Journal of Statistical Software

October 2008, Volume 28, Issue 3. *http://www.jstatsoft.org/*

lawstat: An R Package for Law, Public Policy and Biostatistics

Wallace Hui
University of Waterloo

Yulia R. Gel
University of Waterloo

Joseph L. Gastwirth
George Washington University

Abstract

We present a new R software package **lawstat** that contains statistical tests and procedures that are utilized in various litigations on securities law, antitrust law, equal employment and discrimination as well as in public policy and biostatistics. Along with the well known tests such as the Bartels test, runs test, tests of homogeneity of several sample proportions, the Brunner-Munzel tests, the Lorenz curve, the Cochran-Mantel-Haenszel test and others, the package contains new distribution-free robust tests for symmetry, robust tests for normality that are more sensitive to heavy-tailed departures, measures of relative variability, Levene-type tests against trends in variances etc. All implemented tests and methods are illustrated by simulations and real-life examples from legal cases, economics and biostatistics. Although the package is called **lawstat**, it presents implementation and discussion of statistical procedures and tests that are also employed in a variety of other applications, e.g., biostatistics, environmental studies, social sciences and others, in other words, all applications utilizing statistical data analysis. Hence, name of the package should not be considered as a restriction to legal statistics. The package will be useful to applied statisticians and "quantitatively alert practitioners" of other subjects as well as an asset in teaching statistical courses.

Keywords: goodness-of-fit tests, robust measures of location and scale, tests for symmetry, tests for randomness, tests for homogeneity of proportions, tests for equality of variances.

1. Introduction

Rapid developments in information technology have led to the fact that the nature of many legal cases, especially on antitrust law, securities law, equal employment, intellectual property and product liability, became very statistically oriented. This, in turn, implies an increasing demand for statisticians to serve as expert witnesses and, thus, a need for flexible, reliable and user-friendly software that provides modern tests and procedures in statistical science. More

importantly, the same statistical tests and procedures can be used with the same success in legal settings, biostatistics, finance, environmental studies and many other fields requiring data analysis. Unfortunately in many cases there is a minor or no interaction between practitioners of statistics in law, e.g., professional consultants, and applied statisticians working in other fields, which creates a gap between up-to-date findings and advances in theoretical statistics and methods used by practitioners in legal contexts. Hence, the goal of this package and illustrating it paper is not only to develop **a free** and **publicly available** software routines for statistical methods utilized in legal settings but more broadly, try to link similar statistical issues and problems arisen in a variety of cross-disciplinary research: from legal statistics to atmospheric sciences. Although the package is called **lawstat**, it presents implementation and discussion of statistical procedures and tests that are also employed in a broad range of other applications, e.g., biostatistics, environmental studies, social sciences and others, in other words, all applications utilizing statistical data analysis. Hence, name of the package should not be considered as a restriction to legal statistics. The package and illustrating it paper are intended for applied statisticians and "quantitatively alert practitioners" of law and other subjects as well as an asset in teaching applied statistical courses. Remarkably that since creation of the package **lawstat** in October 2006, most questions were received from researchers in medicine, epidemiology and genetics, with occasional feedback from practitioners in law and finance. In other words, the package is meant to be a *tool* for practical data analysis while the current paper can be considered as an extended manual for this tool. If one is interested in advanced discussion on statistics in the courtroom, use and misuse of statistical methodology in large areas of the law, legal doctrines and different standards of proof, we suggest to consult a number of specialized books on this subject, for example, "Statistical Reasoning in Law and Public Policy" by Gastwirth (1988), "Statistics for Lawyers" by Finkelstein and Levin (1990), "Prove It with Figures: Empirical Methods in Law and Litigation" by Zeisel and Kaye (1997), "Statistical Science in the Courtroom" by Gastwirth (2000), "Science In The Law: Standards, Statistics, and Research Issues" by Faigman, Kaye, Saks, and Saunders (2002).

We start from a short historical excurse on creating the package. The name and development of an R package **lawstat** (for more details on R see R Development Core Team 2007) is motivated by participation of J. L. Gastwirth and Y. R. Gel as statistical experts in a security law case on profit sharing allegation of an Investment Firm (IF). In most of the "profit sharing" cases, customers were alleged to have shared their profit by giving the broker or IF increased commission business on the day of the IPO or 1–2 days before or later. The analysis performed by the regulator relied on statistical procedures and tests, assuming that the observed data under consideration are independent and/or normally distributed. For example, the regulator compared the daily commissions on the IPO days to those on non-IPO days using the Wilcoxon test for equality of two means and found a statistically significant difference (the z-score of 2.37 yielding the p value of 0.02). However, the Wilcoxon test is applicable only under assumption that the observed data are uncorrelated. The effect of dependence on many non-parametric and goodness-of-fit tests is well known in statistical literature for long time, see, for example, Wolff, Thomas, and Gastwirth (1967); Serfling (1968); Gastwirth and Rubin (1975); Moore (1982); Keller-McNulty and McNulty (1987); El-Shaarawi and Damsleth (1988), but still is frequently overlooked in practical data analysis. We show that the commissions on IPO and Non-IPO days exhibit temporal (or serial) correlation, which leads to the inflation of the original p value. After taking into account temporal dependence among commission data, we obtained a much higher p value of 0.09, which contradicts with

the claim of a regulator. The second example refers to the violation of the assumption of normality. We can compare the hypothetical profit (HP), i.e., stock's opening price minus the price when the shares are resold, and the ratio of commissions to HP (C/HP). If there were some profit sharing agreement between a broker and a client, correlation between HP and C/HP is likely to stay around 0 or be positive. The Pearson correlation coefficient between HP and C/HP yields non-significant result of -0.088 with the p value of 0.569. However, both the HP and C/HP data are not normally distributed and, thus, the Pearson coefficient is not an appropriate measure of correlation. If we apply the Spearman distribution-free coefficient, we obtain a highly statistically significant negative correlation of -0.88 with a p values less than 0.0001. Negative Spearman's ρ implies that on days customers made most money, commissions formed a consistently lower percentage than on other IPO days (Gel, Miao, and Gastwirth 2005). Hence, the statistical findings are inconsistent with the profit sharing accusation.

As a result of the analysis performed for the security law case, we accumulated a substantial collection of various statistical tests and procedures, e.g., robust tests for normality and symmetry, robust graphical assessment of normality, the Bartels and runs randomness tests and the Brunner-Munzel test for equality of two means (also known as the Generalized Wilcoxon test), the Cochran-Mantel-Haenszel test for association among two categorical variables in each stratum while adjusting for control variables, which eventually became the R package **lawstat**. The original backbone collection of the R code presented in **lawstat** is constantly enriched by other statistical methods and procedures from legal cases, biostatistics and public policy, e.g., the Levene-typed tests for homogeneity of variances, the Levene directional tests against trends in volatility, the Lorenz curve, and others. Some of those methods are known and accepted in statistical analysis while others constitute new research ideas. The **lawstat** package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=lawstat`.

The paper is organized as follows. Each section is devoted to some statistical question and a test or a procedure associated with this question; each implemented test or procedure is illustrated by an application to a real-world data set. In particular, the Section 2 discusses tests for normality based on a robust estimator of scale. Such tests can be particularly useful if one is concerned with the impact of "very-low-probability-very-high-consequence events". Section 3 focuses on tests for homogeneity of variances in a few groups of observations. Section 4 deals with the tests for symmetry about an unknown median. In Section 5 we discuss tests for randomness that can be useful if a serial correlation among observations is suspected. The paper is concluded by final remarks in Section 6.

2. Robust procedures for testing normality

Many statistical tests and methods are based on the assumption that the data follow a normal distribution. Violation of the normality assumption can lead to unreliable or incorrect conclusions. A problem is known for a long time and there exists an extensive literature on how to assess a normal distribution. There are two types of methods: graphical assessment and quantitative goodness-of-fit tests. In the next subsections we discuss how various robust estimators of location and scale can be used in goodness-of-fit procedures for normality.

2.1. Robust quantile-quantile (RQQ) plots

One of the most popular graphical tests for normality is the quantile-quantile (QQ) plot. The QQ plot displays the sample data quantiles vs. the expected quantiles from a normal distribution. If the data are in fact normally distributed, the QQ plot shows a linear relationship between the observed and expected quantiles. To simplify "by eye" evaluation of a linear relationship, many software packages allow to add the best fitted line to the QQ plot, or so called quantile-quantile (QQ) line. The QQ line in base R (R Development Core Team 2007) is the line connecting the first and third observed quartiles. QQ plots are very popular exploratory technique that allows an immediate insight into how well the observed data satisfy the normality assumption and potential causes of non-normality, e.g., outliers, skewness, and short/long tails. However, it might not be easy to explain meaning of sample quartiles and the choice of such a QQ line to a person who has little background in statistics, which is frequently met in court hearings. Our approach is to standardize the observed data by subtracting the measure of cental location and dividing by the measure of scale, and then to produce a QQ plot. Hence, if the observed data are normally distributed then the standardized data follow $N(0, 1)$ and the sample quantiles should lie along the 45 degree line. Standardization of observed data prior to producing a QQ plot appeared before in other statistical software packages, e.g., SPSS (SPSS Inc. 2006). However, our idea is to utilize robust estimators of location and scale that are less sensitive to atypical observations rather than classical sample mean and standard deviation employed in other software. In particular, we use median as a measure of location and Hubert's median absolute deviation from a sample median (MAD) (Hall and Welsh 1985) or average absolute deviation from a sample median (MAAD) (Gastwirth 1982) as a measure of scale. MAAD is used to evaluate the fairness of tax assessments and defined by $J_n = (C/n)\sum_{i=1}^{n} |X_i - M|$, where $C = \sqrt{\pi/2}$, and $\{X\}_{i=1}^{n}$ are observations. We call such an approach a robust quantile-quantile (RQQ) plot (Gel et al., 2005). Our results show that in many practical situations RQQ plots provide clearer insight into the possible causes of non-normality than the usual QQ plots. The RQQ method is implemented as a generic function `rqq`. The user can choose various standardization schemes, i.e., mean or median as a location estimator and classical standard deviation, MAD or MAAD as a scale estimator. There also exists a choice for different QQ lines, i.e., a 45 degree line (the default option) or the QQ line adopted in base R which passes through the first and third quartiles. The default choice is to standardize data by median and MAD and produce a square RQQ plot. The function `rqq` has also an option to list all left- and right-tailed outliers with a user-prespecified α-significance level (by default $\alpha = 0.05$).

Remark. Note that the goal of the rqq function is to present a straightforward and easy-to-run graphical analysis for normality. To simplify utilization of this function, a user can choose to list which and how many observations fall outside of the expected 95% interval. In no way, this option can be considered as an advanced procedure for outlier detection, e.g., addressing multiple comparison issues etc, and is meant *only* as a simple "quick-and-dirty" graphical representation of *potential* outliers.

Let us illustrate the RQQ plot by assessing normality of a data set coming from the Zuni legal case (Zuni 2002; Gastwirth 2006; James 2007). In 1998 two Public School Districts from the state of New Mexico filed a lawsuit disputing the finance system for public schools. The lawcase is known as *Zuni Public School District No. 89 v. US Department of Education*. The two School Districts alleged the US Department of Education in misinterpreting the Federal Impact Aid Program and claimed that there was a shortage of about USD 180 millions since

1999 to educate Native American students due to incorrect calculation how the federal and state funding is to be distributed among schools. In April 2007 the Supreme Court sided with the federal government and the School Districts asked for a rehearing.

Here we analyze 89 measurements of revenue per pupil in each school district in New Mexico and assess whether the revenue follows a normal distribution. The Zuni data on the number of students and revenue per pupil in each school district are available in **lawstat**. We use the following R code to produce the RQQ plot and choose to list all potential outliers that fall outside of the expected standard normal $z_{0.025}$ and $z_{0.975}$ quantiles:

```
R> data("zuni")
R> rqq(zuni[,"Revenue"], scale = "J", line.it = TRUE,
+    outliers = TRUE, alpha = 0.025)

[1] left.tail.outliers <0 rows> (or 0-length row.names)
  right.tail.outliers
1            4.790642
2            7.728014
3            8.273093
4           10.480665
```

Figure 1 presents the usual QQ plot (left) and the RQQ plot (right) standardized by median and J (MAAD). Clearly there exist four very extreme outliers in the right tail. Besides those four outliers, the usual QQ plot suggests that the Zuni revenue data are almost normally distributed. In contrast, the RQQ plot also indicates substantial deviations from the 45 degree line in the left tail.

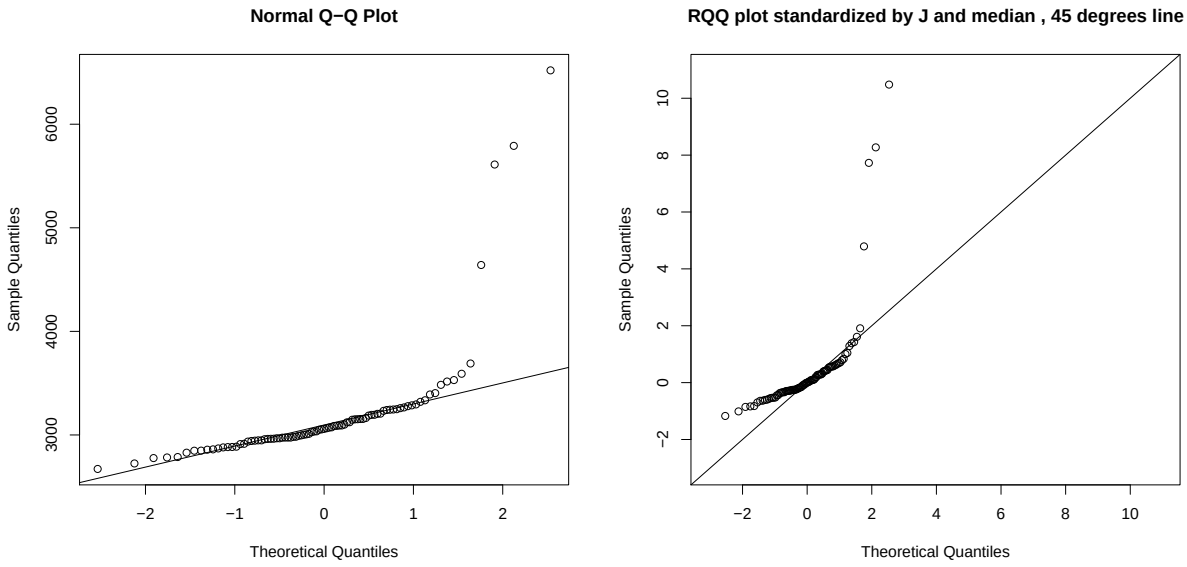

Figure 1: QQ and RQQ plots of the revenue data from the Zuni case. (The RQQ plot is standardized by median and J.)

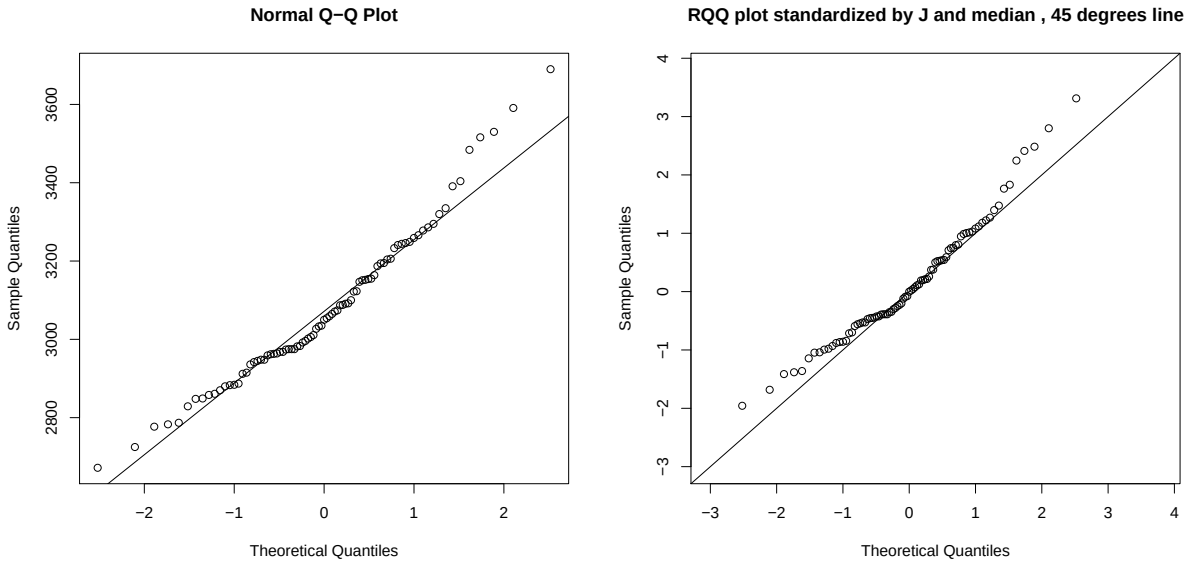

Figure 2: QQ and RQQ plots of the Zuni revenue data without the four largest outliers. (The RQQ plot is standardized by median and J.)

Figure 2 shows the usual QQ and RQQ plots of the revenue data after deleting the four extreme outliers. Now both QQ and RQQ plots present similar patterns and suggest that even after trimming the outliers, the revenue data are not normally distributed. In particular, there exist noticeable deviations from normality in both tails, i.e., the right tail is heavier and the left tail is shorter than expected for normally distributed data. The findings are consistent with the original results provided by the RQQ plot for the raw (untrimmed) revenue data and supported by histograms of the raw and trimmed data (see Figure 3).

2.2. Directed test against heavy tailed alternatives

Though graphical methods for testing normality provide an immediate idea on how closely the data of interest follow a normal distribution, frequently a quantitative estimate of such a closeness is required. One of the most popular goodness-of-fit methods for normality is the Shapiro-Wilk (SW) test. Though being a very powerful test that is applicable against all types of alternatives, i.e., being an omnibus test, SW does not provide information on specific causes of non-normality, i.e., outliers, skewness or heavy/short tails (D'Agostino and Stephens 1986; Shapiro and Wilk 1965). However, in many practical situations, especially in financial data modelling, a data analyst is particularly interested in detecting heavy tailed deviations from normality and a more specialized or directed testing procedure is needed. We propose two tests for normality that particularly focus on detecting heavy tailed alternatives. In both tests we utilize the average absolute deviation from the sample median (MAAD). Let $X_1, X_2, \ldots, X_n$ be a sample of independent and identically distributed random variables. Let μ, ν and σ be the population mean, median and standard deviation respectively. Let $\bar{X}$, M

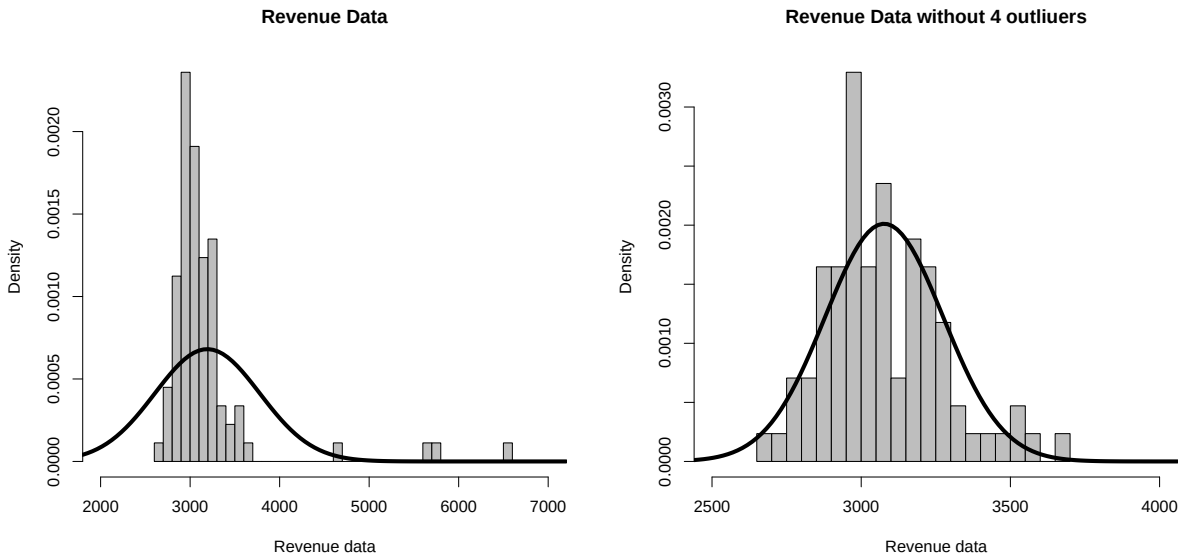

Figure 3: Histograms of the Zuni revenue data with and without the four largest outliers. A normal density curve (black line) with the same mean and standard deviation is superimposed.

and s_n be the corresponding sample estimates of μ, ν and σ. Then MAAD is defined as

$$J_n = \frac{C}{n}\sum_{i=1}^{n}|X_i - M|, \qquad C = \sqrt{\frac{\pi}{2}}, \tag{1}$$

which is shown to be a consistent estimate of scale under the null hypothesis of normality (Gastwirth 1982). The robust estimator J is implemented as `j.maad`.

Our first test statistic is based on the ratio of a classical estimate of a standard deviation s and the robust estimator J, i.e.,

$$R = \frac{s_n}{J_n}, \tag{2}$$

The intuitive idea of such a test statistic is that under the null hypothesis of normality, the ratio s/J should be close to 1. In contrast, for symmetric heavy tailed data, a less sensitive to deviations in the tails estimator J should be less than s. Hence, the ratio s/J is to be greater than 1. The test statistic R is asymptotically normally distributed if $X \sim N(\mu, \sigma)$ (Gel, Miao, and Gastwirth 2007), which implies the following one-sided rejection region

$$\text{reject } H_0 : \text{normality, if } \frac{\sqrt{n}(R-1)}{\sigma_R} \geq z_\alpha, \tag{3}$$

where z_α is the upper α-percentile of a standard normal distribution.

The proposed new test (SJ) is essentially a directed test focusing on detecting heavy tails and outliers and is not expected to have good power against skewed alternatives. The SJ test has a similar idea as the Bonnet-Seier (BS) test (Bonett and Seier 2002) that is based on the log-transformation of the Geary G-kurtosis. The Bonett-Seier test is implemented in the

R package **moments** as `bonett.test`; see Komsta and Novomestky (2007). The G-kurtosis utilizes an absolute moment around a sample mean while our SJ test is based on a more robust estimator J. As shown by Gel *et al.* (2007), the SJ test typically yields equal or higher power then the SJ test; hence, the BS test is omitted from further power comparison. The new SJ test is implemented as `sj.test` with an option to use approximation of critical values by a t-distribution (by default) or to estimate the exact critical values from Monte Carlo (MC) simulations. Number of MC simulations can be selected by a user. The authors recommend to use the exact critical values for small sample sizes (≤ 30 observations). The output of `sj.test` includes a standardized test statistic, the ratio s/J and a p value. Here is the example of the SJ test applied to the Zuni revenue data. The critical values are calculated using 1000 MC simulations.

```
R> sj.test(zuni[,"Revenue"], crit.values = "empirical", N = 1000)

        Test of Normality - SJ Test

data:  zuni[, "Revenue"]
Standardized SJ Statistic = 27.4543, ratio of S to J = 1.774, p-value <
2.2e-16
```

The second test is a robust modification of the Jarque-Bera (JB) test for normality (Jarque and Bera 1980; Bowman and Shenton 1975). In finance and econometrics, JB is the most widely used test for detecting heavy tailed alternatives with a minor degree of skewness. The original JB test statistic is a sum of squared sample measures of skewness and kurtosis

$$JB = \frac{n}{6} b_1 + \frac{n}{24}(b_2 - 3)^2. \tag{4}$$

Here $\sqrt{b_1} = \hat{\mu}_3/\hat{\mu}_2^{3/2}$ is the sample skewness and $b_2 = \hat{\mu}_4/\hat{\mu}_2$ is the sample kurtosis. The JB test statistic is asymptotically χ^2_2-distributed.

Since it is well known that the sample moments are even more sensitive to outliers than the sample mean, we propose to utilize the robust estimator of scale J in the denominators of skewness and kurtosis. Hence, we obtain the following new test statistic

$$RJB = \frac{n}{C_1}\left(\frac{\hat{\mu}_3}{J_n^3}\right)^2 + \frac{n}{C_2}\left(\frac{\hat{\mu}_4}{J_n^4} - 3\right)^2, \tag{5}$$

where C_1 and C_2 are positive constants. Gel and Gastwirth (2008) show that the new RJB test statistic follow an asymptotic χ^2-distribution with two degrees of freedom and propose to use the normalizing constants $C_1 = 6$ and $C_2 = 64$. Consequently, the one-sided rejection region is

$$\text{reject } H_0 : \text{normality, if } RJB \geq \chi^2_{1-\alpha,2}, \tag{6}$$

where $\chi^2_{1-\alpha,2}$ is the upper α-percentile of the χ_2-distribution with 2 degrees of freedom.

Our studies indicate that the new RJB and the classical JB test statistics converge to the asymptotic χ^2_2-distribution very slowly. Hence, the exact (empirically calculated using the Monte Carlo simulation) critical values are to be utilized for small and moderate sample

sizes, which is also supported by the results of Poitras (2005) and Thadewald and Buning (2006). The new test is implemented as `rjb.test` with an option for the robust Jarque-Bera (RJB) or classical Jarque-Bera (JB) tests. The function `rjb.test` extends the code for the classical Jarque-Bera test `jarque.bera.test` from the **tseries** in R (see Trapletti and Hornik 2007). The user also can choose to use either the χ^2_2-approximated critical values (by default) or the exact (Monte Carlo simulated) critical values with an optional number of simulations. We suggest to use the empirical critical values for small and moderate sample sizes for both JB and RJB tests. The `rjb.test` output contains the name of a chosen test, i.e., RJB or JB, the test statistic and the corresponding p value. Below we illustrate application of `rjb.test` to the Zuni revenue data.

```
R> rjb.test(zuni[,"Revenue"], crit.values = "empirical", N = 1000)

     Robust Jarque Bera Test

data:  zuni[, "Revenue"]
X-squared = 59035.41, df = 2, p-value < 2.2e-16

R> rjb.test(zuni[,"Revenue"], option = "JB", crit.values = "empirical",
+    N = 1000)

      Jarque Bera Test

data:  zuni[, "Revenue"]
X-squared = 1340.501, df = 2, p-value < 2.2e-16
```

Our findings indicate that all three tests, i.e., SJ, RJB and JB, agree that the raw Zuni revenue data are not normally distributed, and yield highly statistically significant p values. The most popular omnibus Shapiro-Wilk (SW) test `shapiro.test` provides a similar p value of $1.149 \cdot 10^{-15}$. If we apply all four tests to the revenue data without four extreme outliers, then RJB, JB and SW agree in their conclusions and yield p values of 0.03, 0.02 and 0.02 respectively, while a p value of SJ is 0.25. The result is expected since the SJ test focuses on symmetric deviations in the tails, while the trimmed revenue data are somewhat skewed. The conclusions of RJB, JB and SW to reject normality of the trimmed revenue data confirm findings of the RQQ plot and histograms in Figures 1 and 3 respectively.

Next we perform the power comparison study among the SJ, RJB, JB and SW tests applied to normal inverse Gaussian (NIG) distributions (Atkinson 1982; Barndorff-Nielsen and Blaesild 1983). The flexible closed form of such distributions makes NIG very attractive for a variety of applications and, in particular, for modelling heavy-tailed financial processes. For more on applications of NIG distributions see, for example, Barndorff-Nielsen (1997), examples from the R package **fBasics** (Wuertz 2007) and references therein. The NIG distribution is completely determined by the four parameters (α, β, μ and $\delta \in R$). The parameters can be naturally interpreted in terms of a shape of the resulting probability density function, and appropriate tuning of α, β, μ and δ can describe a wide range of continuous probability distributions (see Figure 4 and Table 1). To simulate samples from a NIG distribution, we utilize the R package **fBasics**.

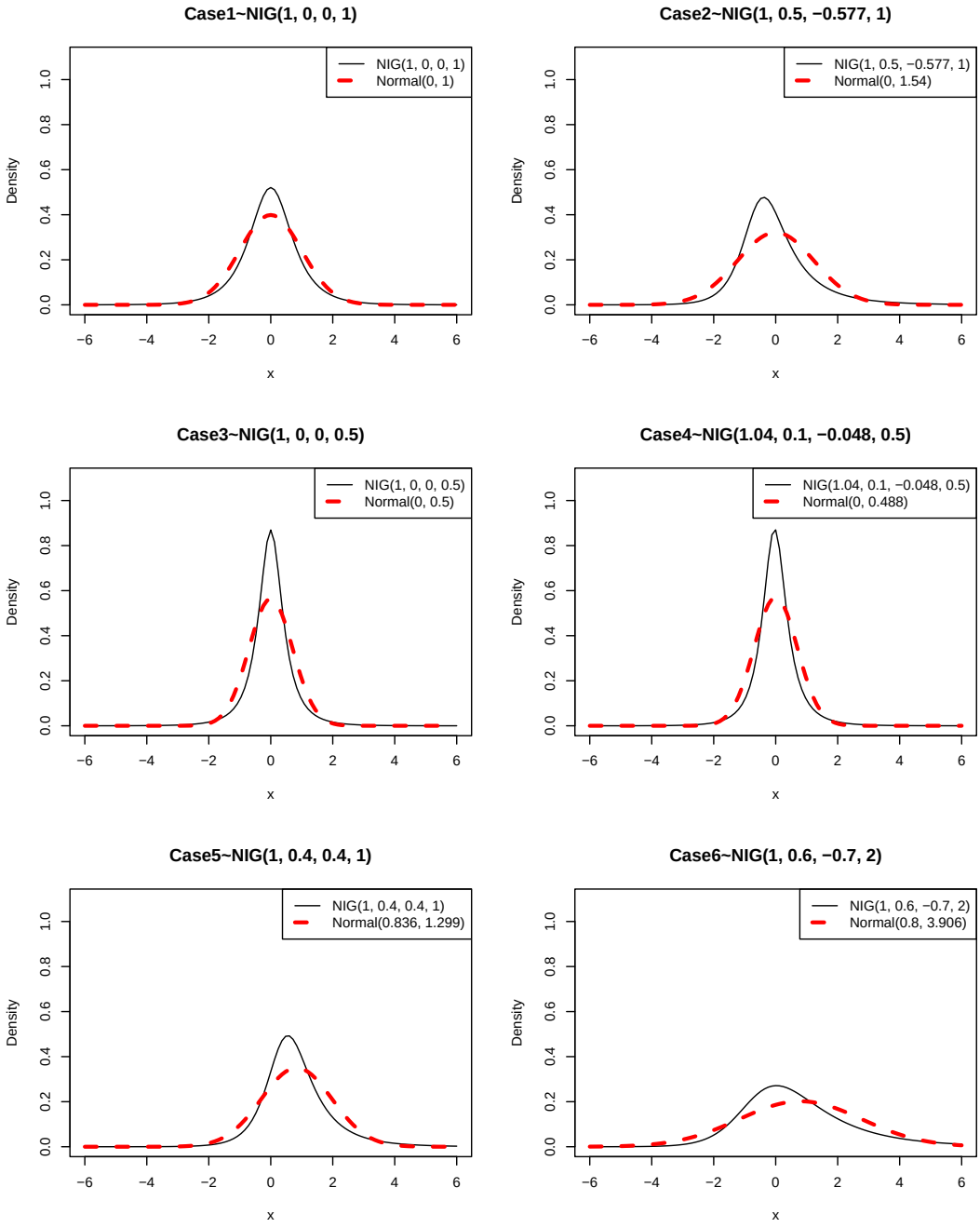

Figure 4: Plot of probability density functions of NIG distributions with various parameters α, β, μ and δ, with the superimposed normal probability density curves with the same mean and standard deviation.

α	=	shape parameter, $\alpha \in \mathcal{R}$	Mean	=	$\mu + \frac{\beta\delta}{\gamma}$
β	=	skewness parameter, $0 \leq \lvert\beta\rvert \leq \alpha,\ \beta \in \mathcal{R}$	Variance	=	$\frac{\delta\alpha^2}{\gamma^3}$
μ	=	location parameter, $\mu \in \mathcal{R}$	Skewness	=	$\frac{3\beta}{\alpha\sqrt{\delta\gamma}}$
δ	=	scale parameter, $\delta > 0,\ \delta \in \mathcal{R}$	Kurtosis	=	$\frac{3\left(1+\frac{4\beta^2}{\alpha^2}\right)}{\delta\gamma}$
γ	=	$\sqrt{\alpha^2-\beta^2},\ \gamma \in \mathcal{R}$	$M_X(z)$	=	$e^{\mu z+\delta(\gamma-\sqrt{\alpha^2-(\beta+z)^2})}$

Table 1: Mean, variance, skewness, kurtosis and moment generating function of a normal inverse Gaussian (NIG) distribution.

$NIG(\alpha,\beta,\mu,\delta)$		Mean	Var	$\sqrt{b_1}$	b_2	Size	Percentage of Rejections			
						n	RJB	JB	SJ	SW
Case 1:	$\alpha = 1$					30	0.559	0.497	0.577	0.454
	$\beta = 0$	0	1	0	3	50	0.776	0.707	0.794	0.650
	$\mu = 0$					70	0.857	0.809	0.888	0.773
	$\delta = 1$					100	0.936	0.903	0.961	0.885
Case 2:	$\alpha = 1$					30	0.756	0.742	0.715	0.741
	$\beta = 0.5$	0	1.540	1.612	6.928	50	0.923	0.907	0.890	0.914
	$\mu = -0.577$					70	0.979	0.976	0.971	0.975
	$\delta = 1$					100	0.995	0.991	0.985	0.994
Case 3:	$\alpha = 1$					30	0.722	0.658	0.753	0.627
	$\beta = 0$	0	0.500	0	6	50	0.901	0.853	0.926	0.834
	$\mu = 0$					70	0.962	0.932	0.976	0.927
	$\delta = 0.5$					100	0.981	0.975	0.989	0.979
Case 4:	$\alpha = 1.04$					30	0.731	0.661	0.773	0.631
	$\beta = 0.1$	0	0.488	0.401	6.010	50	0.899	0.851	0.925	0.823
	$\mu = -0.048$					70	0.965	0.931	0.976	0.922
	$\delta = 0.5$					100	0.991	0.975	0.998	0.981
Case 5 :	$\alpha = 1$					30	0.679	0.665	0.657	0.644
	$\beta = 0.4$	0.836	1.299	1.253	5.368	50	0.873	0.854	0.862	0.849
	$\mu = 0.4$					70	0.941	0.921	0.937	0.924
	$\delta = 1$					100	0.988	0.983	0.984	0.986
Case 6:	$\alpha = 1$					30	0.721	0.75	0.617	0.779
	$\beta = 0.6$	0.800	3.906	1.423	4.575	50	0.906	0.914	0.806	0.929
	$\mu = -0.7$					70	0.967	0.971	0.903	0.975
	$\delta = 2$					100	0.993	0.994	0.965	0.996

Table 2: The power comparison of the RJB, JB, SJ and SW tests for $\alpha = 0.05$ and 1000 Monte Carlo simulations. The exact critical values are utilized for RJB, JB and SJ. Measures of skewness and kurtosis are denoted as $\sqrt{b_1}$ and b_2 respectively.

The results in Table 2 show that for symmetric or slightly skewed heavy tailed alternatives (Cases 1, 3 and 4), the best performance for all samples is provided by the SJ test. The next best result is yielded by the RJB test which is followed by the JB and SW tests. If the degree of skewness increases and also there exist heavy-tailed deviations (Cases 2 and 5), then the RJB test provides the best results for all sample sizes. In case 2, JB and SW perform similarly

while SJ is shown to be the least powerful. In case 5 that is less skewed than the case 2, SJ and JB provide similar performance and are followed by the SW test. With higher increase of skewness (Case 6), the SW test becomes the most powerful and is followed by JB, RJB and finally SJ.

Our conclusion is that for symmetric or slightly skewed alternatives with heavy tails, the SJ test is preferred. For moderately skewed heavy-tailed deviations, the RJB test becomes the most preferable test. If the degree of skewness increases, the most powerful procedure is the SW test. Note that the RJB test typically outperforms the JB test in detection of all symmetric or moderately skewed heavy tailed alternatives that are of particular importance for applications in finance and econometrics. Overall, in many practical situations it makes sense to apply a few goodness-of-fit tests and see how close the obtained findings are; if there are any doubts on a specific cause of non-normality, graphical methods such as QQ or RQQ plots can be of substantial help for applied data analysis.

3. Levene's family of tests for equality of variances

In a variety of applications, a data analyst needs to assess whether variances of different samples are equal. For example, it is more desirable to use a lab with equipment that provides the least variability during calibration experiments; volatilities of various stocks need to be compared to manage investment risks; among a number of qualifying tests, an instructor might wish to choose the exam providing the highest spread of the results for further classification purposes.

The problem of assessing homogeneity of variances has a long history and there exists a substantial number of related procedures. Many such tests rely on the assumption of normality and are not robust to its violation. In 1960, Prof. Howard Levene proposed a new approach for testing equality of variances which is essentially the F test computed on the absolute deviations of observations from the group mean. Levene's approach is shown to be a powerful and robust to non-normality test and quickly became a very popular tool for assessing equality of variances in various applications, e.g., clinical trials, astronomy, marine pollution, business, auditing and law cases.

Consider k random samples $x_{i1}, x_{i2}, \ldots, x_{in_i}$ from the i-th population with unknown mean μ_i, variance σ_i^2 and distribution function $F(\cdot)$, $i = 1, 2, \ldots, k$. The null hypothesis is that all group variances are equal

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2,$$

vs. the alternative hypothesis

$$H_1 : \sigma_i^2 \neq \sigma_j^2, \text{ for at least one } i \neq j.$$

Levene (1960) considered various monotonic smooth transformations $G(\cdot)$ of absolute differences between each observations and the corresponding sample group mean, i.e., $d_{ij} = |x_{ij} - \bar{x}_{i.}|$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$. Here $\bar{x}_{i.}$ denotes the sample mean in the ith group. For example, G can be square, log, square-root or identity functions. Choosing finally $G(\cdot)$ to be an identity function and treating d_{ij} as independent normally distributed random variables, Levene applied the classical F test to d_{ij}, which leads to the test statistic

$$L = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k}(\bar{d}_{i.}-\bar{d}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(d_{ij}-\bar{d}_{i.})^2},$$
$$\bar{d}_{i.} = \sum_{j=1}^{n_j} d_{ij}/n_j, \quad \bar{d}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_j} d_{ij}/N. \tag{7}$$

The F test is applicable to independent and normally distributed observations. Clearly, d_{ij} are not $N(\mu_i, \sigma_i^2)$ and correlation among d_{ij} in the same ith group is shown to be of order $1/n_i^2$. However, Levene's test statistic is revealed to be robust to these violations.

Levene's test is implemented as `levene.test`. The user can choose various estimators of a group center, i.e., a sample mean suggested in the original work of Levene (1960) (the default option); a sample median proposed by Brown and Forsythe (1974), which is also known as a modified Levene's test; or a trimmed sample mean with a user prespecified α-level of trimming. The function `levene.test` is modified from the response posted by Brian Ripley to the R-help mailing list. A similar function for Levene's test is included in the R package **car** (Fox 2002). However, Levene's function in **car** enables to use only the classical Levene test with group centers being a sample mean. Such Levene's test can yield an incorrect size of the test for skewed distributions (Miller 1968; Carroll and Schneider 1985). Hence, the modified Brown-Forsythe-Levene test based on a sample median or robustly trimmed Levene test are frequently preferred. Our `levene.test` allows to utilize all three options for a group center estimate. If the trimmed mean is selected, the authors suggest to use a heavily trimmed estimator with $\alpha = 0.25$ applied to both tails (Gastwirth, Gel, and Miao 2008). The output includes a value of Levene's test statistic and a corresponding p value.

Frequently a particular pattern in group variances is suspected, e.g., a monotonic increasing or decreasing trend. Hence, the test hypothesis takes the form

$$H_1 : \sigma_1^2 \le \sigma_2^2 \le \ldots \le \sigma_k^2, \quad \text{increasing trend}$$
$$H_1 : \sigma_1^2 \ge \sigma_2^2 \ge \ldots \ge \sigma_k^2, \quad \text{decreasing trend.}$$

Assign all observations in group i a score w_i, $i = 1, \ldots, k$. Now regress d_{ij} on w_i and consider the regression slope

$$\hat{\beta} = \frac{\sum_{i=1}^{k} n_i(w_i - \bar{w})(\bar{d}_{i.} - \bar{d}_{..})}{\sum_{i=1}^{k} n_i(w_i - \bar{w})^2}, \quad \bar{w} = \sum_{i=1}^{k} w_i.$$

Under the null hypothesis, $\hat{\beta} = 0$ and the new Levene-type trend statistic follows a t-distribution with $(N-1)$ degrees of freedom, i.e., $\hat{\beta}/\text{s.e.}(\hat{\beta}) \sim t_{N-1}$.

Various scores can be chosen to reflect the monotonic increasing or decreasing behavior of group variances, e.g., linear trend $w_i = i$, non-linear trends $w_i = i^2$ or $w_i = \ln(i)$, in a similar way as proposed by Neuhauser and Hothorn (2000). If no information is known about the most appropriate scores, the authors suggest to use the linear scores of $w_i = i$. Gel and Gastwirth (2008) show that linear scores yield satisfactory performance even for detection in non-linear trends in variances.

Levene's trend test is implemented as a function `ltrend.test` with the same set of options for group center estimators, i.e., a sample mean, median or α-trimmed mean (the trimming

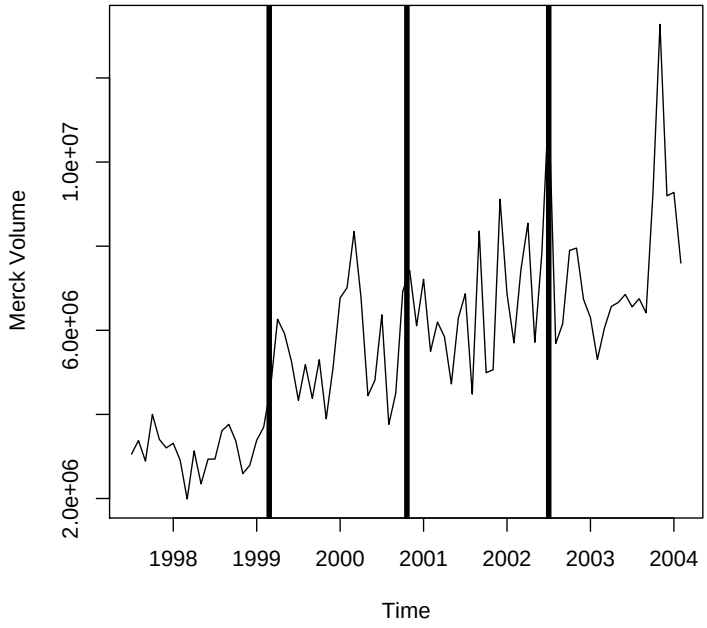

Figure 5: The Merck trade volume data from July 1997 to February 2004.

α of 0.25 is the default choice). In addition, the user can pre-specify the scores w_i to be utilized in the testing procedure using the option `score`. The default choice is to use the linear scores that are computed from a factor variable `group`; in this case, a user does not need to provide an additional input to `score`. The output includes Levene's trend statistic and the corresponding p value.

We illustrate application of Levene's and Levene's trend test by analyzing temporal volatility of Merck's monthly trade volume data. Merck & Co., Inc. (Merck) is a global pharmaceutical company that discovers, develops, manufactures and markets a range of products to improve human and animal health. It is now one of the top 7 largest pharmaceutical companies in the world both by capital and revenue.

The time series plot of Merck's trade volume monthly data from July 1997 to February 2004 is shown in Figure 5. The data consists of 80 observations. We divide the data into four temporal periods and apply Levene's and Levene's trend tests with a group center being a 25% trimmed mean. Here `mrk.vol` is the Merck trading volume and `temp.period` is a group variables consisting of 4 factors, i.e., 4 temporal periods. The scores are not specified and the default option for linear scores is utilized.

```
R> levene.test(mrk.vol, temp.period, option = "trim.mean",
+    trim.alpha = 0.25)

    Modified Robust Levene-type test based on the absolute deviations from
        the trimmed mean
```

Test	$(\cdot)^2$	BF	$trim_{0.25}$
Levene's	0.0004	0.0252	0.0081
LTrend	4.486e-05	0.0038	0.0011

Table 3: Levene's and Levene's trend tests with the linear scores $1, 2, 3$ for the Merck volume data from July 1997 to February 2004.

```
data:  mrk.vol
Test Statistic = 4.2227, p-value = 0.008129

R> ltrend.test(mrk.vol, temp.period, option = "trim.mean",
+    trim.alpha = 0.25)

      ltrend test based on the modified Levene-type procedure using the
        group trimmed means

data:  mrk.vol
Test Statistic = 313656.8, p-value = 0.001092
```

The summary of Levene's and Levene's trend tests with three options for a group center estimate are presented in Table 3. All tests support the alternative hypothesis of unequal variances. Remarkably, all three p values yielded by various versions of Levene's trend test are noticeably smaller than p values of Levene's tests. These findings indicate that there very likely exists a monotonic increase in volatility of Merck's trading volume, which is also supported by the Figure 5.

Remark. An important issue is how legitimate is to use Levene's and Levene's trend tests for the Merck volume example, since both tests assume that the data under consideration are independent, while the volume data can be serially correlated in time. Hence, we need to investigate dependence structure for the Merck data. Figure 6 presents plots of sample autocorrelations (acf) for each considered time segment. Except of the period between November, 2000 and June, 2002, which shows a somewhat significant correlation at lag 3, there exists no evidence of correlation among the data. To verify our findings, we also run the Box-Pierce (BP) test for all the four periods up to the tenth lag; for discussion on the Box-Pierce test see Cromwell, Walter, and Terraza (1994). The BP tests fails to reject the null hypothesis of the Merck volume data being uncorrelated in time, for all test runs. In particular, for somewhat questionable correlation at lag 3 for the period November, 2000–June, 2002, the BP test yields a p value of 0.09. Thus, we can conclude that for this particular set of the data, employment of Levene's and Levene's trend tests is legitimate. Note that with an increase of a sample size, the dependence structure is to be re-investigated.

4. Tests for symmetry about an unknown median

Suppose that we get a sample of independent observations $X_1, X_2 \ldots, X_n$ from a continuous distribution F with the probability density function f, unknown mean μ, median ν and

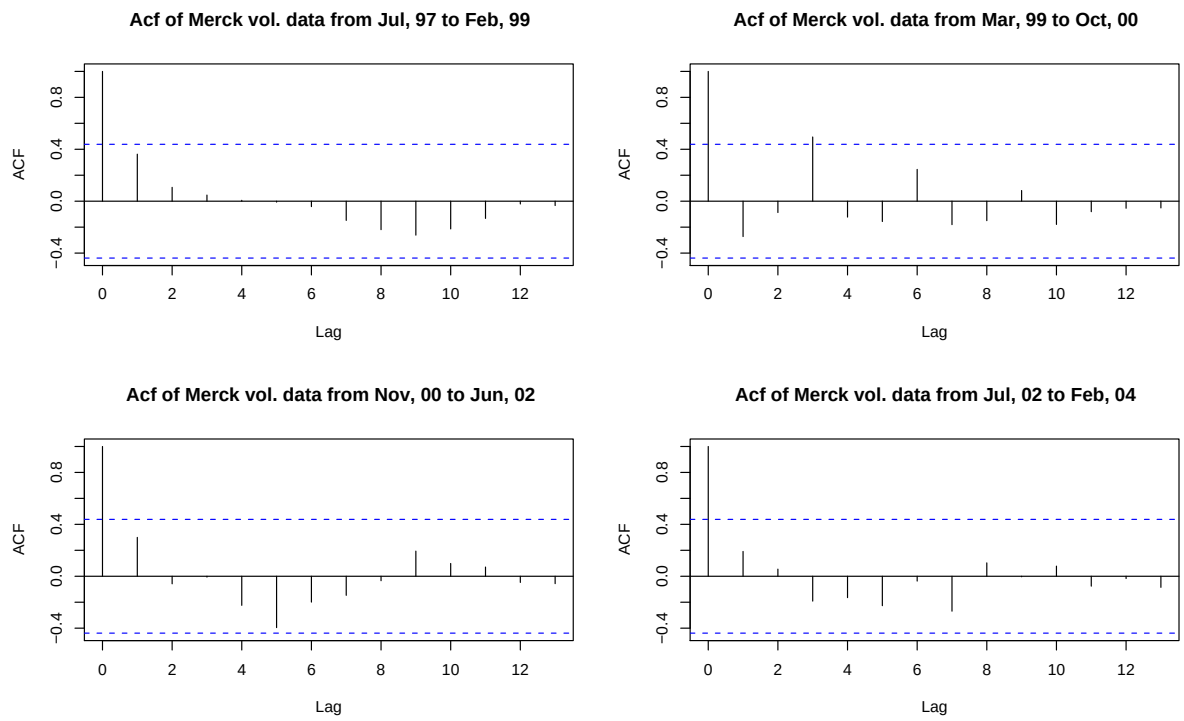

Figure 6: The plots of sample autocorrelation functions for four considered periods: July, 1997–February, 1999, March, 1999–October, 2000, November, 2000–June, 2002, and July, 2002–February, 2004.

standard deviation σ. Our target is to test whether F is symmetric about ν. Hence, the test hypothesis is given by

$$\begin{aligned} H_0 : f(\nu - x) &= f(\nu + x) \\ H_A : f(\nu - x) &\neq f(\nu + x). \end{aligned}$$

There exists a substantial number of papers discussing how to test symmetry if the median ν is given (see Einmahl and McKeague (2003); Annaert, Brys, and Ceuster (2005), for a review). If a median ν is unknown, the testing problem becomes much more complicated and still remains an active area of research in statistics.

In this paper, we discuss implementation of three recent tests for symmetry about the unknown median ν. The first procedure, proposed by Cabilio and Masaro (1996), is based on the test statistic

$$C = \frac{\bar{X} - M}{s_n},$$

where $\bar{X}$, M and s_n are the sample mean, median and standard deviation respectively. The statistic $\sqrt{n}C$ is asymptotically normally distributed. However, the asymptotic variance of C depends on the underlying distribution F that is generally unknown. Cabilio and Masaro (1996) propose to utilize the asymptotic variance of 0.5708 that is derived for F being a standard normal distribution, and argue that the effect of misspecification is relatively minor for most practical purposes.

Mira (1999) suggests an alternative test statistic based on the Bonferroni measure of skewness γ_1, i.e.,

$$\gamma_1 = 2(\bar{X} - M).$$

The asymptotic distribution of $\sqrt{n}\gamma_1$ is also normal and depends on F. Mira (1999) adopted the same approach of approximating an asymptotic variance of $\sqrt{n}\gamma_1$ by an asymptotic variance under assumption of F being a standard normal distribution.

Miao, Gel, and Gastwirth (2006) modify the Cabilio-Masaro procedure and propose a test statistic

$$T = \frac{\bar{X} - M}{J_n}, \qquad J_n = \frac{C}{n}\sum_{i=1}^{n}|X_i - M|, \quad C = \sqrt{\frac{\pi}{2}}.$$

Similarly to statistics $\sqrt{n}C$ and $\sqrt{n}\gamma_1$, $\sqrt{n}T$ is normally distributed as $n \to \infty$ and the asymptotic variance of $\sqrt{n}T$ also depends on a generally unknown F. Based on a study of the size of the test, Miao *et al.* (2006) conclude that approximation of a true asymptotic variance by an asymptotic variance of T under $X \sim N(0, 1)$ is feasible for practical purposes and the impact of misspecification is relatively minor (Miao *et al.* 2006).

All three tests are implemented as a function `symmetry.test` with a choice for the Cabilio-Masaro test, the Mira test and the T test denoted as MGG. The MGG procedure is the default option.

Here is the example on testing symmetry of the prediction errors of 48-hour ahead MM5 forecasts of surface temperature, measured at 96 different locations in the US Pacific Northwest on January 3, 2000. Data have been kindly provided by the research group of Professor Clifford Mass in the Department of Atmospheric Sciences at the University of Washington. The prediction error, or "bias", is the difference between the forecasted and observed surface temperature. MM5 is the fifth-generation Pennsylvania State University and National Center for Atmospheric Research Mesoscale Model. The data are available from **lawstat** as `bias`. If it turns out that the distribution of the forecasting errors is skewed, it will imply that the predicted surface temperature is consistently over/underestimated.

```
R> data("bias")
R> symmetry.test(bias)

      Test of Symmetry - MGG Test

data:  bias
Test Statistic = 1.3779, p-value = 0.1682

R> symmetry.test(bias, option = "cabilio.masaro")

         Test of Symmetry - Cabilio-Masaro Test

data:  bias
Test Statistic = 1.2898, p-value = 0.1971

R> symmetry.test(bias, option = "mira")
```

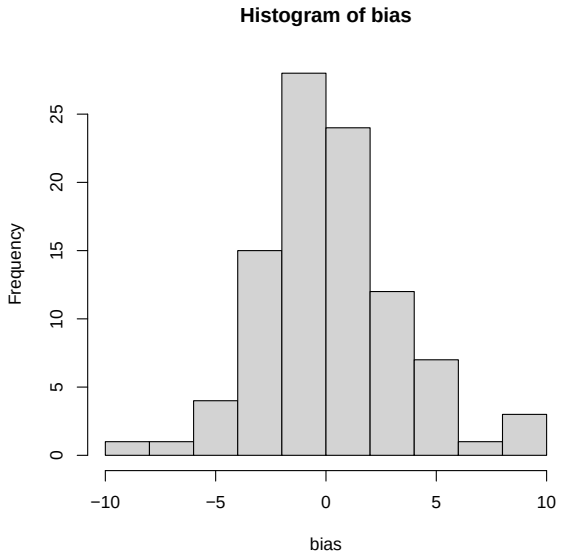

Figure 7: The histogram of the surface temperature error forecasts.

```
        Test of Symmetry - Mira Test

data:  bias
Test Statistic = 1.3209, p-value = 0.1865
```

Here all three tests fail to reject the null hypothesis for symmetry, which coincides with findings of the histogram plot for the bias data (see Figure 7). Thus, we can conclude that the forecasting errors are symmetric.

The Table 4 presents a small power comparison study of the three tests for symmetry applied to NIG distributions with various degrees of skewness. For more extensive simulation study see Miao *et al.* (2006). Our findings indicate that for the distributions considered, the MGG test statistic is generally preferred for all cases and sample sizes; however, none of the tests have power against alternative 3.

5. Tests for randomness

The assumption that the observed data form a random sample is the key condition for validity of a variety of statistical procedures, from classical regression analysis to non-parametric tests. In fact, one of the main assumptions of the least squares (LS) method in linear regression is randomness (uncorrelatedness) of the residuals. If the data are normally distributed, there exists a number of powerful tests based on sample autocorrelation functions (acf), e.g., the portmanteau class of tests. However, if the underlying distribution is not normal, the acf based tests may be substantially affected by deviations from normality (Bartels 1977). The alternative is to use the non-parametric (distribution-free) tests for randomness.

$NIG(\alpha, \beta, \mu, \delta)$		Mean	Var	$\sqrt{b_1}$	b_2	Size	Percentage of Rejections		
						n	MGG	Mira's	CM
Case 1:	$\alpha = 1.14$					30	0.103	0.044	0.040
	$\beta = 0.2$	0	0.919	0.497	3.002	50	0.153	0.089	0.091
	$\mu = -0.178$					70	0.181	0.109	0.110
	$\delta = 1$					100	0.245	0.152	0.146
Case 2	$\alpha = 1$					30	0.445	0.190	0.241
	$\beta = 0.5$	0	1.540	1.612	6.928	50	0.642	0.412	0.461
	$\mu = -0.577$					70	0.785	0.593	0.637
	$\delta = 1$					100	0.903	0.792	0.817
Case 3	$\alpha = 1.04$					30	0.103	0.023	0.031
	$\beta = 0.1$	0	0.488	0.401	6.010	50	0.134	0.047	0.052
	$\mu = -0.048$					70	0.155	0.062	0.068
	$\delta = 0.5$					100	0.149	0.066	0.069
Case 4	$\alpha = 1$					30	0.182	0.055	0.076
	$\beta = 0.25$	0	0.551	1.078	7.746	50	0.245	0.093	0.108
	$\mu = -0.129$					70	0.316	0.144	0.170
	$\delta = 0.5$					100	0.439	0.204	0.243

Table 4: The power comparison of the Mira, Cabilio-Masaro and MGG tests of symmetry for $\alpha = 0.05$ and 1000 Monte Carlo simulations. Measures of skewness and kurtosis are denoted as $\sqrt{b_1}$ and b_2 respectively.

Bartels (1982) proposed a test statistic based on von Neumann's ratio (RVN)

$$RVN = \sum_{i=1}^{T-1}(R_i - R_{i+1})^2 / \sum_{i=1}^{T}(R_i - \bar{R})^2, \tag{8}$$

where R_i is the rank of the i-th observation in a sample $X_1, \ldots, X_T$. The asymptotic range of the test statistic (8) is between zero and four and the expected value is two (under the null hypothesis of randomness), which is similar to the Durbin-Watson (DW) test statistic (Cromwell *et al.* 1994). The Durbin-Watson test is available as `dwtest` in the R package **lmtest** (Zeileis and Hothorn 2002). As shown by Bartels (1982), the rank-based statistic $\sqrt{T}RVN$ asymptotically follows a normal distribution and is more powerful than the distribution-free runs test proposed by Wald and Wolfovitz (1943). Note that by construction the test statistic (8) concerns the first order serial correlation among the data, i.e., whether the two consecutive observations are dependent, and is not intended for detecting higher order correlations among the data points. The Bartels test is implemented as `bartels.test`. The output contains the Standardized Bartels statistic that follows $N(0, 1)$, the RVN statistic (8) and a corresponding p value. The user can choose to test against positive correlation (the test statistic (8) is close to 0 and the Standardized Bartels statistic is far in the left tail of $N(0, 1)$), against negative correlation (the statistic (8) is close to 4 and its standardized version is far in the right tail of $N(0, 1)$) or against general existence of correlation (the default option).

Along with the Bartels test, we implement the the Wald-Wolfowitz runs test (Wald and Wolfovitz 1943; Cromwell *et al.* 1994). The runs test is also available from the R package **tseries** (see Trapletti and Hornik 2007). However, the input to `runs.test` in **tseries** is a a dichotomous factor where factoring is arbitrarily defined by a user. In contrast, in our

case the input is a numerical vector which is later factored into two groups in such a way that each group contains observations either above or below a sample median. A run is a consecutive sequence of observations that are all above (or below) the reference line (median). The number of observations in each run defines the length of a run. One can compare how many runs are observed vs. the expected number of runs. If there are two many runs, then a negative serial correlation is likely to present. If there are too few runs, the data are likely to exhibit a positive correlation. The standardized runs test statistic asymptotically follows $N(0, 1)$. Similarly to the Bartels test, the runs test focuses only on the first order serial correlation, i.e., linear dependence among neighboring observations. The function `runs.test` enables to choose either a two-sided alternative (the default option) or a specific testing against positive (negative) correlation, i.e., left (right) tailed alternatives respectively. Choosing an option `plot.it`, the user can also visualize runs as a sequence of A and B, where A and B corresponds to observations above or below median respectively.

Let us illustrate application of the Bartels and runs tests to the forecasting errors of 48-hour ahead MM5 forecasts of surface temperature, measured at 96 different locations in the US Pacific Northwest on January 3, 2000. The errors, sometimes referred as "bias", are differences between the predicted and observed surface temperature and are available from **lawstat** as `data(bias)`. It is important to check that forecasting errors are uncorrelated, since if forecasting errors turn out to be correlated, it implies that the employed numerical weather prediction (NWP) model does not filter well all the available information and future weather forecasts are hence biased and inaccurate.

```
R> data("bias")
R> runs.test(bias, plot.it = TRUE)

        Runs Test - Two sided

data:  bias
Standardized Runs Statistic = -0.8208, p-value = 0.4117

R> bartels.test(bias)

        Bartels Test - Two sided

     Bartels Test - Two sided

data:  bias
Standardized Bartels Statistic = -1.1188, RVN Ratio = 1.772, p-value =
0.2632
```

As we see from the output of both tests and Figure 8, the null hypothesis of randomness is not rejected and there is no noticeable indication of a first order serial correlation in the surface temperature prediction errors, which is a good sign that MM5 weather forecasting procedures perform reasonably well. If we plot the sample autocorrelation (acf) functions (see Figure 8), we notice that all acf are insignificant except of an acf at the third lag. Since both the Bartels and runs test do focus only on detecting first order serial correlations, we

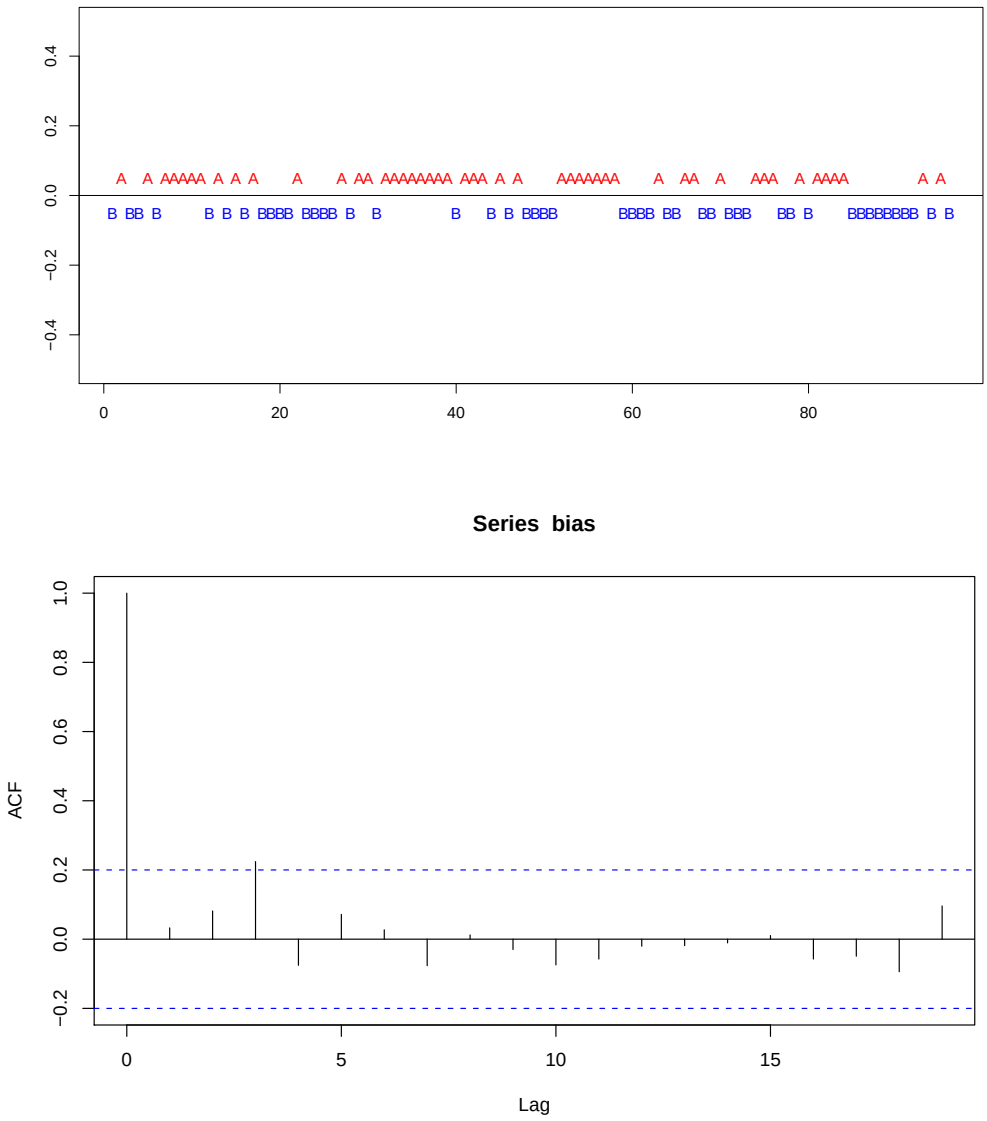

Figure 8: The plots of the runs and sample autocorrelation functions of the surface temperature error forecasts.

may suspect that our data are still linearly dependent, i.e., observations at three lags apart are correlated. Let us run the Box-Pierce (BP) test for the third lag. The Box-Pierce test belongs to the class of non-distribution-free portmanteau tests and needs an assumption of normality for observations (Cromwell *et al.* 1994). On the other hand, the advantage of the BP test is that it allows to test for higher order serial correlations. The BP test does not reject the null hypothesis of residuals being uncorrelated at the third lag. Hence, we can conclude that the errors of surface temperature forecasts are in fact uncorrelated.

```
R> Box.test(bias, lag = 3)
```

```
        Box-Pierce test

data:  bias
X-squared = 5.5634, df = 3, p-value = 0.1349
```

6. Conclusion

This paper discusses implementation in R of some newly developed and already well known non-parametric and goodness-of-fit tests and procedures that are utilized in litigations, environmental studies and biostatistics. Besides of the methods discussed in the paper, **lawstat** contains such procedures as the Brunner-Munzel test, which is also known as the generalized Wilcoxon test (Brunner and Munzel, 2000); the Cochran-Mantel-Haenszel test (Agresti 2002; Gastwirth 1984; Hall, Woolson, Clarke, and Jones 1999); the Lorenz curve, the coefficient of dispersion and the Gini index (Gastwirth 1988; Bonett and Seier 2008), as well as a number of illustrative data sets from law cases, environmental and archeological studies. We plan to constantly update **lawstat** with new statistical methods and tests.

Acknowledgments

The authors would like to thank the research group of Professor Clifford Mass in the Department of Atmospheric Sciences at the University of Washington for proving weather data, Saad Zaman for providing Merck's monthly trade volume data and Kimihiro Noguchi for detecting bugs in the code. The term "very-low-probability-very-high-consequence events" was suggested by an anonymous referee. The research of Professor Gel was supported by a grant from the National Science and Engineering Research Council (NSERC) of Canada, and the research of Professor Gastwirth was supported in part by grant SES-0317956 from the National Science Foundation. Wallace Hui was supported by the NSERC Undergraduate Research Award, 2006. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET, http://www.sharcnet.ca/).

References

Agresti A (2002). *Categorical Data Analysis*. 2nd edition. John Wiley & Sons, New York.

Annaert J, Brys G, Ceuster MD (2005). "A New Large Sample Test of Univariate Symmetry: A Comparative Size-Power Study." *Advances and Applications in Statistics*, **5**, 353–370.

Atkinson AC (1982). "The Simulation of Generalized Inverse Gaussian and Hyperbolic Random Variables." *SIAM Journal on Scientific and Statistical Computing*, **3**, 502–515.

Barndorff-Nielsen O (1997). "Processes of Normal Inverse Gaussian Type." *Finance and Stochastics*, **2**, 41–68.

Barndorff-Nielsen O, Blaesild P (1983). *Hyperbolic Distributions*. John Wiley & Sons, New York.

Bartels R (1977). "Estimation of a First Order Autoregressive Scheme with Non-Normal Stable Disturbances." *Journal of Statistical Computation and Simulation*, **6**, 35–48.

Bartels R (1982). "The Rank Version of Von Neumann's Ratio Test for Randomness." *Journal of the American Statistical Association*, **77**, 40–46.

Bonett DG, Seier E (2002). "A Test of Normality with High Uniform Power." *Computational Statistics & Data Analysis*, **40**, 435–445.

Bonett DG, Seier E (2008). "Confidence Interval for a Coefficient of Dispersion in Nonnormal Distributions." *Biometrical Journal*, **48**, 144–148.

Bowman K, Shenton L (1975). "Omnibus Test Contours for Departures from Normality Based on b_1 and b_2." *Bimetrika*, **62**, 243–250.

Brown MB, Forsythe AB (1974). "Robust Tests for Equality of Variances." *Journal of the American Statistical Association*, **60**, 364–367.

Cabilio P, Masaro J (1996). "A Simple Test of Symmetry About an Unknown Median." *The Canadian Journal of Statistics*, **24**, 349–361.

Carroll RJ, Schneider H (1985). "A Note on Levene's Tests for Equality of Variances." *Statistics and Probability Letters*, **3**, 191–194.

Cromwell JB, Walter CL, Terraza M (1994). *Univariate Tests for Time Series Models*. Sage Publications Inc.

D'Agostino R, Stephens M (1986). *Goodness-of-fit Techniques*. Marcel Dekker, New York.

Einmahl JHJ, McKeague I (2003). "Empirical Likelihood Based Hypothesis Testing." *Bernoulli*, **9**, 267–290.

El-Shaarawi A, Damsleth E (1988). "Parametric and Nonparametric Tests for Dependent Data." *Water Resources Bulletin*, **24**, 513–519.

Faigman DL, Kaye DH, Saks MJ, Saunders J (2002). *Science in The Law: Standards, Statistics, and Research Issues*. West Group.

Finkelstein MO, Levin BA (1990). *Statistics for Lawyers*. Springer-Verlag.

Fox J (2002). *An R and S-PLUS Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA.

Gastwirth JL (1982). "Statistical Properties of a Measure of Tax Assessment Uniformity." *Journal of Statistical Planning and Inference*, **6**, 1–12.

Gastwirth JL (1984). "Statistical Methods for Analyzing Claims of Employment Discrimination." *Industrial and Labor Relations Review*, **1**, 75–86.

Gastwirth JL (1988). *Statistical Reasoning in Law and Public Policy*. Academic Press, Boston, Toronto.

Gastwirth JL (2000). *Statistical Science in the Courtroom*. Springer-Verlag, New York.

Gastwirth JL (2006). "A Sixty Million Dollar Statistical Issue Arising in the Interpretation and Calculation of a Measure of Relative Disparity Mandated by Law: Zuni Public School District 89 v. US Department of Education." *Law, Probability and Risk*, **5**, 33–61.

Gastwirth JL, Gel YR, Miao W (2008). "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice." Working paper, Department of Statistics, George Washington University.

Gastwirth JL, Rubin H (1975). "The Behavior of Robust Estimators on Dependent Data." *The Annals of Statistics*, **3**, 1070–1100.

Gel YR, Gastwirth JL (2008). "A Robust Modification of the Jarque-Bera Test of Normality." *Economics Letters*, **99**, 30–32.

Gel YR, Miao W, Gastwirth JL (2005). "The Importance of Checking the Assumptions Underlying Statistical Analysis: Graphical Methods for Assessing Normality." *Jurimetrics*, **46**, 3–29.

Gel YR, Miao W, Gastwirth JL (2007). "Robust Directed Tests of Normality Against Heavy Tailed Alternatives." *Computational Statistic & Data Analysis*, **51**, 2734–2746.

Hall DB, Woolson RF, Clarke WR, Jones MF (1999). *Cochran-Mantel-Haenszel Techniques: Applications Involving Epidemiologic Survey Data*. John Wiley & Sons, New York.

Hall P, Welsh AH (1985). "Limit Theorems for the Median Deviation." *Annals of the Institute of Statistical Mathematics*, **37**, 27–36.

James O (2007). "Breaking Free of Chevron's Constraints: Zuni Public School District et al. v. US Department of Justice." University of Wisconsin, Legal Studies Research Paper No. 1042.

Jarque C, Bera A (1980). "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals." *Economics Letters*, **6**, 255–259.

Keller-McNulty S, McNulty M (1987). "The Independent Pairs Assumption in Hypothesis Tests Based on Rank Correlation Coefficients." *The American Statistician*, **41**, 40–41.

Komsta L, Novomestky F (2007). ***moments***: *Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.11, URL http://CRAN.R-project.org/package=moments.

Levene H (1960). *Robust Tests for Equality of Variances*. Stanford University Press, Palo Alto.

Miao W, Gel YR, Gastwirth JL (2006). "A New Test of Symmetry About an Unknown Median." In A Hsiung, CH Zhang, Z Ying (eds.), "Random Walk, Sequential Analysis and Related Topics – A Festschrift in Honor of Yuan-Shih Chow," World Scientific.

Miller RG (1968). "Jacknifing Variances." *The Annals of Mathematical Statistics*, **39**, 567–582.

Mira A (1999). "Distribution-Free Test for Symmetry Based on Bonferroni's Measure." *Journal of Applied Statistics*, **26**, 959–972.

Moore DS (1982). "The Effect of Dependence on Chi Squared Tests of Fit." *The Annals of Statistics*, **10**, 1163–1171.

Neuhauser M, Hothorn LA (2000). "Parametric Location-Scale and Scale Trend Tests Based on Levene's Transformation." *Computational Statistics & Data Analysis*, **33**, 189–200.

Poitras G (2005). "More on the Correct Use of Omnibus Tests of Normality." *Economics Letters*, **90**, 304–309.

R Development Core Team (2007). *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 RJ (1968). "Contributions to Central Limit Theory for Dependent Variables." *The Annals of Mathematical Statistics*, **39**, 1158–1175.

Shapiro S, Wilk M (1965). "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, **52**, 591–611.

SPSS Inc (2006). *SPSS for Windows, Release 15*. SPSS Inc., Chicago, IL. URL http://www.spss.com/.

Thadewald T, Buning H (2006). "Jarque-Bera Test and its Competitors for Testing Normality – A Power Comparison." *Journal of Applied Statistics*, **34**, 87–105.

Trapletti A, Hornik K (2007). ***tseries**: Time Series Analysis and Computational Finance*. R package version 0.10-11., URL http://CRAN.R-project.org/package=tseries.

Wald A, Wolfovitz J (1943). "The Exact Test for Ramdomess in the Nonparametric Case Based on Serial Correlation." *The Annals of Mathematical Statistics*, **14**, 378–388.

Wolff SS, Thomas JB, Gastwirth JL (1967). "The Effect of Autoregressive Dependence on a Nonparametric Test." *Professional Group on Information Theory*, **13**, 311–313.

Wuertz D (2007). ***fBasics**: **Rmetrics** – Markets and Basic Statistics*. R package version 260.72, URL http://CRAN.R-project.org/package=fBasics.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. URL http://CRAN.R-project.org/doc/Rnews/.

Zeisel H, Kaye D (1997). *Prove It with Figures: Empirical Methods in Law and Litigation*. Springer-Verlag.

Zuni (2002). "Zuni Public School District et al. v. State of New Mexico et al. Case No. CV98-14-22." Eleventh Judical District Court. Report of the Special Master, January 14, 2002.

Affiliation:

Wallace Hui, Yulia R. Gel
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
Fax: +1/519/746-1875
E-mail: wlwhui@uwaterloo.ca, ygl@math.uwaterloo.ca

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 28, Issue 3 *Submitted:* 2007-11-12
October 2008 *Accepted:* 2008-01-11