EVALUATING THE SMIRNOV DISTRIBUTION FUNCTION

by

John F. Monahan

Department of Statistics

Institute of Statistics Mimeo Series 1965
Center for Research in Scientific Computation #TR-89-2

November, 1989

EVALUATING THE SMIRNOV DISTRIBUTION FUNCTION

John F. Monahan
Department of Statistics
North Carolina State University
Raleigh, North Carolina 27695 - 8203

November, 1989

Abstract

The limiting distribution of the Kolmogorov-Smirnov goodness of fit statistic D_n is usually given by the formula

$$\Pr(\sqrt{n}\, D_n \leq z) \xrightarrow{d} L(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}.$$

This expression works well as a computing formula for L(z) for large z, say z>1. However, for small values of z, evaluating this slowly decaying alternating series respresentation of L(z) is a numerical nightmare; every route appears doomed to slow convergence and catastrophic cancellation. However, an adventure into the world of theta functions leads to a simple, safe computing formula.

1. Introduction

Kolmogorov and Smirnov in the 1930's devised goodness of fit tests for the one and two sample problems based on the empirical distribution function. For the one sample problem, observed are $X_1, X_2, ..., X_n$ independent and identically distributed (iid) from some distribution; the hypothesis to be tested is that the distribution is some specified df F(x). The empirical distribution function is simply described by

$$F_n(x) = (\text{ number of } X_i\text{'s} \le x) / n.$$

Kolmogorov's (1933) test statistic is $D_n = \sup | F_n(x) - F(x) |$, and its distribution does not depend upon F when the hypothesis is true. While in small samples distribution of D_n is difficult, in large samples the distribution of D_n can be described by the asymptotic result $\Pr(\sqrt{n}\, D_n \le z) \xrightarrow{d} L(z)$, where the distribution function L(z) is given by the series

$$L(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2}. \tag{1}$$

Although Kolmogorov did the earliest work on this subject, the distribution L(z) is often called the Smirnov distribution.

The two sample problem works similarly. Observed are $X_1^{(1)}, X_2^{(1)}, ..., X_m^{(1)}$ iid from the first sample with distribution function $F^{(1)}(x)$, and $X_1^{(2)}, X_2^{(2)}, ..., X_n^{(2)}$ iid $F^{(2)}(x)$ for the second sample. Form the two empirical distribution functions

$$F_m^{(1)}(x) = (\text{number of } X_i^{(1)}\text{'s} \le x)/m \quad \text{and} \quad F_n^{(2)}(x) = (\text{number of } X_i^{(2)}\text{'s} \le x)/n$$

and compute the test statistic $D_{m,n} = \sup | F_m^{(1)}(x) - F_n^{(2)}(x) |$ for testing the hypothesis H: $F^{(1)} = F^{(2)}$. Smirnov (1939) showed that the limiting distribution of $D_{m,n}$ takes the same form as in the one sample problem:

$$\lim \Pr(\sqrt{\tfrac{1}{m}+\tfrac{1}{n}}\, D_{m,n} \le z) = L(z) \quad \text{as } m,n \to \infty \text{ and } m/n \to r \text{ nonzero and finite}$$

where L(z) has the series expansion as previously given.

Gnedenko (1962, pp.393-401) and Wilks (1962, pp.454-459) give a lucid analyses of the two sample problem where m=n for both the exact small sample case and the asymptotic case. Feller (1948) derives both the one and two sample asymptotic distributions from a counting argument. In the same issue (1948), Smirnov's table of L(z) is reprinted. Note that both papers in this issue give the formula for L(z) missing the "2" in the exponent.

Both of these tests are omnibus tests and not very powerful when compared to others designed for testing against specific alternatives, such as a location shift. In simulation studies for verifying certain distributions, the sample sizes may be very large and the Kolmogorov-Smirnov tests become practical and effective just because of consistency. The onmibus nature is preferred to any specific alternatives, since the kinds of departures anticipated are the unpredictable effects of programming errors.

In common use, the rejection regions are large values of the statistics D_n and $D_{m,n}$, and so the usual computational problem is to evaluate L(z) for moderate to large values of z. However, especially in simulation studies, small values of D_n or $D_{m,n}$ may lead to a rejection of the iid hypothesis on the grounds that the independence hypothesis has been violated. Moreover, there are other applications for which L(z) may be required for any value of z, and motivate the problem to be addressed here.

2. Computation

For z large, the problem of evaluating L(z) is really the evaluation of tail probabilities 1-L(z). In such a case, the alternating series expansion (1) can be rewritten into a sum of positive terms by pairing up:

$$1 - L(z) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2z^2} = 2 [\, e^{-2z^2} - e^{-8z^2} + e^{-18z^2} - e^{-32z^2} + \ldots \,] \qquad (2)$$

$$= 2 [\sum_{j=1}^{\infty} (e^{-2(2j-1)^2z^2} - e^{-2(2j)^2z^2})]$$

$$= 2 \sum_{j=1}^{\infty} e^{-2(2j-1)^2z^2} (1 - e^{(2-8j)z^2})$$

leaving a sum of rapidly decreasing positive terms. In fact, for evaluating at z = 1, only the first two terms are needed:

$$1 - L(1) = 2 [.135335(1-.002479) + 1.5x10^{-8}(1 - 8x10^{-7})] = .26999967$$

where Smirnov' s table gives L(1) = .730000 to 6 digits. Such a simple calculation indicates that z=1 is large z, and the tail formula works well.

For small values of z, the story looks quite different. Let us denote the partial sum of the series representation by $L_m(z)$, that is,

$$L_m(z) = 1 - 2 \sum_{k=1}^{m} (-1)^{k-1} e^{-2k^2z^2}.$$

Now examine the convergence of the partial sum for a small value of z, say z=0.1:

m	$L_m(0.1)$
0	1.000000
1	- .980100
2	.885835
3	- .804407
4	.572720
5	- .640340
6	.333163

Needless to say, the convergence appears to be very slow for z as small as 0.1 . In fact, k=9 is the first term which is less than 1/2, and k=26 is the first term less than 10^{-6}. Taking z even smaller to z = 0.01, k=83 is the first term less than 1/2 and k=263 is the first term less than 10^{-6}. Clearly, the convergence is ponderously slow for small z. Even more disturbing is the uncertainty in the value of L(z) for z as small as 0.1 and 0.01, since the accuracy needed in the arithmetic to compute L(z) cannot be determined without knowledge of the magnitude of L(z). For if the value of L(z) is 10^{-m} for some z, then at least m digit base ten arithmetic is needed. And if only d (d<m) decimal digits are used, then the cancellation can be catastrophic, leaving at best the first d digits all zero and no sense where m may be. In practice, with d digit arithmetic, the cancellation and rounding of finite d digit arithmetic may leave even the d^{th} digit is error, not zero and perhaps even negative. For the case of z=0.1 with d=6 base B=16 (approximately d=6 decimal) arithmetic, going out to 26 terms yields a result of .000000477, while using d=14 (double precision) the value is -.000000234, neither of which is appealing.

The first response to this problem is to find another way of addition so that the terms are all positive, and since decreasing, would give a good sense of the magnitude of L(z). One route is to pair up in a fashion previously used for large z and 1-L(z). For small z and L(z), one route is the following pairing:

$$L(z) = (1-2e^{-2z^2}) + 2(e^{-8z^2} - e^{-18z^2}) + 2(e^{-32z^2} - e^{-50z^2}) + \ldots$$

but all of the terms are positive only if the first one is positive. And $1 > 2e^{-2z^2}$ only if $z > z_{1/2} = \sqrt{ln\ 2/2} = .5887$, so that the range of applicability does not extend anywhere near z = 0.1 . Note that at $z_{1/2}$ is a small value of z, since $L(z_{1/2}) = .121124$, indicating that L(z) is quite small at z = 0.1 . A second attempt, similar to the pairing is to take sets of three terms:

$$L(z) = (1 - 2e^{-2z^2} + e^{-8z^2}) + (e^{-8z^2} - 2e^{-18z^2} + e^{-32z^2}) + (e^{-32z^2} - 2e^{-50z^2} + e^{-72z^2}) + \ldots$$

$$a_k = \int_0^1 g(x)\, e^{-2\pi ikx} dx = \int_0^1 [\ \sum_{n=-\infty}^{\infty} f(x+n)\ ]\, e^{-2\pi ikx} dx = \sum_{n=-\infty}^{\infty} \int_n^{n+1} f(x)\, e^{-2\pi ikx} dx$$

$$= \int_{-\infty}^{\infty} f(x)\, e^{-2\pi ikx} dx.$$

Substituting these formulas into the Poisson summation formula from the beginning, the formula for g(x) takes a different look:

$$g(x) = \sum_{n=-\infty}^{\infty} f(x+n) = \sum_{k=-\infty}^{\infty} a_k e^{2\pi ikx} = \sum_{k=-\infty}^{\infty} [\ \int_{-\infty}^{\infty} f(u)\, e^{-2\pi iku} du\ ]\, e^{2\pi ikx} \qquad (5)$$

To allow some of the steps just carried out, some restrictions apply:

3) to allow the interchange of infinite sum and integral, $\int |\ f(x)\ |\ dx < \infty$.

4) absolute summability of the Fourier series is needed, requiring $\sum |\ a_k| < \infty$

Now at this point, this appears to be just an exercise in analysis and much ado about nothing. But this exercise becomes more interesting when the form of the function f is determined to be $f(x) = \exp\{-tx^2\}$. Clearly some of the conditions are easily satisfied:

1) $f(x) = \exp\{-tx^2\}$ is continuous in x

2) $\sum f(x+n)$ is uniformly convergent

3) $\int |\ f(x)\ |\ dx < \infty$

4) the absolute summability ($\sum |\ a_k\ | < \infty$) can be determined from the coefficients

$$a_k = \int_{-\infty}^{\infty} e^{-tu^2 - 2\pi iku}\, du = e^{-k^2\pi^2/t} \int_{-\infty}^{\infty} e^{-t(u + \frac{k\pi i}{t})^2} du = \sqrt{\tfrac{\pi}{t}}\ e^{-k^2\pi^2/t}$$

The result of this special case for f(x) can then be substituted into (5) to produce a messy, but enlightening formula:

$$\sum_{n=-\infty}^{\infty} e^{-t(x+n)^2} = \sum_{k=-\infty}^{\infty} a_k\, e^{2\pi ikx} = \sqrt{\tfrac{\pi}{t}} \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2/t}\, e^{2\pi ikx} \qquad (6)$$

The usefulness of this entire exercise can be found by setting $x = 1/2$ and $t = \pi^2/(2z^2)$. But both sides of (6) require some manipulation; the left hand side will be first, then the right.

$$\sum_{n=-\infty}^{\infty} e^{-t(\frac{1}{2}+n)^2} = 2\sum_{j=1}^{\infty} e^{-t\frac{(2j-1)^2}{4}} = 2\sum_{j=1}^{\infty} e^{-(2j-1)^2\pi^2/8z^2}$$

$$\sqrt{\tfrac{\pi}{t}} \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2/t}\, e^{2\pi ik(\frac{1}{2})} = \sqrt{\frac{\pi 2z^2}{\pi^2}} \sum_{k=-\infty}^{\infty} (-1)^k\, e^{-2k^2z^2} = z\sqrt{\tfrac{2}{\pi}} \sum_{k=-\infty}^{\infty} (-1)^k\, e^{-2k^2z^2}$$

Setting the right and left hands sides equal again yields

The number of terms that are used in any part of computing are powers of previous ones and rarely are many required. As a result, much of the work will be done in computing

$$u = e^{-\pi^2/8z^2}, \qquad v = e^{-2z^2}, \qquad \text{and} \quad w = \frac{\sqrt{2\pi}}{z}.$$

Only in a few cases are many multiplications needed, and the corruption of the last digits of the product due to roundoff will be lost in the subsequent additions.

Smirnov Distribution Function L(z)

Single Precision

Interval	Method	Formula
(0.0, .083)	0	0
(0.83, .7702497)	$K_1(z)$	w u
(.7702497, 1.334112)	$K_2(z)$	$w\ u\ (1 + u^8)$
(1.334112, 1.471763)	$L_2(z)$	$1 - 2\ v\ (1 - v^3)$
(1.471763, 2.943525)	$L_1(z)$	1 - 2 v
(2.943525, ∞)	1	1

Double Precision

Interval	Method	Formula
(0, 0.83)	0	0
(.083, .5042468)	$K_1(z)$	w u
(.5042468, .873381)	$K_2(z)$	$w\ u\ (1 + u^8)$
(.873381, 1.054708)	$K_3(z)$	$w\ u\ (1 + u^8(1 + u^{16}))$
(1.054708, 1.111156)	$L_4(z)$	$1 - 2\ v\ (1 - v^3(1 - v^5(1 - v^7)))$
(1.111156, 1.481542)	$L_3(z)$	$1 - 2\ v\ (1 - v^3(1 - v^5))$
(1.481542, 2.222313)	$L_2(z)$	$1 - 2\ v\ (1 - v^3)$
(2.222313, 4.444625)	$L_1(z)$	1 - 2 v
(4.444625, ∞)	1	1

Smirnov Tail Function 1 - L(z)

Single Precision

Interval	Method	Formula
(0.0, .2553579)	1	1
(.2553579, .7787835)	$1 - K_1(z)$	1 - w u
(.7787835, .9613513)	$1 - L_3(z)$	$2\ v\ (1 - v^3(1 - v^5))$
(.9613513, 1.665109)	$1 - L_2(z)$	$2\ v\ (1 - v^3)$
(1.665109, 9.437659)	$1 - L_1(z)$	2 v
(9.437659, ∞)	0	0

Double Precision

Interval	Method	Formula
(0, .1724320)	1	1
(.1724320, .5042468)	$1 - K_1(z)$	$1 - w\ u$
(.5042468, .873381)	$1 - K_2(z)$	$1 - w\ u\ (1 + u^8)$
(.873381, .899262)	$1 - K_3(z)$	$1 - w\ u\ (1 + u^8(1 + u^{16}))$
(.899262, 1.137486)	$1 - L_4(z)$	$2\ v\ (1 - v^3(1 - v^5(1 - v^7)))$
(1.137486, 1.557567)	$1 - L_3(z)$	$2\ v\ (1 - v^3(1 - v^5))$
(1.557567, 2.543496)	$1 - L_2(z)$	$2\ v\ (1 - v^3)$
(2.543496, 9.436593)	$1 - L_1(z)$	$2\ v$
(9.436593, ∞)	0	0

5. Acknowledgement

The author wishes to thank Dr. Ed Battiste for leading him on the right track.

6. References

R. Bellman (1961) A Brief Introduction to Theta Functions, Holt, Rhinehart, and Winston, New York.

W. Feller (1948) "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions," Annals of Mathematical Statistics, Volume 19, pp. 177-189.

B. V. Gnedenko (1962) The Theory of Probability, translated by B. D. Steckler, Chelsea, New York.

A. N. Kolmogorov (1933) "Sulla Determinazione Empirica di una Legge di Distributione," Giornale dell'Instituto Italiano Attuari, Volume 4, pp. 83-91.

N. Smirnov (1939) "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples," Bulletin Mathematique de l'Universite de Moscou, Volume 2, pp. 3-14.

N. Smirnov (1939) "Sur les Ecart de la Courbe de Distribution Empirique," (in Russian) Recueil Mathematique, Volume 6, pp. 3-26.

N. Smirnov (1948) "Table for Estimating the Goodness of Fit of Empirical Distributions," Annals of Mathematical Statistics, Volume 19, pp. 279-281.

S. S. Wilks (1962) Mathematical Statistics, Wiley, New York.