New Trends in Mathematical Sciences
http://www.ntmsci.com

Asymptotic efficiencies of the goodness-of-fit test based on linear statistics

Muhammad Naeem

Deanship of Preparatory Year Program Umm al Qura University, Makkah Mukarramah, KSA

Received: 15 February 2015, Revised: 23 March 2015, Accepted: 2 May 2015
Published online: 9 October 2015

Abstract: A huge literature about the Pitmans asymptotic efficiencies (AE) of the goodness-of-fit tests based on higher-order non-overlapping spacings is available. The performance of the linear random variable based on higher-order non-overlapping spacings is measured asymptotically. Since the linear test satisfy the Cramrs condition so the probability of Large Deviation results are applicable for this statistics. It is observed that just like Pitmans sense, Linear test in intermediate (all three cases) is most efficient as well because it satisfies the above mentioned condition.

Keywords: Asymptotic efficiency, goodness of fit, spacing, cramers condition, large deviation.

1 Introduction

Let $W_1, W_2, ..., W_{n-1}$ be an ordered (ascending form) sample from a population having continuous cumulative distribution function (cdf) F and probability density function (pdf) $f(W)$. The goodness-of-fit problem is to test if this distribution is equal to a specified one. A common approach to these problems is to transform the data via the probability integral transformation $U = F(W)$ so that the support of F is reduced to [0,1] and the specified cdf reduces to that of a uniform random variable on [0,1].Therefore, without loss of generality, one may deal with the problem of testing the null hypothesis .

$$H_0 : f(w) = 1, \quad 0 \leq w \leq 1 \tag{1.1}$$

against alternative that f is a pdf of some other random variable (different from uniform) having support on [0,1]. From now on, we deal with this reduced problem. There are two basic approaches for the goodness-of-fit problem: tests based on observed frequencies and those, based on spacings. It is known that while the tests based on frequencies perform better in detecting differences between the distribution functions, the tests based on spacings are useful to detect differences between the corresponding densities. It is worth noticing too that Jammalamadaka and Tiwari (1987) have shown that comparable test based on α-spacings is better than chi-squared test in terms of local power. With notations $W_0 = 0$ and $W_n = 1$,the non-overlapping α-spacings are defined as $D_j^{(\alpha)} = W_{j\alpha} - W_{(j-1)\alpha}$, $j = 1, 2, ..., N'$, $D_{N'+1}^{(\alpha)} = 1 - W_{N'\alpha}$, where integer $\alpha \ \varepsilon \ [1, n]$, $N' = [n/\alpha]$ is the greatest integer less than or equal to n/α. Consider $N = N'$ if n/α is an integer and $N = N' + 1$ otherwise. Note that all $D_j^{(\alpha)}$ and W_j depend on n also but the extra suffix is omitted for notational simplicity. We are testing hypothesis (1.1) against the sequence of alternatives

$$H_{1,n} : f(w) = 1 + d\, l(w)\delta(n), \quad 0 \leq w \leq 1 \tag{1.2}$$

* Corresponding author e-mail: naeemtazkeer@yahoo.com

where $\delta(n) \to \infty$ as $n \to \infty$, $d > 0$ is a distance between H_1 and H_0 and $l(w)$ is a direction of H_1 such that

$$\int_0^1 l(w)dw = 0, \quad \int_0^1 l^2(w)dw = 1. \tag{1.3}$$

Assume that α may tends to infinity as $n \to \infty$, we consider test based on the statistic

$$\Lambda_N = \sum_{j=1}^{N} \zeta_{jn} D_{jN}^{(\alpha)}, \quad \zeta_{jn} \, \varepsilon \, R, \quad j = 1,2,...N. \tag{1.4}$$

The large value of Λ_N rejects the hypothesis. Tests based on simple spacing, i.e.1-spacings, have been proposed by many authors (see, for example, Pyke (1965) and the references contained therein). Distribution theory of such statistics and their asymptotic efficiencies have been studied, for instance by Rao and Sethuraman (1975), Holst and Rao (1981). For the first time the asymptotic normality of the statistics based on disjoint α -spacings was discussed by Del Pino (1979) and has shown that it is more efficient in Pitman sense than simple spacings statistics, see also, Mirakhmedov and Naeem (2008 a,b). We also refer to a series of papers by Jammalamadaka Rao and co-authors (see, for example, Morgan Kuo and Jammalamadaka (1981) and Jammalamadaka et al (1989). Here statistics (1.4) is called Linear statistics based on α -spacings and it was studied by Holst and Rao (1981) with $\alpha = 1$. The linear statistics belongs to the class of non symmetric statistics based on spacings and it is well known that tests based on such type of statistics can detect alternative (1.2) with $\delta(n) = n^{-1/2}$. The present paper discusses the test of goodness of fit based on (1.4) with $\alpha \geq 1$ which may increase jointly with n. It is shown that the Kallenberg intermediate efficiency coincides with Pitman efficiency of statistic (1.4).

2 Asymptotic normality of Λ_N

In the following discussion, $0 = U_0 \leq U_1 \leq ... \leq U_{n-1} \leq U_n = 1$ be an ordered sample from uniform [0,1] distribution and $T_j^{(\alpha)}$ their non-overlapping α-spacings. Let $h_j(w,u)$, where $u = j/N$ and $j = 1,2,...,N$, be measurable functions. Consider the statistics

$$R_N = \sum_{j=1}^{N} h_j(NT_j^{(\alpha)}, u) \tag{2.1}$$

Let X and $X_{J,\alpha}$, $\alpha = 1,2,...,N$ be independent and identically distributed random variables with common density function $\gamma_\alpha(t) = t^{\alpha-1}e^{-t}(\Gamma(\alpha))^{-1}$, $t > 0$, where $\Gamma(\alpha)$ is well known gamma function. We supose that the following moments exists

$$S_{N,\alpha} = X_{1,\alpha} + ... + X_{N,\alpha},$$

$$Q_N = \sum_{j=1}^{N} h_j(X_{N,\alpha}, j/N),$$

$$\rho_N = corr(Q_N, S_{N,\alpha}),$$

$$f_j(t,u) = h_j(t,u) - E(h_j(X,u)) - (t-\alpha)\rho_N \sqrt{\frac{VarQ_N}{N\alpha}}, \tag{2.2}$$

$$A_N = \sum_{j=1}^{N} E(h_j(X_{j,\alpha}, u)).$$

$$\sigma_N^2 = Var(f_j(X_{j,\alpha}, u)).$$

The following Assertion is the Corollary 2 of Mirakhmedov (2005).

Assertion: If $\frac{1}{\sigma_n^3}\sum_{j=1}^{n} E\left|f_j(X_{j,\alpha}, u)\right|^3 \to 0$, as $N \to \infty$ and $u = j/N$ then the random variable R_N has asymptotically normal distribution with expectation A_N and variance σ_N^2. The statistic Λ_N is a special case of (2.1) with

$$h_j(w, y) = \zeta(y)w.$$

Therefore, from Assertion , by putting, $\Psi_{m,N} = \frac{1}{N}\sum_{j=1}^{N} \zeta_{j,n}^m$ and $\Upsilon_{m,N} = \frac{1}{N}\sum_{j=1}^{N}(\zeta_{j,n} - \Psi_{1,N})^m$ we have the following theorem,

Theorem 2.1. If $(\Upsilon_{4,N}/N\Upsilon_{2,N}^2) \to \infty$ as $N \to \infty$ then the random variable Λ_N has asymptotically normal distribution with expectation $n\Psi_{1,N}$ and variance $n\Upsilon_{2,N}$.

Proof. The r. v. Λ_N is included in the family of statistic mentioned in (2.1) with kernel function $h_j(u) = \zeta_{j,N}u$. It is well known that
$E(X)^s = \alpha(\alpha+1)...(\alpha+s-1)$, $s \geq 1$; $E(X) = Var(X) = \alpha$

$$E(X-\alpha)^4 = 3\alpha(\alpha+2). \tag{2.3}$$

By using these in the notation (2.2) we have

$$E(h_j(X, j/N)) = \zeta_{j,N}\alpha,$$

$$Var(Q_N) = \alpha N\Psi_{2,N},$$

$$\rho_N = \Psi_{1,N}/\sqrt{\Psi_{2,N}},$$

$$f_{j,N}(X_{j,\alpha}) = (\zeta_{j,N} - \Psi_{1,N})(X_{j,\alpha} - \alpha),$$

$$\sigma_N^2 = \left(1 - \frac{\Psi_{1,N}^2}{\Psi_{2,N}}\right)\alpha N\Psi_{2,N} = \alpha N(\psi_{2,N} - \Psi_{1,N}^2) = \alpha N\Upsilon_{2,N} = n\Upsilon_{2,N}. \tag{2.4}$$

Now by using (2.3)

$$\sum_{j=1}^{N} Ef_j^4(X_{j,\alpha}) = N\Upsilon_{4,N}E(X_{j,\alpha}-\alpha)^4 = 3\alpha(\alpha+1)N\Upsilon_{4,N}. \tag{2.5}$$

By applying relations (2.4), (2.5) in the above mentioned Assertion, and well known inequality $\beta_{3,N} \le \beta_{4,N}^{1/2}$ one can get

$$\left|P\{\Lambda_N < x\sqrt{n\Upsilon_{2,N}} + n\Psi_{1,N}\} - \Phi(w)\right| \le C\left[\frac{3}{N}\left(1+\frac{1}{\alpha}\right)\frac{\Upsilon_{4,N}}{\Upsilon_{2,N}^2}\right]^{1/2}. \tag{2.6}$$

It proves Theorem 2.1.

Remark 2.1. Actually in (2.6) we have a more general result as compared to Theorem 2.1, namely estimation of the remainder term in central limit theorem for Λ_N.

Remark 2.2. The r. v. Λ_N coincides with linear combination of order statistics, in our case, with sample from uniform (0, 1) distribution. As usually, consider a r. v. of type Λ_N , it is assumed that

$$\sum_{j=1}^{N} \zeta_{j,n} = 0, \quad \sum_{j=1}^{N} \zeta_{j,n}^2 = 1. \tag{2.7}$$

This condition does not lose generality because otherwise, instead of $\zeta_{j,n}$, one can take $\xi_{j,n} = (\zeta_{j,n} - \Psi_{1,N})/\sqrt{\Upsilon_{2,N}}$. Under condition (2.7) if, additionally, $\Upsilon_{4,N} \to \infty$ as $N \to \infty$ then r. v. Λ_N has asymptotically normal distribution with zero expectation and variance α .

3 Probability of large deviation.

The Cramer's condition: there exists $\Pi > 0$ such that $E\exp\{\Pi\,|g(W,u)|\} < \infty$ where $u = j/N$, $j = 1,2,...,N$, is obviously satisfied by the statistic Λ_N . Therefore, by the Theorem, Mirakhmedov et. al. (2011), it follows

Theorem 3.1. For all $w \ge 0$, $w = o(\sqrt{n})$

$$P_0\{\Lambda_N \ge x\sqrt{n\Upsilon_{2,N}} + n\Psi_{1,N}\} = \Phi(-w)\exp\left\{-\frac{w^3}{\sqrt{N}}K_N\left(-\frac{w}{\sqrt{N}}\right)\right\}\left\{1+O\left(\frac{w+1}{\sqrt{N}}\right)\right\},$$

where $K_N(u) = \kappa_{0,N} + \kappa_{1,N}u + ...$ is a special Cramer's type power series for N large enough and $\kappa_{j,N} \le \Pi < \infty$, $j = 0,1,2,....$

4 Asymptotic relative efficiency

From Theorem 2.1 and well known theorem on convergence of moments (see, for example Theorem 6.14 by Moran (1948b) it follows that

$$E_1\Lambda_N = N\Psi_{1,N} + o(N) \;\; and \;\; Var_1\Lambda_N = N\Upsilon_{2,N}^2(1+o(1)), \tag{4.1}$$

Let P_j, E_j and Var_j are probability, expectation and variance accounted under H_j, j=0,1. We assume that $E_1\Lambda_N - E_0\Lambda_N > 0$. For non-symmetric statistics,

$$w_N(g) \equiv \sqrt{N}(E_1\Lambda_N - E_0\Lambda_N)/\sqrt{Var_1\Lambda} = -\sqrt{n}\delta(n)d\ \rho(g)(1+o(1)),$$

with

$$\rho(g) = \int_0^1 l(u)corr(g(W,u),W)du \tag{4.2}$$

because $Var(W) = \alpha$. Due to Holder's inequality $\rho(g) \le 1$ since (1.3) and equality can be achieved if and only if $g(y,t) = l(t)y$. Therefore, non-symmetric tests discriminate alternatives H_1 (1.2) with $\delta(n) = n^{-1/2}$. Asymptotically optimal test in Pitman's sense is a linear test, which rejects H_0 if

$$\Lambda_N = \sum_{j=1}^{N} \zeta_{j,N} D_{j,N}^{(\alpha)} > \Pi.$$

This result was also obtained by Rao and Sethuraman (1975) and Holst and Rao (1981) for $\alpha = 1$. Asymptotically most powerful test among non symmetric tests is a linear test based on linear statistic $\Lambda_N = \sum_{j=1}^{N} l(j/N)D_{j,N}^{(\alpha)}$ critical region of which is given by $\{w : w \ge t_\omega\sqrt{n}\}$ and asymptotic power is $\Phi(d - t_\omega)$ and it can be applied to detect alternatives with given direction $l(w)$, because all statistical characteristics of this test depend on $l(w)$. The Linear test, being non symmetric, can detect alternatives at a distance $\sqrt{n}$ and sequence of alternatives, (1.2) with $\delta(n) = n^{-1/2}$, are called Pitman's alternatives. The Pitman's alternatives, when the sequence of alternatives $H_{1,n}$ converges with a rate $\delta(n) = n^{-1/2}$, is one of extreme cases. Another extreme case arises in Bahadur's concept which proposes that alternative is fixed i.e.$H_{1,n}$ does not approach to null hypothesis. One can say, there seems no need to use statistical methods in the case of alternatives far from the null hypothesis. Between these extreme cases there is intermediate approach. For our case intermediate alternatives determine in (1.2) $\delta(n)$ such that

$$\delta(n) \to 0\ ,\ \ \sqrt{n}\delta(n) \to \infty. \tag{4.3}$$

These situations give rise to the concepts of intermediate asymptotic efficiency (IAE) due to Kallenberg (1983), see also, Inglot (1999), Ivchenko and Mirakhmedov (1995). According to the classification of Kallenberg (1983) this efficiency will be weak $\omega - IAE$ if $\sqrt{n}\delta(n) = O\left(\sqrt{\ln n}\right)$, middle $\omega - IAE$ if $\sqrt{n}\delta(n) = o\left(n^{1/6}\right)$ and considered to be strong $\omega - IAE$ when $\sqrt{n}\delta(n) = o\left(n^{1/6}\right)$. Asymptotic relative efficiency of one test is defined as the ratio of its asymptotic slopes. For Pitman's alternatives it is equal to Pitman's asymptotic relative efficiency, see for example Fraser (1957). We have,

Theorem 4.1. Let alternative H_1 be specified by (1.2) and (4.3). If $\int_0^1 E\exp\{\Pi|g(W,t)|\}\,dt < \infty$ for some $\Pi > 0$, then $\frac{e_n^\omega(g)}{n\delta^2(n)} = \frac{d^2}{2}\rho^2(g)(1+o(1))$.

Proof. Following Mirakhmedov (2006), under the conditions of the theorem, we have $e_n^\omega(g) = -\log\Phi(-w_N(g)) + o\left(w_N^2(g)\right)$. Since $-\log\Phi(-w) = \frac{1}{2}w^2(1+o(1))$ as $w \to \infty$, from (4.2) we have $e_n^\omega(g) = \frac{d^2}{2}n\delta^2(n)\rho^2(g)(1+o(1))$. The Theorem 4.1 is proved.

It follows from Theorem 4.1 that $\rho^2(g)$ should be taken as a measure of IAE of the g-test. Thus, for intermediate alternatives (1.2) and (4.3) with direction $l(w)$ within the class of non-symmetric tests, the linear test (based on statistics Λ_N) is most efficient in the IAE sense, since it satisfies condition of Theorem 4.1. Due to Kallenberg's classification one can say that the linear test is optimal in the strong IAE sense.

5 Conclusions

1.We shall measure the performance of Linear test by the asymptotic value of slope $e_N^{\omega}(\Lambda_N) = -\log P_0\{\Lambda_N \geq N\Psi_{1,N} + o(N)\}$.

2.Since the Cramer's condition, for Λ_N , is satisfied so the probability of Large Deviation results are applicable for this statistics.

3.Just like Pitman's sense, Linear test is most efficient in intermediate (all three cases) because it satisfies the conditions of Theorem 4.1.

References

[1] Del Pino,G.E.(1979). On the asymptotic distribution of k-spacings with applications to goodness-of-fit tests. Ann. Statist.,7, 1058-1065.

[2] Fraser, D.A.S. (1957). Nonparametric methods in Statistics. John Wiley, New York.

[3] Inglot, T. (1999). Generalized intermediate efficiency of goodness-of-fit tests. Math. Methods Statist., 8, 487-509.

[4] Ivchenko, G.I. and Mirakhmedov M.A. (1995). Large deviations and intermediate efficiency of decomposable statistics in a multinomial scheme. Math. Methods of Statistics.,4(3), 294-311.

[5] Holst, L. and Rao, J.S. (1981). Asymptotic spacings theory with applications to the two sample problem. Canadian J. Statist. 9,603-610.

[6] Jammalamadaka,S, R. and Tiwari,R. C. (1987). Efficiencies of some disjoint spacing tests relative to a chi- square test. In Perspectives and New Directions in Theoretica and Applied Statistics l(Madan Puri, J.P. Valaplana, and Wolgang Wertz), John Wiley, 311-317.

[7] Jammalamadaka S. R. Zhou X. and Tiwari R.C. (1989). Asymptotic efficiencies of spacings tests for goodness of fit, Metrika, 36, 355-377.

[8] Kallenberg,W.C.M. (1983). Intermediate efficiency. Ann. Statist.; 11, 170-182.

[9] Mirakhmedov, Sh.M. (2005). Lower estimation of the remainder term in the CLT for a sum of the functions of k spacings. Statist. and Probab. Letters. 73, 411-424.

[10] Mirakhmedov Sh.M. (2006). Probability of large deviations for the sum of functions of spacings. Inter. J. Math. and Math. Sciences. V., Article ID 58738, 1-22.

[11] Mirakhmedov, Sh. M. and Muhammad Naeem (2008 a). Asymptotic Properties of the Goodness -Of -Fit Tests Based on Spacings. Pak. J. statistics. Vol. 24(4), 253-268.

[12] Mirakhmedov Sh.M. and Naeem, M. (2008 b). Asymptotical Efficiency of the Goodness of Fit Test Based on Extreme k-spacings Statistic. J. Appl. Probabl. Statist. Dixie W publishing corporation, USA, V. 3. No.1, 65-75.

[13] Mirakhmedov Sh.M., Tirmizi, S.I. and Naeem M. (2011). Cramer-Type large deviation theorem for the sum of functions of higher ordered spacings. Metrika 74 (1),33-54. .

[14] Moran, P. A. P. (1948b). Some theorems on time series II. The significance of the serial correlation coefficient. Biometrika 35, 25560.

[15] Morgan, K. and Jammalamadaka S.R.(1981). Limit theory and efficiencies for tests based on higher ordered spacings. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions., Calcutta, 333-352.

[16] Pyke, R. Spacings. j.Roy. stat. Soc.Ser. B 27, 395-449(1965).

[17] Rao, J.S. and Sethuraman J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. Ann. Statist. 3,299-313.