An Investigation of the Kolmogorov-Smirnov Two Sample Test using SAS®

Tison Bolen, Dawit Mulugeta, Jason Greenfield, Lisa Conley, Cardinal Health,

Advanced Analytics Team, Dublin, Ohio 43017, USA

ABSTRACT

The Kolmogorov-Smirnov (K-S) test is one of the most useful and general nonparametric methods for comparing two samples. It is sensitive to all types of differences between two populations (shift, shape, etc.). In this paper, we will present a thorough investigation into the K-S test including: discussion of the formal test procedure, practical demonstration of the test, large sample approximation of the test statistic and ease of use in SAS® using the NPAR1WAY procedure.

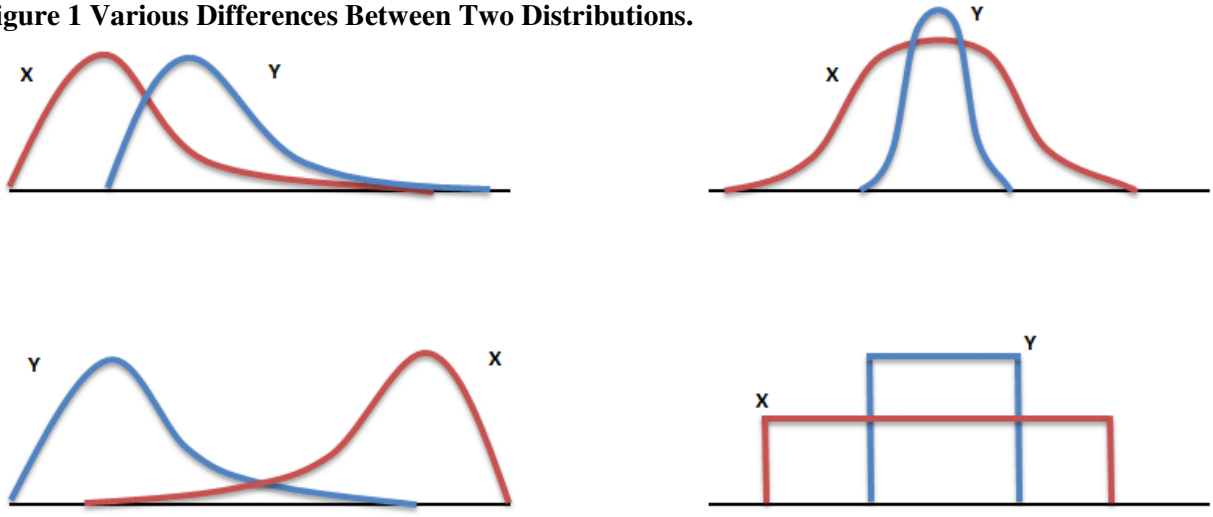
INTRODUCTION

The Kolmogorov-Smirnov test is a nonparametric procedure used to test for the equality of continuous, one-dimensional probability distributions which can be extended for the comparison of two independent samples. The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function (cdf) of the reference distribution [3,4,5,6, 7].

Our investigation of the K-S test will focus on the two sample two-sided version. This is the simplest form which detects all cases of the alternative hypothesis [1, 2]. Variations of the K-S test can be used as general and flexible goodness-of-fit tests, specifically for situations when specific tests are yet to be developed [3], these situations will not be addressed in this paper.

The K-S test is sensitive to all types of differences between two populations and can detect any of these differences (Figure 1).

Figure 1 Various Differences Between Two Distributions.



Consider two independent samples from different populations

$X = \{x_1, x_2, \cdots, x_m\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$.

The X population (or probability distribution) is completely described by the cumulative distribution function, defined, $\forall t \in \Re$ by $F(t) = P(X \le t)$ [2].

Similarly, the Y population (probability distribution) is completely described by is cdf $G(t) = P(Y \le t)$.

If the X and Y populations are identical, then $F(t) \equiv G(t), \forall t$ [1,2].

If the two distributions are not identical, then $F(t) \neq G(t)$, for some t [1,2].

The most useful version of the K-S test is the two-sided version for testing, provided below, because it is the simplest test that detects all cases of H_1 [1,2,3,4,5,6,7].

$H_0 : F(t) \equiv G(t), \forall t \in \Re$ versus $H_1 : F(t) \neq G(t),$ for some t [1, 2].

MATERIALS AND METHODS

The idea of the K-S two sample two-sided test is to estimate the function F by a function F_m determined from the X sample. In a similar fashion, estimate G by a function G_n from the Y sample. In order to decide whether $F(t) \equiv G(t), \forall t$, look at how closely the estimated functions $F_m(t)$ and $G_n(t)$ match up [1, 2].

In order to estimate the empirical distribution functions we consider the following:

$$F_m(t) = \frac{\#\ of\ \ observed\ \ x's \le t}{m} \text{ and } G_n(t) = \frac{\#\ of\ \ observed\ \ y's \le t}{n} \text{ defined } \forall t \in \Re.$$

The empirical cdf $F_m(t)$ is always a "step" function with jumps of size $1/m$ at the observed values of X. If the true probability distribution of X is continuous, then the true cdf $F(t)$ is a continuous graph which is approximated by the step function $F_m(t)$ [1]. Same goes for $G_n(t)$ and $G(t)$ for the Y distribution. If the null hypothesis is true and $H_0 : F(t) \equiv G(t), \forall t \in \Re$, then $F_m(t)$ and $G_n(t)$ should be close $\forall t \in \Re$. So we will look at the test based on $\max_t \left| F_m(t) - G_n(t) \right|$ [1, 2].

Formal test procedure for a level α test of:

$H_0 : F(t) \equiv G(t), \forall t \in \Re$ versus $H_1 : F(t) \neq G(t),$ for some t.

Compute $J = \frac{mn}{d} \bullet \max_t \left| F_m(t) - G_n(t) \right|$, where d is the greatest common divisor of m and n [1,2].

If $J \geq j_{\alpha,m,n}$ conclude H_1, otherwise H_0, where $j_{\alpha,m,n}$ is the upper α quartile of the distribution of J under H_0 [1,2]. Since the difference of the empirical cdf's is a step function, the maximum over all $t \in \Re$ will only occur at an observed value of X or Y. These are the only values for which $|F_m(t) - G_n(t)|$ will change. In the case of ties we still compute the maximum and therefore we only need to evaluate the difference at these distinct values of $t \in \Re$ [1, 2].

PRACTICAL EXAMPLE

A company studied two techniques of assembling a part. Forty workers were randomly selected from the worker population and eight were randomly assigned to each technique. The worker assembled a part and the measurement was the amount of time in seconds required to complete the assembly. Some workers did not complete the task [8].

Table 1 Data for Comparing Techniques of Assembling a Part.

Technique 1		Technique 2	
Worker	Time	Worker	Time
1	41.0	6	45.6
2	49.1	7	41.0
3	49.2	8	46.4
4	54.8	9	50.7
5	45.0	10	47.9
		11	44.6

We have $j(\alpha = 0.1082, m = 5, n = 6) = 20$ [1, 2]. So a level $\alpha = 0.1082$ level test is to conclude H_1 if $J \geq 20$, else fail to reject H_0. Order the observations and compute $|F_m(t) - G_n(t)|$ for each observed value [1, 2, 8].

Table 2 Calculation of the Absolute Maximum Difference of the Empirical Cumulative Distributions.

X	Y	$\lvert F_m(t) - G_n(t) \rvert$
41.0	41.0	|1/5 - 1/6| = 1/30
	44.6	|1/5 - 2/6| = 4/30
45.0	45.6	|2/5 - 3/6| = 3/30
	46.4	|2/5 - 4/6| = 8/30
	47.9	|2/5 - 5/6| = 13/30
49.1		|3/5 - 5/6| = 7/30
49.2		|4/5 - 5/6| = 1/30
	50.7	|4/5 - 6/6| = 6/30
54.8		|5/5 - 6/6| = 0

From (Table 1) [8] and (Table2), we have $\max_t \left|F_m(t) - G_n(t)\right| = 13/30$ and d = {the greatest common divisor of m and n } = 1 [1, 2].

Therefore, $J = \frac{5(6)}{1} \bullet \frac{13}{30} = 13 < 20 \Rightarrow$ fail to reject H_0. We consult our reference [1,2] to determine the p-value, however, the table only provides limited critical values. The p-value is given by $P_0(J \geq 13) > 0.1082$. The value of J depends only on the relative rank ordering of the $X's$ and $Y's$ which implies that J is distribution free under the null hypothesis and explains why this is a nonparametric test [1, 2].

LARGE SAMPLE APPROXIMATION EXAMPLE

In order to perform the large sample approximation for the K-S problem let's first consider the following: $J^* = \frac{d}{\sqrt{mn(m+n)}} J$, where the upper α quartiles of the probability distribution of J^* can be found from the $Q(s)$ values [1, 2].

The large sample approximation level α test is given by:

If $J^* \geq q^*_\alpha$ conclude H_1, otherwise H_0, where q^*_α is the upper α quartile of the distribution of J^* under H_0 [1, 2]. Using the data from the example that we discussed earlier [8], we have $J^* = \frac{13}{\sqrt{5(6)(11)}} = 0.7156 \Rightarrow H_0$ and the P-value is $P_0(J^* \geq 0.72) \approx Q(0.72) = 0.6851$ [9].

EXAMPLE USING PROC NPAR1WAY in SAS®

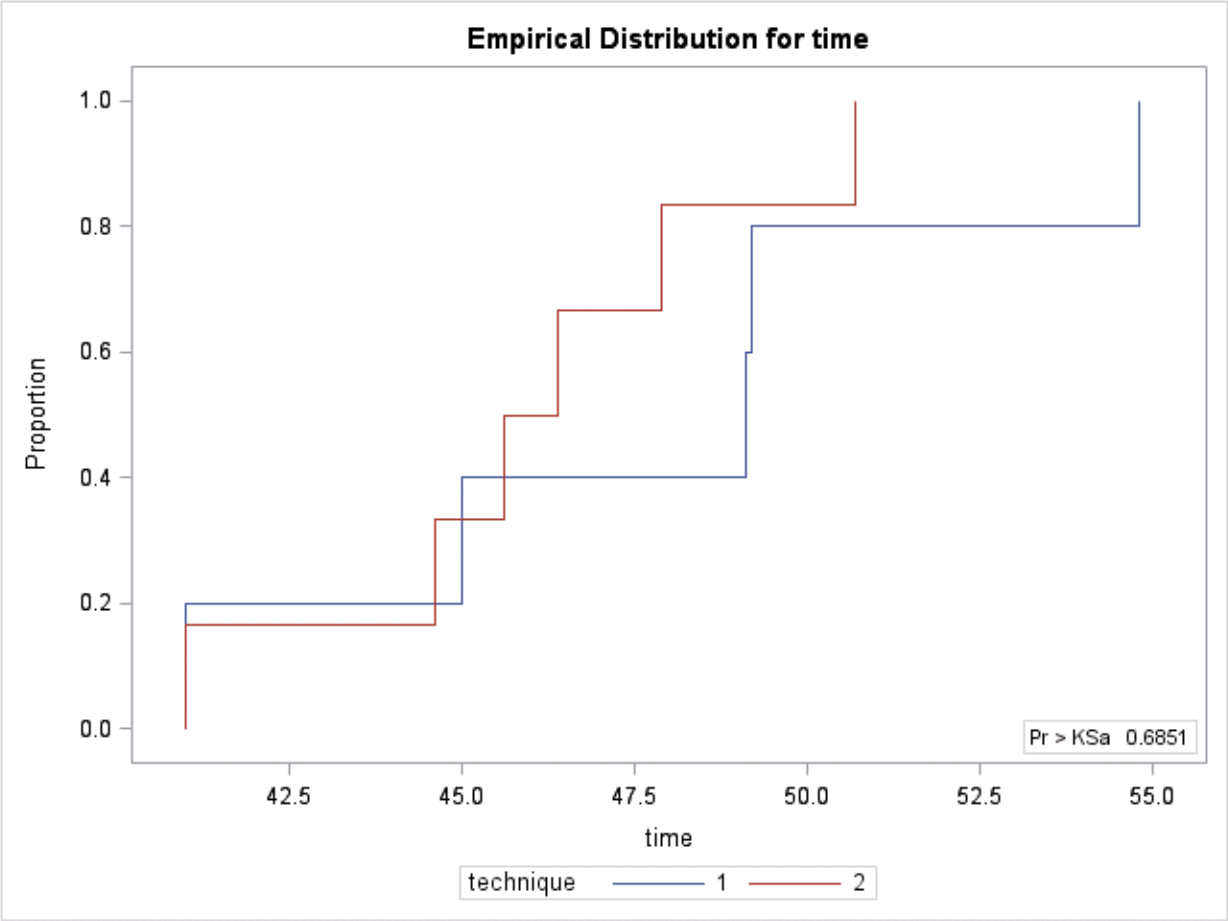
Now that we have investigated the K-S two sample test manually, let us demonstrate how easily the example presented in (Table 1) [8] can be handled using the SAS® procedure NPAR1WAY. First we read in the data using a SAS® datastep (Figure 2).

Figure 2 SAS® Datastep and NPAR1WAY Procedure Code.

```
data work.technique ;
    input technique worker time @@;
    datalines ;
    1  1  41.0 1  2  49.1 1  3  49.2 1  4  54.8 1  5  45.0  2  6  45.6
    2  7  41.0 2  8  46.4 2  9  50.7 2 10  47.9 2 11  44.6
run ;
ods graphics on ;
proc npar1way edf plots = edfplot data = work.technique ;
class technique ;
var time ;
exact ks ;
run ;
ods graphics off ;
```

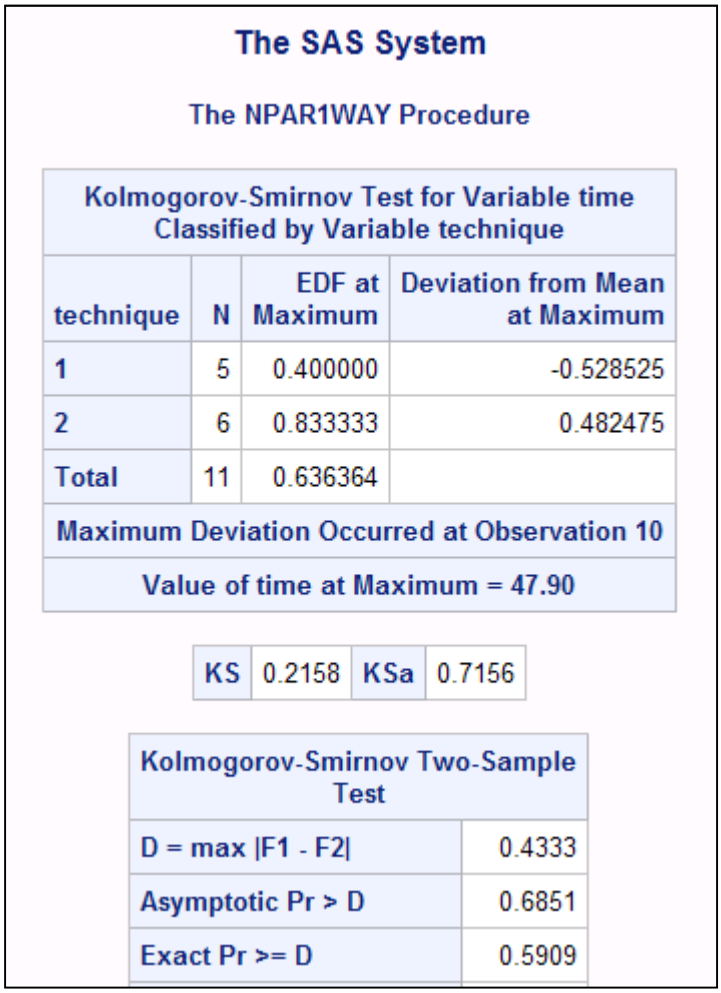
Next we inspect the empirical cumulative distribution functions (edf) and calculate the exact and the large sample approximation of J and the associated p-values using proc NPAR1WAY. The SAS® code (Figure 2) is short and simple to follow. The "edf plots" command will produce a graph (Figure 3) of the two empirical cumulative distributions. From the plot we can visually inspect that there appears to be a rather significant divergence between the two cdf's somewhere between time 47 and 48, which is consistent with where we determined the maximum to be in the practical example.

Figure 3 SAS® Plot of the Empirical Distributions for time.



A closer inspection into the output provided by proc NPAR1WAY (Figure 4) confirms that the maximum deviation occurs at time 47.9 which agrees with our earlier discovery. We also see from the output (Figure 4), that our large sample p-value approximation calculation agrees with proc NPAR1WAY. Recall the exact p-value that we looked up, and we were able to determine $P_0(J \geq 13) > 0.1082$, but were unable to produce a more precise probability due to the limitations of the table [1]. The "exact ks" command will produce the exact p-value. We can see from the output (Figure 4) that the exact p-value is 0.5909. Of course since 0.5909 > 0.1082 this, too, agrees with our earlier discovery and the fact that we failed to reject H_0.

Figure 4 SAS® NPAR1WAY K-S Output.



The SAS System

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable time Classified by Variable technique			
technique	N	EDF at Maximum	Deviation from Mean at Maximum
1	5	0.400000	-0.528525
2	6	0.833333	0.482475
Total	11	0.636364	
Maximum Deviation Occurred at Observation 10			
Value of time at Maximum = 47.90			

KS	0.2158	KSa	0.7156

Kolmogorov-Smirnov Two-Sample Test	
D = max |F1 - F2|	0.4333
Asymptotic Pr > D	0.6851
Exact Pr >= D	0.5909

CONCLUSION

Taking advantage of making no assumptions about the distribution of the data [10], the K-S two-sided test is one of the most useful and general tests for making inferences about two independent samples [2,3,4,5,6]. It is important to note, that this generality comes at some cost, other parametric tests may be more sensitive to differences if the data meet the distributional requirements (for example, Student's t-test) [10]. As most nonparametric tests, the K-S test is extremely practical and easy to understand [1, 2].
We conducted our inspection of the K-S test using the part assembly data [8], both from a practical (by-hand) aspect and using SAS®. The NPAR1WAY procedure is very robust and provides excellent output and plots. It is a quick and easy way to perform a variety of nonparametric tests, including the K-S test. It is our opinion that if one wishes to compare two independent samples, for which the distributional assumptions of other tests cannot be met, then the K-S test is an excellent alternative.

REFERENCES

[1] Myles Hollander and Douglas A.Wolfe. (1999) Nonparametric Statistical Methods, 2nd Edition. Wiley Series in Probability and Statistics

[2] Douglas Critchlow, (2005) STAT 661 Applied Nonparametric Statistics Course Notes, The Ohio State University

[3] Justel, A., Peña, D. and Zamar, R. (1997) *A multivariate Kolmogorov-Smirnov test of goodness of fit*, Statistics & Probability Letters, 35(3), 251-259

[4] Eadie, W.T.; D. Drijard, F.E. James, M. Roos and B. Sadoulet (1971). *Statistical Methods in Experimental Physics*. Amsterdam: North-Holland. pp. 269–271

[5] Stuart, Alan; Ord, Keith; Arnold, Steven [F.] (1999). *Classical Inference and the Linear Model*. Kendall's Advanced Theory of Statistics **2A** (Sixth ed.). London: Arnold. pp. 25.37–25.43

[6] Corder, G.W., Foreman, D.I. (2009).*Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach* Wiley

[7] Stephens, M.A. (1979) *Test of fit for the logistic distribution based on the empirical distribution function*, Biometrika, 66(3), 591-5

[8] George A. Milliken and Dallas E. Johnson. (2009)Analysis of Messy Data, Volume 1, Designed Experiments, 2nd Edition. CRC Press, New York.

[9] Frank J. Massey, Jr., (1951), *The distribution of maximum deviation between two sample cumulative step functions,* Annals of Math. Stat., Vol. 22, pp. 125-128

[10] http://www.physics.csbsju.edu/stats/KS-test.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the authors at:

Tison Bolen
Manager, Advanced Analytics Team
Cardinal Health, Dublin, OH
Email: tison.bolen@cardinalhealth.com

Dawit Mulugeta, Ph.D.
Manager, Advanced Analytics Team
Cardinal Health, Dublin, OH
Email: dawit.mulugeta@cardinalhealth.com

Jason Greenfield
Consultant, Advanced Analytics Team
Cardinal Health, Dublin, OH
Email: jason.greenfield@cardinalhealth.com

Lisa Conley
Director, Advanced Analytics Team
Cardinal Health, Dublin, OH
Email: lisa.conley@cardinalhealth.com