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. Despite this, 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 serve to 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 an estimation procedure proposed by Cramér and von Mises (tests sometimes called Cramér-von Mises-Smirnov tests, or simply Cramér-von Mises tests).

*** why "estimation procedure" above? *** Citations in the paragraph above? Also check the body of paper. ***

In their most basic forms, these nonparametric goodness-of-fit tests are intended for continuous null distributions, but they have also been adapted for use with discrete null distributions. Unfortunately, most modern statistical software have failed to incorporate these discrete versions, leaving researchers with the χ^2 test or in the unfortunate position of applying a nonparametric test designed for a continuous null distribution and hoping for the best. As we will see, this latter course of action can be particularly dangerous with small sample sizes. 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.

*** Later, make sure we point out how using one of the "old" tests and hoping for the best is a problem. ***

Kolmogorov-Smirnov Test

Overview

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

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

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 (1) is easily adapted for one-sided tests. For these, the absolute value 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 (*** year ****).

**** check this above, both do away with the absolute value, I think? ***

The extension of the Kolmogorov-Smirnov test to non-continuous null distributions does not have such a clean solution. The formula of the test statistic D remains unchanged, but the distribution of the test statistic is much more difficult to obtain. Unlike the continuous case, the null 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 (**** citation?) when the null distribution is discontinuous (*** or should this be discrete? ***). In the early 1970's, Conover (1972) developed the method implemented here for computing exact p-values in the case of discrete null distributions.

Implementation

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

The test statistic is, in terms of the theory, the same as in the continuous case; it would seem that no additional work would be required, but unfortunately 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 $\{x_1, \ldots x_N\}$ and g is continuous (the classical Kolmogorov-Smirnov situation). In order to determine the supremum of the difference between these two functions, notice that

$$\begin{aligned}\sup_x |f(x) - g(x)| \\ &= \max_i \left(|g(x_i) - f(x_i)|, \lim_{x \to x_i} |g(x) - f(x_{i-1})| \right) \qquad (2) \\ &= \max_i \left(|g(x_i) - f(x_i)|, |g(x_i) - f(x_{i-1})| \right). \qquad (3)\end{aligned}$$

Computing the maximum over these $2N$ (*** why 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 $\lim_{x \to x_e} g(x)$ with the value $g(x_i)$.

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

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

where the discontinuities in g are at least a 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). We implement the Kolmogorov-Smirnov test statistic for discrete null distributions by forcing the user to completely specify the null distribution.

Having obtained the test statistic, the p-value must then be calculated. For larger sample sizes, the null distribution coincides (***should this be stated precisely as an asymptotic result???) with the standard null distribution (CITATION???). 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.

** code hasn't previously existed? I deleted this stuff, not sure if it fits here. Is it obvious (since we bothered to do this in the first place)? Or move it?

Cramér-von Mises Tests

Overview

While the Kolmogorov-Smirnov test may be the most popular of the nonparametric goodness-of-fit tests, there is another family of tests which has been shown to be more powerful against a large class of alternatives hypotheses. The original test was developed jointly (*** or was it simultaneously? ***) 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:

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

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

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

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 relative powers of these tests to different alternatives are studied in depth in Stephens (1974). In general, the W^2 statistic is recommended unless there is a compelling reason to consider one of the others. The A^2 statistic was developed by Anderson in the process of generalizing the test for the two-sample case, but was shown to perform very similarly to the original statistic in the one-sample setting (*** so why, then would we ever use it??? ***). Watson's U^2 statistic was developed for distributions which are cyclic (with an ordering to the support but no natural starting point *** how best to describe this? ***). For example, a distribution on the months of the year could be considered cyclic. Watson's statistic is invariant to cyclical reordering (giving preference to one ordering would be unnatural). **** is this last point necessary? ***

*** Is a second paragraph on Stephen's results really necessary? Can we pick and choose and get away with one paragraph? ***

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).

*** I edited, below, make sure you check it. ****

Generalizations of the Cramér-von Mises tests to discrete distributions were developed in V Choulakian (1994). As with the Kolmogorov-Smirnov test, the forms of the test statistics are unchanged, although the discreteness allows for slightly simpler representations. The null distribution of the test statistics are again hypothesis-dependent. The methods do not offer finite sample results, but rather show that the asymptotic null distribution is equal to a weighted sum of independent chi-squared variables (with the weights depending on the particular null distribution). (*** is this true for all 3? ***) We implement this asymptotic distribution here; the original papers (*** really, all of them? ***) show that this approximation is conservative and asymptotically equivalent to the true null distribution.

*** What does it mean for a distribution to be equal to a weighted sum of independent chi-squared variables? I think this needs to be stated more precisely, above. A formal definition of Q is needed someplace, too, as I'll probably note below. ***

Implementation

Calculation of the three test statistics is done using the matrix algebra given by V Choulakian (1994). The form of the asymptotic null distribution is obtained using R's eigenvalue decomposition functions.

The only notable difficulty in the process involves calculating the percentiles of the weighted sum of chi-squares,

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

Imhof (1961) provides a method for obtaining the distribution of a weighted sum of independent chi-squared variables. It is 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 actual weights used.

There is no analytic solution to the integral, 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. We resolve this problem by using a simple conservative approximation to avoid the numerical instability. Consider the following inequality:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{p} \lambda_i \chi^2_1 \geq x\right) &\leq \mathbb{P}\left(\lambda_{max} \sum_{i=1}^{p} \chi^2_1 \geq x\right) && (10) \\ &= \mathbb{P}\left(\chi^2_p \geq \frac{x}{p\,\lambda_{max}}\right) && (11) \end{aligned}$$

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 approximation is useful for smaller p-values (smaller than 0.001, based on our observations of the numerical instablity of the original formulaton); it should not greatly affect the interpretation of the results.

*** How is the cutoff established in the code? ***

Figure 1 shows the non-monotonicity of the function (*** which function, reference it? ***) as the test statistic grows, compared to the conservative approximation.

*** stuff deleted, redundant ***.

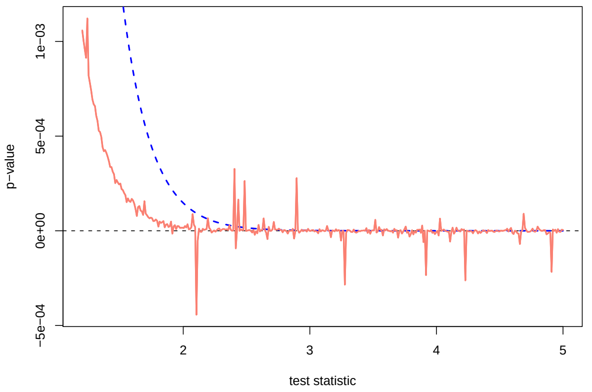

Figure 1: Plot of p-value for given test statistics using numerical integration (pink) versus a conservative chi-squared bound (dashed blue).

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`); they 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. 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 imprecise but conservative, even for small samples (CITATION).

*** Discussion of what we might have broken: cases where the user provided a discrete distribution to the original ks.test() even though it wasn't intended to be supported. ***

The function **cvm.test()** is similar 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`.

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:

```
> 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 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 p-values, the upper bound is very close to the true p-value).

```
> library(ks.test)
> 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 employ two of the Cramér-von Mises tests. Need short discussion, and we could use a good cyclical example:

```
> 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

> 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
```

TAYLOR: good cyclical example, relating to new material about to go into earlier section.

Discussion

*** I commented out a bunch of stuff. I think the discussion can be short and sweet. ***

*** Hmm... if you claim the following, you want to make sure it is true. Did you read it somewhere? ***

In the end, for methods such as those presented in this paper where it is possible, there are enough positives to suggest that we attempt to implement exact p-value calculations. This does lead to an interesting direction for future research in non-parametric goodness of fit tests: the methods presented here were historically chosen because they could be shown to have fixed null distributions which could easily calculated in an era without fast computers for carrying out permutation tests. It is quite possible that variants without the property, which could be easily used today, have greater power to certain alternative distributions.

In the continuous setting, both of the Kolmogorov-Smirnov and the Cramér-von Mises tests have two-sample analogues. Here data are observed from two processes, and the hypothesis tested is whether they came from the same (but unspecified) distribution. There does not exist an analogous theory for discrete distributions. 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 (*** is this true? ***), which are generally not even stated in the two-sample case.

While we have implemented the two most popular variants of goodness-of-fit tests, there are several more exotic varieties to be found. For further generalizations of tests see the extended study done in de Wet and Venter (1994).

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.

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.

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.

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

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