

# On the Use of the Shapiro-Wilk Test in Two-Stage Adaptive Inference for Paired Data from Moderate to Very Heavy Tailed Distributions

B. FREIDLIN<sup>1,\*</sup>, W. MIAO<sup>2</sup>, and J. L. GASTWIRTH<sup>3</sup>

<sup>1</sup> Biometric Research Branch, National Cancer Institute, Bethesda, Maryland, 20892 USA

<sup>2</sup> Macalester College, St. Paul, Minnesota, USA

<sup>3</sup> George Washington University, Washington, DC, USA

## Abstract

Paired data arises in a wide variety of applications where often the underlying distribution of the paired differences is unknown. When the differences are normally distributed, the  $t$ -test is optimum. On the other hand, if the differences are not normal, the  $t$ -test can have substantially less power than the appropriate optimum test, which depends on the unknown distribution. In textbooks, when the normality of the differences is questionable, typically the non-parametric Wilcoxon signed rank test is suggested. An adaptive procedure that uses the Shapiro-Wilk test of normality to decide whether to use the  $t$ -test or the Wilcoxon signed rank test has been employed in several studies. Faced with data from heavy tails, the U.S. Environmental Protection Agency (EPA) introduced another approach: it applies both the sign and  $t$ -tests to the paired differences, the alternative hypothesis is accepted if either test is significant. This paper investigates the statistical properties of a currently used adaptive test, the EPA's method and suggests an alternative technique. The new procedure is easy to use and generally has higher empirical power, especially when the differences are heavy-tailed, than currently used methods.

*Key words:* Adaptive tests; Paired data; Power robustness; Shapiro-Wilk test.

## 1. Introduction

In many applications paired data control for other relevant variables by using them in the matching process. Basic textbooks analyze such data by applying the paired  $t$  or the Wilcoxon signed rank tests to the differences. The  $t$ -test is optimal when the differences are normally distributed while the Wilcoxon test is a distribution-free alternative often recommended when the differences may not be normal. A natural adaptive procedure would test for normality and use a distribution-free method when normality is rejected. One such procedure is to first apply the Shapiro-Wilk (SW) test (SHAPIRO and WILK, 1965) to the differences: if normality is accepted, the  $t$ -test is used; otherwise the Wilcoxon signed rank test is used. This approach, denoted by *STW* hereafter, is commonly used in the medical literature

\* Corresponding author: Boris Freidlin freidlinb@ctep.nci.nih.gov

(LEACH et al., 1994; WINDUS et al., 1997) although sometimes a different test of normality (LILLIEFORS, 1967) is employed at the first stage. The *STW* procedure is also implemented and illustrated in SAS JMP (SALL et al. 1996).

Differences having long-tailed distributions arise in studies of the environmental equivalence of fuels. To protect against heavy tailed non-normality, the EPA requires that both the sign and *t*-tests yield non-significant results when applied to the differences in order for the null hypothesis to be accepted. After promulgating this method, the EPA deviated from it in analyzing data comparing the pollution levels of an alternative fuel to regular gasoline. The auto manufacturers were concerned that a new fuel with higher emission levels might lead to their cars not being in compliance with the environmental standards and successfully sued the agency (GASTWIRTH, 1988; FINKELSTEIN and LEVIN, 2001).

The use of the *SW* test as the preliminary one may be useful in other settings for which adaptive methods have been developed. For example, measures of skewness or tailweight calculated from a sample have been used to select the second stage test in one and *k*-sample problems, by HOGG (1974); RANGLES and WOLFE (1979); RUBERG (1986); HILL, PADMANABHAN and PURI (1988); O'GORMAN (1997); BÜNING and KÖSSLER (1998). An alternative use of the adaptive approach occurs in multi-stage experiments. Here an interim analysis is carried out in order to modify the test or design used at a later stage of the experiment. These methods are discussed in BAUER and KOHNE (1994), LANG, AUTERITH and BAUER (2000), NEUHAUSER (2001) and WASSMER (2000).

Because the Wilcoxon test is nearly as powerful as the *t*-test on data from the normal distribution, there is little need to use a preliminary test of normality before employing it. The Wilcoxon signed rank test, however, does not have high power when the differences are heavy-tailed. We propose an alternative two-stage adaptive procedure that takes advantage of the information about the heaviness of the tails in the selection of a nonparametric-paired test. As a low *p*-value of the *SW* test indicates that the distribution is "far" from normality, the *p*-value of the *SW* test is used to select the second stage test. This paper examines both the theoretical and small sample properties of the new adaptive method and compares it to the commonly used *STW* method and the EPA's technique. The proposed test has greater power robustness than either of the other two tests.

Section 2 presents the requisite background information as well as our new alternative procedure. The large sample theory of the *STW* and the new tests is summarized in Section 3. Simulation results are presented in section 4. Several real world data sets are reanalyzed in Section 5.

## 2. Description of the procedures

Suppose random variables *Z* and *X* describe measurements in two groups and have a joint distribution function  $F(z, x)$ . If *Z* and *X* are exchangeable, that is

$F(z, x) = F(x, z)$ , (e.g., group membership is assigned independently and at random) the paired differences  $Y = Z - X$  have a symmetric distribution (RANGLES and WOLFE 1979, P. 58). Let  $Y_1, Y_2, \dots, Y_n$  be the paired differences with mean  $\Delta$ . We are testing  $H_0: \Delta = 0$  vs.  $H_a: \Delta \neq 0$ . The  $p$ -value of the  $SW$  test of normality on  $Y$  is utilized to select the non-parametric test to be used to analyze the differences. This differs from the ordinary application of the  $SW$  test to check normality (e.g. GAN and KOEHLER, 1990; RAMSEY and RAMSEY, 1990). We examine the power characteristics of the  $SW$  test for data from several alternative symmetric distributions in order to determine how it should be used to select the second stage test. Figure 1 presents graphs of the power of the  $SW$  test (5% test) to detect four alternative distributions (logistic, double exponential,  $t_2$  and Cauchy) as a function of sample size. For samples of 20 or more, the  $SW$  test has power over .85 to distinguish normal data from Cauchy. Samples of at least 60 (310) are needed to have the same power to distinguish between the normal and  $t_2$  (double-exponential). On the other hand, the power to detect a logistic distribution from a normal is never over .2 and does not increase for sample sizes between 50 and 300. This result is somewhat surprising and may reflect the slowness of the  $SW$  statistic to approach its limiting distribution. Because it is difficult to distinguish data from a logistic or double exponential distributions from a normal in a moderate sized sample, we replace the  $t$ -test by the Wilcoxon test as the “default” test in our adaptive method.

When the differences follow a normal distribution, the ordinary  $t$ -test is optimum (LEHMANN, 1986). It is well known, however, that if the data are not normal,

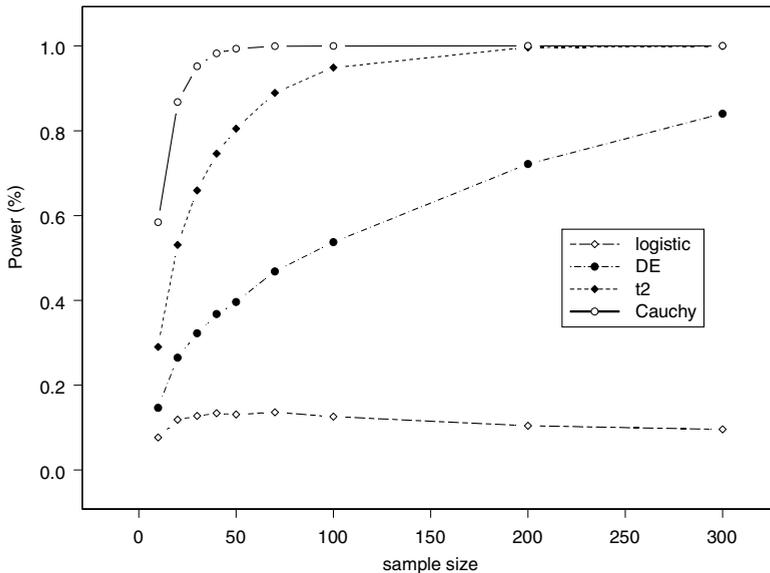


Fig. 1. Power of  $SW$  (5%) statistic to distinguish Logistic, DE,  $t_2$  and Cauchy from a normal distribution

the  $t$ -test need not be the most powerful procedure. Nonparametric and robust methods were developed with the aim of having relatively high power when the data comes from non-normal as well as normal populations. Let  $Y_1, \dots, Y_n$  be the paired differences,  $R_i^+$  be the absolute rank of  $Y_i$ , i.e. the rank of  $|Y_i|$  among  $|Y_1|, \dots, |Y_n|$ . Let  $\delta(y)$  be an indicator function taking value of 1 if  $y$  is positive and 0 otherwise. For symmetric distributions, most of the commonly used non-parametric tests can be written in the form of a linear signed rank test:

$$T = \frac{1}{n} \sum_{i=1}^n a(R_i^+) \delta(Y_i) \quad \text{with} \quad a(i) = J\left(\frac{i}{n+1}\right), \quad (1)$$

and  $J(u)$  is a normalized score function defined on  $(0,1)$  satisfying  $\int_0^1 J(u) du = 0$  and  $\int_0^1 J^2(u) du = 1$ . When  $J(u) = 2\sqrt{3}(u - \frac{1}{2})$ , the test  $T$  is the Wilcoxon test which has asymptotic relative efficiency (ARE) .955 relative to the  $t$ -test on normal data and over the entire family of symmetric distributions its minimum ARE relative to the  $t$ -test is at least .864 (HODGES and LEHMAN, 1956).

Symmetric distributions are characterized by their tailweight as this reflects the probability of obtaining a fairly extreme observation. We adopt a minor variant of a tailweight criterion due to RANDELES and HOGG (1973);

$$\theta = \frac{10 \left[ \int_{F^{-1}(0.95)}^{\infty} t dF(t) - \int_{-\infty}^{F^{-1}(0.05)} t dF(t) \right]}{\int_{F^{-1}(0.5)}^{\infty} t dF(t) - \int_{-\infty}^{F^{-1}(0.5)} t dF(t)}.$$

As the EPA was concerned with the differences from distributions with tails at least as heavy as the normal this study considers normal, logistic, contaminated normal, double exponential,  $t_2$ , Cauchy and slash distributions. The slash distribution is the ratio of a standard normal over a uniform random variable (MORGENTHALER and TUKEY, 1991). This family is well representative of the entire range of possible symmetric alternatives (HALL and JOINER, 1982; MORGENTHALER and TUKEY, 1991). The value of  $\theta$  for the light tailed uniform distribution and those used in our study are: uniform (1.9), normal (2.585), logistic (2.864), contaminated normal (3.192), double exponential (3.303),  $t_2$  (4.359), Cauchy (10), slash (10).

In the situation where the family of alternative distributions is believed to be limited to Normal, Logistic or the Double exponential, the Wilcoxon signed rank test is highly correlated with the maximin efficiency robust test (GASTWIRTH, 1966). Thus, the Wilcoxon test can be used without a preliminary test when it is reasonable to assume the data come from any one of those three distributions. Consequently, if the data have moderate tails, the Wilcoxon test is used at the second stage. On the other hand, if the data indicates that the distribution has heavy tails, one should use non-parametric tests that have high power for data

from such distributions at the second stage. After examining the sampling distribution of the *SW* test under various distributions and exploring a variety of selection rules for sample sizes of 10–300, the following 3 linear signed rank tests, with the score functions specifying them, were chosen for use at the second stage with score functions:

$$\begin{aligned}
 J_1(u) &= 2\sqrt{3}\left(u - \frac{1}{2}\right), \\
 J_2(u) &= \sqrt{30}\left(u - \frac{1}{2}\right)\sqrt{1 - 4\left(u - \frac{1}{2}\right)^2}, \\
 J_3(u) &= \frac{2 \tan\left[\pi\left(u - \frac{1}{2}\right)\right]}{1 + \tan^2\left[\pi\left(u - \frac{1}{2}\right)\right]}.
 \end{aligned}$$

Notice that the first linear signed rank test *T1* is just the Wilcoxon test, the test *T2* is the most powerful rank test for data from a *t*-distribution with 2 degrees of freedom (GASTWIRTH, 1970), and the third test, *T3*, is the Cauchy scores test (CAPON, 1961). Within the family of symmetric, unimodal distributions, the lower the *p*-value of the *SW* test, the heavier is the tail of the distribution. Hence, we use this *p*-value to choose the second stage test. Let  $p_{SW}$  be the *p*-value of the *SW* test. The new adaptive procedure, denoted by *S3*, is:

- If  $p_{SW} > 0.01$ , choose the test *T1* with score function  $J_1(u)$ ;
- If  $0.01 \geq p_{SW} > 0.0001$ , choose the test *T2* with score function  $J_2(u)$ ;
- If  $p_{SW} \leq 0.0001$ , choose the test *T3* with score function  $J_3(u)$ .

### 3. Asymptotic properties of the *SW* statistic

In order to guarantee the type I error, adaptive methods use a selector statistic that is independent of the second stage test (see RANGLES and WOLFE 1979; HOGG, 1974; BÜNING and KÖSSLER, 1998). The rationale underlying the new adaptive procedure is that the *SW* test and linear signed rank test of form (1) are asymptotically uncorrelated. Thus, one expects their degree of dependence in moderate-sized samples will be quite small and have little effect on the significance level of the two-stage procedure. Since the standard condition of independence is not met a simulation study was carried out. The results indicate that the inflation of the overall significance level of the adaptive method *S3* was very small (less than .002 for samples of size (*n*) greater than 8) and diminishes as *n* increases. The main results are stated here and their derivations are outlined in the appendix.

First, one shows that up to terms of  $o_p(n^{-1/2})$ , the Shapiro-Wilk test statistic, *SW*, is equivalent to the statistic

$$W^* = \left(\frac{1}{n} \tilde{Y}'c\right)^2 / s^2, \tag{2}$$

where  $Y'$  is the vector of i.i.d. observations  $Y_1, \dots, Y_n$  from distribution  $F$ ,  $\tilde{Y}'$  is the vector of the order statistics  $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ ,  $c$  is the vector with  $c_i = \Phi^{-1}(i/(n+1))$ ,  $\Phi(\cdot)$  is standard normal c.d.f., and  $s^2 = (1/n) \sum (Y_i - \bar{Y})^2$ . From CHERNOFF et al. (1967),  $W^*$  is seen to be the square of the ratio of the optimal L-estimator of  $\sigma$  for normal data to  $s$ , i.e. the numerator and the denominator of the  $W^*$  are asymptotically equivalent.

When the data come from a normal distribution, it can be shown that the statistic  $SW$  approaches 1 at a rate faster than  $n^{-1/2}$ . The precise result is given in DE WET and VENTER (1972) and LESLIE et al. (1986). For non-normal data, e.g., the underlying density is logistic or double exponential, the following asymptotic result holds:

**Theorem 1:** When the sample statistic  $Y$  comes from double exponential or logistic distribution,  $\sqrt{n}(SW - \mu) \Rightarrow N(0, \tau^2)$  for some  $\mu$  and  $\tau^2$ . For the double exponential  $\mu = .963$  and  $\tau^2 = .145$  and for the logistic  $\mu = .991$  and  $\tau^2 = .058$ .

Since the means of the asymptotic distributions of the  $SW$  statistic on data from the double exponential and logistic distributions are near 1, large samples will be needed to distinguish those models from the normal one (see Figure 1).

We next discuss the joint asymptotic distribution of the  $SW$  statistic and the signed rank statistics of form (1). The key result, which follows from HOLLANDER (1968) and RANGLES and HOGG (1973), is:

**Theorem 2:** For symmetric distributions the SHAPIRO-WILK test is asymptotically uncorrelated with any signed rank tests of form of (1) as well as with the  $t$ -test.

For samples from a normal distribution, we have:

**Corollary 3:** For samples from normal distribution, the  $SW$  and the signed rank tests of form (1) are asymptotically independent.

#### 4. Power simulation results

In order to compare the power properties of the procedures, samples of size 8, 15, 25, 50 and 300 paired differences were simulated. We considered the following possible distributions for them: (1) normal, (2) logistic, (3) contaminated normal, (4) double exponential, (5)  $t_2$ , (6) Cauchy, and (7) slash. The results for the different sample sizes are given in Table 1. For each sample size, empirical power estimates for the one sample  $t$ -test, the signed rank tests optimal for each of the distributions 1–7, and the adaptive procedures are presented. Two versions of the  $STW$  procedure, denoted by  $STW$  5% and  $STW$  10%, respectively, are examined depending on whether the 5% or 10% level is used to reject normality. Two-sided .05 level tests are used at the second stage. In order to ensure that the levels of the non-adaptive tests were within .002 of .05, randomized versions were used, when necessary. Notice that in each case, the alternative  $\Delta$  was chosen so that the power of the optimal test for each underlying distribution was near 85%. Since the EPA's

procedure uses two tests at the same size  $\alpha$ , the overall type I error exceeds  $\alpha$  (GASTWIRTH, 1988). Our simulation indicated that this inflation is substantial. Thus, we reduced the size of each test so that the overall test had the pre-specified level (.05). Results for this adjusted version of the EPA's method are reported.

Table 1 indicates that the new procedure, *S3*, has better power robustness than the three other methods (*STW 5%*, *STW 10%*, *EPA*) over the set of distributions for sample sizes of 50 or less, which are most relevant for the motivating application. Only for sample size 300 from contaminated normal distribution was the power of the new procedure noticeably less than the other methods. Such sample sizes of paired data, however, are not realistic. In almost all other cases, the *S3* procedure had power within a few percentage points of the best performing test for each distribution. For heavier tailed distributions, its power was noticeably higher than the Wilcoxon, both *STW* tests and the *EPA* procedure. For example, when  $N = 25$ , for Cauchy or slash distributions, the power of *S3* exceeded that of either *STW* test and the *EPA*'s test by over 15%. This large increase in power for heavier tailed distributions cost less than 2% power when the differences are truly normally distributed. The pattern for the other sample sizes  $\leq 50$  was similar. Since the new procedure is easy to implement, it should be used in practice.

Table 1  
Empirical size and power estimates

	Null (normal)	Normal	Logistic	Contam. Normal*	DE	t2	Cauchy	Slash	Mixture DE**
<i>N</i> = 8									
<i>t</i> -test	.050	.856	.865	.856	.860	.826	.653	.660	.849
Normal scores	.050	.838	.849	.848	.843	.828	.717	.721	.804
Wilcoxon	.050	.837	.850	.850	.845	.831	.719	.724	.796
Sign	.050	.671	.721	.769	.747	.799	.767	.770	.555
<i>T</i> 2 scores	.050	.827	.846	.857	.846	.849	.763	.768	.774
Cauchy scores	.050	.711	.760	.823	.790	.857	.856	.855	.627
<i>STW 5%</i>	.050	.856	.864	.859	.857	.839	.722	.728	.843
<i>STW 10%</i>	.051	.855	.863	.859	.855	.837	.722	.727	.835
<i>S3</i>	.051	.838	.852	.857	.849	.855	.806	.807	.797
<i>EPA</i> procedure	.050	.837	.854	.857	.853	.844	.735	.740	.826
<i>N</i> = 15									
<i>t</i> -test	.050	.856	.845	.792	.793	.655	.357	.406	.835
Normal scores	.050	.847	.841	.814	.799	.736	.585	.633	.834
Wilcoxon	.051	.842	.853	.846	.829	.786	.650	.700	.847
Sign	.049	.678	.742	.750	.783	.786	.771	.798	.683
<i>T</i> 2 scores	.050	.793	.836	.852	.849	.846	.786	.824	.823
Cauchy scores	.050	.654	.735	.765	.796	.823	.845	.865	.701
<i>STW 5%</i>	.051	.856	.852	.837	.818	.771	.641	.689	.849
<i>STW 10%</i>	.051	.855	.854	.841	.821	.776	.645	.693	.850
<i>S3</i>	.052	.845	.856	.856	.840	.828	.801	.846	.850
<i>EPA</i> procedure	.050	.831	.840	.825	.824	.787	.715	.748	.825

Table 1 (Continued)

	Null (normal)	Normal	Logistic	Contam. Normal*	DE	t2	Cauchy	Slash	Mixture DE**
<i>N</i> = 25									
<i>t</i> -test	.050	.853	.836	.767	.749	.572	.219	.263	.798
Normal scores	.050	.847	.843	.822	.785	.728	.554	.602	.832
Wilcoxon	.050	.837	.854	.850	.825	.793	.659	.703	.856
Sign	.050	.679	.749	.746	.804	.790	.781	.783	.749
<i>T</i> 2 scores	.050	.768	.825	.840	.848	.850	.801	.829	.848
Cauchy scores	.050	.603	.708	.724	.802	.816	.850	.852	.746
<i>STW</i> 5%	.051	.852	.848	.838	.799	.773	.652	.696	.840
<i>STW</i> 10%	.051	.851	.850	.843	.807	.779	.654	.698	.845
<i>S</i> 3	.051	.837	.855	.852	.834	.830	.832	.841	.856
EPA procedure	.050	.831	.832	.805	.805	.755	.687	.696	.817
<i>N</i> = 50									
<i>t</i> -test	.050	.859	.850	.751	.756	.482	.112	.150	.722
Normal scores	.049	.854	.861	.827	.814	.715	.520	.598	.798
Wilcoxon	.049	.842	.874	.854	.860	.790	.641	.713	.838
Sign	.050	.677	.766	.735	.855	.777	.761	.769	.775
<i>T</i> 2 scores	.050	.752	.837	.830	.884	.845	.792	.836	.847
Cauchy scores	.050	.566	.708	.688	.852	.805	.841	.847	.777
<i>STW</i> 5%	.050	.857	.861	.838	.824	.774	.641	.711	.809
<i>STW</i> 10%	.051	.856	.864	.844	.835	.779	.641	.712	.818
<i>S</i> 3	.051	.842	.872	.834	.868	.820	.838	.847	.836
EPA procedure	.050	.827	.849	.799	.859	.761	.716	.725	.803
<i>N</i> = 300									
<i>t</i> -test	.050	.848	.815	.732	.633	.354	.035	.043	.639
Normal scores	.050	.847	.832	.825	.731	.721	.512	.594	.774
Wilcoxon	.050	.831	.847	.853	.794	.805	.650	.723	.832
Sign	.049	.662	.732	.727	.845	.788	.772	.762	.852
<i>T</i> 2 scores	.050	.713	.796	.814	.831	.857	.807	.843	.861
Cauchy scores	.050	.509	.649	.647	.815	.813	.853	.844	.836
<i>STW</i> 5%	.050	.846	.822	.845	.778	.805	.650	.723	.831
<i>STW</i> 10%	.050	.845	.824	.847	.783	.805	.650	.723	.831
<i>S</i> 3	.050	.829	.843	.722	.819	.813	.853	.844	.841
EPA procedure	.050	.815	.814	.777	.828	.756	.716	.708	.838

\* Contaminated normal: *N* (mean = 0 SD = 1) with probability .9 and *N* (mean = 0, SD = 3) with probability .1.

\*\* Mixture of double exponentials: DE(mean = 0) with probability .75 and DE(mean = Δ) with probability .25

Values of parameter Δ for each of the alternative distributions for *n* = 8, 15, 25, 50, 300.

Normal: 1.25, 0.84, 0.627, 0.438, 0.173

Logistic: 2.31, 1.50, 1.11, 0.785, 0.30

Contaminated Normal : 1.63, 1.00, 0.735, 0.505, 0.20

Double exponential: 1.80, 1.508, 0.765, 0.538, 0.188

t2: 2.46, 1.25, 0.878, 0.578, 0.228

Cauchy: 4.10, 1.65, 1.073, 0.661, 0.25

Slash: 5.30, 2.40, 1.57, 1.02, 0.391

Mixture of double exponentials: 5.4, 2.68, 1.8, 1.1, 0.4

At the suggestion of a referee an asymmetric alternative was also explored. We assumed a mixture of double exponential with mean 0 and another double exponential with mean  $\Delta > 0$  (with probabilities .75 and .25, respectively). In the motivating example this corresponds to a situation where the new fuel is equivalent to the old value for a sizeable fraction of cars but produces more pollution for other type of cars. The results in the last column of Table 1 indicate that the adaptive procedures perform well.

## 5. Examples

To illustrate the comparative performance of the new procedure (*S3*), we use it and other methods to reanalyze three data sets. We applied the EPA's original procedure so it has an overall level exceeding .05, which should give it a power advantage over the *STW* and *S3* methods. The first data set refers to paired electrical measurements on 24 wiring boards, each board was measured right after soldering and after three weeks of exposure to a high temperature environment (IMAN, 1995). These data are used in the SAS JMP guidebook (SALL et al. 1996). The results of the adaptive and non-adaptive tests are reported in the first column of Table 2. Notice that normality is clearly rejected ( $p_{SW} < .01$ ). While both adaptive procedures reject the null hypothesis, the *p*-value (.0008) obtained using the new procedure is more significant than the one (.012) from *STW*. The *p*-value of *S3* equals that yielded by the *T2* test. This is sensible as the *S3* procedure selects that test after rejecting normality. It should be noted that the *t*-test does not reject the null hypothesis while the normal scores test (CHERNOFF and SAVAGE, 1958) just barely rejects it at the .05 level. The EPA's procedure will also reject the null hypothesis as the sign test is significant.

The second data set consists of the differences between the emission levels of 16 cars when they were driven with two different fuels. The data arose in a law case questioning the EPA's decision that they were equivalent. The case and the data set are described in GASTWIRTH (1988, p. 611) and FINKELSTEIN and LEVIN (2001, p. 187). The results of several tests are given in the second column of Table 2. As the *SW* test rejects normality at the .05 level but not at the .01 level, both *S3* and the *STW* use the Wilcoxon and yield the same *p*-value. This may be a function of the small sample size, which did not yield a sufficiently low *p*-value of the *SW* test to have *S3* applying the *T2* or Cauchy scores test to the data. These tests gave even more significant results. The EPA's method also detected a significant difference. But the EPA did not follow its own result, however, triggering the lawsuit.

The third example deals with differences in corn yields from sprayed vs. unsprayed strips from 14 farms in Bone County, Iowa (SNEDECOR and COCHRAN, p. 71 1989). The test results are presented in the last column of Table 2. The differences clearly are not normally distributed so both *STW* and *S3* chose a non-

Table 2  
Test results for the three data sets

Test	<i>p</i> -value		
	SAS JMP data 0.05, 0.06, -0.06, -0.11, -0.16, -0.17, -0.3, -0.3, -0.31, -0.54, -0.74, -0.82, -0.83, 0.83, -0.88, -0.9, -1.42, -1.45, -1.67, -1.82, -1.92, -2.66, 2.8, 2.92 tailweight = 3.204	EPA/Petro. Data 0.0, -0.015, 0.02, -0.029, 0.043, 0.045, 0.06, 0.09, 0.18, 0.18, 0.188, 0.19, 0.219, 0.231, 0.343, -0.4 tailweight = 3.133	Boone County Data -0.5, -0.9, 1.3, -1.5, 1.6, 1.7, 2.3, 2.5, 2.8, 3.0, 4.4, 6.8, -7.7, 22.0 tailweight = 4.1747
Shapiro-Wilk	.0076	.0258	.0018
<i>t</i> -test	.1107	.0626	.1429
Normal score	.0458	.0511	.0553
Wilcoxon	.0123	.0175	.0419
Sign	.0066	.0352	.1796
<i>T</i> <sub>2</sub> scores	.0008	.0030	.0234
Cauchy scores	.0002	.0028	.0266
<i>STW</i> *	.0123	.0175	.0419
<i>S</i> <sub>3</sub>	.0008	.0175	.0234
EPA procedure	.0066	.0352	.1429

\* both *STW* 5% and *STW* 10% have identical *p*-values and are reported in the same row

parametric test and found a significant difference. As *S*<sub>3</sub> select the *T*<sub>2</sub> scores test, it yields a lower *p*-value than *STW*, which applies the Wilcoxon test at the second stage. Notice that the EPA's method did not detect this difference as the *p*-value of both the *t* and sign tests exceeded .10.

## 6. Discussion

Although the Wilcoxon test has high relative efficiency on normal data, it is not very efficient, relative to the optimum tests, on heavier tailed distributions, e.g., *t*<sub>2</sub>, and Cauchy. Thus, the *STW* procedure, which uses the Wilcoxon test when the *SW* test rejects normality, does not have high power when the differences are heavy tailed. When normal or near normal data is anticipated, the Wilcoxon test provides a method with good power properties (LEHMANN, 1986). When the differences may come from a family of symmetric distributions that includes both moderate and heavy tailed ones, an adaptive procedure, such as *S*<sub>3</sub> that can select an optimal rank test for a heavy tailed distribution is preferable. The results on both simulated and actual data demonstrate that one can obtain an adaptive test that has high power relative to the best test for a reasonably large class of moderate to heavy tailed symmetric distributions.

Acknowledgements

This Research was supported in part by a grant from National Science Foundation and was completed while Prof. Gastwirth was visiting the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics at the National Cancer Institute. It is a pleasure to thank the referees for their thoughtful and useful comments.

Appendix

Let  $m$  and  $V$  be the expected value and covariance matrix of the order statistics from a standard normal distribution and  $g = V^{-1}m$ . The Shapiro-Wilk statistic ( $SW$ ) is:

$$SW = \frac{(\tilde{Y}'g)^2}{ns^2 \|g\|^2} = \left[ \frac{\|g\|^2}{4n} \right]^{-1} \frac{\left( \frac{1}{2n} \tilde{Y}'g \right)^2}{s^2} = a_n W^* + b_n. \tag{A.1}$$

with  $a_n = \left[ \frac{\|g\|^2}{4n} \right]^{-1}$ ,  $W^* = \left( \frac{1}{n} \tilde{Y}'c \right)^2 / s^2$  and  $b_n = a_n \left[ \frac{\left( \frac{1}{2n} \tilde{Y}'g \right)^2 - \left( \frac{1}{n} \tilde{Y}'c \right)^2}{s^2} \right]$ .

First, we show that up to terms of  $o_p(n^{-1/2})$ , the  $SW$  is equivalent to the statistic  $W^*$ . Using the results  $\frac{\|m\|^2}{n} = 1 + o(n^{-1/2})$ ,  $\|g - 2m\|^2 = O((\log n)^{-1})$  (lemmas iii, iv and equation (1) of LESLIE, STEPHENS and FOTOPULUS, 1986) and  $\|m - c\|^2 = O((\log n)^{-1})$  (VERRILL, 1987), we get:

$$\frac{\|g\|^2}{4n} = 1 + o(n^{-1/2}), \quad \frac{\|c\|^2}{n} = 1 + o(n^{-1/2})$$

and  $\|g - 2c\|^2 = O((\log n)^{-1})$ . (A.2)

For distributions with finite variance, the Law of Large Number implies  $\frac{1}{n} \|\tilde{Y}\|^2 = \frac{1}{n} \sum Y_{(i)}^2 \rightarrow \mu_2$ , for some  $\mu_2$ . So,  $\frac{1}{\sqrt{n}} \|\tilde{Y}\| \rightarrow \sqrt{\mu_2}$ .

$$\left| \frac{1}{2n} \tilde{Y}'g - \frac{1}{n} \tilde{Y}'c \right| = \frac{1}{2n} |\tilde{Y}'(g - 2c)| \leq \frac{1}{2\sqrt{n}} \left( \frac{1}{\sqrt{n}} \|\tilde{Y}\| \right) \|g - 2c\| = o(n^{-1/2}). \tag{A.3}$$

Furthermore, Theorem 3 of CHERNOFF et al. (1967) implies that when sample comes from double exponential or logistic distributions, the L-statistic  $\frac{1}{n} \tilde{Y}'c$  con-

verges to some constant, so does the  $\frac{1}{2n} \tilde{Y}'g$ , according to (A.3). From (A.2) and (A.3), we have:

$$a_n = 1 + o(n^{-1/2}) \quad \text{and} \quad b_n = o(n^{-1/2}).$$

Hence, the statistics  $SW$  and  $W^*$  have the same  $\sqrt{n}$  - asymptotic distribution.

**Proof of Theorem 1:** It is sufficient to show that  $\sqrt{n}(W^* - \mu) \Rightarrow N(0, \tau^2)$ . Let  $T_1 = (1/n) \tilde{Y}'c$ ,  $T_2 = (1/n) \sum Y_{(i)}^2$  and  $T_3 = \bar{Y}$ . Then  $W^*$  can be written as

$$W^* = \frac{\left(\frac{1}{n} \tilde{Y}'c\right)^2}{\frac{1}{n} \sum Y_{(i)}^2 - \bar{Y}^2} = \frac{T_1^2}{T_2 - T_3^2}.$$

Because  $T_1$  and  $T_2$  are linear combinations of functions of order statistics, Theorem 3 of CHERNOFF et al. (1967) can be used to show that when the sample comes from a double exponential or logistic distributions,  $\sqrt{n}(T_i - \mu_i)/\sigma_i \Rightarrow N(0, 1)$ ,  $i = 1, 2$ . Statistic  $T_3$  is the sample mean and by the central limit theorem  $\sqrt{n}(T_3 - 0)/\sigma_3 \Rightarrow N(0, 1)$ . Using the delta-method and the Remark 9 of CHER-

NOFF et al. (1967), it can be shown that  $\sqrt{n}(W^* - \mu) \Rightarrow N(0, \tau^2)$  with  $\mu = \frac{\mu_1^2}{\mu_2}$  and  $\tau^2 = \frac{4\mu_1^2\sigma_1^2}{\mu_2^2} + \frac{\mu_1^4\sigma_2^2}{\mu_2^2} - 4\frac{\mu_1^3}{\mu_2^3} \text{cov}(T_1, T_2)$ .

**Proof of Theorem 2:** It is sufficient to show that  $\text{cov}(SW, T) = 0$  asymptotically, where  $T$  refers to a signed rank test of form (1) or the  $t$ -test. The proof follows along those outlined in HOLLANDER (1968) and RANDES and HOGG (1973).

Let  $Z = (Z_1, \dots, Z_n)'$ , be an i.i.d. sample, independent of  $Y$ , where  $Z = 0$  with probability 1. Then a signed rank test of  $Y$  can be written as:  $T(Y) = T(Y, Z) = \sum \delta(Y_i - Z_i) a(R_i^+)$  where  $\delta(x) = 1$  if  $x > 0$ ,  $\delta(x) = 0$  if  $x \leq 0$ ,  $R_i^+$  is the rank of  $|Y_i - Z_i|$ , and  $a(i)$  is the centered score function. Obvious, for any scale  $b = (b, b, \dots, b)'$ ,  $T(Y + b, Z + b) = T(Y, Z)$ . The fact that score function  $a(i)$  is centered implies that asymptotically  $T(-Y, -Z) = -T(Y, Z)$ . In other words, asymptotically  $T(Y) = T(Y, Z)$  is odd and translation invariant. The  $t$ -test statistic of  $Y$  can be written as:  $t(Y) = t(Y, Z) = \frac{\bar{Y} - \bar{Z}}{ns/((n-1)\sqrt{n})}$ . Since  $s$  is even and translation invariant,  $t(Y) = t(Y, Z)$  is also odd and translation invariant.

Recall that the  $g = V^{-1}m$ . Since standard normal distribution is symmetric,  $g_i = -g_{n+1-i}$ . Consequently, for any scale  $b = (b, b, \dots, b)'$ ,  $(\tilde{Y} + b)'g = \tilde{Y}'g + bg = \tilde{Y}'g$ . Furthermore, let  $(-\tilde{Y}) = ((-Y)_{(1)}, \dots, (-Y)_{(n)})'$  be the ordered statistics of  $-Y$ . Then

$$(-\tilde{Y})'g = \sum_{i=1}^n (-Y)_{(i)}g_i = \sum_{i=1}^n (-Y)_{(n+1-i)}(-g_{n+1-i}) = \tilde{Y}'g.$$

Which means that  $(\tilde{Y}'g)^2$  is even and translation invariant, so does

$$SW(Y) = SW(Y, Z) = \frac{(\tilde{Y}'g)^2}{ns^2\|g\|^2}.$$

Obviously, the joint distribution of  $Y$  and  $Z$  is symmetric about  $(0, 0)'$ . According to Theorem 1 of HOLLANDER (1968), asymptotically,  $\text{Cov}(SW, T) = 0$ .

**Proof of Corollary 3:** When the distribution is normal,  $SW$  converges to 1 at a rate faster than  $\sqrt{n}$  (e.g. LESLIE et al., 1986). Hence, by Slutsky's Theorem,  $\sqrt{n}(SW - 1) \rightarrow 0$  in probability.

It is well known that for any signed rank test  $T$ ,  $\sqrt{n}(T - \mu) \Rightarrow N(0, \sigma^2)$ . So, for any constant  $a$  and  $b$ , using the Slutsky's Theorem again, we have:

$$a\sqrt{n}(SW - 1) + b\sqrt{n}(T - \mu) \Rightarrow N(0, b^2\sigma^2),$$

which means that  $\sqrt{n}(SW - 1)$  and  $\sqrt{n}(T - \mu)$  are asymptotically joint normal. Theorem 1 implies that  $\sqrt{n}(SW - 1)$  and  $\sqrt{n}(T - \mu)$  are also asymptotically uncorrelated. Consequently,  $SW$  and  $T$  are  $\sqrt{n}$ -asymptotically independent.

## References

- BAUER, P. and KOHNE, K., 1994: Evaluation of experiments with adaptive interim analysis. *Biometrics* **50**, 1029–1041.
- BÜNING, H. and KÖSSLER, W., 1998: Adaptive tests for umbrella alternatives. *Biometrical Journal* **40**, 573–587.
- CAPON, J., 1961: Asymptotic efficiency of certain locally most powerful rank tests. *The Annals of Mathematical Statistics* **32**, 88–100.
- CHERNOFF, H., and SAVAGE, I. R., 1958: Asymptotic normality and efficiency of certain nonparametric test statistics. *The Annals of Mathematical Statistics* **29**, 972–94.
- CHERNOFF, H., GASTWIRTH, J. L., and JONES, M. V., 1967: Asymptotic distribution of linear combinations of function of order statistics with applications to estimation. *The Annals of Mathematical Statistics* **38**, 352–72.
- DE WET, T. and VENTER, J. H., 1973: Asymptotic distributions for quadratic forms with applications to tests of fit. *The Annals of Statistics* **1**, 380–387.
- FINKELSTEIN, M. O. and LEVIN, B., 2001: *Statistics for lawyers*, 2nd edition. Springer, New York.
- GAN, F. F. and KOEHLER, K. J., 1990: Goodness-of-fit tests based on p-p probability plots, *Technometrics* **32**, 289–303.
- GASTWIRTH, J. L., 1966: On robust procedures. *Journal of the American Statistical Association* **61**, 929–948.
- GASTWIRTH, J. L., 1970: On robust rank tests. In M. L. Puri (ed.): *Non-parametric techniques in statistical inference*. Cambridge University Press, Cambridge.
- GASTWIRTH, J. L., 1988: *Statistical Reasoning in Law and Public Policy: (Vol. 2) Tort Law, Evidence and Health*. San Diego, CA: Academic Press.
- HILL, N. J., PADMANABHAN, A. R., PURI, M. L., 1988: Adaptive nonparametric procedures and applications. *Applied Statistics* **37**, 205–218.
- HALL, D. L., and JOINER, B. L., 1982: Representations of the space of distributions useful in robust estimation of location. *Biometrika* **69**, 55–59.

- HODGES J. L. and LEHMANN, E. L., 1956: The efficiency of some nonparametric competitors of the  $t$ -test. *The Annals of Mathematical Statistics* **27**, 324–335.
- 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.
- HOLLANDER, M., 1968: Certain uncorrelated nonparametric test statistics. *Journal of The American Statistical Association* **63**, 707–714.
- IMAN, R. L., 1995: *A Data-Based Approach to Statistics*. Duxbury Press: Belmont.
- LANG, T., AUTERITH, A., and BAUER, P., 2000: Trend tests with adaptive scoring. *Biometrical Journal* **42**, 1007–1020.
- LEACH, C. S., LANE H. W., and KRAUHS, J. M., 1994: Short-term space flight on nitrogenous compounds, lipoproteins, and serum proteins. *Journal of Clinical Pharmacology* **34**, 500–509.
- LEHMANN, E. L., 1986: *Testing Statistical Hypotheses*. Chapman & Hall, New York.
- LESLIE, J. R., STEPHENS, M. A., and FOTOPOULOS, S., 1986: Asymptotic distribution of the Shapiro-Wilk statistic for testing for normality. *The Annals of Statistics* **14**, 1497–1506.
- LILLIEFORS, H. W., 1967: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62**, 399–402
- MORGENTHALER, S. and TUKEY, J. W., 1991: *Configural Polysampling: a route to practical robustness*. Wiley, New York.
- NEUHAUSER, M., 2001: An adaptive location-scale test. *Biometrical Journal* **43**, 809–819
- O’GORMAN, T. W., 1997: A comparison of an adaptive two-sample test to the  $t$ -test and the rank-sum test. *Communications in Statistics – Simulation and Computation* **26**, 1393–1411.
- RAMSEY, P. and RAMSEY P., 1990: Simple Tests of Normality in Small Samples. *Journal Of Quality Technology* **22**, 299–309.
- RANDLES, R. H. and HOGG, R. V., 1973: Adaptive distribution-free test. *Communications in Statistics* **2**, 337–356.
- RANDLES, R. H. and WOLFE, D. A., 1979: *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.
- RUBERG, S. J., 1986: A continuously adaptive nonparametric two-sample test. *Communications in Statistics – Theory and Methods* **15**, 2899–2920.
- SALL, J., LEHMAN, A., and SAUL, J., 1996: *JMP Start Statistics: A Guide to Statistical and Data Analysis Using JMP*. Duxbury, New York.
- SHAPIRO, S. S. and WILK, M. B., 1965: An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611.
- SNEDECOR, G. W. and COCHRAN, G. C., 1989: *Statistical Methods*, 8<sup>th</sup> ed. Iowa State University Press, Ames.
- VERRILL, S. and JOHNSON, R. A., 1987: The asymptotic equivalence of some modified Shapiro-Wilk statistics: complete and censored sample cases. *The Annals of Statistics* **15**, 413–419.
- WASSMER, G., 2000: Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers* **41**, 253–279.
- WINDUS, W. D., SANTORO, S. A., ATKINSON, R., and ROYAL, H. D., 1997: Effects of antiplatelet drugs on dialysis-associated platelet deposition in polytetrafluoroethylene grafts. *American Journal of Kidney Diseases* **29**, 560–564.

Received July 2002

Revised April 2003

Accepted June 2003