18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Evaluating the statistical power of goodness-of-fit tests for health and medicine survey data

Steele, M. [1,2], **N. Smart** [1], **C. Hurst** [3] **and J. Chaseling** [4]

[1] *PHCRED, Faculty of Health Science and Medicine, Bond University*
[2] *Faculty of Business, Technology and Sustainable Development, Bond University*
[3] *Faculty of Health, Queensland University of Technology*
[4] *Faculty of Health Science and Medicine, Bond University*
Email: misteele@bond.edu.au

Abstract: Goodness-of-fit test statistics are widely used in health and medicine related surveys however little regard is usually given to their statistical power. This paper investigates the simulated power of five categorical goodness-of-fit test statistics used to analyze health and medicine survey data collected on a 5-point Likert scale.

The test statistics used in this power study are Pearson's Chi-Square, the Kolmogorov-Smirnov test statistic for discrete data, the Log-Likelihood Ratio, the Freeman-Tukey and the special case of the Power Divergence statistic defined by Cressie and Read (1984). Recommendations based on these simulations are provided on which of these goodness-of-fit test statistics is the most powerful overall and which is the most powerful for the predefined uniform null against the four general shaped alternative distributions (see Figure 1) investigated in this paper.

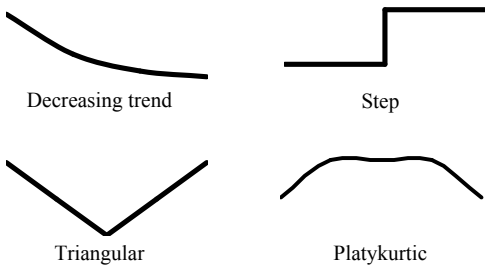

Figure 1. Type of alternative distributions used in the power studies.

Keywords: *goodness-of-fit, power, chi-square tests, discrete*

1. INTRODUCTION

Although widely used in the analysis of health and medicine related survey data and other types of categorical data, only a limited number of published studies are available on the power of discrete goodness-of-fit test statistics. Those available typically compare the power of the Pearson (1900) Chi-Square with test statistics based on the empirical distribution function (e.g. Choulakian et al., 1994, Pettitt and Stephens, 1977, From, 1996, Steele and Chaseling, 2006) however the recent papers by Steele et al. (2008) and Ampadu (2008) do compare several goodness-of-fit test statistics that are asymptotically equivalent to the χ^2 distribution.

Table **1**. Test statistics used in health and medicine surveys using Likert data

Test Statistic	Equation
Pearson's Chi-Square (Pearson, 1900)	$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
Discrete Kolmogorov-Smirnov (Pettitt and Stephens, 1977)	$KS = \max_{1 \le i \le k} \lvert O_i - E_i \rvert$
Log-Likelihood Ratio (Wilks, 1935)	$G = 2\sum_{i=1}^{k} O_i \ln\left(\frac{O_i}{E_i}\right)$
Freeman-Tukey (Freeman and Tukey, 1950)	$FT = 4\sum_{i=1}^{k}\left(\sqrt{O_i} - \sqrt{E_i}\right)^2$
Power Divergence with λ=⅔ (Cressie and Read, 1984)	$PD = \frac{2}{\lambda(1+\lambda)}\sum_{i=1}^{k} O_i\left[\left(\frac{O_i}{E_i}\right)^{\lambda} - 1\right]$

where k is the number of cells, O_i is the observed frequency in cell i, E_i is the expected frequency in cell i.

This paper considers discrete distributions on a 5-point Likert (1932) scale as commonly used in health and medicine surveys and compares the Monte Carlo simulated power of Pearson's Chi-Square, the discrete Kolmogorov-Smirnov, the Log-Likelihood Ratio, the Freeman-Tukey and the Power Divergence statistic with λ=⅔ (see Table 1).

As with any power study suitable null and alternative hypotheses need to be defined. Goodness-of-fit tests based on multinomial data have been undertaken in a number of currently unpublished physiotherapy projects at Bond University and some of these have been interested in testing for uniformity on the 5-point Likert scale (that is p_i=0.2 for i=1, 2, 3, 4 and 5). In this study the Monte Carlo simulated power of the test statistics given in Table 1 will be simulated for a uniform null distribution against the four alternative distributions given in Table 2. In Section 2 the calculation of the simulated power is discussed and the results of the power studies for each alternative distribution are given in Section 3. Concluding comments are made in Section 4 on which are the most powerful of these five test statistics for health and medicine related Likert type data.

Table **2**. Distributions used in the power study

Distribution	Cell Probabilities				
	1	2	3	4	5
Uniform	0.20	0.20	0.20	0.20	0.20
Decreasing	0.45	0.19	0.14	0.12	0.10
Step	0.15	0.15	0.20	0.25	0.25
Triangular	0.27	0.18	0.10	0.18	0.27
Platykurtic	0.095	0.27	0.27	0.27	0.095

2. CALCULATION OF THE SIMULATED POWER

For consistency this power study uses similar null and alternative distributions and sample size to those used by Steele and Chaseling (2006) and Steele et al. (2008). The major difference being that the number of cells is five as is common in Likert type data. The sample sizes are 10, 20, 30, 50, 100 and 200, representing 2, 4, 6, 10, 20 and 40 observations per cell under the uniform null distribution.

The power of each test statistic is estimated from 10000 simulated random samples. In some cases the discrete nature of the data precludes a critical value at the nominated 5% level of significance. In such cases linear interpolation is used to enable meaningful power comparison.

3. RESULTS OF THE POWER STUDY

3.1. Decreasing Alternative Hypothesis

Figure 2 confirms the common recommendations made in the power studies identified in Section 1 that Empirical Distribution Function (EDF) test statistics such as the discrete Kolmogorov-Smirnov are generally more powerful for these trend type alternative distributions. Although all five test statistics are shown to have very high power for the larger sample sizes there are some other differences in power identified. It is clear for this alternative distribution that the Freeman-Tukey test has relatively lower power than the other four test statistics and that the Log-Likelihood Ration test has slightly lower power than Pearson's Chi-Square and the Power Divergence test with $\lambda=\frac{2}{3}$. Based on these results for a 5-point Likert scale it appears that the discrete Kolmogorov-Smirnov is more powerful at identifying the trend alternative should one exist. Should a Chi-Square type test statistic be desired, then either the Power Divergence with $\lambda=\frac{2}{3}$ or Pearson's Chi-Square should be used with higher power.

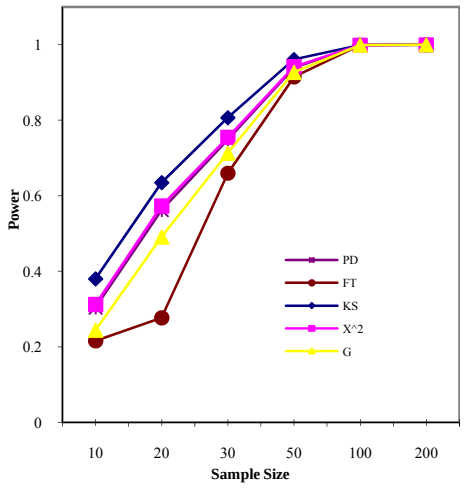

Figure **2**. Simulated power for a uniform null against a decreasing alternative distribution

3.2. Step alternative hypothesis

As was shown to occur with the decreasing alterative in Section 3.1 the power of the Kolmogorov-Smirnov test statistic is shown in Figure 3 to be greater than the power of the other four test statistics for this step type increasing alternative distribution. Although the powers are calculated for a 5-point Likert scale it again agrees with the general comments about EDF test statistics when testing a uniform null against a trend type alternative distribution. It is interesting to note that all of the Chi-Square type test statistics produce approximately the same power. Such results were not observed in the more comprehensive power study of Chi-Square type test statistics with 10 cells by Steele and Chaseling (2007). It is important to note that even for the sample size of 50, that is 10 observations per cell under the uniform null distribution, the power of all five test statistics is below 0.35 which indicates that a very large sample size is required to correctly reject this uniform null in favor of this step type alternative when using a Likert scale.

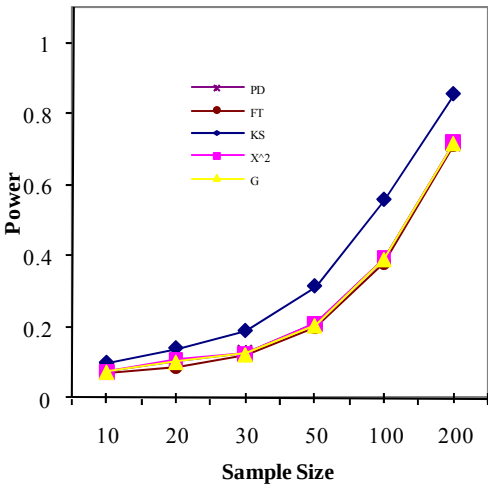

Figure **3**. Simulated power for a uniform null against a step alternative distribution

3.3. Triangular alternative hypothesis

For this particular triangular alternative distribution Figure 4 shows that the powers of all the Chi-Square type test statistics are generally much higher than the discrete Kolmogorov-Smirnov test statistic. As was the case with the step type alternative in Section 3.2 the power of all

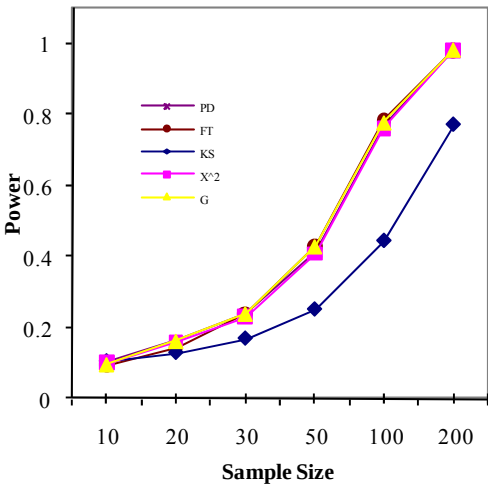

Figure **4**. Simulated power for a uniform null against a triangular alternative distribution

the test statistics were very low even for sample sizes as high as 10 per cell under the uniform null distribution. Clearly a very large sample size is required to increase to power of these test statistics for a triangular alternative distribution when the data are collected on a 5-point Likert scale

3.4. Platykurtic alternative hypothesis

Figure 5 shows that the powers of all four Chi-Square type test statistics are relatively similar, and quite low for sample sizes up to and including 6 per cell under the uniform null distribution. The power of the discrete Kolmogorov-Smirnov only becomes competitive for the very large sample size of 40 per cell under the uniform null distribution.

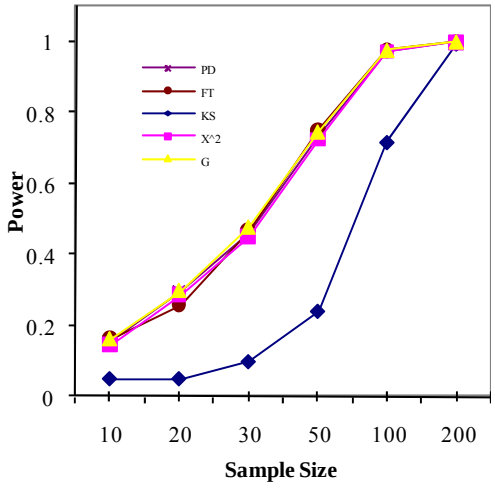

Figure 5. Simulated power for a uniform null against a platykurtic alternative distribution

4. CONCLUSION AND RECOMMENDATIONS

As shown by Steele and Chaseling (2006) and Steele et al. (2008), albeit for a larger number of cells, it is difficult to make general recommendations as to the most powerful goodness-of-fit test statistic for the specific alternative distributions used in this study. Given the widespread use in the health and medical survey research community of Pearson's Chi-Square for Likert type data some comments relating to power are needed:

- For sample sizes less than six per cell under the uniform null distribution, the simulated powers of all four test statistics were very poor for all alternative distributions with the exception of the decreasing trend distribution.
- The simulated power of the Freeman-Tukey test statistic is generally shown to be relatively less than the power of all the other investigated test statistics.
- There is generally no improvement in the simulated power for the Power Divergence test statistic with $\lambda=⅔$ over either Pearson's Chi-Square or the Log-Likelihood Ratio test statistics.

REFERENCES

Ampadu, C. (2008), On the powers of some new Chi-Square type statistics. *Far East Journal of Theoretical Statistics*, 26, 59-72.

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

Cressie, N., and Read, T.R.C. (1984), Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B*, 46, 440-464.

Freeman, M.F., and Tukey, J.W. (1950), Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21, 607-611.

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, 3167-3183.

Likert, R. (1932), A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.

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*, 157-175.

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

Steele, M., and Chaseling, J. (2006), 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, 1067-1075.

Steele, M., Hurst, C., and Chaseling, J. (2006), The power of Chi-Square type goodness-of-fit test statistics. *Far East Journal of Theoretical Statistics*, 26, 109-119.

Wilks, S.S. (1935), The likelihood test of independence in contingency tables. *Annals of Mathematical Statistics*, 6, 190-196.