Comparing the Simulated Power of Discrete Goodness-of-fit Tests for Small Sample Sizes

Michael Steele[1,*], Janet Chaseling[2] and Cameron Hurst[1]

[1]*James Cook University,* [2]*Griffith University*

[*]*Corresponding author: Mike.Steele@jcu.edu.au*

Abstract

Although a variety of goodness-of-fit test statistics are used by applied researchers, studies of their power have been limited. This paper investigates the simulated power of six goodness-of-fit test statistics for discrete data for small sample sizes. The null distribution is uniform and the simulated power of each of the test statistics is calculated for a number of alternative distributions including trend, triangular, flat/platykurtic type, sharp/leptokurtic type and bimodal.

KEY WORDS: Goodness-of-fit, simulation.

Introduction

The limited amount of published research into the power of discrete goodness-of-fit (GOF) test statistics suggest that for small sample sizes the discrete Kolmogorov-Smirnov (*S*), Cramér-von Mises (W^2) and Anderson-Darling (A^2) test statistics are more powerful than Pearson's Chi-Square (χ^2) for a uniform null against alternative trend type distributions (Choulakian *et al.,* 1994; Pettitt and Stephens 1977). Only a limited amount of research has been conducted into the powers of these, and other discrete GOF test statistics, for other alternative distributions (Steele and Chaseling, 2006a and 2006b); From, 1994; Steele *et al.,* 2005). This paper investigates the powers of the six discrete GOF test statistics in Table 1 for a uniform null against several alternative distributions given in Table 2 for small sample sizes. The simulation methods used are discussed in Section 2, the results of the power studies are discussed for each alternative distribution in Section 3 and concluding recommendations based on power for small sample sizes are given in Section 4.

Table 1: Test statistics for the power study

Name of Test Statistic	Equation	Author
Pearson's Chi-Square	$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$	Pearson (1900)
Discrete Kolmogorov-Smirnov	$S = \max_{1 \le i \le k} \lvert Z_i \rvert$	Pettitt and Stephens (1977)
Discrete Cramér-von Mises	$W^2 = N^{-1} \sum_{i=1}^{k} Z_i^2 p_i$	Choulakian *et al*. (1994)
Discrete Watson	$U^2 = N^{-1} \sum_{i=1}^{k} (Z_i - \bar{Z})^2 p_i$	Choulakian *et al*. (1994)
Discrete Anderson-Darling	$A^2 = N^{-1} \sum_{i=1}^{k} \frac{Z_i^2 p_i}{H_i (1 - H_i)}$	Choulakian *et al*. (1994)
Nominal Kolmogorov-Smirnov	$NS = \frac{1}{2} \sum_{i=1}^{k} \lvert O_i - E_i \rvert$	Pettitt and Stephens (1977)

where k is the number of cells, N is the sample size, p_i is the probability of an event occurring in cell i, E_i is the expected frequency in cell i, O_i is the observed frequency in cell i, $Z_i = \sum_{j=1}^{i} (O_j - E_j)$,

$H_i = \sum_{j=1}^{i} E_j$ *and* $\bar{Z} = \sum_{j=1}^{k} Z_j p_j$.

Table 2: Distributions used in the power study

Description	Cell Probabilities									
	1	2	3	4	5	6	7	8	9	10
Uniform	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Decreasing	0.32	0.13	0.10	0.08	0.07	0.07	0.06	0.06	0.05	0.05
Step	0.05	0.05	0.05	0.05	0.05	0.15	0.15	0.15	0.15	0.15
Triangular (∨)	0.17	0.13	0.10	0.07	0.03	0.03	0.07	0.10	0.13	0.17
Platykurtic	0.04	0.11	0.11	0.12	0.12	0.12	0.12	0.11	0.11	0.04
Leptokurtic	0.05	0.05	0.05	0.05	0.30	0.30	0.05	0.05	0.05	0.05
Bimodal	0.05	0.11	0.17	0.11	0.06	0.06	0.11	0.17	0.11	0.05

Simulations and Methods for the Power Study

So that meaningful comparisons can be made with other published work this power study uses the same simulation techniques as Steele and Chaseling (2006). The sample sizes are 1, 2, 3 and 5 per cell under the uniform null distribution. The power of each test statistic is approximated from 10000 simulated random samples from the null and alternative distributions. A problem arises in that the exact 5% significance level selected is rarely achieved as the simulated distribution of the test statistic is discrete. To attempt to overcome this problem, the powers are obtained from both sides of the 5% level and linearly interpolated.

Power Study Results

Decreasing Trend Alternative for Small Sample Sizes

For very small sample sizes it is shown in Figure 1 that A^2, W^2 and S have greater powers than the two nominal test statistics χ^2 and NS, and the circular test statistic U^2. The order of the powers of these test statistics are similar to those obtained by Choulakian *et al*. (1994) for a decreasing alternative with 12 cells. It should be noted that these authors did not consider the NS test statistic. Of importance for the applied researcher is that for sample sizes of approximately 5 per cell under the null distribution the powers of all these test statistics, with the exception of NS, are quite close to 1 for this decreasing alternative distribution.

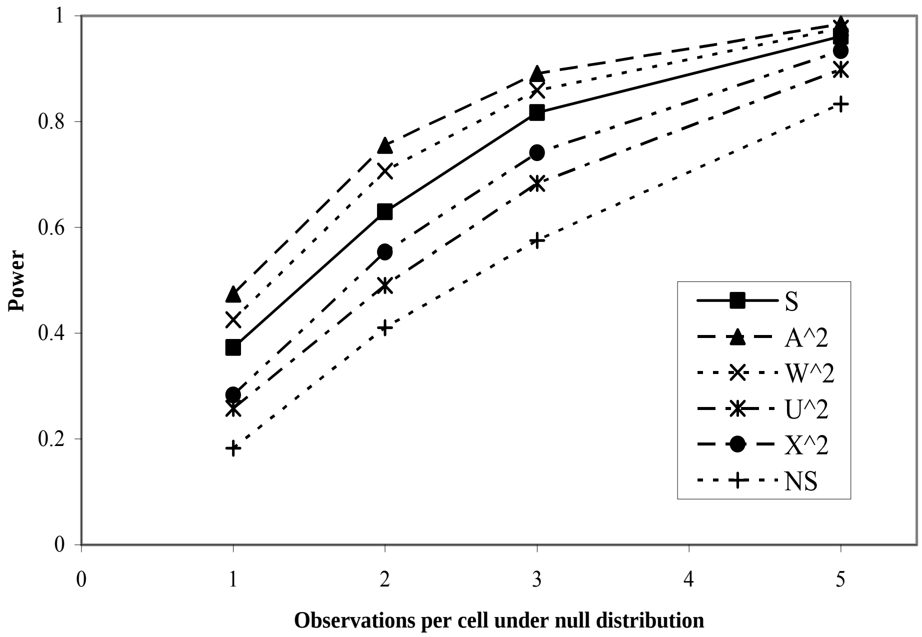

Figure 1: The power of the six test statistics for a uniform null against a decreasing alternative when the same size is small

Step-type Trend Alternative for Small Sample Sizes

The results from Figure 2 show that the powers of the nominal type test statistics χ^2 and *NS* are substantially smaller than the other four test statistics used in this power study. Importantly these differences still exist when the sample size is 5 per cell under the null distribution. Although this paper is concerned with smaller sample sizes it is worth noting that Steele and Chaseling (2006) showed that these differences were negligible for sample sizes of at least 10 per cell under the null distribution.

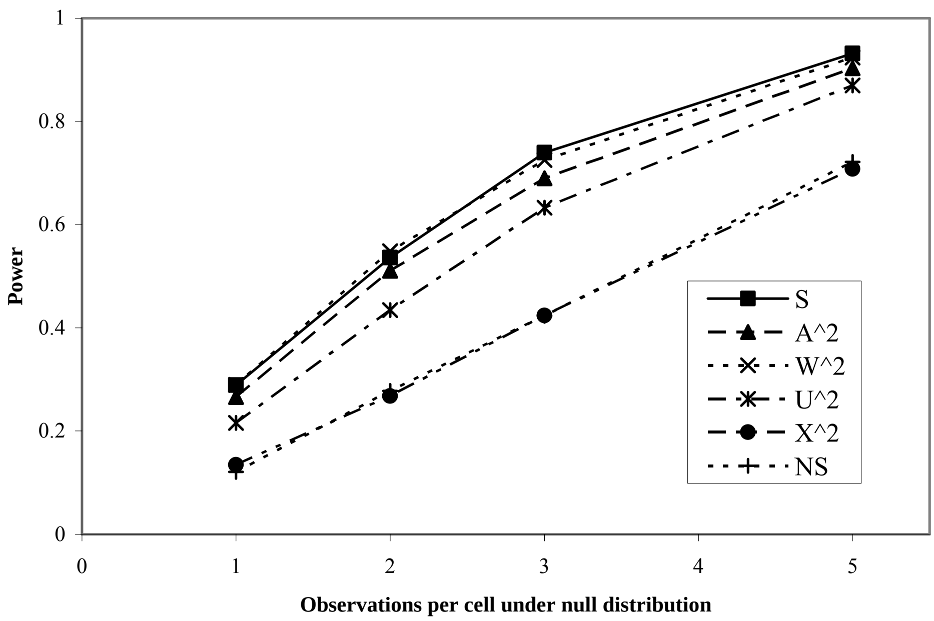

Figure 2: The power of the six test statistics for a uniform null against a step type alternative when the same size is small

Triangular Alternative for Small Sample Sizes

It is quite obvious in Figure 3 that the power of U^2 is substantially greater than all other test statistics considered for this particular triangular alternative distribution. With the exception of the very small sample size of 1 per cell under the uniform null, the power of U^2 is shown to be about double that of the next most powerful test statistic. Based on these results only U^2 can be recommended as a suitably powerful test statistic for such situations.

Platykurtic Alternative for Small Sample Sizes

The powers of all 6 test statistics are shown in Figure 4 to be very poor for all sample sizes considered for this platykurtic alternative distribution and none of these test statistics can be recommended to applied researchers based on power for the smaller sample sizes. It is also interesting to note that the powers of A^2, W^2 and S have not noticeably differed from their starting value of 0.05 for all sample sizes considered.

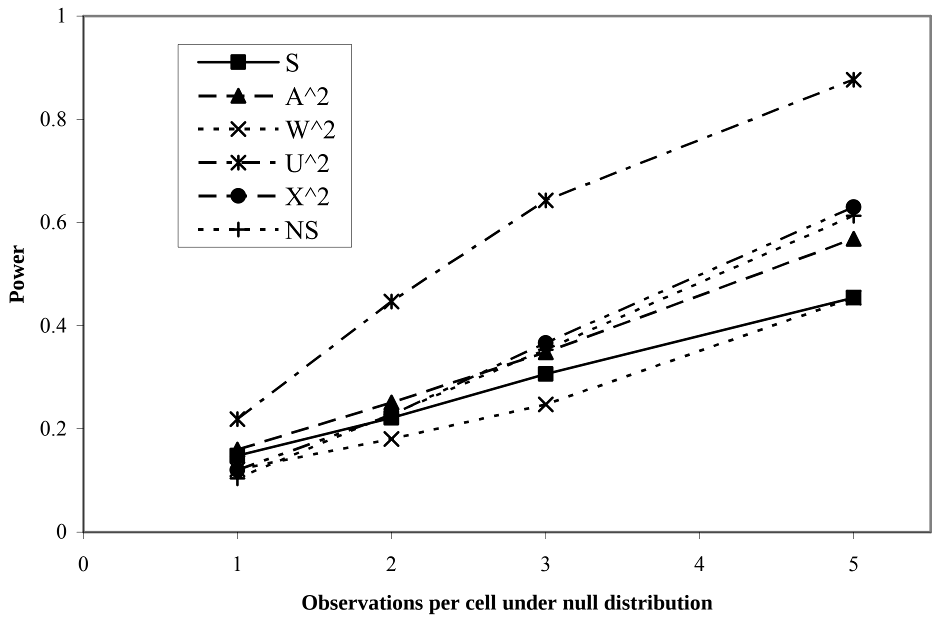

Figure 3: The power of the six test statistics for a uniform null against a triangular alternative when the same size is small

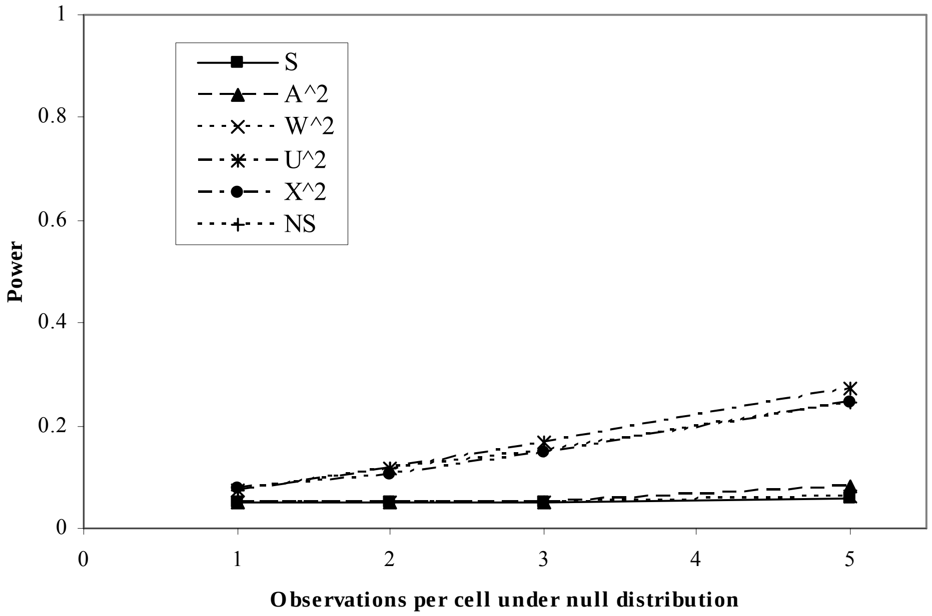

Figure 4: The power of the six test statistics for a uniform null against a platykurtic alternative when the same size is small

Leptokurtic Alternative for Small Sample Sizes

The χ^2, U^2 and, to a lesser extent, the *NS* test statistics are shown in Figure 5 to have greater power than the other three test statistics for the uniform null against this particular leptokurtic alternative distribution for all sample sizes. Importantly the powers of χ^2 and U^2 are approximately equal for all sample sizes and are quite high (approximately 0.50) for the very small sample size of 1 per cell under the uniform null distribution. Also the powers of all six test statistics are quite close to 1 when the sample sizes reach 5 per cell under the uniform null distribution.

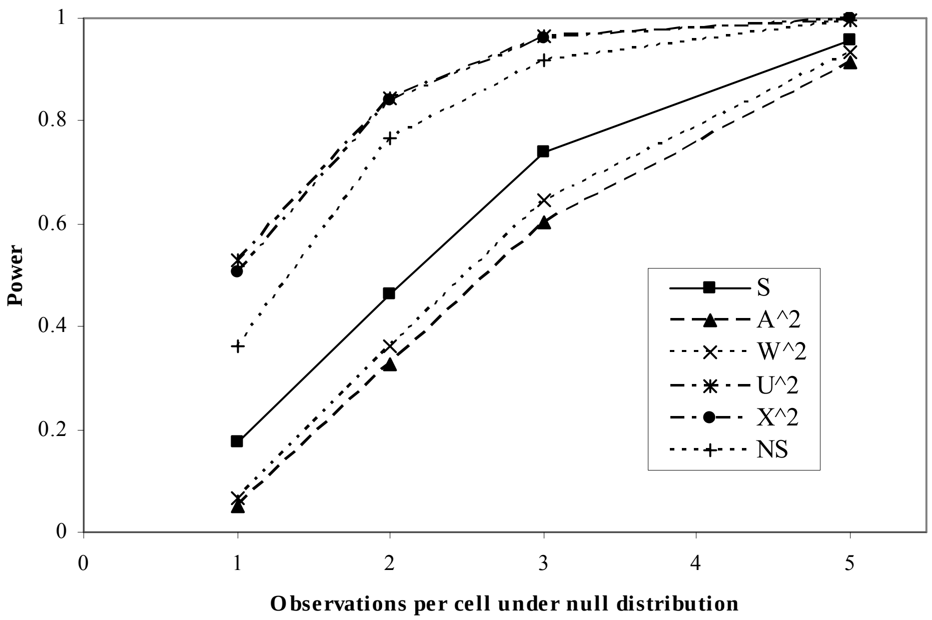

Figure 5: The power of the six test statistics for a uniform null against a leptokurtic alternative when the same size is small

Bimodal Alternative for Small Sample Sizes

In general it is shown in Figure 6 that the powers of all six test statistics are quite poor for this bimodal alternative distribution with the exception of the two nominal test statistics χ^2 and *NS*. The four other test statistics are shown to have power not substantially changing from the initial power value of 0.05. Clearly none of the six test statistics can be recommended for small sample sizes for this alternative distribution.

Conclusions and Recommendations

As is common in similar power studies it is difficult to recommend one particular goodness-of-fit test statistic as being overall more powerful than the other test statistics. This makes it difficult for non-statistical users of goodness-of-fit tests who will persevere with Pearson's χ^2 for simplicity. While further studies into the powers of goodness-of-fit tests for categorical data are needed for many other test statistics, null and alternative distributions some general recommendations can be made:

i) for trend type alternatives discussed in Sections 3.1 and 3.2 it is clear that the test statistics based on the cumulative sum of the differences between the observed and expected frequencies (A^2, W^2 and S) are the more powerful

ii) for the specific non-trend alternatives discussed in this power study Pearson's χ^2 is as powerful as any other test statistic with the exception of the triangular alternative distribution where U^2 was much more powerful.

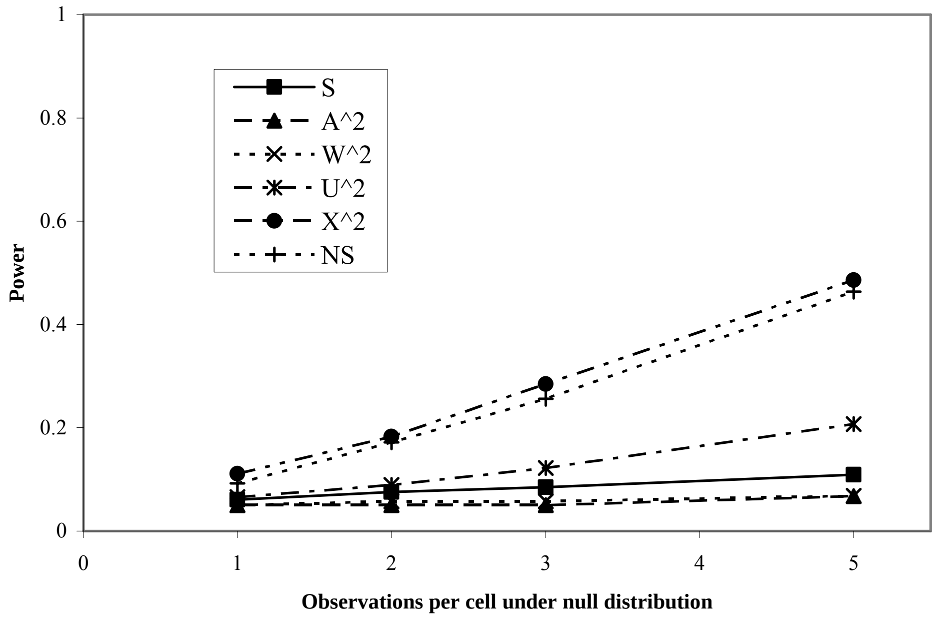

Figure 6: The power of the six test statistics for a uniform null against a bimodal alternative when the same size is small

References

Choulakian, V., Lockhart, R., and Stephens, M. 1994. Cramér-von Mises statistics for discrete distributions. The Canadian Journal of Statistics 22: 125-137.

From, S. 1996. A new goodness-of-fit test for the equality of multinomial cell probabilities verses trend alternatives. Communications in Statistics – Theory and Methods 25: 3167-3183.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 5: 157-175.

Pettitt, A., and Stephens, M. 1977. The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. Technometrics 19: 205-210.

Steele, M., Chaseling, J., and Hurst, C. 2005. Simulated power of the discrete Cramér-von Mises goodness-of-fit tests. p 1300-1304. In: Zerger, A., Argent, R. (eds.). MODSIM05. Proceedings of the International Conference on Modeling and Simulation Advances and Applications for Management and Decision Making, Melbourne, Australia.

Steele, M., and Chaseling, J. 2006a. Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions. Communications in Statistics – Simulation and Computation 35: (IN PRESS).

Steele, M., and Chaseling, J. 2006b. A comparison of the powers of the Chi-Square test statistic with the discrete Kolmogorov-Smirnov and Cramér-von Mises test statistics. p 615-621. In: Rizzi, A., and Vichi, M. (Eds.). COMPSTAT2006. Proceedings of the Computational Statistics Conference. Rome, Italy.