On Bootstrap Tests of Symmetry About an Unknown Median

Tian Zheng[†§] Joseph L. Gastwirth[‡]

[†]Department of Statistics, Columbia University, New York, New York 10027. tzheng@stat.columbia.edu

[‡]Department of Statistics, George Washington University, Washington, DC 20052. jlgast@gwu.edu

[§]To whom correspondence should be addressed. Tel: (212) 851-2134; Fax: (212) 851-2164.

Abstract

It is important to examine the symmetry of an underlying distribution before applying some statistical procedures to a data set. For example, in the *Zuni School District* case, a formula originally developed by the Department of Education trimmed 5% of the data symmetrically from each end. The validity of this procedure was questioned at the hearing by Chief Justice Roberts. Most tests of symmetry (even nonparametric ones) are not distribution free in finite sample sizes. Hence, using asymptotic distribution may not yield an accurate type I error rate or/and loss of power in small samples. Bootstrap resampling from a symmetric empirical distribution function fitted to the data is proposed to improve the accuracy of the calculated p-value of several tests of symmetry. The results show that the bootstrap method is superior to previously used approaches relying on the asymptotic distribution of the tests that assumed the data

come from a normal distribution. Incorporating the bootstrap estimate in a recently proposed test due to Miao, Gel and Gastwirth (2006) preserved its level and shows it has reasonable power properties on the family of distribution evaluated.

Keywords: Parametric bootstrap; Resampling; Testing symmetry about an unknown center; *Zuni school district* case

Running title: Bootstrap tests of symmetry.

1 Introduction

As noted by Lehmann and Romano (2005, page 248) the problem of testing whether data comes from a symmetric distribution when the center is unknown is more difficult than the corresponding problem when the center is known. In January 2007, the U.S. Supreme Court heard the *Zuni School District 89 v. U.S. Department of Education* case that concerned the formula used to determine whether the school districts in a state have approximately equal funds available for the education of their students. A summary of the issues and previous administrative and federal court proceedings related to the case is given elsewhere (Gastwirth, 2006, 2008). Here we focus on an important topic, implicitly raised by Chief Justice Roberts. In the process of calculating the measure of relative disparity used to determine whether the expenditures in the school districts of a state are sufficiently equal for a state, rather than the local districts, to receive most of the Federal Impact Aid money, the Department of Education deletes the largest and smallest 5% of the data. The Justice asked why *none* of the three parties involved in the case discussed the issue of outliers in their briefs. His question is quite important as trimming the upper and lower 5 or 10% of the data is a well accepted method when the objective is to estimate the center of the data

(Staudte and Sheather, 1990, pg. 68-70). It is questionable when estimating the spread as measures based on trimming may systematically understate the variability in the population (Stuart and Ord, 1994, pg. 59-60). Furthermore, symmetric trimming is not appropriate even when estimating the location or centrality parameter when the data come from a skewed distribution (Collins, 1976; Clarke, Gamble and Bednarski, 2000).

When the center of the data is unknown, the p-values of most tests of symmetry are obtained using the large sample null distribution of the test statistic to their small sample distribution. For most tests of symmetry, the null distributions, even the asymptotic ones (Miao, Gel and Gastwirth, 2006), depend on the form of the underlying distribution. Although in large samples, the parameters of the limiting distribution can be estimated, educational funding data typically refers to school districts in a state. Hence, the available sample sizes often are too small to rely on asymptotic theory. In the *Zuni School District* case, there were only 57 districts and in the amicus brief submitted by Alaska, there were only 44 districts. This paper uses the bootstrap method (Efron, 1979, 1982; Efron and Tibshirani, 1993) to estimate the null distributions of the test statistics for several easy to explain tests of symmetry (e.g. Boos, 1982; Cabilio and Masaro, 1996; Mira, 1999; Miao, Gel and Gastwirth, 2006). Simulation studies are used to determine when the corresponding tests have reliable properties (level and power) in samples of modest size. Educational funding data for local school systems for the states of New Mexico and Alaska from the Zuni School District case as well as data on per-pupil teaching expenditure of individual schools from a classic school segregation case, *Hobson v. Hansen* are then analyzed with several tests of symmetry. The asymmetry of the funding data supports the need for justification of the trimming method adopted by the Department of Education or the development of a

more appropriate method.

In section 2, the symmetry test statistics examined are described and their asymptotic distributions presented. Then the bootstrap method of estimating their distributions in small sample sizes is given. Results on the sample sizes required to preserve the nominal level are given in Section 3.1 and power results are given in Section 3.2. The test proposed by Miao, Gel and Gastwirth (2006), hereafter called MGG, is shown to preserve the level and have good power detecting asymmetry. The tests are then applied to educational funding data in Section 5. The last section summarizes our results and their implications.

2 Methods

In this paper, we consider and compare six simple to explain tests of symmetry about an unknown median. Such tests are very important in applications, especially in the legal setting, as judges and juries rely on their intuitive understanding of the evidence as well as the testimony of expert witnesses.

2.1 Testing symmetry about an unknown median.

Consider n i.i.d. observations, $\{X_1, \ldots, X_n\}$, from an unknown distribution F with an *unknown* median ν, mean μ and standard deviation σ. Denote the sample mean, median and standard deviation by $\bar{X}$, M and s, respectively. If F is symmetric, the distribution $F(z)$ of $X - \nu$ should be the same as that of $\nu - X$, i.e., $1 - F(-z)$. Therefore, testing symmetry is equivalent to testing

$$H_0 : F(x - \nu) = 1 - F(\nu - x)$$

versus

$$H_a : F(x - \nu) \neq 1 - F(\nu - x).$$

Many tests of symmetry are based on the fact that, under the null hypothesis, $\mu = \nu$.

Existing tests of symmetry include tests based on sample measures of skewness (Oja, 1981), differences between the sample mean ($\bar{X}$) and median (M) (Hotelling and Solomon, 1932; Gastwirth, 1971) and differences between the empirical distribution functions of the $(X_i - \nu)$'s and $(\nu - X_i)$'s (Smirnov, 1947; Butler, 1969). When ν is unknown, an estimate $\hat{\nu}$ of ν can replace it in calculation of the expression coefficients. For example, $\hat{\nu}$ can be the sample median M. Other proposed tests include tests based on triplets of observations (Randles et al., 1980), modified sign tests on deviations from a specified center (e.g. sample mean) (Gastwirth, 1971) and modified signed-rank test (Bhattacharya, Gastwirth and Wright, 1982). Readers are referred to Lehmann and Romano (2005) and Hollander (2006) for reviews of this topic.

In this paper, we evaluate two types of simple tests of symmetry, one uses the difference between sample mean ($\bar{X}$) and median (M) and the other compares the distributions of the $(X_i - M)$'s and $(M - X_i)$'s.

2.2 Tests of symmetry based on the difference between $\bar{X}$ and M.

Cabilio and Masaro (1996) (CM1996) studied a simple test of symmetry that compares the sample mean and sample median, standardized by the sample standard deviation, i.e.,

$$C = \frac{\bar{X} - M}{s}.$$

This is a sample version of the measure of skewness,

$$S = \frac{\mu - \nu}{\sigma},$$

which was proposed earlier by Hotelling and Solomon (1932). Cabilio and Masaro (1996) showed that the distribution of C under the null hypothesis of symmetry is asymptotically normal, and derived the asymptotic variance for a set of symmetric distributions (e.g., the asymptotic variance of $\sqrt{n}C$ for a normal distribution is 0.5708). The authors suggested the use of the asymptotic variance derived under normality to obtain critical values for the test of symmetry based on C.

Mira (1999) (M1999) studied a similar test using the difference between the sample mean and median,

$$\hat{\gamma}_1 = \bar{X} - M,$$

which essentially is the sample version of the skewness measure proposed by Bonferroni (1930), $2(\mu - \nu)$. Following Cabilio and Masaro (1996), Mira (1999) used the critical values from the asymptotic normal distribution of $\hat{\gamma}_1$ with F being the standard normal distribution to conduct the test.

Miao, Gel and Gastwirth (2006) proposed a modification of C (Cabilio and Masaro, 1996) that uses a robust estimate of standard deviation. Specifically, Miao, Gel and Gastwirth (2006) used the mean deviation from the median, a robust measure of dispersion (Stuart and Ord, 1994, pg 52-53),

$$J = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \sum_{i=1}^{n} |X_i - M|.$$

Their test statistic is

$$T = \frac{\bar{X} - M}{J}.$$

They showed $\sqrt{n}T$ is asymptotically normal with mean 0 and variance σ_f^2, which depends on the underlying density. When f is normal, $\sigma_f^2 = 0.5708$. For purpose of comparing the MGG test to those of Cabilio and Masaro (1996) and Mira (1999), this value of σ_f^2 was used. Miao, Gel and Gastwirth (2006) also showed that the asymptotic variance σ_f^2 varies from 0.5708 to 0.9689 as the underlying distribution changes from a normal to a t distribution with three degrees of freedom. This range indicates that a more accurate estimation of the null sampling distribution of the test statistic should improve its statistical properties on data from a symmetric distribution that is not very close to normal.

2.3 Tests of symmetry based on difference between the distributions of the $(X_i - M)$'s and $(M - X_i)$'s.

For testing symmetry about a specified center, Kolmogrov-Smirnov type tests such as the Smirnov test (Smirnov, 1947; Butler, 1969) were proposed to compare the distributions of the left and the right deviations: $\{M - X_1, \ldots, M - X_n\}$ versus $\{X_1 - M, \ldots, X_n - M\}$. The test statistic is the supremum distance between the two distribution functions as used in the Kolmogrov-Smirnov tests. It is equivalent to define two sets of *positive* deviations, that is, $\{M - X_i; X_i < M\}$ and $\{X_i - M; X_i > M\}$.

Alternatively, one can focus on the difference in the location parameters of the two distribution. Therefore, location tests such as the two-sample t-test can also be applied. Modified Wilcoxon signed-rank tests have also been considered to test for symmetry (Gupta, 1967; Bhattacharya et al., 1982). If t-test is applied directly on $(M - X_i)$'s and $(X_i - M)$'s, the resulting test statistic equals $-\sqrt{2n}C$ where C is the test statistics of Cabilio and Masaro (1996). The sample means of $(M - X_i)$'s and $(X_i - M)$'s are $M - \bar{X}$ and $\bar{X} - M$ respectively.

The standard deviations of these departures are both s. Therefore the t statistic is $2(M - \bar{X})/s\sqrt{2/n} = \sqrt{2n}(M - \bar{X})/s$. Inspired by the Kolmogrov-Smirnov test described above, we apply two-sample location tests to $\{M - X_i; X_i < M\}$ and $\{X_i - M; X_i > M\}$ instead.

For testing the difference between the distributions of the $(X_i - M)$'s and $(M - X_i)$'s, in this paper, we evaluate the Kolmogrov-Smirnov symmetry test (KS), the two-sample t-test (t-test) and the Wilcoxon signed-rank test (Wilcoxon) between the left positive deviations and the right positive deviations, which are compared to the tests described in Section 2.2.

2.4 Bootstrap estimation

Most tests of symmetry are not distribution-free. As observed in Miao, Gel and Gastwirth (2006), the asymptotic variance of their test statistic depends on the *unknown* underlying distribution of X. Hence, basing the distribution of the symmetry test statistic on a specific underlying distribution, e.g., the normal, may lead to an incorrect Type I error. Resampling and permutation methods are known to yield reliable an estimated sampling distribution of a test statistic under the null hypothesis, while utilizing the natural variation and structure of the observed data.

Schuster and Barker (1987) proposed the use of bootstrap in formulating a nonparametric test of symmetry about a known center. Their strategy involved fitting a "smoothed version" of the closest symmetric distribution to the observed data's distribution. As pointed out by Modarres (2002), the symmetrized empirical distribution function $\hat{F}^s$, which puts mass $1/2n$ at each point of $\{X_1, \ldots, X_n\}$ and $\{2\nu - X_1, \ldots, 2\nu - X_n\}$, is the nonparametric maximum likelihood estimate of the underlying symmetric distribution F. Stine (1985) used a similar strategy to symmetrize residuals. When ν is unknown as in our case, we

estimate ν by M in the symmetrized ECDF $\hat{F}^s$. The resulting ECDF is asymptotically similar to the nonparametric MLE (Efron, 1979; Hinkley, 1976). Due to this connection with the nonparametric MLE of the distribution function, we use bootstrap samples that were drawn from $\hat{F}^s$ to simulate the null distribution of the test statistics investigated. For each bootstrap sample, the set of test statistics outlined in the previous section are computed. After the bootstrap is repeated a large number of times, for a given sample of observed data, the bootstrap p-value for each test is calculated as the proportion of the bootstrap samples that have values further away from the null hypothesis than the observed sample.

3 Simulation studies

This section reports a simulation study assessing the performance of the tests described in Sections 2.2 and 2.3. Their distribution is estimated using the symmetrized bootstrap. Both the level and power of the tests for sample sizes (30, 50, 100, 300) are examined.

3.1 Simulation setup

We first choose four symmetric distributions with different tail characteristics to examine the accuracy of the type I error rate of the tests. They are the standard normal distribution, the student's t distribution with 3 df, Beta distribution with parameters (2,2), and the uniform distribution on $[0, 1]$.

To evaluate the power performance of the tests described in Sections 2.2 and 2.3, we simulate data from five asymmetric distributions. The first is a bimodal mixture of two normal distributions, see Figure 1. The other four distributions come from the generalized lambda distribution family (Ramberg and Schmeiser, 1974), which were also used by Miao,

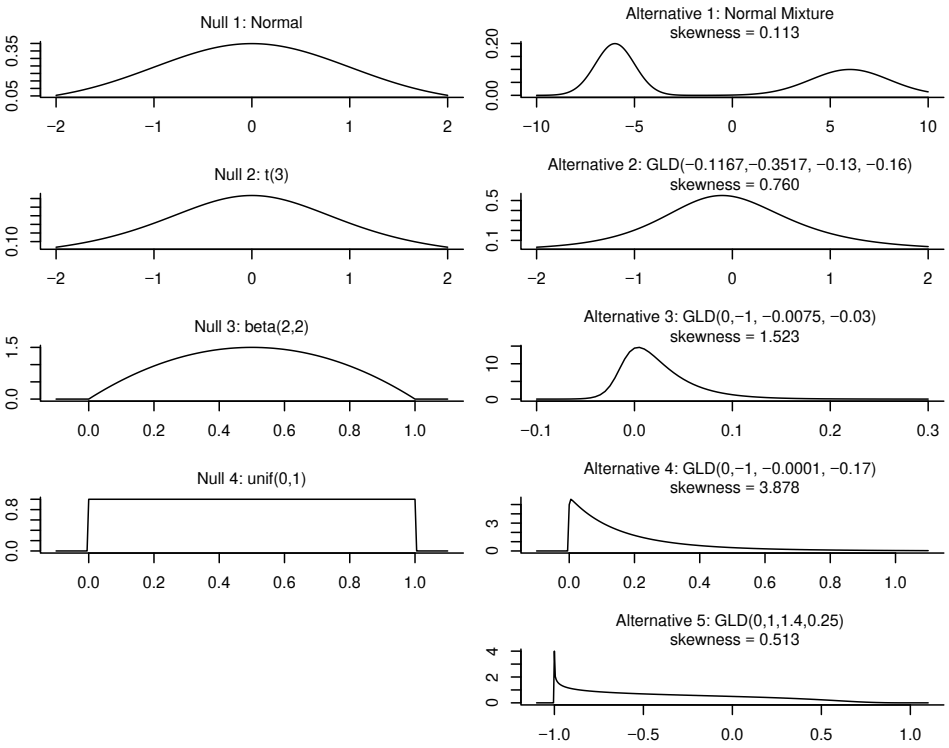

Figure 1: Distribution models used in simulation studies.

Gel and Gastwirth (2006). From Figure 1, one sees that these distributions have different degrees of asymmetry. The first two do not deviate substantially from symmetry with the 3^{rd}, 4^{th} and 5^{th} become increasingly more asymmetric. The measure of skewness $\gamma_1 = \mu_3/\mu_2^{3/2}$ based on the central moments μ_3 and μ_2 (Stuart and Ord, 1994, page 109) is calculated for each of these distributions and noted in Figure 1.

Several sample sizes (n = 30, 50, 100 and 300) are considered in order to determine the sample size required to assure the reliability of the bootstrap method. For each distribution in Figure 1 and each sample size, 1000 data sets are simulated. The tests were then applied to these data sets with p-value evaluated by bootstrapping from the symmetrized $\hat{F}^s$. The nominal significance level $\alpha = 0.05$ is used in this paper.

3.2 Size of the bootstrap tests

Table 1: Size of the bootstrap tests. Number of simulations is 1000.

dist.	n	CM1996	M1999	MGG2006	KS	t-test	Wilcoxon
normal	30	0.052	0.047	0.047	0.051	0.048	0.051
	50	0.068	0.063	0.061	0.047	0.066	0.058
	100	0.045	0.039	0.043	0.029	0.041	0.044
	300	0.056	0.059	0.055	0.039	0.060	0.055
t3	30	0.038	0.032	0.033	0.029	0.034	0.037
	50	0.038	0.036	0.039	0.032	0.039	0.039
	100	0.041	0.043	0.040	0.036	0.040	0.041
	300	0.051	0.044	0.049	0.029	0.040	0.049
beta(2,2)	30	0.061	0.058	0.053	0.056	0.069	0.056
	50	0.080	0.077	0.073	0.057	0.086	0.076
	100	0.082	0.078	0.080	0.055	0.084	0.071
	300	0.068	0.065	0.064	0.056	0.070	0.070
uniform	30	0.094	0.094	0.086	0.101	0.104	0.093
	50	0.092	0.087	0.084	0.102	0.098	0.079
	100	0.076	0.069	0.069	0.063	0.082	0.080
	300	0.070	0.071	0.069	0.069	0.074	0.074

Table 1 reports the estimated size of the selected tests at nominal significance level 0.05.

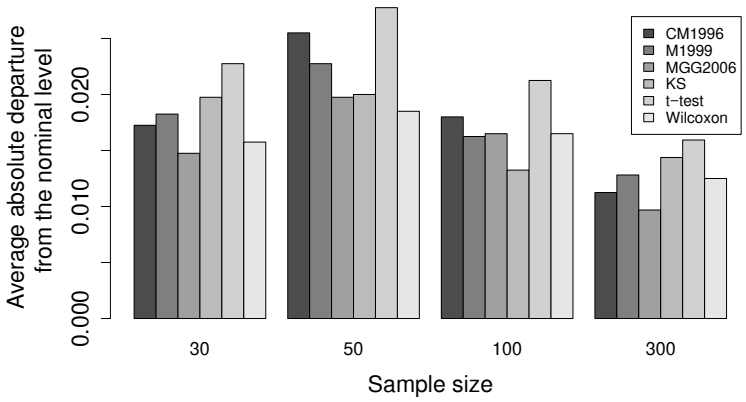

Figure 2: Average absolute departure from the nominal type I error rate.

The performance of these tests is also summarized by their average absolute departure from the nominal level 0.05 at each sample size (see Figure 2). The test of Miao, Gel and Gastwirth (2006) using a robust estimate of the standard deviation has the best performance among the tests based on the difference between $\bar{X}$ and M. In their original paper, Miao, Gel and Gastwirth (2006) reported type I error rates when the test is carried out using the asymptotic null distribution assuming a normal underlying distribution. The bootstrap method works well for normal and t_3 data. However, even in a sample size of 300, the level of all the tests deviated from 0.05 for data from a uniform distribution and only the KS test had a near 0.05 level for the Beta(2,2) distribution. Hence, further research is needed if "light-tailed" distributions occur in a particular application.

The left panel of Figure 3 displays the estimated size of the test by Miao, Gel and Gastwirth (2006) based on 10000 simulations. The broken line indicates the nominal level 0.05 and the gray dotted lines are two standard deviations away from the nominal level to allow for some perturbation due to simulation randomness. We use the letters "n", "t", "b"

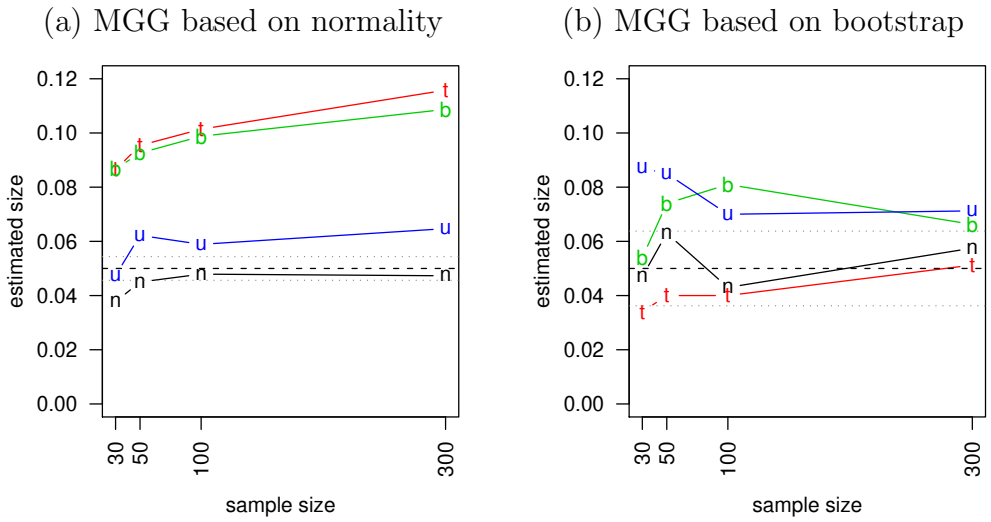

Figure 3: Comparison between the MGG2006 test using the asymptotic distribution assuming the data come from a normal distribution and the bootstrap MGG2006 test.

and "u" to label the results for the following underlying distributions: normal, t distribution with 3 df, beta(2,2) and uniform. Since the asymptotic distribution is correct when the data come from normal distributions, the Type I error is well approximated even when $n = 50$ and approaches the nominal level as the sample size increases. For the other distributions, the asymptotic variance for normal distribution used is an underestimate of the variance of the test statistic which results in an inflated type I error. Moreover the magnitude of the excess Type I error increases as the sample size increases.

The right panel in Figure 3 shows the performance of the bootstrap MGG2006 test. The computation is based on 1000 simulations, therefore the dotted lines are further away from the nominal level to indicate that results in this panel may have more uncertainty due to the smaller number of simulations. As is clear from comparing the right panel to the left one, the size of the bootstrap test approaches the nominal level for all four underlying distributions

considered, while the approximation assuming normality will not.

For tests that compare the difference between the distributions of the $(X_i - M)$'s and $(M - X_i)$'s, the test that modifies the Wilcoxon rank sum test preserves the Type I error best. With respect to preserving the nominal level, these three tests (KS, t and Wilcoxon) are roughly comparable to the tests comparing $\bar{X}$ and M.

3.3 Power comparison

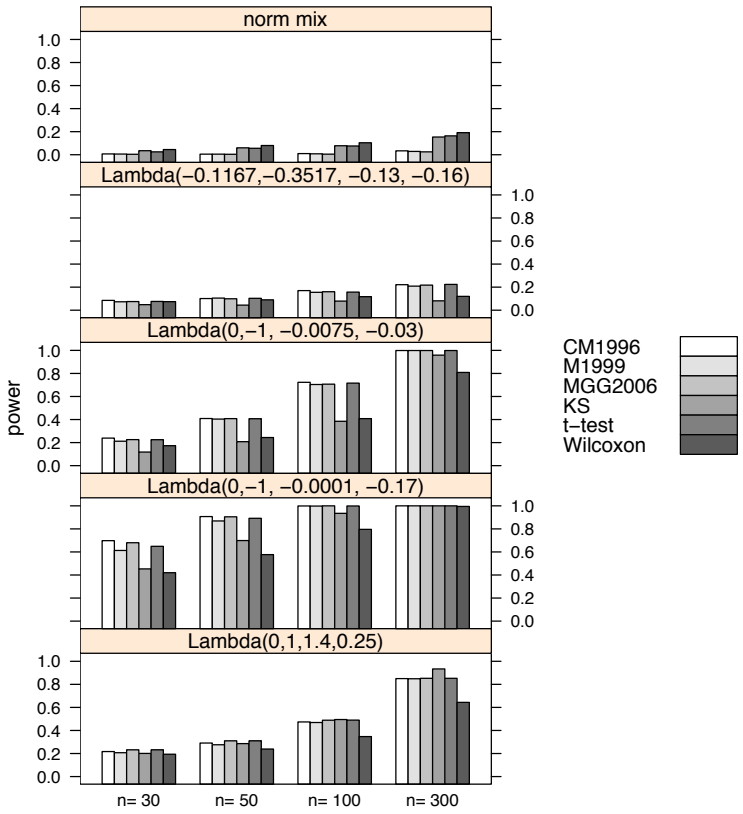

Figure 4: Power of the bootstrap tests.

Figure 4 (Table 2 provides more detail) displays the probability for these bootstrap tests to detect asymmetry of five different underlying distributions using data with different sample

Table 2: Power of the bootstrap tests. Number of simulations is 1000.

dist.	n	CM1996	M1999	MGG2006	KS	t-test	Wilcoxon
normal	30	0.007	0.006	0.004	0.035	0.025	0.045
mixture	50	0.005	0.005	0.004	0.060	0.056	0.080
	100	0.010	0.008	0.005	0.078	0.076	0.104
	300	0.034	0.029	0.025	0.154	0.164	0.191
$\lambda_1 = -0.1167$	30	0.085	0.073	0.075	0.048	0.076	0.074
$\lambda_2 = -0.3517$	50	0.101	0.105	0.099	0.044	0.103	0.090
$\lambda_3 = -0.13$	100	0.170	0.155	0.160	0.079	0.157	0.117
$\lambda_4 = -0.16$	300	0.221	0.209	0.218	0.081	0.224	0.120
$\lambda_1 = 0$	30	0.239	0.212	0.226	0.118	0.225	0.173
$\lambda_2 = -1$	50	0.409	0.404	0.408	0.208	0.407	0.244
$\lambda_3 = -0.0075$	100	0.723	0.704	0.707	0.385	0.716	0.408
$\lambda_4 = -0.03$	300	0.999	0.999	0.999	0.959	0.999	0.809
$\lambda_1 = 0$	30	0.698	0.613	0.680	0.453	0.649	0.420
$\lambda_2 = -1$	50	0.907	0.869	0.905	0.699	0.892	0.577
$\lambda_3 = -0.0001$	100	0.999	0.998	1.000	0.935	0.999	0.796
$\lambda_4 = -017$	300	1.000	1.000	1.000	1.000	1.000	0.995
$\lambda_1 = 0$	30	0.217	0.207	0.232	0.201	0.232	0.194
$\lambda_2 = 1$	50	0.291	0.276	0.310	0.286	0.310	0.239
$\lambda_3 = 1.4$	100	0.474	0.469	0.489	0.495	0.490	0.347
$\lambda_4 = 0.25$	300	0.850	0.849	0.853	0.934	0.853	0.644

sizes.

Overall, the normal mixture distribution is the hardest to detect due to its bimodality. The tests that based on the the difference between the distributions of the $(X_i - M)$'s and $(M - X_i)$'s have better power compared to the tests based on $(\bar{X} - M)$, however no test had power greater than 20% at sample size 300.

For the unimodal generalized lambda distributions, the tests based on $(\bar{X} - M)$ are generally more powerful. The performance of the t-test comparing the differences between $(X_i - M)$'s and $(M - X_i)$'s is similar to the tests based on $\bar{X} - M$. This is not surprising since the t-test statistic can be rewritten (approximately) as $(\bar{X} - M)$ normalized by a variability measure computed using the left positive departures and the right positive departures.

Among these tests based on $(\bar{X} - M)$, the test of Mira (1999) generally has slightly lower power that the CM1996 and MGG2006 tests. This agrees with the study in Efron (1979) showing that a standardized statistic is more likely to benefit from the bootstrap procedure than a non-standardized one.

In terms of power performance, the tests of Cabilio and Masaro (1996), Miao, Gel and Gastwirth (2006) and the t-test are similar and superior to the other tests. As the results in Section 3.2 demonstrate that the CM1996 (Cabilio and Masaro, 1996) and t-test produce more false positives than MGG2006 (Miao, Gel and Gastwirth, 2006), the test MGG2006 is the best choice among the tests considered here.

4 Application to education funding allocation data

This study was motivated in part by the questionable use of symmetric trimming of the per-pupil expenditure data for a state's school districts when calculating a measure of relative disparity used in the allocation of federal funds. The tests of symmetry will now be used to examine several educational funding data sets. The first two, from New Mexico and Alaska, actually were submitted to the U.S. Supreme Court in the *Zuni School District* case but were hardly mentioned in the decision. The third data set refers to the 1967 per-pupil expenditures on teachers in the elementary schools in Washington D.C. The data were used in court from the classic school segregation case, *Hobson v. Hansen*. As the three previous data sets were of small to moderate size, the final data set reports the per-pupil expenditures in Missouri's 522 school districts during the 2001-2002 academic year.

The first data set reports revenues available per-pupil in the 89 Educational Agencies or

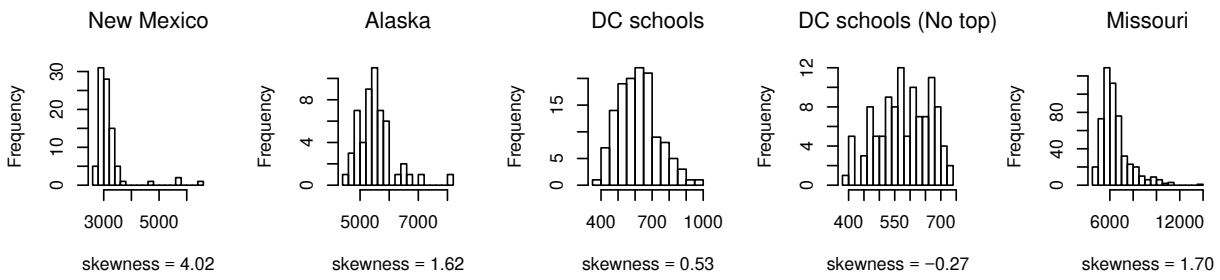

Figure 5: Histograms of the educational funding data sets. The measure of skewness γ_1 (Stuart and Ord, 1994, page 109) is also included.

school districts in New Mexico and is reproduced in Gastwirth (2006) and available in the R-package, `lawstat`. The results in Table 3 below show that the CM, Mira, MGG and t-tests all reject the hypothesis of symmetry with p-values in the neighborhood of .01-.02. These results are consistent with the right skewness of the data seen in the histogram in Figure 5. The modified KS and rank-sum tests, however, did not detect the lack of symmetry of the data. This highlights the importance of power study in Section 3.3 as those two tests were generally less powerful than the others.

The second data set reporting the per-pupil expenditures of the 44 Educational Agencies in Alaska was submitted to the Supreme Court by the state of Alaska in its amicus brief to the Court in the *Zuni School District* case. All the tests accepted the symmetry hypothesis, although the histogram suggests it might be right skewed. The small sample size ($n = 44$) naturally diminishes the power of any statistical test. All the tests detected a similar degree of skewness in the much larger data set from Missouri ($n = 522$).

The data from the *Hobson v. Hansen* case is given in Figure 5 (DC schools). Here the histogram looks more symmetric with a small right-skew (which was primarily due to the

Table 3: P-values from bootstrap tests of symmetry on education fund allocation data. P-values are estimated based on 10000 bootstraps.

	New Mexico	Alaska	DC	DC (no top)	Missouri
MGG2006	0.0077	0.474	0.292	0.761	< 0.0001
CM1996	0.0083	0.472	0.297	0.759	< 0.0001
Mira1999	0.0192	0.471	0.291	0.757	< 0.0001
KS	0.2881	0.543	0.628	0.596	< 0.0001
Wilcoxon	0.1608	0.574	0.643	0.494	< 0.0001
t-test	0.0103	0.495	0.281	0.756	< 0.0001

thirteen predominantly white schools). All tests of symmetry accept that hypothesis. It is also worth noting that the p-values of the t-test and the three tests based on the difference between the mean and median of the data were quite close. This is consistent with the findings of the simulation study reported in Sections 3. The subset of 110 majority black elementary schools was considered by Gastwirth (2008) in order to eliminate the effect of segregation (DC schools (No top) in Figure 5). The p-values of all the tests on this subset are larger than on the entire data set, implying that the restricted data are more symmetric than the entire data set. This is consistent with the histograms in Figure 5 and the fact that at the time of the case predominantly white schools were better funded than schools with larger minority proportions.

To explore the properties of the tests in a larger sample, data for the 522 districts in Missouri were examined. The histogram of the Missouri data in Figure 5, the data of Missouri is very right-skewed. All tests considered here estimated the p-value to be < 0.0001 when $10,000$ bootstrap samples were taken.

5 Discussion and conclusion

This paper shows that the sampling distribution of several tests of symmetry can be estimated using the bootstrap when re-samples are taken from the symmetrized empirical CDF about the sample median.

For commonly occurring sample sizes, the method provides more accurate distributions for these intuitive tests of symmetry based on the difference between sample mean and median than a previously used asymptotic approximation. The bootstrap symmetry tests preserved the nominal level, especially for the heavy-tailed t_3 distribution, better than the previously used asymptotic procedure. Since the nominal level is preserved, the comparative powers of the various symmetry tests can be evaluated more reliably. The results showed that the test of Miao, Gel and Gastwirth (2006) (MGG2006) is a reasonable overall test of symmetry about an unknown center, especially when the underlying distribution is not "light-tailed." The bootstrap method can sometimes give unreliable estimates of probability of rejection if the statistic of interest is not asymptotically pivotal. In that case, recent research on improving the reliability of bootstrap tests (e.g. Davidson and Mackinnon, 2007) can be applied to extend the approach used in this paper.

The methods are applied to the actual data submitted to the U.S. Supreme Court in an educational funding case. Although the U.S. Department of Education used a method of trimming developed for symmetric data, often the data are not symmetric. All the tests of symmetry with reasonable power rejected the null hypothesis that the main data set in the legal case came from a symmetric distribution. This result raises an important question concerning the "outlier" deletion procedure specified in the Federal Impact Aid Act. Since

Chief Justice Roberts expressed concern about the lack of discussion of outliers by the lawyers in the *Zuni School District 89 v. U.S. Department of Education* case, hopefully the legal community will pay more attention to statistical issues in the future. The development of statistical methods based on intuitive understandable statistic measures should assist courts in assessing the meaning and implications of statistical evidence.

6 Acknowledgement

The research of Professor Zheng was supported in part by grant R01 GM070789 from the National Institutes of Health and the research of Professor Gastwirth was supported in part by grant SES-0317956 from the National Science Foundation.

References

Bhattacharya, P. K., Gastwirth, J. L. and Wright, A. L. (1982). Two modified Wilcoxon tests for symmetry about an unknown location parameter. *Biometrika*, **69** 377–382.

Bonferroni, C. E. (1930). Elementi di statistica generale. *Seeber, Firenze*.

Boos, D. D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator. *Journal of the American Statistical Association*, **77** 647–651.

Butler, C. C. (1969). A test for symmetry using the sample distribution function. *The Annals of Mathematical Statistics*, **40** 2209–2210.

Cabilio, P. and Masaro, J. (1996). A simple test of symmetry about an unknown median. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, **24** 349–361.

Clarke, B. R., Gamble, D. K. and Bednarski, T. (2000). A note on robustness of the-trimmed mean. *Australian & New Zealand Journal of Statistics*, **42** 113–117.

Collins, J. R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *The Annals of Statistics*, **4** 68–85.

Davidson, R. and Mackinnon, J. G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics & Data Analysis*, **51** 3259–3281.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7** 1–26.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans, monograph 38. *Philadelphia: SIAM*.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York, NY.

Gastwirth, J. L. (1971). Sign test for symmetry. *Journal of the American Statistical Association*, **66** 821–828.

Gastwirth, J. L. (2006). A 60 million dollar statistical issue arising in the interpretation and calculation of a measure of relative disparity: Zuni Public School District 89 v. U.S. Department of Education. *Law Probablity and Risk*, **5** 33–61.

Gastwirth, J. L. (2008). The U.S. Supreme Court finds a statutes description of a simple statistical measure of relative disparity ambiguous allowing the Secretary of Education to interpret the formula: *Zuni Public School District 89 v. U.S. Department of Education II. Law Probablity and Risk*, **7** 225–248.

Gupta, M. K. (1967). An asymptotically nonparametric test of symmetry. *The Annals of Mathematical Statistics*, **38** 849–866.

Hinkley, D. (1976). On estimating a symmetric distribution.

Hollander, M. (2006). Testing for symmetry. *Encyclopedia of Statistical Sciences*.

Hotelling, H. and Solomon, L. M. (1932). The limits of a measure of skewness. *Annals of Mathematical Statistics*, **3** 141–142.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

Miao, W., Gel, Y. R. and Gastwirth, J. L. (2006). A new test of symmetry about an unknown median. In *Random Walk, Sequential Analysis and Related Topics—A Festschrift in Honor of Yuan-Shih Chow* (A. Hsiung, C.-H. Zhang and Z. Ying, eds.). World Scientific, Singapore.

Mira, A. (1999). Distribution-free test for symmetry based on Bonferroni's measure. *Journal of Applied Statistics*, **26** 959–972.

Modarres, R. (2002). Efficient nonparametric estimation of a distribution function. *Computational Statistics & Data Analysis*, **39** 75–95.

Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, **8** 154–168.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Commun. ACM*, **17** 78–82.

Randles, R. H., Fligner, M. A., Ii, G. E. P. and Wolfe, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, **75** 168–172.

Schuster, E. F. and Barker, R. C. (1987). Using the bootstrap in testing symmetry versus asymmetry. *Communications in Statistics-Simulation and Computation*, **16** 69–84.

Smirnov, N. V. (1947). On criteria for the symmetry of distribution laws of random variables. *Rossiiskaya Akademiya Nauk*, **56** 13–16.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*. John Wiley, New York, NY.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, **80** 1026–1031.

Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics, Vol I: Distribution Theorem*. 6th ed. Oxford Univ Press.