Ann. Inst. Statist. Math.
38 (1986), Part A, 485-493

RELATIVE EFFICIENCIES OF GOODNESS OF FIT PROCEDURES FOR ASSESSING UNIVARIATE NORMALITY

JAMES A. KOZIOL

(Received Sept. 13, 1984; revised Feb. 25, 1985)

Summary

Efficiency properties of the Cramér-von Mises, Anderson-Darling, Watson, and DeWet-Venter statistics for assessing normality are investigated. For these statistics, the approximate slopes are determined, and the equivalence of ratios of limiting approximate slopes to limiting Pitman efficiencies is established. From relative efficiency comparisons, the Cramér-von Mises and Watson statistics perform rather poorly; choice between the Anderson-Darling and DeWet-Venter statistics should be made on the basis of anticipated alternatives.

1. Introduction

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent, identically distributed random variables with underlying continuous cumulative distribution function F. We wish to test the null hypothesis that the X_i are normally distributed, that is,

$$(1.1) \qquad H_0:\ F(x)=\Phi(x-\mu/\sigma)\ , \qquad -\infty<\mu<\infty,\ \sigma>0\ ,$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, and μ and σ are unknown.

A number of omnibus tests have been proposed for assessing the composite hypothesis (1.1). The Shapiro-Wilk statistic and its variants (Shapiro and Wilk [19], Shapiro and Francia [18], Filliben [12]) are essentially correlation-type statistics: one would reject the null hypothesis (1.1) for sufficiently small values of the correlation between the ordered sample and the corresponding percentiles (or, expected order statistics) of the standard normal distribution, for example. Versions of the quadratic goodness-of-fit statistics Cramér-von Mises, Anderson-Darling, and Watson have also been proposed for testing (1.1); these

Key words and phrases: Bahadur efficiency, Pitman efficiency, Anderson-Darling, Cramér-von Mises, DeWet-Venter, Shapiro-Francia, Watson.

are described in some detail in Pearson and Hartley [15].

We shall focus attention on these three quadratic procedures, and on the DeWet-Venter statistic (DeWet and Venter [7]), a statistic closely related to that of Shapiro and Francia. In the composite hypothesis setting (1.1) wherein the functional form of the cumulative distribution is assumed known but parameters are estimated, these various procedures are not nonparametric in nature, but have limiting null distributions that depend explicitly on the assumption of normality. (The limiting distributions are nevertheless parameter-free, however.) We refer the reader to DeWet and Venter [7], [8], Durbin [9], [10], Durbin, Knott, and Taylor [11], and Stephens [22] for results pertaining to asymptotic distribution theory of the procedures.

Only limited comparisons have been made among these test criteria, on the basis of simulation studies. Shapiro, Wilk, and Chen [20] claimed that the Shapiro-Wilk statistic provides a generally superior omnibus measure of non-normality, but their conclusion was later somewhat discounted by Stephens [21]. Pearson, D'Agostino, and Bowman [16] have also undertaken a rather extensive Monte Carlo power study, of both omnibus tests and directional tests.

The purpose of this paper is to effect a comparison among the statistics using the criterion of Bahadur efficiency. In Section 2, we show that the goodness of fit procedures may be related to Bahadur's "standard sequences" of test statistics, and that the approximate slopes of these sequences are readily calculable. In Section 3, we note that the conditions stipulated by Wieand [24] are satisfied in our particular setting, thereby allowing us to conclude that limiting Bahadur efficiencies (that is, ratios of slopes) are equivalent to limiting Pitman efficiencies. We use this fact in order to reexamine the skewness and kurtosis alternatives to normality originally investigated by Durbin, Knott and Taylor [11]. Not surprisingly, our asymptotic calculations are in accord with the findings of Durbin, Knott and Taylor, of Stephens, and of Pettitt [17]: the Shapiro-Francia statistic, which places heaviest emphasis on tail behavior, is most sensitive to alternatives wherein departures from normality are most pronounced in the tails; the Anderson-Darling statistic is intermediate, being sensitive not only to those alternatives previously mentioned, but also to alternatives to normality determined by behavior in the central part of the range; both of these statistics tend to outperform the Cramér-von Mises and Watson statistics.

2. The test statistics and their approximate slopes

Given the sample $X_1, X_2, \cdots, X_n$, let $Y_i=(X_i-\bar{X}_n)/s_n$, $i=1, 2, \cdots, n$, where $\bar{X}_n = n^{-1}\sum_{j=1}^{n} X_j$ and $s_n^2 = n^{-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$. Let $F_n(\cdot)$ and $Q_n(\cdot)$ denote the empirical distribution function and the quantile function respectively of the Y_i:

$$F_n(Y) = n^{-1}\sum_{j=1}^{n} I(Y_i \leqq Y)\,, \qquad -\infty < Y < \infty\,,$$

$$Q_n(t) = Y_{(k)}^n \qquad \text{if } \frac{k-1}{n} < t \leqq \frac{k}{n},\ k=1, 2, \cdots, n,\ 0 < t \leqq 1\,.$$

Here, $I(\cdot)$ is the usual indicator function, and $Y_{(1)}^n < Y_{(2)}^n < \cdots < Y_{(n)}^n$ are the order statistics of the Y_i. We shall investigate the relative performances of four goodness of fit statistics for assessing the null hypothesis (1.1). These four statistics are:

(i) the Cramér-von Mises statistic

$$W_n^2 = n\int_{-\infty}^{\infty} [F_n(Y) - \Phi(Y)]^2 d\Phi(Y)\,;$$

(ii) the Anderson-Darling statistic

$$A_n^2 = n\int_{-\infty}^{\infty} [F_n(Y) - \Phi(Y)]^2/[\Phi(Y)(1 - \Phi(Y))] d\Phi(Y)\,;$$

(iii) the Watson statistic

$$U_n^2 = n\int_{-\infty}^{\infty}\left[F_n(Y) - \Phi(Y) - \int_{-\infty}^{\infty}[F_n(Y) - \Phi(Y)]d\Phi(Y)\right]^2 d\Phi(Y)\,;$$

(iv) the DeWet-Venter statistic $D_n^2 = L_n - a_n$, where

$$L_n = n\int_0^1 [Q_n(t) - \Phi^{-1}(t)]^2 dt\,,$$

and a_n is a centering constant computed by them ($a_n = O(\log n)$).

The asymptotic null distributions of W_n^2, A_n^2, and U_n^2 are described by Stephens [22], and that of D_n^2 is given by DeWet and Venter [7]; however, few facts are known concerning their asymptotic power properties. Indeed, such power comparisons are complicated by the lack of methodology for direct Pitman efficiency calculations.

We shall, therefore, compare the asymptotic performances of these various goodness of fit statistics against particular classes of alternatives, by using a criterion of efficiency introduced by Bahadur [2]. Bahadur considers the situation in which the probability distribution of X_i is determined by a parameter θ which takes values in a set Θ. It is required to test the null hypothesis that some θ in Θ_0 obtains, where Θ_0 is a given subset of Θ. For each n, let T_n be a test statistic such that large values of T_n are significant. Suppose that T_n has an asymptotic null distribution, that is, there exists a probability distribution function G such that, for each θ in Θ_0, $P_\theta(T_n<t)=G_n(t,\theta)\to G(t)$ as $n\to\infty$ for each t. For given $s=(x_1,x_2,\cdots,x_n)$, the approximate level attained by T_n is defined as $L_n(s)=1-G[T_n(s)]$. The rate at which L_n tends to zero when a given nonnull θ obtains is regarded by Bahadur as a measure of the asymptotic efficiency of the sequence of test statistics $\{T_n\}$ against that θ. If for each nonnull θ there exists a $c(\theta)$, $0<c<\infty$, such that $n^{-1}\log L_n\to -c(\theta)/2$ as $n\to\infty$ with probability one when θ obtains, the value $c(\theta)$ is called the approximate slope of $\{T_n\}$. Given two sequences of test statistics $\{T_n^{(1)}\}$, $\{T_n^{(2)}\}$ with approximate slopes $c^{(1)}(\theta)$, $c^{(2)}(\theta)$ respectively, the ratio $c^{(1)}(\theta)/c^{(2)}(\theta)$ is known as the approximate Bahadur efficiency of $\{T_n^{(1)}\}$ compared with $\{T_n^{(2)}\}$. The theory of approximate slopes and of a related concept, that of exact slopes, is discussed more extensively in Bahadur [2], [3], [4].

It is in general a nontrivial problem to determine the approximate slope of a given sequence $\{T_n\}$. One useful method was described by Bahadur [2], who defined $\{T_n\}$ to be a standard sequence if the following three conditions are satisfied:

(i) T_n has an asymptotic null distribution, G, which is continuous;

(ii) there exists a constant a, $0<a<\infty$, such that

$$\log[1-G(t)]=-(at^2/2)[1+o(1)]\ , \qquad \text{as } t\to\infty\ ;$$

(iii) there exists a real-valued function $b(\theta)$ on $\Theta-\Theta_0$ with $0<b(\theta)<\infty$, such that, for each θ in $\Theta-\Theta_0$,

$$\lim_{n\to\infty} P_\theta\{|T_n/n^{1/2}-b(\theta)|>t\}=0 \qquad \text{for all } t>0\ .$$

The approximate slope of the standard sequence $\{T_n\}$ is then

$$c(\theta)=ab^2(\theta)\ .$$

We shall apply this method for the determination of the approximate slopes of the goodness of fit statistics, by showing that $\{W_n\}$, $\{A_n\}$, $\{U_n\}$, and $\{D_n\}$ are all standard sequences; here, $D_n=\mathrm{SGN}(L_n-a_n)|L_n-a_n|^{1/2}$. First, note that W_n^2, A_n^2, and U_n^2 are asymptotically distributed as $\sum_{i=1}^\infty \lambda_i Z_i^2$, where $Z_1, Z_2,\cdots$ are i.i.d. $N(0,1)$ random variables, and $\lambda_1>\lambda_2>\cdots>0$ are the eigenvalues devolving from the integral equations associated with their respective covariance kernels; furthermore, this sum converges in mean square and with probability one (Durbin [9]). Similarly, D_n^2 is asymptotically distributed as $\sum_{i=1}^\infty \lambda_i(Z_i^2-1)$ (DeWet and Venter [7]). It follows that each of the statistics W_n, A_n, U_n, and D_n has a continuous limiting distribution under (1.1); hence we turn to the determination of the large deviation probabilities.

Under the circumstances just described, the following large deviation result obtains (Zolotarev [25]; Hoeffding [14]; Abrahamson [1]; Beran [5]):

$$\log P\left(\sum_{i=1}^{\infty} \lambda_i Z_i^2 > t\right) = -(t/2\lambda_1)[1+o(1)] \qquad \text{as } t \to \infty .$$

Clearly, this result remains true with the Z_i^2 centered about their expectation.

It is now straightforward to find the approximate slopes of the standard sequences $\{W_n\}$, $\{A_n\}$, $\{U_n\}$, and $\{D_n\}$ under various alternatives. In this regard, we remark that Stephens [22] describes numerical techniques for the computation of the eigenvalues λ_i associated with the statistics W_n^2, A_n^2, and U_n^2, and indeed calculates the requisite λ_1 values; DeWet and Venter [7] explicitly provide the λ_i for D_n^2. Suppose under a particular alternative, $F_n(\cdot) \to G_\theta(\cdot)$. Then the approximate slopes $c(\theta)$ of the standard sequences may be determined from Table 1. In the following section we carry out this evaluation numerically for various alternatives approaching the null case, and relate these values to limiting Pitman efficiencies.

Table 1. Values of a and $b^2(\theta)$ for calculation of approximate slopes

Statistic	a	$b^2(\theta)$
W_n	54.53	$\int_{-\infty}^{\infty} [G_\theta(y)-\Phi(y)]^2 d\Phi(y)$
A_n	10.17	$\int_{-\infty}^{\infty} [G_\theta(y)-\Phi(y)]^2/[\Phi(y)(1-\Phi(y))] d\Phi(y)$
U_n	63.57	$\int_{-\infty}^{\infty} [G_\theta(y)-\Phi(y)]^2 d\Phi(y) - \left\{\int_{-\infty}^{\infty} [G_\theta(y)-\Phi(y)] d\Phi(y)\right\}^2$
D_n	3.	$\int_0^1 [G_\theta^{-1}(t)-\Phi^{-1}(t)]^2 dt$

3. Comparison of efficiencies under local alternatives

Bahadur [3] emphasizes that the most important property of a slope is its value in the immediate vicinity of the null hypothesis. Because the approximate slope and the exact slope of a test sequence typically coincide in a neighborhood of the null parameter, the main conclusions relevant to power considerations available from exact slopes (cf. Bahadur [4]) also pertain to approximate slopes. Also of relevance is Bahadur's comment [2] that in one-sided testing problems, the limiting approximate Bahadur efficiency of two asymptotically normal test sequences as the alternative parameter converges to the null value coincides with their limiting Pitman efficiency as the alpha level approaches zero. Wieand [24] has generalized Bahadur's remark to include test sequences with asymptotic distributions other than normal, and those used in two-sided testing problems. In particular, for the simple goodness of fit problem, Wieand computes the limiting approximate slopes of the square roots of the Cramér-von Mises and Watson statistics against location and scale alternatives; since these statistics satisfy his Condition III*, Wieand can invoke his theorem equating limiting Bahadur efficiency to limiting Pitman efficiency for these statistics (see also Wieand [23], for further details).

That the square root of the simple Anderson-Darling statistic also satisfies Wieand's Condition III* for appropriate alternatives follows from Gregory ([13], Theorems 4.1 and 3.1). Further, it can be shown that Condition III* holds for the version of the goodness of fit statistic with estimated parameters given that it is satisfied by the simple statistic, since limiting distributions do not depend on values of the parameters, and rates of convergence remain unaltered with asymptotically efficient estimates. Similarly, it is not difficult to show that the statistic D_n also satisfies Wieand's Condition III*. We conclude, therefore, that Wieand's theorem remains true for the goodness of fit statistics with particular alternatives considered here.

Following Durbin, Knott, and Taylor [11], we consider two classes of alternatives. The first class is based on an Edgeworth series for the density $g_\theta(\cdot)=G'_\theta(\cdot)$, specifically,

$$g(y;\theta_1,\theta_2)=\phi(y)\left[1+\frac{1}{6}\theta_1H_3(y)+\frac{1}{24}\theta_2H_4(y)\right], \tag{3.1}$$

where $\phi(\cdot)=\Phi'(\cdot)$, the standard normal density function, and $H_j(\cdot)$ is the j-th Hermite polynomial. As Durbin, Knott, and Taylor note, nonzero values of θ_1 and θ_2 indicate departures from normality characterized by skewness and kurtosis respectively, which are heavily dominated by behavior in the tails. (Clearly, the nonpositivity of $g(y;\theta_1,\theta_2)$ for certain values of θ_1 and θ_2 preclude it from representing a density function globally. Nevertheless, the results in this section concerning g remain valid upon restricting its range to where positive and renormalizing as needed, and careful attention to limiting arguments.)

The second class of alternatives is specified by

$$G(y;\theta_3,\theta_4)=\Phi(y)+\theta_3\sin[3\pi\Phi(y)]+\theta_4\sin[4\pi\Phi(y)]. \tag{3.2}$$

Here, nonzero values of θ_3 and θ_4 indicate skewness and kurtosis—like departures from normality respectively, but the departures should be determined more by behavior in the central part of the range than (3.1).

In Table 2 we list values of $\lim_{\theta\to\theta_0}[c(\theta)/\theta^2]$ for the four goodness-of-fit statistics; we consider four alternatives, obtained from allowing one

Table 2. Limiting values of $c(\theta)/\theta^2$ for the goodness of fit criteria against sine and Edgeworth alternatives

Statistic	Sine alternative (3.2)		Edgeworth alternative (3.1)	
	Shift in skewness ($\theta_4=0$)	Shift in kurtosis ($\theta_3=0$)	Shift in skewness ($\theta_2=0$)	Shift in kurtosis ($\theta_1=0$)
W_n	27.27	27.27	3.34	7.79
A_n	35.76	38.68	3.95	9.82
U_n	28.92	30.17	2.63	9.09
D_n	27.78	33.41	6.0	18.0

nonzero θ in either (3.1) or (3.2). Recall that, the ratio of any two values in a column represents a limiting (as $\theta \to \theta_0$) approximate Bahadur efficiency, which by application of Wieand's theorem is also a limiting (as $\alpha \to 0$) Pitman efficiency.

Note that W_n and U_n are rather similar in terms of relative efficiency, with perhaps W_n to be preferred for skew alternatives and U_n for kurtic alternatives. However, this issue is moot, since each statistic is convincingly dominated by A_n. Note also the strikingly good performance of D_n against Edgeworth alternatives, where it is clearly the statistic of choice. On the other hand, the Anderson-Darling statistic dominates it against sine alternatives, where tail behavior is of less prominence. These findings are thus complementary to, and congruent with, those of Pettitt [16], who examined the relative performances of these procedures at alpha levels more typically encountered in practice. On the basis of these studies, we would recommend the Anderson-Darling statistic as an omnibus procedure: in the absence of prior information concerning the alternative of interest, it exhibits relatively good performance against a wide range of departures from normality.

We conclude with the remark that the notion of Bahadur efficiency can shed considerable light on the relative performances of goodness of fit statistics in simple (parameters known) versus composite (parameters estimated) hypothesis testing problems. In the present context, the simple gof hypothesis $H_s: F=\Phi$ (*no* estimated parameters) ought to be distinguished from the composite null hypothesis H_c given in (1.1). The former hypothesis might be tested with the Cramér-von Mises statistic

$$V_n^2 = n \int [F_n(x) - \Phi(x)]^2 d\Phi(x) ,$$

where F_n here denotes the empirical distribution function of the original sample $X_1, X_2, \cdots, X_n$; and the latter, with W_n^2. We shall now use Bahadur efficiency to assess the relative merits of V_n^2 and W_n^2 under

various alternatives. Consider, then, the following possibilities:
(i) If H_s obtains, then $b^2(\theta)=0$ for both V_n^2 and W_n^2, and both can provide level-α tests when compared with their appropriate null distributions.
(ii) If F is normal, but with either nonzero mean or nonunit variance (or both), then $b^2(\theta)>0$ for V_n^2, and V_n^2 will asymptotically reject H_s with probability 1. V_n^2 is an omnibus statistic here, but W_n^2 remains a level-α test of H_c, and its corresponding value of $b^2(\theta)$ is zero.
(iii) Most interestingly, suppose F is a non-normal distribution. Clearly, it is possible to envisage alternatives for which $b^2(\theta)$ is identical for $\{V_n\}$ and $\{W_n\}$. (A key fact here is the observation that, from the construction of a standard sequence of gof test statistics—e.g. $\{W_n/\sqrt{n}\}$ —the insertion of an $O_p(1)$ estimator for a parameter value does not alter the limiting value of $b^2(\theta)$.) However, since the null distribution of V_n^2 (under H_s) is stochastically larger than that of W_n^2 (under H_c), it has a smaller large deviation probability (the "a" term) than W_n^2, and hence will have a smaller slope than W_n^2 against these particular alternatives. That is, the limiting Bahadur efficiency (equivalently, limiting Pitman efficiency) of V_n^2 relative to W_n^2 is less than one. This provides theoretical support for an empirical finding of Stephens [21]: namely, that against certain alternatives to normality, V_n^2 can be considerably less powerful than W_n^2. Professor Stephens concludes that "it is better *not* to have the true mean and variance available but to estimate it from the data", in these circumstances. Thus, although one might argue that the goals and purposes of assessing H_s and H_c may not be altogether congruent, both theoretical and empirical evidence suggest that closer attention be paid to the relative merits of statistics derived for one hypothesis but used in the other setting.

Acknowledgments

This research was supported in part by grants K04-CA00687 and R01-CA26666 from the U.S. National Cancer Institute. The author thanks the editor and referee for their helpful remarks.

RADIATION EFFECTS RESEARCH FOUNDATION*

REFERENCES

[1] Abrahamson, I. G. (1965). On the stochastic comparison of tests of hypotheses, Unpublished doctoral dissertation, University of Chicago.

[2] Bahadur, R. R. (1960). Stochastic comparison of tests, *Ann. Math. Statist.*, **31**, 276-295.

* Now at Scripps Clinic and Research Foundation, La Jolla, California.

[3] Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics, *Ann. Math. Statist.*, **38**, 303-324.

[4] Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*, SIAM, Philadelphia.

[5] Beran, R. J. (1975). Tail probabilities of noncentral quadratic forms, *Ann. Statist.*, **3**, 969-974.

[6] Csorgö, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*, Academic Press, New York.

[7] DeWet, T. and Venter, J. H. (1972). Asymptotic distributions of certain test criteria of normality, *S. Afr. Statist. J.*, **6**, 135-149.

[8] DeWet, T. and Venter, J. H. (1973). Asymptotic distributions for quadratic forms with applications to tests of fit, *Ann. Statist.*, **1**, 380-387.

[9] Durbin, J. (1973a). *Distribution Theory for Tests Based on the Sample Distribution Function*, SIAM, Philadelphia.

[10] Durbin, J. (1973b). Weak convergence of the sample distribution function when parameters are estimated, *Ann. Statist.*, **1**, 279-290.

[11] Durbin, J., Knott, M. and Taylor, C. C. (1975). Components of Cramér-von Mises statistics, II, *J. R. Statist. Soc.*, B, **37**, 216-237.

[12] Filliben, J. J. (1975). The probability plot correlation coefficient test of normality, *Technometrics*, **17**, 111-117.

[13] Gregory, G. G. (1980). On efficiency and optimality of quadratic tests, *Ann. Statist.*, **8**, 116-131.

[14] Hoeffding, W. (1964). On a theorem of V. M. Zolotarev, *Theor. Prob. Appl.*, **9**, 89-91.

[15] Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians*, Vol. II, Cambridge University Press, London.

[16] Pearson, E. S., D'Agostino, R. B. and Bowman, K. D. (1977). Tests for departure from normality: Comparison of powers, *Biometrika*, **64**, 231-246.

[17] Pettitt, A. N. (1977). A Cramér-von Mises type goodness of fit statistic related to $\sqrt{b_1}$ and b_2, *J. R. Statist. Soc.*, B, **39**, 364-370.

[18] Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality, *J. Amer. Statist. Ass.*, **67**, 215-216.

[19] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, **52**, 591-611.

[20] Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for normality, *J. Amer. Statist. Ass.*, **63**, 1343-1372.

[21] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons, *J. Amer. Statist. Ass.*, **69**, 730-737.

[22] Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters, *Ann. Statist.*, **4**, 357-369.

[23] Wieand, H. S. (1975). Computation of Pitman efficiencies using the Bahadur approach, *Report No. 7*, Department of Mathematics, University of Pittsburgh.

[24] Wieand, H. S. (1976). A condition under which the Pitman and Bahadur approaches to efficiency coincide, *Ann. Statist.*, **4**, 1003-1011.

[25] Zolotarev, V. M. (1961). Concerning a certain probability problem, *Theor. Prob. Appl.*, **6**, 201-204.