SFB
823

Discussion Paper

Robust nonparametric tests for the two-sample location problem

Roland Fried, Herold Dehling

Nr. 19/2011

Robust nonparametric tests for the two-sample location problem

Roland Fried[1] **and Herold Dehling**[2]

Abstract We construct and investigate robust nonparametric tests for the two-sample location problem. A test based on a suitable scaling of the median of the set of differences between the two samples, which is the Hodges-Lehmann shift estimator corresponding to the Wilcoxon two-sample rank test, leads to higher robustness against outliers than the Wilcoxon test itself, while preserving its efficiency under a broad range of distributions. The good performance of the constructed test is investigated under different distributions and outlier configurations and compared to alternatives like the two-sample t-, the Wilcoxon and the median test, as well as to tests based on the difference of the sample medians or the one-sample Hodges-Lehmann estimators.

Keywords Distribution-free tests, Wilcoxon test, Outliers, Heavy tails, Skewed distributions.

[1] Corresponding author. Department of Statistics, University of Dortmund, 44221 Dortmund, Germany. e-mail: fried@statistik.tu-dortmund.de. Tel: +49 231 755 3119. Fax: +49 231 755 3454.

[2] Department of Mathematics, Ruhr-University of Bochum, 44780 Bochum, Germany

1 Introduction

Consider the classical two-sample problem with independent observations,

$$\begin{aligned} X_1, \ldots, X_m &\sim F \\ Y_1, \ldots, Y_n &\sim G, \end{aligned}$$

where F and G are distribution functions corresponding to continuous distributions. We focus on the situation where G is a shifted version of F, $G(x) = F(x - \Delta)$ for all $x \in \mathbb{R}$ and an unknown $\Delta \in \mathbb{R}$. We denote the density of F by f as usual.

The two-sample location problem has received considerable attention in the past. The robustness of tests of the null hypothesis of equality of F and G, which can be expressed as $H_0 : \Delta = 0$ in our context, against violations of the assumptions is still under discussion. The two-sample t-test is sometimes regarded as robust against deviations from normality, because the central limit theorem guarantees its asymptotic validity, if the common variance of F and G exists. Reed and Stark (2004) verify the usefulness of this approximation in case of sample sizes between 10 and 40. For a discussion of weaknesses of the t-test see e.g. Wilcox and Keselman (2003).

The most prominent nonparametric competitors to the t-test are the Wilcoxon two-sample rank sum test and the median test. The Wilcoxon test rejects the null hypothesis H_0 if the sum W of the ranks $R_1, \ldots, R_n$ of $Y_1, \ldots, Y_n$ in the sample of all observations $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is too large or too small. Critical values are determined by permutational arguments since all possible $\binom{m+n}{n}$ assignments of the ordered ranks $R_{(1)} < \ldots < R_{(n)}$ are equiprobable under H_0. As opposed to this, the median test uses the number of values in $Y_1, \ldots, Y_n$ larger than the global median of all $m + n$ observations, which follows a hypergeometric distribution under H_0. Randomization can be applied to achieve exact significance levels.

A general approach to the construction of tests is standardization of an estimator of Δ and rejection of H_0 if this test statistic is too far from zero. The two-sample t-test is derived from this idea. Alternatively, we can replace the sample mean used in the t-test by a robust estimator like the sample median, as recommended e.g. by Bovik and Munson (1986) and Fried (2007). The resulting estimator is the difference between the medians of the two samples,

$$\hat{\Delta}^{(3)}_{m,n} = \tilde{Y}_n - \tilde{X}_m = \text{med}\{Y_1, \ldots, Y_n\} - \text{med}\{X_1, \ldots, X_m\}.$$

Following Lehmann (1963b), we construct tests of H_0 using the Hodges-Lehmann two-sample estimator $\hat{\Delta}^{(2)}_{m,n}$ of Δ (Hodges and Lehmann 1963). It corresponds to that value which we need to subtract from $Y_1, \ldots, Y_n$ to align the samples, meaning that the Wilcoxon rank sum statistic of the aligned samples becomes equal to its expected value under H_0, which is $n(m+n+1)/2$. For calculations we have the formula

$$\hat{\Delta}^{(2)}_{m,n} = \mathrm{med}\{Y_j - X_i, i = 1, \ldots, m, j = 1, \ldots, n\} \ .$$

$\hat{\Delta}^{(2)}_{m,n}$ is symmetrically distributed about the location difference Δ whenever the underlying distribution F is symmetric, or if the sample sizes m and n are equal. As opposed to its normal theory competitor, the difference of the sample means,

$$\hat{\Delta}^{(0)}_{m,n} = \bar{Y}_n - \bar{X}_m,$$

$\hat{\Delta}^{(2)}_{m,n}$ cannot be expressed as a difference of a statistic based on $Y_1, \ldots, Y_n$ and a statistic based on $X_1, \ldots, X_m$, and it is less affected by outliers.

In case of symmetric distributions, Lehmann (1963a) proposes estimation of Δ by

$$\hat{\Delta}^{(1)}_{m,n} = \hat{Y}_n - \hat{X}_m \ .$$

Here, $\hat{X}_m$ and $\hat{Y}_n$ are the Hodges-Lehmann one-sample location estimators (Hodges and Lehmann 1963) for the center of the distribution of the X and the Y sample, respectively, obtained from the signed rank test for hypothesis about the median of a single sample. $\hat{X}_m$ and $\hat{Y}_n$ can be calculated as the median of $\{(X_i + X_j)/2, 1 \leq i < j \leq m\}$ and $\{(Y_i + Y_j)/2, 1 \leq i < j \leq n\}$, respectively.

If F is symmetric, $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$ have the same asymptotic relative efficiency as compared to $\hat{\Delta}^{(0)}_{m,n}$, which equals the Pitman efficiency of the two-sample Wilcoxon test relative to the two-sample t-test, namely $12\sigma^2[\int f^2(x)dx]^2$, where σ^2 is the variance of F, see Lehmann (1963a). This asymptotic efficiency becomes $3/\pi \approx 0.955$ at the normal distribution, it never drops down below 86.4% and it can become arbitrarily high. Hoyland (1965) conjectures that $\hat{\Delta}^{(2)}_{m,n}$ is to be recommended in case of heavy tailed distributions. Moreover, he shows for the case of shifted asymmetric distributions $F = G(\cdot - \Delta)$ that the asymptotic efficiency of $\hat{\Delta}^{(2)}_{m,n}$ is always larger than that of $\hat{\Delta}^{(1)}_{m,n}$, and it can even become arbitrarily large. The asymptotic efficiency of $\hat{\Delta}^{(1)}_{m,n}$ relative to $\hat{\Delta}^{(0)}_{m,n}$ is $\sigma^2[\int f(x)f(-x)dx]^2/[\int F^2(-x)f(x)dx - 0.25]$.

Section 2 constructs tests for $H_0 : \Delta = 0$ based on the robust estimators $\hat{\Delta}^{(j)}_{m,n}$, $j = 1, 2, 3$, as alternatives to the two-sample t-test, Wilcoxon test and median test. Section 3 compares small sample versions of the tests, which are based on permutational arguments, in a simulation study. Section 4 studies large sample versions of the tests, which are based on the asymptotic distributions of the test statistics. Section 5 concludes.

2 Tests for the two-sample location problem

We want to test $H_0 : \Delta = 0$ against the alternative $H_1 : \Delta \neq 0$. The classical procedure for this testing problem under the assumption that F and G are shifted normal distributions is the two-sample t-test. It is obtained standardizing $\hat{\Delta}^{(0)}_{m,n}$ by the pooled sample variance $\hat{S}^2 = \frac{1}{m+n-2}\Big[\sum_{i=1}^{m}(X_i - \bar{X}_m)^2 + \sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2\Big]$. Among its drawbacks are its reliance on the normality assumption in small samples and the possibly large loss of power caused already by a few outliers even in rather large samples, see e.g. Fried (2007). Therefore it seems worthwhile to investigate alternatives. Generalizing the idea underlying the two-sample t-test, it is intuitive to reject H_0 if a robust estimator of Δ like $\hat{\Delta}^{(j)}_{m,n}$, $j \in \{1, 2, 3\}$, is far away from zero, scaling it by an estimate of the variability.

2.1 Tests based on the Hodges-Lehmann estimators

For scaling the Hodges-Lehmann two-sample estimator (HLE2) $\hat{\Delta}^{(2)}_{m,n}$ (or the difference between the one-sample estimators HLE, $\hat{\Delta}^{(1)}_{m,n}$) we look for an adequate robust estimator of the variability in the data. Whereas $\hat{\Delta}^{(2)}_{m,n}$ measures the variability between the two samples, a related measure of the variability within the two samples is the median of the absolute set of differences in the samples,

$$S^{(1)}_{m,n} = \text{med}\{|X_i - X_j| : 1 \leq i < j \leq m, |Y_i - Y_j| : 1 \leq i < j \leq n\} .$$

Another measure of the variability within the samples is the median of the absolute set of differences within the joint median-corrected sample,

$$S^{(2)}_{m,n} = \text{med}\{|Z_i - Z_j| : 1 \leq i < j \leq m + n\},$$

where $(Z_1, \ldots, Z_{m+n})' = (X_1 - \tilde{X}_m, \ldots, X_m - \tilde{X}_m, Y_1 - \tilde{Y}_n, \ldots, Y_n - \tilde{Y}_n)'$, and $\tilde{X}_m = \text{med}\{X_1, \ldots, X_m\}$ and $\tilde{Y}_n = \text{med}\{Y_1, \ldots, Y_n\}$ are the respective sample medians.

The distribution of $T^{(k,l)}_{m,n} = \hat{\Delta}^{(k)}_{m,n}/S^{(l)}_{m,n}$, $k, l \in \{1, 2\}$, is unknown in finite samples, but distribution-free tests can be constructed by deriving critical values for the test statistics using the permutation principle: we split the total $N = m + n$ observations repeatedly into two groups of m and n observations, and calculate the absolute value of the test statistic for each of the permuted samples in two-sided testing. We reject the null hypothesis if $|T^{(k,l)}_{m,n}|$ is among the largest 100α percent of these values.

If both m and n are small, we consider all possible splits. Under H_0, since all observations are exchangeable then, the rank of the observed value $|T^{(k,l)}_{m,n}|$ among all random splits has a discrete uniform distribution. An exact p-value is thus obtained as $p = A/B$, where A is the number of splits leading to absolute test statistics at least as large as $|T_{m,n}^{(k,l)}|$, and $B = \binom{N}{m}$ is the total number of splits. Randomization can be applied to achieve exact significance levels, similarly as for the Wilcoxon test. However, discreteness of the test statistic poses less severe problems than for the latter, since different splits will almost surely lead to different values of the test statistic. The test based on comparing p to the significance level α without further randomization will be only a little conservative except if m and n are very small.

If not both m and n are small we suggest not to consider all possible splits, and generate a large number b of random splits instead, including additionally the observed 'true' split. Already in case of $m = n = 10$ there are $\binom{20}{10} = 184756$ possible splits, so that evaluation of all of these would lead to large computation times because of the computational needs for the median. Under H_0, the rank of $|T_{m,n}^{(k,l)}|$ within the total $b + 1$ splits considered follows again a discrete uniform distribution, so that we can obtain an exact p-value as $\hat{p} = (a+1)/(b+1)$, where a is the number of randomly selected splits leading to absolute test statistics at least as large as the observed one (e.g. Edgington, 1995, p. 41f). $\hat{p}$ can be seen as a slightly positively biased estimate for the p-value p arising from all possible splits. In our implementation, we restrict the total number of splits to at most 10000 if $\alpha = 0.05$, so that the standard error of $\hat{p}$ is about 0.0022 if $p \approx 0.05$ and $b = 10000$.

If m and n are large, we can construct asymptotical tests based on $\hat{\Delta}_{m,n}^{(2)}$, see Lehmann (1963c). If $m, n \to \infty$ with $m/N \to \lambda \in (0, 1)$, $N = m + n$, then we have under H_0

$$\sqrt{12\lambda(1-\lambda)} \int f^2(x)dx \ \sqrt{N}\hat{\Delta}_{m,n}^{(2)} \overset{a,H_0}{\sim} N(0,1) \ . \tag{1}$$

This leads us to an asymptotically $N(0, 1)$ distributed test statistic under the null hypothesis $F = G$ if we plug in a consistent estimator of the value of the density h of $X - Y$ at 0, $h(0) = \int f^2(x)dx$. It is obvious from the above asymptotics that the resulting test has the same relative asymptotic efficiency of $12\sigma^2[\int f^2(x)dx]^2$ relatively to the two-sample t-test as the Wilcoxon test, see also Lehmann (1963c).

For estimation of $h(0)$ we apply a kernel density estimator to the set of all pairwise differences within the two samples, using $X_2 - X_1, \ldots, X_m - X_1, \ldots, X_m - X_{m-1}, Y_2 - Y_1, \ldots, Y_n - Y_{n-1}$ (i.e. $m(m-1)/2 + n(n-1)/2$ differences altogether). We could also use the set of all pairwise differences within the full sample consisting of $m + n$ observations instead, possibly correcting by the median within each sample, but this did not lead to generally better results in our simulations. Note that the differences, which our kernel density estimator is based on, overlap and are thus not independent. This does not impose problems with consistency because the resulting estimator can be written as a U-statistic, which are consistent under general conditions.

2.2 Tests based on the difference of medians

Another possibility is to construct tests from the difference between the medians of the samples. If $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent and the density f is continuous and strictly positive at the median $F^{-1}(0.5)$, then we have asymptotically

$$\begin{aligned} \sqrt{n}(\mathrm{med}\{Y_1, \ldots, Y_n\} - G^{-1}(0.5)) &\overset{a}{\sim} N\left(0, \frac{1}{4f^2(F^{-1}(0.5))}\right) \text{ and} \\ \sqrt{m}(\mathrm{med}\{X_1, \ldots, X_m\} - F^{-1}(0.5)) &\overset{a}{\sim} N\left(0, \frac{1}{4f^2(F^{-1}(0.5))},\right) \end{aligned}$$

see Serfling (1980, p. 77). Since $\mathrm{med}\{X_1, \ldots, X_m\}$ and $\mathrm{med}\{Y_1, \ldots, Y_n\}$ are independent, an asymptotically standard normal random variable under the null hypothesis $F = G$ then is

$$\sqrt{\frac{mn}{m+n}} 2f\left(F^{-1}(0.5)\right) \hat{\Delta}^{(3)}_{m,n} \overset{H_0,a}{\sim} N(0,1) \; .$$

To construct tests from $\hat{\Delta}^{(3)}_{m,n}$, we estimate $f(F^{-1}(0.5))$ applying a kernel density estimator to the combined median-corrected sample $X_1 - \tilde{X}_m, \ldots, X_m - \tilde{X}_m, Y_1 - \tilde{Y}_n, \ldots, Y_n - \tilde{Y}_n$.

In case of small sample sizes, we again apply the idea of permutation tests and derive critical values from all splits of the joint sample, or from a random selection of them if there are too many. Natural choices for standardization of $\hat{\Delta}^{(3)}_{m,n}$ are $S^{(3)}_{m,n} = 2\mathrm{med}(|X_1 - \tilde{X}_m|, \ldots, |X_m - \tilde{X}_m|, |Y_1 - \tilde{Y}_n|, \ldots, |Y_n - \tilde{Y}_n|)$, or the sum of the median absolute deviations about the respective median (MADs) in the two samples. Fried (2007) also designs tests for shift detection based on differences of sample medians scaled by robust estimators of variability like the MAD, but under the assumption of observing normal distributions contaminated by outliers. As opposed to this, we aim at tests which are both robust and nonparametric and work under mild assumptions. The permutation tests based on $\hat{\Delta}^{(i)}_{m,n}$, $i = 1, 2, 3$, constructed here are distribution free under the null hypothesis, or at least approximately so if not all possible splits or the asymptotical versions are used. In the next section we will investigate the power of these tests under different scenarios.

3 Performance of the tests in small samples

First we investigate the small sample versions of the tests, which use critical values derived from the permutation principle. We consider sample sizes $m, n \in \{5, 10\}$ and generate 1000 pairs of samples for each of different data situations to study the performance of the tests under different conditions. The power of the tests is approximated by calculating the frequency of cases for which the null hypothesis is rejected among the 1000 repetitions for each combination of error distribution and location difference. The resulting power curves are smoothed a little by a weighted moving average with weights (0.25,0.5,0.25). To compare the results for different distributions, we choose the sizes of the location differences as multiples of the difference between the 84.13% and the 50% percentile of the distribution, which is 1 in case of the standard normal. We report the results for $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$ scaled by $S^{(2)}_{m,n}$ as this gave slightly larger powers than scaling by $S^{(1)}_{m,n}$. For $\hat{\Delta}^{(2)}_{m,n}$ we use $S^{(3)}_{m,n}$, since the test with standardization by the sum of the MADs of the two samples resulted in more rejections under H_0 than indicated by the nominal significance level in case of unequal sample sizes, i.e. it was found to be oversized (anti-conservative) in this situation.

Figure 1 depicts the simulation results for $m = n = 10$. We use randomized versions of the Wilcoxon and the median test, which keep the chosen significance level $\alpha = .05$ exactly in case of clean samples from the same distributions. The randomization tests based on the estimates of the location difference $\hat{\Delta}^{(k)}_{m,n}$, $k = 0, 1, 2, 3$, are almost exact. In case of shifted normal distributions, the two-sample t-test of course offers the largest power, closely followed by the Wilcoxon test and the tests based on $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$. The test based on $\hat{\Delta}^{(3)}_{m,n}$ is worse but offers larger power than the median test here.

The t-test loses its superiority in case of the heavy-tailed t_3-distribution and performs not much better than the median test then. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$ show the largest power and outperform the Wilcoxon test and the one based on $\hat{\Delta}^{(3)}_{m,n}$. The randomization test based on the t-statistic (denoted by mean diff.) provides relatively large power against a small location difference Δ, but is outperformed if Δ is large. In case of the t_1-distribution, $\hat{\Delta}^{(3)}_{m,n}$ leads to the most powerful test, followed by $\hat{\Delta}^{(2)}_{m,n}$, $\hat{\Delta}^{(1)}_{m,n}$ and the median test. The Wilcoxon test is worse than these under these conditions. The randomization test based on $\hat{\Delta}^{(0)}_{m,n}$ performs better than the t-test then, which becomes rather conservative, but considerably worse than all the other tests considered here. In case of the skewed χ^2_3-distribution, the randomization tests based on $\hat{\Delta}^{(2)}_{m,n}$ or $\hat{\Delta}^{(0)}_{m,n}$ achieve the largest power.

The two-sample t-test and the randomization test based on $\hat{\Delta}^{(0)}_{m,n}$ lose all their power because of a single large outlier in small samples, and the Wilcoxon test is also affected considerably, although much less than the others. As opposed to this, the robust estimators $\hat{\Delta}^{(1)}_{m,n}$, $\hat{\Delta}^{(2)}_{m,n}$ and $\hat{\Delta}^{(3)}_{m,n}$ offer better protection against outliers in small samples, with $\hat{\Delta}^{(2)}_{m,n}$ offering the largest power.

We got results very similar to those before also for $m = 10$ and $n = 5$ (Figure 2) and $m = n = 5$ (not shown here). A main difference is that a single large outlier can disguise even a rather large shift when using the Wilcoxon or the median test in case of small samples, i.e. the advantages of the tests based on the robust estimators $\hat{\Delta}^{(j)}_{m,n}$, $j = 1, 2, 3$, increase.

4 Performance of the asymptotical tests

In order to compare the performance of the asymptotic versions of the tests, we first inspect their sizes in case of different sample sizes $m = n \in \{3, 6, \ldots, 75\}$ and different distributions. We generate 10000 data sets for each setting and derive the empirical rejection rates under the null hypothesis. The normal, t_3, t_2, t_1, χ^2_1 and two contaminated normal distributions $(1-\epsilon)N(0,1)+\epsilon N(0,25)$ with $\epsilon \in \{0.05, 0.2\}$ are considered here. Besides the two-sample t, the Wilcoxon and the median test we include the tests based on $\hat{\Delta}^{(j)}_{m,n}$, $j = 1, 2, 3$, scaled by a kernel density estimate, as described in Sections 2.1 and 2.2. For kernel density estimation we use the function 'density' of R (R Development Core Team, 2009), version 2.9.2, with the default Gaussian kernel and bandwidth.

Figure 3 indicates that a sample size of $m = n = 30$ observations is sufficient for the asymptotical critical values leading to approximately valid statistical tests at a 5% significance level if the underlying distributions are normal, contaminated normal or t-distributions with at least two degrees of freedom. The tests based on $\hat{\Delta}^{(3)}_{m,n}$ are oversized up to $m = n = 20$, and those using $\hat{\Delta}^{(2)}_{m,n}$ up to $m = n = 12$ in these situations. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ perform similarly to those using $\hat{\Delta}^{(2)}_{m,n}$ except for the 20% contaminated normal, for which $\hat{\Delta}^{(1)}_{m,n}$ scaled by the kernel density estimator becomes liberal. The t-test and the test based on $\hat{\Delta}^{(3)}_{m,n}$ become conservative in case of large samples from heavy-tailed distributions like the t_1 and t_2.

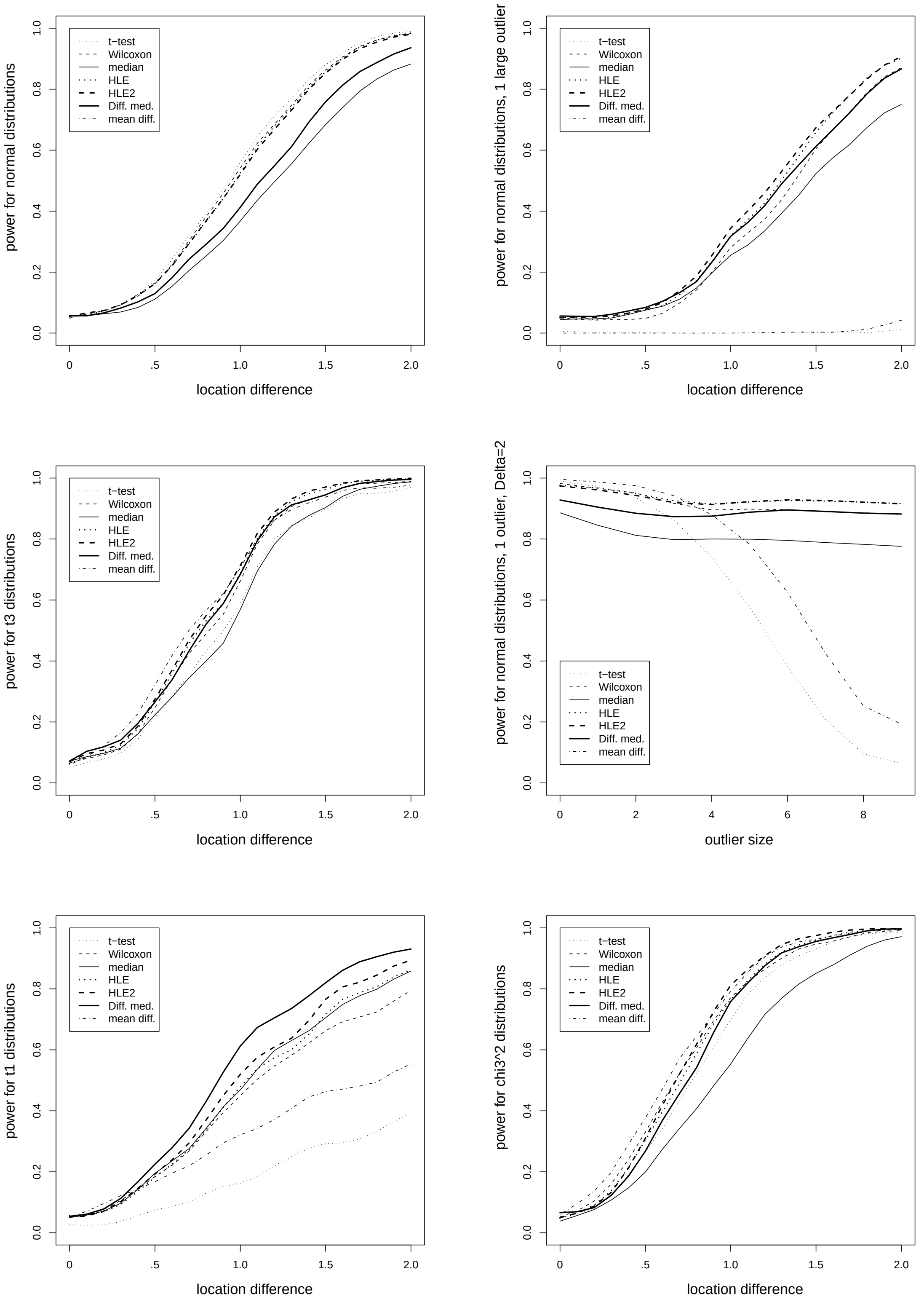

Figure 1: Power of the small sample tests as a function of Δ, $m = n = 10$. Left: Normal (top), t_3 (center), and t_1 distribution (bottom). Right: Normal with 1 outlier of size 10 (top) or 1 outlier of increasing size, $\Delta = 2$ (center), χ^2_3 distribution (bottom). t-test (dotted), Wilcoxon (dashed), median test (solid), HLE $\hat{\Delta}^{(1)}_{m,n}$ (bold dotted), HLE2 $\hat{\Delta}^{(2)}_{m,n}$ (bold dashed), $\hat{\Delta}^{(3)}_{m,n}$ (bold solid) and mean diff. (dash-dot).

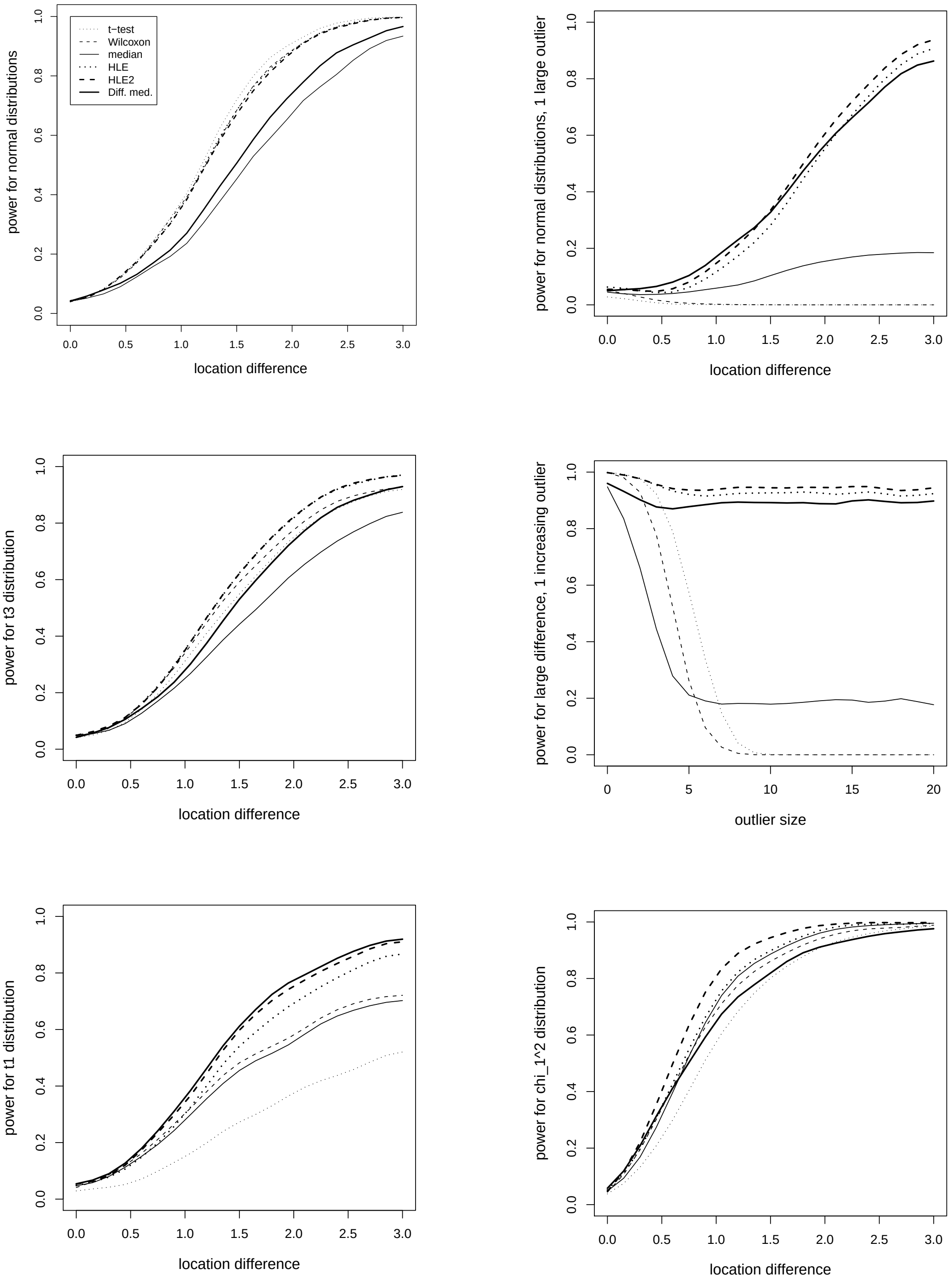

Figure 2: Power of the tests in small samples as a function of Δ, $m = 10$, $n = 5$. Left: Normal (top), t_3 (center), and t_1 distribution (bottom). Right: Normal with an outlier of size 10 (top), with an outlier of increasing size and $\Delta = 3$ (center), χ_1^2 distribution (bottom). t-test (dotted), Wilcoxon test (dashed), median test (solid), $\hat{\Delta}_{m,n}^{(1)}$ (bold dotted), $\hat{\Delta}_{m,n}^{(2)}$ (bold dashed) and $\hat{\Delta}_{m,n}^{(3)}$ (bold solid).

In case of the very heavy-tailed t_1-distribution, $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$ scaled by the density estimate lead to severe violations of the significance level. This phenomenon seems strange given that the kernel density estimator should estimate the density at 0 consistently given that the derivatives of the density exist and are bounded, cf. Simonoff (1996, p. 42). Varying the bandwidth by changing the adjustment factor improves the results, but different factors are needed for different sample sizes. We report only the results for the default bandwidth since they correspond to routine and automatic application of the tests. In case of the very right-skewed χ^2_1-distribution, $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(3)}_{m,n}$ lead to largely oversized tests.

We also inspected situations where one of the samples is twice the size of the other one. A total sample size of $m + n = 60$ observations is again sufficient for the asymptotical critical values leading to approximately valid tests at a 5% significance level for most of the distributions considered here. The χ^2_1-distribution affords a total of about 80 observations for the test based on $\hat{\Delta}^{(2)}_{m,n}$ then. The two-sample t-test is slightly oversized in this situation. The results when one of the two samples stems from a contaminated normal were very different for contamination in the larger and in the smaller sample: if the larger sample was contaminated, most of the tests except the median test were conservative, particularly the two-sample t-test and the tests based on $\hat{\Delta}^{(1)}_{m,n}$ or $\hat{\Delta}^{(2)}_{m,n}$. If the smaller sample was contaminated, many of the tests became anti-conservative, except the median test and the tests based on $\hat{\Delta}^{(3)}_{m,n}$.

Next we investigate the power of the asymptotical tests, generating 1000 samples for each of several data situations and different sample sizes. Figure 4 illustrates the results for $m = n = 50$ and location differences $j \cdot 0.04 \cdot (F_{0.841} - F_{0.5})$, $j = 1, 2, \ldots, 25$, where F_p denotes the $100p\%$ percentile of the underlying distribution. In case of normal distributions, the t-test is of course the most powerful procedure, followed by the Wilcoxon. There is a substantial gap to the median test. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ or $\hat{\Delta}^{(2)}_{m,n}$ offer about the same power as the Wilcoxon test, and the same applies to $\hat{\Delta}^{(3)}_{m,n}$ with respect to the median test.

The findings are almost the same, if there is one outlier of size 10 in one of the samples, since all tests are robust against a single outlier in moderately large samples, except for the two-sample t-test, which loses a lot of power and becomes the worst test then (not shown here). To challenge the methods more, we consider situations with a location difference $\Delta = 2$ and additional $N(2k, k - 1)$-distributed noise added

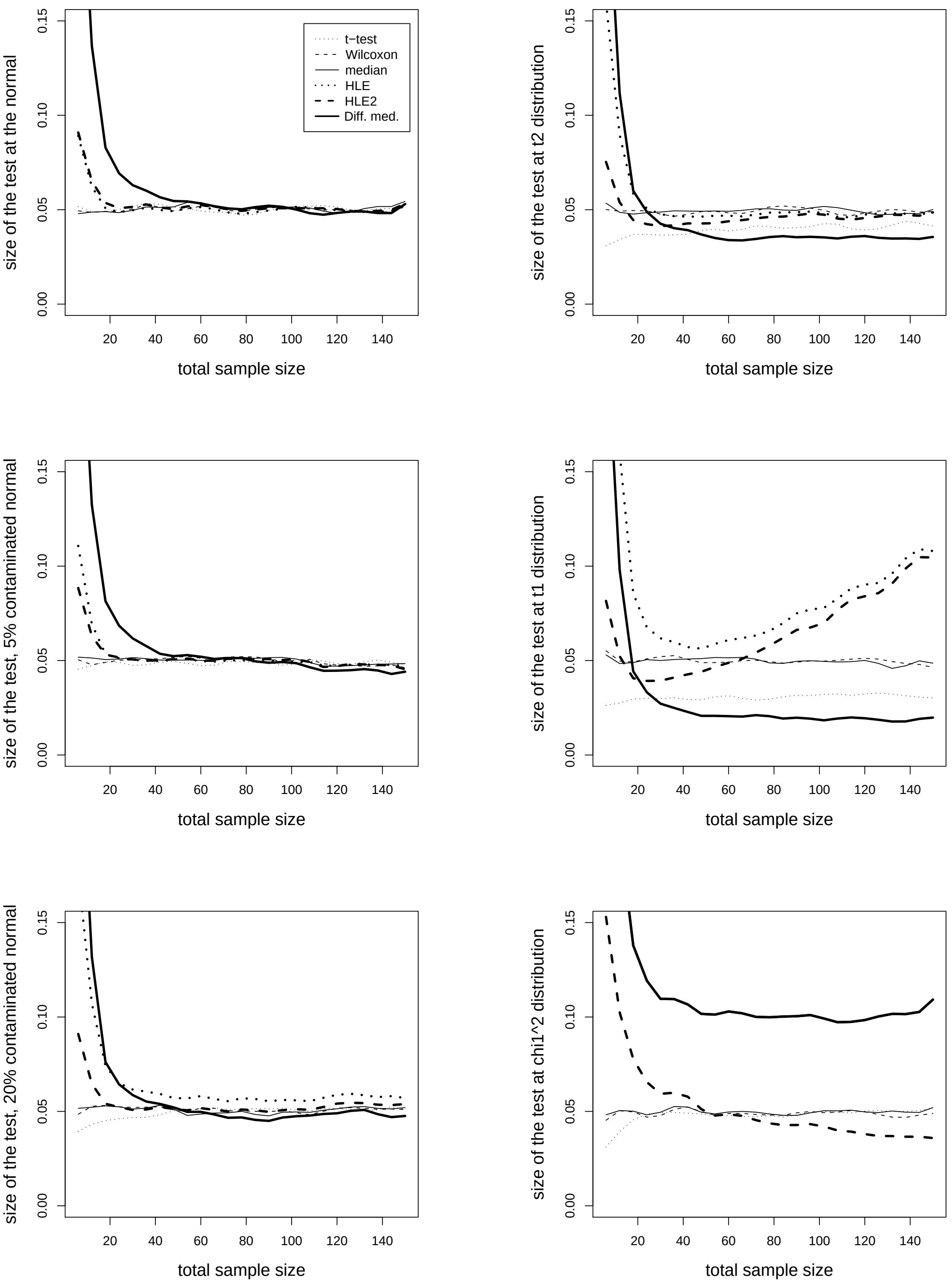

Figure 3: Sizes of the asymptotical tests as a function of the total sample size $m+n$ for different distributions, $m = n$. Left: normal (top), normal with 5% (center) or 20% contamination (bottom). Right: t_2 (top), t_1 (center), and χ_1^2-distribution (bottom). t-test (dotted), Wilcoxon test (dashed), median test (solid), test based on $\hat{\Delta}_{m,n}^{(1)}$ (bold dotted), $\hat{\Delta}_{m,n}^{(2)}$ (bold dashed) and $\hat{\Delta}_{m,n}^{(3)}$ (bold solid).

to an increasing number k of observations in one of the samples, i.e. the size and the number of outliers increase simultaneously. We consider outliers of random size with increasing mean and variance, since outliers of identical size would harm the kernel density estimators used for standardizing $\hat{\Delta}^{(1)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$ less. The power of the t-test resists a few small outliers, but breaks down thereafter. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ and the Wilcoxon test perform somewhat better, but are dominated by the median test and the tests based on $\hat{\Delta}^{(3)}_{m,n}$ and $\hat{\Delta}^{(2)}_{m,n}$.

In case of the t_3-distribution we get almost the same ordering of the methods as for the normal distribution. The Wilcoxon test possesses the same power as the tests based on $\hat{\Delta}^{(1)}_{m,n}$ or $\hat{\Delta}^{(2)}_{m,n}$. The two-sample t-test is outperformed by the other tests. As the tails of the distribution become heavier, in case of a t_1-distribution, the two-sample t-test loses almost all its power, according to the non-existence of any moments. The most powerful test is the median test then, followed by the test based on $\hat{\Delta}^{(3)}_{m,n}$ and the Wilxocon test. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ or $\hat{\Delta}^{(2)}_{m,n}$ become anti-conservative here as noted before.

In case of right-skewed (shifted) χ^2_3-distributions, the Wilcoxon test and the test based on $\hat{\Delta}^{(2)}_{m,n}$ are more powerful than the t-test. The tests based on $\hat{\Delta}^{(1)}_{m,n}$ are quite anti-conservative and not to be recommended in case of asymmetric distributions. The tests based on $\hat{\Delta}^{(3)}_{m,n}$ are more powerful than the median test but become anti-conservative as the skewness increases, i.e. for (shifted) χ^2_1-distributions. The median test is somewhat more powerful than the t-test then. The Wilcoxon test is the most powerful level α-test considered here in case of small values of Δ, whereas the test using $\hat{\Delta}^{(2)}_{m,n}$ is more powerful if Δ is large.

We confirmed these results for sample sizes $m = n = 30$ (not shown here). The main difference was that the tests using $\hat{\Delta}^{(1)}_{m,n}$ or $\hat{\Delta}^{(2)}_{m,n}$ were less anti-conservative for these smaller sample sizes in those scenarios for which these tests had problems before, see Figure 3. We checked these results also in case of unequal sample sizes $m = 2n = 50$ (not shown here). The results were very similar to those reported above for equal sample sizes, except for the test using $\hat{\Delta}^{(3)}_{m,n}$, which had severe problems with asymmetric distributions in case of unequal sample sizes.

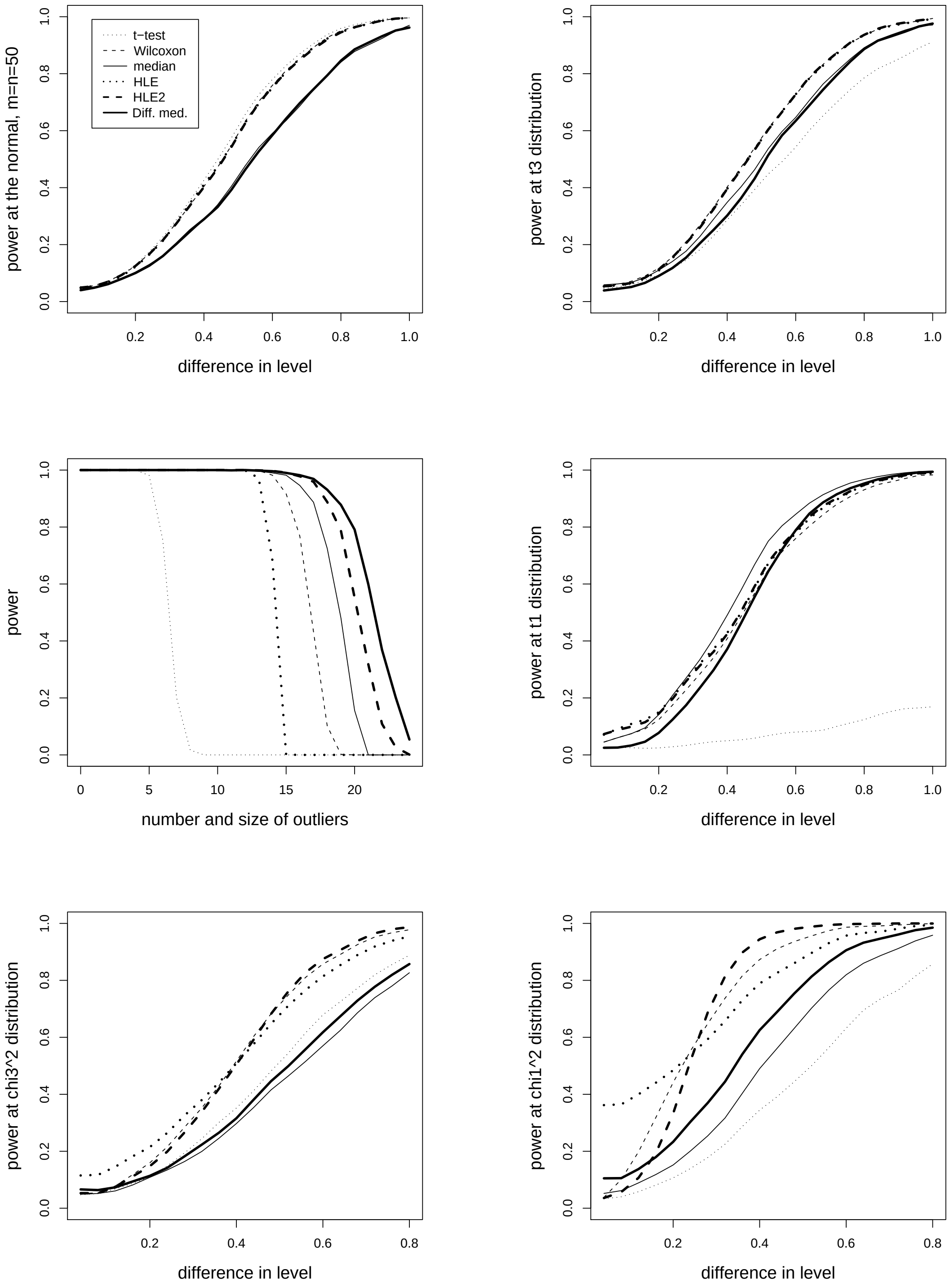

Figure 4: Power of the asymptotic tests in case of increasingly shifted normal (top left), t_3 (top right), t_1 (center right), χ_1^2 (bottom right), χ_3^2 (bottom left), and normal distributions with an increasing number and size of outliers, $\Delta = 2$ (center left), $m = n = 50$. t-test (dotted), Wilcoxon test (dashed), median test (solid), test based on $\hat{\Delta}_{m,n}^{(1)}$ (bold dotted), $\hat{\Delta}_{m,n}^{(2)}$ (bold dashed) and $\hat{\Delta}_{m,n}^{(3)}$ (bold solid).

5 Conclusions

Nonparametric tests for a location difference between two samples based on the properly scaled Hodges-Lehmann two-sample shift estimator, the median difference, can be constructed using the permutation principle in small samples, or the asymptotical distribution otherwise. The resulting tests are distribution free in small samples and at least approximately so in large samples. They perform very similarly to the Wilcoxon test, from which this estimator can be derived, under a broad variety of distributions, but they offer higher robustness against outliers. This is an advantage particularly in routine application, where we cannot check all data points carefully. Furthermore, we found a somewhat better performance as compared to the Wilcoxon test in case of asymmetric distributions. These advantages of the test based on the median difference as compared to the Wilcoxon test correspond to the stronger effects of extreme observations on the test statistic of the latter: a single outlier changes the sum of the ranks substantially in small samples, while the median difference is little affected. This explains the higher vulnerability of the Wilcoxon test against outliers, and also its smaller power in case of heavy tailed and skewed distributions, since these cause extreme observations.

The main disadvantage we observed was a violation of the significance level of the asymptotical test in case of the very heavy-tailed t_1-distribution, which does not possess any moments. The reason might be an inadequate choice of the bandwidth by the applied kernel density estimator. This problem could be overcome by a manual choice of the bandwidth, what is not possible in automatic application. A closer investigation of the effects of the dependencies between the pairs of observations on the resulting kernel density estimation seems worthwhile, particularly with respect to the suitable choice of the bandwidth.

We also constructed test statistics from the difference of the two sample medians. The resulting tests obtain even larger robustness against outliers in normal samples and perform similar to the median test otherwise. The main drawback of these tests, as compared to those above, are the reduced power in case of normal and moderately heavy-tailed distributions and problems in case of large skewness, or in case of skewness in combination with different sample sizes.

Initially, we also considered the 20%-trimmed two-sample t-test since these tests are often recommended in the literature (e.g. Keselman et al. 2002; Reed and Stark 2004), but did not find relevant advantages with respect to the criteria and data situations considered here. Moreover, we found these tests to be oversized even in case of moderate to large sample sizes, and not to improve the ordinary two-sample t-test substantially in case of a moderate number of medium-sized outliers, the heavy-tailed t_1-distribution and the skewed χ_1^2-distribution, under which the ordinary t-test has little power.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 823, "Statistical modelling of nonlinear dynamic processes") is gratefully acknowledged.

References

Bovik AC, Munson Jr DC (1986) Edge detection using median comparisons. Comput Vision, Graphics and Image Processing 33: 377–389

Edgington, E.S. (1995) Randomization Tests. 3rd Edition. Marcel Dekker, inc., New York

Fried R (2007) On robust shift detection in time series. Computational Statistics and Data Analysis 52: 1063-1074

Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. Ann Math Statist 34: 598–611

Hoyland A (1965) Robustness of the Hodges-Lehmann estimates for shift. Ann Math Statist 36: 174–197

Keselman HJ, Wilcox RR, Kowalchuk RK, Olejnik S (2002) Comparing trimmed or least squares means of two independent skewed populations. Biometrical Journal 44: 478–489

Lehmann EL (1963a) Robust estimation in analysis of variance. Ann Math Statist 34: 957–966

Lehmann EL (1963b) Asymptotically nonparametric inference: An alternative approach to linear models. Ann Math Statist 34: 1494–1506

Lehmann EL (1963c) Nonparametric confidence intervals for a shift parameter. Ann Math Statist 34: 1507–1512

Reed III. JF, Stark (2004) Robust two-sample statistics for equality of means: a simulation study. Journal of Applied Statistics 31: 831–854

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York

Simonoff JS (1996) Smoothing methods in Statistics. Springer, New York

Wilcox RR, Keselman HJ (2003) Modern robust data analysis methods: measures of central tendency. Psychological Methods 8: 254–274