Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions

Taylor B. Arnold
Yale University

John W. Emerson
Yale University

Abstract

A general theory for extending nonparametric goodness-of-fit tests to discrete null distributions has existed for several decades. Despite this, modern statistical software have generally failed to provide this methodology to users. The packages **ks.test** and **cvm.test** serve to fill this need in the R language for the two most popular nonparametric tests. The structure of these two packages are explained and examples to their specific usage are presented. Particular attention is given to the various numerical issues that arise in their implementation.

Keywords: Cramer von Mises, goodness-of-fit tests, Kolmogorov-Smirnov, R.

1. Introduction

Given an observed sequence of data, the typical question of statistical inference is to determine the underlying process from which it came. Given a particular null distribution, goodness-of-fit tests are used in order to test whether it is likely that the data were generated via the null distribution.

While almost any hypothesis test can be viewed as a variant on a goodness-of-fit test, the term is typically applied to those tests which are nonparametric in nature. That is, they do not conduct the hypothesis testing through the explicit calculation of the parameters generating the data. Instead, they generally attempt to calculate differences between the empirical distribution of the observed data and the overall distribution of the null model. In many cases these tests are preferential since they tend to have an increased power for interesting deviations from the null model, whilst allowing for more robustness due to uninteresting deviations such as measurement error. By far the most popular of these nonparametric tests are due to Kolmogorov and Smirnov, and several variants on the estimated proposed by Cramer von Mises.

While the original aim of nonparametric goodness of fit tests was more continuous null distributions, discrete versions have existed since the early 1970s. Unfortunately, current statistical software has largely failed to incorporate them. This has left the end-user to either use only parametric tests, such as Pearson's Chi-Squared test, or to incorrectly apply the functions meant for continuous distributions. As shown in section 5, the latter can be particularly dangerous in the small sample setting. The packages **ks.test** and **cvm.test** serve to fill this void for users operating in the R environment.

2. Kolmogorov-Smirnov Test

2.1. Overview

Of all the methods for nonparametric goodness-of-fit tests, by far the most popular is the method devised by Kolmogorov and Smirnov. It is the only such test built into the base of R. The idea behind it is fairly simple. Given the cumulative distribution function $F_0(x)$ of the continuous null distribution, and the empirical distribution function $F_{data}(x)$ of the observed data, one constructs the statistic:

$$D = \sup_x |F_0(x) - F_{data}(x)| \tag{1}$$

The difficult part of the theory is showing that under the null model D has a fixed distribution. Perhaps most interestingly is the fact that this distribution does not depend on which null model was specified. Given this, it is easy to read p-values for the test statistic off of a table or computer function.

The extension of this result to non-continuous null distributions does not have such a clean solution. While the calculation of the test statistic D is still just as easy, the distribution of the testing statistic is much more difficult. Unlike in the continuous case, it ultimately depends on which null model was chosen; this makes it impossible to simply read p-values directly off of a fixed table. It was known since at least the 1950's that using the tables for continuous distributions resulted in conservative p-values; it was not until Conover (1972) that a method for computing exact p-values in this case was developed.

Two common alternatives of the above test statistic exit, and are also encoded in R. The absolute value is discarded and the rest is either left alone (the 'greater' testing alternative) or the supremum is replaces with a negative infimum (the 'lesser' hypothesis alternative). These can be helpful depending on the nature of the alternative hypotheses for which the test is desired to be powerful towards.

2.2. Implementation

The implementation of the discrete Kolmogorov-Smirnov function consists of two parts. First the particular test statistic needs to be calculated, and then the p-value for that particular statistic must be computed. These are the same general procedures as for the default function **ks.test** in R. Therefore, the basic implementation here attempts to incorporate the discrete test into the default function. There is a default built-in class for step-functions and therefore it is easy to write code which checks for the input null distribution being a step-function and

to divert these cases to the implementation below. Other null distributions are handled as they were previously.

Given that the test statistic is, theoretically, the same as in the continuous case it would seem that the first part could be directly taken from the existing code. This is unfortunately not the case. Given the quantity x, which is a vector consisting of the difference between the cumulative null distribution and cumulative empirical distribution at each observed data point, the default in R to determine the Kolmogorov-Smirnov test statistic is (with modifications made for readability):

```
STATISTIC <- max(c(abs(x), abs(x-1/n)))
```

This method utilizes the fact that the supremum in equation 1 is the absolute difference between a continuous function and stepwise constant function; this supremum can be reached either at an observed data point, or for a sequence of points approaching an observed data point from the left. This particular fact is still true if both functions are stepwise constant, but the problem comes from the fact that this code assumes, with the code `abs(x-1/n)`, that the cumulative null distribution is continuous at the observed data points. This should never be the case for discrete null distributions. This gives the incorrect default behavior:

```
R>  ks.test(c(1,2), ecdf(c(1,2)))

        One-sample Kolmogorov-Smirnov test

data:  c(1, 2)
D = 0.5, p-value = 0.5
alternative hypothesis: two-sided
```

Where clearly the test statistic should have been 0, since the null and empirical cumulative distributions are the same.

Getting around this issue is easy to do by computing the test statistic as `max(abs(x))` in the discrete case. The important point here is that R's default behavior for computing the test statistic was dependent on the continuity of the null distribution, even though in theory the discreteness should not have affected the calculations.

Once the test statistic is determined, the p-value for this value needs to be computed. For high sample sizes, the test distribution becomes the same as in the continuous case. The default options for **ks.test** already allow the user to specify if they would like approximate asymptotic distributions (since these exist in the continuous setting as well). So when this approximation flag is turned on, this distribution is used for all situations. When the user requires exact p-values, the methodology in Conover (1972) needs to be used. Code for this procedure in the R language, or in any other open source options, did not previously exist and is included in the revised **ks.test**. The calculations are complex but straightforward; the full details are contained in the package source and in the original Conover paper.

The interesting part of the implementation of Conover's method, from a computational standpoint, are the difficult numerical issues that arise when calculated the p-values for larger sample sizes.

3. Cramer von Mises Tests

3.1. Algorithm

While the Kolmogorov-Smirnov test is the most well known of the non-parametric goodness of fit tests, there is another family of tests which has been shown to be more powerful to a large class of alternatives distributions. The original was developed jointly by Harald Cramer and Richard von Mises (Cramer 1928; von Mises 1928), and further adapted in Anderson and Darling (1952), and Watson (1961). The test statistics are, respectively, given as:

$$W^2 = n \cdot \int_{-\infty}^{\infty} \left[F_{data}(x) - F_0(x)\right]^2 dF_0(x)$$

$$A^2 = n \cdot \int_{-\infty}^{\infty} \left[F_0(x) - F_0(x)^2\right]^{-1} \left[F_{data}(x) - F_0(x)\right]^2 dF_0(x)$$

$$U^2 = n \cdot \int_{-\infty}^{\infty} \left[F_{data}(x) - F_0(x) - W^2\right]^2 dF_0(x)$$

Where F is either the cumulative distribution of the null model or the empirical cumulative distribution of the observed data. As in the Kolmogorov-Smirnov test statistic, these all have distribution free null distributions in the continuous case.

It has been shown that these tests can be more powerful than Kolmogorov-Smirnov tests to certain deviations. As they all involve integration over the whole range of data, rather than one supremum, it is not surprising that they are generally best when the true alternative distribution deviates a little over the whole range of data rather than deviating a lot over a small range. For a complete analysis of the relative powers of these tests see Stephens (1974).

Generalizations of the Cramer-von Mises tests were developed in V Choulakian (1994). Much like for the Kolmogorov-Smirnov test, the theoretical form of the test statistics are unchanged; although the discreteness allows for a slightly simpler representation. The null distribution of the test statistics are again distribution dependent, unlike the continuous version. The methods do not suggest finite sample results, but rather show that the asymptotic null distribution is equal to a weighted sum of independent chi-squared variables (the weights depending on the particular distribution). This asymptotic distribution is what we implement here; the original papers shows that this approximation

3.2. Implementation

For the user, the function **cvm.test** was designed to work in exactly the same manner as **ks.test** in R. This alleviates the need to learn a new set of input and output instructions and hopefully will increase the usage of these tests. As there is no default functionality in R for continuous Cramer-von Mises, subsequent versions of the package should also originally implement the continuous case even though this is not what we concentrate on here.

Calculation of the three test statistics is done by straightforward matrix algebra as given in V Choulakian (1994). Determining the form of the asymptotic null distribution is also easy using the built-in eigenvalue decomposition functions. The only difficulty in the process involves actually calculating the percentiles for these weighted chi-squares.

The method used for calculating the distribution of a weighted sum of independent chi-squared variables is given in Imhof (1961). A general method for computing any quadratic form of

normals is presented, which is easily adapted for our case since each chi-squared variable has only one degree of freedom. The exact formula given for the distribution function of Q, the weighted sum of chi squares, is:

$$\mathbb{P}\{Q \geq x\} = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\sin \theta(u, x)}{u\rho(u)} du$$

For continuous functions $\theta(\cdot, x)$ and $\rho(\cdot)$ which depend on the actual weights used.

There is no analytic solution to the integral, but integration can be carried out using numerical techniques. This seems fine in most situations, but numerical issues do become a problem in the regime of large test statistics x. The function $\theta(\cdot, x)$ is linear in x, and thus as the the test statistic grows the corresponding period of the integrand decreases. As the function acquires too many inflection points, the approximation becomes unstable. This is further magnified by this occurring when p-values should be very small; thus tiny fluctuations which would be undetectable elsewhere are quite prominent. Figure ? shows the non-monotonicity of the function as the test statistic grows.

There is fortunately a simple conservative approximation which can get around this numerical problem. Given the following inequality:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{p} \lambda_i \chi_1^2 \geq x\right) &\leq \mathbb{P}\left(\lambda_{max} \sum_{i=1}^{p} \chi_1^2 \geq x\right) \\ &= \mathbb{P}\left(\chi_p^2 \geq \frac{x}{p\,\lambda_{max}}\right) \end{aligned}$$

We see that the values for the weighted sum can be bounded by a simple transformation and a chi-squared distribution of a higher degree of freedom. Calculations for the cumulative distribution of a chi-squared distribution are stable and already implemented in R, which makes this an even easier fix. Therefore, whenever the above approximation yields a small p-values, we suspect numerical issues and use the preceding conservative approximation instead.

4. Discussion

Go into more detail about reasons for using these versus chi-squared type tests.

Issues revolving around incorrect usage of default ks.test() in R because of continuity error (i.e. ks.test(1:2, ecdf(1:2))

Further discuss numerical issues; ks -> n over 30 has problems, CvM -> integrate function can sometimes refuse to run because of failure to calculate integral.

Benefits and drawbacks of using simulation to gather p-values.

Possibly go into a brief discussion of classes and class inheritance in R. The use of 'htest' in our functions; benefits and issues with using it. Particularly troublesome in ks.test where we have a range of p-values.

Two-sample cases; no theory for discrete distributions. Why? Distribution dependence messes things up. Discuss this for sure.

For further generalizations of tests see de Wet and Venter (1994)

5. Examples

References

Anderson TW, Darling DA (1952). "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes." *Ann. Math. Statist.*, **23**, 193–212.

Conover WJ (1972). "A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions." *Journal of the American Statistical Association*, **67**(339), 591–596.

Cramer H (1928). "On the composition of elementary errors: II, Statistical applications." *Skand. Akt.*, **11**, 141–180.

de Wet T, Venter J (1994). "Asymptotic distributions for quadratic forms with applications to tests of fit." *Annals of Statistics*, **2**, 380–387.

Imhof J (1961). "Computing the distribution of quadratic froms in normal variables." *Biometrika*, **48**, 419–426.

Stephens MA (1974). "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association*, **69**(347), 730–737.

V Choulakian R A Lockhart MAS (1994). "Cramer-von Mises statistics for discrete distributions." *The Canadian Journal of Statistics*, **22**(1), 125–137.

von Mises RE (1928). *Wahrscheinlichkeit, Statistik und Wahrheit.* Julius Springer, Vienna, Austria.

Watson GS (1961). "Goodness of fit tests on the circle." *Biometrika*, **48**, 109–114.

Affiliation:

Taylor B. Arnold
24 Hillhouse Ave.
New Haven, CT 06511, USA
E-mail: taylor.arnold@yale.edu
URL: http://euler.stat.yale.edu/~tba3

John W. Emerson
24 Hillhouse Ave.
New Haven, CT 06511, USA
E-mail: john.emerson@yale.edu
URL: http://euler.stat.yale.edu/~jay

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd