Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data

Author(s): Marc S. Weiss

Source: *Journal of the American Statistical Association*, Dec., 1978, Vol. 73, No. 364 (Dec., 1978), pp. 872-875

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2286297

Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data

MARC S. WEISS*

The one-sample Kolmogorov-Smirnov goodness-of-fit test (K-S test) is designed for use with independent data. Critical values of the K-S statistic can be highly sensitive to correlated data. For data representing sampling of a second-order autoregressive process, an empirically derived correction for the K-S statistic provides correct critical values as well as increased power. The correction is a simple linear function of a parameter τ which is derived from the second and the fourth moments of the normalized spectrum of the autoregressive process.

KEY WORDS: Kolmogorov-Smirnov statistic; Autoregressive process; Correlated data.

1. INTRODUCTION

The Kolmogorov-Smirnov (K-S) statistic for testing goodness-of-fit between a theoretical distribution and a sample set of observations is well-known and widely used. If $F(x)$ is a population distribution function and $S_n(x)$ is the sample distribution step-function, the K-S statistic is defined as:

$$KS = \max|F(x) - S_n(x)| \ . \tag{1.1}$$

For continuous $F(x)$, the sampling distribution of KS is known and is independent of $F(x)$.

For the particular case of the normal distribution with mean and variance estimated from the sample, Monte Carlo corrections for critical values of the K-S statistic have been obtained (Lilliefors 1967, Stephens 1974). Recently, some attention has been devoted to applying the K-S test to sampled electroencephalographic (EEG) data (McEwen and Anderson 1975; Persson 1974; Weiss 1973; Jones, Crowell, and Kapuniai 1969). Sampled EEG data, as most time series data, have a spectral content and autocorrelation function which is dependent on the sampling rate. Often this rate is chosen on the basis of spectral considerations, sampling being carried out at twice the highest frequency of interest. Consequently, the sampled data can be highly correlated. If one wishes to apply the K-S test (or any other statistic that demands independent data) correctly, one must discard data by reducing the effective sampling rate to ensure an uncorrelated ("white noise") sample (Persson 1974). Since in many practical instances the original sample length is of fixed duration, this means a decrease in sample size and a corresponding reduction in the power of the test. Ideally, one would like to be able to estimate the autocorrelation properties of the data and use this estimate to correct the test statistic so that all the data may be used.

This article presents the results of the first step in an attempt to develop such a general procedure. These results demonstrate a method of applying the K-S test to correlated data which corrects for the influence of the correlation on the distribution of the K-S statistic. This correction enables all the data to be used as originally sampled, thereby increasing the power of the test. The results are applicable to those time series data which can be modeled as being generated by a second-order autoregressive (AR) process (Box and Jenkins 1970, Jenkins and Watts 1968), and require knowledge of the underlying parameters of the process. The problem of estimating these parameters and extending the results to more general types of time series is the subject of work in progress and is briefly discussed in this report.

Although the AR model has been extensively developed and used, some pertinent ideas and results will be briefly outlined. Measurement of a stochastic process at fixed intervals of time results in a sample of sequential observations. The resulting time series is said to be autoregressive of order p if each observation X_t can be expressed as:

$$X_t = a_1X_{t-t_s} + a_2X_{t-2t_s} + \ldots + a_pX_{t-pt_s} + e_t \ , \tag{1.2}$$

where the e_t's are independent and identically distributed random variables, each with zero mean and distribution function $F(e)$, and t_s is the time between the successive observations X_{t-t_s} and X_t. The second-order AR process is of particular theoretical and practical importance and can be written as:

$$X_t = a_1X_{t-t_s} + a_2X_{t-2t_s} + e_t \ . \tag{1.3}$$

For the process (1.3) to be covariance stationary, it is required that:

$$\begin{aligned} |a_2| &< 1 \ , \\ |a_1| &< 1 - a_2 \ . \end{aligned} \tag{1.4}$$

The first-order process can be considered a special case of (1.3) and (1.4) with $a_2 = 0$. The one-sided, normalized power density spectrum corresponding to (1.3) is:

$$P(\omega) = t_sK_1/(K_2 + K_3\cos\omega t_s + K_4\cos 2\omega t_s) \ , \quad 0 \le \omega t_s < \pi \ . \tag{1.5}$$

* Marc S. Weiss is a neurophysiologist at the Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA 70189. Opinions or conclusions contained in this report are those of the author and do not necessarily reflect the views or the endorsement of the Navy Department.

© Journal of the American Statistical Association
December 1978, Volume 73, Number 364
Theory and Methods Section

The constants in (1.5) are:

$$K_1 = 2((1 - a_2)^2 - a_1^2)(1 + a_2)/(1 - a_2) ;$$
$$K_2 = 1 + a_1^2 + a^2 ;$$
$$K_3 = -2a_1(1 - a_2) ;$$
$$K_4 = -2a_2 .$$

Examination of (1.5) shows that for $|a_1(1 - a_2)| < |4a_2|$ and $a_2 < 0$, the spectrum will have a peak at ω_0 if $\cos \omega_0 t_s = a_1(a_2 - 1)/4a_2$.

Thus, processes such as EEG activity, which have peaked spectra, can often be approximately represented by a second-order AR series (Gersch 1970; Jones, Crowell, and Kapuniai 1969; Pfurtscheller and Haring 1972).

The results described in this article demonstrate the effect of spectra of the form (1.5) on the K-S test for normality, with mean and variance estimated from the observed sample.

2. METHOD

All Monte Carlo simulations described in this report were done on a small, general-purpose digital computer. The basic uniform random number generator used an algorithm described by Lewis and Payne (1973). Random normal deviates were generated by a table look-up procedure using Hastings, Hayward, and Wong's (1955) approximation for the 32,767 table values. The cycle length of the pseudorandom normal sequence finally used was slightly greater than 2^{22} (4,718,592).

Sequences of the form (1.3) were generated for different choices of a_1, a_2, and $F(e)$. Preliminary testing indicated that discarding the first 40 values of the generated series was adequate to eliminate initialization (i.e., transient or nonstationary) effects. In addition to a_1, a_2, and $F(e)$, each simulation involved a choice of sample size (n), number of samples (1,000 in most cases), and number of independent replications (usually three). The details for each set of simulations will be described.

3. RESULTS

3.1 Preliminary Test (a_1, $a_2 = 0$)

As a preliminary methodological check, 1,000 samples of size 25 were generated with a_1 and a_2 equal to zero and $F(e)$ a standard normal distribution. The K-S statistic was computed (assuming a standard normal population distribution), and excellent agreement was obtained with the critical values tabulated by Miller (1956).[1]

This test was then repeated using 10,000 samples of sizes 16 through 256, this time assuming that the population distribution has a mean and standard deviation estimated by the sample mean and standard deviation. Using these results, a formula was developed empirically for computing critical values of the K-S statistic using the sample mean and standard deviation:

$$KS(\alpha, n) = M(\alpha, n)(0.7 - 1.8n^{-\frac{3}{2}} + 3.15n^{-2} + 0.075\alpha + 0.01245 \log_e \alpha) ;$$
$$n = \text{sample size } (>8) ;$$
$$\alpha = \text{significance level } (\geq 0.01, \leq 0.20) ; \quad (3.1)$$

where $M(\alpha, n)$ is the expression given by Miller (1956) for computing critical values of the K-S statistic with mean and variance known:

$$M(2\alpha, n) = (-0.5 \log_e \alpha)^{\frac{1}{2}} n^{\frac{1}{2}} - 0.16693n^{-1} - 0.025864(-\log_e \alpha)^{\frac{3}{2}} n^{-\frac{3}{2}} + 0.0028575(\log_e \alpha)^2 n^{-\frac{3}{2}} - 0.08467\alpha n^{-\frac{3}{2}} - 0.11143n^{-\frac{3}{2}} . \quad (3.2)$$

When tested in a second independent simulation, formulas (3.1) and (3.2) yielded values with average errors of less than one percent with respect to the Monte Carlo values. These values were also consistent with those reported by Stephens (1974) (worst case difference less than one percent).

3.2 Critical Values of Modified K-S Statistic

For the model described by (1.3)–(1.5), 1,000 samples of sizes (n) 16 through 256 were generated for N different values of a_1 and a_2 using the normal distribution for $F(e)$. Each 1,000-sample simulation was replicated at least three times. For each set of parameters (a_1, a_2, n), the K-S statistic was computed using the mean and variance estimates from the sample. Critical values thus obtained were averaged over replications and an attempt was made to relate these values to the AR parameters. The region where $a_1 > 0$ and $a_2 < 0$ was investigated in detail, as was the subregion for which $a_1^2 + 4a_2 < 0$. In this subregion the effect of correlation is most marked and spectra have low-frequency peaks.

The results of these simulations indicate that there was no simple relationship between the empirical K-S statistic and the parameters a_1 and a_2. However, series with narrow, low-frequency peaks in the spectra (1.5) yielded larger K-S values than series with broader or higher frequency spectral peaks. Two spectral parameters reflected this effect: f_2, the square root of the second moment, and f_4, the fourth root of the fourth moment of (1.5). In particular, the ratio $(f_4 - f_2)/f_4 f_2$ was monotonically related to the empirical K-S statistic. Using this result, a dimensionless parameter was defined as:

$$\tau = ((1/f_2 - 1/f_4)/t_s)^{\frac{1}{2}} . \quad (3.3)$$

For the uncorrelated case ($a_1 = 0$, $a_2 = 0$), τ has a value of 0.69. The value of τ increases with increasing magnitude of $a_1 + a_2$, taking on values greater than 1.0 when $a_1 + a_2 > 0.8$.

Table 1a and 1b summarizes the results of the simulations which led to the validation of (3.3) as a correction parameter. There was a strong linear relationship (as evidenced by the high correlations) between the parameter τ computed using (3.3) and (1.5), and the ratio

[1] Values from Miller's table for KS critical values at significance levels of 0.10, 0.05, 0.02, and 0.01, respectively, are: 0.23768, 0.26404, 0.29516, 0.31657. The corresponding simulated values obtained were: 0.2386, 0.2658, 0.2954, 0.3136.

1. Correlation of Empirical KS_e Statistic and the Parameter τ[a]

n	τ Mean	τ Range	KS_e/KS Mean	KS_e/KS Range	N	Correlation coefficient
			a. Second-order AR			
16	2.18	1.01, 3.25	.944	.868, 1.036	9	.982
25	1.62	.99, 2.74	.947	.874, 1.111	14	.967
32	2.20	.90, 3.43	1.075	.860, 1.300	51	.988
64	2.21	1.00, 4.29	1.151	.891, 1.586	5	.995
128	3.32	.92, 6.70	1.452	.890, 2.180	21	.996
256	3.40	1.00, 6.70	1.522	.910, 2.374	18	.998
			b. First-order AR			
16	2.22	1.52, 3.19	.906	.881, .940	5	.991
25	1.82	1.15, 3.15	.934	.868, 1.042	7	.994
32	1.89	1.03, 3.15	.984	.879, 1.118	9	.978
64	1.94	1.03, 3.15	1.070	.897, 1.343	11	.992
128	1.89	1.03, 3.15	1.123	.901, 1.540	9	.992

[a] These results are for the empirical KS_e values at the .05 significance level. For each sample size n, N different choices of a_1 and a_2 were used to generate time series with values of τ as indicated. The time series and the N values of the empirical KS_e statistic were obtained by simulation procedures described in the text. The corresponding values for the known uncorrelated KS statistic were computed by using equation (3.1).

Note: The values of a_1 and a_2 were such that for the second-order AR, $a_1 > 0$, $a_2 < 0$, and $a_1^2 + 4a_2 < 0$. For the first-order AR, $a_1 > 0$.

2. Power Comparisons for Second-Order AR[a]

a_1	a_2	n	τ	F(e)[b] U_1	t_1	t_2	X_1^2	X_2^2	Normal
1.81	−.845	192	4.02	.05	.87	.34	.19	.13	.04
		48	1.49	.06	.84	.27	.16	.11	.04
		32	1.09	.05	.80	.27	.14	.14	.06
		16	.65	.05	.69	.19	.10	.08	.04
1.65	−.69	192	3.94	.03	.90	.38	.18	.12	.05
		48	1.62	.03	.86	.30	.17	.11	.03
		32	1.24	.05	.83	.28	.15	.09	.04
		16	.82	.04	.71	.20	.10	.07	.05
1.60	−.8	192	1.99	.04	.97	.47	.48	.27	.04
		48	.58	.04	.90	.37	.41	.16	.04
		32	.43	.04	.82	.27	.19	.12	.04
		16	.80	.04	.70	.21	.11	.09	.05
1.3	−.5	192	2.09	.02	1.00	.68	.83	.52	.04
		48	.81	.04	.99	.56	.48	.34	.04
		32	.67	.04	.95	.42	.44	.25	.04
		16	.68	.04	.79	.26	.21	.13	.05

[a] This table gives the fraction of samples yielding KS_e values significant at the .05 level when the sample size (n) and the underlying noise distribution (F(e)) are as shown.

[b] t_1, t_2: Student's t with 1 and 2 df; X_1^2, X_2^2: chi-squared with 1 and 2 df; U_1: uniform.

KS_e/KS. Empirically, this can be expressed as:

$$KS_e(\alpha, n)/KS(\alpha, n) = R(n)(\tau - 1) + 1 \ , \quad (3.4)$$

where n is the sample size, $KS_e(\alpha, n)$ is the critical value at the α level of the empirical K-S statistic, and $KS(\alpha, n)$ is the critical value computed from (3.1). Examination of the simulation results expressed in the form of (3.4) indicated that a function of the form:

$$R(n) = c_1 - c_2/n + c_3/n^2 \quad (3.5)$$

provided a reasonably good fit to the data. The results of empirically fitting (3.5) to the data yielded:

$$R(n) = 0.30 - 4.3/n + 19.0/n^2 \quad (3.6)$$

when $a_1 > 0$, $a_2 < 0$, and $a_1^2 + 4a_2 < 0$; and

$$R(n) = 0.45 - 14/n + 120/n^2 \quad (3.7)$$

when $a_2 = 0$ (first-order model) and $a_1 > 0$.

The results (3.4)–(3.7) are valid for $\tau \geq 1.0$ and $n \geq 16$ and can be used to compute the critical values of the K-S statistic when $\tau \geq 1.0$. For smaller values of τ, the uncorrelated statistic $KS(\alpha, n)$ can be used without correction, but may yield slightly conservative estimates for critical values. It is apparent from (3.6) and (3.7) that the constants in the correction equation differ in the two regions of the (a_1, a_2) plane which were selected for investigation.

3.3 Power of Modified K-S Statistic

As a final series of simulations, the power of the modified K-S statistic, computed by using (3.4)–(3.7), was investigated. Several values of AR coefficients were used as well as several noise distributions, as indicated in Table 2. The following procedure was used to determine power empirically: For a given set of parameters $(a_1, a_2, F(e))$ where $F(e)$ was nonnormal, an AR series of length 192 was generated. The power for this original series was computed using (3.4) to determine critical values of the K-S statistic at the 0.05 significance level. The power was then recomputed for new series of smaller sizes obtained from the original series by selecting every jth point, where j ranged from 4 to 12. For each of these new series, a new τ was computed as well as a new modified K-S critical value. Though strictly speaking, (3.4) does not apply in the skipped case (since the samples no longer represent a second-order AR series), results in the last column of Table 2 indicate that (3.4) and (3.6) yielded correct critical values. Similar results hold for the first-order AR series, as indicated in Table 3. These

3. Power Comparisons for First-Order AR[a]

a_1	n	τ	F(e)[b] U_1	t_2	X_1^2	X_2^2	Normal
.95	192	3.19	.04	.30	.13	.08	.04
	48	1.90	.04	.22	.09	.07	.04
	32	1.62	.05	.24	.09	.07	.05
	16	1.21	.04	.15	.05	.06	.06
.90	192	2.45	.03	.43	.26	.15	.04
	48	1.42	.04	.31	.18	.12	.04
	32	1.20	.04	.28	.16	.10	.04
	16	.91	.05	.17	.11	.08	.04
.85	192	2.08	.02	.54	.54	.25	.03
	48	1.19	.04	.42	.34	.17	.04
	32	1.01	.04	.35	.26	.17	.06
	16	.79	.04	.24	.15	.10	.05
.65	192	1.41	.03	.86	.99	.88	.04
	48	.82	.05	.61	.76	.48	.04
	32	.74	.05	.45	.57	.36	.04
	16	.69	.04	.30	.32	.20	.06

[a] This table gives the fraction of samples yielding KS_e values significant at the .05 level when the sample size (n) and the underlying noise distribution (F(e)) are as shown.

[b] t_2: Student's t with 2 df; X_1^2, X_2^2: chi-squared with 1 and 2 df; U_1: uniform.

results indicate that except for $F(e)$ uniform, power is increased by using the original sample in conjunction with (3.4)–(3.7).

Note that except for the normal case, the distribution of the resultant AR series was, in general, different from that of $F(e)$ (the underlying noise). This alternative distribution against which the power of the K-S test was measured was a weighted sum of independent $F(e)$'s, the weights being a function of a_1 and a_2. The poor performance when $F(e)$ was uniform illustrated this effect. This reflected the weakness of the K-S test for normality when the underlying distribution is a sum of independent uniform distributions. For example, for independent samples from a triangular distribution (sum of two independent uniform distributions), simulations as described previously indicate that the power of the K-S test for normality is only 0.06 ($n = 100$, 0.05 significance level).

4. DISCUSSION

For data which can be approximated by a second-order AR series with a low-frequency peaked spectrum, the power of the K-S test for normality can be increased by applying equations (3.4)–(3.7). This correction for the effect of the nonzero correlation requires knowing two spectral moments which reflect the influence of the two AR coefficients. Since this is a parametric correction, the actual power of the procedures suggested may depend on the manner in which these moments are estimated.

There are several ways in which this can be done. First, a spectral estimate can be computed for the time series and from this, the two moments and an estimate, τ_e, for τ can be computed. A second, more direct approach is to use average zero-crossing counting techniques. The average zero-crossing rate of the original time series and the average zero-crossing rate of the first derivative of the time series are estimates of the second and fourth moments, respectively, of the power density spectrum (Saltzberg and Burch 1971). Thus τ_e can be computed directly without transforming the data via Fourier analysis. The relative merits and shortcomings of these two approaches are currently being investigated.

If one is interested in the AR modeling of time series, then, of course, one could estimate the coefficients, remove the AR component from the original series, and apply the standard K-S test to the residual series. However, if the question of primary interest is the normality of the process (such as EEG), this approach involves unnecessary computation. It should be further noted that (3.4) may have more general applicability, as the results in Tables 2 and 3 indicate. For the reduced sample sizes (skipped data), the correction for the KS critical values appears valid despite the fact that the samples are no longer strictly autoregressive. Thus corrections to the K-S statistic of the form (3.4) may also provide the basis for a direct, computationally efficient test for normality of nonautoregressive time series. This possibility is also currently being investigated.

[*Received July 1976. Revised May 1978.*]

REFERENCES

Box, George E.P., and Jenkins, Gwilym M. (1970), *Time Series Analysis; Forecasting and Control*, San Francisco: Holden-Day.

Gersch, Will (1970), "Spectral Analysis of EEG's by Autoregressive Decomposition of Time Series," *Mathematical Biosciences*, 7, 205–222.

Hastings, Cecil, Jr., Hayward, Jeanne T., and Wong, James P., Jr. (1955), *Approximations for Digital Computers*, Princeton, N.J.: Princeton University Press.

Jenkins, Gwilym M., and Watts, Donald G. (1968), *Spectral Analysis and Its Application*, San Francisco: Holden-Day.

Jones, Richard H., Crowell, David H., and Kapuniai, Linda E. (1969), "A Method for Detecting Change in a Time Series Applied to Newborn EEG," *Electroencephalography and Clinical Neurophysiology*, 27, 436–440.

Lewis, T.G., and Payne, W.H. (1973), "Generalized Feedback Shift Register Pseudorandom Number Algorithm," *Journal of the Association for Computing Machinery*, 20, 456–468.

Lilliefors, Hubert W. (1967), "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399–402.

McEwen, James A., and Anderson, Grant B. (1975), "Modeling the Stationarity and Gaussianity of Spontaneous Electroencephalographic Activity," *IEEE Transactions on Biomedical Engineering*, BME-22, 361–369.

Miller, Leslie H. (1956), "Table of Percentage Points of Kolmogorov Statistics," *Journal of the American Statistical Association*, 51, 111–121.

Persson, Jan (1974), "Comments on Estimations and Tests of EEG Amplitude Distributions," *Electroencephalography and Clinical Neurophysiology*, 37, 309–313.

Pfurtscheller, G., and Haring, G. (1972), "The Use of an EEG Autoregressive Model for the Time-Saving Calculation of Spectral Power Density Distributions with a Digital Computer," *Electroencephalography and Clinical Neurophysiology*, 33, 113–115.

Saltzberg, Bernard, and Burch, Neil R. (1971), "Period Analytic Estimates of Moments of the Power Spectrum: A Simplified EEG Time Domain Procedure," *Electroencephalography and Clinical Neurophysiology*, 30, 568–570.

Stephens, M.A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737.

Weiss, Marc S. (1973), "Non-Gaussian Properties of the EEG During Sleep," *Electroencephalography and Clinical Neurophysiology*, 34, 200–202.