September 1, 2011

Dear Martyn,

We appreciate the opportunity to revise and resubmit this article. The input from you, the reviewers, and a reviewer of a separate paper making use of the testing routines have been helpful. Taylor also enjoyed talking with you at UseR!

Below, we reply to each item of yours and the reviewers, marking our replies in bold. I apologize for the loose formatting in places due to copy-and-paste from the letter PDFs.

Regards,

Jay and Taylor

From the Associate Editor:

I must apologise for the long delay since you submitted this article. I am attaching two reviewers reports. Both reviewers are positive and have some constructive suggestions. Based on this I would like you to submit a revised manuscript after considering some changes to the functions. Your resubmission should include a point-by-point reply to the two reviews.

We appreciate the feedback and have done this.

Reviewer 2 has a suggestion for tighter bounds in the cvm.test function than the one shown in equations 10 + 11, which I suggest you implement.

We have added this bound (which may or may not be tighter depending on the situation), and use the lesser of the two. We will return to this point elsewhere in our response.

Reviewer 1 wants to see some simulations showing the accuracy of this approximation and I think there is a good case for doing this.

There may be some confusion in the wording, and we don't think a simulation was requested for the purpose of the paper (we think the reviewer was requesting the estimation of p-values via simulation). We did add this option to both ks.test() and cvm.test(), following the convention of fisher.test(), for example, and added a simple example to the paper.

Reviewer 1 goes further and suggests that you use Monte Carlo simulation to calculate p-values in the range (0.0001, 0.01) in the cvm.test function itself, pointing out that this range is important in a multiple comparison situation. I'm not sure I agree with this latter suggestion, and I would hope that it can be circumvented with the use of the improved bounds.

We agree, and have provided both the improved bounds as well as options for estimating p-values via simulation. We also agree with your perspective that switching between these methods automatically is not desirable, and prefer to let the user make an explicit choice.

I am aware that you have updated the package since your submission. I also saw your message to the R-devel list about incorporating these changes into the stats package. I don't see any major problems with this idea, but I would like to keep this issue separate from the destiny of the article in The R Journal.

Yes, we agree and have modified the wording in an appropriate manner.

I have a few minor comments on the manuscript:
- page 1. Introduction, third paragraph. Articles in The R Journal do not cite R itself.

Done.

- page 4. Examples. Could you please use scoping consistently to make it clear function is being called: your one or the one in stats.

Thank you, we did this.

- page 4, LHS. In the second example, the p-value returned by ks.test is a conservative upper bound. I think this could be mentioned earlier (page 2).

We have reworked the earlier section and make reference to the possible behavior.

- page 4, RHS. "A different toy example shows the dangers of using R's existing ks.test() function [with discrete data]"

Done, thank you.

best regards
Martyn

From Reviewer 1:

Reviewer’s Report on the Paper
“Non-parametric Goodness-of-Fit Tests for
Discrete Null Distributions”
March 02, 2011

In this paper the authors present four non-parametric goodness-of-fit tests for discrete distributions: Kolmogorov-Smirnov test, and three Cram´r-von Mises type tests.

Regarding the Kolmogorov-Smirnov test

Line 18 from bottom on page 2: the author said, “ For large sample sizes, the null distribution can be approximated with the null distribution from the classical Kolmogorov-Smirnov test”.

This statement is very ambiguous. What is the “null distribution from the classical Kolmogorov-Smirnov test”? Do the authors mean the corresponding test for continuous data? However, a discrete distribution function is near to a continuous function only if it has too many discontinuous points. So, if there are only a few discontinuous points for the null distribution function, the classical K-S test (for continuous data, as the reviewer understands) can not be applied, no matter how large the sample is. The authors should clarify this point.

We changed this paragraph to better make our intended point, and thank the reviewer for catching the error. We agree that the use of the word “null” is particularly challenging here. And there was no intended convergence result that may have been implied by our previous use of the word “approximated”.

Line 15 from bottom on page 2: When the sample size is small and an exact p-value is required, the authors’ contribution is the implementation of the idea of Conover (1972) and the revision of the corresponding function ks.test(). The revised function ks.test() has obvious advantages over the existing version when the null distribution is discrete. However, it seems that the authors should mention that the number of discontinuous points should be small.

There may be several issues to address here. First, we really are talking about discrete distributions and not discontinuous distributions. So we will try to clarify our intended vocabulary, first. If there are P elements in the support of the discrete distribution, and we have a sample size N, having N small relative to P is in essence much like having an (admittedly small) sample from a continuous distribution. There is another issue at play, but we think it is beyond the scope of this paper: the degree to which the discrete distribution is close to uniform (or very much non-uniform). Use of the classical KS test really wouldn't be terrible when N is small relative to P and the distribution is close to uniform; and nor would Conover/Gleser's methods (though both would have low power we expect). If N is large relative to P (or if the distribution is very much non-uniform with smaller N relative to P), then yes, this is where Conover/Gleser's methodology is more likely to be helpful. We are hesitant to add too much detail on this point in the paper.

The basic point is that this methodology is useful when the null distribution is not “close to” continuous, and this “close to” may relate to a variety of things, including the size of the support, the sample size, and the degree of non-uniformity, in particular. We have tried to avoid excessive discussion of the finer points of the methodology, which we don't think is appropriate for this particular paper; many of the references provided are very helpful on these issues.

Regarding the three Cram´r-von Mises Type tests

Asymptotically, the three tests have the general equivalent form of weighted sum of chi-squares:

p

$\lambda i\ \chi 2 = 1$,

i,df

$Q=$

$i=1$

and the p-value for a given value x of the test statistic is given by

$\Pr \{Q \geq x\} =$

$1\ 1$

$+$

$2\ \pi$

∞

0

$\sin \theta(u, x)$

du

$u\rho(u)$

“for continuous function $\theta(\cdot, x)$ and $\rho(\cdot)$ depending on the weights λi ”. This result is due to Imhof (1961). The computation of the p-value is achieved numerically. This is fine in most situations. But when x is extremely large or the p-value is extremely small, the numerical process is instable, resulting weird results. To avoid the numerical instability, the authors’ solution is to compute the upper bound of the p-value:

p

$\Pr$

$\chi 2$

p

p

$\chi 2$

1

$\geq x/p\lambda max \equiv \Pr\ \lambda max$

$\lambda i\ \chi 2 = 1 \geq x \equiv \Pr \{Q \geq x\}$.

i,df

$\geq x \geq \Pr$

$i=1$

$i=1$

According to the authors, this upper bound is a “conservative approximation”.
The conservative approximation is useful when the p-value is extremely small, e.g. < 10^{-5} . However, for p-values in the range (0.0001, 0.01), the authors should evaluate the performance of the conservative approximation by using the Monte-Carlo simulation. According to Figure 1 on page 3, it seems that the approximation may be poor in that range. In fact,

in the computer age, one can easily use the Monte Carlo simulation to get a more accurate estimate for the p-value for a given value of x, (note the fact that all λi ’s are known!), unless the p-value is really tiny. The reviewer’s experience is, when the p-value is larger than 10−4 , the Monte-Carlo simulation works very well.

Reviewer 2 provided an alternative upper bound which we incorporate. We also provide for estimation of the p-value via Monte Carlo simulation, and appreciate this suggestion. However, we decided not to automatically simulate the p-value, and instead provide functionality more consistent with that of fisher.test(), for example -- requiring the user to specify simulate.p.value=TRUE. We also note that Figure 1 is for one specific example, and would change significantly in other examples.

In multiple testing, where a good portion of p-values are usually in the range (0.00001, 0.001), it is very important to ensure the accuracy of the estimated p-values since it significantly impacts the set of null hypotheses to be rejected. The authors should consider how to improve the results for the p-values in that range. Perhaps, a more reasonable way to compute the p-value for a given value x of the test statistic is,
• to use the Imhof’s method if the p-value is moderately large, (e.g., larger than 0.01);
• to apply the Monte-Carlo simulation if the p-value is in the range (0.0001, 0.01); (This will give better results, unless the authors can give convincing evidence that this is not necessary.)
• to use the conservative approximation if the p-value is really tiny. (Monte Carlo simulation does not work and, the Imhof’s method can not be applied due to numerical instability.)
Suggestions. The authors should revise the paper according to the above comments.

See above, and we appreciate the reviewer's comments on this issue.

From Reviewer 2:

Referee’s report to Authors on
R Journal Manuscript
Nonparametric goodness-of-fit tests for discrete null distributions
by
Taylor B. Arnold & John W. Emerson
Summary:
This article describes implementation of empirical distribution function tests for fully specified discrete distributions. I found the article clear, well written and pretty comprehensive. I have three comments or suggestions:
1. I think you ought to remind readers that the hypotheses being tested must not have any uknown parameters; it is not ok to estimate the parameters and then use the fitted distribution as the null hypothesis. The relevant large sample theory for Cram´r-
e
von Mises statistics is in Lockhart, Spinelli, and Stephens, CJS, 2007. Those authors should have written R code to implement their tests but they never did.

Thank you for the suggestion; we added a note in the discussion addressing this point, along with this reference.

2. In the 2007 paper just cited the authors suggest alternative test statistics because they noticed that in non-uniform cases if they reversed the order of the cells they got a different statistic value. The modifications are not great. Suppose you were testing the hypothesis that a sample of X values come from the binomial distribution with k trials and success probability 3/4. The values k − X would be a sample from the binomial distribution with success probability 1/4 and you would probably want the test statistics to be the same in both cases. (Switching the meaning of success and failure should not change the conclusion about whether or not the binomial model is appropriate.)

An interesting issue, we agree. One of the cvm.test() options is invariant to circular re-orderings, but not to your example of course. We choose not to delve further into this matter in this paper; existing papers are more appropriate references.

3. The upper bound on the tail of a linear combination of chi-squares can likely be improved by Markov’s inequality:
λi Zi2 ≥ x) ≤ E(exp(t
P(
i
λi Zi2)) exp(−tx) =
e−tx
(1 − 2tλi)
.
Here the Zi are standard normal and I hope I have the formula for the MGF of a
chi-square righHere the Zi are standard normal and I hope I have the formula for the

mgf of a chi-square right. The final quantity on the right should be minimized over $0 < t < (2\lambda\mathrm{max})-1$ to get a pretty tight bound as in the fashion of extreme value theory.
Overall, well done.

We appreciate this suggestion, and your memory was correct. We now calculate this bound in addition to the one we originally proposed, and choose the lesser of the two in cases where we suspect numerical instabilities. We found examples where each was useful. We also add an option to allow estimation of the p-value via Monte Carlo simulation.