GENERALIZED TWO SAMPLE KOLMOGOROV-SMIRNOV TEST INVOLVING A POSSIBLY CENSORED SAMPLE

Tommaso Gastaldi

Dipartimento di Statistica, Probabilitá e Statistiche Applicate

Universitá degli Studi di Roma "La Sapienza"

and

Department of Statistics

University of North Carolina at Chapel Hill

Abstract: Let $\underset{\sim}{x}$ be a sample from an unknown population X, and $\underset{\sim}{x}'$ be another sample from which some of the smallest and/or the largest observations may have been removed (*censored*). A nonparametric procedure, to test the null hypothesis stating that $\underset{\sim}{x}'$ also comes from X, against the alternative hypothesis stating that this is false, is provided. Simulation results are included.

Key Words and Phrases: censoring; two sample Kolmogorov-Smirnov test.

1. Introduction

We have a sample $\underset{\sim}{x}$ of size n ($n \geq 1$), chosen from an unknown population **X**, for which the order statistics are

$$\underset{\sim}{x} = \{x_1, \cdots, x_n\} \qquad \text{with} \qquad x_1 \leq x_2 \leq \cdots \leq x_n,$$

and another sample $\underset{\sim}{x}'$ of size r ($r \geq 1$)

$$\underset{\sim}{x}' = \{x'_1, \cdots, x'_r\} \qquad \text{with} \qquad x'_1 \leq x'_2 \leq \cdots \leq x'_r \, .$$

We suspect that $\underset{\sim}{x}'$ is a sample chosen from X or the "remainder" of a sample, say $\underset{\sim}{z}$, chosen from X, from which the s_o smallest and the s_r largest observations have been possibly removed (*censored*). Thus, we desire to test such a hypothesis against the opposite statement that this is not true.

Notation. We will refer to the following hypotheses:

H_X: $\underset{\sim}{x}'$ is a, possibly censored, sample chosen from X,

$\bar{H}_X$: $\underset{\sim}{x}'$ is not a, possibly censored, sample chosen from X,

H_C: $s_o+s_r \geq 1$, H_X (i.e., $\underset{\sim}{x}'$ is a censored sample from X),

H_L: $s_o \geq 1$, $s_r=0$, H_C (i.e., $\underset{\sim}{x}'$ is a left-censored sample from X),

H_R: $s_o=0$, $s_r \geq 1$, H_C (i.e., $\underset{\sim}{x}'$ is a right-censored sample from X),

$H_{L,R}$: $s_o \geq 1$, $s_r \geq 1$, H_C (i.e., $\underset{\sim}{x}'$ is a bilaterally censored sample from X),

H_{n_L,n_R}: $s_o=n_L$, $s_r=n_R$, $n_L+n_R \geq 0$, H_X.

Observation. The test $\bar{H}_X$ vs. $H_{0,0}$ is the well-known test of the hypothesis that two given samples $\underset{\sim}{x}$ and $\underset{\sim}{x}'$ come from the same population X.

Note. In a series of paper (1977, 1978a, 1978b), N. L. Johnson has studied the following test: *given that* H_X *is true*, $\underset{\sim}{x}$ is a censored sample vs. $\underset{\sim}{x}'$ is not a censored sample; clearly, we are considering different problems.

2. Tests of Hypotheses

We will focus on the following test comparisons

$$\bar{H}_X \text{ vs. } H_X,\ \bar{H}_X \text{ vs. } H_C,\ \bar{H}_X \text{ vs. } H_L,\ \bar{H}_X \text{ vs. } H_R,\ \bar{H}_X \text{ vs. } H_{L,R},$$
$$\left\{\bar{H}_X \text{ vs. } H_{n_L,n_R};\ n_L=0, 1, \cdots ;\ n_R=0, 1, \cdots\right\}.$$

Assumption. For each of the above tests, we assume $\overline{H}_X$ as the alternative hypothesis H_1 and the second one as the null hypothesis H_O.

Goal. The concern of this paper is to provide a general nonparametric methodology for all the above tests.

Under any of the above null hypotheses, we intend to test if the hypothetical sample $\underset{\sim}{z}$

$$\underset{\sim}{z}=\{u_1, \cdots, u_{s_O}, x'_1, \cdots, x'_r, v_1, \cdots, v_{s_r}\}$$

where the set $\{u_1, \cdots, u_{s_O}, v_1, \cdots, v_{s_r}\}$ *is subject to some restrictions*, say $R(H_O)$, *depending on the null hypothesis*, is chosen from the same population as the sample $\underset{\sim}{x}$. Obviously, the above mentioned restrictions are:

$$R(H_X)=\{s_O\geq 0,\ s_r\geq 0\},\quad R(H_C)=\{s_O+s_r\geq 1\},$$
$$R(H_L)=\{s_O\geq 1,\ s_r=0\},\quad R(H_R)=\{s_O=0,\ s_r\geq 1\},$$
$$R(H_{L,R})=\{s_O\geq 1,\ s_r\geq 1\},\quad R(H_{n_L,n_R})=\{s_O=n_L,\ s_r=n_R\}.$$

We can perform any of the above tests using, for instance, the Kolmogorov-Smirnov statistic (e.g., Kolmogorov (1933)) and accept the null hypothesis at level α if there exists $\underset{\sim}{z}$, *subject to the restrictions* $R(H_O)$, such that

$$(1)\qquad D_{n,r+s_O+s_r} \equiv \max_{x} \left|F_{\underset{\sim}{z}}(x)-F_{\underset{\sim}{x}}(x)\right| \leq K_\alpha$$

where $F_{\underset{\sim}{x}}(x)$ is the *empirical* c.d.f. of $\underset{\sim}{x}$, $F_{\underset{\sim}{z}}(x)$ is the *empirical* c.d.f. that $\underset{\sim}{z}$ assumes under H_O, and K_α is the α-th quantile of the $D_{n,r+s_O+s_r}$ distribution.

We now consider explicit tests for any of the null hypotheses of concern, distinguishing four cases.

Case I: when $x'_1\leq x_1$ and $x_n\leq x'_r$.

In this case, under any of the considered null hypotheses, we have

$$(2)\qquad F_{\underset{\sim}{z}}(x) = \frac{s_O}{r+s_O+s_r} + \frac{r}{r+s_O+s_r} F_{\underset{\sim}{x}'}(x) \qquad (x'_1\leq x\leq x'_r).$$

Thus, we will accept H_O at level α against the hypothesis $\bar{H}_X$, stating that $\underset{\sim}{x}'$ is not a, possibly censored (*in the way that* H_O *specifies*), sample coming from the same population as $\underset{\sim}{x}$, if there exist s_O and s_r, subject to the restrictions $R(H_O)$, such that

$$(3) \qquad D_{n,r+s_O+s_r} \equiv \max_{x'_1 \le x \le x'_r} \left| \frac{s_O}{r+s_O+s_r} + \frac{r}{r+s_O+s_r} F_{\underset{\sim}{x}'}(x) - F_{\underset{\sim}{x}}(x) \right| \le K_\alpha.$$

Otherwise we will reject H_O.

Case II: when $x_1 < x'_1$ and $x_n \le x'_r$.

In this case, it is straightforward to show that the following inequality holds

$$(4) \qquad I(n,r,s_O,s_r) \le D_{n,r+s_O+s_r} \le S_O(n,r,s_O,s_r)$$

where

$$(5) \qquad I(n,r,s_O,s_r) \equiv \max_{x'_1 \le x \le x'_r} \left| \frac{s_O}{r+s_O+s_r} + \frac{r}{r+s_O+s_r} F_{\underset{\sim}{x}'}(x) - F_{\underset{\sim}{x}}(x) \right|$$

$$(6) \qquad S_O(n,r,s_O,s_r) \equiv \max\left\{ \frac{s_O+1}{r+s_O+s_r}, F_{\underset{\sim}{x}}(x'_1), I(n,r,s_O,s_r) \right\}.$$

We will accept H_O at level α if there exist s_O and s_r, satisfying $R(H_O)$, such that

$$(7) \qquad S_O(n,r,s_O,s_r) \le K_\alpha.$$

We will reject H_O at level α if, for each choice of s_O and s_r in the set defined by $R(H_O)$, we have

$$(8) \qquad I(n,r,s_O,s_r) > K_\alpha.$$

If none of the two above circumstances occurs, the test is indeterminate. Similar results apply in the remaining two cases.

Case III: when $x'_1 \le x_1$, $x'_r < x_n$.

In this case, we have

$$(9) \qquad I(n,r,s_O,s_r) \le D_{n,r+s_O+s_r} \le S^r(n,r,s_O,s_r)$$

where

(10) $$S^r(n,r,s_O,s_r) = \max\left\{\frac{s_r}{s_O+r+s_r},\ 1-F_{\underline{x}}(x'_r),\ I(n,r,s_O,s_r)\right\}.$$

Case IV: when $x_1 < x'_1$, $x'_r < x_n$.

In this case, we have

(11) $$I(n,r,s_O,s_r) \le D_{n,r+s_O+s_r} \le S_O^r(n,r,s_O,s_r)$$

where

(12) $$S_O^r(n,r,s_O,s_r) = \max\left\{S_O(n,r,s_O,s_r),\ S^r(n,r,s_O,s_r)\right\}.$$

Note. The above test procedure can be considered a generalization of the two sample Kolmogorov-Smirnov test.

3. Simulation Results

In order to investigate the behaviour of the statistics $I(n,r,s_O,s_r)$, $D(n,r,s_O,s_r)$ and $S(n,r,s_O,s_r)$, we have done numerous simulations. In Tables 1-4, the results of some of the simulations are summarized. Each table refers to a generation of 1,000 pairs of random samples from a Uniform(0,1) population. In each simulation, the four possible cases we considered in section 2. are sorted. (Note: In Case I, we have put S=I=D.)

In most of the situations in which realistic values of s_O and s_r have been considered (i.e., not very large values), the empirical distributions of $I(n,r,s_O,s_r)$ and $S(n,r,s_O,s_r)$ are nearly coincident. In practice, this ensures that the chance of the test being indeterminate is very small (at least for the hypotheses H_{n_L,n_R}; $n_L=0,1,\cdots$; $n_R=0,1,\cdots$).

Table 1

n= 10 , r= 9 , s0= 1 , sr= 0
# of simulated pairs of samples= 1000

Means

	I	D	S	S-I	# of occurrences
Case I	0.3538	0.3538	0.3538	0.0000	106
Case II	0.3282	0.3411	0.3616	0.0333	411
Case III	0.3694	0.3694	0.3694	0.0000	124
Case IV	0.3178	0.3270	0.3423	0.0245	359
Overall	0.3323	0.3409	0.3548	0.0225	1000

Variances

	I	D	S	S-I	# of occurrences
Case I	0.0114	0.0114	0.0114	0.0000	106
Case II	0.0143	0.0145	0.0166	0.0038	411
Case III	0.0178	0.0178	0.0178	0.0000	124
Case IV	0.0127	0.0130	0.0131	0.0029	359
Overall	0.0138	0.0141	0.0150	0.0026	1000

	Maxima of S-I	Percentage of (S-I=0)
Case I	0.0000	100.0000
Case II	0.2000	69.5864
Case III	0.0000	100.0000
Case IV	0.2000	75.2089
Overall	0.2000	78.6000

Table 2

n= 10 , r= 8 , s0= 1 , sr= 1
# of simulated pairs of samples= 1000

Means

	I	D	S	S-I	# of occurrences
Case I	0.3478	0.3478	0.3478	0.0000	46
Case II	0.3518	0.3627	0.3793	0.0275	193
Case III	0.3543	0.3579	0.3761	0.0218	197
Case IV	0.3172	0.3340	0.3626	0.0454	564
Overall	0.3326	0.3449	0.3678	0.0352	1000

Variances

	I	D	S	S-I	# of occurrences
Case I	0.0073	0.0073	0.0073	0.0000	46
Case II	0.0125	0.0131	0.0148	0.0035	193
Case III	0.0143	0.0147	0.0164	0.0017	197
Case IV	0.0119	0.0118	0.0134	0.0039	564
Overall	0.0123	0.0124	0.0140	0.0032	1000

	Maxima of S-I	Percentage of (S-I=0)
Case I	0.0000	100.0000
Case II	0.2000	77.2021
Case III	0.1000	73.6041
Case IV	0.2000	53.1915
Overall	0.2000	64.0000

Table 3

n= 30 , r= 29 , s0= 1 , sr= 0
# of simulated pairs of samples= 1000

Means

	I	D	S	S-I	# of occurrences
Case I	0.2108	0.2108	0.2108	0.0000	102
Case II	0.2059	0.2064	0.2074	0.0015	410
Case III	0.2236	0.2236	0.2236	0.0000	117
Case IV	0.2034	0.2039	0.2055	0.0021	371
Overall	0.2075	0.2079	0.2089	0.0014	1000

Variances

	I	D	S	S-I	# of occurrences
Case I	0.0047	0.0047	0.0047	0.0000	102
Case II	0.0043	0.0042	0.0043	0.0001	410
Case III	0.0057	0.0057	0.0057	0.0000	117
Case IV	0.0039	0.0039	0.0039	0.0001	371
Overall	0.0043	0.0043	0.0043	0.0001	1000

	Maxima of S-I	Percentage of (S-I=0)
Case I	0.0000	100.0000
Case II	0.0667	95.3659
Case III	0.0000	100.0000
Case IV	0.0667	94.6092
Overall	0.0667	96.1000

Table 4

n= 50 , r= 46 , s0= 2 , sr= 2
# of simulated pairs of samples= 1000

Means

	I	D	S	S-I	# of occurrences
Case I	0.1600	0.1600	0.1600	0.0000	10
Case II	0.1619	0.1623	0.1633	0.0013	104
Case III	0.1724	0.1725	0.1730	0.0007	123
Case IV	0.1600	0.1603	0.1629	0.0029	763
Overall	0.1617	0.1620	0.1642	0.0024	1000

Variances

	I	D	S	S-I	# of occurrences
Case I	0.0004	0.0004	0.0004	0.0000	10
Case II	0.0018	0.0018	0.0018	0.0001	104
Case III	0.0029	0.0029	0.0029	0.0000	123
Case IV	0.0025	0.0025	0.0024	0.0001	763
Overall	0.0024	0.0024	0.0024	0.0001	1000

	Maxima of S-I	Percentage of (S-I=0)
Case I	0.0000	100.0000
Case II	0.0600	96.1538
Case III	0.0400	97.5610
Case IV	0.0600	88.5976
Overall	0.0600	90.6000

ACKNOWLEDGEMENTS

This paper was written while I was visiting the Department of Statistics of the University of North Carolina at Chapel Hill, in 1990.

I worked under the helpful guidance of Prof. N. L. Johnson, whose suggestions and advice are incorporated in this article, and to whom I would like to express my gratitude.

REFERENCES

Johnson, N. L.(1977). The study of possibly incomplete samples. *Proc. Fifth Brasov Conference, Sept. 1974,* 59-73, Bucharest: Acad. Rep. Soc. Rom.

Johnson, N. L.(1978a). Completeness comparison among sequences of samples. In *Contributions to survey Sampling and Applied Statistics (Papers in Honor of H. O. Hartley)* (Ed. H.A. David), 259-275, New-York: Academic Press.

Johnson, N. L.(1978b). Completness comparisons among sequences of samples, II. Censoring from above or below and general censoring. *Journal of Statistical Planning and Inference*, 2, 107-123.

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

Kozioł J. A. - Byar D. P.(1975). Percentage Points of the Asymptotic Distributions of One and Two Sample K-S Statistics for Truncated or Censored Data. *Technometrics*, vol. 17, No. 4, 507-510.