Behavior Research Methods, Instruments, & Computers
1998, 30 (2), 227-231

STATISTICS: RESEARCH AND TEACHING

Reexamining the goodness-of-fit problem for interval-scale scores

RICHARD A. CHECHILE
Tufts University, Medford, Massachusetts

A classic data-analytic problem is the statistical evaluation of the distributional form of interval-scale scores. The investigator may need to know whether the scores originate from a single Gaussian distribution or from a mixture of Gaussian distributions or from a different probability distribution. The relative merits of extant goodness-of-fit metrics are discussed. Monte Carlo power analyses are provided for several of the more powerful goodness-of-fit metrics.

The goodness-of-fit problem in statistics is a general issue that has relevance for research both in psychology and in other disciplines. In this paper, a review of this problem will be outlined. In part, the need for such a review is to close the gap that exists between current practice in psychological research and what is known about the goodness-of-fit problem in statistics. Recent developments now make possible the utilization of more powerful tools for evaluating the goodness-of-fit. In particular, the focus of this paper will be on cases in which the researcher has interval- or ratio-scale measures.

Given interval- or ratio-scale measurements, $x_1, \ldots, x_n$, the goodness-of-fit problem is concerned with the question of whether or not these scores originate from a particular probability distribution function. For example, a set of n reaction time measurements might be modeled by a log-Gaussian distribution or an ex-Gaussian distribution or a mixture distribution of different stochastic processes (see Luce, 1986). We need an answer to this type of question in order to understand the underlying psychological processes. Moreover, if the reaction time values are to be used to fit a psychological model, the likelihood function of the reaction times needs to be specified because all methods of parameter estimation require knowledge of the likelihood function (i.e., the distributional nature of the interval-scale scores). Consequently, the goodness-of-fit evaluation of any proposal about the distribution is a very important initial step toward understanding the psychological processes.

There is a large body of statistical research on the goodness-of-fit problem, but many of these developments unfortunately have yet to be exported to psychology. There are a large number of goodness-of-fit procedures. All of these methods control for the Type I error rate, but there are considerable differences among the methods in regard to power. Generally, current practice in psychology employs only the older goodness-of-fit metrics, which have relatively low power. It is not convincing to argue that the hypothesized model (i.e., the null hypothesis in a goodness-of-fit test) is reasonable because of a nonsignificant test statistic that is known to be low in power.

Correspondence concerning this article should be addressed to R. A. Chechile, Psychology Department, Tufts University, Medford, MA 02155 (e-mail: rchechil@emerald.tufts.edu).

Nominal-Scale Goodness-of-Fit Measures

The two old goodness-of-fit warriors are the χ^2 (K. Pearson, 1900) and the G^2 (Neyman & E. S. Pearson, 1928) statistics. The χ^2 and G^2 statistics are defined as

$$\chi^2 = \sum_{i=1}^{k} (f_{oi} - f_{ei})^2 / f_{ei},$$

$$G^2 = \sum_{i=1}^{k} 2 f_{oi} \log(f_{oi} / f_{ei}),$$

where f_{oi} and f_{ei} are the respective observed and expected frequencies in the ith category. Somewhat more recently, Cressie and Read (1984) developed the power divergence statistic, which is a function of a parameter λ and is given as

$$2nI^{\lambda} = 2\lambda^{-1}(\lambda+1)^{-1} \sum_{i=1}^{k} f_{oi} [(f_{oi}/f_{ei})^{\lambda} - 1].$$

The power divergence statistic has only rarely been acknowledged in psychology (see Batchelder, 1991).

All of these statistics were designed for multinomial or categorical data, although they can also be used when there are interval-scale data. To do so, k observational categories or k intervals need to be created for the continuous random variable. There is no unique way to form the intervals, and there are many possible values for the number of intervals (i.e., the k value). Furthermore, only frequency information is utilized, so even the ordinal infor-

mation inherent in the interval-scale score is lost. For both of these reasons, the nominal-scale goodness-of-fit statistics are relatively low power tests. Despite their low power, the χ^2 and G^2 statistics (with interval-scale data) are still the most commonly used goodness-of-fit measures in psychological research.

Ordinal-Scale Metrics

A number of ordinal-scale statistics are available for goodness-of-fit tests. The ordinal-scale tests are based on the empirical cumulative distribution and on the corresponding theoretical cumulative distribution function, $F(x)$. These tests are generally referred to as "empirical distribution function," or EDF, procedures (see Stephens, 1974, for a review of EDF procedures). With EDF tests, all of the sample scores are included in the empirical step-function cumulative distribution (i.e., the scores are not grouped into arbitrary categories).

The best known of the EDF tests is described by many writers as the Kolmogorov–Smirnov test (e.g., Siegel, 1956; Siegel & Castellan, 1988). Actually this test is based on an original theorem by Kolmogorov (1933); an alternative proof was provided later by Smirnov (1939). The table of critical values was computed by Massey (1951). With this procedure, the investigator computes for each score $D_i^+ = |(i/n) - F(x_i)|$ and $D_i^- = |F(x_i) - (i-1)/n|$—that is, the absolute-value deviations between the theoretical cumulative probability and either of the extremes of the sample step-function cumulative proportion. The test statistic is the maximum of these $2n$ absolute-value deviations. It is commonly assumed in psychological statistics that this test statistic is distribution free. However, Lilliefors (1967, 1969) showed that the Massey tabled values are incorrect and that the test is not distribution free. Lilliefors pointed out that the underlying Kolmogorov theorem assumes that there is knowledge of the population mean and variance of the hypothesized distribution. Yet in virtually all applications the investigator uses the sample data to estimate the population mean and variance (i.e., the hypothesized distribution is fitted by the sample data). This fitting has been shown by Lilliefors (1967) to change the critical value of the test statistic substantially. Moreover, Lilliefors (1969) has shown that the critical value depends on the nature of the theoretical distribution. Lilliefors's results have been replicated, and other writers have come to refer to the D statistic as the Lilliefors statistic (e.g., Dallal & Wilkinson, 1986).

The contributions of Lilliefors have not apparently been recognized in psychological statistics heretofore. Besides the Lilliefors statistic, a number of other EDF statistics have been developed. Stephens (1974) discusses four other EDF statistics (i.e., the Cramér–von Mises statistic, the Kuiper statistic, the Watson statistic, and the Anderson–Darling statistic). These statistics also have not yet been exported to psychological statistics. All of these EDF statistics have greater power than the original Kolmogorov–Smirnov test with the Massey (1951) tabled values. Certainly the EDF tests have more power than the nominal-scale test statistics. However, there are reasons to expect that the EDF statistics are lower power methods in comparison with interval-scale tests. The EDF statistics are based on a probability transformation from the measured interval-scale scores. Probability transforms cannot in general preserve the interval-scale properties of the original measured quantities (except in the case of a rectangular distribution). Hence, it is reasonable to expect that this probability transformation results in lower power relative to interval-scale goodness-of-fit statistics.

Interval-Scale Metrics

Several goodness-of-fit procedures utilize the interval-scale information directly. Nearest neighbor spacings methods (e.g., Bickel & Breiman, 1983; Pyke, 1965) can be used as a goodness-of-fit test. However, Schilling (1983) has found that there is considerably lower power for these spacing-based statistics than for the EDF Lilliefors statistic. The nearest neighbor statistics do not directly compare the entire pattern of the interval-scale data with a hypothesized set of values. A more successful approach was that of the Shapiro and Wilk (1965, 1972) W statistic and the Shapiro and Francia (1972) W' statistic. These statistics have been shown to result in high power (see E. S. Pearson, D'Agostino, & Bowman, 1977; Shapiro, Wilk, & Chen, 1968; and Stephens, 1974). Unfortunately, these statistical tests require the evaluation of expected-order statistics, and those calculations are computationally elaborate. The expected-order statistics are the n means (across repreated samples) of the ith ranked score for $i = 1, \ldots, n$. To find these values requires the numerical evaluation of difficult integrals. The expected-order statistics have been determined for many values of n and for many of the common distributions (see Harter & Balakrishnan, 1996). Expected-order statistics for mixture models are not available, however, and the W and W' statistics are not a feasible method for exploring the type of question raised in the introduction regarding mixture models for reaction times.

Probability plots provide yet another interval-scale approach to assess the goodness of fit. In a probability plot, the sample ranked scores, $x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$, are plotted against corresponding $y_{(i)}$ values, where $F(y_{(i)}) = (i - \gamma)/(n + 1 - 2\gamma)$ and where γ is the plotting-point position. There is little consensus about the value for γ, and eight different values have been proposed: $\gamma = 0$ (Gumbel, 1964); $\gamma = .3$ (Benard & Bos-Levenbach, 1953); $\gamma = .3175$ (Filliben, 1975); $\gamma = .375$ (Blom, 1958); $\gamma = .4$ (Larsen, Curran, & Hunt, 1980); $\gamma = .44$ (Gringorten, 1963); $\gamma = .5$ (Hahn & Shapiro, 1967); $\gamma = .567$ (Larsen et al., 1980). The choice of the plotting point (i.e., the γ value) is known to affect the power of the subsequent goodness-of-fit test, but there has not yet been a good rationale for any particular γ value (Looney & Gulledge, 1985). For example, Bowker and Lieberman (1972) use $\gamma = .5$, but go on to state that the choice is arbitrary.

Recently, Chechile (1997) has proposed a vector-based dissimilarity metric to assess the goodness-of-fit. This is a general approach that can result in identical results to those obtained with the W and W' statistics; however, Chechile recommends two practical methods for defining the dissimilarity statistic, and these methods result in somewhat different measures than do the W and W' statistics. Moreover, these procedures can be used even in the case of nonstandard distributions such as probabilistic mixture models. Because Chechile found only slight differences in the two recommended procedures, only one of these procedures will be discussed here. This method has also been shown by Chechile to be related to a probability-plot procedure, but now with a strong rationale for a particular plotting-point position.

A Dissimilarity Statistic as a Goodness-of-Fit Measure

The fundamental question of goodness-of-fit pertains to whether the data have a distributional shape that is probable, given random sampling from the hypothesized distribution. Distributional shape is invariant to the location (i.e., the mean) and scale factor (i.e., the standard deviation of the scores). Consequently, in this paper all sample scores, x_i, will be first transformed to a standard form such that the mean is m_* and the standard deviation is S_*. The transformed sample values are $v_i = m_* + (x_i - m)S^{-1}S_*$, where m and S are the respective mean and standard deviation of the untransformed scores. Values of m_* and S_* are the respective mean and standard deviation of a set of n idealized points or markers that best represent the hypothesized distribution. The v_i scores still are interval-scale measures. Hence the transformation above is designed to focus the goodness-of-fit evaluation on detecting departures from the hypothesized "distributional shape" by matching the mean and standard deviation of the v_i scores to a corresponding set of idealized u_i scores. The transformed scores are sorted in order to obtain the sample order statistics; that is, $v_{(1)} \leq v_{(2)} \leq \ldots \leq v_{(n)}$. Let **V** be a column matrix (vector) of these sample order statistics.

The goodness-of-fit dissimilarity metric is a function of the mismatch between the vector **V** and a comparable vector **U** of n idealized $u_{(i)}$ scores, where $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$. Because of the transformation to match the mean and standard deviation, the vectors **V** and **U** have the same scalar magnitude. However, if the transformed sampled scores that make up the **V** vector differ in distributional shape from the idealized scores, which are represented by the **U** vector, there will be a nonzero angle between the **V** and **U** vectors in an n-dimensional vector space. The cosine of the angle between the vectors is a measure of the similarity of the sampled scores to the idealized representation (i.e., the similarity would be 1.0 for an angle of 0). If the angle between the vectors is θ, the goodness-of-fit dissimilarity metric, h, is defined as $1 - \cos\theta$. Hence, a perfect match would yield a dissimilarity score of 0. The h statistic is more readily computed by the following equation:

$$h = (1 - r)\,[1 + \{m_*^2 n / S_*^2 (n - 1)\}]^{-1}, \qquad (1)$$

where r is the product-moment correlation between the corresponding $v_{(i)}$ and $u_{(i)}$ values. If the mean m_* of the $u_{(i)}$ markers is 0, the h statistic is $1 - r$. For many distributions (e.g., the exponential distribution), the mean of the markers will not be 0, so the general form in Equation 1 is required. For a fix n, it follows from Equation 1 that $h = (1-r)/K$, where the constant K is greater than 1.0. Because of the monotonic ordering of both the $v_{(i)}$ and $u_{(i)}$ values, the r value is positive. Hence, it follows that $0 \leq h < 1$ and that the angle θ must be less than 90°.

The whole problem now centers on what shall be taken as the idealized markers $u_{(i)}$. For the normal distributions, $u_{(i)}$ values can be selected so that the h statistic is linked directly with the W and W' statistics. With other choices for the $u_{(i)}$ values, the h statistic can be linked to any of the probability-plot statistics. However, a simpler procedure is recommended for generating the idealized markers. If the $u_{(i)}$ values are selected such that $F(u_{(i)}) = (i-.5)/n$, $(i = 1, \ldots, n)$, there is a sense in which these values are an optimal way to represent the hypothesized continuous distribution with only n values. Such chosen $u_{(i)}$ values are each medians of a zone that has a probability of $1/n$. For each zone, the zonal median minimizes an absolute-value error function within the zone. This choice for the idealized markers is directly linked to the plotting position of $\gamma = .5$ used with probability plots, but now there is a strong rationale for that plotting convention.

An Example of Computing the *h* Statistic

To demonstrate the calculations, let us consider a case with low n. For this example, let $n = 8$ and let us assume that the scores originate from a normal distribution. To find the eight idealized scores, we need to find the z scores that correspond to the cumulative proportion of .0625, .1875, .3125, .4375, .5625, .6875, .8125, and .9375; that is, these values are $(i - .5)/n$. The desired z scores can be readily found from tables of the normal distribution that give z for a given cumulative probability (see Kelley, 1948). The resulting values of $u_{(i)}$ are −1.53414, −.88715, −.48878, −.15731, .15731, .48878, .88715, and 1.53414. Although tabled values were used, the resulting values are nevertheless very close to the results obtained by using the Bagby (1995) approximation formula. The mean m_* of the $u_{(i)}$ markers is 0, and the standard deviation S_* is .98622990. Now suppose that the interval-scale scores $(x_1, \ldots, x_n)$ obtained by the investigator are 9.388, 11.756, 13.406, 16.437, 18.067, 20.538, 25.665, and 30.393. The sample mean is 18.20625, and the sample standard deviation is 7.12286238. To test whether the experimental scores originate from a normal distribution, we transform the measurements to v_i values in order to match the mean and standard deviation of the observed scores to that of the set of $u_{(i)}$ idealized markers of an eight-point

representation of a standard normal distribution. The transformation is $m_* + (x_i - m)S^{-1}S_*$. The sorted v_i values are the $v_{(i)}$ scores, and these scores are -1.22097, $-.89310$, $-.66464$, $-.24497$, $-.01928$, $.32285$, 1.03274, and 1.68737. The correlation between the $u_{(i)}$ and $v_{(i)}$ values is .98237. According to Equation 1, when the mean m_* of the $u_{(i)}$ markers is 0, the dissimilarity metric is $1 - r$; hence, the h statistic for the set of observations is .01763. The critical ($\alpha = .05$) h statistic value is .094 for the normal distribution with $n = 8$ (Chechile, 1997). Therefore, the hypothetical data did not demonstrate a significant goodness-of-fit departure from the hypothesized normal distribution.

Power Studies

All of the goodness-of-fit statistics control for the Type I error rate. However, there can be important differences in the statistical power among the measures. To demonstrate the relative sensitivity of some of the more successful goodness-of-fit measures, a set of power studies was performed. In particular, the Lilliefors D statistic, the Shapiro and Francia W' statistic, and the h statistic were utilized in goodness-of-fit tests (at the $\alpha = .05$ level) for detecting a departure from a normal distribution. All samples in these analyses were with $n = 50$. The $\alpha = .05$ critical values for each test statistic were determined by a separate set of 15,000 Monte Carlo samples from a normal distribution. The Monte Carlo sampling was based on the inverse transform method (see Fishman, 1996). For the h statistic, the determination of the 50 idealized markers (i.e., the $u_{(i)}$ values) for the normal distribution only needs to be calculated once and stored.

For the power analyses, separate sets of 15,000 random samples were taken for each combination of the test statistic with each alternative distribution. A power value is estimated by the proportion of the 15,000 random samples from the alternative distribution that would be detected as different from a normal distribution at the $\alpha = .05$ level. Three different Weibull distributions were used as alternative distributions. The cumulative probability function for the Weibull is given as $F(x) = 1 - \exp(-x^{-\beta})$ for $x > 0$. The β parameter affects the shape of the distribution, and the three types of Weibull distributions examined corresponded to $\beta = 1$, $\beta = 1.5$, and $\beta = 2$. The resulting power values are shown in Table 1. For each distribution, the highest power was obtained with the h statistic and the lowest was for the D statistic.

Table 1
Power for the Lilliefors D, Shapiro–Francia W', and the Dissimilarity Statistic h When There is a False Gaussian Null With $n = 50$ and Type I Error Set at .05

	Power		
Alternative Distribution	D	W'	h
Weibull $\beta = 1.0$	.945	.995	.998
Weibull $\beta = 1.5$	.509	.718	.779
Weibull $\beta = 2.0$	.192	.232	.291

Discussion

Heretofore psychological research has utilized low-power goodness-of-fit tools to evaluate the distributional structure of the data. The χ^2 and G^2 statistics are appropriate if the data are categorical, but these nominal-scale statistics result in low statistical power when the data are interval-scale measurements. The use of EDF statistics such as the Lilliefors D statistic would greatly improve the statistical power in psychological statistics (see Stephens, 1974, for a discussion of other EDF procedures). Yet even higher levels of power can be achieved by using the dissimilarity statistic h. Moreover, the h statistic is no more difficult to execute than the EDF statistics. For these statistics, the investigator needs either to find the x value that yields a given cumulative probability, $F(x)$, or to find the $F(x)$ value for a given x.

The Shapiro–Francia W' statistic also results in higher power than the Lilliefors D statistic, but W' is, in general, harder to compute because of the necessity of first obtaining the computationally difficult expected-order statistics. Importantly, however, the h statistic achieved higher power than the W' statistic, so there is no compelling reason to use the W' statistic.

Chechile (1997) provided critical values for the h statistic for the normal and exponential distributions. However, there are many possible distributions of potential interest, especially when one considers mixture distributions. Provided that the investigator can obtain the n idealized markers for the hypothesized distribution, the sample h statistic is still a simple computation via Equation 1. However, in order to understand the implications of the sample h value for these nonstandard distributions, the investigator will need to determine the critical value of the h statistic by means of Monte Carlo sampling. In fact, Monte Carlo sampling will also be required for the other goodness-of-fit measures (e.g., the EDF statistics, and W') when one is evaluating any nonstandard hypothesized distribution, because the critical values are generally dependent on the distributional form. There are no truly distribution-free tests if the sample data are used to fit the hypothesized distribution.

In the case of Monte Carlo sampling, it is recommended that the investigator carefully consider the computational method used to sort the data. Sorting is a necessary step for the calculation of the h statistic per Monte Carlo sample. Sorting is also required in order to determine an EDF statistic for each sample. Inefficient sorting can add unnecessary time to the determination of the critical values of the goodness-of-fit statistic. For Monte Carlo samples, however, there is an efficient method of approximately presorting the samples, and this presorting can greatly reduce the sorting time. For each randomly generated score there is an $F(x)$ value associated with that score. In fact, if the inverse-transform method was used to generate the random score in the first place, a random number on the (0,1) interval was selected and that number was treated as the $F(x)$ value. With the inverse-transform

method, the random score is the x value that yields the $F(x)$ value for the probability distribution. The random x value can be initially stored in an empty slot in a large array. The approximate neighborhood of the empty slot is an integer and a function of $F(x)N$, where N is the size of the large array. With this method, the value of N must be much larger than the sample size. The large array is then compressed to another array of size n, and this array is approximately sorted. Any sorting algorithm can now be used on the presorted array to rapidly obtain the final listing. With a sufficiently large initial array, the sorting of a set of 1,000 scores can typically be accomplished in only a few iterations with this method.

Finally, current computer hardware has enabled a host of statistical calculations that would be difficult or impossible to execute with simple paper-and-pencil calculations. The h statistic and the EDF goodness-of-fit statistics are now easy calculations to obtain for a set of interval-scale data. The more computationally challenging problem is the determination of the critical levels for these statistics when one is testing a nonstandard distribution. However, modern computers and efficient sorting algorithms provide a means to determine the critical values.

REFERENCES

BAGBY, R. J. (1995). Calculating normal probabilities. *American Mathematical Monthly*, **102**, 46-49.

BATCHELDER, W. H. (1991). Getting wise about minimum distance measures. *Journal of Mathematical Psychology*, **35**, 267-273.

BENARD, A., & BOS-LEVENBACH, E. C. (1953). The plotting of observations on probability paper. *Statisica Neerlandica*, 7, 163-173.

BICKEL, P. J., & BREIMAN, L. (1983). Sums of functions of nearest-neighbor distances, moment bounds, limit theorem, and a goodness-of-fit test. *Annals of Probability*, **11**, 185-214.

BLOM, G. (1958). *Statistical estimates and transformed beta variables*. New York: Wiley.

BOWKER, A. H., & LIEBERMAN, G. J. (1972). *Engineering statistics* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

CHECHILE, R. A. (1997). *A vector-based goodness-of-fit metric for interval data*. Manuscript submitted for publication.

CRESSIE, N., & READ, T. R. C. (1984). Multinominal goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B*, **46**, 440-464.

DALLAL, G. E., & WILKINSON, L. (1986). An approximation to the distribution of Lilliefors's test statistic for normality. *American Statistician*, **40**, 294-296.

FILLIBEN, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, **17**, 111-118.

FISHMAN, G. S. (1996). *Monte Carlo: Concepts, algorithms, and applications*. New York: Springer-Verlag.

GRINGORTEN, I. I. (1963). A plotting rule for extreme probability paper. *Journal of Geophysical Research*, **68**, 813-814.

GUMBEL, E. J. (1964). *Statistical theory of extreme values*. Washington, DC: National Bureau of Standards.

HAHN, G. J., & SHAPIRO, S. S. (1967). *Statistical models in engineering*. New York: Wiley.

HARTER, H. L., & BALAKRISHNAN, N. (1996). *CRC handbook of tables for the use of order statistics in estimation*. Boca Raton, FL: CRC Press.

KELLEY, T. L. (1948). *The Kelley statistical tables*. Cambridge, MA: Harvard University Press.

KOLMOGOROV, A. N. (1933). Sulla determinazione empirica di una legge di distribuizione. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 83-91.

LARSEN, R. I., CURRAN, T. C., & HUNT, W. F. (1980). An air quality data analysis system for interrelating effects, standards, and needed source reduction: Part 6. Calculating concentration reductions needed to achieve the new national ozone standard. *Journal of the Air Pollution Control Association*, **30**, 662-669.

LILLIEFORS, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**, 399-402.

LILLIEFORS, H. W. (1969). On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, **64**, 387-389.

LOONEY, S. W., & GULLEDGE, T. R. (1985). Use of the correlation coefficient with normal probability plots. *American Statistician*, **39**, 75-79.

LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

MASSEY, F. J. (1951). The Kolmogorov–Smirnov test for goodness-of-fit. *Journal of the American Statistical Association*, **46**, 68-78.

NEYMAN, J., & PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika*, **20A** (Pt. I), 175-240; (Pt. II), 263-294.

PEARSON, E. S., D'AGOSTINO, R. B., & BOWMAN, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, **64**, 231-246.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine*, **50**, 157-172.

PYKE, R. (1965). Spacings. *Journal of the Royal Statistical Society: Series A*, **109**, 85-110.

SCHILLING, M. F. (1983). Goodness-of-fit testing in R^m based on the weighted empirical distribution of certain nearest-neighbor statistics. *Annals of Statistics*, **11**, 1-12.

SHAPIRO, S. S., & FRANCIA, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215-216.

SHAPIRO, S. S., & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.

SHAPIRO, S. S., & WILK, M. B. (1972). An analysis of variance test for the exponential distribution (complete samples). *Technometrics*, **14**, 355-370.

SHAPIRO, S. S., WILK, M. B., & CHEN, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, **63**, 1343-1372.

SIEGEL, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

SIEGEL, S., & CASTELLAN, N. J., JR. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

SMIRNOV, N. (1939). Ob uklonenijah empiričeskoi krivoi raspredelenija [On the deviations of the empirical distributional curve]. *Matematiceskii Sbornik*, **48**, 3-26.

STEPHENS, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730-737.

(Manuscript received October 15, 1997;
revision accepted for publication March 23, 1998.)