JIRSS (2012)
Vol. 11, No. 2, pp 191-202

A Goodness of Fit Test For Exponentiality Based on Lin-Wong Information

M. Abbasnejad, N. R. Arghami, M. Tavakoli

Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Iran.

Abstract. In this paper, we introduce a goodness of fit test for exponentiality based on Lin-Wong divergence measure. In order to estimate the divergence, we use a method similar to Vasicek's method for estimating the Shannon entropy. The critical values and the powers of the test are computed by Monte Carlo simulation. It is shown that the proposed test are competitive with other tests of exponentiality based on entropy.

Keywords. Divergence measure, entropy, exponentiality test, order statistics, Vasicek's sample entropy.

MSC: Primary: 62G10; Secondary: 94A17.

1 Introduction

In social studies, engineering, medical sciences, reliability studies and management science, it is very important to know whether the underlying data follow a particular distribution. So many authors were interested in goodness of fit tests.

Let X be a continuous random variable with distribution function $F(x)$ and probability density function $f(x)$. Consider the following hy-

M. Abbasnejad(✉)(ma_abbasnejad@yahoo.com),N. R. Arghami(arghami_nr@yahoo.com), M. Tavakoli(mahsa_tavakoli88@yahoo.com)
Received: March 5, 2012; Accepted: April 5, 2012

potheses

$$\begin{cases} H_0: & f(x) = f_0(x) \\ H_1: & f(x) \neq f_0(x), \end{cases}$$

where $f_0(x) = \theta e^{-\theta x}$, $x > 0$, $\theta > 0$, and θ is unknown.

Many authors including Lilliefors (1969),Van-Soest (1969), Finkelstein and Schafer (1971), Stephens (1974) and Harris (1976) presented different test statistics for exponentiality. For the first time, Ebrahimi *et al.* (1992) introduced an exponentiality test based on entropy.

The entropy of X is defined by Shannon (1948) as

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{1}$$

The problem of estimation of $H(f)$ has been considered by many authors including Ahmad and Lin (1976), Vasicek (1976), Dudewicz and Van der Meulen (1987), Joe (1989), Van Es (1992), Correa (1995), Wieczorkowski and Grzegorewski (1999), Yousefzadeh and Arghami (2008) and Alizadeh (2010).

Many researchers presented the goodness of fit tests based on various entropy estimators. Among these various entropy estimators, Vasicek's sample entropy has been most widely used in goodness of fit tests.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution F. Using $F(x) = p$, Vasicek expressed equation (1) as

$$H(f) = \int_0^1 \log\{\frac{d}{dp} F^{-1}(p)\} dp,$$

and by replacing the distribution function F by the empirical distribution function F_n and using a difference operator instead of the differential operator, the derivative of $F^{-1}(p)$ was estimated by

$$\frac{X_{(i+m)} - X_{(i-m)}}{2m/n}.$$

Therefore $H(f)$ was estimated as

$$HV_{n,m} = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \frac{n}{2m} (X_{(i+m)} - X_{(i-m)}) \right\},$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ are the order statistics and m is a positive integer smaller than $n/2$. For $i < 1$, $X_{(i)} = X_{(1)}$ and for $i > n$, $X_{(i)} =$

$X_{(n)}$.

The asymmetric Kullback-Leibler distance of f from f_0 is:

$$D(f, f_0) = \int_0^{+\infty} f(x) \log \frac{f(x)}{f_0(x)} dx = -H(f) - \ln \theta + \theta E(X).$$

It is well known that $D(f, f_0) \geq 0$ and the equality holds if and only if $f(x) = f_0(x)$ almost everywhere. Ebrahimi *et al.* (1992) introduced an exponentiality test based on Kullback-Leibler information and estimated the test statistic by Vasicek entropy estimator. So, they introduced the test statistic as

$$TV_{n,m} = \frac{\exp(HV_{n,m})}{\exp(\log(\overline{X})) + 1}.$$

Ebrahimi *et al.* (1994) proposed a modified sample entropy and Park and Park (2003) derived test statistics for normality and exponentiality based on the modified estimation of entropy. Yousefzadeh and Arghami (2008) estimated Shannon entropy based on a new estimator of distribution function and used it to obtain a test statistic for exponentiality. Gurevich and Davidson (2008) proposed a test statistic which is the standardized version of the test statistic of Ebrahimi *et al.* (1992). Vexler and Gurevich (2010) and Gurevich and Vexler (2011) developed empirical likelihood ratio tests for goodness of fit and demonstrated that the well-known goodness of fit tests based on sample entropy and Kullback-Leibler information are a product of the proposed empirical likelihood methodology. Alizadeh and Arghami (2011a) compared five exponentiality tests using different entropy estimators like Vasicek (1976), Van Es (1992), Correa (1995) and Alizadeh (2010).

Another distance of f from f_0 is introduced by Renyi (1961) as

$$D_r(f, f_0) = \frac{1}{r-1} \log \int_0^{+\infty} (\frac{f(x)}{f_0(x)})^{r-1} f(x) dx, \quad r > 0 \ (\neq 1).$$

$D(f, f_0) \geq 0$ and the equality holds if and only if $f = f_0$. Abbasnejad (2011) introduced a test based on Renyi information for normality and exponentiality.

The rest of the paper is organized as follows. In Section 2, we proposed a test statistic for exponentiality based on Lin-Wong information. In Section 3, a simulation study is performed to analyze the behavior of the test statistic. We compare the proposed test with the other tests of exponentiality based on information measures.

2 Test statistics

Lin and Wong (1990) introduced a new divergence distance of two density functions $f(x)$ and $g(x)$ as

$$D_{LW}(f,g) = \int_{-\infty}^{\infty} f(x) \log \frac{2f(x)}{f(x)+g(x)} dx.$$

Since Lin-Wong information belongs to Csiszer family, we have $D_{LW}(f,g) \geq 0$ and the equality holds if and only if $f(x) = g(x)$ (See Kapur and Kesavan, 1992). So, it motivates us to use Lin-Wong information as a test statistic for exponentiality.

Lin-Wong information in favor of $f(x)$ against $f_0(x)$ is

$$D_{LW}(f,f_0) = \int_{0}^{\infty} f(x) \log \frac{2f(x)}{f(x)+\theta e^{-\theta x}} dx. \tag{2}$$

Under the null hypothesis $D_{LW}(f,f_0) = 0$ and large values of $D_{LW}(f,f_0)$ favor H_1.

To estimate $D_{LW}(f,f_0)$, we use two following methods.

In the first method, using $F(x) = p$, similar to Vasicek's method we express equation (2) as

$$\int_0^1 \log \frac{2(\frac{dF^{-1}(p)}{dp})^{-1}}{(\frac{dF^{-1}(p)}{dp})^{-1} + \theta e^{-(\theta F^{-1}(p))}} dp.$$

Now, replacing F by F_n and using difference operator in place of the differential operator, we get an estimator L_V of $D_{LW}(f,f_0)$ as

$$L_V = -\frac{1}{n}\sum_{i=1}^{n} \log\{\frac{1}{2} + \frac{n}{4m\bar{X}}(X_{(i+m)} - X_{(i-m)})e^{-\frac{X_{(i)}}{\bar{X}}}\}, \tag{3}$$

where $X_{(i)} = X_{(1)}$ for $i < 1$ and $X_{(i)} = X_{(n)}$ for $i > n$.

In equation (3), we used maximum likelihood estimator, $1/\bar{X}$ instead of θ. It is obvious that L_V is invariant with respect to scale transformation.

It must be noted that to estimating $D_{LW}(f,f_0)$ in the second method we use a method similar to Bowman (1992) for estimating the Shannon entropy by Kernel density function estimation. However, we do not take it into consideration here since its performance is poor in terms of powers.

Now, similar to the proof of Theorem 2 of Alizadeh and Arghami (2011b), we prove that the test based on L_V is consistent.

Theorem 2.1. *Let F be an unknown continuous distribution with a positive support and F_0 be the exponential distribution with unspecified parameter. Then under H_1 the test based on L_V is consistent.*

Proof. As $n, m \to \infty$ and $m/n \to 0$, we have

$$\begin{aligned}\frac{2m}{n} &= F_n(X_{(i+m)}) - F_n(X_{(i-m)}) \simeq F(X_{(i+m)}) - F(X_{(i-m)}) \\ &\simeq \frac{f(X_{(i+m)}) + f(X_{(i-m)})}{2}(X_{(i+m)} - X_{(i-m)}),\end{aligned}$$

where $F_n(a) = (\#x_i \leq a)/n = (1/n)\sum I_{(-\infty, X_i]}(a)$, and I is the indicator function. Therefore noting that $\frac{1}{\overline{X}}$ is the MLE of θ and it is consistent, we have

$$\begin{aligned}L_V &= -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{1}{2} + \frac{1}{\overline{X}}e^{-\frac{X_{(i)}}{\overline{X}}}.\frac{1}{2}.\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\right\} \\ &\simeq -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{1}{2} + \theta e^{-\theta X_{(i)}}.\frac{1}{2}.\frac{2}{f(X_{(i+m)}) + f(X_{(i-m)})}\right\} \\ &\simeq -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{1}{2} + \theta e^{-\theta X_{(i)}}.\frac{1}{2f(X_{(i)})}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{2f(X_i)}{f(X_i) + \theta e^{-\theta X_i}}\right\} \longrightarrow E(\log\left\{\frac{2f(X_i)}{f(X_i) + \theta e^{-\theta X_i}}\right\}) \\ &= \int_0^\infty f(x)\log\frac{2f(x)}{f(x) + f_0(x)}dx = D(f, f_0),\end{aligned}$$

where the last limit holds by the low of large numbers. So, the test based on L_V is consistent.

Remark 2.1. It may be noted that Lin-Wong information can be used to constructing general goodness of fit tests (not just for exponentiality). One can consider any known density function (with known or unknown parameters) under the null hypothesis and put it instead of the function $g(x)$ in the definition of Lin-Wong distance to obtain the test statistic. For example, a test of normality had been considered by the authors, however, it had a poor performance.

3 Simulation study

A simulation study is performed to analyze the behavior of the proposed test statistic.

Table 1. Critical values of L_V for $\alpha = 0.01$ and $\alpha = 0.05$.

	α	
n	0.01	0.05
5	0.4762	0.3937
6	0.4277	0.3459
7	0.4114	0.3252
8	0.3664	0.2945
9	0.3256	0.2574
10	0.3080	0.2349
15	0.2218	0.2255
20	0.1759	0.1780
25	0.1533	0.1482
30	0.1296	0.1282

We determine the critical points using Monte Carlo simulation with 10000 replicates. For choice of m we use the formula, $m = [\sqrt{n} + 0.5]$, which was used by Wieczorkowski and Grzegorzewski (1999). Table 1 gives the critical values of L_V for various sample sizes.

We compute the powers of the test based on L_V statistic by Monte Carlo simulation. To facilitate comparison of the power of the proposed test with powers of the tests published, we selected the same three alternatives listed in Ebrahimi *et al.* (1992), Gurevich and Davidson (2008) and Alizadeh and Arghami (2011a) and their choices of parameters:
(a) the Weibull distribution with density function

$$f(x;\lambda,\beta) = \beta\lambda^{\beta}x^{\beta-1}\exp\{-(\lambda x)^{\beta}\}, \quad x > 0,\ \beta > 0,\ \lambda > 1,$$

(b) The gamma distribution with density function

$$f(x;\lambda,\beta) = \frac{\lambda^{\beta}x^{\beta-1}\exp\{-\lambda x\}}{\Gamma(\beta)}, \quad x > 0,\ \beta > 0,\ \lambda > 1,$$

(c) The Log-Normal distribution with density function

$$f(x;\nu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp -\frac{1}{2\sigma^2}(\ln(x)-\nu)^2, \quad x > 0,\ -\infty < \nu < \infty,\ \sigma^2 > 0.$$

We also chose the parameters so that $E(X) = 1, i.e.\ \lambda = \Gamma(1 + (1/\beta))$ for the Weibull, $\lambda = \beta$ for the gamma and $\nu = -\sigma^2/2$ for the log-normal family of distributions.

The test statistics of competitor tests are as follows:
(1) Ebrahimi *et al.* (1992)

$$TV_{n,m} = \frac{\exp(HV_{n,m})}{\exp(\log(\bar{X})) + 1}.$$

(2) Abbasnejad (2011)

$$ED_r^V = \log \bar{X} + \frac{1}{r-1} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[(\frac{2m/n}{X_{(i+m)} - X_{(i-m)}}) e^{\frac{X_{(i)}}{\bar{X}}} \right]^{r-1} \right\}.$$

(3) Gurevich and Davidson (2008)

$$MKL_n^1 = \max_{1 \le m < n/2} \left\{ \frac{n \left[\prod_{j=1}^{n} (X_{(j+m)} - X_{(j-m)}) \right]^{1/n}}{2me\bar{X}} \right\}.$$

(4) Alizadeh and Arghami (2011a)

$$TA_{n,m} = \frac{\exp(HA_{n,m})}{\exp(\log(\bar{X}) + 1)},$$

where

$$HA_{n,m} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\hat{f}(X_{(i+m)}) + \hat{f}(X_{(i+m)})}{2} \right),$$

where $\hat{f}(X) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{X - X_{(i)}}{h})$, and the Kernel function is chosen to be the standard normal density function and the bandwidth h is chosen to be the normal optimal smoothing formula, $h = 1.06sn^{-1/5}$, where s is the sample standard deviation.

The goodness of fit test based on entropy involves choosing the best integer parameter m. Unfortunately, there is no choice criterion of m, and in general it depends on the alternative. Ebrahimi *et al.* (1992) tabulated the values of m, which maximize the power of the test. Similar table is given by Abbasnejad (2011). Gurevich and Davidson (2008) obtained their test statistic by maximizing the test statistic of Ebrahimi *et al.* (1992) over the various values of m and so they did not need to choose the best values of m. It is shown that by Alizadeh and Arghami (2011a), there is no m that is optimal for all alternatives. We suggest the value of m similar to Ebrahimi *et al.* (1992) based on simulation results. Tables 2-4 show the estimated power of the test L_V and those of the competing tests, at the significance level $\alpha = 0.05$ and $\alpha = 0.01$ based on the result of 10000 simulation (of sample size 10,20). For the

Table 2. Power comparisons against the gamma distribution at the significance levels $\alpha = 0.01, 0.05$.

n	β	α	ED_r^V	$TV_{n,m}$	MKL_n^1	$TA_{n,m}(m)$	L_V
10	2	0.01	0.086	0.136	0.118	0.122(3)	**0.137**
		0.05	0.295	0.355	0.324	0.362(3)	**0.365**
	3	0.01	0.245	0.348	0.300	0.345(3)	**0.355**
		0.05	0.584	0.637	0.601	0.692(3)	**0.698**
	4	0.01	0.434	0.577	0.501	0.563(3)	**0.590**
		0.05	0.788	0.859	0.790	0.882(3)	**0.885**
20	2	0.01	0.150	0.244	0.281	**0.360**(5)	0.342
		0.05	0.421	0.485	0.550	**0.646**(5)	0.629
	3	0.01	0.513	0.690	0.707	**0.817**(5)	0.791
		0.05	0.817	0.873	0.902	**0.961**(5)	0.942
	4	0.01	0.787	0.924	0.921	**0.972**(5)	0.958
		0.05	0.962	0.986	0.988	**0.998**(6)	0.995

Table 3. Power comparisons against the Weibull distribution at the significance levels $\alpha = 0.01, 0.05$.

n	β	α	ED_r^V	$TV_{n,m}$	MKL_n^1	$TA_{n,m}(m)$	L_V
10	2	0.01	0.320	0.425	0.364	0.421(3)	**0.425**
		0.05	0.668	0.702	0.662	0.759(1)	**0.759**
	3	0.01	0.832	0.900	0.831	0.904(3)	**0.906**
		0.05	0.982	0.986	0.962	0.992(2)	**0.993**
	4	0.01	0.985	0.993	0.978	0.995(3)	**0.995**
		0.05	1.000	1.000	0.999	1.000(3)	**1.000**
20	2	0.01	0.650	0.783	0.789	**0.860**(5)	0.854
		0.05	0.901	0.929	0.941	**0.977**(1)	0.969
	3	0.01	0.996	1.000	0.999	1.000(2)	**1.000**
		0.05	1.000	1.000	1.000	1.000(1)	**1.000**
	4	0.01	1.000	1.000	1.000	1.000(1)	**1.000**
		0.05	1.000	1.000	1.000	1.000(1)	**1.000**

Table 4. Power comparisons against the log-normal distribution at the significance levels $\alpha = 0.01, 0.05$.

n	β	α	ED_r^V	$TV_{n,m}$	MKL_n^1	$TA_{n,m}(m)$	L_V
10	-0.3	0.01	0.073	0.109	0.100	0.097(3)	**0.117**
		0.05	0.272	0.308	0.298	0.302(4)	**0.317**
	-0.2	0.01	0.186	**0.302**	0.254	0.262(3)	0.280
		0.05	0.511	0.602	0.554	0.593(3)	**0.606**
	-0.1	0.01	0.637	0.784	0.704	0.775(3)	**0.796**
		0.05	0.921	0.954	0.918	0.968(3)	**0.96**1
20	-0.3	0.01	0.128	0.223	0.236	0.276(7)	**0.278**
		0.05	0.397	0.478	0.499	**0.569**(8)	0.534
	-0.2	0.01	0.378	0.591	0.621	**0.727**(5)	0.679
		0.05	0.735	0.827	0.848	**0.921**(8)	0.872
	-0.1	0.01	0.938	0.991	0.991	0.992(6)	**0.993**
		0.05	0.997	0.999	0.999	1.000(6)	**1.000**

test statistic $TA_{n,m}$, the value of m which maximizes the power of the test for each alternative is given in parentheses.

According to Tables 2, 3 and 4, the L_V test behaves better than other tests for $n = 10$ and for all of the alternatives (except log-normal(-0.2)). However, for $n = 20$, the $TA_{n,m}$ test is better than the other tests and the L_V test has greater or equal powers for some alternatives. So we can suggest the L_V test statistic for small sample sizes. Also, for large sample sizes, the L_V test has the advantage of having fixed m, in comparison with $TA_{n,m}$ and one may prefer the proposed test.

4 Conclusion

In this paper, we introduced a goodness of fit test for exponentiality based on Lin-Wong divergence measure. To construct the test statistic we estimated the Lin-Wong distance similar to Vasicek's method for estimating of the Shannon entropy. By a simulation study the powers of the proposed test were computed under several alternatives and different sample sizes. It is shown that, L_V test compares favorably with the leading competitors specially for small sample sizes.

Acknowledgements

The authors would like to thank the referees for their attention to the topic and useful comments and suggestions. Partial support from Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad is acknowledged.

References

Abbasnejad, M. (2011), Some goodness of fit tests based on Renyi information. Applied Mathematical Sciences, **5**, 1921-1934.

Ahmad, I. A. and Lin, P. E. (1976), A nonparametric estimation of the entropy for absolutely continuous distributions. IEEE. Trans. Inform. Theo., **22**, 327-375.

Alizadeh Noughabi, H. (2010), A new estimator of entropy and its application in testing normality. J. Statist. Comput. Simul., **80**, 1151-1162.

Alizadeh Noughabi, H. and Arghami, N. R. (2011a), Monte Carlo comparison of five exponentiality tests using different entropy estimates. J. Statist. Comput. Simul., **81**, 1579-1592.

Alizadeh Noughabi, H and Arghami, N. R. (2011b), Testing exponentiality based on characterizations of the exponential distribution. J. Statist. Comput. Simul., **81**, 1641-1651.

Bowman, A. W. (1992), Density based tests for goodness of fit. J. Statist. Comput. Simul., **40**, 1-13.

Correa, J. C. (1995), A new estimators of entropy. Commun. Statist. Theory Meth., **24**, 2439-2449.

Dudewicz, E. and Van der Meulen, E. (1987), The empiric entropy, a new approach to nonparametric entropy estimation. In New Perspectives in Theoretical and Applied Statistics, ML Puir, JP Vilaplana and W Wentz, eds, New York: Wiley, 207-227.

Ebrahimi, N., Habibullah, M., and Soofi, E. S. (1992), Testing exponentiality based on Kullback-Leibler information. J. R. Statist. Soc., **54**, 739-748.

Ebrahimi, N., Pflughoeft, K., and Soofi, E. S. (1994), Two measures of sample entropy. Statist. Prob. Lett., **20**, 225-234.

Finkelstein, J. and Schafer, R. E. (1971), Imported goodness of fit tests. Biometrika, **58**, 641-645.

Gurevich, G. and Davidson, A. (2008), Standardized forms of Kullback-Leibler information based statistics for normality and exponentiality. Computer Modelling and New Technologies, **12**, 14-25.

Gurevich, G. and Vexler, A. (2011), A two-sample emprical likelihood ratios test based an samples entropy. Statist. Comput., **21**, 657-670.

Harris, C. M. (1976), A note on testing for exponentiality. Nav. Res. Logist. Q., **28**, 169-175.

Joe, H. (1989), Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math., **41**, 683-697.

Kapur, J. N. and Kesavan, H. K. (1992), Entropy optimization principles with application. Academic Press.

Lin, J. and Wong, S. K. M. (1990), A new directed divergence measure and its characterization. Int. J. General Systems, **17**, 73-81.

Lilliefors, H. W. (1969), On the Kolmogorov-Smirnov test for exponential distribution with mean unknown. J. Am. Statist. Assoc., **64**, 387-389.

Park, S. and Park, D. (2003), Correcting moments for goodness of fit tests based on two entropy estimates. J. Statist. comput. Simul., **73**, 685-694.

Renyi, A. (1961), On measures of entropy and information. In Proceeding of the Fourth Berkeley Symposium I, UC Press, Berkeley, 547-561.

Shannon, C. E. (1948), A mathematical theory of communication. Bell System Technical Journal, **27**, 379-423.

Stephens, M. A. (1974), EDF statistics for goodness of fit and some comparisons. J. Am. Statist. Assoc., **69**, 730-737.

Van Es, B. (1992), Estimating functional related to a density by a lass of statistic based on spacings. Scand. J. Statist., **19**, 61-72.

Van-Soest, J. (1969), Some goodness of fit tests for the exponential distribution. Statist. Neerland., **23**, 41-51.

Vasicek, O. (1976), A test foe normality based on sample entropy. J. R. Statist. Soc., **38**, 54-59.

Vexler, A. and Gurevich, G. (2010), Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. Comput. Statist. Data Anal., **54**, 531-545.

Wieczorkowski, R. and Grzegorzewski, P. (1999), Entropy estimators - Improvements and comparisons. Commun. Statist. Comput. Simul., **28**, 541-567.

Yousefzadeh, F. and Arghami N. R. (2008), Testing exponentiality based on typeII censored data and a new cdf estimator. Commun. Statist. Simul. Comput. **37**, 1479-1499.