Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions

by Taylor B. Arnold and John W. Emerson

Abstract Methodology extending nonparametric goodness-of-fit tests to discrete null distributions has existed for several decades. However, modern statistical software has generally failed to provide this methodology to users. We offer a revision of R's `ks.test()` function and a new `cvm.test()` function that fill this need in the R language for two of the most popular nonparametric goodness-of-fit tests. This paper describes these contributions and provides examples of their usage. Particular attention is given to various numerical issues that arise in their implementation.

Introduction

Goodness-of-fit tests are used to assess whether data are consistent with a hypothesized null distribution. The χ^2 test is the best-known parametric goodness-of-fit test, while the most popular nonparametric tests are the classic test proposed by Kolmogorov and Smirnov followed closely by several variants on Cramér-von Mises tests.

In their most basic forms, these nonparametric goodness-of-fit tests are intended for continuous null distributions, but they have also been adapted for discrete null distributions. Unfortunately, most modern statistical software packages and programming environments have failed to incorporate these discrete versions. As a result, researchers would typically rely upon the χ^2 test or use of a nonparametric test designed for a continuous null distribution. For smaller sample sizes, in particular, both of these choices can produce misleading inferences.

This paper presents a revision of R's `ks.test()` function and a new `cvm.test()` function to fill this void for researchers and practitioners in the R environment (R Development Core Team, 2010). We first present overviews of the theory and general implementation of the discrete Kolmogorov-Smirnov and Cramér-von Mises tests. We discuss the particular implementation of the tests in R and provide examples. We conclude with a short discussion, including the state of existing continuous and two-sample Cramér-von Mises testing in R.

Kolmogorov-Smirnov Test

Overview

The most popular nonparametric goodness-of-fit test is the Kolmogorov-Smirnov test. 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, the test statistic is given by

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

The distribution of D does not depend on the hypothesized distribution, making this a computationally attractive method. Slakter (1965) offers a standard presentation of the test and its performance relative to other algorithms. The test statistic is easily adapted for one-sided tests. For these, the absolute value in (1) is discarded and the tests are based on either the supremum of the remaining difference (the 'greater' testing alternative) or by replacing the supremum with a negative infimum (the 'lesser' hypothesis alternative). Tabulated p-values have been available for these tests since 1933.

The extension of the Kolmogorov-Smirnov test to non-continuous null distributions is not straightforward. The formula of the test statistic D remains unchanged, but its distribution is much more difficult to obtain; unlike the continuous case, its distribution depends on the null model. It has been known since at least the 1950's that using the tables associated with continuous null distributions results in conservative p-values when the null distribution is discontinuous (see *** or *** or Stephens (1974), for example). In the early 1970's, Conover (1972) developed the method implemented here for computing p-values in the case of discrete null distributions.

Implementation

The implementation of the discrete Kolmogorov-Smirnov test involves two steps. First, the particular test statistic is calculated (corresponding to the desired one-sided or two-sided test). Then, the p-value for that particular test statistic may be computed.

The form of the test statistic is the same as in the continuous case; it would seem that no additional work would be required for the implemenetation, but this is not the case. Consider two non-decreasing functions f and g, where the function f is a step function with jumps on the set $\{x_1, \dots x_N\}$ and g is continuous (the classical Kolmogorov-Smirnov situation). In order to determine the supremum of the difference between these two functions, notice that

$$\sup_x |f(x) - g(x)|$$
$$= \max_i \left[\max\left(|g(x_i) - f(x_i)|, \lim_{x \to x_i} |g(x) - f(x_{i-1})| \right) \right] \quad (2)$$
$$= \max_i \left[\max\left(|g(x_i) - f(x_i)|, |g(x_i) - f(x_{i-1})| \right) \right]. \quad (3)$$

Computing the maximum over these $2N$ values (with f equal to $F_{data}(x)$ and g equal to $F_0(x)$ as defined above) is clearly the most efficient way to compute the Kolmogorov-Smirnov test statistic for a continuous null distribution. When the function g is not continuous, however, equality (3) does not hold in general because we cannot replace $\lim_{x \to x_i} g(x)$ with the value $g(x_i)$.

If it is known that g is a step function, it follows that for some small ϵ,

$$\sup_x |f(x) - g(x)| = \max_i (|g(x_i) - f(x_i)|, |g(x_i - \epsilon) - f(x_{i-1})|) \quad (4)$$

where the discontinuities in g are more than some distance ϵ apart. This, however, requires knowledge that g is a step function as well as of the nature of its support (or break-points). As a result, we implement the Kolmogorov-Smirnov test statistic for discrete null distributions by requiring the complete specification of the null distribution.

Having obtained the test statistic, the p-value must then be calculated. For larger sample sizes, the null distribution can be approximated with the null distribution from the classical Kolmogorov-Smirnov test. When an exact p-value is required for smaller sample sizes, the methodology in Conover (1972) is used. Full details of the calculations are contained in source code of our revised function `ks.test()` and in Conover's original paper.

Cramér-von Mises Tests

Overview

While the Kolmogorov-Smirnov test may be the most popular of the nonparametric goodness-of-fit tests, Cramér-von Mises tests have been shown to be more powerful against a large class of alternatives hypotheses. The original test was developed by Harald Cramér and Richard von Mises (Cramer, 1928; von Mises, 1928) and further adapted by Anderson and Darling (1952), and Watson (1961). The original test statistic, W^2, Anderson's A^2, and Watson's U^2 are:

$$W^2 = n \cdot \int_{-\infty}^{\infty} [F_{data}(x) - F_0(x)]^2 \, dF_0(x) \quad (5)$$

$$A^2 = n \cdot \int_{-\infty}^{\infty} \frac{[F_{data}(x) - F_0(x)]^2}{F_0(x) - F_0(x)^2} dF_0(x) \quad (6)$$

$$U^2 = n \cdot \int_{-\infty}^{\infty} \left[F_{data}(x) - F_0(x) - W^2\right]^2 dF_0(x) \quad (7)$$

As with the original Kolmogorov-Smirnov test statistic, these all have null distributions of their test statistics which are independent of the specified null models.

The W^2 statistic was the original test statistic. The A^2 statistic was developed by Anderson in the process of generalizing the test for the two-sample case. Watson's U^2 statistic was developed for distributions which are cyclic (with an ordering to the support but no natural starting point); it is invariant to cyclic reordering of the support. For example, a distribution on the months of the year could be considered cyclic.

It has been shown that these tests can be more powerful than Kolmogorov-Smirnov tests to certain deviations from the hypothesized distribution. They all involve integration over the whole range of data, rather than use of a supremum, so they are best-suited for situations where the true alternative distribution deviates a little over the whole support rather than having large deviations over a small section of the support. Stephens (1974) offers a complete analysis of the relative powers of these tests.

Generalizations of the Cramér-von Mises tests to discrete distributions were developed in Choulakian et al. (1994). As with the Kolmogorov-Smirnov test, the forms of the test statistics are unchanged, and the null distributions of the test statistics are again hypothesis-dependent. Choulakian et al. (1994) does not offer finite-sample results, but rather shows that the asymptotic distributions of the test statistics under the null hypothesis each involve consideration of a weighted sum of independent chi-squared variables (with the weights depending on the particular null distribution). We implement the asymptotic distributions here; Stephens (1974) showed that these approximations are conservative and asymptotically equivalent to the true null distributions.

**** Why would Stephens (1974) have this, when they were developed in 1994? ****

Implementation

Calculation of the three test statistics is done using the matrix algebra given by Choulakian et al. (1994). The only notable difficulty in the implementation of the discrete form of the tests involves calculating the percentiles of the weighted sum of chi-squares,

$$Q = \sum_{i=1}^{p} \lambda_i \chi^2_{i,1df}, \quad (8)$$

where p is the number of elements in the support of the hypothesized distribution. Imhof (1961) provides a method for obtaining the distribution of Q, easily adapted for our case because the chi-squared variables have only one degree of freedom. The exact formula given for the distribution function of Q is given by

$$\mathbb{P}\{Q \geq x\} = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin\theta(u,x)}{u\rho(u)}du \tag{9}$$

for continuous functions $\theta(\cdot,x)$ and $\rho(\cdot)$ depending on the weights λ_i.

There is no analytic solution to the integral in (9), so the integration is accomplished numerically. This seems fine in most situations we considered, but numerical issues appear in the regime of large test statistics x (or, equivalently, small p-values). The function $\theta(\cdot,x)$ is linear in x; as the test statistic grows the corresponding periodicity of the integrand decreases and the approximation becomes unstable. As an example of this numerical instability, the red plotted in Figure 1 shows the non-monotonicity of the numerical evaluation of equation (9) for a null distribution that is uniform on the set $\{1,2,3\}$.

We resolve this problem by using a simple conservative approximation to avoid the numerical instability. Consider the following inequality:

$$\mathbb{P}\left(\sum_{i=1}^{p}\lambda_i\chi_1^2 \geq x\right) \leq \mathbb{P}\left(\lambda_{max}\sum_{i=1}^{p}\chi_1^2 \geq x\right) \tag{10}$$

$$= \mathbb{P}\left(\chi_p^2 \geq \frac{x}{p\lambda_{max}}\right) \tag{11}$$

The values for the weighted sum can be bounded using a simple transformation and a chi-squared distribution of a higher degree of freedom. The original formulation is preferable for most p-values, while this upper bound is used for smaller p-values (smaller than 0.001, based on our observations of the numerical instability of the original formulation). Although it would be preferable to determine the use of the bound based on values of the test statistic rather than the p-value, the range of "extreme" values of the test statistic varies with the hypothesized distribution.

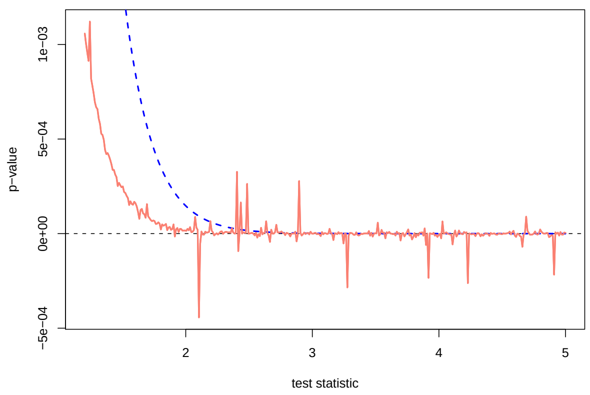

Figure 1: Plot of calculated p-values for given test statistics using numerical integration (red) compared to the conservative chi-squared bound (dashed blue). The null distribution is uniform on the set $\{1,2,3\}$ in this example. The instability in the calculated p-values is a result of numerical instabilities.

Kolmogorov-Smirnov and Cramér-von Mises Tests in R

Functions `ks.test()` and `cvm.test()` are provided for convenience in packages **ks.test** and **cvm.test**, respectively. Function `ks.test()` offers a revision of R's Kolmogorov-Smirnov function `ks.test()` from base package **stats**, while `cvm.test()` is a new function for Cramér-von Mises tests. Both are available from the authors or from R-Forge (`https://r-forge.r-project.org/R/?group_id=802`); `ks.test()` will be proposed for inclusion in **stats** in late 2010.

The revised `ks.test()` function supports one-sample tests for discrete null distributions by allowing the second argument, `y`, to be an empirical cumulative distribution function (an R function with class `ecdf`) or an object of class `stepfun` specifying a discrete distribution. As in the original version of `ks.test()`, the presence of ties in the data (the first argument, `x`) generates a warning unless `y` describes a discrete distribution. When a discrete distribution is specified, exact p-values are not available for two-sided alternative hypotheses, but the reported p-values will be conservative; very tight bounds for interesting (small) p-values are presented in Conover (1972). For one-sided tests, exact p-values are calculated using Conover's method (when `exact = NULL` or `exact = TRUE`) if the sample size is less than or equal to 30; otherwise, asymptotic distributions are used which are reliable in such cases (CITATION? Is this correct?). When `exact = FALSE` the asymptotic distribution is used which is known to be conservative though imprecise for small samples (CITATION).

The function **cvm.test()** is similar in design to

ks.test(). Its first two arguments specify the data and null distribution; the only extra option, `type`, specifies the variant of the Cramér-von Mises test:

- `x`: a numerical vector of data values.
- `y`: an `ecdf` or step-function (`stepfun`) for specifying the null model
- `type`: the variant of the Cramér-von Mises test; `W2` is the default and most common method, `U2` is for cyclical data, and `A2` is the Anderson-Darling alternative.

As with `ks.test()`, `cvm.test()` returns an object of class `htest`. Although we designed `cmv.test()` primarily for cases with discrete distributions `y`, we implemented the continuous forms of the test for variants `W2` and `A2` (as noted in the Discussion, below).

Examples

Consider a toy example with observed data of length 2 (specifically, the values 0 and 1) and a hypothized null distribution that places equal probability on the values 0 and 1. With the current `ks.test()` function in R (which, admittedly, doesn't claim to handle discrete distributions), the reported p-value, 0.5, is clearly incorrect:

```
> stats::ks.test(c(0, 1), ecdf(c(0, 1)))

        One-sample Kolmogorov-Smirnov test

data:  c(0, 1)
D = 0.5, p-value = 0.5
alternative hypothesis: two-sided
```

Instead, the value of D given in equation (1) should be 0 and the associated p-value should be 1. Our revision of `ks.test()` fixes this problem when the user provides a discrete distribution:

```
> library(ks.test)
> ks.test(c(0, 1), ecdf(c(0, 1)))

        One-sample Kolmogorov-Smirnov test

data:  c(0, 1)
D = 0, p-value = 1
alternative hypothesis: two-sided
```

Next, we simulate a sample of size 25 from the discrete uniform distribution on the integers $\{1,2,\ldots,10\}$ and show several variants of the new `ks.test()` implementation. The first is the default two-sided test, where the reported p-value is a conservative upper bound for the actual p-value. In this case, the approximation may not be that tight, but this is irrelevant for such large p-values (for more interesting (small) p-values, the upper bound is very close to the true p-value as shown in Conover (1972)).

```
> set.seed(1)
> x <- sample(1:10, 25, replace = TRUE)
> x

 [1]  3  4  6 10  3  9 10  7  7  1  3  2  7
[14]  4  8  5  8 10  4  8 10  3  7  2  3

> ks.test(x, ecdf(1:10))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.08, p-value = 1
alternative hypothesis: two-sided
```

Next, we conduct the default one-sided test, where Conover's method provides the exact p-value (up to the numerical precision of the implementation):

```
> ks.test(x, ecdf(1:10), alternative = "g")

        One-sample Kolmogorov-Smirnov test

data:  x
D^+ = 0.04, p-value = 0.7731
alternative hypothesis:
the CDF of x lies above the null hypothesis
```

In contrast, the option `exact=FALSE` results in the p-value obtained by applying the classical Kolmogorov-Smirnov test, resulting in a conservative p-value:

```
> ks.test(x, ecdf(1:10), alternative = "g",
+         exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D^+ = 0.04, p-value = 0.9231
alternative hypothesis:
the CDF of x lies above the null hypothesis
```

A different toy example shows the dangers of using R's existing `ks.test()` function:

```
> ks.test(rep(1, 3), ecdf(1:3))

        One-sample Kolmogorov-Smirnov test

data:  rep(1, 3)
D = 0.6667, p-value = 0.04938
alternative hypothesis: two-sided
```

If, instead, either `exact=FALSE` is used with the new `ks.test()` function, or if the original `stats::ks.test()` is used, the reported p-value is 0.1389.

Finally, we demonstrate the Cramér-von Mises tests with the same simulated data.

```
> library(cvm.test)
> cvm.test(x, ecdf(1:10))

        Cramer-von Mises - W2

data:  x
W2 = 0.057, p-value = 0.8114
alternative hypothesis: Two.sided
```

 ISSN 2073-4859

```
> cvm.test(x, ecdf(1:10), type = "A2")

        Cramer-von Mises - A2

data:  x
A2 = 0.3969, p-value = 0.75
alternative hypothesis: Two.sided
```

*** Cyclical example? ***

```
> library(cvm.test)
> x <- sample(1:4, 20, replace=T)
> cvm.test(x, ecdf(1:4), type='U2')

        Cramer-von Mises - U2

data:  x
U2 = 0.075, p-value = 0.4027
alternative hypothesis: Two.sided

> y <- x%%4 + 1
> cvm.test(y, ecdf(1:4), type = 'U2')

        Cramer-von Mises - U2

data:  y
U2 = 0.075, p-value = 0.4027
alternative hypothesis: Two.sided
```

Discussion

This paper presents the implementation of a range of nonparametric goodness-of-fit tests for discrete null distributions. In some cases, the p-values are known to be exact. In others, conservativeness and tightness of bounds have been established in the literature. As a result, no simulations are necessary for these methods (which were developed during an era when extensive simulations may have been prohibitively expensive or time-consuming). However, this does raise the possibility that alternative tests relying upon modern computational abilities could provide even greater power in certain situations, a possible avenue for future work.

In the continuous setting, both of the Kolmogorov-Smirnov and the Cramér-von Mises tests have two-sample analogues. When data are observed from two processes or sampled from two populations, the hypothesis tested is whether they came from the same (unspecified) distribution. An analogous theory for discrete distributions does not exist. This comes from the fact that the discrete null distributions of the test statistics depend on the exact null distribution; therefore the two-sample case would surely have to depend on the exact distributions used as well which are generally not even stated in the two-sample case. *** I don't think I buy this. Couldn't this just be done as a chi-squared test on a 2xp table, at least? *** Also, see http://ideas.repec.org/p/mtl/montde/2001-23.html.

While we have implemented the two most popular variants of goodness-of-fit tests for discrete distributions, further generalizations of tests are described in the extended study of de Wet and Venter (1994). We also provide general one-sample Cramér-von Mises testing for continuous distributions simply to help complement the abilities provided by several other R packages, although these tests are not the focus of this paper. In particular, functions implemented in package **nortest** focus on the composite hypothesis of normality, while package **ADGofTest** provides the Anderson-Darling variant of the test for continuous distributions. Packages **CvM2SL1Test** and **CvM2SL1Test** provide support for two-sample Cramér-von Mises tests with continuous distributions. Package **cramer** offers a multivariate Cramér test for the two-sample problem. Possible reference, though it isn't clear if this particular one is really part of this literature segment? http://en.scientificcommons.org/39157548

*** Need to be added to the bibliography, whatever we choose to do ***

Bibliography

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23:193–212, 1952.

V. Choulakian, R. A. Lockhart, and M. A. Stephens. Cramér-von mises statistics for discrete distributions. *The Canadian Journal of Statistics*, 22(1):125–137, 1994.

W. J. Conover. A kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67(339):591–596, 1972.

H. Cramer. On the composition of elementary errors: Ii, statistical applications. *Skand. Akt.*, 11:141–180, 1928.

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

J. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426, 1961.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org/. ISBN 3-900051-07-0.

M. J. Slakter. A comparison of the pearson chi-square and kolmogorov goodness-of-fit tests with respect to validity. *Journal of the American Statistical Association*, 60(311):854–858, 1965.

M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

R. E. von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer, Vienna, Austria, 1928.

G. S. Watson. Goodness of fit tests on the circle. *Biometrika*, 48:109–114, 1961.

Taylor B. Arnold
Yale University
24 Hillhouse Ave.
New Haven, CT 06511 USA
taylor.arnold@yale.edu

John W. Emerson
Yale University
24 Hillhouse Ave.
New Haven, CT 06511 USA
john.emerson@yale.edu