STATISTICA, anno LXV, n. 3, 2005

TRANSFORM METHODS FOR TESTING THE NEGATIVE BINOMIAL HYPOTHESIS

Simos G. Meintanis

1. INTRODUCTION

The negative binomial distribution (NBD) is one of the most popular models for overdispersed data, i.e. for data under which, contrary to the Poisson assumption, the variance exceeds the corresponding mean. Therefore testing the null hypothesis that the data at hand follow the NBD is of considerable interest. Some standard procedures for testing goodness-of-fit based on the cumulative distribution function (CDF), have been extended to also include discrete models. For the Poisson distribution, Székely and Rizzo (2004) and Gürtler and Henze (2000) studied the finite-sample performance of the Kolmogorov-Smirnov and the Cramér-von Mises test, in comparison to new tests based on the probability generating function (PGF), and found that the latter tests compare favorably to tests based on the CDF. In addition, since for many discrete models, the PGF, unlike the CDF, can be written in closed form, the PGF-tests present a computationally more convenient solution for goodness-of-fit testing. The only PGF-test for the NBD was proposed by Rueda and O'Reilly (1999), where however the corresponding test statistic appears to be computationally complex, and at least one parameter is assumed to be known. Here we propose a computationally simple goodness-of -fit test for the NBD and generalize the approach of Rueda and O'Reilly (1999), in that both parameters of the NBD are assumed unknown. As an additional motivation for adapting the PGF-techniques in the present situation is an exchange of ideas with N.C. Weber, including a privately communicated manuscript by Binnie and Weber (2003).

Assume $X \geq 0$ is a non-degenerate (discrete) random variable and let $P(s) = \mathrm{E}(s^X)$, $0 < s \leq 1$, be the corresponding probability generating function (PGF). If X has a negative binomial distribution with parameters $\alpha > 0$, $p \in (0,1)$, i.e. if

$$\mathrm{P}(X = x) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) x!}\, p^{\alpha}(1-p)^{x}, \quad x = 0,1,2,...,$$

then its PGF is given by $\pi(s)=[1+\rho(1-s)]^{-\alpha}$, where $\rho=(1-p)/p$. Hence, subject to the condition $P(1)=1$, $\pi(s)$ is the unique solution of the differential equation

$$[1+\rho(1-s)]P'(s)-\mu P(s)=0\,, \tag{1}$$

with $\rho>0$ and $\mu=\mathrm{E}(X)>0$. On the basis of independent observations $X_1,X_2,...,X_n$ on the random variable X, we wish to test the null hypothesis,

$H_0: X$ follows a NBD for some $\rho,\mu>0$.

Our test procedure employs (1) with $P(s)$ replaced by the empirical PGF

$$P_n(s)=\frac{1}{n}\sum_{j=1}^{n}s^{X_j}\,,$$

and (ρ,μ) replaced by the moment estimator $(\hat{\rho}_n,\bar{X}_n)$, where $\hat{\rho}_n=(S_n^2-\bar{X}_n)/\bar{X}_n$, with $\bar{X}_n=n^{-1}\sum_{j=1}^{n}X_j$, and $S_n^2=n^{-1}\sum_{j=1}^{n}X_j^2-(\bar{X}_n)^2$, denoting the sample mean and sample variance, respectively. In particular let $D_n(s)=[1+\hat{\rho}_n(1-s)]P'_n(s)-\bar{X}_nP_n(s)$. The test rejects H_0 for large values of $T_{n,a}$ where,

$$T_{n,a}=n\int_0^1 D_n^2(s)\,s^a\,ds\,, \tag{2}$$

$a>0$ being a parameter the role of which we discuss later. From (2) straightforward calculations yield

$$T_{n,a}=\frac{1}{n}[\bar{X}_n^2\sum_{j,k=1}^{n}I(X_{jk}^{+}+a)-2\bar{X}_n\sum_{j,k=1}^{n}X_j\{(1+\hat{\rho}_n)I(X_{jk}^{+}+a-1)-\hat{\rho}_nI(X_{jk}^{+}+a)\}$$

$$+\sum_{j,k=1}^{n}X_jX_k\{(1+\hat{\rho}_n)^2I(X_{jk}^{+}+a-2)+\hat{\rho}_n{}^2I(X_{jk}^{+}+a)-2\hat{\rho}_n(1+\hat{\rho}_n)I(X_{jk}^{+}+a-1)\}],$$

with $X_{jk}^{+}=X_j+X_k$, $I(\beta)=\int_0^1 s^{\beta}ds=(1+\beta)^{-1}$, $\beta>-1$ and $a>1$. Obviously this expression is appropriate for computational purposes.

2. THEORETICAL RESULTS

From (2) and letting $s=\exp(-t)$, the test statistic may be written as

$$T_{n,a} = \int_0^\infty \Delta_n^2(t)\, e^{-at}\, dt\,, \quad \text{with} \quad \Delta_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \delta(X_j, \hat{\rho}_n, \bar{X}_n; t) \tag{3}$$

where

$$\delta(X, \rho, \mu; t) = \{1 + \rho(1 - e^{-t})\} X\, e^{-tX} - \mu\, e^{-t(X+1)}\,.$$

Representation (3) puts the test statistic in the framework of testing goodness-of-fit by the empirical Laplace transform, for which the technical details have been sufficiently explored. See for example, Henze and Meintanis (2002), and Meintanis and Iliopoulos (2003). Therefore most of the arguments will only be sketched.

We first show that the family of test statistics $\{T_{n,a}, 0 < a < \infty\}$, is closed at the boundary $a = \infty$. Specifically from (3) and letting $g(t) = \Delta_n^2(t)$ a Taylor expansion yields

$$g(t) = n[\overline{X_n^3} - (2\hat{\rho}_n + \bar{X}_n)\overline{X_n^2} - (\hat{\rho}_n + 2\bar{X}_n + 1)\bar{X}_n]^2 \frac{t^4}{4} + o(t^4)\,,$$

where $\overline{X_n^k} = n^{-1}\sum_{j=1}^{n} X_j^k$, $k = 2, 3$. An Abelian theorem for Laplace transforms (Zayed, 1996, Section 5.11) yields

$$\lim_{a\to\infty} a^5 T_{n,a} = 6n[\overline{X_n^3} - (2\hat{\rho}_n + \bar{X}_n)\overline{X_n^2} - (\hat{\rho}_n + 2\bar{X}_n + 1)\bar{X}_n]^2\,.$$

Therefore as $a \to \infty$, the test statistic when suitable rescaled approaches a limit value. Notice that under H_0, the stochastic limit of $\overline{X_n^3} - (2\hat{\rho}_n + \bar{X}_n)\overline{X_n^2} - (\hat{\rho}_n + 2\bar{X}_n + 1)\bar{X}_n$ is equal to zero. Kyriakoussis *et al.* (1998) arrived at a somewhat similar moment-based test statistic via a characterization of the NBD. Their statistic apart from being straightforward to compute, and although it has a simple limiting normal distribution, assumes the value of the parameter α to be a known integer, and is consistent only within the family of power-series distributions. In the following theorem the asymptotic null distribution of $T_{n,a}$ is derived.

Theorem 1. Assume that X has a NBD with parameters ρ_0 and μ_0. Then, there is a zero-mean Gaussian element Δ having covariance kernel $K(s,t) = \mathrm{E}\,\Delta(s)\Delta(t)$, such that, for the process $\Delta_n(\cdot)$ defined in (3), $\Delta_n \xrightarrow{D} \Delta$ as $n \to \infty$. Moreover,

$$T_{n,a} = \int_0^\infty \Delta_n^2(t) e^{-at}\, dt \xrightarrow{D} \int_0^\infty \Delta^2(t) e^{-at}\, dt \text{ , as } n \to \infty .$$

Proof. The main idea of the proof is to successively approximate the process $\Delta_n(\cdot)$ by asymptotically equivalent processes $\Delta_n^*(\cdot)$ and $\tilde{\Delta}_n(\cdot)$, with the last process admitting the representation

$$\tilde{\Delta}_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \tilde{\delta}(X_j, \rho_0, \mu_0; t),$$

where $\tilde{\delta}(X_1, \cdot, \cdot; t)$, $\tilde{\delta}(X_2, \cdot, \cdot; t)$,..., are independent and identically distributed random variables, satisfying $\mathrm{E}(\tilde{\delta}) = 0$ and $\mathrm{E}(\tilde{\delta}^2) < \infty$. To this end, a Taylor expansion of $\delta(\cdot, \hat{\rho}_n, \bar{X}_n; t)$ around the true values yields

$$\Delta_n^*(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \delta(X_j, \rho_0, \mu_0; t) - \sqrt{n}(\bar{X}_n - \mu_0) e^{-t} \frac{1}{\sqrt{n}} \sum_{j=1}^n e^{-tX_j}$$

$$+ (1 - e^{-t}) \sqrt{n}(\hat{\rho}_n - \rho_0) \frac{1}{n} \sum_{j=1}^n X_j e^{-tX_j} .$$

By invoking the law of large numbers it follows that the last process is asymptotically equivalent to the process resulting from $\Delta_n^*(\cdot)$ with $n^{-1}\sum_{j=1}^n e^{-tX_j}$ replaced by $\mathrm{E}(e^{-tX})$ and $n^{-1}\sum_{j=1}^n e^{-tX_j}$ replaced by $\mathrm{E}(Xe^{-tX})$.

Now recall that the moment estimator of ρ is given by $\hat{\rho}_n = g(S_n^2, \bar{X}_n)$, where $g(x, y) = (x / y) - 1$. Expanding $g(S_n^2, \bar{X}_n)$ around the true values σ_0^2 and μ_0, and utilizing the asymptotic equivalence of S_n^2 to $n^{-1}\sum_{j=1}^n (X_j - \mu_0)^2$ we arrive at $\tilde{\Delta}_n(t)$ with

$$\tilde{\delta}(X_j, \rho_0, \mu_0; t) = \delta(X_j, \rho_0, \mu_0; t) + [(1 - e^{-t})\mu_0^{-1} \mathrm{E}(Xe^{-tX})][(X_j - \mu_0^2) - \sigma_0^2]$$

$$- [(1 - e^{-t}) \frac{\sigma_0^2}{\mu_0^2} \mathrm{E}(Xe^{-tX}) + e^{-t} \mathrm{E}(e^{-tX})](X_j - \mu_0) .$$

An application of the central limit theorem and the continuous mapping theorem concludes the proof.

The following result implies the consistency of the goodness-of-fit test that rejects H_0 for large values of $T_{n,a}$ against general alternatives.

Theorem 2. Let $X \geq 0$ be a nondegenerate random variable with mean $\mu < \infty$. In addition assume that $\hat{\rho}_n \to \rho > 0$, almost surely. Then

$$\liminf_{n\to\infty} \frac{T_{n,a}}{n} \geq 0, \tag{4}$$

almost surely.

Proof. Starting with (3) we have that

$$\frac{T_{n,a}}{n} = \int_0^\infty \tilde{D}_n^2(t)\, e^{-at}\, dt,$$

where $\tilde{D}_n(t) = n^{-1/2}\Delta_n(t)$. By the strong law of large numbers it follows that $\tilde{D}_n(t) \to \tilde{D}(t)$, almost surely, with $\tilde{D}(t) = [1+\rho(1-e^{-t})]\mathrm{E}(Xe^{-tX}) - \mu e^{-t}\mathrm{E}(e^{-tX})$. Hence by Fatou's lemma,

$$\liminf_{n\to\infty} \frac{T_{n,a}}{n} \geq \tilde{\Delta}_a := \int_0^\infty \tilde{D}^2(t)\, e^{-at}\, dt,$$

almost surely, which finishes the proof of the theorem.

It may be easily seen from (1), that $\tilde{\Delta}_a$ is positive unless X follows a NBD with PGF equal to $\pi(s)$. Consequently, a level α-test that rejects H_0 for large values of $T_{n,a}$ is consistent against each fixed alternative distribution satisfying the conditions of Theorem 2. Unfortunately, and since $\tilde{\Delta}_a$ is a decreasing function of a, there is no way of choosing a "good" value of a based on Theorem 2. Therefore the choice for this parameter has to be made on the basis of finite-sample results. Also, notice that $\tilde{\Delta}_a = \infty$ when the second moment of the underlying distribution is infinite. Therefore the test that rejects the null hypothesis for large values of $T_{n,a}$ is consistent even against models with infinite variance.

3. SIMULATION RESULTS

This section presents the results of a Monte Carlo study conducted at a 10% nominal level with sample size $n = 100$. Our procedure rejects the null hypothesis either when the value of the test statistic exceeds the critical point, or when the estimates of the parameters lie outside the parameter space $(0,\infty)\times(0,1)$ for (α, p).

Since the null distribution of the test statistic depends on the (unknown) values of the parameters (α, p) we perform a parametric bootstrap to obtain the critical

point of the test as follows: Conditionally on the observed value of $(\hat{\alpha}_n, \hat{p}_n)$, generate $B=200$ bootstrap samples from $\mathrm{NBD}(\hat{\alpha}_n, \hat{p}_n)$. Calculate the value of the test statistic, say T_j^*, (j=1,2,..., B), for each bootstrap sample. Obtain the critical point as $T^*_{(0.90B)} = T^*_{(180)}$, where $T^*_{(j)}$, j=1,2,..., B, denote the ordered T_j^*-values. The parametric bootstrap is a well established technique, both on theoretical and empirical grounds. For instance it is known (Henze, 1996) that as $n, B \to \infty$, the distribution of the bootstrap statistic converges to the asymptotic distribution of $T_{n,a}$. For further theoretical justification of the parametric bootstrap the reader is referred to Stute *et al.* (1993), Henze (1996) and Babu and Rao (2004). Extensive finite-sample results are provided by Gürtler and Henze (2000), Garren *et al.* (2001), and Székely and Rizzo (2004).

It should be noted here that typically, in cases of goodness-of-fit for families involving one or more unknown shape parameters, the asymptotic distribution of the test statistic depends on the true value(s) of these parameters (Henze, 1996 and Székely and Rizzo, 2004). Moreover, the asymptotic distribution of these statistics is highly non-standard, and therefore calculation of percentage points is by no means a trivial issue. This disadvantage is shared by both, classical CDF tests as well as recently developed PGF tests. However even when these percentage points have been calculated, and assuming a large sample size, the practitioner needs to employ detailed look-up tables, each table being appropriate for a single combination of true-parameter value(s), and interpolate in that table using the current estimate(s) of the parameter value(s). Therefore, and as described above, the parametric bootstrap, despite the fact that it is computationally intensive, it is otherwise easily programmed in a computer, and provides a method that avoids the use of look-up tables, as well as reliance on doubtful asymptotic critical values.

In Table 1 and Table 2, the percentage of rejection of H_0 in 1000 replications is shown rounded to the nearest integer. The distributions considered are:

1) Negative binomial distributions with parameters α and p

2) Two-component mixtures of Poisson distributions denoted by $\mathrm{MP}(\lambda, p)$, where $X \sim \mathrm{P}(1)$ or $X \sim \mathrm{P}(\lambda)$, with probability p and $(1-p)$, respectively

3) Two-component mixtures of NBD, where $X \sim \mathrm{NBD}(1, 0.25)$ or $X \sim \mathrm{NBD}(\alpha, p)$, each with probability 0.50, denoted by $\mathrm{MNB}(\alpha, p)$

From the figures in Tables 1 and 2 we may infer that the bootstrap version of $T_{n,a}$ recovers the nominal level of significance to a satisfactory degree, and exhibits significant power against some popular alternatives to the NBD. Also, despite the fact that $T_{n,a}$ is seen to be fairly robust with respect to the parameter a, it appears that a "moderate" value of a, say $a=5$, is preferable.

TABLE 1

Empirical level for $T_{n,a}$ *in 1000 samples of size* $n = 100$ *at 10% nominal level*

DISTRIBUTION	$a = 2$	$a = 3$	$a = 5$	$a = 10$
$\alpha = 1, p = 0.25$	12	12	12	13
$\alpha = 1, p = 0.50$	11	11	10	9
$\alpha = 1, p = 0.75$	13	13	11	11
$\alpha = 2, p = 0.25$	11	10	11	12
$\alpha = 2, p = 0.50$	11	11	12	12
$\alpha = 2, p = 0.75$	14	14	13	12
$\alpha = 4, p = 0.25$	11	11	11	11
$\alpha = 4, p = 0.50$	11	10	10	9
$\alpha = 4, p = 0.75$	13	13	12	12
$\alpha = 5, p = 0.25$	10	10	10	10
$\alpha = 5, p = 0.50$	11	11	10	11
$\alpha = 5, p = 0.75$	14	13	12	11
$\alpha = 10, p = 0.25$	10	10	10	10
$\alpha = 10, p = 0.50$	9	9	10	10
$\alpha = 10, p = 0.75$	13	13	11	11

TABLE 2

Empirical power for $T_{n,a}$ *in 1000 samples of size* $n = 100$ *at 10% nominal level*

DISTRIBUTION	$a = 2$	$a = 3$	$a = 5$	$a = 10$
MP(2,0.25)	33	34	34	33
MP(2,0.50)	24	25	24	25
MP(2,0.75)	31	31	30	30
MP(3,0.25)	23	24	24	24
MP(3,0.50)	14	14	14	14
MP(3,0.75)	12	11	11	10
MP(4,0.25)	55	56	57	56
MP(4,0.50)	36	38	41	43
MP(4,0.75)	9	8	10	6
MP(5,0.25)	89	89	90	89
MP(5,0.50)	58	62	67	71
MP(5,0.75)	8	6	9	7
MP(10,0.25)	100	100	100	100
MP(10,0.50)	100	100	100	100
MP(10,0.75)	29	32	36	43
MNB(3,0.25)	11	11	10	9
MNB(3,0.50)	37	37	38	38
MNB(3,0.75)	95	96	95	94
MNB(4,0.25)	14	14	13	12
MNB(4,0.50)	66	66	65	62
MNB(4,0.75)	100	100	100	100

Department of Economics SIMOS G. MEINTANIS
National and Kapodistrian University of Athens

REFERENCES

G. BABU, C.R. RAO (2004), *Goodness-of-fit tests when parameters are estimated*, "Sankhyā", 66, pp. 63-74.

E. BINNIE, N.C. WEBER (2003), *An estimator for the shape parameter in the negative binomial distribution based on the empirical probability generating function*, Unpublished manuscript.

S. GARREN, R.L. SMITH, W. PIEGORSCH (2001), *Bootstrap goodness-of-fit test for the beta-binomial model*, "Journal of Applied Statistics", 28, pp. 561-571.

N. GÜRTLER, N. HENZE (2000), *Recent and classical goodness-of-fit tests for the Poisson distribution*, "Journal of Statistical Planning and Inference", 90, pp. 207-225.

N. HENZE (1996), *Empirical-distribution-function goodness-of-fit tests for discrete models*, "Canadian Journal of Statistics", 24, pp. 81-93.

N. HENZE, S. MEINTANIS (2002), *Tests of fit for exponentiality based on the empirical Laplace transform*, "Statistics", 36, pp. 147-161.

A. KYRIAKOUSSIS, G. LI, A. PAPADOPOULOS (1998), *On characterization and goodness-of-fit test of some discrete distribution families*, "Journal of Statistical Planning and Inference", 74, pp. 215-228.

S. MEINTANIS, G. ILIOPOULOS (2003), *Tests of fit for the Rayleigh distribution based on the empirical Laplace transform*, "Annals of the Institute of Statistical Mathematics", 55, pp. 137-151.

R. RUEDA, F. O'REILLY (1999), *Tests of fit for discrete distributions based on the probability generating function*, "Communications in Statistics-B", 28, pp. 259-274.

W. STUTE, W. MANTEIGA, M. QUINDIMIL (1993), *Bootstrap based goodness-of-fit tests*, "Metrika", 40, pp. 243-256.

G. SZÉKELY, M. RIZZO (2004), *Mean distance test of Poisson distribution*, "Statistics and Probability Letters", 67, pp. 241-247.

A.I. ZAYED (1996), *Handbook of Function and Generalized Function Transformations*, CRC Press, New York.

RIASSUNTO

Metodi basati su trasformate per saggiare l'ipotesi di distribuzione Binomiale Negativa

La funzione generatrice di probabilità empirica è utilizzata per costruire test di adattamento a distribuzioni Binomiali Negative. Si mostra che i test proposti, formulati come integrali ponderati, sono consistenti e si studia la loro distribuzione asintotica per specificate ipotesi nulle. Valori limite delle statistiche test sono valutati al tendere a zero della funzione ponderante. È infine presentato uno studio di simulazione.

SUMMARY

Transform methods for testing the negative binomial hypothesis

We employ the empirical probability generating function in constructing a goodness-of-fit test for negative binomial distributions. The proposed tests, which are formed as weighted integrals, are shown to be consistent and their asymptotic null distribution is investigated. As the decay of the weight function tends to infinity, limit statistics are obtained. A small simulation study is presented.