Journal of Statistical Software

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

Fuzzy P-values and Ties in Nonparametric Tests

Charles J. Geyer
University of Minnesota

Abstract

This article introduces the use of fuzzy P-values and confidence intervals (Geyer and Meeden 2005) with classical nonparametric tests, specifically with the sign test and the one-sample and two-sample Wilcoxon tests. The method presented can be applied to any test and provides simple exact fuzzy tests at any level in the presence of ties. It also provides the corresponding fuzzy confidence intervals.

Keywords: sign test, Wilcoxon rank sum test, Mann-Whitney-Wilcoxon signed rank test.

1. Introduction

1.1. Fuzzy P-values

Geyer and Meeden (2005) introduce the notion of a *fuzzy P-value*, which for our purposes can be described as follows. If P is a random variable that has a $\text{Uniform}(0, 1)$ distribution unconditionally under the null hypothesis H_0, then a random variable having the *conditional distribution* of P given observed data Y is an *exact fuzzy P-value*. The test that rejects H_0 when $P \leq \alpha$ rejects with probability α under H_0 because

$$E\{\Pr[P \leq \alpha | Y]\} = \Pr(P \leq \alpha) = \alpha \tag{1}$$

[compare with equation (1.7) in Geyer and Meeden (2005)].

Geyer and Meeden (2005) recommend that this *distribution* of P given Y and not a number deemed to be a random realization of it is what a statistician or scientist should report in a situation where classical randomized tests are appropriate. The reason is that, as has long been recognized, it is absurd that two statisticians use exactly the same procedure on exactly the same data but report different results, and this has been (we think) the main reason why randomized tests are not used in practice, despite some of them being uniformly most powerful and recommended by the widely accepted theory theory of hypothesis testing (Lehmann 1959). The solution proposed by Geyer and Meeden (2005) is to not report a realization of the random variable (because every realization is different) but to report the abstract random variable (a description of its probability distribution, which is always the same). We follow the same principle here. A fuzzy P-value is not a single number but a random variable having a probability distribution and described by a picture of its probability density function (PDF) or its cumulative distribution function (CDF) or by some other mathematical characterization of the distribution.

1.2. Interpretation

We know how to interpret conventional P-values. Low P-values are good (if you are on the side that wants to reject the null hypothesis), the lower the better. High P-values are bad (for your side). Somewhere in the middle, traditionally around 0.05, P-values are equivocal.

Fuzzy P-values are somewhat harder to interpret because they are fuzzy, smeared out over a range of values. But it is still the case that extremes are easy to interpret. If the whole distribution of a fuzzy P-value is concentrated below 0.01, then this is strong evidence against the null hypothesis. If the whole distribution of a fuzzy P-value is concentrated above 0.2, then this is evidence against the null hypothesis so weak as to be practically nonexistent. The strength of evidence goes from one extreme to the other as the distribution of the fuzzy P-value shifts. In the middle it is equivocal.

For concreteness, suppose we are doing a lower-tailed sign test with sample size 10 and observe $T = 2$ for the test statistic. The conventional P-value is $P = 0.055$, which is borderline statistically significant but not actually below 0.05. The possible significance levels for the conventional test are (using R)

```
> round(pbinom(0:4, 10, 1 / 2), 5)
[1] 0.00098 0.01074 0.05469 0.17188 0.37695
```

We see that the next smallest possible P-value is P = 0.011 (for $T = 1$). So this is not analogous to a t-test or any other hypothesis test with a continuous reference distribution. Intuitions that that come from experience with t-tests are not transferable.

We can state this point more pedantically by saying the test is *exact* only if one is using a significance level (α level) that is a value the cumulative distribution function of the test statistic (0.00098, 0.01074, etc. in our example). Otherwise the test is *conservative* (the power under the null hypothesis is greater than the significance level), and the conventional nonrandomized test is only *conservative-exact* rather than exact.

For the fuzzy test for $T = 2$, the fuzzy P-value is uniformly distributed on the interval $(0.011, 0.055)$. It is mostly below 0.05, hence this result is more analogous to a t-test with $P < 0.05$ than one with $P > 0.05$.

Although this procedure was only proposed in 2005, it has always been implicit in the classical theory of randomized hypothesis tests (Lehmann 1959). If one generates a realization of this fuzzy P-value and says “reject H_0” if $P < \alpha$ and “accept H_0” if $P > \alpha$, then this is the classical randomized test.

The conventional P-value is the upper endpoint of the support of the fuzzy P-value. So the fuzzy P-value provides more information than the conventional P-value. Given how easy it

Table 1: Endpoints of Supports of Fuzzy P-values. T is a random variable having the null distribution of the test statistic, t is the observed value of the test statistic, and τ is the center of symmetry of the null distribution of the test statistic. The fuzzy P-value is uniformly distributed on the interval with the endpoints indicated in the table.

	lower endpoint	upper endpoint
upper-tailed	$\Pr(T > t)$	$\Pr(T \geq t)$
lower-tailed	$\Pr(T < t)$	$\Pr(T \leq t)$
two-tailed	$\Pr(\lvert T-\tau\rvert > \lvert t-\tau\rvert)$	$\Pr(\lvert T-\tau\rvert \geq \lvert t-\tau\rvert)$

is to provide the fuzzy P-value, there seems no reason to prefer the conventional P-value. Readers who don't like fuzzy P-values can ignore all but their upper endpoints.

2. The sign test

The sign test (Hollander and Wolfe 1999, Section 3.4) concerns the value of the population median μ of data X_1, X_2, …, X_n assumed to be independent and identically distributed (IID) from some (completely unspecified) distribution. The test statistic of the sign test is the number T of X_i that are greater than the value of μ assumed under the null hypothesis.

2.1. No ties

In this section we assume $\Pr(X_i = \mu)$ is zero, as happens when the distribution of the X_i is continuous. Under the null hypothesis the distribution of T is Binomial$(n, 1/2)$.

Fuzzy P-values allow an exact test at any significance level. Geyer and Meeden (2005) discuss uniformly most powerful (UMP) and UMP unbiased (UMPU) tests for binomial data. Although these tests are different from the sign test, they have the same null distribution of their test statistic, and hence can use the same fuzzy P-values. If we follow the same logic as Geyer and Meeden (2005) in the special case where the test statistic has a *symmetric* binomial distribution, the fuzzy P-value is uniformly distributed on the intervals shown in Table 1 (where for the sign test $\tau = n/2$).

This fuzzy test is UMP (one-tailed) or UMPU (two-tailed) given the decision to use T as the test statistic. It is, of course, not UMP or UMPU for the original data. For example, a conventional z-test would be UMP or UMPU if the data were normal with known variance, and no test is most powerful regardless of the distribution of the data.

2.2. Ties

Now we allow for the case where $\Pr(X_i = \mu)$ is non-zero. Before we advance our proposal, we say a few things about currently available methods.

Currently available methods

The simplest and most popular method of dealing with ties, which we call the *standard* *method*, simply ignores them (Hollander and Wolfe 1999, pp. 62 and 67). The resulting test is a valid conditional test (conditioning on the observed number of ties), but that validity is something that could please only a theoretician.

Consider the following data: we observe 1 data point below μ, 90 equal to μ, and 9 above μ. The "standard method" simply ignores the 90 ties and proceeds as if we had seen 10 data points, (1 below and 9 above). The conventional two-tailed P-value is then $P = 0.021$, which implies fairly strong evidence against the null hypothesis that the true median is μ. But the 90% of the data that are exactly at μ strongly supports the null hypothesis that μ is the median. So here we have a case where the data are overwhelmingly in favor of the null hypothesis and yet the "standard method" rejects it.

To be fair to Hollander and Wolfe (1999), they do say "this approach is satisfactory as long as [ties] do not represent a sizable percentage of the total" (we would argue it is never "satisfactory" to ignore data favoring the null hypothesis). Moreover, they outline two other procedures. One we call the *conservative method* counts all ties as favoring the null hypothesis. The other we call the *randomized method* counts ties as favoring the null or alternative at random with equal probabilities.

Fuzzy methods

Here is a simple alternative method using fuzzy P-values. "Unrandomize" the randomized method. Add conceptual infinitesimal jitter to the data to break ties. Suppose the observed data have l, t, and u points below, tied with, and above (respectively) the hypothesized value of μ. After jittering we have $l + K$ points below, $u + t - K$ above, and no ties, where K is a Binomial$(t, 1/2)$ random variable. If we actually observed K we would know how to compute the fuzzy P-value. When we don't observe K, we calculate the distribution of the fuzzy P-value mixing over K.

Here is an example. Suppose we observe 2, 3, and 12 data points below, tied with, and above the hypothesized value of μ (respectively), and wish to do an upper-tailed test. Figure 1 shows the PDF of the fuzzy P-value. The areas under the four steps are in ratios $1 : 3 : 3 : 1$, the ratios of probabilities for the Binomial$(3, 1/2)$ distribution of K in this case. The step boundaries are the relevant probabilities of the Binomial$(17, 1/2)$ distribution of the sign test statistic, which ranges from 12 to 15 as K ranges from 0 to 3, calculated by the expression

```
   1 - pbinom(11:15, 17, 1/2)
```

in R (R Core Team 2013). Figure 1 was made by the R contributed package **fuzzyRankTests** (Geyer 2013), available from CRAN, which does all calculations and plots discussed in this article.

So much for computation, now for the interpretation. The evidence against the null hypothesis is not overwhelming, since some of the distribution of the fuzzy P-value is above 0.05 (about six percent). It is, however, fairly strong. The very liberal "standard method" ignores the three ties, giving $P = 0.006$. The very conservative "conservative method" counts the ties as evidence in favor of the null hypothesis giving $P = 0.072$ (the upper bound of the support of the fuzzy P-value).

We hope the reader will agree liberal is too liberal and conservative too conservative and fuzzy is a happy medium.

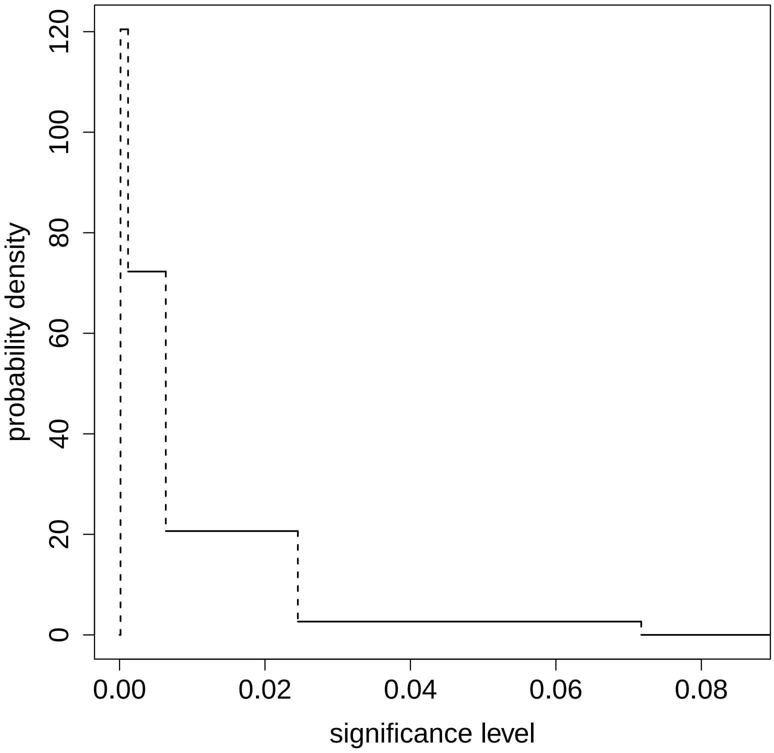

Figure 1: Fuzzy P-value for upper-tailed sign test. Data are 2 below, 3 tied with, and 12 above the hypothesized value of the median under the null hypothesis.

Disclaimer

No method is beyond criticism. The procedure outlined in the preceding section is hard to justify scientifically if it is hard to imagine a hidden continuous variable that has been rounded or grouped to form the observed data.

We agree with Hollander and Wolfe (1999, p. 67) where they say that nonparametrics (regardless of what method one uses for dealing with ties) is a poor substitute for methods specifically designed for ordered categorical data. If one thinks an ordered categorical model is scientifically appropriate, then that is what one should use.

Our proposed fuzzy P-values are exact only when the null hypothesis is about the median of the *actual data, however rounded or grouped.* The test cannot be exact when the null hypothesis is about the median of the unobserved unrounded data, nor could it be without a specific model for the joint distribution of the unrounded data and the rounded or grouped data. If one did have such a specific model, then the method for obtaining fuzzy P-values in latent variable situations proposed by Thompson and Geyer (2007) could be used.

3. Other rank tests

We give only a brief discussion of fuzzy P-values for the Wilcoxon rank tests. Implementation details are extensively discussed in the design document found in the `doc` directory of the **fuzzyRankTests** package (Geyer 2013).

3.1. Mann-Whitney-Wilcoxon rank sum test

The methodology for other classical rank tests is similar. So long as the null distribution of the test statistic is symmetric, Table 1 is still valid. Suppose we are doing a two-tailed Wilcoxon rank sum test with data

X	1.2	3.7	5.1	8.3	12.1	12.1	
Y	8.3	12.1	14.3	15.2	18.7	19.7	21.2

The standard methodology uses tied ranks, but we do not. We use

X	1	2	3	4–5	6–8	6–8	
Y	4–5	6–8	9	10	11	12	13

the notation 4–5 meaning tied for ranks 4 and 5 and similarly for 6–8. When we break the ties by infinitesimal jittering, then the possibilities 4 and 5 for the smallest Y value are equally likely as are the possibilities 6, 7, and 8 for the next Y value. These possible values lead to possible values of the Mann-Whitney form of the test statistic, the number of (X, Y) pairs for which $X > Y$, between 2 and 5 and an elementary probability calculation gives corresponding probabilities in the ratios $1 : 2 : 2 : 1$. Figure 2 shows the PDF of the fuzzy P-value. The areas under the four steps are in ratios $1 : 2 : 2 : 1$ and the step boundaries are calculated by the expression

```
2 * pwilcox(1:5, 6, 7)
```

in R (the 2 is for two tails).

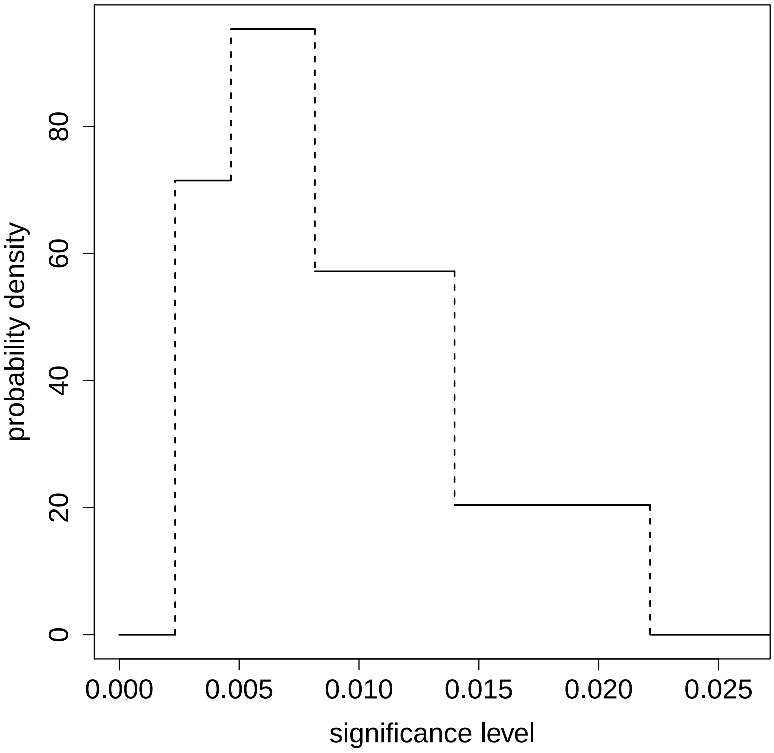

Figure 2: Fuzzy P-value for two-tailed rank sum test.

In general, the contribution of each tied class to the latent test statistic also has the Mann-Whitney distribution but with sample sizes that are the number of x and y values in the tied class. Then one performs a convolution operation to find the distribution of the sum of the contributions (or uses the R package **fuzzyRankTests**).

3.2. The Wilcoxon signed rank test

Other rank tests are similar to those already discussed. We mention only one particular detail about the signed rank test. When there are ties, there are two kinds of tied classes. The contribution to the latent test statistic of the class tied at the value μ hypothesized under the null hypothesis has the Wilcoxon signed rank distribution but with sample size the number of data values tied with μ. The contributions to the latent test statistic of all other tied classes have the Mann-Whitney distribution but with sample sizes that are the number of data values in the tied class above and below μ.

4. Discussion

The methods discussed in this article are not the only methods for obtaining what some call exact tests in the presence of ties. The network and shift algorithms (Mehta and Patel 1983; Streitberg and Röhmel 1987) as implemented in the proprietary software StatXact (`http://www.cytel.com`) and the free software R with the **exactRankTests** contributed package (R Core Team 2013; Hothorn 2001; Hothorn and Hornik 2013) also perform exact rank tests, but the P-values they produce are not exact in the sense used in this article. They are "conservative-exact" being exact only for a few significance levels and conservative for others because they do not randomize. Of course, fuzzy versions of those procedures could also be derived. (The **exactRankTests** package still supported but no longer under development. The **coin** package (Hothorn, Hornik, van de Wiel, and Zeileis 2006, 2008) is its successor.)

Infinitesimal jittering has been used in Bayesian inference, apparently originated by Hill (1968). Formal theory for that idea needs finitely additive probability (Lane and Sudderth 1978), but we do not see the need for that here, where the infinitesimal jittering story merely motivates a randomization scheme.

As mentioned in Section 2.2.3, another approach to rounded or grouped data is to explicitly model the distribution of the unobservable unrounded data and the rounding-grouping mechanism. We have a method for that approach (Thompson and Geyer 2007). In the case of the rank sum test, with the standard null hypothesis that the unrounded data are independent and identically distributed from a common continuous distribution and are rounded to obtain the observed data, the method of Thompson and Geyer (2007) and the method of Section 3 give the same fuzzy P-values. Other situations or other assumptions lead to other fuzzy P-values, but this more complicated procedure would have little impact on the fuzzy P-value with most latent data models.

We claim that the fuzzy P-value procedures described here are by far the easiest to understand. They are simple enough so that non-expert users can understand and implement them.

A. Theory

This section proves the basic ideas without going through UMP theory. Readers who have taken or taught a theoretical probability and statistics course should recognize a slight variant of a standard exercise.

Let X be a random variable and F its CDF. Let F_- be the left continuous version of F, that is

$$\begin{aligned} F(x) &= \mathrm{pr}\{X \leq x\} \\ F_-(x) &= \mathrm{pr}\{X < x\} \end{aligned}$$

Let U be a Uniform$(0, 1)$ random variable independent of X. We claim the random variable

$$V = UF(X) + (1 - U)F_-(X)$$

is Uniform$(0, 1)$.

Fix v such that $0 < v < 1$ and x such that $F_-(x) \leq v \leq F(x)$. Write $f(x) = F(x) - F_-(x)$. Then

$$\begin{aligned} \mathrm{pr}\{V \leq v\} &= \mathrm{pr}\{X < x\} + \mathrm{pr}\{V \leq v \mid X = x\} \cdot \mathrm{pr}\{X = x\} \\ &= F_-(x) + v - F_-(x) \\ &= v \end{aligned}$$

the second term on the right hand side of the first line being zero if $f(x) = 0$ and otherwise the conditional probability in that term being

$$\mathrm{pr}\{F_-(x) + Uf(x) \leq v\} = \frac{v - F_-(x)}{f(x)}$$

That finishes the proof of the claim and shows that our recommended procedure (based on Table 1) does indeed obey equation (1) when there are no ties.

The main text of the article implicitly gives the argument for ties. We have three (correlated) random objects, the original data, the "infinitesimally jittered" data, and the fuzzy P-value P. By the theorem above P is unconditionally Uniform$(0, 1)$ because it is based (theoretically) on the infinitesimally jittered data, which has no ties. Thus we get a valid fuzzy P-value when we condition on the original data.

B. Confidence intervals

Some readers may be wondering about fuzzy confidence intervals dual to our fuzzy tests. They are simple mixtures of standard confidence intervals. We give one example.

Suppose we have sorted data

3.13 3.48 3.50 4.70 4.76 4.82 5.28 5.67 5.82 8.67

and one wants a confidence interval for the population median dual to the sign test, which (Hollander and Wolfe 1999, Section 3.6) has end points that are symmetric pairs of order statistics. The three intervals with largest confidence levels are

interval	level
(3.13, 8.67)	99.8
(3.48, 5.82)	97.9
(3.50, 5.67)	89.1

The 95% fuzzy confidence interval formed from these has membership function that is 1.0 on the interval $(3.50, 5.67)$, is 0.676 on the part of $(3.48, 5.82)$ not included in the narrower interval, and 0.0 elsewhere. The corresponding randomized confidence interval is $(3.48, 5.82)$ with probability 0.676 and $(3.50, 5.67)$ with probability 0.324.

Without giving details of the argument, we assure the reader that if one goes through the argument of Geyer and Meeden (2005, Section 1.4) one does indeed find that the critical function $\phi(x, \alpha, \mu)$ associated with the randomized sign test described in Section 2.1 above is a constant function of μ on intervals between order statistics and does give the fuzzy confidence intervals just described. Detailed proofs of this and many other theoretical points in this article can be found in the design document found in the `doc` directory of the **fuzzyRankTests** package (Geyer 2013).

When the data distribution is continuous so no ties are possible, the values at discontinuities of (the membership function of) the fuzzy confidence interval do not matter. If ties can occur, then values at discontinuities do matter. The correct values are found by inverting the fuzzy test (and done by the R package **fuzzyRankTests**).

Another issue that arises when ties can occur is that some of the intervals on which the fuzzy confidence interval is constant (which have order statistics as end points) can collapse because their end points are tied. The R package **fuzzyRankTests** also deals with this situation correctly. Implementation details are discussed in the aforementioned design document.

Acknowledgments

We thank Gary Oehlert and Bill Sudderth for helpful comments.

References

Geyer CJ (2013). ***fuzzyRankTests**: Fuzzy Rank Tests and Confidence Intervals.* R package version 0.3-5 (first CRAN version 2005), URL http://CRAN.R-project.org/package=fuzzyRankTests.

Geyer CJ, Meeden GD (2005). "Fuzzy and Randomized Confidence Intervals and P-Values (with discussion)." *Statistical Science*, **20**, 358–387.

Hill BM (1968). "Posterior Distribution of Percentiles: Bayes' Theorem for Sampling from a Population." *Journal of the American Statistical Association*, **63**, 677–691.

Hollander M, Wolfe DA (1999). *Nonparametric Statistical Methods.* second edition. Wiley-Interscience, New York.

Hothorn T (2001). "On Exact Rank Tests in R." *R News*, **1**, 11–12.

Hothorn T, Hornik K (2013). ***exactRankTests**: Exact Distributions for Rank and Permutation Tests.* R package version 0.8-24, URL `http://CRAN.R-project.org/package=exactRankTests`.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**, 257–263.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**, 1–23. URL `http://www.jstatsoft.org/v28/i08/`.

Lane DA, Sudderth WD (1978). "Diffuse Models for Sampling and Predictive Inference." *The Annals of Statistics*, **6**, 1318–1336.

Lehmann EL (1959). *Testing Statistical Hypotheses.* John Wiley & Sons, New York. 2nd edition, Wiley, 1986 and Springer, 1997; 3rd edition, co-authored with Joseph P. Romano, Springer 2010.

Mehta CR, Patel NR (1983). "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables." *Journal of the American Statistical Association*, **78**, 427–434.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Streitberg B, Röhmel J (1987). "Exakte Verteilungen für Rang- und Randomisierungstests im allgemeinen c-Stichprobenfall." *EDV in Medizin und Biologie*, **18**, 12–19.

Thompson EA, Geyer CJ (2007). "Fuzzy p-Values in Latent Variable Problems." *Biometrika*, **94**, 49–60.

Affiliation:

Charles J. Geyer
School of Statistics
University of Minnesota
313 Church Hall
224 Church Street SE
Minneapolis, MN 55455, USA
E-mail: `geyer@umn.edu`
URL: `http://users.stat.umn.edu/~geyer/`

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