Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions

by Taylor B. Arnold and John W. Emerson

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.
*** try to clean up, avoid package/function issue, make it a general contribution if possible.

Introduction

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, in exchange for failing to detect less interesting deviations due to factors such as measurement error. By far the most popular of these nonparametric tests are due to Kolmogorov and Smirnov, followed closely by several variants on an estimation procedure proposed by Cramér-von Mises.

While the original aim of nonparametric goodness of fit tests was meant for 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.

** Ditto on wordsmithing the contribution (package/function, etc...).

Kolmogorov-Smirnov Test

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 distribution of D under the null model does not depend on which null distribution is being used, making this a computationally attractive method. For a standard treatment of the test and its performance relative to other algorithms see Slakter (1965). Two common alternatives of the above test statistic exist. 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.

The extension of this result to non-continuous null distributions does not have such a clean solution. The formula of the test statistic D remains unchanged, however 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.

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.

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 procedures. This is, unfortunately, often 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. If we want to determine the supremum of the difference between these two functions notice that:

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

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 when given a continuous null distribution. When the function g is not continuous, notice that this formula no longer works since in general we cannot replace the limit of g with its value at x_i. If it is known that g is also a step function, we could replace the formula for some small ϵ by:

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

Where the discontinuities in g are at least a distance ϵ apart. But this requires knowing that g is a step function. In the case of not knowing, or having a g which is in neither a step function nor a continuous function, the only method for computing the supremum is to take the numerical limit in (1). This clearly takes more computational time than simply cycling over $2N$ values.

Therefore, in order to implement the discrete Kolmogorov-Smirnov test statistic, the user must either indicate the points of discontinuity of the null distribution's cdf or live with a much smaller.

*** Check Cramér-von Mises for hyphen, and let's get the accent done properly.

Cramér-von Mises Tests

Overview

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 Cramér and Richard von Mises (Cramer, 1928; von Mises, 1928), and further adapted by Anderson and Darling (1952), and Watson (1961). The test statistics are, respectively, given as:

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

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

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

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.

*** New paragraph, perhaps here, on the differences between these three tests.

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 Cramér-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 (*** NOT FINISHED?)

Implementation

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 1 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.

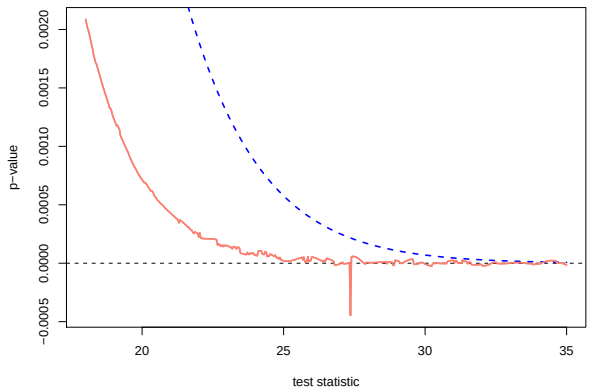

Figure 1: Plot of p-value for given test statistics using numerical integration (pink) versus a conservative chi-squared bound (blue).

Kolmogorov-Smirnov and Cramér-von Mises Tests in R

Package **ks.test** contains function `ks.test()`, a revision of R's Kolmogorov-Smirnov function `ks.test()` from base package **stats**. Package **cvm.test** contains function `cvm.test()`, a new function for Cramér-von Mises tests. Both are available from the authors as of October, 2010, 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 sample (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

Hmm... I thought the package had a trivial example where the standard `ks.test()` was a problem?

```
> library(ks.test)
> x <- sample(1:10, 25, replace = TRUE)
> x

 [1]  3  4  6 10  3  9 10  7  7  1  3  2  7  4  8  5  8 10
[19]  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

> 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

> 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
```

Discuss this, nice example. Relate to new material that is about to go into the earlier section.

```
> 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

> 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
```

JAY: simulation showing original ks.test() was conservative, Conover exact.

```
> 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

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)). Will be earlier, needed here?

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

The computational capabilities of modern computers provide an alternative to using a complex formula to calculate the p-value for a test statistic. By drawing random samples from the null distribution, in many cases the p-values for a given statistic can be calculated accurately in a relatively short time span. This can be quite useful in some cases, but using the exact formulas from our methodology has several distinct advantages. First of all, if the function is called many times (say, within another simulation study), the computational benefits of using a formula can quickly become substantial. There is also a greater need for supervision in a simulation study, with the exact number of runs needed to reach convergence (or other similar threshold). Additionally, while simulations are generally a well received accepted method, it is often important in applied data analysis to be sure that the calculated *p*-values are not the result of a numerical oddity in a particular run of a simulation.

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.

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. More precisely, explain that we could provide upper and lower bounds for the p-value but are limited by class `htest`.

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)

Future work: continuous CVM.

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 froms 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. Cramer-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
24 Hillhouse Ave.
New Haven, CT 06511 USA
taylor.arnold@yale.edu

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