Statistics and Probability Letters 116 (2016) 65–71

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

An exact Kolmogorov–Smirnov test for whether two finite populations are the same

Jesse Frey

Department of Mathematics and Statistics, Villanova University, Villanova, PA 19085, United States

ARTICLE INFO

Article history:
Received 26 February 2016
Accepted 19 April 2016
Available online 27 April 2016

Keywords:
Nonparametric hypothesis testing
Recursion
Sampling without replacement

ABSTRACT

We develop an algorithm for finding exact critical values for the two-sample Kolmogorov–Smirnov test in the finite-population case. We then compare these exact critical values to the asymptotic values that are available in the literature. The asymptotic critical values work well for equal sample sizes, but can be excessively conservative when the sample sizes differ.

1. Introduction

Given simple random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from two infinite populations, one often wishes to test for a difference between the two populations. Such tests can be done either in parametric fashion, as with the familiar two-sample t test, or in nonparametric fashion. One commonly used nonparametric test is the two-sample Kolmogorov–Smirnov test.

Let $F(t)$ and $G(t)$ be the distribution functions for the two populations. The null and alternative hypotheses for the two-sample Kolmogorov–Smirnov test are then $H_0 : F(t) = G(t)$ *for all* t and $H_1 : F(t) \neq G(t)$ *for some* t. One computes the empirical distribution functions $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$ and $\hat{G}(t) = \frac{1}{m}\sum_{j=1}^{m} I(Y_j \leq t)$, where $I(A)$ is the indicator function that is one if A holds and zero otherwise. The test statistic is then $D_{KS} = \sup_t |\hat{F}(t) - \hat{G}(t)|$, and one rejects H_0 when D_{KS} is excessively large. Exact critical values were tabled by Kim and Jennrich (1974), and accurate asymptotic approximations have also been developed (see Kim, 1969).

Suppose now that the two populations are finite. The two populations can then be the same only if both are of the same size, say N, and the simple random sampling would be done without replacement to maximize the representativeness of the sample. Since the populations are finite, the population distribution functions are no longer continuous. Instead, they are step functions given by $F(t) = \frac{1}{N}\sum_{i=1}^{N} I(x_i \leq t)$ and $G(t) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \leq t)$, where $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_N\}$ are the values for the two populations. We may still test $H_0 : F(t) = G(t)$ *for all* t versus $H_1 : F(t) \neq G(t)$ *for some* t by using the test statistic $D_{KS} = \sup_t |\hat{F}(t) - \hat{G}(t)|$, but the critical values from the infinite-population case are no longer appropriate. Instead, since the population is gradually exhausted as the sample size increases, the critical value for a given α level tends to be smaller here than in the infinite-population case. O'Neill and Stern (2012) showed how to obtain asymptotic critical values by multiplying the infinite-population critical values by appropriate constants. In this paper, we show how to obtain critical values that are exact even for small samples.

We first, in Section 2, develop an algorithm for computing probabilities of the form $P(D_{KS} \leq c)$ under the null hypothesis in the finite-population case. Probabilities of the form $P(D_{KS} > c)$ can be then obtained via the complement rule, and these probabilities can be used to obtain exact critical values appropriate for testing at any desired α level. In Section 3, we use

E-mail address: jesse.frey@villanova.edu.

http://dx.doi.org/10.1016/j.spl.2016.04.016
0167-7152/

the algorithm from Section 2 to compare the sizes for level-α tests that use exact critical values and the sizes for level-α tests that use the asymptotic critical values of O'Neill and Stern (2012). What we find is that while the asymptotic critical values perform well when $n = m$, they can be excessively conservative when n and m differ. We conclude with a discussion in Section 4.

2. The algorithm

Assume known population size N and sample sizes n and m. Suppose that we wish to compute $P(D_{KS} \le c)$ for specified $c \ge 0$ under the null hypothesis. The two populations are then exactly the same, and we assume for now that all N population values are distinct, though we relax this assumption later on. Since the test statistic $D_{KS} = \sup_t |\hat{F}(t) - \hat{G}(t)|$ is unchanged when we apply the same monotone transformation to both samples, we may assume without loss of generality that the N population values are the integers 1 to N.

Choosing the first sample requires selecting n values without replacement from the N population values, and choosing the second sample requires independently choosing m values without replacement from the N population values. Thus, there are $\binom{N}{n}\binom{N}{m}$ ways to choose the two samples, and all such choices are equally likely. As a result, the probability $P(D_{KS} \le c)$ may be written as $T/\left(\binom{N}{n}\binom{N}{m}\right)$, where T is the number of possible pairs of samples such that $D_{KS} \le c$.

In order for a sample to satisfy $D_{KS} \le c$, we need $|\hat{F}(t) - \hat{G}(t)| \le c$ at all points t. Let K be the set of all integer pairs $(i, j),\ 0 \le i \le n,\ 0 \le j \le m$, such that $|i/n - j/m| \le c$, and define count functions $I(t) = \#\{X_i \le t\}$ and $J(t) = \#\{Y_j \le t\}$. A sample then satisfies $D_{KS} \le c$ if and only if, for every real number t, the ordered pair $(I(t), J(t))$ is in the set K.

To develop an efficient algorithm for computing T, we first consider a small example. Suppose that $n = m = 3$ and that we wish to compute $P(D_{KS} \le 0.5)$. The set K of allowable pairs (i, j) is then as shown in Table 1. Specifically, Table 1 shows values for the indicator function $K(i, j)$ that is one if $(i, j) \in K$ and zero otherwise. Assuming $N \ge 5$, one pair of samples that gives $D_{KS} \le 0.5$ is $\{1, 2, 4\}$ (first sample) and $\{2, 3, 5\}$ (second sample). With this pair of samples, we have

$$(I(t), J(t)) = \begin{cases} (0,0), & t < 1, \\ (1,0), & 1 \le t < 2, \\ (2,1), & 2 \le t < 3, \\ (2,2), & 3 \le t < 4, \\ (3,2), & 4 \le t < 5, \\ (3,3), & t \ge 5. \end{cases}$$

Thus, as t increases, the ordered pair $(I(t), J(t))$ goes from $(0, 0)$ to $(n, m) = (3, 3)$ in five steps, with the consecutive steps being of sizes $(1, 0)$, $(1, 1)$, $(0, 1)$, $(1, 0)$, and $(0, 1)$, respectively. A step of size $(1, 0)$ occurs at any value t that is in the first sample only; a step of size $(0, 1)$ occurs at any value t that is in the second sample only; and a step of size $(1, 1)$ occurs at any value t that appears in both samples. Only these three step sizes are possible.

Any choice of samples that leads to the same sequence of five steps that we obtained with our samples $\{1, 2, 4\}$ and $\{2, 3, 5\}$ would also give $D_{KS} \le 0.5$, and one such choice of samples is determined by choosing the five values (1, 2, 3, 4, and 5 in our case) where the steps occur. Thus, the number of samples where $(I(t), J(t))$ goes through the same steps that we saw in our example is $\binom{N}{5}$. Had the path for the sample we wrote down required four steps rather than five, the number of samples with the same sequence of steps would be $\binom{N}{4}$, and in general, the number of samples that give a particular sequence of k ordered steps is $\binom{N}{k}$. The minimum possible number of steps required to go from $(0, 0)$ to (n, m) is $\max\{n, m\}$, which occurs when the units in the smaller sample all also appear in the larger sample, thus maximizing the amount of overlap between the two samples, and the maximum possible number of steps is $n + m$, which occurs when the two samples do not overlap at all. Thus, taking $\binom{a}{b} = 0$ if $a < b$, we have that

$$T = \sum_{k=\max\{n,m\}}^{n+m} C_k \binom{N}{k},$$

where C_k is the number of k-step paths from $(0, 0)$ to (n, m) such that every step is of size either $(0, 1)$, $(1, 0)$, or $(1, 1)$ and the path never leaves the set K.

To obtain the counts $C_{\max\{n,m\}}, \ldots, C_{n+m}$, we use a recursion. Define $C_k(i, j)$ to be the number of k-step paths from $(0, 0)$ to (i, j) where each step is of size either $(0, 1)$, $(1, 0)$, or $(1, 1)$ and where the path never leaves the set K. It then follows that $C_k = C_k(n, m)$. The kth step in any path that contributes to $C_k(i, j)$ must have been of size either $(0, 1)$, $(1, 0)$, or $(1, 1)$. Thus, for $k \ge 1$,

$$C_k(i, j) = K(i, j)\left\{C_{k-1}(i-1, j) + C_{k-1}(i, j-1) + C_{k-1}(i-1, j-1)\right\}, \tag{1}$$

where $C_k(i, j)$ is taken to be zero if i, j, or k is negative. This leads to the following algorithm for computing $P(D_{KS} \le c)$ under the null hypothesis.