Simulated Power of the Discrete Cramér-von Mises Goodness-of-Fit Tests

[1]Steele, M., [2]Chaseling, J. and [3]Hurst, C.

[1]School of Mathematical and Physical Sciences, James Cook University, [2]Australian School of Environmental Studies, Griffith University, [3]School of Information Technology and Mathematical Sciences, University of Ballarat. E-Mail: Mike.Steele@jcu.edu.au

Keywords: *Goodness-of-fit; Power; Empirical distribution function.*

EXTENDED ABSTRACT

The use of goodness-of-fit test statistics for discrete or categorical data is widespread throughout the research community with the Chi-Square the most popular when a researcher aims to determine if observed categorical data differs from a hypothesized multinomial distribution. Even for ordinal categorical data, the use of empirical distribution function (EDF) test statistics such as the Kolmogorov-Smirnov, the three Cramér-von Mises (A^2, W^2 and U^2 as defined below) and various modifications of these are limited in the literature. Power studies of the EDF type test statistics are even more limited.

This paper compares the simulated power of the three Cramér-von Mises test statistics with that of the Chi-Square test statistic for a uniform null hypothesis against a variety of alternative distributions which are summarized in Figure 1. Recommendations are made on which is the most powerful test statistic for the predefined alternative distributions.

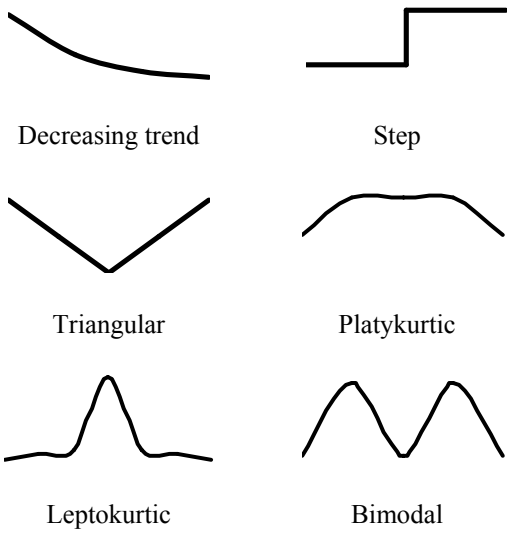

Figure 1. Type of alternative distributions used in the power studies.

The results of the simulated power studies in this paper lead to the following general recommendations:

- For trend type alternatives A^2 and W^2 appear much more powerful than U^2 and χ^2. (See Figure 2 for a uniform null against a decreasing trend alternative distribution).

- For all the other investigated alternative distributions U^2 and χ^2 appear much more powerful than A^2 and W^2. (See Figure 3 for a uniform null against a leptokurtic type alternative distribution).

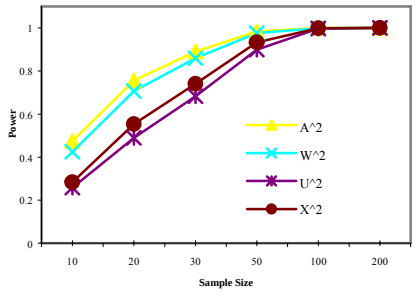

Figure 2. Powers for a uniform null and a decreasing alternative distribution.

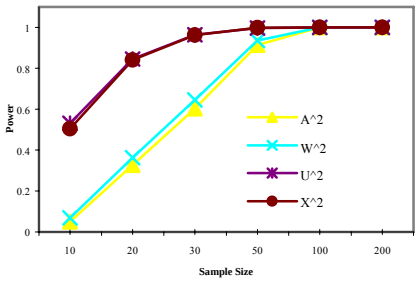

Figure 3. Powers for a uniform null and a leptokurtic alternative distribution.

1. INTRODUCTION

Although designed for ordinal categorical data, the empirical distribution function (EDF) type goodness-of-fit test statistics Cramér-von Mises (W^2), Anderson-Darling (A^2) and Watson (U^2) as defined by Choulakian *et al.* (1994) are not widely used in the applied literature. These authors have used simulation studies to show that A^2 and W^2 are relatively more powerful than the Chi-Square (χ^2) test statistic (Pearson 1900) when the null distribution is uniform and the alternative distribution follows a trend. The test statistics are specified in Table 1.

Table 1. Test statistics used in the power study.

Test Statistic	Equation
Discrete Cramér-von Mises	$W^2 = N^{-1}\sum_{i=1}^{k} Z_i^2 p_i$ (1)
Discrete Anderson-Darling	$A^2 = N^{-1}\sum_{i=1}^{k} \frac{Z_i^2 p_i}{H_i(1-H_i)}$ (2)
Discrete Watson	$U^2 = N^{-1}\sum_{i=1}^{k} (Z_i - \bar{Z})^2 p_i$ (3)
Pearson's Chi-Square	$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ (4)

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

There have been limited investigations of the powers of these particular EDF type test statistics. This paper uses simulated powers to extend the studies of Choulakian *et al.* (1994) and From (1996) by comparing the powers of the three Cramér-von Mises type test statistics with the χ^2 test statistic for a uniform null distribution (A_0) against the fully specified alternative distributions summarized in Figure 1 (Decreasing A_1, Step A_2, Triangular or 'bath-tub type' A_3, Platykurtic A_4, Leptokurtic A_5 and Bimodal A_6) and fully defined in Table 2. The uniform null distribution was used because most similar published power studies of discrete goodness-of-fit tests have used such a null distribution however further work on non-uniform null distributions has been undertaken by Steele (2002). For a small number of categories some of the alternative distributions do not clearly exhibit the shapes illustrated in Figure 1. Also the distributions become quite similar for a small number of categories. For this reason a larger number of categories (k=10) was used.

Table 2. Distributions used in the power study.

	Cell Probabilities									
	1	2	3	4	5	6	7	8	9	10
A_0	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
A_1	0.32	0.13	0.10	0.08	0.07	0.07	0.06	0.06	0.05	0.05
A_2	0.05	0.05	0.05	0.05	0.05	0.15	0.15	0.15	0.15	0.15
A_3	0.17	0.13	0.10	0.07	0.03	0.03	0.07	0.10	0.13	0.17
A_4	0.04	0.11	0.11	0.12	0.12	0.12	0.12	0.11	0.11	0.04
A_5	0.05	0.05	0.05	0.05	0.30	0.30	0.05	0.05	0.05	0.05
A_6	0.05	0.11	0.17	0.11	0.06	0.06	0.11	0.17	0.11	0.05

In Section 2 the simulation and linear interpolation techniques used to approximate power are discussed with sample size considerations. The results of the power studies are presented in Section 3 and a summary table of the most powerful test statistic for each alternative distribution is presented in the concluding Section 4.

2. CALCULATION OF THE SIMULATED POWER

For a uniform null distribution over ten cells against the alternative distributions defined in Table 2 the powers of the test statistics are approximated for sample sizes of 10, 20, 30, 50, 100 and 200. The sample sizes represent expected frequencies of 1, 2, 3, 5, 10 and 20 per cell under the uniform null distribution and by selecting these expected frequencies researchers who use goodness-of-fit tests with a minimum requirement of 5 observations per cell can make power comparisons for different minimum number of observations per cell. It is also shown in the results that in most of the situations discussed below that sample sizes of around 20 per cell produce power approximations very close to 1. The powers are estimated using 10000 simulated random samples. The simulated null distribution of each test statistic is discrete which means that a critical value and corresponding power at a significance level of exactly 5% may not be possible. To enable meaningful comparisons of the powers of each test statistic, the powers are obtained for critical values either side of the 5% level, and linearly interpolated to produce the approximate power for the 5% level.

3. POWER STUDY RESULTS

3.1. Uniform Null with a Decreasing (A_1) Alternative

For small sample sizes Figure 4 shows that A^2 and W^2 have powers greater than χ^2 and U^2. The largest cumulative difference between the uniform null and the decreasing alternative distribution occurs at the second cell and as A^2 and W^2 are affected by large cumulative differences at the earlier cells this is one reason why they have larger power under these circumstances. Also χ^2 generally has higher power than U^2. For sample sizes of at least 5 per cell (ie $N \geq 50$ in this example), the powers of all the test statistics are very high.

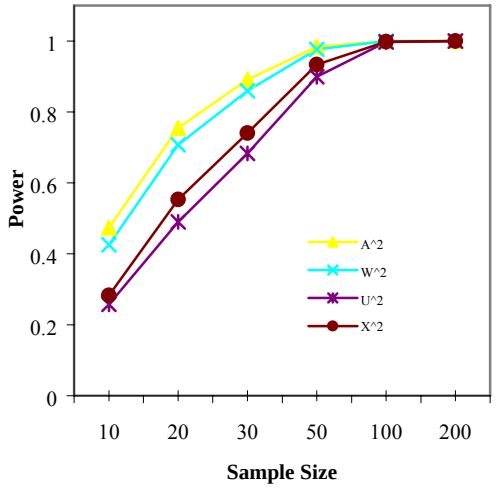

Figure 4. Powers for uniform null and decreasing (A_1) alternative.

3.2. Uniform Null with a Step Type (A_2) Alternative

For the step type distribution the cumulative difference between a uniform null and the step type A_2 distribution increases up to the fifth cell. Because they are more able to detect larger cumulative differences in the earlier cells the test statistics A^2 and W^2 are shown in Figure 5 to be more powerful. It should be noted that the power of U^2 is almost as good as A^2 and W^2 while the power of χ^2 is noticeably less than the three Cramér-von Mises type test statistics. For larger sample sizes of ten or more per cell (ie $N \geq 100$ in this situation) the powers of all four test statistics are very high and approximately the same.

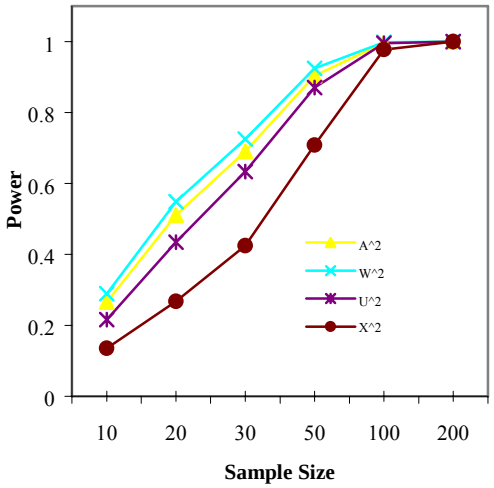

Figure 5. Powers for a uniform null and step type (A_2) alternative.

3.3. Uniform Null with a Triangular (A_3) Alternative

The major cumulative differences between a uniform distribution and the A_3 triangular alternative distribution do not occur in the earlier cells as was the case in Sections 3.1 and 3.2. For this reason it is expected that A^2 and W^2 are less likely to detect a difference and hence have lower power. The U^2 statistic is circular in that although it can be used on ordinal type data, calculation of the test statistic does not depend on which cell is defined as the first. This circular test statistic is shown in Figure 6 to be much more powerful than the other three test statistics. However for larger sample sizes the powers of all the test statistics are approximately the same and high. This result also corresponds to a similar triangular type alternative distribution based on 12 cells by Choulakian *et al.* (1994).

3.4. Uniform Null with a Platykurtic (A_4) Alternative

As the cumulative differences between a uniform null and the A_4 platykurtic alternative distribution are not large the A^2 and W^2 test statistics are expected to have lower power. The power of W^2 is shown in Figure 7 to be very poor for all sample sizes however for the smaller sample sizes of five per cell (that is $N \leq 50$) under the uniform null all the test statistics have poor power. For larger sample sizes χ^2 and U^2 are shown to have much higher power.

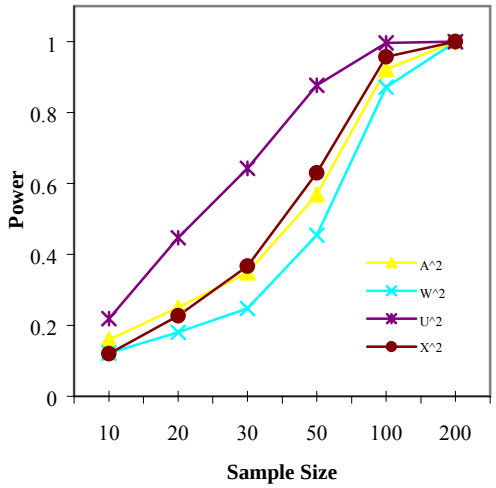

Figure 6. Powers for uniform null and triangular (A_3) alternative.

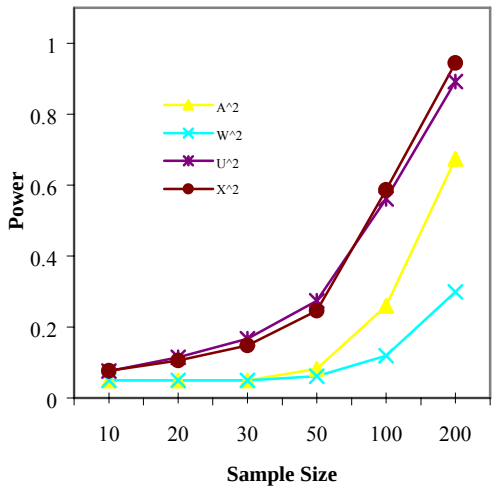

Figure 7. Powers for uniform null and platykurtic (A_4) alternative.

3.5. Uniform Null with a Leptokurtic (A_5) Alternative

As was also the case in Section 3.4, the cumulative differences between the uniform null and the leptokurtic A_5 alternative are quite small for earlier cells and the low powers of A^2 and W^2 in Figure 8 show this to be true for smaller sample sizes. The powers of U^2 and χ^2 are shown to be approximately equal for all sample sizes. It appears that due to its circular nature, U^2 is more able to detect the large cumulative differences which occur at the middle cells.

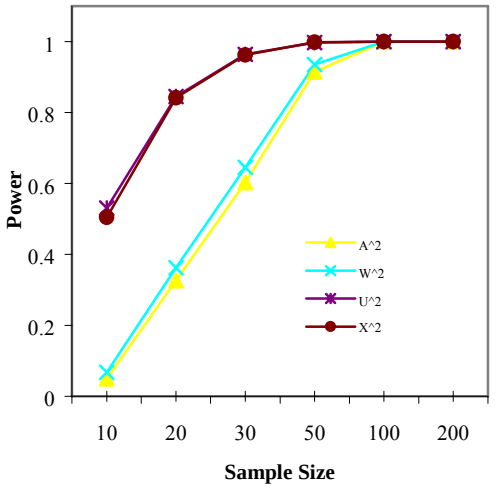

Figure 8. Powers for uniform null and leptokurtic (A_5) alternative.

3.6. Uniform Null with a Bimodal (A_6) Alternative

The powers of the test statistics are shown in Figure 9 to be quite diverse. The power of χ^2 is shown to be approximately double those of the other test statistics for smaller sample sizes. Although the power of U^2 is quite low it is still much larger than the very weak powers of A^2 and W^2.

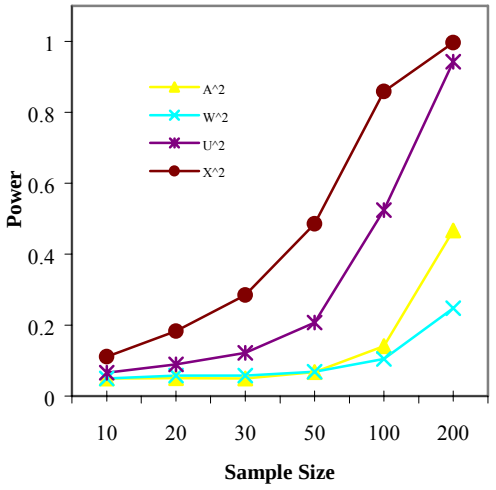

Figure 9. Powers for uniform null and bimodal (A_6) alternative.

4. CONCLUSIONS

Although it is not possible to recommend one of these test statistics as being the most powerful for all situations a very broad summary of the simulated powers in this paper suggests that, particularly for smaller sample sizes:

- For trend type alternatives A^2 and W^2 appear much more powerful than U^2 and χ^2.

- For all the other investigated alternative distributions U^2 and χ^2 appear much more powerful than A^2 and W^2.

Importantly, when considering the power of the test statistic, the simulated results presented in this and other papers suggests that the applied researcher should not blindly use one particular test statistic. However the broad summary above may assist an applied researcher to at least consider alternatives to the χ^2 test when testing whether their observed ordinal data differs from that expected under a multinomial null distribution.

5. REFERENCES

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

From, S.G. (1996), A new goodness of fit test for the equality of multinomial cell probabilities verses trend alternatives, *Communications in Statistics-Theory and Methods,* 25(1996) 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 Series 5*, 50(1900) 157-175.

Steele, M., Chaseling, J. and Hurst, C. (2005), A power study of goodness-of-fit tests for categorical data, 55th Session of the International Statistical Institute, Proceedings, Sydney, Australia.

Steele, M (2002), *The power of categorical goodness-of-fit test statistics*, PhD thesis, Griffith University, Brisbane, Australia.