LARGE SAMPLE RESULTS FOR KOLMOGOROV-SMIRNOV STATISTICS FOR DISCRETE DISTRIBUTIONS

By Constance L. Wood and Michele M. Altavela

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York

SUMMARY

Considered are the asymptotic distributions of the Kolmogorov-Smirnov goodness-of-fit statistics when the hypothesized distribution is discrete. Each of these statistics is shown to have the same distribution as a continuous functional of an associated empirical process on the unit interval. Using known weak-convergence properties for the empirical process, the asymptotic distributions of the Kolmogorov-Smirnov statistics are derived. A discussion and example concerning the use of these results is included.

Some Key Words: Kolmogorov-Smirnov statistics; Goodness-of-fit tests; Limit distributions; Discrete distributions.

1. INTRODUCTION

Several authors have recently recommended the use of Kolmogorov-Smirnov statistics for testing goodness-of-fit to a completely specified discrete distribution. See Conover (1972), Horn (1977), and Pettitt and Stephens (1977). In particular, Coberly and Lewis (1972) and Conover (1972) both

---

Paper No. BU-595-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

give formulas for calculating the exact distributions of the one-sided Kolmogorov-Smirnov statistics. Conover (1972) also gives an approximation to the distribution of the Kolmogorov-Smirnov statistic. These computations, however, are not feasible for large sample sizes. Considered here are the asymptotic distributions of not only the one-sided Kolmogorov-Smirnov statistics but also the Kolmogorov-Smirnov statistic when the underlying distribution is discrete.

Schmid (1958) first examined the asymptotic null distributions of these statistics when the hypothesized cumulative distribution function possessed a finite number of discontinuities and was increasing between the discontinuities. It was conjectured by Schmid that his results could be extended to purely atomic distributions and distributions with a countable infinite number of discontinuities by appropriate limiting procedures. Applying a result due to Billingsley (1968) on the weak convergence of the sample distribution function, we derive the limiting distributions of the Kolmogorov-Smirnov statistics directly and thereby circumvent these limiting procedures and, hence, their justification. The limiting distributions, while not given in closed form, are presented in Section 2. A discussion and an example of how to use these results for computing significance levels is found in Section 3.

2. RESULTS

Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with common cumulative distribution function, F. We wish to test $H_0 : F(x) = H(x)$, $-\infty < x < \infty$, where H is the hypothesized, discrete distribution with all parameters, if any, specified against alternatives of the form

$$H_{11} : F(x) \geq H(x), \quad \text{with} \quad F(x) > H(x) \quad \text{for some } x ,$$

$$H_{12} : F(x) \leq H(x), \quad \text{with} \quad F(x) < H(x) \quad \text{for some } x ,$$

or

$$H_{13} : F(x) \neq H(x), \quad \text{for some } x .$$

Our test statistics are based on the sample cumulative distribution function

$$F_n(x) = (\# \text{ of } X_i\text{'s} \leq x)/n, \quad -\infty < x < \infty .$$

In particular, for a fixed x_0, $F_n(x_0)$ is a binomial proportion with probability of success equal to $F(x_0)$. Further, F_n is a strongly consistent estimator of F, uniformly in x .

Corresponding to each of the alternatives given above, an appropriate measure of discrepancy between the observed sample distribution function and the hypothesized distribution and, hence between F and H, is

$$D_n^+ = \sup_x n^{\frac{1}{2}}[F_n(x) - H(x)] , \tag{2.1}$$

$$D_n^- = \sup_x n^{\frac{1}{2}}[H(x) - F_n(x)] , \tag{2.2}$$

or

$$D_n = \sup_x n^{\frac{1}{2}}|F_n(x) - H(x)| , \tag{2.3}$$

the Kolmogorov-Smirnov statistics for testing H_{11}, H_{12} and H_{13} respectively. Since F_n and H are both step-functions,

$$D_n^+ = \max_{x \in J} n^{\frac{1}{2}}[F_n(x) - H(x)] , \tag{2.4}$$

$$D_n^- = \max_{x \in J} n^{\frac{1}{2}}[H(x) - F_n(x)] , \tag{2.5}$$

and

$$D_n = \max_{x \in J} n^{\frac{1}{2}} |H(x) - F_n(x)| , \qquad (2.6)$$

where J is the set of discontinuity points of H .

Historically, these distance measures have only been used for goodness-of-fit tests to absolutely continuous distributions, while the chi-square test has commonly been employed for discrete data. Horn (1977) gives a comprehensive review of both and their competitors. The chi-square test statistic may also be written as a measure of discrepancy between F_n and H . However, it is one which does not take into account the natural ordering among the observations, a fact exploited in analysis of attribute data. To be more specific, the chi-square test statistic is invariant to permutations of the cell labels. In contrast, the Kolmogorov-Smirnov test statistics are sensitive to the overweighting or underweighting of any tail or segment of the empirical distribution relative to the hypothesized distribution. It is from this fact that the Kolmogorov-Smirnov test statistics derive their greater powers.

One advantage of the chi-squared test is that for a fixed number of cells, it is asymptotically distribution free. It is well known that the Kolmogorov-Smirnov statistics for absolutely continuous distributions are strictly distribution-free. However, the relationships between (2.1)-(2.3) and (2.4)-(2.6) indicate that this is not true for discrete distributions. In particular, letting W° denote the tied-down Wiener Process on [0,1]; i.e., for every k and $0 \le t_1, \cdots, t_k \le 1$, $\{W^\circ(t_1), \cdots, W^\circ(t_k)\}'$ has a multivariate normal distribution with zero mean vector and

$$E\{W^\circ(t_i) \cdot W^\circ(t_j)\} = \min(t_i, t_j) - t_i t_j ,$$

we have that under the null hypothesis

THEOREM. *The limiting distributions of* D_n^+ (D_n^-) *and* D_n *are given by* $\max_{x \in J} [W^\circ\{H(x)\}]$ *and* $\max_{x \in J} |W^\circ\{H(x)\}|$ *respectively.*

As an example, suppose that the number of discontinuities of H is finite, say r . Then, for any $\lambda > 0$,

$$\lim_{n\to\infty} \Pr(D > \lambda) = \Pr[\max_{x\in J} |W^\circ\{H(x)\}| > \lambda]$$

$$= 1 - \Pr[\max_{x\in J} |W^\circ\{H(x)\}| \leq \lambda]$$

$$= 1 - \Pr[|Z_1| \leq \lambda, \; \cdots, \; |Z_{r-1}| \leq \lambda] ,$$

where $(Z_1, \cdots, Z_{r-1})'$ is a multivariate normal vector with

$$E(Z_i) = 0$$

and

$$E(Z_i \cdot Z_j) = \min\{H(x_i), \; H(x_j)\} - H(x_i)H(x_j) \quad . \tag{2.7}$$

Even though this multivariate normal probability is neither known in closed form nor computationally tractable, for a given λ it can be estimated quite readily by Monte Carlo simulation. See Section 3.

Now we will present the proof of the theorem.

PROOF OF THEOREM. Letting $J^* = \{t : t = H(x), \; x \in J\}$, we can write

$$D_n^+ = \max_{t\in J^*} n^{\frac{1}{2}} [F_n\{H^{-1}(t)\} - t] ,$$

$$D_n^- = \max_{t\in J^*} n^{\frac{1}{2}} [t - F_n\{H^{-1}(t)\} - t] ,$$

and

$$D_n = \max_{t \in J^*} n^{\frac{1}{2}} | F_n\{H^{-1}(t)\} - t | ,$$

where $H^{-1}(t) = \inf\{x : H(x) \geq t\}$. Note that for every $t \in J^*$, $F_n \circ H^{-1}(t)$ is equal to the sample distribution function of $\xi_i = H(x_i)$, $i = 1, \cdots, n$, say $H_n(t)$. Denoting the distribution function of ξ_i by H^* we also have that $H^*(t) = t$, for every $t \in J^*$. Since the maximums are to be taken only over points in J^*,

$$D_n^+ = \max_{t \in J^*} n^{\frac{1}{2}} \{H_n(t) - H^*(t)\} ,$$

$$D_n^- = \max_{t \in J^*} n^{\frac{1}{2}} \{H^*(t) - H_n(t)\} ,$$

and

$$D_n = \max_{t \in J^*} n^{\frac{1}{2}} | H_n(t) - H^*(t) | .$$

From Billingsley (1968), Theorem 16.4, we find that $[n^{\frac{1}{2}}\{H_n(t) - H^*(t)\} : 0 \leq t \leq 1]$ converges weakly to $W^\circ H^*$ in $D[0,1]$. It immediately follows from the continuous mapping theorem that the limiting distributions of D_n^+ (D_n^-) and D_n are given by

$$\sup_{t \in J^*} W^\circ\{H^*(t)\} = \sup_{x \in J} W^\circ\{H(x)\}$$

and

$$\sup_{t \in J^*} |W^\circ\{H^*(t)\}| = \sup_{x \in J} |W^\circ\{H(x)\}|$$

respectively.

The approach which we have taken to derive the limiting distribution of the Kolmogorov-Smirnov statistics can also be used to derive asymptotic results for other EDF goodness-of-fit test statistics commonly used only with continuous distributions. For a discussion of EDF statistics, see Stephens (1974). We only require that the statistic can be written as a continuous functional of the empirical process $n^{\frac{1}{2}}\{F_n(x) - x\}$, $-\infty < x < \infty$, which then can be replaced by a corresponding functional of $n^{\frac{1}{2}}\{H_n(t) - H^*(t)\}$, $0 \leq t \leq 1$.

3. AN EXAMPLE

Horn (1977) recommends the use of the one-sided Kolmogorov-Smirnov statistic D_n^- to test goodness-of-fit of health impairment scale data for insulin-dependent diabetes patients to a maximum acceptable standard distribution. Table 1 gives the relative cumulative frequencies both observed and expected under the maximum acceptable standard distribution, where the ordered categories range from "no impairment" to "death". Since the Kolmogorov-Smirnov statistics are independent of the spacing between discontinuity points, it is unnecessary to assign numerical values to these categories.

Table 1. *Insulin Dependent Diabetes Data*

	Health Impairment Level					
	1	2	3	4	5	6
Observed C.D.F.	0.000	0.500	0.633	0.867	0.933	1.000
Hypothesized C.D.F.	0.033	0.600	0.833	0.933	0.961	1.000

The one-sided Kolmogorov-Smirnov statistic, D_n^-, applied to this data gives an observed value of 1.095 with exact significance level ($n = 30$) of

.026 . To calculate the asymptotic significance level of this value, say P, we need to estimate

$$\lim_{n\to\infty} P(D_n^- \geq 1.095) = 1 - P(Z_1 < 1.095, \cdots, Z_5 < 1.095) , \qquad (3.1)$$

where $(Z_1, \cdots, Z_5)'$ is a multivariate normal (MVN) vector with zero mean vector and covariance matrix Σ given by (2.7) with H as shown in Table 1.

Ten thousand independent MVN vectors were generated with zero mean and this covariance structure. Each vector was checked to see if it fell in the region (3.1). The estimated significance level $\hat{P}$ was found to be .0143. Noting that $n^{\frac{1}{2}}F_n$ takes jumps of size $n^{-\frac{1}{2}}$, an obvious finite sample correction factor is to reduce the observed value of D_n^- by $\frac{1}{2}n^{-\frac{1}{2}}$. This results in an estimated significance of .022 which is closer to the exact significance level.

In this case, as in many situations, the order of Σ is small. This means that MVN vectors are relatively easily generated. One commonly used method is to decompose Σ into UU' where U is a lower triangular matrix; i.e., Cholesky Decomposition. For details see Forsythe and Moler (1967), Section 23. Then if $\underline{Y}$ is MVN($\underline{0}$, $I_{5\times 5}$), UY is MVN(0, Σ) .

For the example discussed in this section, ten thousand vectors were generated. In many hypothesis testing situations, less accuracy in the estimate of P is required; e.g., P is much less than .90 . Therefore we suggest at least a two-staged procedure to estimate P . The first stage yielding a rough estimate $\hat{P}$ from which it can be decided if a more accurate estimate is needed and, if so, the number of simulations required.

The author would like to thank Professor Robert J. Serfling for suggesting a simplified proof of the theorem and the referee for many helpful suggestions which led to improvements in presentation. This research was supported in part by Hatch Project 402, Biometrics Unit, Cornell University.

REFERENCES

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Coberly, W. A. and Lewis, T. O. (1972). A note on a one-sided Kolmogorov-Smirnov test of fit for discrete distribution functions. *Ann. Inst. Statist. Math. 24*, 183-187.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *J. Am. Statist. Assoc. 67*, 591-596.

Forsythe, G. and Moler, C. B. (1967). *Computer Solution of Linear Algebraic Systems*. Prentice-Hall, Englewood Cliffs, N. J.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics 33*, 237-248.

Pettitt, A. N. and Stephens, M. A. (1977). The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics 19*, 207-210.

Schmid, P. (1958). On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. *Ann. Math. Statist. 29*, 1011-1027.

Stephens, M. A. (1974). EDF statistics for goodness-of-fit and some comparisons. *J. Am. Statist. Assoc. 69*, 730-737.