Statistics and Its Interface Volume 2 (2009) 213–221

A new two stage adaptive nonparametric test for paired differences

Weiwen Miao* and Joseph Gastwirth†

This paper proposes a new adaptive procedure for analyzing paired data. The procedure uses a function of ordered absolute values of the differences to measure tail heaviness of the underlying distribution. The value of the measure is then used to choose an appropriate signed rank test. The new adaptive procedure is shown to preserve the size of the test at its nominal level for all continuous distributions and typically has nearly the same power as the best signed rank test for a wide range of distributions.

Keywords and phrases: adaptive paired test, signed rank test, tail-heaviness.

1. MOTIVATION AND INTRODUCTION

Paired data arises in a wide variety of applications (e.g. Sprent and Smeeton, 2001). In a recent security law case where both authors were statistical consultants to the defendant lawyers, a Wall Street firm was charged of favoring a few customers who "shared profits" with them. Since the "profit sharing" funds usually sold their shares through a different firm than the one they acquired them, the regulators need to estimated the profit. Unfortunately, two authorities, the SEC (Securities and Exchange Commission) and the NASD (National Association of Securities Dealers) used different methods to estimate the profits customers made. A natural question is whether the two estimates of profits are essentially the same. As a firm could be examined by each regulator and be accused of "profit sharing" by one regulator and exculpated by the other. Since the two different measurements are for the same stock, the data is naturally paired. Data of 157 IPOs (Initial Public Offering) during an 18-month period were made available to us. Our analysis shows that the differences have heavier tail than the normal distribution. It's better to use the t_2 or Cauchy scores test to detect the difference.

Paired data also arise in environmental studies where duplicate observations are taken. In 2003, the Michigan Department of Environmental Quality Remediation and Redevelopment Division conducted off-site samplings for the Detroit lead assessment project. Twenty-four soil samples were collected for lead analysis in vicinity of the former Great Lakes Smelting Company in Detroit, Michigan. These 24 samples are from 12 different locations near the plant with two samples from each location. Our interest is to test whether the two measurements from the same location are the same. Analysis shows that the differences are not normal. (The p-value of the Shapiro-Wilk test is 0.029.) Hence, the paired t-test is not appropriate for the analysis of this data.

Motivated by these two examples, we develop an adaptive procedure that uses a preliminary test to first analyze the tail-heaviness of the underlying distribution of the differences. That information is then used to choose an appropriate signed rank test to analyze the paired differences. For the sample of 157 estimated profits, our adaptive procedure chooses the t_2 scores test which yields a p-value of 0.009, about one-third of the one obtained from the paired t-test (p-value = 0.027). For the environmental data, the new adaptive procedure also chooses the t_2 scores test which yields a p-value of 0.038, showing that the two measurements are actually significantly different. For comparison purposes, the p-values for the t-test and the Wilcoxon test are 0.46 and 0.11, respectively. This example illustrates that in small sample sizes, the new adaptive procedure is able to detect a significant difference when both the t-test and the Wilcoxon test could not.

Typically, when the paired differences follow a normal distribution, the t-test is used; and when the normality assumption fails, the nonparametric Wilcoxon test is recommended. However, Freidlin, Miao and Gastwirth (2003) showed that this approach does not have high power when the paired differences are heavy-tailed. Randles and Hogg (1973) introduced an adaptive procedure that strictly maintains the size of the test and has high power when the distribution is heavy-tailed. Unfortunately, their procedure does not have high power for light to moderate tailed distributions. Freidlin et al. (2003) proposed an adaptive procedure (denoted by FMG) that uses the p-values of the Shapiro-Wilk normality test to choose an appropriate signed rank test to analyze the pairs. Their adaptive procedure has high power for a reasonable large class of moderate to heavy tailed distributions. However, their preliminary test is only asymptotically uncorrelated with the signed rank test used to analyze the paired sample, and the type I error for the procedure is slightly inflated. The FMG test was motivated by an environmental law case with sample size 16 and designed for

*Corresponding author.

†The research of Professor Gastwirth was supported in part by NSF grant SES-0317956.

relatively small samples. When the sample size is large, the Shapiro-Wilk and other tests of normality will detect minor deviations from normality. This may lead the FMG procedure to select a test that is good for a very heavy tailed distribution even if the differences are not that far from normal. The new method relying on a measure of "heavy-tailness" rather than the p-value of a preliminary test is applicable in large samples.

Mathematically, let $X_1, \ldots, X_n$ be i.i.d. *paired* differences from a continuous distribution F. The X_i are naturally symmetric about a center μ. We are interested in testing whether this center is equal to a *known* value μ_0. Without loss of generality, assume that $\mu_0 = 0$, i.e. we are interested in testing:

$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0.$$

It's well known that when the X_is follow a normal distribution, the optimal test is the t-test (Lehmann, 1986). However, if we *know* that X comes from a heavy-tailed distribution, e.g. t_2 or Cauchy, then the signed rank tests with t_2 (Gastwirth, 1970) or Cauchy scores (Capon, 1961) have higher power than the t-test. We propose an adaptive procedure (denoted by MG) that first uses functions of order statistics of $|X_i|$ to obtain the information about the tail-heaviness of the underlying distribution, and then use it to choose an appropriate signed rank test to analyze the pairs. The procedure strictly keeps the size of the test at the nominal level for all sample sizes, and has about the same power as the best signed rank test in a wide range of distributions, including the t family.

The paper is organized as follows: Section 2 introduces the preliminary test and its properties; Section 3 presents the percentiles of the preliminary test and how to use them to form the adaptive procedure. Power simulations for a variety of distributions are given in Section 4. The securities profit case data and the environmental data are analyzed in Section 5, using the new adaptive test and the results are compared with the paired t-test, Wilcoxon test and the FMG test.

2. PROPERTIES OF THE PRELIMINARY TESTS

2.1 The preliminary tests

We want an adaptive procedure to have high power over a wide range of distributions, from the light tailed uniform to the heavy-tailed Cauchy distributions. The tests used to analyze data at the second stage are signed rank tests. Those tests are based on the absolute ranks. It's well known (Lemma 8.3.11, Randles and Wolfe, 1979) that the ranks and the order statistics of the absolute values of the X_i are independent. We choose functions of the ordered $|X_i|$ as the preliminary test, to measure the tail-heaviness. Doing so, the preliminary test and the tests used to analyze data are independent, which guarantees that the overall size of the adaptive procedure is kept at the nominal level.

The most commonly used measure of spread is the sample standard deviation, s. However, some other statistics, such as the median absolute deviation from the median are more robust measures of the spread. Under the null hypothesis of symmetric about 0, the corresponding sample statistics are:

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \quad \text{and} \quad \widetilde{M} = \frac{median(|X_1|, \ldots, |X_n|)}{\Phi^{-1}(0.75)},$$

where $\Phi(x)$ is the c.d.f. of standard normal. Note that the constant $1/\Phi^{-1}(0.75)$ in $\tilde{M}$ is chosen to make $\tilde{M}$ a consistent estimator for the standard deviation when the underlying distribution is normal.

When the data are normally distributed, the ratio of two statistics, namely,

$$sM = \frac{\tilde{s}}{\widetilde{M}}$$

is near 1. They will be greater (or less) than 1.0 when data has heavier (or lighter) tails than normal. The statistic sM provides us indication of heaviness of the tails. We then use it as the preliminary test to choose the appropriate signed rank tests to analyze the paired data.

If one estimates the center by the sample median rather than the assumed 0, the statistic $\tilde{M}$ becomes the median absolute deviation from the median, which is asymptotically equivalent to the semi-interquartile range, a robust estimate of spread (Hall and Welsh, 1985).

2.2 Asymptotic properties of the preliminary test

The following theorem shows that the statistic sM is asymptotically normally distributed.

Theorem 2.1. *Let X be a symmetric continuous random variable with distribution function F. Let f and σ^2 be its density function and the variance. Let $F(0) = \frac{1}{2}$ and $\xi = F^{-1}(0.75)$, i.e. 0 and ξ are the median of X and $|X|$, respectively. Let*

$$\gamma_1 = \int_0^\infty x f(x)dx, \qquad \gamma_2 = \int_0^\infty x^2 f(x)dx,$$
$$\gamma_3 = \int_0^\infty x^3 f(x)dx, \qquad \gamma_4 = \int_0^\infty x^4 f(x)dx.$$

If the distribution F satisfies the conditions (A)–(C) stated in the Appendix, then statistic sM is asymptotically normal. In other words,

$$\sqrt{n}\Big[sM - \frac{\sigma\Phi^{-1}(0.75)}{\xi}\Big] \Rightarrow N(0, \sigma^2_{sM}),$$

Table 1. Asymptotic means and standard deviations of the statistic sM

Distribution	Mean	Standard Deviation
$U(-1,1)$	0.7788	0.5225
Normal	1	0.9276
Logistic	1.1136	1.2181
$0.95N(0,1)+0.05N(0,3^2)$	1.1388	1.5853
t_5	1.1983	1.8542
$0.9N(0,1)+0.1N(0,3^2)$	1.2406	1.7730
Double Exponential	1.3761	1.9388
$0.9N(0,1)+0.1N(0,5^2)$	1.6741	3.3517

where

$$\sigma_{sM}^2 = \Phi^{-2}(0.75) \times \left[\frac{2\gamma_4-\sigma^4}{4\xi^2\sigma^2} + \frac{\sigma^2}{16\xi^4 f^2(\xi)} + \frac{\int_{-\xi}^{\xi} x^2 f(x)dx - \frac{1}{2}\sigma^2}{2\xi^3 f(\xi)}\right].$$

The proof of the Theorem is given in Appendix.

Table 1 lists the means and standard deviations of the asymptotic distributions of the sM statistic for some commonly used symmetric distributions.

3. THE ADAPTIVE PROCEDURES

3.1 The signed rank tests

The signed rank tests used to analyze paired samples are expressed as:

$$S = \sum_{i=1}^{n} a(R_i^+)\delta(X_i), \quad \text{with } a(i) = J(\frac{i}{n+1})$$

where R_i^+ is the rank of $|X_i|$ among $|X_1|, |X_2|, \ldots, |X_n|$,

$$\delta(X_i) = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{otherwise} \end{cases},$$

and J is a score function appropriate for the data.

When the data comes from a light-tailed normal distributions or short-tailed uniform distributions, the normal scores test is known to have high power; when the tail-heaviness of the underlying distributions is somewhat medium, like logistic, double exponential or contaminated normal distributions, the Wilcoxon test is highly correlated with the maximin efficiency robust test (Gastwirth, 1966) and the Wilcoxon test should be used. Furthermore, if the data is heavy-tailed, e.g. from a t_2 or Cauchy distribution, then the appropriate signed rank test is the t_2 scores or the Cauchy scores test. We choose the following 4 score functions which include the extreme members of the t family of distributions:

- $J_1(u) = \Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)$, (Normal scores)
- $J_2(u) = u$, (Wilcoxon)
- $J_3(u) = \frac{3\sqrt{2}}{2}u\sqrt{1-u^2}$, ($t_2$ scores)
- $J_4(u) = \frac{2\tan(\frac{1}{2}\pi u)}{1+\tan^2(\frac{1}{2}\pi u)}$. (Cauchy scores)

3.2 Percentiles of the preliminary tests and adaptive procedures

In order to form the cut-off of the adaptive procedure, we looked at both the empirical and asymptotic percentiles for statistic sM for different distributions. Table 2 lists the empirical percentiles for statistic sM, for sample size 100. The distributions in the tables are arranged according to their tail heaviness, from the very heavy-tailed Cauchy to the light-tailed normal and uniform distributions. In the tables, C. Norm(5) (or C. Norm(3)) represents Tukey's 10% contaminated normal $0.9N(0,1)+0.1N(0,5^2)$ (or $0.9N(0,1)+0.1N(0,3^2)$). The results are based on 10^5 simulations.

The general idea of the adaptive procedure is to use the sampling distribution of the preliminary test on a variety of underlying distributions to create cut-off values to classify the sample as light, medium, heavy or very heavy tailed. Then one chooses the signed rank test that *is known* to have high power for a distribution with tails of that type. We start with the heavy tailed Cauchy distribution, then move down the tail-heaviness to the light-tailed normal distribution, and finally to the short-tailed uniform distribution. Table 2 shows that more than 95% of Cauchy samples have sM values bigger than 2.7 (5^{th} percentile), and more than 80% of t_2 samples have $sM < 2.7$. Thus 2.7 is chosen as the cut-off value of sM between the Cauchy scores test and the t_2 scores test. In other words, if $sM \geq 2.7$, the Cauchy scores test is used. If sM is slightly less than 2.7, the t_2 scores test should be used. Furthermore, Table 2 indicates that more than 97.5% t_2 samples, more than 90% contaminated normal samples ($0.9N(0,1)+0.1N(0,5^2)$) and more than 80% double exponential samples have sM values larger than 1.2 (2.5^{st}, 10^{th} and 20^{th} percentiles, respectively), while about 80% of logistic samples have sM values smaller than 1.2 (80^{th} empirical percentile). Therefore we choose cut-off 1.2 between the t_2 scores test and the Wilcoxon test. This means that if sM is between 1.2 and 2.7, the t_2 scores test is used. Similarly, we choose 1.02 as the cut-off between the Wilcoxon test and the normal scores test. In other words, if sM is between 1.02 and 1.2, the Wilcoxon test is used, and if $sM < 1.02$, the normal scores test is used. The cut off 1.02 was chosen because for sample size 100, all the uniform samples have sM values smaller than 1, but for sample size 25, about 95% of uniform samples have sM values less than 1.02. Furthermore, the Wilcoxon test has high power on both normal and logistic distributions. Consequently, the new adaptive procedure, denoted by MG, is the following:

the MG Adaptive Test:

- $sM \geq 2.7$, use the Cauchy scores test;
- $2.7 > sM \geq 1.2$, use the t_2 scores test;
- $1.2 > sM \geq 1.02$, use the Wilcoxon test;
- $sM < 1.02$, use the Normal scores test.

Table 2. Percentiles of the statistic sM on a Variety of Distributions (n=100)

	Cauchy	t_2	C. Norm(5)	D. Exp.	C. Norm(3)	Logistic	Normal	Uniform
1%	2.0944	1.1450	1.0218	1.0165	0.9089	0.8827	0.8237	0.6834
2.5%	2.3968	1.2135	1.0894	1.0587	0.9439	0.9106	0.8458	0.6951
5%	2.7533	1.2804	1.1563	1.0974	0.9789	0.9365	0.8661	0.7062
10%	3.2813	1.3696	1.2426	1.1461	1.0219	0.9688	0.8903	0.7197
20%	4.1983	1.4984	1.3588	1.2110	1.0803	1.0107	0.9230	0.7377
25%	4.6705	1.5564	1.4065	1.2372	1.1053	1.0278	0.9360	0.7450
30%	5.1899	1.6129	1.4504	1.2614	1.1282	1.0434	0.9481	0.7517
40%	6.3855	1.7309	1.5335	1.3079	1.1712	1.0736	0.9712	0.7646
50%	7.9517	1.8615	1.6167	1.3540	1.2140	1.1020	0.9934	0.7772
60%	10.219	2.0171	1.7036	1.4026	1.2607	1.1330	1.0169	0.7907
70%	13.914	2.2263	1.8008	1.4583	1.3133	1.1676	1.0439	0.8057
75%	16.798	2.3644	1.8554	1.4911	1.3439	1.1879	1.0591	0.8145
80%	21.114	2.5508	1.9191	1.5284	1.3787	1.2106	1.0768	0.8246
90%	43.597	3.2484	2.0970	1.6356	1.4768	1.2753	1.1265	0.8531
95%	87.661	4.2296	2.2534	1.7303	1.5655	1.3336	1.1703	0.8790
97.5%	173.56	5.6334	2.4012	1.8198	1.6458	1.3874	1.2116	0.9033
99%	436.05	8.5566	2.5813	1.9310	1.7480	1.4548	1.2657	0.9340

4. SIZE AND POWER OF THE NEW ADAPTIVE PROCEDURE

4.1 Size

The preliminary test sM is a function of ordered absolute values of the differences, and the tests used to analyze data are the signed rank tests based on absolute ranks. For continuous distributions, the signed ranks and the absolute order statistics are independent (Lemma 8.3.11 of Randles and Wolfe, 1979). Hence the adaptive procedure MG maintains the size at the nominal level (Lemma 11.6.3 of Randles and Wolfe, 1979). Our simulation results confirm that the adaptive test MG keeps the size at its nominal level for sample sizes as small as 10, when the asymptotic normal distribution is used to calculate the p-value for the signed rank tests.

Freidlin, Miao and Gastwirth (2003) proposed an adaptive procedure (denoted by FMG test) for paired data, using the p-value of the Shapiro-Wilk test to choose appropriate signed rank tests. They proved that the Shapiro-Wilk test is asymptotically uncorrelated with the signed rank test. Simulations showed that their method had nearly nominal level in sample sizes 10-25. The adaptive test MG strictly maintains the size at its nominal level and is a slightly improvement over the FMG test in this regard.

In contrast to the situation under the null hypothesis, under the alternative hypothesis, the sM statistic and the signed rank tests are somewhat correlated. Simulation studies, suggested by a referee, show that the magnitude of the correlations are usually less than 0.2 at meaningful alternatives. Furthermore, under the alternatives, the sM statistic and the p-value of the chosen test are also related. Simulations show that for light-tailed (heavy-tailed) underlying distributions, smaller values of the sM statistic are associated with smaller (larger) p-values. This is consistent with theory, as the expected values of the sM statistic are small on light-tailed distributions and small values of the sM lead to the selection of either the Normal scores test or the Wilcoxon test at the second stage. Those two tests are known to have high power for light-tailed normal and logistic distributions.

4.2 Power

The following graphs are the size and power comparisons for 3 different procedures: The adaptive procedure proposed by Freidlin, Miao and Gastwirth (2003), the optimal signed rank tests for the particular distribution (denoted by Best SRT) and the new MG adaptive procedure. For contaminated normal distributions $(0.9N(0,1) + 0.1N(0,3^2))$, the Wilcoxon test is considered to be highly efficient (Hodges and Lehmann, 1961), and for uniform distributions, the normal score test is taken to be the best signed rank test. The graphs are based on 10^4 simulations for samples of size 100.

Figures 1 and 2 indicate that the two adaptive procedures: FMG and MG, have about the same power as the best signed rank tests for normal and logistic distributions. It's interesting to note that for double exponential distributions (Figure 3) the two adaptive procedures MG and FMG have slightly higher power than the the best signed rank test (the sign test). For heavy-tailed t_2 and contaminated normal distributions $(0.9N(0,1) + 0.1N(0,3^2))$, the MG adaptive test has about the same power as the best signed rank test; their powers are slightly higher than the FMG procedure. For sample size 100, the power of the MG procedure is about 2-6% higher than the FMG test for contaminated normal distributions. But when the sample size increases to 200, the power of the MG test is significantly higher than the FMG test (about 10–15% higher). (Also see Table 3 below.) This occurs because in large samples, the Shapiro-Wilk test will yield a low p-value even when the

Table 3. Power comparison for the FMG, MG and the Best SRT tests

Distribution	Sample Size	Alternative	MG	FMG	Best SRT
Normal	25	0.627	0.8268	0.8377	0.8460
	100	0.31	0.8565	0.8523	0.8700
	200	0.215	0.8415	0.8324	0.8497
Logistic	25	1.11	0.8288	0.8487	0.8479
	100	0.545	0.8554	0.8629	0.8691
	200	0.374	0.8478	0.8379	0.8566
Double Exponential	25	0.805	0.8375	0.8577	0.8496
	100	0.365	0.8733	0.8747	0.8558
	200	0.25	0.8700	0.8652	0.8478
t_2	25	0.878	0.8134	0.829	0.8497
	100	0.4	0.8432	0.8145	0.8524
	200	0.275	0.8369	0.7997	0.8427
$0.9N(0,1) + 0.1N(0,3^2)$	25	0.735	0.8374	0.8563	0.8512
	100	0.35	0.8379	0.7703	0.8526
	200	0.245	0.8316	0.6953	0.8557

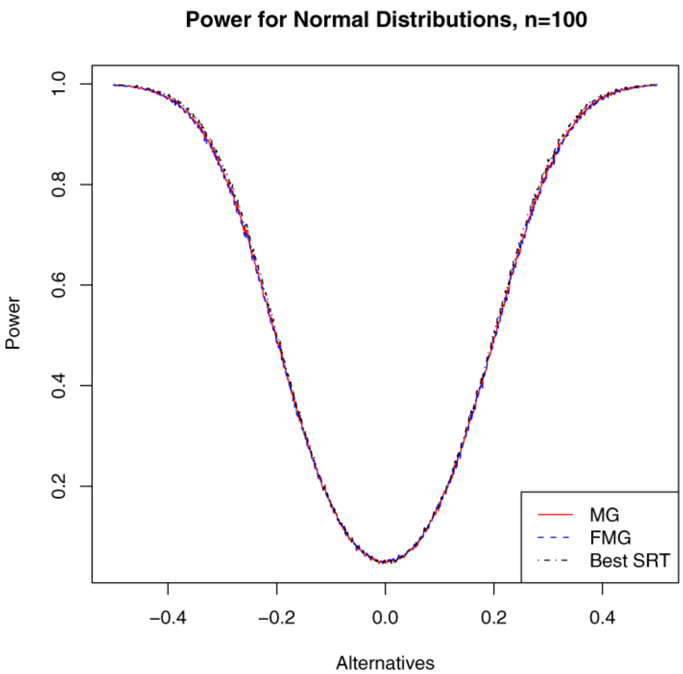

Figure 1. Power comparison for normal distributions.

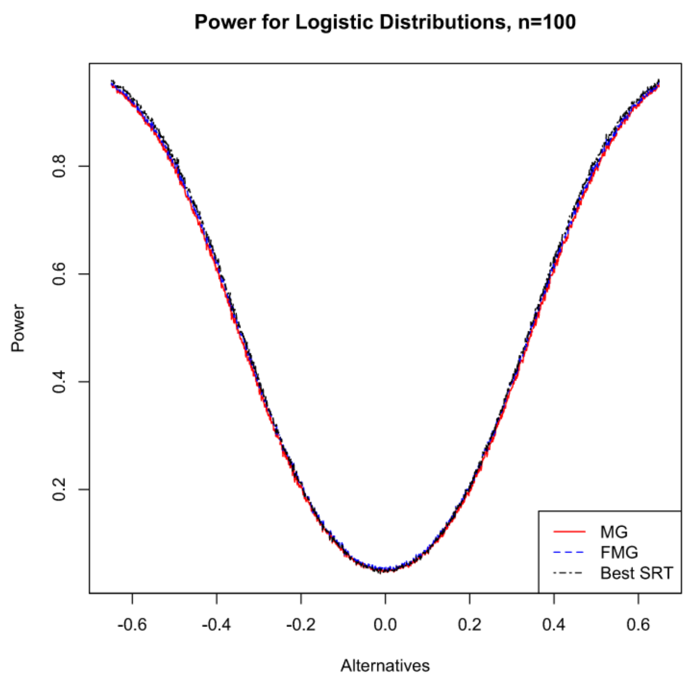

Figure 2. Power comparison for logistic distributions.

data are nearly normal, e.g. contaminated normal, and the FMG procedure will choose the t_2 or Cauchy scores test too often on contaminated normal distributions. For uniform distributions, again the MG adaptive test has about the same power as the normal scores test, but the FMG adaptive procedure has significantly lower power compared to the other two procedures. This is not surprising as the normal scores test, which is known to have high power on uniform distributions, is not one of the second stage choices in the FMG test.

The following table presents the powers of the FMG, the MG and the best signed rank test for sample sizes at 200, for Normal, Logistic, Double exponential, t_2 and contaminated Normal distributions. The powers of the tests for samples of size 25 and 100 are also included. The purpose of the table is to illuminate the roles of the sample sizes in determining the properties of the adaptive test. The alternatives are chosen so that the best signed rank test has about 85% power. The results are based on 10^4 simulations. The table clearly shows that for t_2 and contaminated normal distributions, when the sample size gets larger, the new adaptive test MG is better than the FMG test.

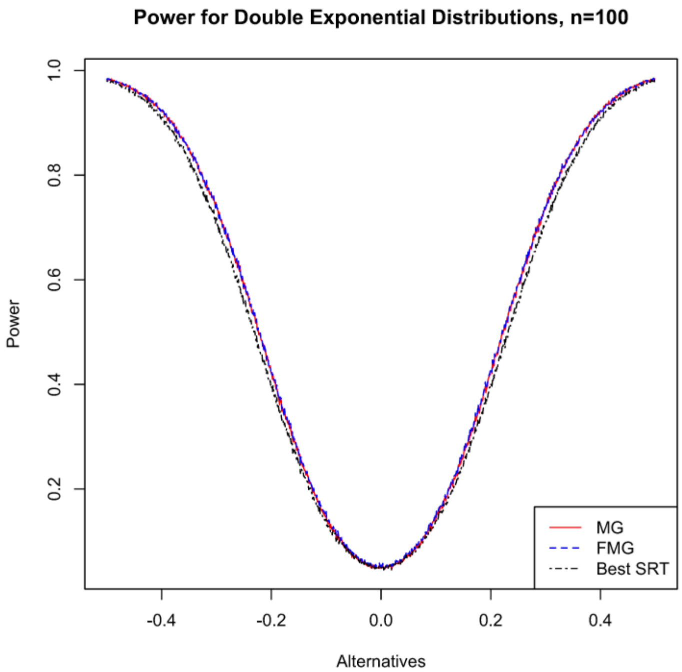

Figure 3. Power comparison for double exponential distributions.

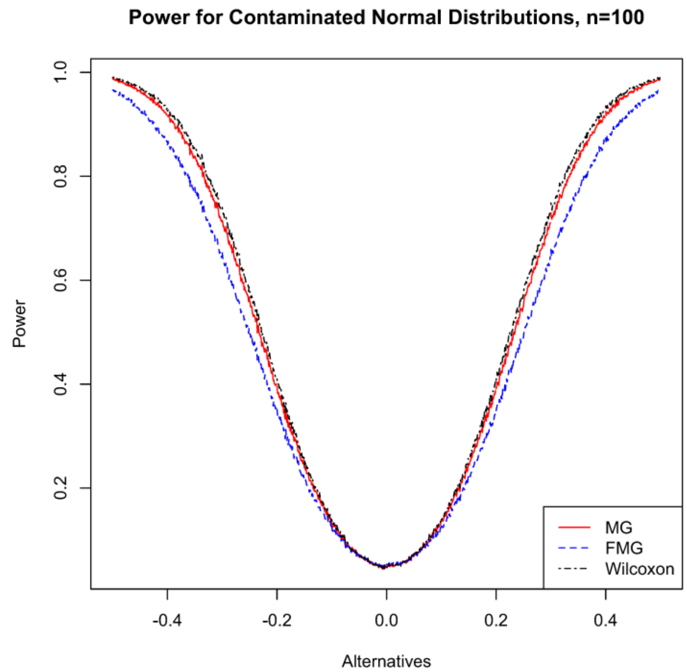

Figure 5. Power comparison for contaminated normal distributions ($0.9N(0,1) + 0.1N(0,3^2)$).

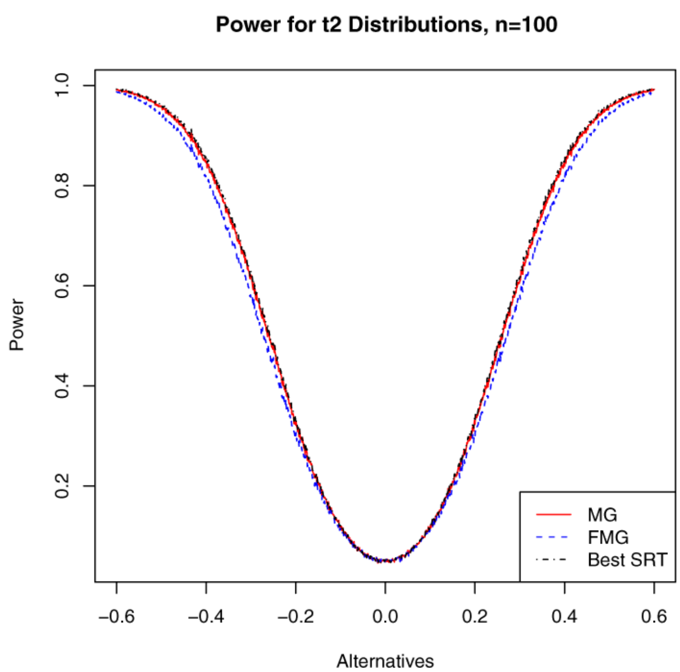

Figure 4. Power comparison for t_2 distributions.

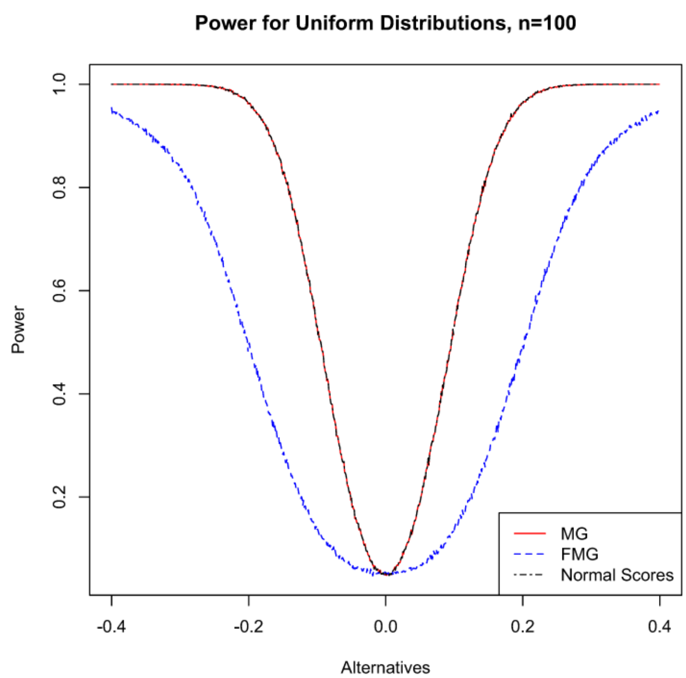

Figure 6. Power comparison for uniform distributions.

Simulations were also performed on the Cauchy distribution for samples of size 25, 100 and 200. The powers of the FMG, the MG and the Cauchy scores tests are about the same.

The main reason that the adaptive procedure MG is powerful is that the preliminary test chooses either the best or the second best signed rank test most of the time. Simulations show that for normal distributions, our adaptive test chooses the normal scores test about 62% of the time, the Wilcoxon test about 35% of the time, and only chooses the

t_2 scores test about 3% of the time. For samples from a t_2 distributions, the MG procedure chooses the t_2 scores test about 82% of the time, the Cauchy scores test about 14% of time, and the Wilcoxon test about 4% of time.

These power simulations indicate that the adaptive procedure MG keeps the size at the nominal level and generally has the same power as the best signed rank test for a variety of symmetric distributions. This procedure is recommended for use in practice.

5. EXAMPLES

5.1 The securities hypothetical profit data

This data concerns the hypothetical profit estimates for 157 IPOs during an 18-month period. The p-value of paired t-test is 0.027, indicating that the two measurements from two different regulators are statistically different. But the data is clearly not normal (the p-value of the Shapiro-Wilk test is $1.4 * 10^{-15}$). Applying the adaptive procedure, the preliminary test statistic sM = 2.44. Accordingly, the MG procedure chooses the t_2 scores test. The two-sided p-value for the t_2 scores test is 0.009, noticeably smaller than 0.027, the p-value of the paired test. In this example, the FMG adaptive test chooses Cauchy scores test, which yields a p-value of 0.064.

5.2 The environmental data

This data consists of 24 soil samples for lead analysis, 2 from each of the 12 different locations in the Detroit area. We are interested in whether the two measurements from the same location are the same. The p-value of the Shapiro-Wilk test is 0.029, so the differences are not normally distributed. Applying the new adaptive procedure, the preliminary test statistic sM is 2.07. So, the new adaptive test chooses the t_2 scores test, which yields a p-value of 0.0377, indicating that the two measurements are not the same.

In contrast to the above analysis, the p-value for the paired t-test is 0.46, which is not even close to statistical significant. In this example, the p-value of the Wilcoxon test is 0.11, much smaller than the 0.46, but still not significant at the commonly accepted 5% level. The FMG procedure chooses the Wilcoxon test (based on the 0.029 p-value of the Shapiro-Wilk test) and does not detect the significant difference, but performs better than the paired t-test.

This example shows that for small heavy-tailed non-normal samples, the new adaptive procedure MG is able to detect the significant difference while the paired t-test, the Wilcoxon test and the FMG adaptive test cannot.

6. FINAL REMARKS

In addition to the MG adaptive procedure, we also explored the adaptive test based on the ratio of $\tilde{s}/\tilde{J}$, where $\tilde{J} = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i|$, the mean absolute deviation from the *known* center 0. Combing the Central Limit Theorem and the delta-method, it can be shown that $\tilde{s}/\tilde{J}$ is also asymptotical normal. However, the asymptotical means and coefficient of variations of the $\tilde{s}/\tilde{J}$ are less spread out than the $sM = \tilde{s}/\tilde{M}$ for most commonly occurring symmetric distributions. This means that the statistic $\tilde{s}/\tilde{M}$ is better to distinguish different distributions. Power simulation also confirmed that the adaptive procedure based on $\tilde{s}/\tilde{M}$ has slightly higher power than the procedure based on the preliminary test $\tilde{s}/\tilde{J}$.

Studies showed that adaptive procedures for one-sample and two-sample problems work well over a wide range of distributions (Hogg 1974; Hogg, Fisher and Randles, 1975 and Hill, Padmanabhan and Puri, 1988). This article proposes an adaptive procedure for paired data. This adaptive test, MG, keeps its size at the nominal level and has nearly the same power as the best signed rank test for a wide range of distributions, including the family of t distributions. It can be recommended to use in practice.

ACKNOWLEDGEMENT

The authors would like to thank reviewers for several helpful suggestions.

APPENDIX

The following are the conditions (A)–(C) in Theorem 2.1:

(A) $F^{-2}(x)$ is continuous on the open unit interval $(0,1)$ and satisfies a first order Lipschitz condition on every interval bounded away from 0 and 1; $(F^{-2})'(x)$ exists and is continuous except on a set of Jordan content 0;

(B) $f(\xi) \neq 0$ and the Riemann integral $\int_0^1 (F^{-2})'(\frac{u+1}{2})[u(1-u)]^{\frac{1}{2}}du$ converges absolutely;

(C) For $i = 1, 2$ there exist $\delta_i \in (0,1)$ such that $\forall k_i > 0$ $\exists M_i < \infty$ such that

(C1) if $0 < u_1, u_2 < \delta_1$ and $k_1^{-1} < u_1/u_2 < k_1$, then

$$M_1^{-1} < \frac{F^{-1}(\frac{u_1+1}{2})(F^{-1})'(\frac{u_1+1}{2})}{F^{-1}(\frac{u_2+1}{2})(F^{-1})'(\frac{u_2+1}{2})} < M_1;$$

(C2) if $1-\delta_2 < u_1, u_2 < 1$ and $k_2^{-1} < (1-u_1)/(1-u_2) < k_2$, then

$$M_2^{-1} < \frac{F^{-1}(\frac{u_1+1}{2})(F^{-1})'(\frac{u_1+1}{2})}{F^{-1}(\frac{u_2+1}{2})(F^{-1})'(\frac{u_2+1}{2})} < M_2.$$

Proof of Theorem 2.1. As X is symmetric about 0,

$$f_{|X|}(x) = 2f(x), \qquad F_{|X|}(x) = 2F(x) - 1, \quad \text{and}$$
$$F_{|X|}^{-1}(u) = F^{-1}(\frac{u+1}{2}).$$

It can be shown that

$$E(\tilde{s}^2) = \sigma^2, \qquad E(\tilde{M}) = \frac{\xi}{\Phi^{-1}(0.75)},$$
$$Var(\tilde{s}^2) = \frac{1}{n}[2\gamma_4 - \sigma^4],$$
$$Var(\tilde{M}) = \frac{1}{16nf^2(\xi)\big[\Phi^{-1}(0.75)\big]^2}.$$

Let

$$M = median(|X_1|, \ldots, |X_n|), \quad I_i = \begin{cases} 1 & \text{if } |X_i| < \xi \\ 0 & \text{otherwise.} \end{cases}$$

Consider the interval (ξ, M). Note that this interval contains $\frac{n}{2} - \sum I_i$ observations as M is the $\frac{n+1}{2}$th observation and $\sum_{i=1}^n I_i$ is the # of observations less than ξ. Using the Bahadur representation, $f_{|X|}(\xi)(M-\xi) = 2f(\xi)(M-\xi) \approx P\big(|X| \in (\xi, M)\big)$. Consequently,

$$2nf(\xi)(M-\xi) \sim \frac{n}{2} - \sum I_i = -\sum_{i=1}^n (I_i - \frac{1}{2}).$$

This implies that

$$M - \xi \sim -\frac{1}{2nf(\xi)} \sum_{i=1}^n (I_i - \frac{1}{2}).$$

Hence,

$$\begin{aligned} Cov(\tilde{s}, M) &= Cov\Big(\frac{1}{n}\sum_{i=1}^n X_i^2, -\frac{1}{2nf(\xi)}\sum_{i=1}^n (I_i - \frac{1}{2})\Big) \\ &= -\frac{n}{2n^2 f(\xi)} Cov(X_1^2, I_1 - \frac{1}{2}) \\ &= -\frac{1}{2nf(\xi)}\big[E(X^2 I_1) - E(X_1^2)E(I_1)\big] \\ &= -\frac{1}{2nf(\xi)}\Big[\int_{-\xi}^{\xi} x^2 f(x)dx - \frac{1}{2}\sigma^2\Big] \end{aligned}$$

and

$$Cov(\tilde{s}, \tilde{M}) = -\frac{1}{2nf(\xi)\Phi^{-1}(0.75)}\Big[\int_{-\xi}^{\xi} x^2 f(x)dx - \frac{1}{2}\sigma^2\Big].$$

Let $|X|_{(i)}$ be the order statistics of $|X_i|$. Since both $\tilde{s}^2$ and $\tilde{M}$ are L-statistics, they can be represented as:

$$\frac{1}{n}\sum_{i=1}^n c_i h\Big(|X|_{(i)}\Big) + a|X|_{[\frac{n+1}{2}]}$$

with appropriate function $h(\cdot)$ and constants c_i and a, where $[\frac{n+1}{2}]$ is the largest integer smaller than $\frac{n+1}{2}$. The function h and constants c_i and a corresponding to the statistics $\tilde{s}^2$ and $\tilde{M}$ are:

(i) For $\tilde{s}^2$, $c = 1$, $h(x) = x^2$, $a = 0$;
(ii) For $\tilde{M}$, $c_i = 0$, $h(x) = x^2$, $a = \Phi^{-1}(0.75)$.

The joint asymptotic normality of $(\tilde{s}^2, \tilde{M})^t$ is obtained by applying Corollary 3 in Chernoff et al. (1967). It can be checked that the conditions (A)–(C) listed in this Theorem are equivalent to the assumptions A^*, B^{**} and E in Chernoff et al. (1967) with their function $H(u) = F^{-2}(\frac{u+1}{2})$. According to Remark 9 of Chernoff et al. (1967), asymptotically $(\tilde{s}^2, \tilde{M})^t$ is normal with mean $(\sigma^2, \frac{\xi}{\Phi^{-1}(0.75)})^t$.

The Taylor expansion of $sM = \frac{\sqrt{\tilde{s}^2}}{\tilde{M}}$ at $(\sigma^2, \frac{\xi}{\Phi^{-1}(0.75)})^t$ is:

$$\begin{aligned} sM = \frac{\sqrt{\tilde{s}^2}}{\tilde{M}} &= \frac{\sqrt{\sigma^2}}{\xi/\Phi^{-1}(0.75)} + \frac{1}{2\sigma\xi/\Phi^{-1}(0.75)}(\tilde{s}^2 - \sigma^2) \\ &\quad - \frac{\sqrt{\sigma^2}}{\xi^2/\Phi^{-2}(0.75)}(\tilde{M} - \xi) + 0_p(n^{-\frac{1}{2}}) \\ &= \frac{\sigma\Phi^{-1}(0.75)}{\xi} + \frac{\Phi^{-1}(0.75)}{2\sigma\xi}(\tilde{s}^2 - \sigma^2) \\ &\quad - \frac{\sigma\Phi^{-2}(0.75)}{\xi^2}(\tilde{M} - \xi) + 0_p(n^{-\frac{1}{2}}). \end{aligned}$$

Consequently,

$$\sqrt{n}\big[sM - \frac{\sigma\Phi^{-1}(0.75)}{\xi}\big] \Rightarrow N(0, \sigma_{sM}^2)$$

with

$$\begin{aligned} \sigma_{sM}^2 &= \frac{n\Phi^{-2}(0.75)}{4\xi^2\sigma^2}Var(\tilde{s}^2) + \frac{n\sigma^2\Phi^{-4}(0.75)}{\xi^4}Var(\tilde{M}) \\ &\quad - \frac{n\Phi^{-3}(0.75)}{\xi^3}Cov(\tilde{s}^2, \tilde{M}) \\ &= \Phi^{-2}(0.75)\big[\frac{2\gamma_4 - \sigma^4}{4\xi^2\sigma^2} \\ &\quad + \frac{\sigma^2}{16\xi^4 f^2(\xi)} + \frac{\int_{-\xi}^{\xi} x^2 f(x)dx - \frac{1}{2}\sigma^2}{2\xi^3 f(\xi)}\big]. \end{aligned}$$

□

Received 11 November 2008

REFERENCES

[1] Capon, J. (1961). Asymptotic Efficiency of Certain Locally Most Powerful Rank Tests. *The Annals of Mathematical Statistics* **32** 88–100. MR0120739

[2] Chernoff, H., Gastwirth, J. L., and Johns Jr., M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *The Annals of Mathematical Statistics* **38** 352–372. MR0203874

[3] D'Agostino, R. B. and Cureton, E. E. (1973). A Class of Simple Linear Estimators of the Standard Deviation of the Normal Distribution. *Journal of American Statistical Association* **68** 207–210.

[4] Weston Solutions of Michigan, INC. (2004). Phase I Summary Report for Detroit Lead Assessment Project, Great Lakes Smelting – 1640 East Euclid Street, Detroit, Wayne County, Michigan.

[5] FREIDLIN, B., MIAO, W., and GASTWIRTH, J. L. (2003). On the Use of the Shapiro-Wilk Test in Two-Stage Adaptive Inference for Paried Data from Moderate to Very Heavy Tailed Distributions. *Biometrical Journal* **45** 887–900. MR2012347

[6] GASTWIRTH, J. L. (1966). On Robust Procedures. *Journal of the American Statistical Association* **61** 929–948. MR0205397

[7] GASTWIRTH, J. L. (1970). On Robust Rank Tests. *Nonparametric Techniques in Statistical Inference*. Cambridge University Press, Cambridge. MR0277051

[8] HALL, P. and WELSH, H. (1985). Limit Theorems for the Median Deviation. *Ann. Inst. Statist. Math.* **37** 27–36. MR0790373

[9] HILL, N. J., PADMANABHAN, A. R., and PURI, M. L. (1988). Adaptive Nonparametric Procedures and Applications. *Applied Statistics* **37** 205–218. MR0954744

[10] HODGES, J. L. and LEHMANN, E. L. (1961). Comparison of the Normal Scores and Wilcoxon Test. *Proc. Fourth Berkeley Sympo. Math. Stat. Prob.* **1** 307–317. MR0132645

[11] HOGG, R. V. (1972). More Light on the Kurtosis and Related Statistics. *Journal of the American Statistical Association* **67** 422–424.

[12] HOGG, R. V. (1974). Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory. *Journal of the American Statistical Association* **69** 909–923. MR0461779

[13] HOGG, R. V., FISHER, D. M., and RANDLES, R. H. (1975). A Two-Sample Adaptive Distribution-Free Test. *Journal of the American Statistical Association* **70** 656–661.

[14] LEHMANN, E. L. (1986). *Testing Statistical Hypothesis*. Chapman & Hall, New York.

[15] RANDLES, R. H. and HOGG, R. V. (1973). Adaptive Distribution Free Test. *Communications in Statistics* **2** 337–356. MR0331628

[16] RANDLES, R. H. and WOLFE, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Krieger, Florida. MR0547836

[17] SPRENT, P. and SMEETON, N. C. (2001). *Applied Nonparametric Statistical Methods*. Chapman & Hall. MR2297847

Weiwen Miao
Department of Mathematics
Haverford College
Haverford, PA 19041.
E-mail address: wmiao@haverford.edu

Joseph Gastwirth
Department of Statistics
George Washington University
Washington, DC, 20052.
E-mail address: jlgast@gwu.edu