Statistical Powers of Univariate Normality Tests: Comparative Analysis of 2016 Election Process in Uganda

Nafiu Lukman Abiodun[1,*], Muyombya Solomon Matovu[2], Rasaki Olawale Olanrewaju[3]

[1]*Kabale University, Uganda*
lanafiu@kab.ac.ug
[2]*Kampala International University, Uganda*
smmuyombya@kiu.ac.ug
[3]*Pan African University Institute for Basic Sciences, Technology and Innovation, Kenya*
olanrewaju_rasaq@yahoo.com
*Correspondence: lanafiu@kab.ac.ug

ABSTRACT. The study conducted an empirical comparison of powers of the univariate normality tests using real data for consistency with simulated data. The main objective was to compare the empirical power of the normality tests using natural data for consistency. Six normality tests were selected from the Empirical Distribution Function (EDF), the Correlation and Regression family of normality tests and moment-based normality tests were considered in this study. From the EDF family of normality tests, the Kolmogorov-Smirnov test (Lilliefors correction) and Anderson-Darling normality tests were chosen. From the regression and correlation family of distributions, Shapiro-Wilk and Shapiro-Francia normality tests were chosen. Jaque-Bera and D'Agostino Pearson normality tests were chosen from the moment family of normality tests. The data adopted in this study comprised of the 2016 Uganda election results as cited by Solomon (2016). The analysis was done using a combination of different statistical packages (EXCEL, R and STATA). The data was gathered from the Uganda Electoral Commission website and cleaned using EXCEL package. After the data cleaning process, the data analysis was done using a combination of R (S-plus programing). The analysis involved writing computer programs that tested the data for normality of the different normality tests specified in the study. All the graphical visualizations were done in STATA. The results of the analysis indicate that S-W produced the most powerful results, followed by S-F, D-A, K-S, J-B and lastly A-D. The data thus indicates that the moment-based tests results are better than the EDF-based tests due to huge kurtosis and skewness statistics in the data. The study thus recommends that it is vital to deal with outliers before carrying out any further statistical tests, and the moment-based tests should be applied in the instances where the data is known to have instances of kurtosis and skewness.

1. Introduction

There are some empirical tests of normality that have grown in popularity and usage and are readily supplied in most popular statistical packages like SPSS, MINITAB, STATA, and R. Anderson Darling (AD), Smirnov Kolmogorov (S-K), Shapiro–Wilk (SW), Jarque – Bera (JB), Shapiro-Francia (S-F), Lilliefors (LF) and D'Agostino-Pearson test (K^2) have grown in popularity, usage and applicability (Ghasemi and Zahediasl, 2012). These normality tests were been further explored in this study to investigate which of them provides the most powerful measures of normality provided data. Using Monte-Carlo simulation exercise developed from samples selected from particular symmetric distributions like Normal distribution, t and Beta distributions, asymmetric distributions like Gamma and Chi–square distributions and other contaminated normal distributions will be generated (Binder and Heermann, 2010). These were also studied to investigate the effects of sample size on the given measure of normality.

The graphical methods that entail the use of Q–Q plots 0and Box–plots are the easiest and simplest tools for testing and depicting normality figuratively. However, they suffer from the problem of subjectivity, as it is not clear to identify the strength and extent of normality by just looking at these plots. That has given rise to the numerical tests of normality like Skewness and Kurtosis as well as several other formal tests of normality (Gujarati, 2003). There are over 44 formally recognized tests of normality used in statistical procedures whose application heavily depends on several statistical assumptions about the size and symmetry of the data. Any statistician or researcher will often be confused regarding the most appropriate normality test to adopt in the research process. Some tests often reject the null hypothesis of normality in some cases, yet others may prove insignificant in similar cases. That further ushers in confusion re–garding the most robust and powerful test for normality given the nature and distribution of the data.

The study aims to provide a basis for identifying the right test by comparing the empirical power of the most popular statistical tests of normality to help the statistician or researcher know the most appropriate tool to use given the nature of the data. The goal is to obtain a parsimonious test statistic that utilizes the least information to produce the best statistic under extreme and varying conditions of the distribution using the empirical power results. An empirical power of 0.5 would imply that if a given experiment is run multiple times, the existing effects would be discovered at least four times in the entire experiment (Brandt, 2012). Despite the fact that 0.8 may be considered the most preferred power of a given test, this comparative study is mainly interested in investigating the variation in such empirical power other than the magnitude of the strength. Thus, a statistical power of 0.4 would be preferred to a statistical power of 0.2 given two tests with reported statistical power. In its simplest form, the law states that as the number

of randomly generated variables that are identically distributed increases, the sample estimates, for instance, the sample mean will tend towards the exact theoretical estimates of the population mean. The famous Italian Mathematician, called Gerolamo Cardano, initially developed it. Though, first mathematically proved in 1713 by another Italian Mathematician, Jakob Bernoulli (Buckley, 2003).

Probabilistically, the law of large numbers provides a description of the results that would be obtained suppose an experiment is to be performed several times. Since initial assumptions will be made in the Monte Carlo simulation exercise in this study, the law of large numbers will imply that as the sample sizes increase due to repeated iterations, the obtained samples will be closer to the actual values (Peligrad and Utev, 2006). The objective of this study is to assess the adequacy of different normality tests; examine the empirical power of the normality tests used; and to compare the empirical power of these normality tests using natural data (February 2016 Ugandan Election Results as cited by Solomon (2016)) for consistency.

2. Literature Review

Shapiro and Wilk in the 18th century in which after conducting a thorough investigation concluded that the Shapiro–Wilk produced the best results under symmetric distribution carried out initial studies regarding empirical power. Nevertheless, for the distributions that were symmetric and short–tailed, the Student Range Test produced the best results. However, it was associated with very low power in the instances where the simulated distributions were asymmetric in nature. In such tests, b_1 and $\sqrt{b_2}$ worked well but were associated with less power if compared to the S–W normality tests (Shapiro *et al.*, 1968).

Further studies modified the original Shapiro comparative studies of the tests of normality by adjusting the power calculations that were initially used by Shapiro in his original studies. For instance, Stephens (1974) criticized the original Shapiro study by asserting that his use of the critical value for carrying out the tests yet bearing the assumptions of the EDF was inappropriate. Stephens (1974) further contends that Shapiro's computations were not right because they were based on a restrictive assumption that the variance and mean of the population were known, an entirely wrong assumption according to Stephens (Stephens, 1974). Stephens (1974) thus conducted a completely new study that recalculated the critical values associated with the distributions under the assumption that the means and variances were unknown. However, the results of his findings were not so different from those originally obtained by Shapiro in his original studies (Beirlant, 2006). Similar studies also indicate that the Anderson–Darling A^2 statistic produces results that are stronger than those associated with W-test for special gamma distributions Gamma (2, 1), Beta (2, 2), lognormal distributions, t (4) and the Weibull (2, 1)

distributions. However, the JB–test was found to be more potent than the A^2 for the Lognormal distributions (Yazici and Yolacan, 2007). However, when it comes to the short–tailed symmetric distributions like t (1), A^2 produced better results than all the others.

However, for the cases where the alternatives are either logistic or t (3), JB–test produces the most comprehensive results (Noughabi and Arghami, 2010). Additionally, the W–test provides more powerful results if compared to the JB test. Yap and Sim (2011) vehemently claims that the JB statistic would be the most appropriate test for instances where the data is symmetric and generated from distributions that have long tails. Available literature indicates that most of the previously conducted studies regarding empirical power seem to heavily rely on the simulated and empirical power. A handful of such studies have considered the behaviour of these empirical powers under natural data. The study tests the consistency of these empirical powers using real data collected from the Uganda electoral commission for the 2016 general election results, as cited by Solomon (2016), mainly concentrating on the results obtained by the two main presidential contenders, Pr. Yoweri Kaguta Museveni and Col, Kiiza Besigye.

3. Methods and Materials

Six distributions were selected from the Empirical Density Function (Kolmogorov–Smirnov and Anderson–Darling), from the correlation/regression family of distributions (Shapiro–Wilk and Shapiro–Francia) and lastly the moment–based normality test (Jaque–Bera and D'Agostino Pearson) are considered in the study. The comparison of their empirical power is made using natural data. These empirical powers are obtained using R, and advanced statistical software for data analysis. The data adopted for use in this study includes the February 2016 Ugandan election results as cited by Solomon (2016). All the variables of interest in this data set are quantitative in nature to facilitate empirical determination and compilation of several statistical tests like the measures of central tendency and spread. These statistics were obtained prior to estimation of the univariate normality figures to ensure that they conform to the requirements for proceeding with tests of normality.

However, in the instances where the data is too distorted to aid a proper differentiation of the tests, transformation techniques preferably logarithms transformation were used to attempt to make the data approximately normally distributed before the various empirical powers for the normality tests are carried out during the analysis. On the other hand, the instances of outliers in the omnibus data were minimized by the use of winsorizing, a statistical technique that reduces the outlying figures in any data to the closest or nearest approximate from the data. The procedure was preferred to trimming because it does not result into loss of statistical power due to a reduction in the sample size. The power of any given statistical test refers to that probability that given

statistic will be able to reject altogether the null hypothesis that is also false in the statistical analysis. It has an inverse relationship to the value beta also referred to as the probability of making a type II error. It relates to that probability or likelihood that any given study will be able to detect the effect when there is an event to be detected. The statistical power of a test is mostly affected and dependent on the overall size of the sample used in the sampling process for detection. Statistically, the power of a test can be viewed as the probability of accepting the alternative hypothesis (rejecting the null hypothesis) when it is true as represented in the model (1).

$$Power \ of \ a \ Test = P(rejection \ of \ H_0 | H_1 \ is \ true)$$

$$= P(Making \ a \ Type \ II \ Error)$$

$$= 1 - \beta \tag{1}$$

4. Computer Software Used

The analysis was done using a combination of different statistical packages (EXCEL, R, and STATA). The data was gathered from the Uganda Electoral Commission website and cleaned using EXCEL package. After the data cleaning process, the data analysis was done using a combination of R (S-plus programing). The analysis involved writing programs that tested the data for normality of the different normality tests specified in the study. All the graphical visualizations were done in STATA.

5. Numerical Results and Discussion

Despite the fact that simulated data provides novel results for comparison of the normality esti-mates, real life scenarios often present data that could produce very controversial and contradicting results if compared to the results from the simulated data. The 2016 Uganda election results, as cited by Solomon (2016), presents a mix of asymmetry and heavy outlying results, purposely se-lected in this study to investigate the behavior of these tests under conditions other than those inherently simulated in a systematic manner.

The election results data used in this analysis includes the results obtained by the two tight pres-idential contenders President Yoweri Kaguta Museveni from the National Resistance Movement (NRM) and his close rival Col. Dr. Kiiza Besigye from the Forum for Democratic Change Party (FDC). It also includes the data for the overall registered voters in the 2016 general elections and well as the total valid votes that were casted from the 112 districts in Uganda at the time of the elections. The table presents a summary of the normality tests associated with the 2016 Uganda

election results in which Mr. Yoweri Kaguta emerged the winner with over 60% of the valid votes cast. The results indicate that the regression and correlation–based normality tests produce the best results and the EDF based tests produced the least powerful results for the election results (Solomon, 2016).
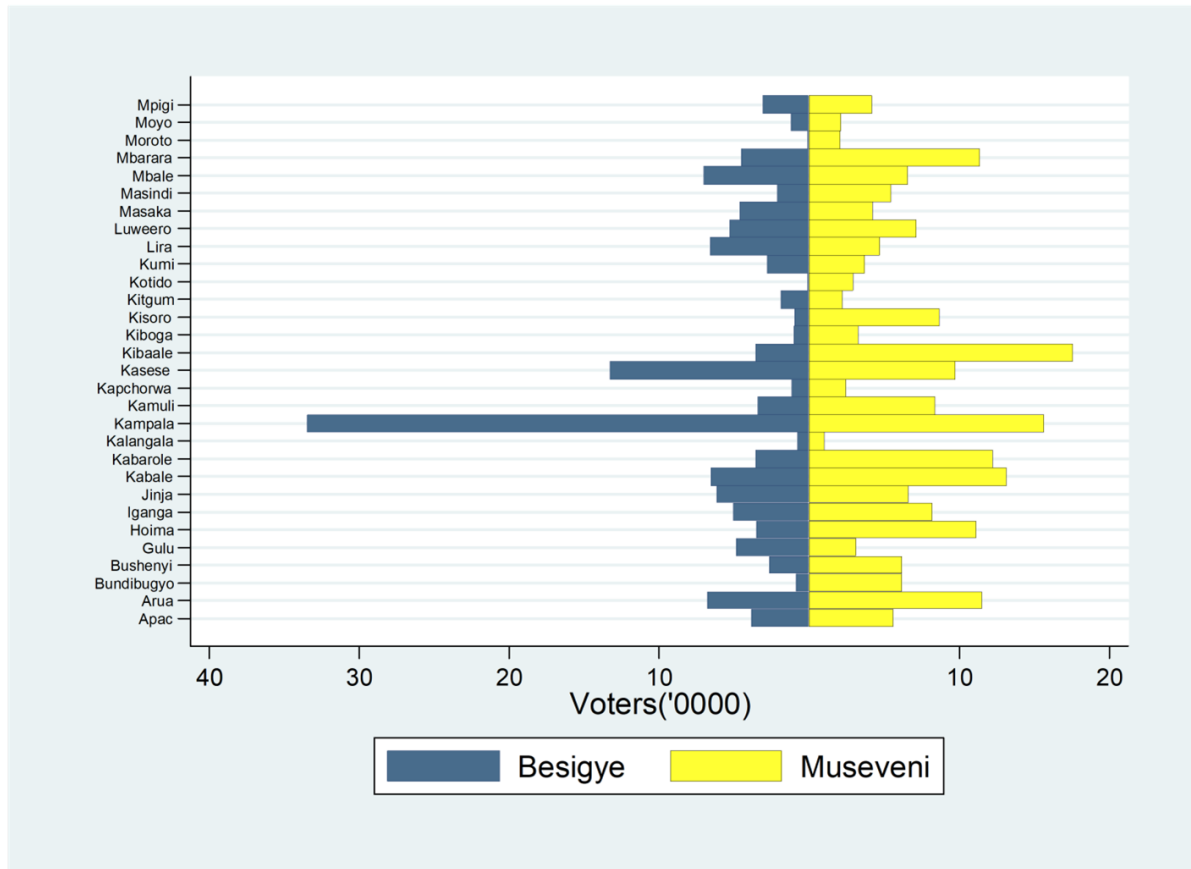

FIGURE 1. Pyramid comparing votes received by Museveni and Besigye

Source: Solomon (2016)

The pyramid in Figure 1 presents a summary of the valid votes counted summarizing those received by the bitterest contenders from a few selected districts that took part in the elections. The Presidential results data contained a whole mix of distributions and inherent randomness partly due to the population distribution by region (district) in Uganda. Figure 1 further shows that these results were too mixed and far from the standard normal distribution as they appear to be skewed to the right, a distribution much similar to a chi distribution discussed earlier. However, due to the huge outliers, generalizations and conclusions regarding the nature of the distributions can be contested on statistical grounds.

TABLE 1. Empirical powers associated with the election results

	K-S.		Shapiro-Wilk		Shapiro- Francia		Anderson-Darling		Jaque-Bera		AD	
	Sig.	power	Sig.	power	Sig.	Power	Sig.	power	Sig.	power	Sig.	Power
Ln_Museveni	.20	80	.03	96.51	.09	91.38	.33	66.64	.42	58.25	.39	60.97
Ln_Besigye	.03	96.82	.00	100.00	.0	100	.00	100	.00	100	.00	100.0
Ln_vvotes	.20	80	.10	90.47	.25	75.24	.49	50.64	.23	76.90	.05	94.87
Ln_rvoters	.20	80	.00	99.82	.01	99.22	.02	97.66	.10	89.66	.01	99.49
Average power		84.2(4)		96.70(1)		91.46(2)		78.73(6)		81.20(5)		88.83(3)

Keys: K-S= Kolmogorov-Smirnov; AD=D'Agostino-Pearson Test
Source:Authors' Computation (2022).

Table 1 presents a summary of the empirical powers that were obtained when the normality tests were run on this data. The results of the analysis indicate that S–W produced the most powerful results (power = 96.7) followed by S–F (power=91.46), D–A (power=88.83), K–S (power=84.2), J–B (power=81.2) and lastly A–D (power=78.73).
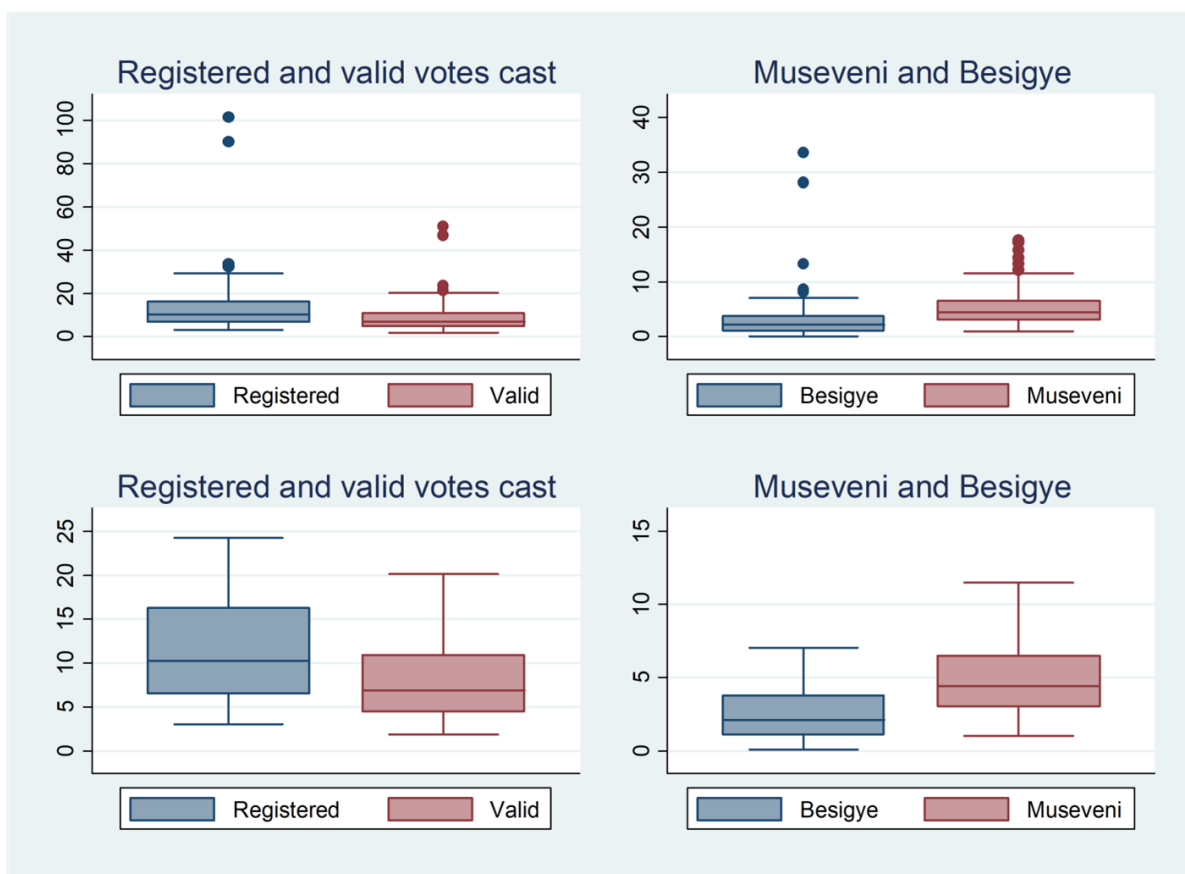

FIGURE 2. Box plot depicting the results of the 2016 general election
Source: Solomon (2016)

Figure 2 presents a summary of the votes received by the two tightest contenders in the 2016 Uganda general elections as cited by Solomon (2016). Preliminary analysis indicates that the data

contains very large outliers that would result into significant but erroneous results if the normality tests were applied with adjustments for both outliers and asymmetry. The boxplots depict the distributions of the total votes cast; the overall registered voters as well as the share received by Besigye and Museveni. The upper section of the drawing canvas on the right shows boxplots of the results including the outliers that were later removed by winsorizing to produce the box plots in the lower region of the plotting canvas. The box plot particularly shows the spread of the data before adjusting for outliers and after adjusting for the outliers through a statistical procedure known as winsorizing. The results indicate that the correlation and regression normality tests are more effective for transformed data compared to the EDF and moment based normality tests. Since most data in every day parlance does not directly conform to any given form of statistical distribution as data obtained from simulation discussed in the previous section, it would be proper to make a suitable choice before choosing any sought off statistical test of normality.

6. Discussion of Findings

Preliminary analysis of the data in its natural form produced results that were indifferentiable and inconsistent with literature due to outliers. To obtain better representative results, the data was transformed after adjusting for outliers. The transformation was meant to remove any implicit outliers and in–data non–uniformities that have significant effect on the result. The transformation also helped in reducing the loss in statistical power that would be associated with a reduction in the degrees of freedom when some variables are dropped from the analysis. The transformation was done to impose some form of normality to the data that would also make it possible to obtain results that would aid ranking of the normality tests. Following transformation and adjustment for the outliers in the data, it possessed a type of distribution so mixed that there was no significant difference in the normality test. Literature indicates that the correlation/regression based normal–ity tests perform better under short–tailed and asymmetric distributions. Removal of the outliers reduced the tails of election results data since the winsorizing process meant that extreme values in the data take on the values of the nearest points that are within the near–normal data. That accorded an added advantage to S–W and S–F, two correlation/regression based normality tests an added strength. Furthermore, since the winsorizing process did not inherently remove all instances of kurtosis and Skewness in the data, two important constructs of the moment–based normality tests, D–A and J–B form that family produced results better than those associated with K–S and A–D. Furthermore, since the sample size was also relatively low (112), an environment known to foster the correlation/regression normality tests, S–W and S–F emerged most powerful.

7. Conclusions

For the election data, a hugely contaminated data set containing a mix–up of outliers, kurtosis, and skewness, S–W produced the most powerful results followed by S–F, D–A, K–S, J–B and lastly

A–D. Due to the low sample size that was associated with data (n =112), the correlation and regression family of normality tests produced the best results. The data also indicates that the moment based tests produced results better that the EDF–based tests due to huge kurtosis and skewness statistics in the data. Furthermore, the tails created due to winsorizing reduced the power of the EDF–based normality tests. The results were less similar to those reported by the asymmetric data due to the huge skewness created by the uneven distribution of the voting population across the districts that participated in the election process.

8. Recommendations

In instances where data of unknown distribution with a relatively small sample size but known to be contaminated by outliers is to be dealt with, it should be transformed to remove the outliers and then the S-W is applied though its produces results similar to the EDF normality test. The moment-based tests should be considered last as they are effective only if the sample size is large enough especially in the J-B test.

References

[1] H.A. Noughabi, N.R. Arghami, A new estimator of entropy, J. Iran. Stat. Soc. 9 (2010) 53-64.

[2] J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, Statistics of extremes: theory and applications, John Wiley and Sons (2006).

[3] K. Binder, D. Heermann, Monte Carlo simulation in statistical physics: an Introduction, Springer Science & Business Media (2010).

[4] S. Brandt, Data analysis: statistical and computational methods for scientists and engineers, Springer Science & Business Media (2012).

[5] J.J. Buckley, Law of large numbers. In Fuzzy Probabilities, Physica–Verlag HD (2003) 145-146.

[6] A. Ghasemi, S. Zahediasl, Normality tests for statistical analysis: a guide for non–statisticians, Int. J. Endocrinol. Metab. 10 (2012) 486–489. https://doi.org/10.5812/ijem.3505.

[7] D.N. Gujarati, Basic Econometrics, Fourth Edition McGraw Hill Gujarati DN (2003).

[8] M. Peligrad, S. Utev, Central limit theorem for stationary linear processes, Ann. Probab. 34 (2006), 1608-1622. https://doi.org/10.1214/009117906000000179.

[9] S.S. Shapiro, M.B. Wilk, H.J. Chen, A comparative study of various tests for normality, J. Amer. Stat. Assoc. 63 (1968) 1343-1372.

[10] M.M. Solomon, Comparative empirical power of univariate normality tests: a case study of the 2016 ugandan general election results, Dissertation, Kampala International University, Uganda (2016).

[11] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Amer. Stat. Assoc. 69 (1974) 730-737.

[12] B.W. Yap, C.H. Sim, Comparisons of various types of normality tests, J. Stat. Comput. Simul. 81 (2011) 2141–2155. https://doi.org/10.1080/00949655.2010.520163.

[13] B. Yazici, S. Yolacan, A comparison of various tests of normality, J. Stat. Comput. Simul. 77 (2007) 175–183. https://doi.org/10.1080/10629360600678310.