

Should the Median Test Be Retired From General Use?

Boris FREIDLIN and Joseph L. GASTWIRTH

Although several authors have indicated that the median test has low power in small samples, it continues to be presented in many statistical textbooks, included in a number of popular statistical software packages, and used in a variety of application areas. We present results of a power simulation study that shows that the median test has noticeably lower power, even for the double exponential distribution for which it is asymptotically most powerful, than other readily available rank tests. We suggest that the median test be "retired" from routine use and recommend alternative rank tests that have superior power over a relatively large family of symmetric distributions.

KEY WORDS: Kolmogorov–Smirnov test; Median test; Nonparametric inference; Power; Two-sample rank tests; Wilcoxon test.

1. INTRODUCTION

Nonparametric tests provide a simple and reliable statistical tool in a variety of applications. They possess the desirable property of having the same sampling distribution for all continuous distributions. For the two-sample location shift setting commonly used nonparametric procedures include the Wilcoxon (Wilcoxon 1945), normal scores, and median (Mood 1954) tests. A number of authors (Ramsey 1971; Conover, Wehmanen, and Ramsey 1978) noted the poor performance of the median test in very small samples. Gastwirth and Wang (1987) showed that the loss of power of the median test relative to the Wilcoxon test increases in the case of highly unbalanced samples. Nevertheless, the median test continues to be used in applications (Hall, Louw, and Joubert 1995; Paternoster et al. 1996; Hernberg et al. 1998). Perhaps because of tradition the median test is presented in basic texts (Riegelman and Hirsch 1989; Sokal and Rohlf 1995; Anderson and Finn 1996; Zar 1998), in virtually all nonparametric texts directed at users

Boris Freidlin is Mathematical Statistician, Biometric Research Branch, National Cancer Institute, MSC7434, Bethesda, MD 20892. Joseph L. Gastwirth is Professor, Department of Statistics, George Washington University, Washington, DC 20052 (E-mail: jlgast@gwu.edu). This research was supported in part by a grant from the National Science Foundation and was completed while the second author was visiting the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics at the National Cancer Institute.

(Daniel 1990; Hollander and Wolfe 1998; Conover 1999) and is included in important statistical software packages (SAS, SPSS, StatXact). In this article we suggest that the median test be "retired" from routine use and recommend alternative rank tests that have superior power. In particular, the asymptotically optimal rank test for the shift alternative for data from a t_2 distribution (Gastwirth 1970) has superior power to the median test over a wide family of distributions in samples of up to 100 in each group.

The median test is valid under weaker conditions than the other rank tests. All that is needed is for the two distributions to have the same median and equal density functions in a neighborhood of that median. Thus, theoretically oriented texts (Randles and Wolfe 1979; Gibbons and Chakraborti 1992; Hettmansperger and McKean 1998) can use it to illustrate the assumptions underlying rank tests as well as their large sample properties.

2. BACKGROUND

Let X_1, \dots, X_m and Y_1, \dots, Y_n be independent random samples from populations with distribution functions $F(x)$ and $G(y)$, respectively. The objective is to test the hypothesis $H_0 : F(x) = G(x)$ against the location shift alternative $H_1 : G(x) = F(x - \theta)$.

All tests considered here are functions of the ranks of the combined sample. Most of the statistics are linear rank tests of the form

$$S_N = \sum_{i=1}^N a_N(i) \delta_i,$$

where $a_N(i)$ is the score function, δ_i is the indicator function which is equal to one if the i th smallest observation in the combined sample is from group 2, and $N = m + n$. For the median (Mood 1954) test $a_N(i) = \text{signum}(2i - N - 1)$. In other words, the test statistic is the number of the members of the second group that exceed the median of the combined sample. The median test is asymptotically most powerful for the double exponential distribution (Hajek and Sidak 1967). This is used to justify its inclusion in some statistical packages (Sall, Lehman, and Saul 1996). We compared the performance of the median test with the following rank tests:

1. The normal scores test (NS) with scores $a_N(i) = E(v_{(i)})$, where $v_{(i)}$ is the i th order statistic from a sample of N standard normal variables. This test is locally most powerful for the normal distribution.

Table 1. Empirical Power Estimates

Test	$n = m = 5$	$n = m = 9$	$n = m = 13$	$n = m = 25$	$n = m = 50$	$n = m = 100$	$n = 10, m = 40$	$n = 75, m = 25$
<i>Under normal distributions</i>								
NS	.811	.788	.801	.803	.803	.800	.797	.792
Wilcoxon	.812	.785	.795	.790	.785	.784	.780	.774
Median	.543	.626	.582	.597	.604	.607	.554	.587
LMPDE	.809	.722	.717	.684	.661	.645	.656	.642
T2	.814	.741	.737	.706	.682	.668	.672	.661
Cauchy	.750	.615	.584	.525	.490	.472	.447	.457
KS	.677	.675	.700	.683	.681	.679	.676	.674
<i>Under logistic distributions</i>								
NS	.800	.791	.791	.780	.788	.780	.793	.792
Wilcoxon	.801	.805	.805	.793	.802	.797	.800	.804
Median	.572	.689	.638	.649	.671	.671	.607	.665
LMPDE	.805	.777	.766	.732	.726	.709	.711	.719
T2	.806	.789	.782	.756	.757	.746	.737	.749
Cauchy	.762	.696	.669	.622	.612	.597	.543	.588
KS	.702	.731	.743	.719	.731	.726	.727	.738
<i>Under double exponential distributions</i>								
NS	.789	.756	.739	.707	.700	.666	.763	.730
Wilcoxon	.792	.791	.781	.758	.757	.731	.798	.782
Median	.610	.740	.698	.727	.766	.770	.661	.748
LMPDE	.801	.809	.804	.793	.806	.797	.764	.794
T2	.798	.810	.804	.788	.793	.771	.791	.803
Cauchy	.783	.765	.756	.747	.763	.749	.666	.747
KS	.734	.772	.781	.760	.772	.756	.790	.795
<i>Under t2 distributions</i>								
NS	.791	.702	.697	.658	.665	.655	.721	.694
Wilcoxon	.794	.757	.760	.730	.744	.741	.779	.768
Median	.690	.736	.694	.688	.716	.719	.657	.721
LMPDE	.804	.801	.801	.761	.765	.753	.760	.770
T2	.800	.804	.810	.784	.800	.798	.802	.813
Cauchy	.851	.782	.779	.742	.754	.750	.694	.750
KS	.799	.767	.778	.732	.746	.742	.783	.775
<i>Under Cauchy distributions</i>								
NS	.641	.573	.536	.508	.487	.475	.601	.520
Wilcoxon	.644	.664	.636	.622	.610	.609	.705	.643
Median	.615	.734	.678	.705	.715	.724	.656	.711
LMPDE	.657	.774	.762	.763	.758	.755	.758	.756
T2	.650	.765	.754	.760	.760	.764	.800	.781
Cauchy	.801	.805	.801	.807	.809	.813	.775	.810
KS	.735	.760	.750	.725	.716	.714	.800	.751
<i>Under Slash distributions</i>								
NS	.642	.592	.579	.560	.549	.544	.600	.592
Wilcoxon	.645	.677	.675	.669	.666	.670	.697	.707
Median	.608	.729	.682	.696	.701	.707	.631	.720
LMPDE	.658	.775	.777	.763	.749	.741	.632	.767
T2	.651	.770	.779	.784	.788	.794	.780	.815
Cauchy	.795	.799	.805	.797	.794	.797	.736	.807
KS	.729	.753	.759	.729	.721	.722	.758	.767

2. The Wilcoxon (W) test with scores $a_N(i) = i$. This test is optimal for data from the logistic distribution.

3. The locally most powerful rank test for the double exponential distribution (LMPDE) with scores $a_N(i) = 2 * Pr(B < i - 1) - 1$, where B is a binomial variable with parameters N and $.5$.

4. The asymptotically most powerful rank test (AMPRT) for the t_2 distribution (T2) with scores $a_N(i) = \sqrt{30}(i/(N+1) - .5)\sqrt{1 - 4(i/(N+1) - .5)^2}$.

5. The Cauchy scores test (C) (Capon 1961) with scores

$$a_N(i) = 2 \tan \pi(i/(N+1) - .5) / (1 + \tan^2 \pi(i/(N+1) - .5)).$$

6. The Kolmogorov-Smirnov test (KS), which is the maximum of the absolute difference between the empirical cdfs of the two samples.

The KS test was shown to be asymptotically optimal for data from a double exponential distribution (Capon 1965). Indeed, Rubin and Sethuraman (1965) showed that for the two-sample shift problem the KS test generally has the same asymptotic relative efficiency as the median test. See Nikitin (1995) for a comprehensive treatment of asymptotic efficiencies of nonparametric tests.

3. POWER SIMULATIONS

The small sample behavior of the six rank tests (NS, W, median, LMPDE, T2, C) and KS test were investigated. Power simulations were conducted for six balanced sample size settings: $n = m = 5, 9, 13, 25, 50, 100$, where the data were generated from one of the following family of distributions: (1) Normal, (2) Logistic, (3) Double exponential, (4) t_2 , (5) Cauchy, and (6) Slash (Morgenthaler and Tukey 1991). We used exact two-sided .05 critical regions, and randomized critical regions when the closest p value was more than .001 from the required critical value. The power estimates were based on 1,000,000 replications. The empirical estimates of the power against a shift, θ , for each of the six underlying densities are given in Table 1. The value, θ , of the shift parameter was set so that the power of the optimal test for each distribution was near .80.

Not surprisingly, the locally optimal tests (NS, W, LMPDE) had the highest power for all sample sizes for the distributions they were designed for. The AMPRT for data from a t_2 distribution (Gastwirth 1970) had the highest power for that distribution once sample sizes in each group were at least 9. On data from a double exponential distribution, however, the median test had substantially less power than the Wilcoxon for samples below or equal 25, and usually had less power than the KS procedure. Moreover, it had less power than the T2 for all sample sizes. Only when group sample sizes equaled 100 were the powers of median and T2 tests indistinguishable. These results indicate that, unlike other AMPRTs, the median test requires quite large samples before its asymptotic optimality properties are realized.

The results for the sample size $n = m = 25$ across all distributions were typical of the balanced setting so only they will be discussed. For normal data the median test had lower power (by at least .08) than W, T2, LNPDE, and KS procedures. The situation is similar for logistic data. For data from the heavier tailed Cauchy or slash distributions the median test did have more power than the NS and W (for sample sizes 9 and higher) but the LMPDE, T2, and KS tests were superior to it. Thus, in virtually all settings the LMPDE, T2, and KS had higher power than the median test, with T2 having the best overall performance. As the comparatively low power of the median test in small unbalanced samples was shown by Gastwirth and Wang (1987) we explored larger samples. The case $n = 75, m = 25$ was typical and the results in the last column of Table 1 indicate that the conclusions from the balanced setting apply to the unbalanced one. It should be noted that while the results in Table 1 are given for values of the shift parameter, θ , corresponding to approximately 80% power of the optimal tests, the same pattern generally holds for other choices of θ . An interesting issue, raised by a reviewer, is the behavior of the tests under asymmetric distributions. A small simulation study (available from authors) indicated that the W, T2, and KS procedures continue to generally have higher

power than the median test under a variety of asymmetric distributions.

There are two approaches to obtaining a single test when data come from one of a finite set of distributions. The maximum efficiency robust test (MERT) due to Gastwirth (1966) and the maximum of the standardized normal versions of the optimal statistics (MX) (Tarone 1981; Fleming and Harrington 1991). For comparison purposes the performance of these robust tests for data from the normal, logistic, double exponential, and Cauchy family was evaluated. They had slightly more power than T2 for normal data, however, for the other distributions considered the powers of these three tests were similar. If one was less concerned about very heavy tailed distributions, for example, one could restrict the possible underlying densities to the normal, logistic, and double exponential distributions, then the Wilcoxon test was nearly as power robust as the MERT and the MX.

4. CONCLUSION

Although the median test is easy to explain its comparatively low power for normal data without a substantial compensating gain on data from the heavier tailed distributions indicates that it should not be recommended for routine use. Of the test statistics examined here, the T2 test performed well and can be used with software allowing user to specify the scores (e.g., StatXact). If one is restricted to the standard tests in a typical software package, the KS procedure is usually superior to the median test across the set of models considered. If one felt that the data could not come from a Cauchy or slash distribution, the Wilcoxon test should be used.

This article has considered the usual two-sample problem where the distributions $F(x)$ and $G(y)$ may differ only in their locations. The levels of all rank tests, including the Wilcoxon (Wetherill 1960) and the median test (Pratt 1964) are somewhat affected when the scale parameters (σ_F, σ_G) of $F()$ and $G()$ differ. When the ratio σ_F/σ_G is noticeably different from 1, one should consider rank tests for the Behrens-Fisher problem (Hettmansperger and McKean 1998). In particular, Fligner and Policello (1981) modified the Wilcoxon test which should suffice for general use. In special situations, where differences in the scale parameters are quite large, the modification of the median test due to Fligner and Rust (1982) may be useful. For most purposes, however, the low power of the median test in conjunction with the existence of sound alternative procedures indicates that basic texts and computer packages could relegate it to a footnote in the future.

[Received June 1999. Revised February 2000.]

REFERENCES

- Anderson, T. W., and Finn, J. D. (1996), *The New Statistical Analysis of Data*, New York: Springer.
- Capon, J. (1961), "Asymptotic Efficiency of Certain Locally Most Powerful Rank Tests," *The Annals of Mathematical Statistics*, 32, 88-100.
- (1965), "On the Asymptotic Efficiency of the Kolmogorov-Smirnov Test," *Journal of the American Statistical Association*, 60, 843-853.

- Conover, W. J. (1999), *Practical Nonparametric Statistics* (3rd ed.), New York: Wiley.
- Conover, W. J., Wehmanen, O., and Ramsey F. L. (1978), "A Note on the Small-Sample Power Functions for Nonparametric Tests of Location in the Double Exponential Family," *Journal of the American Statistical Association*, 73, 188–190.
- CYTEL Software (1995), *StatXact-3, User Manual*, Cambridge, MA: author.
- Daniel, W. W. (1990), *Applied Nonparametric Statistics* (2nd ed.), Boston: PWS-Kent.
- Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.
- Fligner, M. A., and Policello, G. E. (1981), "Robust Rank Procedures for the Behrens–Fisher Problem," *Journal of the American Statistical Association*, 76, 162–168.
- Fligner, M. A., and Rust, S. W. (1982), "A Modification of Mood's Median Test for the Generalized Behrens–Fisher Problem," *Biometrika*, 69, 221–226.
- Gastwirth, J. L. (1966), "On Robust Procedures," *Journal of the American Statistical Association*, 61, 929–948.
- (1970), "On Robust Rank Tests," in *Nonparametric Techniques in Statistical Inference*, ed. M. L. Puri, London: Cambridge University Press.
- Gastwirth J. L., and Wang, J. (1987), "Nonparametric Tests in Small Unbalanced Samples: Application in Employment-Discrimination Cases," *The Canadian Journal of Statistics*, 15, 339–348.
- Gibbons, J. D., and Chakraborti, S. (1992), *Nonparametric Statistical Inference* (3rd ed.), New York: Dekker.
- Hajek, J., and Sidak, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.
- Hall, C. M., Louw, S. J., and Joubert, G. (1995), "Relative Efficacy of Hydrocortisone and Methylprednisolone in Acute Severe Asthma," *South African Medical Journal*, 85, 1153–1156.
- Hernberg, M., Turunen, J. P., von Boguslawsky, K., Muhonen, T., and Pyrhonen, S. (1998), "Prognostic Value of Biomarkers in Malignant Melanoma," *Melanoma Research*, 8, 283–291.
- Hettmansperger, T. P., and McKean, J. W. (1998), *Robust Nonparametric Statistical Methods*, London: Arnold.
- Hollander, M., and Wolfe, D. A. (1998), *Nonparametric Statistical Methods* (2nd ed.), New York: Wiley.
- Mood, A. M. (1954), "On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests," *The Annals of Mathematical Statistics*, 25, 514–522.
- Morgenthaler, S., and Tukey, J. W. (1991), *Configural Polysampling: A Route to Practical Robustness*, New York: Wiley.
- Nikitin, Y. (1995), *Asymptotic Efficiency of Nonparametric Tests*, New York: Cambridge University Press.
- Paternoster, D. M., Stella, A., Simioni, P., Girolami, A., and Plebani, M. (1996), "Fibronectin and Antithrombin as Markers of Pre-eclampsia in Pregnancy," *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 70, 33–39.
- Pratt, J.W. (1964), "Robustness of Some Procedures for the Two-Sample Location Problem," *Journal of the American Statistical Association*, 59, 665–680.
- Ramsey, F. L. (1971), "Small Sample Power Functions for Nonparametric Tests of Location in the Double Exponential Family," *Journal of the American Statistical Association*, 66, 149–151.
- Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: Wiley.
- Riegelman, R. K., and Hirsch, R. P. (1989), *Studying a Study and Testing a Test* (2nd ed.), Boston: Little, Brown, and Company.
- Rubin, H., and Sethuraman, J. (1965), "Probabilities of Moderate Deviations," *Sankhya*, 27, 325–346.
- Sall, J., Lehman, A., Saul, J. (1996), *JMP Start Statistics: A Guide to Statistical and Data Analysis Using JMP*, New York: Duxbury.
- SAS (1990), Cary, NC: SAS Institute.
- SPSS (1997), Chicago: SPSS Inc.
- Sokal, R. P., and Rohlf, F. J. (1995), *Biometry: The Principles and Practice of Statistics in Biological Research*, New York: W.H. Freeman & Co.
- Tarone, R. E. (1981), "On the Distribution of the Maximum of the Log-Rank Statistic and the Modified Wilcoxon Statistic," *Biometrics*, 37, 79–85.
- Wetherill, G. B. (1960), "The Wilcoxon Test and Non-null Hypotheses," *Journal of Royal Statistical Society, Ser. B*, 27, 402–418.
- Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics*, 1, 80–83.
- Zar, J. H. (1998), *Biostatistical Analysis*, New York: Prentice Hall.