Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2011-54

Business indicators of health care quality: outlier detection in small samples

Gaj Vidmar[a,b], Rok Blagus[b], Luboš Střelec[c] and Milan Stehlík

May 2011

[a]University Rehabilitation Institute, Republic of Slovenia
[b]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana
[c]Department of Statistics and Operational Analysis, Mendel University, Brno

SUMMARY

Health care quality monitoring by the Ministry of Health in Slovenia includes over 100 business indicators of economy, efficiency and funding allocation, analysed annually for over 20 hospitals. Most of these indicators are random-denominator same-quantity ratios with a highly correlated numerator and denominator, and the goal is identification of outliers. A large simulation study was performed to study the performance of three types of methods: common outlier detection tests for small samples – Grubbs, Dean & Dixon, and Nalimov test – applied unconditionally and conditionally upon results of Shapiro-Wilk normality test; the boxplot rule; and the double-square-root control chart, for which we introduced regression-through-origin-based control limits. Pert, Burr and 3-parameter-loglogistic distribution, which fitted the real data best, were used with no, one or two outliers in the simulated samples of size 5 to 30. Small (below 0.2; right-skewed) and large (above 0.5, more symmetrical) ratios were simulated. Performance of the methods varied greatly across the conditions. Formal small-sample tests proved virtually useless if applied conditionally upon passed normality pre-test in the presence of outliers. Boxplot rule performed most variedly, but was the only useful one for tiny samples. Our variant of the double-square-root control chart proved too conservative in tiny samples and too liberal for samples of size 20 or more without outliers, but appeared the most useful to detect actual outliers in samples of the latter size. As a possibility for future improvement and research, we propose pre-testing of normality using a class of robustified Jarque-Bera tests.

KEY WORDS: quality indicators; ratios; small samples; outlier detection; simulation

1. INTRODUCTION

Health care quality monitoring by the Ministry of Health in Slovenia includes numerous business indicators of economy, efficiency and funding allocation, which are annually analysed for all the hospitals in the country [1]. Examples of these indicators are the area of a hospital used for a certain service (e.g., dialysis or computed tomography) per total area of the hospital, and the expenses for a certain purpose (e.g., energy consumption or staff education) per total expenses of the hospital.

The essence of associated statistical analyses is identification of outliers, where a compromise between state-of-the-art and wide understandability is desired. Hence, an exploratory approach was adopted that combines three types of methods:

- three common outlier detection tests useful in small samples, namely the Grubbs test [2], the Dean & Dixon test [3] and the Nalimov test [4], which are all based upon assumption of normality and were hence tried unconditionally as well as conditionally upon results of normality tests;
- the Tukey [5] boxplot rule, i.e., identifying as outlier any value more than 1.5-times the inter-quartile range larger or smaller then the third or the first quartile, respectively;
- control charts.

The present study investigates these three approaches through extensive simulations. The paper first introduces a novel proposal regarding the application of appropriate control charts. The simulations setup and the simulation results are then presented, followed by empirical results on robust normality pretesting and outlier detection. Finally a summary is given with a discussion of related work and directions for further research.

2. CONTROL CHARTS

2.1. General considerations

The indicators addressed in this study are same-quantity ratios (thus bound between 0 and 1) which are appropriately treated neither as proportions nor as fixed-denominator ratios. They are random-denominator ratios with highly correlated numerator and denominator. Examples of such indicators are presented in Figure 1. For the two types of indicators, i.e., small and large ratios (as detailed in the section 3), different scales are used in the histograms. The distributions of small ratios are right-skewed, while the distributions of large ratios are roughly symmetric. It is evident that there is very high correlation between the numerator and the denominator, while the ratios tend to be independent of their denominators. To avoid unauthorised disclosure and hospital identification without loosing the information relevant for our study, the actual quantities defining the numerator and the denominator are masked.

It is essential to note that funnel plots, which have rightfully been promoted for monitoring cross-sectional performance indicators in health care [6-9] and also in education [10], may not be the appropriate choice for such data. The reason is that virtually all points would get labelled as outliers in such plots because of huge denominators (thousands of square metres, millions of Euro) yielding excessively narrow confidence intervals for the average proportion (even at 99% confidence level). Even if the whole problem is considered as one of over-dispersion [11, 12], which has been recognised in health-care setting and for which different strategies have been suggested, and given that abandoning indicators is not an

option, neither random-effects models [11] nor the Laney's approach [12] are universally feasible. Therefore, a different choice of the control chart and its control limits is tenable.

2.2. Proposed modification of the Double Square Root Chart

We opted for the double-square-root (Shewart) chart [13], in which the square-root of the numerator is plotted against the square-root of the difference between the denominator and the numerator. Like the funnel plot, this is also an increasingly popular method in statistical health-care quality control [14, 15]. However, it was essential to replace the traditional control limits – based on the underlying assumption of binomial distribution, like in funnel plots, and therefore much too narrow for our data – by newly defined ones. The newly defined control limits were obtained by using linear regression through the origin (the rationale being that, e.g., no costs can be incurred without income, no space can be used for a given purpose without any space, etc.) and estimating control limits using 95% confidence interval for prediction.

Four examples of such charts are presented in Figure 2. They show indicators with no outliers (upper left), one outlier above the control limits (upper right), one outlier below the control limits (lower left), and two outliers (one above and one below the control limits; lower right).

3. SIMULATION SETUP

We studied performance and agreement of the chosen methods through a large simulation study on realistic data. In accordance with the real data (Figure 1), two types of ratios were generated:

- the small ones belonging to the [0, 0.2] interval;
- the large ones belonging to the [0.5, 1] interval.

Samples of size 5, 10, 20, 25 and 30 were drawn from the three distributions that were found to best fit the empirical data. After automated fitting using EasyFit Professional 5.1 software (MathWave Technologies, 2009), the following distributions were chosen: Pert, 3-parameter Burr (2) (referred to henceforth simply as Burr) and 3-parameter loglogistic (3). To further improve resemblance to real data, the modified (4-parameter) Pert distribution (1) with the additional shape parameter (γ) was used [16] (referred to henceforth as 3Ploglog). While Pert is a bounded distribution, Burr and 3Ploglog are only non-negative, but they are nonetheless useful practical models for same-quantity ratios in rejection-based simulations because their parameters can be chosen so that large values are extremely rare (i.e., their right tail can be made extremely thin).

$$f_{\mathrm{Pert}}(x) = \frac{(x-\min)^{\alpha_1-1}(\max-x)^{\alpha_2-1}}{Beta(\alpha_1,\alpha_2)(\max-\min)^{\alpha_1+\alpha_2-1}}$$
$$\alpha_1 = 1+\gamma\left(\frac{\mathrm{mode}-\min}{\max-\min}\right),\ \alpha_2 = 1+\gamma\left(\frac{\max-\mathrm{mode}}{\max-\min}\right) \quad (1)$$

$$f_{\text{Burr}}(x) = \frac{ak\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)\right)^{k+1}} \tag{2}$$

$$f_{\text{3Ploglog}}(x) = \frac{\left(1+\frac{\xi(x-\mu)}{\sigma}\right)^{-(1/\xi+1)}}{\sigma\left[1+\left(1+\frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi}\right]^2} \tag{3}$$

Zero, one or two simulated outliers were included in the samples. The outliers were generated by increasing the relevant parameter (mode, scale parameter and mean for Pert, Burr and 3Ploglog, respectively) by 50 %, 100 %, 150 % and 500 % while holding other parameters (related to dispersion and shape) fixed. The simulation was performed with R [17] using rejection sampling. The *outliers*, *mc2d*, *lmom* and *actuar* R packages were used.

First, data for the ratios were generated (drawn from the given distribution until all data were between 0 in 1) for the base sample and then for the outlier(s). The following parameters were used for the base population, whereby all the drawn values were divided by 1000:

- Small ratios
  - Pert: min = 0, max = 30, mode = 10, $\gamma = 5$
  - Burr: $\alpha = 2$, $k = 3$, $\beta = 30$
  - 3Ploglog: $\xi = 0$, $\mu = \log(10)$, $\sigma = 1.28$
- Large ratios
  - Pert: min = 400, max = 1000, mode = 700, $\gamma = 5$
  - Burr: $\alpha = 3$, $k = 10$, $\beta = 750$
  - 3Ploglog: $\xi = 0$, $\mu = \log(700)$, $\sigma = 1.28$

The outlier was always drawn from the same distribution as the base sample, except that the central tendency parameter of the distribution from which the outlier was drawn was larger. As already mentioned, its value was 150 %, 200 %, 250 % and 600 % of the value for the base population, i.e., larger by a factor of 0.5, 1, 1.5 and 5, respectively. For the outlier population, the following parameters were increased:

- mode for Pert,
- β for Burr and
- μ for 3Ploglog.

Because of complexity and time constraints, only three situations were simulated:

- no outliers;
- one outlier at the right-hand side of the sample distribution;
- two outliers at the right-hand side of the sample distribution.

Since correlation between the numerator and the denominator is required for the double-square-root control chart, once the ratios had been obtained, the numerators were drawn from the uniform distribution with the lower bound set to 0 and the upper bound adjusted so that

the desired correlation was obtained. The desired correlation range was set between 0.2 and 0.6, which is a relatively wide span, in order to avoid convergence problems.

Under each condition, 1000 samples were generated (or slightly fewer in case the 1000 samples were not obtained within the 350 hours of run-time, after which the simulation was stopped).

4. SIMULATION RESULTS

The results of the simulations are summarised in Table I (no-outlier condition), Table II (one simulated outlier) and Table III (two-outlier condition).

When there were no outliers, the methods performed very well – with the estimated accuracy above 90%, except the boxplot rule and especially the Nalimov test, which achieved a 73% accuracy with $n = 5$ and dropped to merely 16% with $n = 30$. Naturally, the results of the formal outlier detection tests were better when we only considered those simulations in which normality was not rejected (i.e., in the conditional case), because otherwise "outliers" were occasionally found simply because of the skewness of the distributions from which the samples were drawn. Overall, the modified double-square-root control chart performed best under this condition.

Under the one-outlier condition, the formal outlier detection tests performed worse in the conditional case than in the unconditional cases. This highlights the problem of normality testing with outliers, which is addressed in the next section. However, accuracy was also low in the unconditional cases and did not depend markedly on the sample size. The boxplot rule proved the most accurate for small ratios, while the modified double-square-root control chart gave the best results for large ratios. It is also noteworthy that the performance of the boxplot rule worsened as sample size increased, while the performance of the control chart improved.

When two outliers were simulated, all methods performed rather poorly. Similarly to the one-outlier situation, the formal tests assuming normality performed much better when all the simulated samples were used (i.e., in the unconditional case). With small ratios, the boxplot rule achieved accuracy comparable to the formal tests applied unconditionally, while the modified double-square-root control chart was the least accurate. However, with large ratios, the control chart was the most accurate, while all other methods proved highly inaccurate.

To summarise, it is apparent that the performance of the methods varied greatly across the conditions. Formal small-sample tests became virtually useless if applied conditionally upon passed normality pre-test in the presence of (especially two) outliers with a sample size of 10. Among the formal tests, the Dean & Dixon test performed worst overall. The simple boxplot method performed the most variedly, but it was the only useful one for tiny samples. Our variant of the double-square-root control chart proved too conservative in tiny samples and too liberal under the no-outlier condition with $n \geq 20$ (both conclusions holding also for the Nalimov test, and for boxplot for small ratios), but it appeared by far the most useful (though still far from perfect) to detect actual outliers with a larger n, especially with large ratios.

Regarding the chosen sample sizes, it should be noted that with a sample size above 30, the simulation procedure failed to converge. However, that did not pose a serious limitation to our study since we focused on small samples, because moderate or large samples are rarely encountered in statistical comparisons of such quality indicators between hospitals or similar organisations. Samples are bound to be of small or moderate size because particularly with

financial indicators, the comparisons can be meaningful only if truly comparable organisations are compared within a specific sector (e.g., hospital type) and/or a very homogenous area (e.g., Slovenia, which is relatively uniformly urbanised).

5. POSSIBILITIES FOR ROBUST NORMALITY PRETESTING

As the starting point for this section, we take the normality test that is commonly attributed to Jarque and Bera [18]. Actually [19], Bowman and Shenton [20] were the first to observe that under normality, the asymptotic means of sample skewness and kurtosis statistics are 0 and 3, respectively, the asymptotic variances of the two statistics are 6/*n* and 24/*n*, respectively, and their asymptotic covariance is 0. Another version of the skewness-kurtosis test for normality was suggested by D'Agostino and Pearson [21].

A class of robust normality tests for small samples possibly containing outliers against Pareto tails has recently been proposed [22]. This class also contains tests that accommodate the kind of alternative distributions that are known to be problematic for the Jarque-Bera test (e.g., bimodal, Weibull and uniform). The proposal can be seen as an extension of the robust modification of the Jarque Bera test [23]. The base for the proposed class of tests is a location functional, denoted by $T(F)$ [24], whereby the relevant location functionals ($T_{(i)}$ for $i = 0..3$) are arithmetic mean ($T_{(0)}$), median ($T_{(1)}$), trimmed mean ($T_{(2)}(s)$) and pseudo-median ($T_{(3)}$). Relaxing the form of j-th theoretical moment estimator $\mu_j = E(X - E(X))^j$ by using $M_j(T(F_n), r) = \frac{1}{n-2r}\sum_{m=1+r}^{n-r}(X_{m:n} - T(F_n))^j$, where $X_{1:n} < X_{2:n} < \ldots < X_{n:n}$ are the order statistics, the new class of test statistics (denoted by RT_{JB}) can be defined as

$$RT_{JB} = \frac{k_1(n)}{C_1}\left(\frac{M_3(r_1, T_{(i_1)}(s_1))}{M_2^{3/2}(r_2, T_{(i_2)}(s_2))} - K_1\right)^2 + \frac{k_2(n)}{C_2}\left(\frac{M_4(r_3, T_{(i_3)}(s_3))}{M_2^{2}(r_4, T_{(i_4)}(s_4))} - K_2\right)^2. \quad (4)$$

While $k_1(n)$ and $k_2(n)$ are theoretical values of proportions for first and second term of the statistics dependent on sample size, the C_1 and C_2 constants can be obtained from Monte Carlo simulations, whereby their values for small samples under trimming ($r > 0$) differ from those without trimming ($r = 0$). The K_1 and K_2 constants are small-sample variants of mean corrections so that asymptotical normality is obtained and thus the χ^2 asymptotical distribution of the test statistics is valid.

Illustrative special cases of this class include the "median robustified *JB* test" (5), the "trimmed-mean robustified *JB* test" with trimming parameter $s = 5$ (6), the "pseudo-median robustified *JB* test" (7) and the "trim-trim robustified *JB* test" with trimming parameters $s = r = 1$ (8):

$$RT_{JB15} = \frac{1}{18}\left(\frac{M_3(0, T_{(1)}(0))}{M_2^{3/2}(0, T_{(1)}(0))} - K_1\right)^2 + \frac{1}{24}\left(\frac{M_4(0, T_{(1)}(0))}{M_2^{2}(0, T_{(1)}(0))} - K_2\right)^2, \quad (5)$$

$$RT_{JB45} = \frac{1}{6}\left(\frac{M_3(0, T_{(2)}(5))}{M_2^{3/2}(0, T_{(2)}(5))} - K_1\right)^2 + \frac{1}{24}\left(\frac{M_4(0, T_{(2)}(5))}{M_2^{2}(0, T_{(2)}(5))} - K_2\right)^2, \quad (6)$$

$$RT_{PMJB} = \frac{1}{18}\left(\frac{M_3(0, T_{(3)}(0))}{M_2^{3/2}(0, T_{(3)}(0))} - K_1\right)^2 + \frac{1}{24}\left(\frac{M_4(0, T_{(3)}(0))}{M_2^{2}(0, T_{(3)}(0))} - K_2\right)^2, \quad (7)$$

$$RT_{TTJB(s=r=1)} = \frac{1}{6}\left(\frac{M_3(1,T_{(2)}(1))}{M_2^{3/2}(1,T_{(2)}(1))} - K_1\right)^2 + \frac{1}{24}\left(\frac{M_4(1,T_{(2)}(1))}{M_2^{2}(1,T_{(2)}(1))} - K_2\right)^2. \quad (8)$$

Some theoretical results on consistency and asymptotical χ^2 distribution of these and other RT_{JB} class tests can be found in [22], where they were introduced. Here, we just briefly summarise some preliminary results of power and size comparisons through simulations with various distributions and sample sizes:

- In samples of size 25, the Jarque-Bera test and its successors [19, 23] had nearly zero power against the Beta(0.5, 0.5) alternative, as did the robust directed test of normality against heavy-tailed alternatives [25] even for sample sizes up to 200. Shapiro-Wilk and Anderson-Darling were the most powerful tests, while some RT_{JB} class tests were almost as powerful.
- Against the Burr (2, 1, 1) alternative in small samples, the power of *JB* test was comparable with the Anderson-Darling test, while the Shapiro-Wilk test and some tests from the RT_{JB} class were more powerful. With a sample size of 100, the power of all tests reached 1. Against logistic alternative, all normality tests had very low power in small samples (because of the resemblance of the logistic to the normal distribution).
- The most powerful test for normality against a mixture of two equally probable normal distributions with means 0 and 5 and unit variance were the D'Agostino, the Anderson-Darling and the Shapiro-Wilk test, while the power of RT_{JB} class tests and of the (original and robust) Jarque-Bera test was very low for $n = 25$, though unlike for the latter, it improved quickly for the RT_{JB} class tests with $n = 50$.
- In simulated samples from the standard normal distribution containing one extreme outlier (from the normal distribution with a mean of 3 and a unit variance), which can be viewed as assessing a particularly defined size of the tests (i.e., defining the proper decision as retaining the null hypothesis of normality despite the outlier), the simpler RT_{JB} class tests did slightly better than the D'Agostino and the Jarque-Bera test, though they were still very much on the liberal side. The Shapiro-Wilk test performed a little less liberally, while the estimated "size" was even closer to nominal for the Anderson-Darling test. The robust medcouple test [26] retained the proper size irrespective of sample size, but it proved the least powerful in the power comparisons mentioned above. Encouragingly, RT_{JB} class tests with trimming applied to both the location and the generalised moment had almost as good power as the simpler RT_{JB} class tests while their estimated "size" was very close to the nominal 5%.

A lot of work remains to be done regarding the RT_{JB} class test, including simulations with other alternatives and under different assumptions. It is inevitably challenging to construct powerful tests that are robust at the same time. There is always a trade-off between power and robustness, which is where the RT_{JB} class tests of normality might offer a useful compromise.

6. SUMMARY AND FUTURE DIRECTIONS

A large simulation study of outlier detection in small samples of random-denominator same-quantity ratios with a highly correlated numerator and denominator was performed. The performance of three types of methods was assessed: the common formal outlier-detection

tests (Grubbs, Dean & Dixon, and Nalimov test) applied unconditionally and conditionally upon results of (Shapiro-Wilk) normality test; the boxplot rule; and the double-square-root control chart (for which we introduced regression-through-origin-based control limits). Pert, Burr and 3-parameter-loglogistic distribution (which fitted the real data best) were used with zero, one or two outliers in the simulated samples of size 5 to 30. Small (below 0.2; right-skewed) and large (above 0.5, more symmetrical) ratios were simulated. The performance of the methods varied greatly across the conditions. Formal small-sample tests became useless if applied conditionally upon passed normality pre-test in the presence of (especially two) outliers with a sample size of 10. Boxplot rule performed most variedly, but it was the only useful one for tiny samples. Our variant of the double-square-root control chart proved too conservative in tiny samples and too liberal for samples of size 20 or more without outliers, but it appeared the most useful to detect actual outliers in samples of the latter size (especially with large ratios).

Following further research on robust normality testing, it might be useful to repeat the first part of our outlier-detection simulations (i.e., the conditionally applied formal outlier tests) with different normality pre-tests. Improved normality pre-testing should improve the feasibility and usability of outlier tests in small samples since the naïve approach of abandoning normality pre-testing resulted in (too) many false alarms in the no-outlier simulations.

Putting our work in a broader context, we should first recognise that the extensive simulation approach to the problems of outliers and robustness owes its main origin to the work of Andrews, Hampel, Huber, Tukey and associates in the context of the Princeton Robustness study [27]. Statistical process control literature has also dealt with outliers, though primarily within the context of time-dependent process data [28]. It should also be noted that the entire outlier detection approach which we assessed through simulations should be viewed as a heuristics rather than as statistical testing or strictly probabilistic decision-making. Constructing a general statistical test for outlier detection is namely an unsolvable problem without many specific substantial assumptions answering the question "What is an outlier?". This problem has some similarity with the problem of choosing the correct number of clusters posed nearly four decades ago [29]. Although the problem of clustering is challenging enough, even in one dimension [30], a clustering approach – which has been mentioned as an option for dealing with over-dispersion [11] – might be worth trying. Another alternative to analysing the kind of data that we addressed, which is better explored and established, are bootstrap tolerance intervals [31].

In conclusion, it may not be surprising that seemingly simplistic methods, exemplified by boxplots and control charts, which combine a robust "eye-balling" approach with "a touch" of implicit inference and vast experience from practical data analysis [5, 32], have yet again proven their value in statistics applied to a real-life industrial and organisational setting.

REFERENCES

1. Robida A. *Uvajanje izboljševanja kakovosti v bolnišnice*. Ministrstvo za zdravje: Ljubljana, 2006.
2. Grubbs FE. Procedures for detecting outlying observations in samples. *Technometrics* 1969; **11**(1):1–21. JSTOR: 1266761
3. Dean RB, Dixon WJ. Simplified statistics for small numbers of observations. Analytical Chemistry 1951; **23**(4):636–638. DOI: 10.1021/ac60052a025
4. Kaiser R, Gottschalk G. *Elementare Tests zur Beurteilung von Messdaten*. Bibliographisches Institut: Mannheim, 1972.
5. Tukey JW. *Exploratory Data Analysis*. Addison-Wesley: Reading, MA, 1977.

6. Harrison WN, Mohammed MA, Wall MK, Marshall TP. Analysis of inadequate cervical smears using Shewhart control charts. *BMC Public Health* 2004; **4**:25. DOI: 10.1186/1471-2458-4-25
7. Guthrie B, Love T, Fahey T, Morris A, Sullivan F. Control, compare and communicate: designing control charts to summarise efficiently data from multiple quality indicators. *Quality and Safety in Health Care* 2005; **14**:450–454. DOI: 10.1136/qshc.2005.014456
8. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; **24**(8):1185–1202. DOI: 10.1002/sim.1970
9. Tu J-K, Gilthorpe MS. The most dangerous hospital or the most dangerous equation. *BMC Health Services Research* 2007; **7**:185. DOI: 10.1186/1472-6963-7-185
10. Farewell VT. Comment: Advancing public sector performance analysis by Professor C. J. Heinrich. *Applied Stochastic Models in Business and Industry* 2008; **24**(5):391–395. DOI: 10.1002/asmb.726
11. Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Quality and Safety in Health Care* 2005; **14**(5):347–351. DOI: 10.1136/qshc.2005.013755
12. Mohammed MA, Laney D. Overdispersion in health care performance data: Laney's approach. *Quality and Safety in Health Care* 2006; **15**(5):383–384. DOI: 10.1136/qshc.2006.017830
13. Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001; **357**:463–467. DOI: 10.1016/S0140-6736(00)04019-8
14. Battersby J, Flowers J, Harvey I. An alternative approach to quantifying and addressing inequity in healthcare provision: access to surgery for lung cancer in the east of England. *Journal of Epidemiology & Community Health* 2004; **58**:623–625. DOI: 10.1136/jech.2003.013391
15. Marshall T, Mohammed MA, Rouse A. A randomized controlled trial of league tables and control charts as aids to health service decision-making. *International Journal for Quality in Health Care* 2004; **16**(4): 309–315. DOI: 10.1093/intqhc/mzh054
16. Vose D. *Risk Analysis: a Quantitative Guide*. Wiley: Chichester, 2008.
17. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2010. http://www.R-project.org [10 November 2010]
18. Jarque CM, Bera AK. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 1980; **6**(3):255–259. DOI: 10.1016/0165-1765(80)90024-5
19. Urzua CM. On the correct use of omnibus tests for normality. *Economics Letters* 1996; **53**(3):247–251. DOI: 10.1016/S0165-1765(96)00923-8
20. Bowman KO, Shenton LR. Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. Biometrika 1975; **62**(2):243–250. DOI: 10.1093/biomet/62.2.243
21. D'Agostino R, Pearson E. Tests for departures from normality. Empirical results for the distribution of $\sqrt{b_1}$ and b_2. *Biometrika* 1973; **60**:613–622. DOI: 10.1093/biomet/60.3.613
22. Stehlík M, Fabián Z, Střelec L. *Small sample robust testing for normality against Pareto tails* (IFAS Research Paper Series 2010-51). Johannes Kepler University: Linz, 2010. http://www.jku.at/ifas/content/e108280/e108491/e108471/e108473/IFasResRep2010-51.pdf [22 April 2011]
23. Gel YR, Gastwirth JL. A robust modification of the Jarque Bera test of normality. *Economics Letters* 2008; **99**(1):30–32. DOI: 10.1016/j.econlet.2007.05.022
24. Bickel PJ, Lehmann EL. Descriptive statistics for nonparametric models II. Location. *Annals of Statistics* 1975; **3**(5):1045–1069. DOI: 10.1214/aos/1176343240

25. Gel YR, Miao W, Gastwirth JL. Robust directed tests of normality against heavy-tailed alternatives. *Computational Statistics & Data Analysis* 2007; **51**(5):2734–2746. DOI: 10.1016/j.csda.2006.08.022
26. Brys G, Hubert M, Struyf A. Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics* 2008;**23**(3):429–442. DOI: 10.1007/s00180-007-0083-7
27. Andrews DF, Bickel PJ, Hampel FR, Huber PJ, Rogers WH, Tukey JW. *Robust Estimates of Location: Survey and Advances*. Princeton University Press: Princeton, NJ, 1972.
28. Davis CM, Adams BM. Robust Monitoring of contaminated data. *Journal of Quality Technology* 2005; **37**(2):163–174.
29. Ling RF. On the theory and construction of of *k*-clusters. *The Computer Journal* 1972; **15**(4):326–332. DOI: 10.1093/comjnl/15.4.326
30. Anderson NH, Titterington DM. A comparison of two statistics for detecting clustering in one dimension. *Journal of Statistical Computation and Simulation* 1995; **53**(1):103–125. DOI: 10.1080/00949659508811699
31. Kenett RS, Zacks S. *Modern industrial Statistics: Design and Control of Quality and Reliability*. Duxbury Press: San Francisco, CA, 1998.
32. Neave HR, Wheeler DJ. *Shewhart's Charts and the Probability Approach* (Manuscript No. 88) [presented at Ninth Annual Conference of the British Deming Association, May 15, 1996]. SPC Press: Knoxville, TN, 2002. http://www.spcpress.com/pdf/DJW088.pdf [10 November 2010]

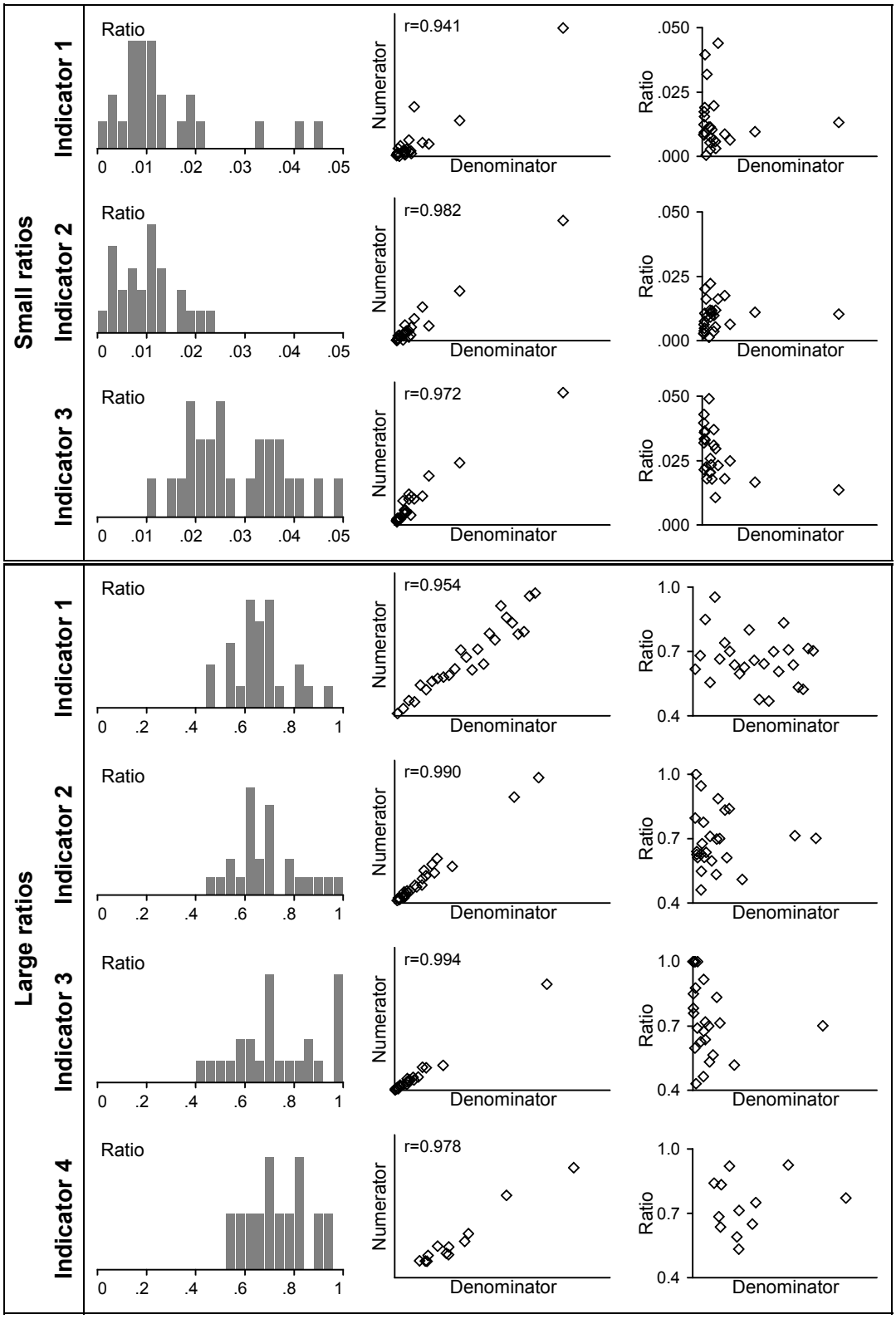

Figure 1. Examples of financial indicators of health care quality. In the left column, the distributions are shown as histograms; in the central column, the numerator is plotted against the denominator and the correlation is listed for each indicator; in the right column, the value of the indicator is plotted against the denominator.

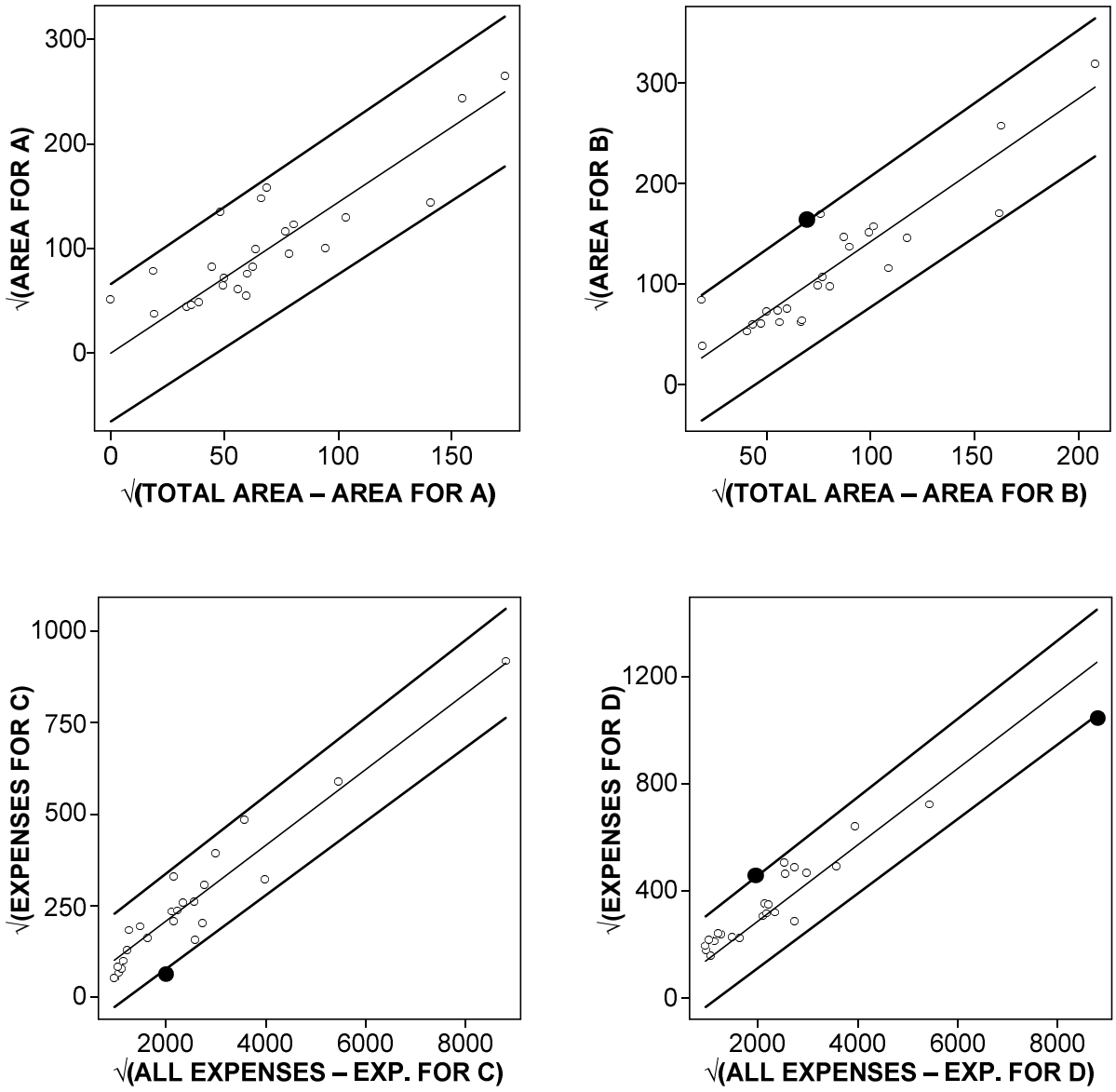

Figure 2. Examples of proposed double-square-root charts. The square-root of the numerator is plotted against the square-root of the difference between the denominator and the numerator, whereby the control limits are obtained by performing linear regression through the origin and estimating the 95% confidence interval for prediction. Outliers are depicted as large filled circles.

Table I. Results of the simulations without outliers. MDSRCC denotes modified double-square-root control chart; Valid denotes the proportion of simulations that passed the Shapiro-Wilk normality test; Mean, Median and Max refer to the number of outliers found; Correct denotes the proportion of simulations in which the test identified no outliers.

Ratio	*n*		5					10					20					25					30				
type	Test		Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct
Small	Grubbs	Conditional		0.09	0	2	91%		0.11	0	3	89%		0.10	0	2	90%		0.08	0	2	92%		0.09	0	3	91%
	Dixon	Conditional	86%	0.02	0	1	98%	69%	0.04	0	1	96%	55%	0.04	0	2	96%	51%	0.04	0	2	96%	49%	0.04	0	2	96%
	Nalimov	Conditional		0.33	0	3	73%		0.70	0	5	52%		1.27	1	6	28%		1.50	1	6	23%		1.69	2	7	16%
	Grubbs	Unconditional		0.23	0	2	80%		0.44	0	5	69%		0.78	0	10	60%		0.96	0	8	57%		1.17	0	10	54%
	Dixon	Unconditional		0.15	0	2	86%		0.22	0	4	82%		0.45	0	6	71%		0.52	0	6	68%		0.58	0	6	66%
	Nalimov	Unconditional		0.46	0	3	63%		1.13	1	6	38%		2.34	2	9	17%		2.91	2	10	13%		3.35	3	11	9%
	Boxplot			0.47	0	2	61%		0.60	0	4	55%		0.97	1	6	47%		1.22	1	6	43%		1.42	1	6	40%
	MDSRCC			0.00	0	0	100%		0.11	0	1	89%		0.71	1	3	35%		0.99	1	3	21%		1.26	1	4	12%
Large	Grubbs	Conditional		0.06	0	2	94%		0.06	0	2	94%		0.07	0	2	93%		0.07	0	3	93%		0.07	0	3	94%
	Dixon	Conditional	87%	0.04	0	2	96%	74%	0.07	0	2	93%	63%	0.04	0	3	96%	63%	0.04	0	3	96%	63%	0.03	0	2	97%
	Nalimov	Conditional		0.23	0	3	81%		0.60	0	7	62%		1.89	1	12	29%		2.38	2	15	21%		3.07	3	21	14%
	Grubbs	Unconditional		0.10	0	2	91%		0.07	0	3	93%		0.07	0	4	94%		0.08	0	3	93%		0.07	0	4	94%
	Dixon	Unconditional		0.09	0	2	92%		0.11	0	2	90%		0.13	0	5	89%		0.13	0	6	90%		0.11	0	5	92%
	Nalimov	Unconditional		0.27	0	3	79%		0.63	0	7	65%		1.62	1	12	47%		1.97	1	15	42%		2.43	1	21	39%
	Boxplot			0.35	0	2	73%		0.29	0	10	80%		0.32	0	8	81%		0.31	0	9	82%		0.34	0	9	80%
	MDSRCC			0.00	0	0	100%		0.10	0	1	90%		0.71	1	3	37%		0.96	1	4	24%		1.30	1	4	12%

Table II. Results of the simulations with one outlier. MDSRCC denotes modified double-square-root control chart; Valid denotes the proportion of simulations that passed the Shapiro-Wilk normality test; Mean, Median and Max refer to the number of outliers found; Correct denotes the proportion of simulations in which the test identified one outlier.

Ratio	*n*		5					10					20					25					30				
type	Test		Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct
Small	Grubbs	Conditional		0.17	0	2	15%		0.22	0	3	20%		0.29	0	2	27%		0.27	0	2	26%		0.28	0	2	27%
	Dixon	Conditional	47%	0.05	0	2	4%	21%	0.07	0	2	6%	10%	0.10	0	2	10%	8%	0.09	0	2	9%	7%	0.09	0	1	9%
	Nalimov	Conditional		0.49	0	3	29%		1.06	1	6	37%		1.93	2	7	32%		2.29	2	7	23%		2.47	2	6	21%
	Grubbs	Unconditional		0.66	1	2	51%		1.00	1	6	53%		1.57	1	8	51%		1.78	1	10	49%		2.00	1	12	47%
	Dixon	Unconditional		0.56	1	2	48%		0.67	1	5	52%		1.16	1	7	53%		1.20	1	7	53%		1.26	1	8	52%
	Nalimov	Unconditional		0.89	1	3	52%		1.82	2	7	38%		3.19	3	11	18%		3.75	3	12	12%		4.17	4	14	9%
	Boxplot			0.76	1	2	60%		1.09	1	4	60%		1.62	1	7	49%		1.88	2	9	45%		2.09	2	8	42%
	MDSRCC			0.00	0	0	0%		0.48	0	1	47%		1.00	1	3	70%		1.18	1	3	64%		1.35	1	4	57%
Large	Grubbs	Conditional		0.12	0	2	10%		0.24	0	3	20%		0.32	0	3	26%		0.33	0	3	27%		0.30	0	3	24%
	Dixon	Conditional	78%	0.06	0	2	5%	62%	0.16	0	4	13%	52%	0.18	0	3	15%	52%	0.20	0	3	16%	52%	0.18	0	3	15%
	Nalimov	Conditional		0.40	0	3	24%		1.02	1	7	29%		2.39	2	14	20%		3.01	3	14	16%		3.62	3	22	13%
	Grubbs	Unconditional		0.18	0	2	14%		0.29	0	4	22%		0.34	0	4	25%		0.35	0	5	24%		0.33	0	5	22%
	Dixon	Unconditional		0.14	0	2	12%		0.24	0	4	18%		0.32	0	6	20%		0.31	0	5	20%		0.29	0	7	19%
	Nalimov	Unconditional		0.45	0	3	24%		1.05	1	7	24%		2.09	1	14	13%		2.63	2	17	10%		3.07	2	22	8%
	Boxplot			0.49	0	2	26%		0.57	0	10	25%		0.62	0	20	25%		0.66	0	25	24%		0.63	0	30	22%
	MDSRCC			0.00	0	0	0%		0.48	0	1	45%		0.89	1	3	57%		1.07	1	3	53%		1.28	1	4	45%

Table III. Results of the simulations with two outliers. MDSRCC denotes modified double-square-root control chart; Valid denotes the proportion of simulations that passed the Shapiro-Wilk normality test; Mean, Median and Max refer to the number of outliers found; Correct denotes the proportion of simulations in which the test identified two outliers.

Ratio	*n*		5					10					20					25					30			
type	Test		Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct	Valid	Mean	Median	Max	Correct
Small	Grubbs	Conditional		0.10	0	2	0%		0.15	0	3	1%		0.24	0	2	1%		0.20	0	2	0%		0.24	0	2	0%
	Dixon		67%	0.03	0	2	0%	18%	0.03	0	2	0%	6%	0.05	0	2	0%	4%	0.06	0	2	0%	3%	0.05	0	1	0%
	Nalimov			0.42	0	3	8%		1.00	1	6	13%		2.07	2	7	21%		2.44	2	7	16%		2.63	2	7	18%
	Grubbs	Unconditional		0.41	0	2	10%		1.13	1	6	33%		2.11	2	9	43%		2.46	2	11	43%		2.78	2	14	43%
	Dixon			0.28	0	2	5%		0.57	0	4	17%		1.45	2	8	42%		1.54	2	9	43%		1.62	2	8	43%
	Nalimov			0.71	0	3	19%		2.19	2	7	34%		3.95	4	12	18%		4.48	4	12	13%		4.98	4	14	9%
	Boxplot			0.43	0	2	0%		1.26	1	4	46%		2.21	2	6	49%		2.52	2	9	45%		2.75	2	8	43%
	MDSRCC			0.00	0	0	0%		0.32	0	1	0%		1.14	1	3	18%		1.40	1	4	29%		1.62	2	4	36%
Large	Grubbs	Conditional		0.02	0	2	0%		0.02	0	4	0%		0.04	0	4	0%		0.04	0	2	0%		0.04	0	3	0%
	Dixon		72%	0.15	0	2	0%	58%	0.27	0	4	1%	45%	0.05	0	3	0%	43%	0.04	0	3	0%	45%	0.04	0	3	0%
	Nalimov			0.11	0	3	0%		0.39	0	6	4%		2.21	2	15	14%		3.04	3	15	13%		3.59	3	16	11%
	Grubbs	Unconditional		0.03	0	2	0%		0.04	0	4	0%		0.06	0	4	1%		0.12	0	5	2%		0.14	0	4	2%
	Dixon			0.16	0	2	0%		0.30	0	4	1%		0.21	0	6	2%		0.20	0	7	2%		0.19	0	5	2%
	Nalimov			0.16	0	3	0%		0.58	0	6	4%		2.16	1	15	11%		2.82	2	17	9%		3.26	3	20	8%
	Boxplot			0.11	0	2	0%		0.61	0	10	15%		0.84	0	7	21%		0.89	0	9	20%		0.83	0	8	18%
	MDSRCC			0.00	0	0	0%		0.37	0	1	0%		1.11	1	3	25%		1.37	1	3	34%		1.59	2	4	37%