Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 443 a 455

On the Use of Ranked Set Samples in Entropy Based Test of Fit for the Laplace Distribution

Uso de muestras de rango ordenado en una prueba de ajuste basada en entropía para la distribución Laplace

Mahdi Mahdizadeh[a]

Department of Statistics, Hakim Sabzevari University, Sabzevar, Iran

Abstract

Statistical methods based on ranked set sampling (RSS) often lead to marked improvement over analogous methods based on simple random sampling (SRS). Entropy has been influential in the development of measures of fit of parametric models to the data. This article develops goodness-of-fit tests of the Laplace distribution based on sample entropy when data are collected according to some RSS-based schemes. For each design, critical values of the corresponding test statistic are estimated, by means of simulation, for some sample sizes. A Monte Carlo study on the power of the new tests is performed for several alternative distributions and sample sizes in order to compare our proposal with available method in SRS. Simulation results show that RSS and its variations lead to tests giving higher power than the test based on SRS.

***Key words*:** Entropy estimation, Goodness-of-fit test, Ranked set sampling.

Resumen

Los métodos estadísticos basados en muestreo de rango ordenado a menudo son una considerable mejora que el muestreo aleatorio simple. La medida de entropía ha sido influencial en el desarrollo de medidas de ajuste de modelos paramétricos. Este artículo propone pruebas de bondad de ajuste de la distribución Laplace basada en la entropía muestral cuando se usan estructuras basadas en muestras de rango ordenado. Para cada diseño, los valores críticos del correspondiente estadístico de prueba son estimados por medio de simulaciones para diferentes tamaños de muestra. Un estudio de Monte Carlo de la potencia de los nuevos tests es implementado para diferentes distribuciones alternas y tamaños de muestra con el fin de comparar el método propuesto con otros disponibles. La simulación muestra que el muestreo de rango ordenado y sus variaciones brindan mayor potencia que los métodos basados en muestreo aleatorio simple.

***Palabras clave*:** entropía, muestreo rango ordenado, prueba de bondad de ajuste.

[a]Assistant Professor. E-mail: mahdizadeh.m@live.com, noniid@yahoo.com

1. Introduction

The ranked set sampling (RSS) was introduced by McIntyre (1952) who built on the sample mean to obtain a more precise estimator of the population mean. In this design, the experimenter exploits inexpensive additional information about the characteristic of interest for ranking randomly drawn sampling units and then quantifies a selected subset of them. Auxiliary information may be provided by, for example, visual inspection, concomitant variables, expert opinion, etc., or some combinations of these methods. This flexibility in the choice of ranking mechanism is an appealing feature which makes RSS a cost-efficient sampling technique potentially applicable in fields such as agriculture, biology, ecology, forestry, etc. As an example, consider the following situation mentioned by Takahasi & Wakimoto (1968). Suppose that the quantity of interest is the height of trees in a orchard. While the actual measurement is going to be laborious, a simple glance can help us to rank a handful of trees locating close to each other.

The RSS method can be summarized as follows:

1. Draw k random samples, each of size k, from the target population.
2. Apply judgement ordering, by any cheap method, on the elements of the ith $(i = 1, \ldots, k)$ sample and identify the ith smallest unit.
3. Actually measure the k identified units in step 2.
4. Repeat steps 1-3, h times (cycles), if needed, to obtain a ranked set sample of size $n = hk$.

The set of n measured observations are said to constitute the ranked set sample denoted by $\{X_{[i]j} : i = 1, \ldots, k\,; j = 1, \ldots, h\}$, where $X_{[i]j}$ is the ith judgement order statistic from the jth cycle. Current literature on RSS reports many statistical procedures, in both parametric and nonparametric settings, which are superior to their counterparts in simple random sampling (SRS). For an excellent review of most previous works on RSS, see the recent book by Chen, Bai & Sinha (2004). The success of RSS can be traced to the fact that a ranked set sample consists of independent order statistics and contains more information than a simple random sample of the same size, whose ordered values are correlated.

A basic version of RSS has been extensively modified to come up with schemes resulting in more accurate estimators of the population attributes. Multistage ranked set sampling (MSRSS) introduced by Al-Saleh & Al-Omari (2002) is such a variation surpassing RSS. The MSRSS scheme can be described as follows:

1. Randomly identify k^{r+1} units from the population of interest, where r is the number of stages.
2. Allocate the k^{r+1} units randomly into k^{r-1} sets of k^2 units each.
3. For each set in step 2, apply 1-2 of RSS procedure explained above, to get a (judgement) ranked set of size k. This step gives k^{r-1} (judgement) ranked sets, each of size k.

4. Without actual measuring of the ranked sets, apply step 3 on the k^{r-1} ranked set to gain k^{r-2} second stage (judgement) ranked sets, of size k each.

5. Repeat step 3, without any actual measurement, until an rth stage (judgement) ranked set of size k is acquired.

6. Actually measure the k identified units in step 5.

7. Repeat steps 1-6, h times, if needed, to obtain an rth stage ranked set sample of size $n = hk$.

Similarly, the rth stage ranked set sample will be denoted by $\{X^{(r)}_{[i]j} : i = 1, \ldots, k; j = 1, \ldots, h\}$. It is to be noted that special case of MSRSS with $r = 2$ is known as double ranked set sampling (DRSS) (Al-Saleh & Al-Kadiri 2000). Clearly, the case $r = 1$ corresponds to RSS.

While testing hypotheses on the parameters of the normal, exponential and uniform distributions under RSS and its variations have been widely investigated, little effort has been made for developing a test of fit based on RSS. Stokes & Sager (1988) characterized a ranked set sample as a sample from a conditional distribution, conditioning on a multinomial random vector, and applied RSS to the estimation of the cumulative distribution function. They proposed the Kolmogorov-Smirnov test in RSS setup and derived the null distribution of the test statistic.

Entropy of a distribution was proposed by Shannon (1948) as a measure of uncertainty in information theory. He found that the entropy of the normal distribution is maximum among all distributions with fixed variance. Based on this result, Vasicek (1976) developed a test for normality and, indeed, introduced a new approach for constructing test of fit. Similar tests have been suggested for other distributions based on their entropy characterization results. See Dudewicz & van der Meulen (1981), Gokhale (1983), Grzegorzewski & Wieczorkowski (1999), Mudholkar & Tian (2002), and Choi & Kim (2006).

The classical Laplace distribution introduced by Laplace in 1774 is one of the basic symmetric distributions often used for modeling phenomena with heavier than the normal tails. It has been applied in steam generator inspection, navigation, reliability, generalized linear regression and Bayesian analysis. For more recent applications refer to Kotz, Kozubowski & Podgórski (2001). In this work, we deal with the problem of developing a goodness-of-fit test for the Laplace distribution when the researcher obtains data using RSS and MSRSS. Mahdizadeh & Arghami (2010) suggested similar procedures for the inverse Gaussian law.

The layout of this article is as follows: In Section 2, entropy estimation is extended to RSS and MSRSS, goodness-of-fit tests for the Laplace distribution based on these designs are introduced. Section 3 contains the results of simulation studies carried out to expose the power properties of the new tests. Section 4 is given to the effect of entropy estimator used in the test statistics on power properties. Some brief conclusions are provided in Section 5.

2. Proposed Tests

To put the procedure into perspective, we first review some concepts from information theory. Suppose that a continuous random variable X has distribution function F_X with density function f_X. Shannon's entropy of f_X is given by

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

It is easy to show that using the quantile function $F_X^{-1}(u) = \inf\{x : F_X(x) \geq u\}$, (1) can be written as

$$H(f_X) = \int_0^1 \log\big(\frac{d}{du} F_X^{-1}(u)\big)\, du \tag{2}$$

This entropy representation was used by Vasicek (1976) to define the sample entropy in terms of order statistics as follows: Let $X_{(1)}, \ldots, X_{(n)}$ be the ordered values of a random sample of size n from F_X. At each sample point $(X_{(i)}, \frac{i}{n})$, the derivative in (2) is estimated by

$$s_i(m, n) = \frac{X_{(i+m)} - X_{(i-m)}}{2\, m/n} \tag{3}$$

where $m \in \{1, \ldots, \frac{n}{2}\}$ is a window size to be determined. Vasicek's entropy estimator is the mean of logarithm of d_i's defined in the above, i.e.,

$$V_{m,n}(f_X) = \frac{1}{n} \sum_{i=1}^{n} \log\Big(\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\Big) \tag{4}$$

where $X_{(i-m)} = X_{(1)}$ for $i \leq m$ and $X_{(i+m)} = X_{(n)}$ for $i \geq n - m$.

Since the entropy estimator (4) is based on spacings, we would need ordered values of the ranked set sample to estimate entropy in RSS. Proceeding as in the SRS case, we first pool the units in all cycles and then form the estimator based on the ordered pooled sample. The MSRSS analogue of $V_{m,n}(f_X)$ becomes

$$V_{m,n}^{(r)}(f_X) = \frac{1}{n} \sum_{i=1}^{n} \log\Big(\frac{n}{2m}(X_{(i+m)}^{(r)} - X_{(i-m)}^{(r)})\Big) \tag{5}$$

where $X_{(a)}^{(r)}$ is the ath $(a = 1, \ldots, n)$ order statistic of the rth stage ranked set sample. The reference to subscript k is not made here for conciseness in notation. From now on, we use $V_{m,n}^{(0)}(f_X)$ to denote the estimator (4). So $X_{(a)}^{(0)}$ represents ath order statistic of a simple random sample of size n. In fact, $\{V_{m,n}^{(r)}(.)\}$ is a sequence of entropy estimators indexed by the stage number in MSRSS.

A Monte Carlo experiment was conducted to compare the proposed estimators of entropy when the underlying distribution is the standard Laplace with mean 0 and variance 2. Generation of random samples is easily done based on a result from

distribution theory; difference a of two standard exponential random variables has the standard Laplace distribution. Figure 1 displays simulated biases and root mean square errors (RMSEs) of $V_{m,n}^{(r)}$ for $r = 0, 1, 2$ based on 50,000 samples with $n = 10, 20, 30$, and $k = 5$ in MSRSS design (this setup will be used in the rest of the paper). An empty circle is used as the plotting symbol, and points corresponding to SRS, RSS and DRSS are connected by solid, dashed and dotted lines, respectively. It is seen that given a sample size, MSRSS improves entropy estimation with respect to SRS. Moreover, the larger stage number r the smaller absolute value of bias, and RMSE of the corresponding estimator. This property is helpful in distinguishing between the results of different designs when the types of connecting lines are not visible because of compactness in Figure 1.

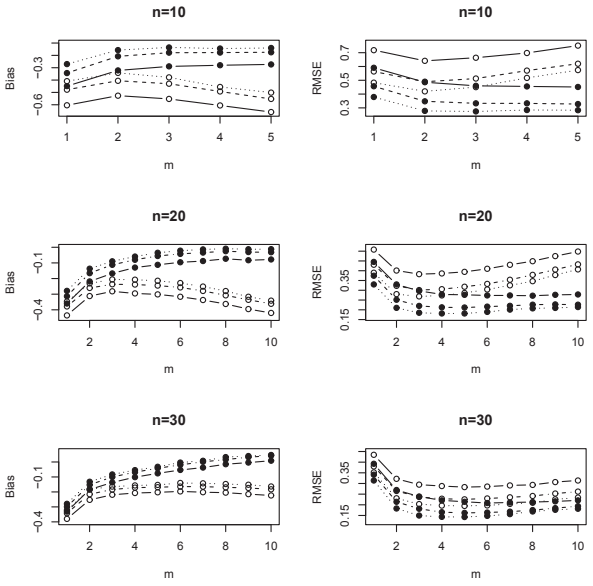

FIGURE 1: Bias and RMSE comparison for the entropy estimators $V_{m,n}$ and $E_{m,n}^{1}$ for the standard Laplace distribution with $H(f) = 1.6931$.

Choi & Kim (2006) presented an entropy characterization of the Laplace distribution and used the following result (Corollary 2) to establish an entropy based test of fit for the Laplace distribution.

Corollary 1. *(Choi & Kim 2006). Suppose X has a Laplace distribution La(μ,θ) with density function*

$$f_X(x;\mu,\theta) = \frac{1}{2\theta}\exp(-|x-\mu|/\theta) \quad \mu \in R\,,\, \theta > 0$$

Then the entropy of f_X *is given by*

$$H(f_X) = \log(2\theta) + 1$$

Corollary 2. *(Choi & Kim 2006). Let X be a random variable with density function $f_X(x)$ satisfying the restriction*

$$E_{f_X}(|X|) = \int_{-\infty}^{\infty} |x|\, f_X(x)\, dx \equiv \theta$$

Under this restriction, the distribution of X maximizing Shannon's entropy is La(0,θ).

Consider a random sample $X_1, \ldots, X_n$ from a population with density function f and suppose it is of interest to test $H_0 : f_X \in \mathcal{L} = \{La(\mu, \theta) : \mu \in R, \theta > 0\}$ against the general alternative $H_1 : f_X \notin \mathcal{L}$. Choi & Kim (2006) proposed rejecting the null hypothesis if

$$T_{m,n}(g_Y) = \exp\big(V_{m,n}(g_Y)\big)/\,\widehat{\theta} \leq T^*_{m,n,\alpha}(g_Y) \tag{6}$$

where

$$V_{m,n}(g_Y) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{n}{2m}(\widetilde{Y}_{(i+m)} - \widetilde{Y}_{(i-m)})\Big)$$

is the estimate of the entropy of $\widetilde{Y} = X - \mu$ based on $\widetilde{Y}_{(i)} = X_{(i)} - \widehat{\mu}$ $(i = 1, \ldots, n)$ with $\widehat{\mu}$ being the median of X_i's, $\widehat{\theta} = \sum_{i=1}^{n} |\widetilde{Y}_i|/n$, and $T^*_{m,n,\alpha}(g_Y)$ is the 100α percentile of the null distribution of $T_{m,n}(g_Y)$.

In order to obtain the percentiles of the null distribution, $T_{m,n}(g_Y)$ was calculated using the estimators $V^{(r)}_{m,n}(g_Y)$ for $r = 0, 1, 2$ based on 50,000 samples of size n generated from the La(0,1) distribution. The values were then used to determine $T^*_{m,n,0.05}(g_Y)$ in different designs and for different sample sizes. To estimate μ and θ in MSRSS, we simply plug the data into the formulae available in SRS. Tables of 0.05 critical points for the tests could be requested from the author. They are not reported here.

To implement the tests, we must first select the window size m associated with a given sample size. In general, there is no unanimous rule to choose the optimal m for each n. Previous studies, however, suggest to of use the window size which leads to the least conservative test. Thus, using the window size giving the largest critical value is advised to achieve higher power. The optimal window size, denoted by m^*, for sample sizes 10, 20 and 30 are approximately 3, 3 and 4, respectively.

3. Simulation Study

In this section, we shall use the Monte Carlo approach to evaluate the entropy tests in terms of power. The distributions considered in the simulation study are as follows: (A) normal(0,1), (B) t(10), (C) logistic(0,1), (D) uniform(0,1), (E) Beta(2,2), (F) chi-square(4), (G) lognormal(0,0.5) and (H) Gamma(1.5,1). We note that (A)-(E) are symmetric and (F)-(H) are asymmetric.

Under each design, 50,000 samples of sizes $n = 10, 20, 30$ were generated from each alternative distribution and the power of the tests were estimated by the

fraction of the samples falling into the corresponding critical region. Figures 2-7 depict the estimated power of the tests in which the same plotting symbol and connecting lines of Figure 1 are employed.

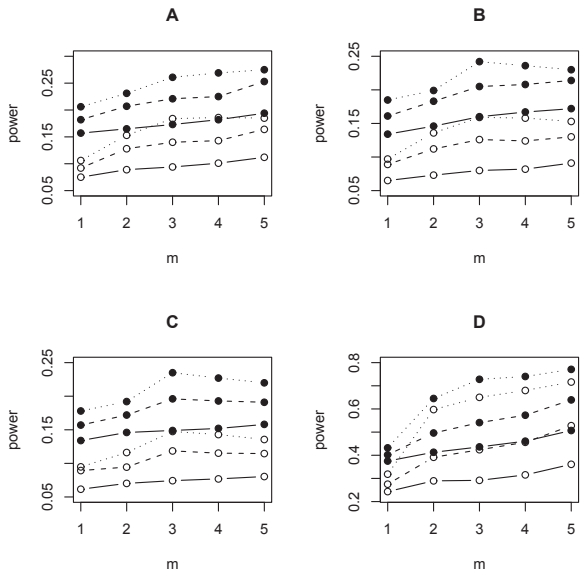

FIGURE 2: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 10$.

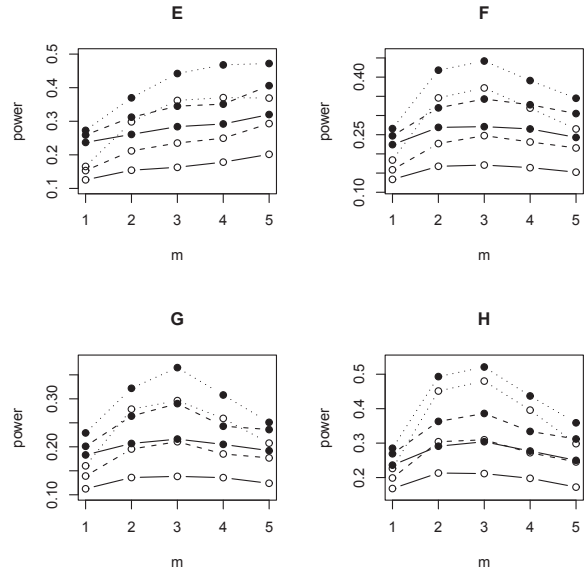

FIGURE 3: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 10$.

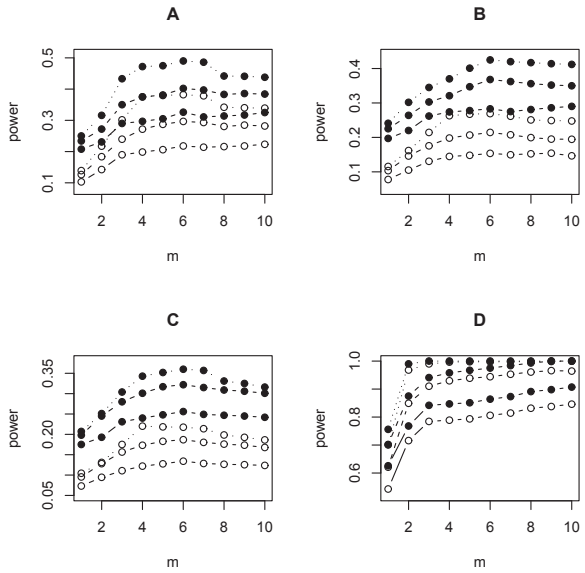

FIGURE 4: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 20$.

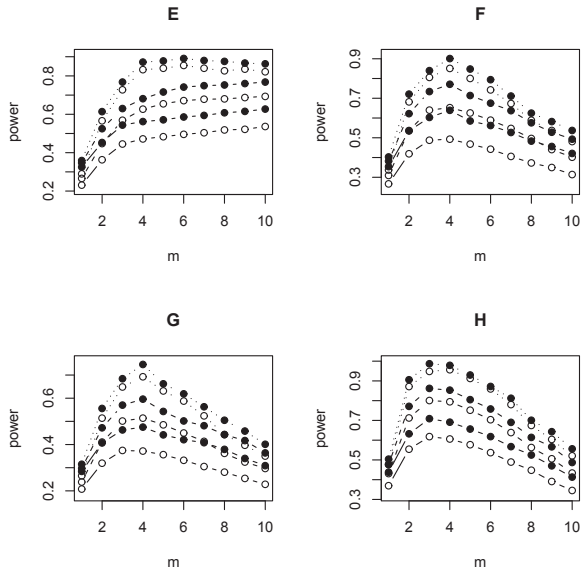

FIGURE 5: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 20$.

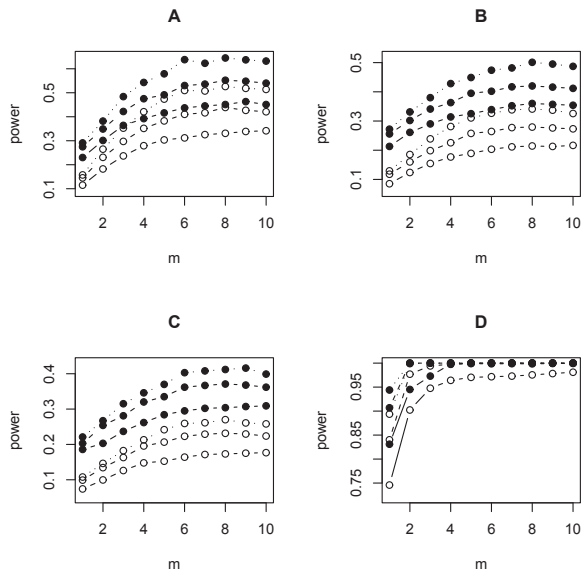

FIGURE 6: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 30$.

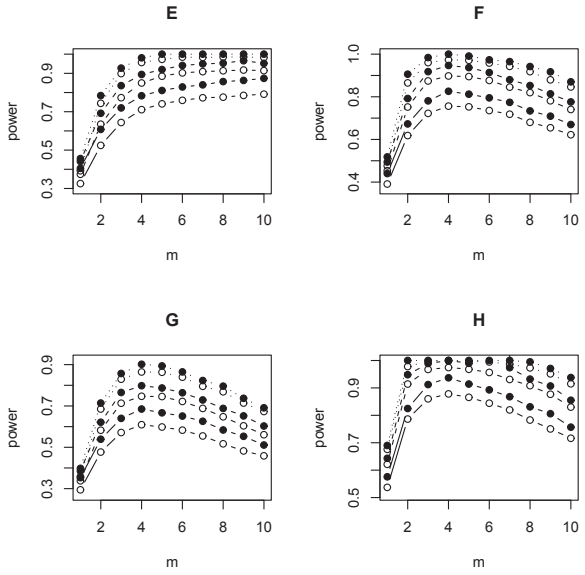

FIGURE 7: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 30$.

It is observed that given a sample size, the entropy tests based on RSS and DRSS are more powerful than that based on SRS regardless of the alternative distribution. More interestingly, the higher sampling effort the more powerful resulting test would be. That is DRSS has the best performance among three considered designs. Remember that a similar trait was reported earlier in the context of entropy estimation. This is fairly expected because the test statistic in each design is constructed based on the corresponding entropy estimator. It should be mentioned that against asymmetric distributions and for each n, maximum power is gained at optimal m or at one of its neighboring values. This trend, however, does not hold for symmetric distributions where maximum power occurs in $m \approx \frac{n}{2}$. Since the best m associated with a sample size varies according to the alternative, we may use a data histogram to decide on the best window size for applying the tests.

It is interesting to examine whether a further increase in power is possible by increasing the number of stages in MSRSS. To this end, testing procedures under MSRSS with $r = 3, 4$ were developed. Figure 8 displays the power of the tests, where alternatives A-H are denoted by integers 1-8 on the X axis, and points corresponding to $r = 2, 3, 4$ are connected by solid, dashed and dotted lines, respectively. Results of DRSS design were included to facilitate comparison. For a given n, the result are provided only for optimal m, not for all $m \in \{1, \ldots, \frac{n}{2}\}$, to save space. From Figure 8, we can see as r increases, some improvement in power happens. Since the differences in results for $r = 2$ and $r = 3, 4$ are not marked for (A)-(C), we may confine ourselves to DRSS against these alternatives.

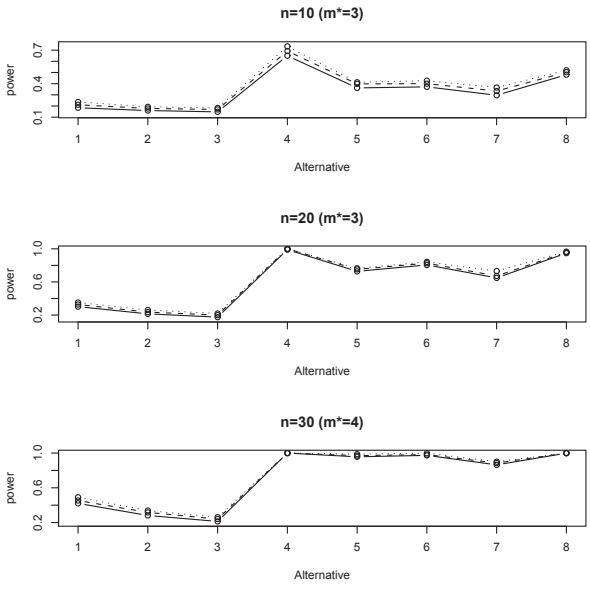

FIGURE 8: Power comparison for the entropy tests of size 0.05 against alternatives A-H under MSRSS designs.

4. Effect of Entropy Estimator

As mentioned before, Vasicek's estimator has been widely used for developing entropy based test of fit. Many authors have modified this test to come up with more efficient estimators. In this section, power behavior of the tests employing such estimators are investigated. To this end, we consider two entropy estimators proposed by Ebrahimi, Pflughoeft & Soofi (1994).

The first estimator which modifies the denominator of (3) is defined as follows

$$E^1_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{X_{(i+m)} - X_{(i-m)}}{c_i m/n}\right) \tag{7}$$

where

$$c_i = \begin{cases} 1 + \frac{i-1}{m} & 1 \le i \le m, \\ 2 & m+1 \le i \le n-m, \\ 1 + \frac{n-i}{m} & n-m+1 \le i \le n \end{cases}$$

The second estimator, obtained by modifying both the numerator and denominator of (3), is given by

$$E^2_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{Z_{(i+m)} - Z_{(i-m)}}{d_i m/n}\right) \tag{8}$$

where

$$d_i = \begin{cases} 1 + \frac{i+1}{m} - \frac{i}{m^2} & 1 \le i \le m, \\ 2 & m+1 \le i \le n-m-1, \\ 1 + \frac{n-i}{m+1} & n-m \le i \le n, \end{cases}$$

the $Z_{(i)}$'s are

$$Z_{(i)} = \begin{cases} a + \frac{i-1}{m}(X_{(1)} - a) & 1 \le i \le m, \\ X_{(i)} & m+1 \le i \le n-m-1, \\ b - \frac{n-i}{m}(b - X_{(n)}) & n-m \le i \le n, \end{cases}$$

and a and b are constants to be determined such that $P(a \le X \le b) \approx 1$. For example, when F has a bounded support, a and b are lower and upper bound, respectively (for uniform(0,1) distribution, $a = 0$ and $b = 1$); if F is bounded below (above), then $a(b)$ is lower (upper) support, $a = \overline{x} - ks$ $(b = \overline{x} + ks)$, where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and k is a suitable number say 3 to 5 (for exponential distribution, $a = 0$ and $b = \overline{x} + ks$); in the case that F has no bound on its support, a and b may be chosen as $a = \overline{x} - ks$ and $b = \overline{x} + ks$.

Simulation results show that both estimators have less bias and less RMSE than Vasicek's estimator (uniformly). Since $E^1_{m,n}$ has simpler form, we focus on

that in the sequel. Simulated biases and RMSEs of $E_{m,n}^{1(r)}$ (The MSRSS analogue of $E_{m,n}^{1}$) for $r = 0, 1, 2$ are given in Figure 1, where a filled circle is used as the plotting symbol, and points corresponding to SRS, RSS and DRSS are connected by solid, dashed and dotted lines, respectively. Again, it is evident that as r increases, $E_{m,n}^{1(r)}$ becomes more efficient. Also, the estimated power of the tests developed using $E_{m,n}^{1(r)}$ for $r = 0, 1, 2$ appear in Figures 2-7 with the same display conventions used for bias and RMSE of the corresponding entropy estimator. In each design, tests based on the new estimator is more powerful than those based on the original estimator for all sample sizes and alternatives.

5. Conclusion

The aim of this paper was to develop goodness-of-fit tests for the Laplace distribution under RSS and MSRSS designs. Motivated by the entropy based test of fit in SRS, we employed the sample entropy based on aforesaid designs to construct the corresponding tests of fit. An extensive simulation study was conducted to provide insight into the finite sample power behavior of the proposed tests. The results indicate that using (multistage) ranked set samples in entropy based test of fit for the Laplace distribution result in higher power as compared with simple random samples. We have developed analogous tests for the uniform, normal, exponential, Weibull and some other distributions using improved entropy estimators whose results will be reported in future articles. Tables of critical points and power of the tests in different designs along with the corresponding computer codes are available on request from the author.

Acknowledgements

We thank the reviewers and the Editor for comments which led to presentational improvements.

[Recibido: noviembre de 2010 — Aceptado: octubre de 2012]

References

Al-Saleh, M. F. & Al-Kadiri, M. (2000), 'Double ranked set sampling', *Statistics & Probability Letters* **48**, 205–212.

Al-Saleh, M. F. & Al-Omari, A. I. (2002), 'Multistage ranked set sampling', *Journal of Statistical Planning and Inference* **102**, 273–286.

Chen, Z., Bai, Z. & Sinha, B. K. (2004), *Ranked set sampling: Theory and Applications*, Springer, New York.

Choi, B. & Kim, K. (2006), 'Testing goodness-of-fit for laplace distribution based on maximum entropy', *Statistics* **40**, 517–531.

Dudewicz, E. J. & van der Meulen, E. C. (1981), 'Entropy-based tests of uniformity', *Journal of the American Statistical Association* **76**, 967–974.

Ebrahimi, N., Pflughoeft, K. & Soofi, E. S. (1994), 'Two measures of sample entropy', *Statistics & Probability Letters* **20**, 225–234.

Gokhale, D. V. (1983), 'On the entropy-based goodness-of-fit tests', *Computational Statistics & Data Analysis* **1**, 157–165.

Grzegorzewski, P. & Wieczorkowski, R. (1999), 'Entropy based goodness-of-fit test for exponentiality', *Communications in Statistics: Theory and Methods* **28**, 1183–1202.

Kotz, S., Kozubowski, T. J. & Podgórski, K. (2001), *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser, Boston, USA.

Mahdizadeh, M. & Arghami, N. R. (2010), 'Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse gaussian law', *Journal of Statistical Computation and Simulation* **80**, 761–774.

McIntyre, G. A. (1952), 'A method of unbiased selective sampling using ranked sets', *Australian Journal of Agricultural Research* **3**, 385–390.

Mudholkar, G. S. & Tian, L. (2002), 'An entropy characterization of the inverse gaussian distribution and related goodness-of-fit test', *Journal of Statistical Planning and Inference* **102**, 211–221.

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

Stokes, S. L. & Sager, T. W. (1988), 'Characterization of a ranked-set sample with application to estimating distribution function', *Journal of the American Statistical Association* **83**, 374–381.

Takahasi, K. & Wakimoto, K. (1968), 'On unbiased estimates of the population mean based on the sample stratified by means of ordering', *Annals of the Institute of Statistical Mathematics* **21**, 249–255.

Vasicek, O. (1976), 'A test of normality based on sample entropy', *Journal of the Royal Statistical Society Series B* **38**, 54–59.