z) = 1 - \exp(-z),\ z > 0$, and $x > 0$. We will use the Cramér-von Mises and Kolmogorov-Smirnov tests. Both of them based on the empirical process $\xi_n(x) = \sqrt{n}(F_n(x) - R((x/\hat\beta)^{\hat\alpha})))$, where $\hat\alpha$ and $\hat\beta$ are the ML estimates of α and β. If the regularity conditions are fulfilled for them we can write the follow covariance function for the transformed to $(0,1)$ limit Gauss process $\xi(t)$:

$$K(t,\tau) = \min(t,\tau) - t\tau - (1/(B_{11}B_{22} - B_{12}^2))$$
$$\times(B_{22}s_1(t)s_1(\tau) - B_{12}(s_1(t)s_2(\tau) + s_2(t)s_1(\tau)) + B_{11}s_2(t)s_2(\tau)).$$

Here, $t, \tau \in (0,\ 1)$,

$$B_{11} = \int_{\mathcal{Z}} \left(\frac{z \log z\, r'(z)}{r(z)} + \log z + 1 \right)^2 r(z)dz,\ B_{22} = \int_{\mathcal{Z}} \left(\frac{z\, r'(z)}{r(z)} + 1 \right)^2 r(z)dz,$$

$$B_{12} = \int_{\mathcal{Z}} \left(\frac{z \log z\, r'(z)}{r(z)} + \log z + 1 \right) \left(\frac{z\, r'(z)}{r(z)} + 1 \right) r(z)dz$$

and

$$s_1(t) = r(R^{-1}(t))R^{-1}(t)\log(R^{-1}(t)),\ \ s_2(t) = r(R^{-1}(t))R^{-1}(t).$$

It follows from these formulae that the limit distributions of the considered statistics do not depend from the parameters α and β. Let β be known. Then the covariance function of the process $\xi(t)$is follow:

$$K(t,\tau) = \min(t,\tau) - t\tau - s_1(t)s_1(\tau)/B_{11}.$$

It does not depend of α in his turn. These results are used in the follow three sections.

3 Pareto distribution

We will consider the Pareto distribution in the form

$$F(x) = 1 - (x/\beta)^{-\alpha},\ x \geq \beta \geq 0,\ \alpha > 0.$$

For this distribution $R(z) = 1 - 1/z$ and $\mathcal{Z} = [\beta, \infty]$. It exists the supereffective unbiased estimate $\hat{\beta}$ of β.

We can transform the sample $X_1, ..., X_n$ to new sample $Y_1, ..., Y_n$, where $Y_i = X_i/\hat{\beta}$. The limit process $\psi(t)\xi(t)$ is equivalent to the process with $\beta = 1$. Hence the covariance function of $\xi(t)$ (without the pound function) is

$$K(t,\tau) = \min(t,\tau) - t\tau - (1-t)\log(1-t))(1-\tau)\log(1-\tau).$$

There $s_1(t) = -(1-t)\log(1-t),\quad B_{11} = 1$. This covariation function coincides with the corresponding covariance function for the exponential family

$$F(x) = 1 - \exp(-x/\beta),\ \beta \geq 0,\ x \geq 0.$$

It can be concluded that the limit distributions of the considered statistics for both families are the same one.

4 Weibull distribution

Consider the two parametric Weibull distribution family

$$F(x) = 1 - e^{-(x/\beta)^{-\alpha}},\ x \geq 0,\ \beta \geq 0,\ \alpha > 0.$$

We can note that $R(z) = 1 - e^{-z}$ and $\mathcal{Z} = [0, \infty]$. Maximum likelihood estimates $\hat{\beta}$ and $\hat{\alpha}$ for β and α can be found by numerical methods from the equation system

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^{\hat{\alpha}} \right)^{1/\hat{\alpha}},\quad \frac{n}{\hat{\alpha}} + \log\left(\frac{X_1 \cdot ... \cdot X_n}{\hat{\beta}^n} \right) - \sum_{i=1}^{n} \left(\frac{X_i}{\hat{\beta}} \right)^{\hat{\alpha}} \log\left(\frac{X_i}{\hat{\beta}} \right) = 0.$$

The covariance function of $\xi(t)$ in this example has the follow elements:

$$s_1(t) = -(1-t)\log(1-t)\log(-\log(1-t)),\quad s_2(t) = -(1-t)\log(1-t),$$

$$B_{11}(t) = \int_0^\infty ((1-z)\log\ z - 1)^2\ e^{-z}dz = (1-C)^2 + \frac{\pi^2}{6},$$

$$B_{12}(t) = \int_0^\infty ((1-z)\log\ z - 1)(1-z)\ e^{-z}dz = 1 - C,$$

$$B_{22}(t) = \int_0^\infty (1-z)^2\ e^{-z}dz = 1,\ \ B_{11}B_{22} - B_{12} = \pi^2/6,$$

where C is the Euler constant. It was found by simulation that the critical levels corresponding to the significance levels 0.1 and 0.05 are approximatively 0.10 and 0.12.

5 Power distribution on $[0,1]$

We consider now the distribution function

$$F(x) = \left(\frac{x-a}{b-a}\right)^{\alpha},\ x \in [a,b],\ b > a,\ \alpha > 0.$$

A supereffective estimates exist for the parameters a and b. Hence, we can transform the sample to the interval $[0,1]$ without changing the limit distribution of the statistics. It is sufficient to consider tests for the hypothetical distribution family $F(x) = x^{\alpha},\ x \in [0,1],\ \alpha > 0$, with $R(z) = z,\ \mathcal{Z} = [0,1]$. It's easy to derive the covariance function of the limit empirical process $\xi(t)$:

$$K(t,\tau) = \min(t,\tau) - t\tau - t\log t\,\tau\log\tau.$$

The limit distribution of the statistics $\omega^{2}{}_{n}$ and D_n for this distribution coincides with the corresponding statistics distributions for the exponential and Pareto distribution and for the Weibull distribution with known parameter α.

References

[1] Beirlant, J., De Wet, T., Goegebeur, Y. (2006) A goodness-of-fit statistic for Pareto-type behaviour. *Journal of Computational and Applied Mathematics.*, **186**, 99–116.

[2] Choulakian, V. Stephens, M.A. (2001) Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics.*, **43**, 478–484.

[3] Chibisov, D. M. (1965) An investigation of the asymptotic power of the test of fit. *Theory of Probability and Applications*, **10**, 421–437.

[4] Deheuvels, P., Martynov, G. (2003) Karhunen-Loève expansions for weighted Wiener processes and Brownian bridges via Bessel functions. *Progress in Probability.*, **55**, 57–93. Birkhäuser, Basel/Switzerland.

[5] Gikhman, I. I. (1954) One conception from the theory of ω^2-test. [in Ukrainian]. *Nauk. Zap. Kiiv Univ.*, **13**, 51–60.

[6] Gulati Sneh, Shapiro, S. (2008) Goodness of fit tests for the Pareto distribution. *Statistical Models and Methods for Biomedical and Technical Systems*, 263–277. Birkhäuser, Boston, (Vonta, F., Nikulin, M., Limnios, N., Huber, C., eds).

[7] Haywood, J., Khmaladze, E. (2008) On distribution-free goodness-of-fit testing of exponentiality. *Journal of Econometrics.*, **143**, 5–18.

[8] Kac, M., Kiefer, J., Wolfowitz, J. (1955) On tests of normality and other tests of goodness-of-fit based on distance methods. *Ann. Math. Statist.*, **30**, 420–447.

[9] Khmaladze, E.V. (1981) A martingale approach in the theory of parametric goodness-of-fit tests. " Theor. Prob. Appl.", **26**, 240–257.

[10] Martynov, G. V. The omega square tests. Moscow, "Nauka" , 1979, 80pp.

[11] Martynov, G. V. (1992) Statistical tests based on empirical processes and related questions. *J. Soviet. Math.*, **61**, 2195–2271.

[12] Martynov, G. V. (1994) Weighted Cramér-von-Mises test with estimated parameters. *LAD'2004: Longevty, Aging and Degradation Models*, StPeterburg, **2**, 207–222

[13] Neuhaus, G. (1974) Asymptotic properties of the Cramér-von Mises statistic when parameters are estimated. *Proc. Prague Symp. Asymptotic Stat.*, **2**, 1973, Prague, Charles Univ., 257–297

[14] Tyurin Yu. N., Savvushkina, N.E. (1984) Goodness-of-fit tests for Weibull-Gnedenko distribution *Izvestia AN SSSR. Tekhnicheskaya kibernetika*,no. 3, 109-112