NEW CRITICAL VALUES FOR THE KOLMOGOROV-SMIRNOV STATISTIC TEST FOR MULTIPLE SAMPLES OF DIFFERENT SIZE

L. FRANK[1]

Recibido: 22/01/10
Aceptado: 05/03/10

SUMMARY

Critical values of the Kolmogorov-Smirnov statistic for two samples of equal size and for three samples of different size are reviewed. The review is carried out by a self written computer program that simulates the extraction of k random uniformly distributed independent samples of size n_i (i=1…k), computes the empirical cumulative distribution function of the test statistic under the null hypothesis, and returns the value with probability 1-α. The findings suggest that published values present conspicuous lack of precision due to discretization, the rounding of figures and non-exhaustive computational search. As a consequence, when comparing two samples of equal size, published tables tend to be rather conservative, although this is not evident when comparing three samples of different size. The program presented also extends the test to the comparison of multiple samples of equal or different size.

Key words. Kolmogorov-Smirnov test, non-parametric statistics.

NUEVOS VALORES CRÍTICOS DEL ESTADÍSTICO DE KOLMOGOROV-SMIRNOV PARA MÚLTIPLES MUESTRAS DE DISTINTO TAMAÑO

RESUMEN

Se revisan los valores críticos del estadístico de Kolmogorov-Smirnov para dos muestras de igual tamaño o tres muestras de distinto tamaño. La revisión se realiza mediante un programa que simula la extracción de k muestras de tamaño n_i (i=1…k) uniformemente distribuidas, obtiene la función de distribución empírica del estadístico bajo la hipótesis nula y devuelve el valor con probabilidad 1-α. Los resultados sugieren que las tablas publicadas presentan una notable falta de precisión debida a discretización, redondeo de números y búsquedas computacionales no exhaustivas. Como consecuencia, al comparar dos muestras de igual tamaño, las tablas ya publicadas son más bien conservadoras, no así las de tres muestras de distinto tamaño. El programa presentado extiende el test a la comparación de múltiples muestras de distinto tamaño.

Palabras clave. Prueba de Kolmogorov-Smirnov, estadística no paramétrica.

INTRODUCTION

The Kolmogorov-Smirnov statistic d_n is the maximum vertical distance between two or more cumulative distribution functions (c.d.f.). In its original version (Kolmogorov, 1933) the statistic $d_n = \max |S_n(x) - F(x)|$ was introduced to test goodness of fit hypothesis, being the null H_0: $S(x) = F(x)$ while the alternative, H_1: $S(x) \neq F(x)$. $S(x)$ is the empirical c.d.f. of the random variable x and $F(x)$ is a completely specified c.d.f.. Kolmogorov demonstrated that for a sample size of $n \to \infty$, $P\{d_n \leq \lambda / n^{1/2}\} \to \Phi(\lambda)$ and computed a table of λ for $\Phi(\lambda) = 1\text{-}\alpha$, where α is the null hypothesis rejection probability. Smirnov (1939) extended the test to the comparison of $k = 2$ samples of size m and n. In this case, the null hypothesis is H_0: $S_m(x) = S_n(x)$ and the test statistic $d_{m,n} = \max |S_m(x) -$

1 Departamento de Métodos Cuantitativos, Facultad de Agronomía. Av. San Martín 4453, (C1417DSE) Buenos Aires, Argentina
Tel.: 54-11-4524-8077, e-mail: lfrank@agro.uba.ar

$S_n(x)|$. Smirnov (1948) also demonstrated that for $m = n$ (as $n\to\infty$), $P\{d_n \le \lambda/n^{1/2}\}\to\Phi(\lambda)$, and for $m \neq n$, $P\{d_{m,n} \le \lambda/[mn/(m+n)]^{1/2}\}\to\Phi(\lambda)$. Conover (1965) demonstrated that for $k>2$ samples of size n, the test statistic $d_\alpha \to \lambda/n^{1/2}$ when $n\to\infty$.

Kolmogorov-Smirnovs' asymptotic distributions $\Phi(\lambda)$ are biased for small samples (*e.g.* $n<40$). Therefore, specific tables of critical values d_α ($0.01\le\alpha\le0.20$) were computed, *e.g.* for $k = 1$ (Miller, 1956), $k = 2$ (Massey, 1951), $k = 3$ (Birnbaum and Hall, 1960) and $k\le10$ (Conover, 1980) samples of equal size, and for $k = 2$ samples of different size (Massey, 1952; Harter and Owen, 1970). However, no tables are available for $k>10$ samples of equal size, or $k>2$ samples of different size, which is a serious flaw for many experimental situations[2]. All these tables have not been reviewed since their early publication in the fifties and sixties, although they have been republished several times.

The objective of the paper is to check the accuracy of published critical values and to extend the Kolmogorov-Smirnov (KS) test to k samples of different size. Due to their widespread, all comparisons are referred to Conover's (1999) tables A19 ($k = 2, m = n$) and A20 ($k = 2, m \neq n$).

MATERIALS AND METHODS

A computer program was written in Euler-Math-Toolbox's (Grothmann, 2009) code as a function of the desired probability $1\text{-}\alpha$ and the sample sizes n_j (for $j=1\ldots k$). The program defines a function $kstat = f(1\text{-}\alpha, \mathbf{x})$, where $\mathbf{x}=[n_1\ldots n_j\ldots n_k]$ and returns the desired test statistic d_α. The algorithm is as follows:

Step 1. Generate k row vectors $\mathbf{y}_j$ (samples) with y_{ij} being uniformly distributed random numbers (observations), $y_{ji} \in [0,100]$.

Step 2. Sort all observations in $\mathbf{y}_j$ increasingly and split their relative frequency in 100 quantiles. Then, compute the cumulative frequencies and gather the k vectors $\mathbf{y}_j^*$ of size 1×10^2 in a matrix $\mathbf{Y}=[\mathbf{y}_1^*, \mathbf{y}_2^*,\ldots, \mathbf{y}_j^*,\ldots, \mathbf{y}_k^*]^T$ of dimension $k\times10^2$

Step 3. Compute the maximum distance $|y_{jj}^*\text{-}y_{ij}^*|$ between quantiles and get the vector of distances $\mathbf{y}_{max}^*$. Next, draw the maximum distance w_h^* (for $h=1,\ldots,100$) of this vector. Store the values w_h^* in vector $\mathbf{w}$.

Step 4. Go to step 1 and repeat q times.

Step 5. Sort $\mathbf{w}$ in ascending order and draw w_{max}^* such that $P\{w_{max}^*\}=1\text{-}\alpha$.

The program was used to check the critical values of the KS statistic in tables of $k = 2$ samples of size m and n (for both $m=n$ and $m\neq n$). The sample sizes were selected so as to match Conover (1999) tables A19 and A20 and the algorithm was repeated 10^5 times for reasons that will become clear later. The precision of the program was evaluated at four levels of q ($q=10^2, 10^3, 10^4, 10^5$). For this purpose, two arbitrary vectors $\mathbf{x}$ were defined, $\mathbf{x}_1=[33,33]$ and $\mathbf{x}_2=[9,15]$, and critical values $d_\alpha\times n_{max}$ (rounded to the closest integer) were computed 10^2 times for each q level. Therefore, the reader should be aware of rounding $d_n\times n_{max}$ too before comparing it with our critical values, despite no warning appears in Conover's footnotes. The computing time was also recorded as a function of the number of loops q. Tables 2 and 3 show the critical values obtained and a slightly modified version of the original code – improved to run faster – is available at http://compute.ku-eichstaett.de/MGF/wikis/euler/.

RESULTS

Table 2 differs from Conover's Table A19 in 13 out of 180 values. In all cases, the values of our table are lower than that of Conover. The same thing happens for $\lambda^*(60\le n\le100)$ with the corresponding asymptotic values λ. Table 3 also differs from Conover's Table A20, but in 89 out of 280 values, although the direction of the differences is not as clear as in the former case.

Table 1. Shows the frequency of d_α as a function of q for vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. Note that our results are nearly exact when $10^4\le q\le10^5$.

2 In reference to this point, Conover (1980) says: «[…] Actually, any of these tests could be applied to any number of samples, and the samples could be of differing sizes if tables of the distributions of the test statistics were available. […] From a practical standpoint, however, this enumeration method of considering all ordered arrangements of the combined sample is too exhaustive even for computers. At least that has been the feeling so far, except for the case of three samples of equal size».

TABLE 1. Simulation results for $P\{d \times n_{max} \leq d_\alpha \times n_{max}\}=0.95$ and different levels of q.

	x_1 = [33,33]					x_2 = [9,15]			
$d_a \times n_{max}$	10^2	10^3	10^4	10^5	$d_a \times n_{max}$	10^2	10^3	10^4	10^5
9	12	0	0	0	**6**	0	0	0	0
10	55	71	99	100	**7**	31	1	0	0
11	31	29	1	0	**8**	64	99	100	100
12	2	0	0	0	**9**	5	0	0	0

Regarding the time of computation, the program needed a few seconds to finish the algorithm for $q=10^2$ repetitions, but more than five hours for $q=10^5$. For $q=10^3$ and $q=10^4$, the times were less than 1 minute and 5 minutes, respectively[3]

TABLE 2. **Critical values of the Kolmogorov-Smirnov test statistic for two samples of equal size *n*.**

n	**0.80**	**0.85**	**0.90**	**0.95**	**0.98**	**0.99**	***n***	**0.80**	**0.85**	**0.90**	**0.95**	**0.98**	**0.99**
1	1	1	1	1	1	1	**21**	6	7	7	8	9	10
2	2	2	2	2	2	2	**22**	7	7	8	8	9	10
3	2	2	2	3	3	3	**23**	7	7	8	9	10	10
4	3	3	3	3	4	4	**24**	7	7	8	9	10	11
5	3	3	3	4	4	4	**25**	7	7	8	9	10	11
6	3	3	4	4	5	5	**26**	7	8	8	9	10	11
7	4	4	4	4	5	5	**27**	7	8	8	9	10	11
8	4	4	4	5	5	6	**28**	7	8	9	9	11	11
9	4	4	5	5	6	6	**29**	8	8	9	10	11	12
10	4	5	5	5	6	7	**30**	8	8	9	10	11	12
11	4	5	5	6	6	7	**31**	8	8	9	10	11	12
12	5	5	5	6	7	7	**32**	8	8	9	10	12	12
13	5	5	6	6	7	8	**33**	8	9	9	10	12	13
14	5	5	6	7	7	8	**34**	8	9	9	11	12	13
15	5	6	6	7	8	8	**35**	8	9	10	11	12	13
16	6	6	6	7	8	9	**36**	8	9	10	11	12	13
17	6	6	7	7	8	9	**37**	9	9	10	11	12	13
18	6	6	7	8	8	9	**38**	9	9	10	11	13	14
19	6	6	7	8	9	9	**39**	9	9	10	11	13	14
20	6	7	7	8	9	10	**40**	9	9	10	11	13	14

DISCUSSION

The newly computed critical values differ roughly 10% and 30% from those of tables A19 and A20, respectively. Nevertheless, this issue has received little attention in the statistical literature. A simple inspection of the algorithm reveals three possible sources of errors that could explain such discrepancies: 1) *discretization* of the random variable d; 2) non-exhaustive simulation of the statistic generating process; 3) rounding of the critical values. Let's consider each case in more detail.

1) All simulations of random experiments imply discretization, as the simulated values must be grouped in classes. For instance, *kstat(.)* splits the cumulative frequencies into 10^2 quantiles.

3 Computations were performed on a machine Hewlett-Packard HP325 uT(PH638LA) AMD Athlon(tm)XP 2800+ 2.08 GHz, 448 MB of RAM.

TABLE 3. **Critical values of the Kolmogorov-Smirnov test statistic for two samples of size *n* and *m*.**

n - *m*	0.80	0.85	0.90	0.95	0.98	0.99	*n* - *m*	0.80	0.85	0.90	0.95	0.98	0.99
1 - 9	8	9	9	9	9	9	**5 - 20**	10	10	11	12	14	15
1 - 10	9	10	10	10	10	10	**6 - 7**	4	4	4	5	5	6
2 - 3	2	3	3	3	3	3	**6 - 8**	4	4	5	5	6	6
2 - 4	3	3	4	4	4	4	**6 - 9**	5	5	5	6	7	7
2 - 5	4	4	4	5	5	5	**6 - 10**	5	5	6	6	7	7
2 - 6	5	5	5	6	6	6	**6 - 12**	6	6	7	7	8	9
2 - 7	5	6	6	7	7	7	**6 - 18**	8	9	9	10	12	12
2 - 8	6	6	7	7	8	8	**6 - 24**	11	11	12	14	15	16
2 - 9	7	7	8	8	9	9	**7 - 8**	4	4	5	5	6	6
2 - 10	7	8	8	9	10	10	**7 - 9**	4	5	5	6	6	7
3 - 4	3	3	3	4	4	4	**7 - 10**	5	5	6	6	7	7
3 - 5	3	3	4	4	5	5	**7 - 14**	6	7	7	8	9	10
3 - 6	4	4	4	5	6	6	**7 - 28**	12	12	13	15	16	18
3 - 7	5	5	5	6	6	7	**8 - 9**	4	5	5	6	6	7
3 - 8	5	5	6	6	7	8	**8 - 10**	5	5	5	6	7	7
3 - 9	6	6	6	7	8	8	**8 - 12**	6	6	6	7	8	8
3 - 10	6	7	7	8	9	9	**8 - 16**	7	7	8	9	10	10
3 - 12	7	8	8	9	10	11	**8 - 32**	12	13	14	16	18	19
4 - 5	3	3	4	4	4	5	**9 - 10**	5	5	5	6	7	7
4 - 6	4	4	4	5	5	5	**9 - 12**	5	6	6	7	7	8
4 - 7	4	4	5	5	6	6	**9 - 15**	6	7	7	8	9	10
4 - 8	5	5	5	6	6	7	**9 - 18**	7	8	8	9	10	11
4 - 9	5	6	6	7	7	8	**9 - 36**	13	14	15	17	19	20
4 - 10	6	6	7	7	8	8	**10 - 15**	6	6	7	8	9	10
4 - 12	7	7	8	8	9	10	**10 - 20**	8	8	9	10	11	12
4 - 16	9	9	10	11	12	13	**10 - 40**	14	15	16	18	20	21
5 - 6	4	4	4	4	5	5	**12 - 15**	6	6	7	7	8	9
5 - 7	4	4	5	5	6	6	**12 - 16**	6	6	7	8	9	9
5 - 8	4	5	5	5	6	6	**12 - 18**	7	7	8	9	10	10
5 - 9	5	5	5	6	7	7	**12 - 20**	7	8	8	9	10	11
5 - 10	5	6	6	7	7	8	**15 - 20**	7	7	8	9	10	10
5 - 15	7	8	9	9	11	11	**16 - 20**	7	7	8	9	10	10

Such discretization may cause a conspicuous upward bias (as seen in A19) in d if q is not big enough. To understand the point consider that

$$max\ |S_1(x)-S_2(x)| = max\ |[S_1(x)-F(x)]-[S_2(x)-F(x)]|,$$

$F(x)$ being a continuos function. It is obvious that $S_j(x)-F(x) \neq 0$ if $q<\infty$, so that $|[S_1(x)-F(x)]-[S_2(x)-F(x)]| \geq 0$ even if the null hypothesis is true. Neither Birnbaum and Hall (1960)–who provided the original figures – nor Conover (1999) reported the magnitude of the bias affecting table A19, although Conover (see footnote attached to A19) was aware of it. Running the program *kstat(.)* several times (setting $q=10^4$) but splitting the frequencies into 10^3 quantiles did not modify our findings. Therefore, we attribute the upward bias of A19 as equivalent to that obtained with a poor simulation of less than 10^2 classes.

2) Regarding the exhaustivity of the simulation, it is clear from Table 1 that no further gains of accuracy may be expected after 10^4 replications. Thus, Tables 2 and 3 (both computed by setting $q=10^5$) should be considered nearly exact. Recall that Massey (1952) computed the original d_α of table A20 by an iterative procedure, which he described in an earlier paper of 1951. Unfortunately, he did not provide further reference on the accuracy of

such procedure. To check the accuracy of Massey's figures we carried out a super exhaustive search of $P\{d \leq 19/45\}$ with a sample of size $\mathbf{x} = [9,15]$. We computed this probability considering 10^3 classes and $q = 10^6$ replications, and obtained $P\{d < 19/45\} = 0.8435$, a difference of order 1×10^{-2} regarding Massey's $P\{d \leq 19/45\} = 0.8342$. On the other hand, Birnbaum and Hall (1960) explicitly warn about some loss of precision in their figures (later Table A19 of Conover) despite the high capability of the computer used at the time. However, all discrepancies between this table and Table 2 go in the same direction, a pattern associated with discretization. So, unlike table A20, Table A19 seems to be less affected by a poor iterative procedure but more affected by discretization.

3) Besides non-exhaustive numerical procedures, Conover's table A20 also shows substantial rounding errors due to editorial adaptation. For example, for $n = 9$ and $m = 15$, $P\{d \leq 19/45\} = 0.8000$ in the adapted version of Conover. Rounding errors are also in line with the observed upward and downward mismatch between our values and those of Table A20, as they typically go in both directions. However, rounding of figures and lack of exhaustivity overlap in Table A20, being the former observable (by comparing Conover's and Massey's tables) and the latter unobservable but implicit in Massey's table. Again, the example of $P\{d \leq 19/45\}$ suggests that at least for some values rounding is more important than the lack of exhaustivity as a source of error.

CONCLUSION

In conclusion, we present evidence that the widely used tables of critical values exhibit clear lack of precision due to discretization, non-exhaustive numerical procedures and rounding of figures. To overcome all this drawbacks we suggest avoiding these tables, using instead a simple program such as *kstat(.)* which provides nearly exact critical values. Nevertheless, we note that *kstat(.)* becomes extremely slow when $q > 10^4$, so that for most practical situations setting q in 10^4 is an acceptable trade-off between precision and computation time.[4]

REFERENCES

BIRNBAUM, Z.W. and HALL, R.A. 1960. Small sample distributions for multisample statistics of the Smirnov type. *The Annals of Mathematical Statistics*, 31: 710-720.

CONOVER, W.J. 1965. Several k-sample Kolmogorov-Smirnov tests. *The Annals of Mathematical Statistics*, 36, 1019-1026.

CONOVER, W.J. 1980. *Practical Nonparametric Statistics*. Second Edition. Wiley.

CONOVER, W.J. 1999. *Practical Nonparametric Statistics*. Third Edition. Wiley.

GROTHMANN, R. 2009. Euler Math Toolbox v8.7. Downloadable from: http://eumat.sourceforge.net/download.html.

HARTER, H.O. and OWEN, D.B. 1970. *Selected Tables in Mathematical Statistics*. Vol. 1. Markham, Chicago.

4 The author did not explored other alternatives, perhaps more efficient, *e.g.* running several times the program but fixing q in 10^2 or 10^3 and drawing the most frequent d_α, though it exceeded the paper objectives.

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

MASSEY, F. 1951. The Distribution of the Maximum Deviation Between Two Sample Cumulative Step Functions. *The Annals of Mathematical Statistics*, 22(1): 125-128.

MASSEY, F. 1952. Distribution Table for The Deviation Between two Sample Cumulatives. *The Annals of Mathematical Statistics*, 23(3), 435-441.

MILLER, L.H. 1956. Table of porcentage point Kolmogorov statistics. *Journal of the American Statistical Association* 51: 111-121.

SMIRNOV, N.V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2(2): 3-16.

SMIRNOV, N.V. 1948. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2): 279-281.