Chlaß, Nadine; Krüger, Jens J.

Working Paper

Small sample properties of the Wilcoxon signed rank test with discontinuous and dependent observations

Jena Economic Research Papers, No. 2007,032

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Chlaß, Nadine; Krüger, Jens J. (2007) : Small sample properties of the Wilcoxon signed rank test with discontinuous and dependent observations, Jena Economic Research Papers, No. 2007,032, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/25598

JENA ECONOMIC RESEARCH PAPERS

2007 – 032

Small Sample Properties of the Wilcoxon Signed Rank Test with Discontinuous and Dependent Observations

by

Nadine Chlaß
Jens J. Krüger

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich-Schiller-University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact m.pasche@wiwi.uni-jena.de.

Impressum:

Friedrich-Schiller-University Jena
Carl-Zeiß-Str. 3
D-07743 Jena
www.uni-jena.de

Max-Planck-Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2007-032

Small Sample Properties of the Wilcoxon Signed Rank Test with Discontinuous and Dependent Observations

by

Nadine Chlaß

Max Planck Institute of Economics

Strategic Interaction Group, Kahlaische Strasse 10, D-07745 Jena, Germany,
Tel.: +49 3641 686 624, Fax: +49 3641 686 623, E-Mail: chlass@econ.mpg.de

and

Jens J. Krüger

Friedrich-Schiller-University Jena

Department of Economics, Carl-Zeiss-Strasse 3, D-07743 Jena, Germany,
Tel.: +49 3641 943 200, Fax: +49 3641 943 202, E-Mail: jens.krueger@wiwi.uni-jena.de

July 2007

Abstract

This Monte-Carlo study investigates sensitivity of the Wilcoxon signed rank test to certain assumption violations in small samples. Emphasis is put on within-sample-dependence, between-sample dependence, and the presence of ties. Our results show that both assumption violations induce severe size distortions and entail power losses. Surprisingly, these consequences do vary substantially with other properties the data may display. Results provided are particularly relevant for experimental settings where ties and within-sample dependence are frequently observed.

JEL classification: C12, C14, C15

Keywords: Wilcoxon signed rank test, ties, dependent observations, size and power

1 Introduction

The Wilcoxon rank tests constitute widely used nonparametric tests of sample dissimilarities based on ranked differences (Feltovic 2003, p. 274). Their virtues comprise a small number of assumptions (for an overview see Hollander/Wolfe 1999, p. 46) as well as the comparability of subjects at two points in time, thus allowing for the comparison of related samples. Wilcoxon rank tests rely on the continuity of investigated variables and therefore absence of ties, and second, the independence of observations within one sample. In what is to follow we investigate the impact of exclusive as well as simultaneous violations of these assumptions on both size and power properties for the Wilcoxon signed rank test.

The Wilcoxon signed rank test is routinely applied in experimental economics where both assumptions are frequently violated. Within experimental settings, there are several threats to continuity of response variables. First, the definition range of responses is not only finite, but frequently quite narrowly restricted, for example within [0,10] (Chlaß et al. 2006) or even within [0,1] (Dittrich et al. 2005). Second, actually observable responses are often furthermore reduced in variety by, for example, the response being restricted in decimals, different subjects using similar decision heuristics, or different subjects simultaneously playing the dominant strategy in each of the two points (situations) in time. Equal differences (ties) may thus frequently result and require consideration. Breaking of ties, e.g. via mid ranks (Hollander/Wolfe 1999, pp. 109), solves the problem only partially, since the question how differences could have been ranked, if more continuous responses had been possible or if subjects had thought to answer in a more continuous way cannot be answered. A more reliable approach therefore seems to consider whether the frequency of ties severely questions the validity of test results and then rather to opt for a test on discrete data (e.g. a χ^2 test), putting up with some loss of information. This, however, requires more detailed knowledge on the impact of ties - knowledge we wish to provide.

Dependence of observations within experimental settings does not only originate from repeatedly measuring the same individual's responses, but also from interaction of participants or latent variables in general. Thus, data possibly remain dependent even on an aggregated level, often rendering only the entire experimental session an independent observation. Only controlling for no interaction at all between several subgroups of subjects provides a larger number of independent observations. Within those subgroups, however, the problem of aggregation remains. Therefore, a considerable amount of data and information is

lost, decreasing overall test power and interpolating temporal trends whose investigation might be of interest as well. There are, of course, empirical methods, e.g. random or mixed effect models which account for dependence within samples and provide efficient and unbiased estimates for such cases. However, these estimates do not allow for any causal statistical interpretation in for which one would need information on the true conditional (in)dependence relations. Obtaining causal information, however, is the very aim of controlled experiments. As a consequence, dependence of observations might provide a valuable indicator of failures within an experimental design and thus serves as a control mechanism that should be retained rather than being accounted for statistically. Hence, it is, from a specification point of view, valuable to assess the impact of dependence on an estimate rather than to provide for it.

Our study investigates the effects of ties and dependent observations on size and power properties of the Wilcoxon signed rank test, determining actual p-values under each of the above-mentioned assumption violations. It aims at providing decision support as to what degree of violations the test remains a reliable tool and when its outcomes cannot be trusted. We proceed by reviewing hitherto obtained results regarding the Wilcoxon test in section 2, detail our simulation design in section 3, discuss our findings in sections 4 and 5 and conclude in section 6.

2 Literature Review

Let us first give an impression of hitherto conducted studies and the insights obtained so far regarding the Wilcoxon tests in general. Starting point of this literature is the study of Hodges and Lehmann (1956) who provide an early comparison of the Wilcoxon rank sum test and some other nonparametric alternatives to the t-test, analyzing asymptotic efficiency. They establish a lower threshold for relative Pitman efficiency of 0.864 in comparing Wilcoxon and t-test. Thus, in testing against shifts, the efficiency loss of the former as compared to the latter is bounded, while efficiency gains on the contrary may be infinite. However, the same comparison shows other rank tests, e.g. by Fisher and Yates (1948) and van der Waerden (1953), to reach relative asymptotic power of unity.

Turning to Monte-Carlo studies, Tanizaki (1997) assesses and compares power properties of the Wilcoxon rank sum and other rank-based tests to the t-test under various distributional assumptions. These comprise normal, Cauchy, logistic, chi-square and uniform distributions.

The results indicate that in small samples, the Wilcoxon test has highest overall power, performs better than other rank-based tests and clearly outperforms the t-test.

Zimmerman (1998) provides a Monte Carlo comparison of the Wilcoxon rank sum test with a modification of the t-test featuring various violations of normality and different standard deviations of the samples. Though for equal standard deviations the Wilcoxon test maintains its exact size as opposed to the t-test, neither test does so for unequal dispersions. In case of heavy-tailed distributions, different standard deviations and different sample sizes the Wilcoxon test may be more severely distorted than the t-test. The study does not investigate power properties.

Zimmerman (2000) provides a detailed Monte-Carlo investigation confined to the size properties of the Wilcoxon rank sum test. Given normally distributed samples and unequal dispersions the test systematically overrejects the null hypothesis if dispersions are unequal while the t-test appears to overreject less pronouncedly. These differences are due to the fact that the Wilcoxon test tests the equality of entire distributions instead of only considering their location parameters.

Freidlin and Gastwirth (2000) compare power properties of the Wilcoxon rank sum test to a variety of competing nonparametric tests for different underlying distributions and situations with equal or unequal sample sizes. They find the Wilcoxon rank sum test to have optimal power properties given logistic distributions. Comparative performance is rather discouraging under fat-tailed slash, Cauchy and t(2) distributions, while the test still demonstrates sound power for normal and double exponential distributions.

Finally, Feltovich (2003) in an extensive study assesses both size and power of the Wilcoxon rank sum test for a variety of situations and additionally provides a comparison to the robust-rank-order test of Fligner and Policello (1981). The situations investigated cover different sample sizes which may be equal or unequal, differing standard deviations and asymmetry of the underlying distributions. Furthermore compared are exact statistics and their normal approximations. A detailed assessment of size properties finds the robust-rank-order test to react more sensitively than the Wilcoxon rank sum test. For asymmetrically distributed data, both tests are found to perform poorly and tend to overreject, with an outperforming Wilcoxon rank sum test for unequal sample sizes. Power analysis is limited to the case with different central tendencies and equal dispersions of samples, no substantial differences being observed between the two tests.

The findings reviewed here find the Wilcoxon test a competitive testing method in many situations. The assumption violations investigated are limited to different dispersions and distributional asymmetry. Two further violations of assumptions required by the test having not yet received any attention remain tied observations and within-sample dependence. While for ties, remedying techniques have been proposed, such as the algorithm by Streitberg and Röhmer (1986, 1987), we show that their consequences persist nevertheless. Within-sample dependence on the other hand, has not yet received any attention at all. Considering the fact that especially in the context of experimental data, both phenomena often occur jointly, we also investigate simultaneous assumption violations.

While previous studies have focused on the Wilcoxon (rank sum) test for unrelated samples, we will concentrate on the procedure for related samples and additionally inquire to what extent the degree of sample relatedness matters.

3 Simulation Design

Our baseline scenario starts by simulating a series $\{v_t\}$ of identically and independently distributed random draws from a normal distribution with mean zero and standard deviation σ_v. The sample size is chosen by taking into account the fraction of ties $\lambda \in [0,1)$ that will be induced later on, so that a total of λn random draws of *v* are simulated to achieve a sample size of *n*. Note that λ and *n* are appropriately chosen so that λn is always integer. From $\{v_t\}$ a series $\{x_t\}$ is constructed via a simple stationary first-order autoregressive process AR (1) with autocorrelation parameter $\rho_x \in [0,1)$ where one period influences only a limited number of future periods. This amounts to simulating $x_t = \rho_x x_{t-1} + v_t$, allowing for a certain burn-in phase. Since we actually want to control the variance of the process $\{x_t\}$, we control for σ_x throughout the simulations and compute σ_v via $\sigma_v = \sigma_x \sqrt{1-\rho_x^2}$.

In a subsequent step a series $\{y_t\}$ is constructed from $\{x_t\}$ using a simple linear relationship $y_t = \delta + \beta x_t + u_t$ where $\{u_t\}$ represents another series of identically and independently distributed draws from a normal distribution with mean zero and standard deviation σ_u. Within these simulations we control for the standard deviation of *y*, σ_y, (in addition to the standard deviation of *x*, σ_x) and the correlation coefficient between *x* and *y*, ρ_{xy}. This is

implemented by computing β by $\beta = \rho_{xy} \cdot \sigma_x / \sigma_y$ and σ_u by $\sigma_u = \sqrt{\sigma_y^2 - \beta^2 \sigma_x^2}$. Thus, we are able to allow for different variances of the series $\{x_t\}$ and $\{y_t\}$ as well as for different degrees of dependence between them. Since $\{y_t\}$ originates from $\{x_t\}$, the properties of the latter carry over to the former so that $\{y_t\}$ equally exhibits stationary properties with a first-order autocorrelation coefficient $\rho_y = \rho_x \rho_{xy}^2$.

The mean difference between $\{x_t\}$ and $\{y_t\}$ is controlled by choosing δ such that the value of Hedges g represents no, small, medium or large differences. Hedges g (Hedges, Olkin 1985) is a measure similar to the more widely known Cohen's *d* (see Cohen 1988, p.23). Taking sample size into account, it allows for a better judgement of small, medium and large effects within comparisons of test power. In our case with equal sample sizes this just amounts to making an adjustment to the overall effect size based on the sample size. Specifically, we control δ by fixing g and computing $\delta = g \cdot \sqrt{(\sigma_x^2 + \sigma_y^2)/2} \Big/ \left(1 - 3/(8n - 9)\right)$, reflecting equal sample sizes for *x* and *y*.

Ties are induced in two different forms. The first variant randomly draws $(1-\lambda)n$ values from the λn simulated values of *x* and *y* with replacement and adds them to the sample, resulting in a total sample size of *n*. The second variant draws a single value out of the λn simulated values of *x* and *y* and adds this value $(1-\lambda)n$ times to the sample, resulting again in a total sample size of *n*. These variants are called the "many-small-ties variant" and the "one-big-tie variant", respectively. Of course, this way of inducing ties results in a total sample of values displaying different time series properties from those originally specified by σ_x, σ_y, ρ_x and ρ_{xy}. Thus, whenever ties are present, all subsequent interpretations of our results e.g. with respect to different degrees of dependence are only suggestive and should be viewed as qualitative statements regarding relatively higher or lower degrees of dependence.

Parameters are specified as follows. We consider four different sample sizes $n \in \{10, 20, 30, 40\}$ and four different fractions of ties $\lambda \in \{0, 0.2, 0.5, 0.8\}$. The standard deviation of *x* is fixed at unity, $\sigma_x = 1$, and that of *y* takes values $\sigma_y \in \{1, 2, 3, 4\}$. For the autocorrelation of *x* we investigate four different values $\rho_x \in \{0, 0.2, 0.5, 0.8\}$ as we do for the correlation between *x* and *y*, $\rho_{xy} \in \{0, 0.2, 0.5, 0.8\}$. Finally, Hedges g is specified in relation

to Cohen's d as $g \in \{0, 0.2, 0.5, 0.8\}$, representing no, small, medium and large effects as suggested by Cohen (1988 p.25).

We simulate 10000 replications for each of the 4096 possible combinations of parameter values and afterwards apply the Wilcoxon signed rank test to each of the simulated samples of x and y. A two-sided test is performed by comparing the test statistic to the respective critical values for a nominal significance level of $\alpha = 0.05$ which will be used throughout this paper. Finally recorded is the fraction of rejections out of the 10000 replications for each parameter combination.

Departing from our baseline scenario introduced above, we specifically modify the underlying distribution of samples $\{v_t\}$ in order to investigate how fat tails or asymmetry of the distribution affect the test properties under assumption violations. As representative examples for distributions with large tails and asymmetry t and gamma distributions were chosen. The first case corresponds to a series being drawn from a t distribution with three degrees of freedom. A subsequent division by its standard deviation $\sqrt{3}$ and scaling by σ_v^2, yields an error process $\{v_t\}$ with mean zero and variance σ_v^2. Regarding parameters for the gamma distribution controlling shape a and scale b we first require the variance being fixed at unity, $\sigma_v^2 = ab^2 = 1$ and second, prespecify skewness via $\gamma = 2/\sqrt{a}$. This results in $a = (2/\gamma)^2$ and $b = \sqrt{1/a}$. A subsequent subtraction of the mean ensures a zero-mean error process $\{v_t\}$. To shed light on the impact of skewness, parameter a is set to take values $a \in \{1, 2, 3, 4\}$.

Furthermore, we investigate the impact of specific kinds of dependence running simulations with a first order moving-average, or MA(1), process $x_t = v_t + \theta v_{t-1}$ where each period influences infinitely many future periods. This specification relies on inverting the first-order autocorrelation coefficient $\rho_x = \theta/(1+\theta^2)$ for computing θ and then using this value to compute the variance of $\{v_t\}$ by $\sigma_v^2 = \sigma_x^2/(1+\theta^2)$.

Finally, all constellations mentioned above are investigated in their impact on size and power properties for the exact version of the Wilcoxon signed rank test which, in the presence of ties, is implemented using the algorithm proposed in (Streitberg/Röhmer 1986, 1987). Our questions are first, whether ties still show a residual impact and second, whether the impact of

within-sample dependence differs anyhow compared to the standard test. To summarize this chapter, figure 1 reviews all treatment combinations investigated.

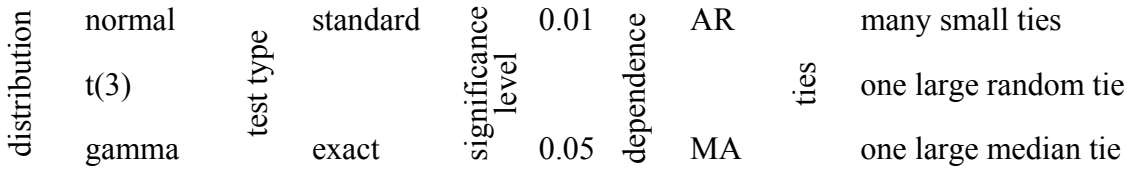

Figure 1: Treatment combinations investigated.

4 Simulation Results

Table 1 presents rejection frequencies for a ***first scenario*** constituting our baseline case. It introduces a normal distribution, an intermediary degree of dependence $\rho_{xy} = 0.5$, equal variances $\sigma_x = \sigma_y = 1$, AR(1) errors and many small ties, drawn with replacement from the simulated values. With the samples showing no difference and in absence of any assumption violation, the test maintains its *size* level quite exactly. *Power* increases with both accruing deviation from the null hypothesis *g* and sample size and becomes quite considerable for larger sample sizes. However, power persists on a sometimes considerably low level for both small and medium size effects, first rising above what is recommended in (Cohen 1988, p. 53) for $\{g = 0.5, n = 40\}$.

As a first assumption violation, let us now investigate the impact of discontinuity in the investigated variables entailing tied observations on both size and power properties. The first four rows with $\rho_x = 0$ and $\lambda \in \{0, 0.2, 0.5, 0.8\}$ indicate that, when within-sample-dependence is absent but ***ties***[1] are present to various degrees, the test rejects a true null hypothesis more often than it should, the problem persisting with increasing sample size. For $\lambda = 0.2$, that is,

[1] The effects of both ties and within-sample dependence on *size* as mentioned here are not immediately obtained by visual inspection of the tables provided. To give an example, the tie effect on size $\Delta^{s}_{t\left|\{\lambda=\bar{\lambda},n=\bar{n}\}\right.}$ is calculated by $\Delta^{s}_{t\left|\{\lambda=\bar{\lambda},n=\bar{n}\}\right.} = (a-(a-b))-(c-(c-d))$ with $a := (f_r\left|\{g=\bar{g},n=\bar{n},\lambda=\bar{\lambda},\rho_x=0\}\right.$, $b := f_r\left|\{g=0,n=\bar{n},\lambda=\bar{\lambda},\rho_x=0\}\right.$, $c := (f_r\left|\{g=\bar{g},n=\bar{n},\lambda=0,\rho_x=0\}\right.$, $d := f_r\left|\{g=0,n=\bar{n},\lambda=0,\rho_x=0\}\right.$ where f_r denotes the respective rejection frequency displayed in the corresponding table. The impact of ties on size- adjusted *power* is calculated the following way: $\Delta^{p}_{t\left|\{\lambda=\bar{\lambda},n=\bar{n}\}\right.} = f_r\left|\{\lambda=\bar{\lambda},n=\bar{n},\rho_x=0\right\} - f_r\left|\{\lambda=0,n=\bar{n},\rho_x=0\right\}$. Intuitively, we subtract from the initial rejection frequency all those parts that are due to a variation in parameters beside the one of interest. We analogously proceed for within-sample dependence.

20% ties within the sample, nominal critical p-values should be lowered by [0.03,0.05][2]. The test may thus still respect a weak significance level (remember the baseline reference here being chosen $\alpha = 0.05$ [3]). Maximal errors in indicated size are of 0.4 $\{g = 0, \lambda = 0.8, n = 10\}$.

Tied observations entail a loss in size-adjusted *power*, this decrease being more pronounced for larger effects. Interestingly, power declines more pronouncedly with increasing sample size for small and medium effects, though in absolute terms, the number of untied observations increases and one therefore would likely expect the reverse. Only for large effects, increasing sample size may actually help reducing power losses. Maximal power losses are of 0.54 $\{g = 0.5, \lambda = 0.8, n = 40\}$.

Within-sample dependence, our second assumption violation investigated, equally leads to an overrejection which turns out to be less pronounced than in the presence of ties only. Interestingly, the problem rather worsens with increasing sample size. For a low level of within-sample dependence $\rho_x = 0.2$, the test may still be applied by lowering critical nominal p-values by $[0.01,0.02]$. Intermediary levels of within-sample dependence $\rho_x = 0.5$ demand a lowering of critical nominal p-values by $[0.04,0.06]$, thus only allowing for weak significance levels (recall the nominal significance level of $\alpha = 0.05$). Maximal errors in indicated size of 0.2 occur for $\{g = 0, \rho_x = 0.8, n = 40\}$.

Power losses entailed are roughly half as important as for tied observations, its dependence on sample size displaying the same pattern. Interestingly, weak degrees of within sample dependence may actually help structuring the sample and thus improve power properties, if the effect itself is small. Maximal power losses of 0.31 occur for $\{g = 0.5, \rho_x = 0.8, n = 40\}$.

When ***both within-sample dependence and ties*** are present a mutual reinforcement of the adverse *size* effects from both violations of the test assumptions can be observed. This tendency towards overrejection declines only very slowly with increasing sample size and recognizably only for large fractions of ties and high degrees of within-sample-dependence.

[2] The interval is defined by the smallest and largest correction necessary for different sample sizes.

[3] For an induced significance level of 0.01, the impact of assumption violations on ***size*** is roughly half as important as for 0.05. The corrections suggested here may thus serve as an upper bound. The impact of assumption violations on power under 0.01, however, is more severe for small and medium effects. Note that effect size and significance level are related: small and medium but yet highly significant effects alone are already difficult to identify for our small sample sizes. For the same degree of assumption violations, such effects are therefore more difficult to identify than effects of intermediate significance. Consistently, this difference in impact of assumption violations under 0.01 and 0.05 disappears for large effect sizes.

Table 1
Baseline Simulation Results

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.051	0.050	0.050	0.046	0.084	0.129	0.183	0.232	0.306	0.554	0.749	0.857	0.642	0.923	0.990	0.998
$\rho_x = 0$	$\lambda = 0.2$	0.085	0.088	0.086	0.091	0.127	0.166	0.206	0.244	0.316	0.503	0.666	0.773	0.584	0.854	0.955	0.986
$\rho_x = 0$	$\lambda = 0.5$	0.170	0.165	0.167	0.164	0.202	0.225	0.253	0.269	0.348	0.476	0.586	0.673	0.547	0.748	0.873	0.930
$\rho_x = 0$	$\lambda = 0.8$	0.460	0.359	0.337	0.336	0.456	0.384	0.374	0.385	0.527	0.516	0.565	0.610	0.628	0.687	0.761	0.837
$\rho_x = 0.2$	$\lambda = 0$	0.058	0.061	0.057	0.064	0.100	0.146	0.193	0.246	0.313	0.564	0.744	0.855	0.635	0.910	0.984	0.998
$\rho_x = 0.2$	$\lambda = 0.2$	0.100	0.101	0.104	0.110	0.135	0.172	0.219	0.250	0.331	0.517	0.658	0.769	0.584	0.835	0.944	0.984
$\rho_x = 0.2$	$\lambda = 0.5$	0.197	0.180	0.181	0.178	0.222	0.238	0.269	0.288	0.352	0.480	0.576	0.672	0.559	0.745	0.858	0.928
$\rho_x = 0.2$	$\lambda = 0.8$	0.487	0.379	0.364	0.350	0.493	0.412	0.406	0.402	0.560	0.531	0.579	0.614	0.645	0.699	0.765	0.828
$\rho_x = 0.5$	$\lambda = 0$	0.092	0.097	0.101	0.104	0.131	0.189	0.238	0.273	0.349	0.557	0.709	0.818	0.637	0.885	0.969	0.994
$\rho_x = 0.5$	$\lambda = 0.2$	0.134	0.143	0.147	0.145	0.164	0.216	0.246	0.280	0.347	0.523	0.643	0.751	0.597	0.829	0.929	0.970
$\rho_x = 0.5$	$\lambda = 0.5$	0.245	0.232	0.231	0.227	0.265	0.286	0.293	0.318	0.404	0.505	0.585	0.665	0.584	0.746	0.853	0.917
$\rho_x = 0.5$	$\lambda = 0.8$	0.525	0.446	0.415	0.411	0.532	0.462	0.446	0.451	0.589	0.557	0.591	0.634	0.688	0.716	0.779	0.821
$\rho_x = 0.8$	$\lambda = 0$	0.211	0.246	0.239	0.246	0.250	0.299	0.332	0.367	0.401	0.555	0.671	0.745	0.619	0.811	0.916	0.964
$\rho_x = 0.8$	$\lambda = 0.2$	0.267	0.282	0.290	0.285	0.300	0.328	0.353	0.383	0.435	0.542	0.633	0.699	0.602	0.772	0.870	0.925
$\rho_x = 0.8$	$\lambda = 0.5$	0.375	0.375	0.368	0.383	0.395	0.404	0.429	0.436	0.487	0.550	0.614	0.653	0.620	0.730	0.812	0.857
$\rho_x = 0.8$	$\lambda = 0.8$	0.618	0.576	0.565	0.556	0.634	0.591	0.582	0.584	0.671	0.647	0.672	0.680	0.731	0.744	0.769	0.807

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$ and $\sigma_x = \sigma_y = 1$; test decisions are based on a significance level $\alpha = 0.05$.

Regarding *power*, large fractions of ties appearing together with high degrees of dependence entail the most severe power losses. Interestingly, power losses under simultaneous assumption violations are an increasing function of the sample size for small and medium effects. Only for large effects, power losses are again U-shapedly dependent on sample size. However, the larger the effect size, the more pronounced overall power losses.

Let us proceed with a second ***scenario of unequal variances*** with $\sigma_y = 4\sigma_x$ displayed in table 2. With no other assumption being violated, the test slightly underrejects the null hypothesis for $n = \{10,20\}$ and starts overrejecting it somewhat for $n \in \{30,40\}$ for all effects. Unequal variances entail power losses, depending U-shapedly on effect size. While for small effect sizes only slight power losses occur though increasing with sample size, large amounts of power are lost for medium effect size. For large effects losses disappear with increasing *n*, being negligible for $n \in \{20,30\}$. The ***tie effect***[4] on *size* is reinforced in this scenario. For levels of $\lambda \in \{0.2, 0.5, 0.8\}$ the test overrejects a null by maximally additional $\{0.014, 0.05, 0.07\}$ as compared to the baseline. Considering actual significance levels under this simultaneous assumption violation, a weak significance level might still be respected for $\lambda = 0.2$ where nominal critical p-values should systematically be lowered by 0.04 up to 0.05. Regarding *size-adjusted power,* unequal sample variances may substantially reduce power losses entailed by ties, this reduction being an increasing function of sample size except for large effects. Furthermore, the larger the fraction of ties, the higher the power loss reduction entailed by differences in variance.

Turning to ***within-sample dependence***, its impact is weakened by unequal variances. Overrejection diminishes, entailing a need for lowering nominal critical p-values by maximally but $\{0.007, 0.014, 0.064\}$ for $\rho_x \in \{0.2, 0.5, 0.8\}$. Regarding *size-adjusted power*

---

[4] The impact of a particular scenario is quantified by including one further condition in our calculations. To analyze the effect of unequal standard deviations on the impact ties show, this requires the following modification:

$$\Delta^{\sigma_y=4\sigma_x}_{\Delta^{s}_{t|\{\lambda=\bar{\lambda},n=\bar{n}\}}} = ((a|\{\sigma_y = 4\sigma_x\} - (a|\{\sigma_y = 4\sigma_x\} - b|\{\sigma_y = 4\sigma_x\})) - (c|\{\sigma_y = 4\sigma_x\} - (c|\{\sigma_y = 4\sigma_x\} - d|\{\sigma_y = 4\sigma_x\}))) - ((a - (a - b)) - (c - (c - d)))$$

for the example above. Changes in the impact ties exert on size-adjusted power, are now calculated as follows:

$$\nabla^{\sigma_y=4\sigma_x}_{\nabla^{p}_{t|\{\lambda=\bar{\lambda},n=\bar{n}\}}} = (f_r|\{\lambda = \bar{\lambda}, n = \bar{n}, \rho_x = 0, \sigma_y = 4\sigma_x\} - f_r|\{\lambda = 0, n = \bar{n}, \rho_x = 0, \sigma_y = 4\sigma_x\}) - (f_r|\{\lambda = \bar{\lambda}, n = \bar{n}, \rho_x = 0\} - f_r|\{\lambda = 0, n = \bar{n}, \rho_x = 0\}),$$

the subtrahend in either equation corresponding to the baseline case with $\sigma_x = \sigma_y$. This latter condition is not mentioned again for reasons of space.

Table 2
Results with Unequal Variances

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.046	0.049	0.051	0.050	0.074	0.103	0.138	0.174	0.214	0.403	0.563	0.693	0.461	0.779	0.924	0.975
$\rho_x = 0$	$\lambda = 0.2$	0.091	0.100	0.096	0.098	0.126	0.157	0.179	0.206	0.263	0.404	0.536	0.640	0.480	0.729	0.874	0.945
$\rho_x = 0$	$\lambda = 0.5$	0.215	0.201	0.197	0.200	0.237	0.234	0.271	0.282	0.352	0.432	0.523	0.599	0.504	0.675	0.807	0.881
$\rho_x = 0$	$\lambda = 0.8$	0.525	0.426	0.400	0.398	0.542	0.436	0.431	0.443	0.575	0.531	0.563	0.593	0.640	0.663	0.721	0.789
$\rho_x = 0.2$	$\lambda = 0$	0.053	0.052	0.056	0.052	0.084	0.109	0.141	0.182	0.214	0.395	0.563	0.696	0.458	0.774	0.922	0.976
$\rho_x = 0.2$	$\lambda = 0.2$	0.099	0.097	0.099	0.102	0.131	0.155	0.182	0.215	0.267	0.409	0.542	0.647	0.476	0.728	0.870	0.947
$\rho_x = 0.2$	$\lambda = 0.5$	0.227	0.209	0.205	0.206	0.252	0.255	0.263	0.288	0.350	0.430	0.516	0.603	0.507	0.677	0.792	0.880
$\rho_x = 0.2$	$\lambda = 0.8$	0.536	0.428	0.408	0.411	0.544	0.453	0.443	0.447	0.593	0.537	0.568	0.602	0.655	0.667	0.730	0.785
$\rho_x = 0.5$	$\lambda = 0$	0.059	0.063	0.065	0.068	0.087	0.121	0.152	0.183	0.230	0.409	0.554	0.682	0.472	0.762	0.920	0.969
$\rho_x = 0.5$	$\lambda = 0.2$	0.114	0.120	0.115	0.117	0.144	0.171	0.200	0.220	0.270	0.418	0.539	0.644	0.491	0.720	0.856	0.933
$\rho_x = 0.5$	$\lambda = 0.5$	0.247	0.228	0.232	0.217	0.253	0.270	0.287	0.301	0.371	0.440	0.537	0.599	0.515	0.670	0.798	0.868
$\rho_x = 0.5$	$\lambda = 0.8$	0.558	0.458	0.440	0.435	0.565	0.477	0.462	0.461	0.607	0.558	0.569	0.612	0.664	0.671	0.731	0.783
$\rho_x = 0.8$	$\lambda = 0$	0.100	0.108	0.115	0.112	0.126	0.160	0.197	0.218	0.265	0.413	0.557	0.669	0.475	0.734	0.871	0.946
$\rho_x = 0.8$	$\lambda = 0.2$	0.165	0.170	0.178	0.181	0.183	0.214	0.245	0.267	0.314	0.433	0.549	0.628	0.502	0.700	0.835	0.908
$\rho_x = 0.8$	$\lambda = 0.5$	0.294	0.287	0.293	0.297	0.309	0.324	0.338	0.334	0.405	0.471	0.531	0.603	0.538	0.677	0.779	0.843
$\rho_x = 0.8$	$\lambda = 0.8$	0.586	0.513	0.494	0.496	0.592	0.519	0.517	0.514	0.628	0.588	0.606	0.632	0.685	0.680	0.737	0.772

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$, $\sigma_x = 1$ and $\sigma_y = 4$; test decisions are based on a significance level $\alpha = 0.05$.

slight improvements are observed for $\rho_x \in \{0.2, 0.5\}$ while under $\rho_x = 0.8$, power gains as compared to the scenario with equal variances reach up to 0.23. Turning to a last constellation where ***both ties and within-sample dependence*** are present, the tie impact on *size* under unequal variances dominates for small levels of within sample dependence. That is, overrejection increases as compared to the baseline.

However, for accruing within-sample dependence the latter effect takes lead and overrejection for simultaneous assumption violations under unequal sample variances diminishes in comparison to the baseline. Regarding the impact on *power*, an overall improvement is observed. Thus, power losses decrease as compared to the reference case and more pronouncedly do so for both increasing within-sample dependence and tie fraction. However, power losses decrease decisively faster with increasing within-sample dependence.

A ***third scenario*** as shown in table 3 allows us to investigate a possible impact of the very ***nature of within-sample dependence*** and for this purpose introduces an MA rather than an AR structure. While for $\rho_x = 0.2$, the impact on *size* is manifested by a slightly increased overrejection, the reverse holds for stronger within-sample dependence: in presence of an MA structure overrejection diminishes by $[0.02; 0.03]$ for $\rho_x = 0.5$ as compared to an AR structure. *Power* improves for the same degree of within-sample dependence up to 0.046 and increasingly does so for larger sample sizes. Again, power improvements become more visible for $\rho_x = 0.5$ and display an increasing dependence on sample size. Thus, the kind of within-sample dependence equally affects the relationship between power and sample size. Furthermore, when ***both ties and within-sample dependence*** are present, power properties improve slightly with increasing *g*. riven by this effect, overrejection entailed by simultaneous assumption violation slightly declines.

A ***fourth scenario*** associated with table 4 subsequently investigates the impact of the ***nature of ties*** by introducing one large tie instead of many small. With respect to *size*, the impact is tremendous. For the smallest fraction $\lambda = 0.2$, overrejection entailed by ties increases by additional $[0.03, 0.05, 0.08, 0.07]$ for $n \in \{10, 20, 30, 40\}$. Thus, weak significance levels cannot even be respected for the smallest fraction of ties investigated. Rejection frequency with samples displaying no difference reaches certainty for $\lambda = 0.8$. In presence of ***both ties and within-sample dependence***, overrejection increases beyond the tie-driven effect by $[0.04,\ 0.07,\ 0.08, 0.09]$ for $n \in \{10, 20, 30, 40\}$.

Table 3
Results for Within-Sample Dependence as MA(1) Instead of AR(1)

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.052	0.049	0.048	0.054	0.093	0.131	0.190	0.227	0.302	0.554	0.746	0.858	0.629	0.922	0.989	0.998
$\rho_x = 0$	$\lambda = 0.2$	0.091	0.083	0.087	0.088	0.123	0.164	0.199	0.247	0.308	0.513	0.665	0.781	0.584	0.846	0.957	0.985
$\rho_x = 0$	$\lambda = 0.5$	0.170	0.166	0.173	0.168	0.204	0.228	0.255	0.285	0.350	0.461	0.584	0.669	0.554	0.751	0.866	0.935
$\rho_x = 0$	$\lambda = 0.8$	0.451	0.358	0.343	0.328	0.462	0.380	0.378	0.384	0.528	0.514	0.560	0.606	0.640	0.695	0.765	0.832
$\rho_x = 0.2$	$\lambda = 0$	0.062	0.059	0.059	0.057	0.097	0.147	0.190	0.240	0.309	0.564	0.737	0.859	0.638	0.911	0.983	0.997
$\rho_x = 0.2$	$\lambda = 0.2$	0.092	0.096	0.102	0.095	0.140	0.181	0.214	0.255	0.324	0.510	0.663	0.766	0.588	0.839	0.943	0.985
$\rho_x = 0.2$	$\lambda = 0.5$	0.186	0.174	0.179	0.176	0.218	0.232	0.255	0.294	0.359	0.470	0.580	0.660	0.555	0.751	0.873	0.936
$\rho_x = 0.2$	$\lambda = 0.8$	0.474	0.382	0.352	0.352	0.486	0.406	0.403	0.397	0.557	0.535	0.576	0.615	0.649	0.700	0.772	0.834
$\rho_x = 0.5$	$\lambda = 0$	0.071	0.073	0.078	0.081	0.114	0.158	0.203	0.256	0.326	0.568	0.730	0.834	0.624	0.902	0.980	0.996
$\rho_x = 0.5$	$\lambda = 0.2$	0.117	0.110	0.118	0.114	0.150	0.192	0.233	0.270	0.335	0.510	0.657	0.760	0.592	0.826	0.937	0.981
$\rho_x = 0.5$	$\lambda = 0.5$	0.210	0.201	0.195	0.199	0.244	0.256	0.272	0.305	0.374	0.480	0.587	0.660	0.565	0.740	0.865	0.921
$\rho_x = 0.5$	$\lambda = 0.8$	0.504	0.407	0.395	0.373	0.524	0.430	0.421	0.414	0.562	0.541	0.576	0.618	0.653	0.694	0.778	0.828

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$, $\sigma_x = 1$ and $\sigma_y = 1$; test decisions are based on a significance level $\alpha = 0.05$.

Table 4
Results for the Variant with One Big (Median) Tie Instead of Many Small Ties

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.054	0.048	0.047	0.048	0.086	0.132	0.180	0.231	0.301	0.558	0.748	0.861	0.636	0.925	0.988	0.998
$\rho_x = 0$	$\lambda = 0.2$	0.119	0.135	0.159	0.161	0.171	0.251	0.314	0.375	0.411	0.664	0.812	0.892	0.720	0.939	0.988	0.997
$\rho_x = 0$	$\lambda = 0.5$	0.355	0.512	0.591	0.668	0.391	0.573	0.668	0.743	0.559	0.801	0.878	0.940	0.773	0.950	0.983	0.995
$\rho_x = 0$	$\lambda = 0.8$	0.714	0.968	0.999	1.000	0.720	0.969	1.000	1.000	0.769	0.978	0.999	1.000	0.836	0.987	1.000	1.000
$\rho_x = 0.2$	$\lambda = 0$	0.057	0.060	0.065	0.062	0.098	0.148	0.192	0.247	0.318	0.551	0.732	0.846	0.629	0.912	0.986	0.997
$\rho_x = 0.2$	$\lambda = 0.2$	0.146	0.165	0.182	0.199	0.191	0.269	0.332	0.391	0.412	0.659	0.806	0.891	0.706	0.931	0.985	0.997
$\rho_x = 0.2$	$\lambda = 0.5$	0.373	0.519	0.597	0.668	0.407	0.592	0.671	0.753	0.572	0.791	0.876	0.930	0.769	0.939	0.980	0.994
$\rho_x = 0.2$	$\lambda = 0.8$	0.713	0.968	0.999	1.000	0.729	0.968	1.000	1.000	0.775	0.978	0.999	1.000	0.837	0.986	1.000	1.000
$\rho_x = 0.5$	$\lambda = 0$	0.093	0.095	0.106	0.101	0.145	0.180	0.226	0.280	0.339	0.554	0.719	0.820	0.635	0.892	0.968	0.991
$\rho_x = 0.5$	$\lambda = 0.2$	0.185	0.217	0.233	0.245	0.227	0.302	0.367	0.417	0.435	0.640	0.780	0.855	0.697	0.907	0.972	0.992
$\rho_x = 0.5$	$\lambda = 0.5$	0.424	0.569	0.646	0.715	0.456	0.620	0.699	0.767	0.595	0.785	0.868	0.921	0.760	0.931	0.971	0.990
$\rho_x = 0.5$	$\lambda = 0.8$	0.734	0.972	0.999	1.000	0.748	0.975	1.000	1.000	0.792	0.979	1.000	1.000	0.847	0.987	1.000	1.000
$\rho_x = 0.8$	$\lambda = 0$	0.209	0.235	0.242	0.242	0.247	0.301	0.328	0.357	0.408	0.560	0.671	0.748	0.618	0.823	0.916	0.959
$\rho_x = 0.8$	$\lambda = 0.2$	0.336	0.371	0.387	0.399	0.362	0.419	0.467	0.495	0.492	0.631	0.724	0.796	0.686	0.846	0.924	0.960
$\rho_x = 0.8$	$\lambda = 0.5$	0.540	0.684	0.736	0.777	0.557	0.703	0.759	0.807	0.634	0.790	0.851	0.896	0.750	0.890	0.941	0.969
$\rho_x = 0.8$	$\lambda = 0.8$	0.784	0.983	1.000	1.000	0.792	0.986	0.999	1.000	0.808	0.988	1.000	1.000	0.855	0.990	1.000	1.000

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$, $\sigma_x = 1$ and $\sigma_y = 1$; test decisions are based on a significance level $\alpha = 0.05$.

The impact on *power* is less unanimous. For weak effects of $g = 0.2$, a small fraction $\lambda = 0.2$ of a larger tie helps structuring the sample, and does so more efficiently than smaller ties improving power by additional $[0.01, 0.03, 0.04, 0.06]$ for $n \in \{10, 20, 30, 40\}$. Power gains for $g = 0.2$ as compared to small ties are furthermore increasing with sample size. However, larger fractions of large ties entail additional power losses for all situations both with increasing sample size and drastically so for larger effect sizes. Maximally incurred additional power losses of 0.5 are observed for $\{g = 0.8, n = 40, \lambda = 0.8\}$. ***Both ties and within-sample dependence*** yield subadditively increased power losses for low fraction of ties, for high fraction of ties however, power losses exceed the sum of the two single effects.

A ***fifth scenario*** displayed in table 5 sheds light on the impact of ties and within-sample dependence under asymmetry of the underlying distributions. In the absence of any further assumption violation and with the samples showing no difference, asymmetry entails a tendency towards underrejection. Its effect on *size* is of approximately -0.02. *Power* under asymmetry increases overall, the improvement being dependent on both effect and sample size. Maximal power gains of 0.07 occur for $\{g = 0.5, n = 30\}$. The ***tie effect*** on *size* is ambiguously influenced by asymmetry. Generally, entailed overrejection decreases, and especially does so for the case of interest with $\lambda = 0.2$ where the effect is slightly weakened but for $n = 10$. Regarding the tie effect on *power*, asymmetry yields power improvements of up to 0.07 $\{g = 0.8, n = 40, \lambda = 0.8\}$. A more pronounced change is observed for the impact of ***within-sample dependence***. Its effect on *size* is overcompensated for $\rho_x = 0.2$ and $\rho_x = 0.5$. Depending on sample size, it diminishes by $[0.004, 0.005, 0.006, 0.008]$ for the former and by $[0.04, 0.05, 0.05, 0.06]$ for the latter case. Furthermore, asymmetry causes ρ_x – entailed overrejection to decline much faster with increasing sample size. Even for $\rho_x = 0.8$, starting with an actual size of 0.12, the test reaches nominal size for $n = 40$. *Power* losses entailed by within-sample dependence decrease quite substantially under asymmetry, and more pronouncedly do so with increasing sample size. Maximal power gains of 0.21 occur for $\{g = 0.8, \rho_x = 0.8, n = 40\}$. Only in the absence of any effect and for very small samples $n \in \{10, 20\}$ slight *power* losses are observed in comparison to the baseline.

Table 5
Results for Gamma-Distributed Instead of Normal-Distributed Series

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.029	0.028	0.029	0.030	0.065	0.114	0.170	0.234	0.289	0.592	0.803	0.915	0.641	0.952	0.997	0.999
$\rho_x = 0$	$\lambda = 0.2$	0.064	0.065	0.065	0.069	0.098	0.157	0.212	0.271	0.312	0.543	0.729	0.843	0.613	0.884	0.977	0.994
$\rho_x = 0$	$\lambda = 0.5$	0.153	0.138	0.147	0.139	0.191	0.219	0.265	0.311	0.358	0.506	0.642	0.736	0.579	0.792	0.909	0.960
$\rho_x = 0$	$\lambda = 0.8$	0.435	0.338	0.311	0.299	0.454	0.367	0.369	0.395	0.528	0.531	0.595	0.659	0.652	0.717	0.819	0.884
$\rho_x = 0.2$	$\lambda = 0$	0.031	0.029	0.030	0.028	0.067	0.116	0.175	0.237	0.284	0.596	0.796	0.911	0.650	0.953	0.996	1.000
$\rho_x = 0.2$	$\lambda = 0.2$	0.064	0.062	0.061	0.058	0.104	0.150	0.212	0.263	0.316	0.549	0.722	0.836	0.618	0.884	0.972	0.993
$\rho_x = 0.2$	$\lambda = 0.5$	0.149	0.139	0.142	0.138	0.191	0.227	0.262	0.302	0.344	0.513	0.636	0.740	0.576	0.789	0.907	0.965
$\rho_x = 0.2$	$\lambda = 0.8$	0.451	0.332	0.314	0.299	0.457	0.378	0.369	0.385	0.548	0.541	0.603	0.653	0.659	0.724	0.811	0.880
$\rho_x = 0.5$	$\lambda = 0$	0.032	0.029	0.027	0.026	0.068	0.108	0.164	0.219	0.295	0.571	0.775	0.890	0.626	0.939	0.995	1.000
$\rho_x = 0.5$	$\lambda = 0.2$	0.070	0.063	0.059	0.061	0.102	0.151	0.195	0.243	0.317	0.522	0.693	0.813	0.597	0.861	0.963	0.992
$\rho_x = 0.5$	$\lambda = 0.5$	0.166	0.142	0.137	0.131	0.199	0.216	0.248	0.280	0.379	0.493	0.602	0.703	0.586	0.773	0.893	0.948
$\rho_x = 0.5$	$\lambda = 0.8$	0.482	0.356	0.317	0.307	0.504	0.393	0.389	0.377	0.564	0.547	0.593	0.644	0.668	0.729	0.804	0.860
$\rho_x = 0.8$	$\lambda = 0$	0.119	0.074	0.059	0.048	0.157	0.162	0.169	0.202	0.342	0.508	0.660	0.784	0.586	0.837	0.955	0.991
$\rho_x = 0.8$	$\lambda = 0.2$	0.176	0.131	0.113	0.091	0.215	0.199	0.214	0.233	0.357	0.476	0.586	0.687	0.576	0.767	0.890	0.954
$\rho_x = 0.8$	$\lambda = 0.5$	0.303	0.257	0.218	0.194	0.335	0.311	0.291	0.284	0.431	0.479	0.535	0.604	0.584	0.705	0.799	0.865
$\rho_x = 0.8$	$\lambda = 0.8$	0.599	0.523	0.484	0.462	0.596	0.531	0.509	0.490	0.641	0.614	0.613	0.633	0.708	0.711	0.751	0.795

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$, $\sigma_x = \sigma_y = 1$ and skewness of one; test decisions are based on a significance level $\alpha = 0.05$.

The impact of a ***simultaneous assumption violation*** on *size* under asymmetry yields a decrease in overrejection beyond a mere addition of the single effects. Superadditivity increases with sample size. Overall nominal significance levels are reached under $\rho_x = 0.2$ and $\rho_x = 0.5$ for $\lambda = 0.2$, indicating that overrejection increases much faster in λ than in ρ_x, though dimensions, of course, lack independence. For less important degrees of simultaneous assumption violations the test remains consistently below indicated significance levels. While for small and medium degrees of within-sample dependence and all fractions of ties *power* losses decrease, the same does not hold for high within-sample dependence.

For $\{\rho_x = 0.8, \lambda = 0.2\}$, power gains are still observed as compared to the baseline case. However, larger fractions of ties incur additional power losses, and severely may do so.

A ***sixth scenario*** as displayed in table 6 investigates performance of the ***exact test***. In presence of ties its computation is performed via an algorithm introduced by Streitberg and Röhmel (1986, 1987). Regarding *size*, the impact is not unanimous, the test displaying slightly lower size levels for $n \in \{10, 20, 30\}$ and slightly higher for $n = 40$ as compared to the standard test. *Power* properties as compared to the baseline case vary inconsistently, though an overall slight improvement for very small samples of $n \in \{10, 20\}$ may be observed. The ***impact of ties*** persists, being affected in an equally unstructured way by this setting. Regarding *size* in the case of interest $\lambda = 0.2$, overrejection entailed by ties decreases by $[0.002, 0.002, 0.01]$ for $n \in \{10, 20, 40\}$. However, it is reinforced by 0.008 for $n = 30$. *Power* in the presence of ties does neither systematically nor substantially vary from the baseline case. With rare exceptions, the impact of ***within-sample dependence*** on *size* is emphasized under the exact test. Given $\rho_x = 0.2$, overrejection increases by $[0.015, 0.011, 0.015, 0.012]$ according to sample size. *Power* varies but slightly and inconsistently as compared to the baseline case. The implications of ***within-sample dependence and ties*** occurring simultaneously are less unanimous regarding size. The exact test shows an increased tendency towards overrejection but for $n = 40$, when overrejection declines as compared to the baseline case. Regarding *power*, no systematic variation can be observed.

A ***seventh*** and last ***scenario*** corresponding to table 7 finally highlights the impact of increased between-sample dependence ρ_{xy}. With the samples showing no difference and in absence of

Table 6
Results With Test Decision Based on Exact P-Values

	$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$ $\lambda = 0$	0.047	0.048	0.049	0.051	0.090	0.136	0.172	0.231	0.301	0.555	0.740	0.861	0.639	0.921	0.988	0.997
$\rho_x = 0$ $\lambda = 0.2$	0.079	0.084	0.093	0.086	0.118	0.165	0.207	0.240	0.316	0.516	0.665	0.775	0.591	0.844	0.953	0.985
$\rho_x = 0$ $\lambda = 0.5$	0.176	0.167	0.167	0.162	0.199	0.224	0.246	0.274	0.350	0.470	0.588	0.669	0.552	0.758	0.872	0.935
$\rho_x = 0$ $\lambda = 0.8$	0.464	0.355	0.342	0.327	0.463	0.383	0.388	0.380	0.527	0.521	0.570	0.618	0.638	0.692	0.768	0.837
$\rho_x = 0.2$ $\lambda = 0$	0.062	0.059	0.064	0.063	0.102	0.149	0.197	0.242	0.320	0.554	0.732	0.852	0.634	0.912	0.983	0.997
$\rho_x = 0.2$ $\lambda = 0.2$	0.098	0.103	0.103	0.106	0.142	0.178	0.220	0.242	0.319	0.512	0.670	0.764	0.586	0.842	0.947	0.984
$\rho_x = 0.2$ $\lambda = 0.5$	0.197	0.183	0.185	0.180	0.227	0.239	0.259	0.295	0.363	0.475	0.577	0.666	0.572	0.743	0.868	0.934
$\rho_x = 0.2$ $\lambda = 0.8$	0.477	0.381	0.357	0.352	0.497	0.417	0.395	0.415	0.554	0.534	0.570	0.614	0.657	0.702	0.773	0.832
$\rho_x = 0.5$ $\lambda = 0$	0.093	0.097	0.104	0.108	0.127	0.182	0.222	0.278	0.336	0.559	0.714	0.825	0.630	0.887	0.970	0.994
$\rho_x = 0.5$ $\lambda = 0.2$	0.142	0.142	0.147	0.146	0.178	0.215	0.251	0.292	0.349	0.527	0.651	0.752	0.594	0.827	0.932	0.972
$\rho_x = 0.5$ $\lambda = 0.5$	0.245	0.240	0.228	0.226	0.267	0.280	0.303	0.319	0.390	0.500	0.586	0.666	0.584	0.748	0.854	0.914
$\rho_x = 0.5$ $\lambda = 0.8$	0.520	0.439	0.416	0.414	0.534	0.464	0.467	0.458	0.586	0.573	0.596	0.639	0.679	0.713	0.774	0.817
$\rho_x = 0.8$ $\lambda = 0$	0.216	0.240	0.243	0.241	0.247	0.297	0.336	0.364	0.406	0.561	0.666	0.745	0.614	0.818	0.912	0.962
$\rho_x = 0.8$ $\lambda = 0.2$	0.266	0.277	0.284	0.292	0.288	0.324	0.362	0.381	0.427	0.537	0.640	0.711	0.609	0.772	0.875	0.931
$\rho_x = 0.8$ $\lambda = 0.5$	0.377	0.377	0.376	0.379	0.402	0.415	0.429	0.428	0.484	0.552	0.612	0.658	0.611	0.730	0.809	0.864
$\rho_x = 0.8$ $\lambda = 0.8$	0.622	0.579	0.568	0.560	0.624	0.587	0.582	0.584	0.675	0.652	0.652	0.671	0.726	0.743	0.776	0.808

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.5$, $\sigma_x = 1$ and $\sigma_y = 1$; test decisions are based on a significance level $\alpha = 0.05$.

any assumption violation, *size* turns out more conservative for all but the largest sample size. Increasing within-sample dependence yields tremendous power gains, a maximal improvement of 0.37 occurring for $\{g = 0.5, n = 20\}$. Power already reaches values close to maximum for precisely this case.

The impact of ***ties*** is equally strongly affected. Starting with *size*, overrejection as compared to the baseline case reduces to roughly half its extent. Thus, even for $\lambda = 0.5$ with actual sizes of $[0.14, 0.12, 0.12, 0.12]$ according to *n*, weak significance levels are only slightly reneged and may thus be respected for differing chosen α. Pursuing with *power*, predications need to be conditioned on effects. For weak effect sizes, increased between-sample dependence reinforces power losses and accruingly does so for larger sample sizes. However, for medium and large effects, power differences under enhanced between-sample dependence become a positive function of sample size. Increasing between-sample dependence may thus yield important power improvements for large effect and sample sizes.

Overrejection entailed by ***within-sample dependence*** decreases for nearly all cases. While for the lowest level $\rho_x = 0.2$, the effect is not yet unanimous with slightly decreased overrejection observed for $n \in \{10, 20\}$ and slightly increased for $n \in \{30, 40\}$, a clear pattern emerges for larger ρ_x. Overrejection decreases importantly, the remedying effect being a positive function of sample size and weak significance levels being within reach for up to $\rho_x = 0.5$. Regarding *power* losses entailed by within-sample-dependence, between-sample dependence likewise exhibits a remedying effect, though kicking in only from medium size effects on. Maximal power gains of 0.22 as compared to the baseline are observed for $\{g = 0.8, n = 20, \rho_x = 0.8\}$. Analyzing the change in impact of both ***within-sample dependence and ties*** on *size*, a decrease in overrejection beyond an addition of the changes due to ties and within-sample dependence is observed. Thus, in simultaneous presence of low or medium within-sample dependence and the lowest fraction of ties, weak nominal significance levels still hold. Turning to *power* in the presence of a simultaneous assumption violation loss variations are dependent on effect size.

Simultaneous assumption violations seem to affect power more importantly for small and medium effect sizes in case of larger ρ_{xy}, while for large effect sizes the reverse holds. However, though more strongly affected, absolute test power for these constellations is superior to the baseline case for all effect sizes.

Table 7
Results with Larger Between-Sample Dependence

		$g = 0$ $n = 10$	$g = 0$ $n = 20$	$g = 0$ $n = 30$	$g = 0$ $n = 40$	$g = 0.2$ $n = 10$	$g = 0.2$ $n = 20$	$g = 0.2$ $n = 30$	$g = 0.2$ $n = 40$	$g = 0.5$ $n = 10$	$g = 0.5$ $n = 20$	$g = 0.5$ $n = 30$	$g = 0.5$ $n = 40$	$g = 0.8$ $n = 10$	$g = 0.8$ $n = 20$	$g = 0.8$ $n = 30$	$g = 0.8$ $n = 40$
$\rho_x = 0$	$\lambda = 0$	0.046	0.048	0.049	0.050	0.149	0.262	0.383	0.486	0.623	0.917	0.985	0.997	0.952	1.000	1.000	1.000
$\rho_x = 0$	$\lambda = 0.2$	0.076	0.073	0.071	0.079	0.157	0.234	0.304	0.385	0.499	0.762	0.911	0.966	0.801	0.981	0.999	1.000
$\rho_x = 0$	$\lambda = 0.5$	0.133	0.113	0.116	0.115	0.181	0.213	0.255	0.297	0.399	0.583	0.732	0.824	0.661	0.878	0.962	0.990
$\rho_x = 0$	$\lambda = 0.8$	0.370	0.254	0.221	0.214	0.390	0.301	0.302	0.308	0.531	0.522	0.615	0.671	0.694	0.776	0.853	0.917
$\rho_x = 0.2$	$\lambda = 0$	0.051	0.053	0.057	0.060	0.151	0.271	0.374	0.490	0.621	0.910	0.986	0.997	0.943	1.000	1.000	1.000
$\rho_x = 0.2$	$\lambda = 0.2$	0.077	0.078	0.083	0.079	0.162	0.236	0.312	0.387	0.500	0.769	0.901	0.964	0.802	0.980	0.999	1.000
$\rho_x = 0.2$	$\lambda = 0.5$	0.133	0.128	0.122	0.123	0.191	0.224	0.268	0.300	0.418	0.594	0.737	0.828	0.673	0.883	0.964	0.989
$\rho_x = 0.2$	$\lambda = 0.8$	0.385	0.261	0.243	0.223	0.410	0.320	0.312	0.329	0.549	0.551	0.613	0.685	0.708	0.787	0.872	0.918
$\rho_x = 0.5$	$\lambda = 0$	0.066	0.068	0.069	0.072	0.173	0.286	0.396	0.490	0.624	0.899	0.980	0.996	0.941	0.999	1.000	1.000
$\rho_x = 0.5$	$\lambda = 0.2$	0.090	0.098	0.098	0.096	0.173	0.257	0.330	0.404	0.522	0.768	0.904	0.960	0.815	0.978	0.997	1.000
$\rho_x = 0.5$	$\lambda = 0.5$	0.160	0.145	0.142	0.135	0.215	0.246	0.285	0.323	0.451	0.603	0.746	0.834	0.719	0.891	0.962	0.989
$\rho_x = 0.5$	$\lambda = 0.8$	0.419	0.302	0.276	0.266	0.460	0.368	0.356	0.363	0.588	0.589	0.649	0.713	0.754	0.816	0.888	0.931
$\rho_x = 0.8$	$\lambda = 0$	0.112	0.132	0.137	0.137	0.218	0.320	0.415	0.500	0.607	0.868	0.955	0.985	0.914	0.995	1.000	1.000
$\rho_x = 0.8$	$\lambda = 0.2$	0.148	0.154	0.156	0.151	0.234	0.292	0.368	0.425	0.549	0.756	0.881	0.940	0.833	0.970	0.995	0.999
$\rho_x = 0.8$	$\lambda = 0.5$	0.228	0.214	0.212	0.206	0.287	0.303	0.333	0.371	0.519	0.645	0.757	0.826	0.767	0.895	0.956	0.983
$\rho_x = 0.8$	$\lambda = 0.8$	0.507	0.419	0.379	0.360	0.539	0.466	0.451	0.448	0.661	0.658	0.695	0.734	0.801	0.853	0.901	0.935

Note: reported are rejection frequencies for the baseline case with $\rho_{xy} = 0.8$, $\sigma_x = 1$ and $\sigma_y = 1$; test decisions are based on a significance level $\alpha = 0.05$.

The next section investigates the phenomena identified in a more general and integrated setting via response surfaces, analyzing a variety of scenarios jointly.

5 Response Surfaces

In this section we summarize our results by estimating response surfaces. Response surfaces can be defined as "a regression model in which each observation corresponds to one experiment, the dependent variable is some quantity that was estimated in the experiments, and the independent variables are functions of the various parameter values, chosen by the experimenter, which characterize each experiment" (Davidson and MacKinnon 1993, pp. 755f.).

Since heteroskedasticity often poses a problem in response-surface regressions, we follow MacKinnon (1994) and employ the form of GMM estimation proposed by Cragg (1983). This estimator allows for an estimation of linear regression models by simultaneously accounting for heteroskedasticity of unknown form. Denoting our response-surface regressions in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \text{ with } E(\mathbf{uu}') = \boldsymbol{\Omega},$$

y represents the vector of rejection frequencies obtained from our experiments and matrix **X** contains parameters such as the degree of dependence, or the fraction of ties set in each experiment. **X** may also include certain nonlinear transformations or interactions of these parameters. Cragg's estimator furthermore requires the specification of a set of auxiliary variables which, together with the columns of **X**, constitute columns of a matrix **Q**. Incorporating these elements the estimator can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\hat{\boldsymbol{\Omega}}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\hat{\boldsymbol{\Omega}}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y},$$

where $\hat{\boldsymbol{\Omega}}$ denotes a diagonal matrix whose nonzero elements are the squared OLS residuals from a regression of **y** on **X**. For the auxiliary variables we choose third powers and two-way interactions of the explanatory variables, as suggested by Cragg (1983). The results appear to be quite robust under different sets of auxiliary variables. Note that R^2 values are obtained from $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ based on the Cragg estimator.

In the regressions reported below the dependent variable is the rejection frequency of our experiments. We opted for this variable instead of its logit transformation since actual

rejection frequencies equal unity. Moreover, response surfaces fit the rejection frequencies themselves much better than their logit transform. Of course, fitted values outside the interval [0,1] emerge, but these are only of minor importance as far as magnitude and frequency are concerned.

We start by discussing the response-surface results for *size* in table 8. All observations with $g = 0$ are pooled together irrespectively of whether they originate from experiments with different tie modes (many small, one big random or median) and/or different distributions (either normal and gamma or t(3) and gamma). Outcomes of test decisions are pooled as well, based on critical values from either the distribution of the test statistic, or the exact p-values following Streitberg and Röhmel (1986, 1987).

The first column of table 8 displays the response surfaces estimates for the normally and gamma distributed experiments. Therein, within-sample dependence is specified by a first-order AR process. R^2 confirms the model by indicating reasonably large explanatory power. This holds although quite diverse sets of experiments are pooled together. In analyzing individual coefficients, we first consider linear terms. They are highly significant except for the within-sample dependence of *x*, ρ_x, and between-sample dependence, ρ_{xy}. Only the induced within-sample dependence of *y*, ρ_y appears to be significant with a negative sign, thus reducing overrejection. Sample size *n* and the fraction of ties λ turn out significantly positive, whereas the degree of skewness γ and the ratio of standard deviations σ_y come out significantly negative (recall that σ_x is fixed at unity throughout).

Quantitatively important quadratic effects can be observed in cases of special interest for this paper, namely within-sample dependence of *x* and fraction of ties. Both coefficients have a positive sign, show reasonably large magnitude, and are associated with large t-statistics. Thus, size distortions induced by these assumption violations become extensively evident in quadratic effects.

The next block of coefficients shows the effects of interactions with sample size. Here, only the variables governing dependence and the fraction of ties significantly interact with *n*. As concerns the latter case, a positive sign reveals size distortions to increase with sample size. This indicates not only the fraction of ties but also their absolute number to play a role. The observed interaction of variables measuring dependence with *n* appears to be more complicated. Significantly positive and negative coefficients appear together so that a clear

predication as to the overall effect of dependence on size cannot be made.

A most important conclusion from the inspection of the tables in the previous section was that

Table 8
Response Surfaces for Size

distribution	normal + gamma		t(3) + gamma	
error process	AR	MA	AR	MA
intercept	-0.133 (-14.420)	-0.166 (-17.810)	-0.154 (-17.731)	-0.174 (-18.665)
n	0.012 (30.431)	0.012 (29.496)	0.012 (30.189)	0.013 (29.744)
σ_y	-0.030 (-5.936)	0.001 (0.220)	0.005 (1.151)	0.006 (1.126)
γ	-0.046 (-18.558)	-0.028 (-11.098)	-0.052 (-21.862)	-0.026 (-10.137)
ρ_x	0.021 (1.706)	0.046 (2.395)	0.016 (1.384)	0.042 (2.193)
ρ_y	-0.171 (-4.914)	-0.024 (-0.468)	-0.198 (-6.623)	-0.023 (-0.440)
ρ_{xy}	-0.009 (-0.755)	0.031 (2.427)	-0.012 (-1.004)	0.033 (2.595)
λ	0.695 (69.257)	0.633 (62.044)	0.681 (72.058)	0.624 (61.482)
n^2	-0.000 (-22.115)	-0.000 (-22.952)	-0.000 (-22.142)	-0.000 (-23.283)
σ_y^2	0.004 (4.675)	-0.000 (-0.229)	-0.002 (-1.946)	-0.001 (-1.026)
γ^2	0.010 (19.438)	0.006 (11.141)	0.011 (21.736)	0.006 (10.660)
ρ_x^2	0.203 (19.488)	-0.036 (-1.320)	0.294 (28.564)	-0.036 (-1.314)
ρ_y^2	0.027 (0.624)	0.156 (1.408)	-0.163 (-4.061)	0.130 (1.184)
ρ_{xy}^2	0.087 (8.192)	-0.027 (-2.398)	0.085 (7.795)	-0.031 (-2.711)
λ^2	0.349 (34.672)	0.469 (47.074)	0.345 (35.505)	0.483 (48.924)
$n \cdot \sigma_y$	-0.000 (-0.131)	-0.000 (-1.392)	0.000 (1.272)	-0.000 (-1.382)
$n \cdot \gamma$	-0.000 (-1.044)	0.000 (0.545)	-0.000 (-0.180)	0.000 (0.432)
$n \cdot \rho_x$	-0.004 (-12.420)	-0.001 (-1.537)	-0.004 (-13.003)	-0.001 (-1.392)
$n \cdot \rho_y$	0.005 (5.120)	0.000 (0.272)	0.004 (5.133)	0.000 (0.351)
$n \cdot \rho_{xy}$	-0.001 (-3.626)	0.000 (0.026)	-0.001 (-3.057)	-0.000 (-0.003)
$n \cdot \lambda$	0.006 (28.215)	0.005 (23.924)	0.005 (27.904)	0.005 (23.797)
small	0.180 (31.250)	0.192 (33.855)	0.177 (31.392)	0.194 (34.589)
small $\cdot$ n	-0.007 (-53.271)	-0.007 (-54.901)	-0.007 (-57.043)	-0.007 (-55.625)
small $\cdot$ σ_y	0.011 (6.224)	0.012 (7.850)	0.007 (4.914)	0.011 (7.420)
small $\cdot$ γ	-0.002 (-1.573)	0.006 (5.483)	-0.002 (-2.183)	0.005 (4.270)
small $\cdot$ ρ_x	0.071 (10.954)	0.006 (0.695)	0.114 (19.350)	0.004 (0.438)
small $\cdot$ ρ_y	-0.087 (-4.068)	-0.024 (-0.859)	-0.160 (-9.705)	-0.025 (-0.932)
small $\cdot$ ρ_{xy}	-0.044 (-6.558)	-0.048 (-7.183)	-0.057 (-8.663)	-0.043 (-6.553)
small $\cdot$ λ	-0.647 (-137.573)	-0.650 (-144.371)	-0.636 (-147.401)	-0.645 (-145.468)
random	0.030 (18.295)	0.031 (17.901)	0.027 (16.940)	0.030 (17.758)
exact	0.003 (2.276)	0.003 (1.816)	0.003 (2.178)	0.003 (1.932)
N	12288	9216	12288	9216
R^2	0.944	0.936	0.935	0.936

Note: t-statistics based on heteroskedasticity-robust standard errors in parentheses.

switching from the many-small-ties to the one-big-tie variant considerably affects the results and worsens size distortions. This is reflected by a dummy variable named “small”, indicating treatments with many small ties, in its interactions with the other variables. Of considerable magnitude appears the coefficient reflecting the interaction of “small” with the fraction of

ties. It reveals the tendency towards overrejection for a larger fraction of ties to be more pronounced in case of one big tie.

The nature of this one big tie being either a deterministic median or a randomly drawn observation also plays a role as indicated by the significantly positive dummy variable for the random ties named “random”.

Our impression regarding the exact test obtained by inspecting the tables is confirmed. The test decision using exact p-values based on a correction for the presence of ties is only slightly affected. The coefficient of the dummy-variable “exact” for these cases is quite small and shows a t-statistic of only slightly larger than two. Although significant at conventional levels, this significance seems negligible considering the quite large sample size available for our response surfaces.

The next column shows our results for within-sample dependence following an MA process instead of the AR process. The major difference to the results for the AR process just discussed is that fewer variables turn out significant. This is especially true for variables representing within-sample dependence, ρ_x and ρ_y. They are neither highly significant individually nor in their quadratic effects, theirs interactions with *n* and their interactions with “small” being not significant at all. Two aspects need to be considered here. On one hand, an MA process can only capture autocorrelations up to 0.5, so that only experiments for lower correlations could be calculated for this variant and fewer data are available. On the other hand, remember our qualitative results having found less pronounced overrejection for the case of an MA process. By contrast, the fraction of ties remains significantly positive with coefficients of considerable magnitude whenever observed in the response surface. Especially the additional effect of the one-big-tie variant shows up again.

The next two columns show corresponding results for experiments with the normal distribution being replaced by the more fat-tailed t-distribution with three degrees of freedom. Most of our results prove to be stable with respect to this change, although some changes in significance and even sign reversals can also be observed. Explanatory power is equally high in this regression.

Let us mention a few further interesting interaction terms with robust characteristics which are not reported above. $\sigma_y \cdot \lambda \cdot \rho_x$ and $\sigma_y \cdot \rho_x$ turn out significantly negative which implies increasing differences in dispersion to weaken the joint impact of ties and within-sample

dependence and of within-sample dependence alone. Skewness of underlying distributions significantly reinforces overrejection entailed by ties $(\gamma \cdot \lambda)$, makes overrejection due to ties decline more slowly with sample size $(\gamma \cdot \lambda \cdot n)$ and reinforces the simultaneous effect of ties and within-sample dependence. However, both skewed and within-sample dependent data $(\gamma \cdot \rho_x)$ cause the test to overreject less often than within-sample dependent data alone.

Turning to *power*, results are shown in table 9. Here, the specification includes Hedge's *g*, a wider set of interactions with *g* and the other variables being introduced. Naturally, *g* is positively associated with the rejection frequency, its quadratic effect showing up both positive and highly significant. Moreover, both interactions of *g* with sample size and the dummy variable for small ties are positive and associated with very high t-statistics. Concerning the interactions with *g* we find all of them significant with power-depressing effects of differences in variance, the within-sample dependence of *x* and the fraction of ties.

Naturally for a consistent test, the interaction of *g* and the sample size *n* is significantly positive with high t-statistics in all regressions. As already suspected in the previous section, increasing between-sample dependence appears to be beneficial for power. The general pattern regarding coefficients tends to parallel our results in table 8. Explanatory power is somewhat weaker in table 9 but R^2 values remain consistently above 0.9 with a single minor exception.

Let us finally mention some interaction terms not reported here. Power losses due to within-sample dependence are significantly stronger the larger the effect size as measured by $\rho_x \cdot g$. For large ties, we find that the larger the effect size, the heavier tie entailed losses of power. While under skewness, power for large effect sizes improves $(\lambda \cdot g)$, this phenomenon decreases with growing sample size $(\lambda \cdot g \cdot n)$. Skewness reinforces the power depressing impact of ties $(\gamma \cdot \lambda)$, within-sample dependence $(\gamma \cdot \rho_x)$ and a joint occurrence of both, $(\gamma \cdot \rho_x \cdot \lambda)$ but does less so with increasing sample size.

Table 9
Response Surfaces for Power

distribution	normal + gamma		t(3) + gamma	
error process	AR	MA	AR	MA
intercept	-0.248 (-40.210)	-0.276 (-41.052)	-0.308 (-50.143)	-0.226 (-30.582)
n	0.020 (72.588)	0.012 (29.496)	0.021 (75.150)	0.018 (55.129)
σ_y	-0.037 (-12.447)	-0.000 (-0.142)	0.011 (3.647)	-0.040 (-10.520)
γ	-0.034 (-19.027)	-0.010 (-5.682)	-0.040 (-23.661)	-0.018 (-9.878)
ρ_x	0.153 (20.649)	0.048 (3.959)	0.233 (31.478)	0.058 (4.169)
ρ_y	-0.183 (-8.241)	0.076 (2.043)	-0.401 (-18.030)	0.020 (0.513)
ρ_{xy}	0.026 (2.963)	0.033 (3.557)	0.012 (1.292)	0.002 (0.210)
λ	0.976 (146.281)	0.990 (136.188)	0.914 (137.899)	0.976 (122.278)
g	0.598 (72.686)	0.459 (51.839)	0.736 (88.773)	0.587 (56.489)
n^2	-0.000 (-53.514)	-0.000 (-52.487)	-0.000 (-51.581)	-0.000 (-41.582)
σ_y^2	0.005 (8.554)	-0.000 (-0.025)	-0.001 (-2.012)	0.013 (18.982)
γ^2	0.006 (15.772)	0.001 (2.205)	0.007 (19.680)	0.002 (6.094)
ρ_x^2	0.065 (10.218)	-0.047 (-2.859)	0.110 (17.230)	-0.065 (-3.243)
ρ_y^2	0.141 (4.955)	-0.085 (-1.136)	0.019 (0.656)	-0.048 (-0.566)
ρ_{xy}^2	0.106 (14.958)	0.010 (1.342)	0.182 (24.784)	0.027 (3.138)
λ^2	0.249 (39.538)	0.299 (43.803)	0.258 (40.403)	0.258 (33.556)
g^2	0.151 (24.316)	0.281 (42.592)	0.061 (9.591)	0.277 (36.718)
$n \cdot \sigma_y$	0.000 (7.858)	0.000 (3.044)	-0.000 (-1.763)	-0.000 (-7.576)
$n \cdot \gamma$	0.000 (5.732)	0.000 (4.459)	0.000 (7.590)	0.000 (3.764)
$n \cdot \rho_x$	-0.004 (-23.112)	-0.000 (-1.693)	-0.005 (-29.454)	-0.001 (-1.778)
$n \cdot \rho_y$	0.003 (5.125)	-0.001 (-0.920)	0.005 (8.874)	0.000 (0.218)
$n \cdot \rho_{xy}$	-0.003 (-12.820)	-0.002 (-6.987)	-0.004 (-15.988)	-0.001 (-3.482)
$n \cdot \lambda$	-0.003 (-22.593)	-0.004 (-24.638)	-0.003 (-19.748)	-0.001 (-6.133)
$n \cdot g$	0.008 (58.954)	0.009 (60.908)	0.006 (46.822)	0.009 (53.325)
$g \cdot \sigma_y$	-0.020 (-13.428)	-0.026 (-15.505)	-0.010 (-6.545)	-0.201 (-87.978)
$g \cdot \gamma$	0.022 (18.523)	-0.008 (-6.005)	0.019 (16.613)	-0.005 (-3.494)
$g \cdot \rho_x$	-0.202 (-34.769)	-0.045 (-4.907)	-0.351 (-60.655)	-0.068 (-6.206)
$g \cdot \rho_y$	0.104 (5.586)	-0.074 (-2.445)	0.426 (23.178)	-0.002 (-0.068)
$g \cdot \rho_{xy}$	0.174 (23.692)	0.199 (26.015)	0.177 (23.489)	0.225 (25.062)
$g \cdot \lambda$	-1.049 (-230.702)	-1.056 (-214.485)	-1.040 (-228.582)	-0.889 (-142.878)
small	-0.013 (-3.020)	0.015 (3.259)	-0.034 (-7.714)	0.083 (16.758)
small $\cdot n$	-0.003 (-38.765)	-0.004 (-39.143)	-0.003 (-35.585)	-0.004 (-41.733)
small $\cdot \sigma_y$	0.001 (0.891)	0.003 (2.635)	0.005 (4.737)	-0.009 (-6.529)
small $\cdot \gamma$	-0.001 (-0.714)	-0.000 (-0.179)	-0.003 (-3.927)	-0.000 (-0.019)
small $\cdot \rho_x$	0.031 (8.155)	-0.004 (-0.749)	0.050 (13.162)	-0.003 (-0.522)
small $\cdot \rho_y$	0.008 (0.623)	-0.014 (-0.665)	-0.055 (-4.257)	-0.017 (-0.740)
small $\cdot \rho_{xy}$	-0.007 (-1.445)	-0.012 (-2.397)	0.004 (0.788)	-0.019 (-3.450)
small $\cdot \lambda$	-0.484 (-151.939)	-0.514 (-152.872)	-0.466 (-144.334)	-0.552 (-138.062)
small $\cdot g$	0.219 (74.805)	0.210 (67.381)	0.228 (78.565)	0.155 (42.234)
random	-0.023 (-22.624)	-0.017 (-15.541)	-0.026 (-24.673)	-0.009 (-6.831)
exact	0.003 (3.015)	0.003 (2.816)	0.003 (2.925)	0.003 (2.491)
N	49152	36864	49152	36864
R^2	0.913	0.924	0.898	0.914

Note: t-statistics based on heteroskedasticity-robust standard errors in parentheses.

6 Conclusion

In what preceded we have investigated the performance of the Wilcoxon signed rank test for related samples under assumption violations hitherto not thoroughly scrutinized. We namely focus on tied observations and within-sample dependence. Our goal was to assess their impact and to additionally shed light on their interaction with further possible characteristics of real data – frequently met on one hand and relevant for the Wilcoxon signed rank test on the other.

Therefore, the scenarios introduced comprise

- a first and baseline scenario, investigating ties and within-sample dependence in their impact on power and size of the Wilcoxon signed rank test assuming normal distributions, medium between-sample dependence, equal standard deviations, small ties and within-sample dependence following an AR(1) process;
- a second scenario, investigating the same impact under inequality in sample standard deviations, a further assumption violation of the test;
- a third, shedding light on the effect of a varying nature of within-sample dependence;
- a forth, inquiring the effect of a varying nature of ties on their impact;
- a fifth, analyzing how asymmetry of the underlying distributions affects the impact of both ties and within-sample dependence;
- a sixth to see how a so-called exact test disposing of a tie breaking algorithm performs;
- and a seventh, analyzing the effect of between-sample dependence on the impact of both ties and within-sample dependence.

As expected, both ties and within-sample dependence were found to entail overrejection and power losses. Ties are found to entail heavier consequences than within-sample dependence, while both assumption violations clearly persist with increasing sample size. Albeit and unexpectedly, slight assumption violations may actually improve the test ability to identify weak effects in small samples. Impacts of both assumption violations hinge on further data characteristics. The extent of this dependence surprises. Both asymmetry of the underlying distributions and increasing between-sample dependence are found to substantially attenuate the impact of both assumption violations. Unequal standard deviations imply miscellaneous consequences for the two assumption violations in question. They weaken the impact of within-sample dependence but enforce the effect of ties with tie correction surprisingly failing

to show any apparent improvement. Finally, unequal standard deviations also influence the impact of ties and within-sample dependence occurring jointly and thereby yield either an additive, subadditive or superadditive combination of the single effects. The results mentioned are confirmed by response surface estimates showing an appealing overall fit.

Our future work aims at providing insights for finer grids of values, that is, degrees of both assumption violations and a more general specification of within-sample dependence. A similar kind of analysis for unrelated samples is already in progress.

References

Chlaß, N., Güth, W., Vanberg, C. (2006), Social Learning of Efficiency Enhancing Trade With(out) Market Entry Costs, Papers on Strategic Interaction 2006-36, Max Planck Institute of Economics.

Cragg, J.G. (1983), More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form, Econometrica, vol. 51, pp. 751-763.

Cohen, J. (1988), Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Davidson, R., MacKinnon, J.G. (1993), Estimation and Inference in Econometrics, Oxford: Oxford University Press.

Dittrich, D.A.V., Güth, W., Kocher, M., Pezanis-Christou, P. (2005), Loss Aversion and Learning to Bid, Papers on Strategic Interaction 2005-03, Max Planck Institute of Economics.

Feltovich, N. (2003), Nonparametric Tests of Differences in Medians: Comparison of the Wilcoxon-Mann-Whitney and Robust Rank-Order Tests, Experimental Economics, vol. 6, pp. 273-297.

Fisher, R.A., Yates, F. (1948), Statistical Tables for Biological, Agricultural and medial Research, 3rd ed., Edinburgh: Oliver and Boyd.

Fligner, M.A., Policello, G.E. (1981), Robust Rank Procedures for the Behrens-Fisher Problem, Journal of the American Statistical Association, vol. 76, p. 162-168.

Freidlin, B., Gastwirth, J.L. (2000), Should the Median Test be Retired from General Use? American Statistician, vol. 54, pp. 161-164.

Hedges, L.V., Olkin, I. (1985) Statistical Methods for Meta-Analysis, Academic Press.

Hodges, J., Lehmann, E.L. (1956), The Efficiency of Some Nonparametric Competitors of the t Test, Annals of Mathematical Statistics, vol. 27, pp. 324-335.

Hollander, M., Wolfe, D.A. (1999), Nonparametric Statistical Methods, 2nd ed., New York: Wiley.

MacKinnon, J.G. (1994), Approximate Asymptotic Distribution Functions for Unit-Root and Cointegration Tests, Journal of Business and Economic Statistics, vol. 12, pp. 167-176.

Streitberg, B., Röhmel, J. (1986), Exact Distributions for Permutations and Rank Tests: An Introduction to Some Recently Published Algorithms, Statistical Software Newsletter vol. 12, pp. 10-17.

Streitberg, B., Röhmel, J. (1987), Exakte Verteilungen für Rang- und Randomisierungstests im allgemeinen *c*-Stichprobenfall, EDV in Medizin und Biologie, vol. 18, pp.12-19.

Tanizaki, H. (1997), Power Comparison of Non-Parametric Tests: Small-Sample Properties from Monte Carlo Experiments, Journal of Applied Statistics, vol. 24, pp. 603-632.

Waerden, van der, B. L. (1953), Ein neuer Test für das Problem in zwei Stichproben, Mathematics Annals, vol. 126, pp. 93-107.

Wilcoxon, F. (1945), Individual Comparisons by Ranking Methods, Biometrics Bulletin, vol. 1, pp. 80-83.

Wilcoxon, F. (1947), Probability Tables for Individual Comparisons by Ranking Methods, Biometrics, vol. 3, pp. 119-122.

Zimmerman, D.W. (1998), Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions, Journal of Experimental Education, vol. 67, pp. 55-68.

Zimmerman, D.W. (2000), Statistical Significance Levels of Nonparametric Tests Biased by Heterogeneous Variances of Treatment Groups, Journal of General Psychology, vol. 127, pp. 354-364.