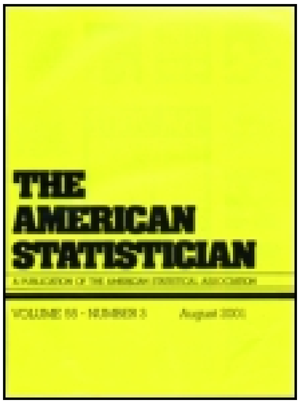

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Use of the Sign Test for the Median in the Presence of Ties

Daniel Y. T Fong[a], C. W Kwan[a], K. F Lam[a] & Karen S. L Lam[a]

[a] Daniel Y. T. Fong is Research Assistant Professor, Clinical Trials Centre, Faculty of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong . C.W. Kwan is Post Doctoral Fellow, Clinical Trials Centre, Faculty of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong. K. F. Lam is Assistant Professor, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. Karen S. L. Lam is Professor, Department of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong. The authors thank the editor, an associate editor, and a referee for their helpful comments on this article. The study used in the example was financially supported by the Li Ka Shing Foundation. Financial support by the RGC (Block Grants) was gratefully acknowledged.
Published online: 01 Jan 2012.

To cite this article: Daniel Y. T Fong, C. W Kwan, K. F Lam & Karen S. L Lam (2003) Use of the Sign Test for the Median in the Presence of Ties, The American Statistician, 57:4, 237-240, DOI: 10.1198/0003130032422

To link to this article: http://dx.doi.org/10.1198/0003130032422

Use of the Sign Test for the Median in the Presence of Ties

Daniel Y. T. FONG, C. W. KWAN, K. F. LAM, and Karen S. L. LAM

The sign test was developed to examine the median difference of paired samples. Ignorance of ties—that is, zero differences—in the sign test can substantially inflate the Type I error rate. A uniformly most powerful test has been developed to test for the symmetry of the positive and negative differences but little attention has been given about inferences concerning the median difference when ties are abundant. This article examines a simple modified sign test which can be implemented in most computer packages and a likelihood ratio test.

KEY WORDS: Likelihood ratio test; Paired samples; Power; Type I error.

1. INTRODUCTION

Hypotheses tests involving medians are important topics of statistical inference when the population distributions are not symmetrical. For a random sample (or paired samples), the inference in this article considers testing the hypothesis

$$\mathrm{H}_0 : M = 0 \quad \text{versus} \quad \mathrm{H}_A : M \neq 0, \tag{1}$$

where M denotes the population median. For this application, the sign test is a valid test as it does not depend on any distributional assumptions about the population. The conventional use of the sign test does not account for ties when zero values are observed. This article examines the problem of using the sign test when ties are present and recommends methods for handling ties.

The sign test was first introduced more than 50 years ago by Dixon and Mood (1946) for testing the median difference of independent paired observations. It has a broad range of applications such as aquatic science (El-Shaarawi, Esterby, Warry, and Kuntz 1985); signal processing (Brikker and Gorbunov 1989); medicine (Bartfield, Crisafulli, Raccio-Robak and Salluzzo 1995); and genetics (Orr 1998). A major assumption of the conventional use of sign test is the absence of ties; however, ties can often occur in practice due to rounding of continuous measurements or discreteness of the data. In the presence of ties, the sign test for testing (1) is often performed only on the untied data which can substantially inflate the Type I error rate. For instance, a set of 20 observations where six values are positive and 14 are zeroes gives a p value (two-sided) of 0.031 from a sign test when all ties are discarded. This results in a conclusion that the median differs from 0 at the 5% level of significance. However, counter to intuition, the sample median is zero.

Some authors modified the sign test in the presence of ties for testing hypotheses other than (1). For example, in an opinion poll, ties can frequently occur when pollsters press those who are undecided. When the objective is to test for the difference of positive and negative responses (directional symmetry), Coakley and Heise (1996) proposed a uniformly most powerful nonrandomized test. Under the same objective, Rayner and Best (1999) proposed a different modification of the sign test by introducing the concept of leakage when the tied responses can be treated as leaning positive or leaning negative. On the other hand, Randles (2001) considered tests of whether positive responses are more than negative responses by the number of tied responses (majority preference) and whether the difference is only greater than a certain proportion of tied responses (intermediate preference). However, when the objective is testing about the median, the above procedures for testing directional symmetry are not applicable (Emerson and Simon 1979). Testing for majority preference is equivalent to a one-sided test of median. However, the treatment of ties considered by Randles (2001) is too conservative for testing about the median, although it is powerful for testing the majority preference hypotheses. Therefore, the manner in which ties should be treated depends on the question being asked (Emerson and Simon 1979).

In the literature, there are two possible approaches to handling ties for the purpose of testing (1). The first is to equally likely assign ties as positive or negative; the second is to assign a positive or negative sign to ties in a way that makes significance less likely (Sprent 1993). The first approach often yields little difference when compared with discarding ties, and the second approach is clearly too conservative. On the other hand, a testing procedure for (1) described by Emerson and Simon (1979) is to test H_0 by using the p value

$$2P(N \geq \max\{n_-, n_+\}), \tag{2}$$

where $N \sim \mathrm{Binomial}(n, 1/2)$, n is the total number of independent observations, and n_- and n_+ are numbers of observations below and above zero, respectively. It can be shown that (2) can attain a maximum value of 2 when all observations are tied. Thus, this procedure has unnecessarily low power.

There is a clear inadequacy in the literature for properly adjusting the sign test for testing (1) in the presence of ties. This motivates the consideration of alternative methods. Two different alternative testing procedures based on the sign test are described in Section 2. The first procedure is a simple modification

Daniel Y. T. Fong is Research Assistant Professor, Clinical Trials Centre, Faculty of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong (E-mail: dytfong@hku.hk).C. W. Kwan is Post Doctoral Fellow, Clinical Trials Centre, Faculty of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong. K. F. Lam is Assistant Professor, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. Karen S. L. Lam is Professor, Department of Medicine, The University of Hong Kong, Pokfulam Road, Hong Kong. The authors thank the editor, an associate editor, and a referee for their helpful comments on this article. The study used in the example was financially supported by the Li Ka Shing Foundation. Financial support by the RGC (Block Grants) was gratefully acknowledged.

of (2) which can be implemented in most statistical packages. The second procedure is constructed based on the likelihood ratio statistics with the critical value obtained by a parametric bootstrap. The performance of the two procedures is compared with existing procedures in a simulation study in Section 3. In Section 4, an example from a clinical trial is used for illustrating the proposed tests. Some concluding remarks are provided in Section 5 about the use of the sign test for the median when ties are present.

2. TWO VERSIONS OF THE SIGN TEST IN THE PRESENCE OF TIES

2.1 A Modified Sign Test

The sign test can be modified by correcting the maximum of (2). With the notation used in Section 1, use

$$P(N \geq \max\{n_-, n_+\})/P\left(N \geq \left[\frac{n - n_0 + 1}{2}\right]\right) \quad (3)$$

as the p value for testing H_0 against H_A, where $[x]$ represents the largest integer smaller than x, and n_0 is the number of observations equal to 0. It can be shown that $2P(N \geq [n - n_0 + 1/2])$ is the maximum of (2) so that the maximum of (3) is 1. Therefore, in general, (3) is strictly smaller than (2) but with a fixed number of ties, the difference tends to 0 as the total number of observations approaches ∞.

When there are no ties, (2) is equivalent to the conventional sign test. However, (3) reduces to that of the conventional sign test only when n is odd. When n is even, (3) is slightly smaller and the denominator becomes $P(N \geq [n/2]) = 1/2 + O(n^{-1/2})$.

2.2 A Likelihood Ratio Test

To fix ideas, let X_1, X_2, ..., X_n be the n independent and identically distributed observations, and $P(X_i < 0) = p_-$, $P(X_i = 0) = p_0$, and $P(X_i > 0) = p_+$ such that $p_- + p_0 + p_+ = 1$. If N_-, N_0, and N_+ denote, respectively, the numbers of X_i's below, equal to, and above 0,(N_-, N_0, N_+) follows a multinomial distribution with parameters n, and (p_-, p_0, p_+). Then, following the definition of a median in Hogg and Craig (1995), the null hypothesis $\mathrm{H}_0 : M = 0$ can be equivalently written as $\mathrm{H}_{10} : p_- \leq 1/2$ and $p_- + p_0 \geq 1/2$. Under H_{10}, the maximum likelihood estimates of p_-, p_0, and p_+ are

$$\tilde{p}_- = \min\left\{\frac{n_-}{n}, \frac{1}{2}\right\},$$

$$\tilde{p}_0 = \max\left\{\frac{1}{2} - \tilde{p}_-, \frac{n_0}{n}\right\},$$

and

$$\tilde{p}_+ = 1 - \tilde{p}_- - \tilde{p}_0,$$

respectively. The corresponding unrestricted estimates are given by $\hat{p}_- = n_-/n$, $\hat{p}_0 = n_0/n$, and $\hat{p}_+ = n_+/n$, and the likelihood ratio statistic is

$$\lambda = \left(\frac{\tilde{p}_-}{\hat{p}_-}\right)^{n_-}\left(\frac{\tilde{p}_0}{\hat{p}_0}\right)^{n_0}\left(\frac{\tilde{p}_+}{\hat{p}_+}\right)^{n_+}.$$

Although $-2\log\lambda$ is asymptotically distributed as χ^2, the corresponding degrees of freedom is difficult to obtain with the composite hypothesis H_{10}. Instead, parametric bootstrap is used to approximate the p value (Davison and Hinkley 1997). Specifically, by taking $\tilde{p}_-$, $\tilde{p}_0$, and $\tilde{p}_+$ as the values of p_-, p_0, and p_+, generate m independent samples of (n_-, n_0, n_+) from the multinomial distribution. By letting λ_i as the likelihood ratio of the ith generated sample, the Monte Carlo p value is

$$\frac{1 + \#\{\lambda_i \leq \lambda\}}{1 + m}.$$

The addition of 1 on both the numerator and denominator takes account the observed sample is an independent sample in addition to the m bootstrapped samples (Davison and Hinkley 1997).

3. A SIMULATION STUDY

This section investigates the Type I error rate and the power of the modified sign test (3) and the likelihood ratio test (LRT) with different amounts of ties. In addition, the procedure (2) from Emerson and Simon (1979) and the conventional sign test on only untied observations are also examined. Without loss of generality, assume $p_- \geq p_+$. Then, by letting $\Delta = p_- - p_+ \geq 0$, we have $M = 0$ if and only if $0 \leq \Delta \leq p_0$.

We compare the four versions of the sign test for handling ties in two scenarios. The first scenario is different nonzero values of M to assess the power of the tests. The second scenario is $M = 0$ with different asymmetric levels of positive and negative signs to assess the Type I error rate of the tests. Samples under the two scenarios are generated as follows.

In scenario 1, a total of 5,000 samples of (n_-, n_0, n_+) are generated from a multinomial distribution with parameters n, p_-, p_0, and p_+ determined from a set of hypothesized values. Values of the sample size n are taken as 80 and 160 and p_0 is taken as any value in steps of 0.1 between 0 and $\min\{\Delta - 0.1, 1 - \Delta\}$, where Δ is taken as a value between 0.1 and 1 in steps of 0.1. Therefore, $p_- = (1 + \Delta - p_0)/2$, and $p_+ = 1 - p_0 - p_-$. Smaller values of p_0 than Δ correspond to larger deviations of M from 0.

In scenario 2, another 5,000 samples are similarly generated with n taken as 80 and 160 with values of p_0 taken between Δ and $1 - \Delta$ in steps of 0.1, where Δ = 0, 0.1, 0.2, and 0.4. Thus, $p_- = (1 + \Delta - p_0)/2$ and $p_+ = 1 - p_0 - p_-$. Larger Δ values give the directional asymmetry.

The conventional sign test, procedure (2) from Emerson and Simon (1979), the modified sign test, and the LRT are applied to all the generated samples. Particularly, the Monte Carlo p value from the likelihood ratio approach is obtained from $m = 500$ bootstrapped samples. The proportion of samples rejected out of 5,000 (the rejection rate), using a 5% level of significance, is computed for each set of hypothesized parameter values. The simulation study was programmed in SAS version 8.2.

Figure 1(a) presents the rejection rates of the four versions of the sign test in samples of size 80 when there are 20% ties. In this case, $M = 0$ is equivalent to $0 \leq \Delta \leq 0.2$, where rejection of H_0 is a Type I error. Note that when $\Delta = 0$, that is, the underlying distribution is directionally symmetric, the conventional sign test leads to a larger Type I error rate than the other three tests. This is due to the difference of sample sizes considered. Indeed, when there are no ties, all tests have the same Type I error rate (not shown here). There is no doubt that the application of the conventional sign test after discarding ties

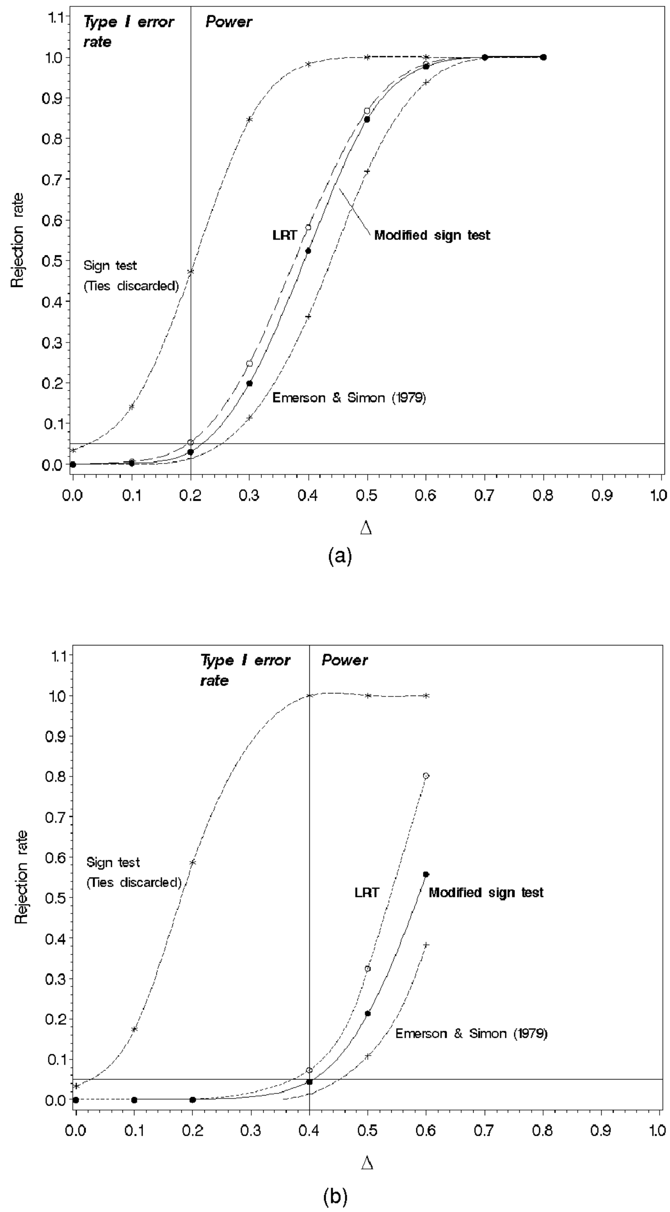

Figure 1. Rejection rates of different versions of sign test in samples of size 80 when (a) p_0 = 0.2, and (b) p_0 = 0.4.

seriously inflates the Type I error rate while the Type I error rate of other three tests is well controlled below 0.05. On the other hand, when $\Delta > 0.2$, the rejection rate corresponds to the power of a test and the modified sign test in (3) is more powerful than (2) from Emerson and Simon (1979) but yet slightly smaller than the LRT.

When the chance of having a tie increases to 40%, the problem with inflated Type I error rate becomes more serious with the conventional sign test (Figure 1(b)). The modified sign test is now much less powerful than the LRT but still outperforms (2). Figure 2 shows the rejection rates of the tests when the sample size is doubled to 160. The Type I error rate is higher when ties are discarded while the powers of the other three tests become closer with negligible difference between the modified sign test and LRT. Moreover, when there are no ties, that is, $p_0 = 0$, the three curves of rejection rates (not shown) coincide at both sample sizes.

4. AN EXAMPLE

We illustrate the use of the modified sign test and LRT, together with the conventional sign test and procedure (2) in a clinical trial. The trial was conducted at a local hospital on patients with Type 2 diabetes mellitus (DM). A random sample of 100 patients consented to participate in the study. Apart from the usual demographic information, each study patient was interviewed with the Chinese version of the Beck Depression Inventory: second edition (BDI-II) at the beginning (baseline) and also after 18 weeks of the intervention. The BDI-II is a questionnaire composed of 21 questions which measures the level of depression of a subject. Each question was rated on a four-point scale (0–3) and the total score was used to measure the depression level. The higher the score, the more severe the depression. Two objectives of the study were to examine the depression level of DM patients at baseline (a level less than 2 is treated as normal) and whether the intervention improved depression after 18 weeks. In this situation, even when more patients with depression improved than deteriorated, the intervention may not be legitimately considered as "effective" if a substantial number of subjects had no (or zero) improvement. Therefore, the median is a more relevant parameter to be estimated. Table 1 summarizes the results of the analysis by the four versions of sign test.

At baseline, the median depression is 0 in the male, female, and total sample. The four versions of the sign test conclude with a significant difference of depression level from 2 at a 5% level of significance. With no ties in the female sample, the p values of the three tests coincide ($n = 63$ is odd). Thus, the DM patients did not appear to have depression problems.

When examining the 18-week change of depression in the total sample, the median change is 0. The conventional sign test gives contradictorily a highly significant p value. In contrast, the other three tests correctly conclude that there was insufficient evidence to show that the intervention improved depression.

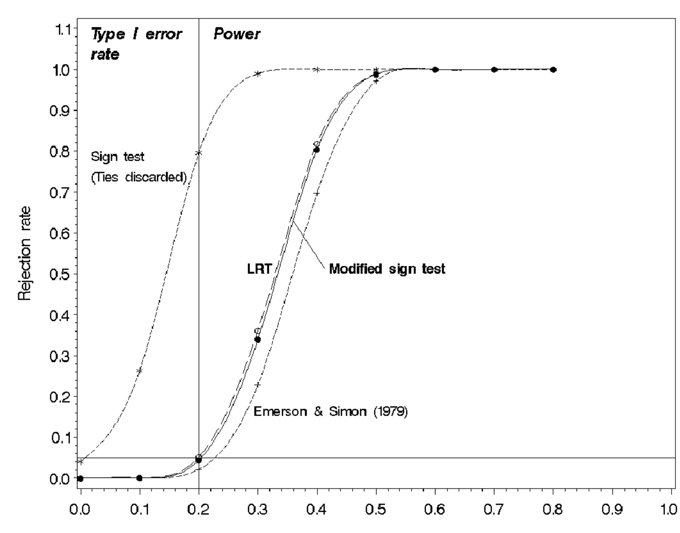

Figure 2. Rejection rates of different versions of sign test in samples of size 160 when p_0 = 0.2.

Table 1. Use of Different Versions of the Sign Test in Examining the Depression Level of Patients with Type 2 Diabetes Mellitus

				p value (sign test)			
Group	n	(n_-, n_0, n_+)	Median	Ties discarded	Emerson & Simon (1979)	Modified	LRT
BDI at Week 0 (H_0: Median = 2)							
Male	37	(25,6,6)	0	0.001	0.047	0.028	0.014
Female	63	(58,0,5)	0	1.66×10^{-12}	1.66×10^{-12}	1.66×10^{-12}	0.002
Total	100	(83,6,11)	0	<0.001	<0.001	<0.001	0.002
Change of BDI (Week 18–Week 0) (H_0: Median = 0)							
Male	34	(6,14,14)	0	0.115	1.771	0.889	1.000
Female	61	(3,25,33)	1	<0.001	0.609	0.305	0.257
Total	95	(9,39,47)	0	<0.001	1.162	0.581	1.000

Here procedure (2) gives a rather unusual p value greater than 1. When we explore further, p values from the Student's t test and the Wilcoxon signed rank test are 0.044 and 0.0002, respectively. The small p value from the Wilcoxon signed rank test is due to the ignorance of ties in SAS while the t test is sensitive to a few extreme values. Indeed, the mean and variance of the change of depression are 0.94 and 20.04, respectively. The mean is relatively small when compared with the range of possible values of depression (i.e., 0–63). Conclusions from the t test and the Wilcoxon signed rank test are not likely to be reasonable in this sample with 41.1% ties. Similar observations can be made for the female sample but all versions of sign test yield the same conclusion on the change of depression in the male sample.

In all cases, the p value from the conventional sign test is consistently the smallest while that of (2) is consistently the largest. In contrast, the modified sign test and LRT consistently provide the same conclusions.

5. CONCLUDING REMARKS

There is no doubt that ties have to be accounted in a careful analysis. In the presence of ties, the inferential question should be clarified before a proper choice of the sign test can be made. This article considers the use of the sign test for the median. The modified sign test and the LRT help to alleviate the problem of inflated Type I error rate with the conventional use of the sign test. With no ties, the simple modification has no effect with an odd number of observations. With an even number of observations, the effect is small and becomes negligible when the sample size gets larger. When ties are abundant, the modification substantially improves the Type I error rate when compared with the conventional sign test, and has good power when compared with the existing procedure of handling ties. The LRT, on the other hand, has good power properties but requires extra computing effort. When the sample size is large (e.g., 160), the power difference between the modified test and LRT becomes negligible but both tests substantially outperform test procedure (2) proposed by Emerson and Simon (1979).

The sign test depends on minimal assumptions about the population and is therefore a valid test in most applications. It remains a popular choice in situations when a symmetric or Gaussian distribution is in doubt. Virtually all statistical packages—including SAS, SPSS, and Stata—are equipped with the conventional sign test without the allowance for ties. Routine use of the conventional sign test in these packages without an awareness of the influence of ties will likely lead to misleading conclusions. In clinical trials, the statistical analysis plan is prepared prior to the data analysis. Any changes of the plan during the analysis have to be justified in the final trial report and are thus often avoided. A testing procedure that demands minimal assumptions is therefore often preferable. We recommend at least the routine use of the modified sign test in a trial analysis to minimize changes of the statistical analysis planduring the analysis. SAS code for computing the modified sign test and LRT is available from the first author upon request.

[Received December 2002. Revised July 2003.]

REFERENCES

Bartfield, J. M., Crisafulli, K. M., Raccio-Robak, N., and Salluzzo, R. F. (1995), "The Effects of Warming and Buffering on Pain of Infiltration of Lidocaine," *Academic Emergency Medicine*, 2, 254–258.

Brikker, A. M., and Gorbunov, A. O. (1989), "The Effect of Methods for Processing Concurrent Observations on the Characteristics of Nonparametric Signal Detectors," *Telecommunications and Radio Engineering*, 44, 71–74.

Coakley, C. W., and Heise, M. A. (1996), "Versions of the Sign Test in the Presence of Ties," *Biometrics*, 52, 1242–1251.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, United Kingdom: Cambridge University Press.

Dixon, W. J., and Mood, A. M. (1946), "The Statistical Sign Test," *Journal of the American Statistical Association*, 41, 557–566.

El-Shaarawi, A. H., Esterby, S. R., Warry, N. D., and Kuntz, K. W. (1985), "Evidence of Contaminant Loading to Lake Ontario from the Niagara River," *Canadian Journal of Fisheries and Aquatic Science*, 42, 1278–1289.

Emerson, J. D., and Simon, G. A. (1979), "Another Look at the Sign Test when Ties are Present: The Problem of Confidence Intervals," *The American Statistician*, 33, 140–142.

Hogg, R. V., and Craig, A. T. (1995), *Introduction of Mathematical Statistics*, New Jersey: Prentice-Hall International Inc.

Orr, H. A. (1998), "Testing Natural Selection vs. Genetic Drift in Phenotypic Evolution Using Quantitative Trait Locus Data," *Genetics*, 149, 2099–2104.

Randles, R. H. (2001), "On Neutral Responses (Zeros) in the Sign Test and Ties in the Wilcoxon-Mann-Whitney Test," *The American Statistician*, 55, 96–101.

Rayner, J. C. W., and Best, D. J. (1999), "Modelling Ties in the Sign Test," *Biometrics*, 55, 663–665.

Sprent, P. (1993), *Applied Nonparametric Statistical Methods*, New York: Chapman and Hall.