Journal of Statistical Research *ISSN 0256 - 422 X*
2007, Vol. 41, No. 2, pp. 73-81
Bangladesh

A COMPARATIVE STUDY OF THE ACCURACY OF THE CHI-SQUARED APPROXIMATION FOR THE POWER-DIVERGENCE STATISTIC AND PEARSON'S CHI-SQUARED STATISTIC IN SPARSE CONTINGENCY TABLES

A. Bere
Department of Mathematics and Statistics, Polytechnic of Namibia, Windhoek, Namibia
Email: abere@polytechnic.edu.na

C. Chimedza
Statistics Department, University of Zimbabwe, Harare, Zimbabwe
Email: cchimedza@science.uz.ac.zw

SUMMARY

The Power divergence family of statistics was introduced by Cressie and Read in 1984. The likelihood ratio statistic and the Pearson's chi-squared statistic are examples of the many members of the power divergency family which are linked through a family parameter λ. We present here the results of a comparative simulation study on the accuracy of the chi-squared approximation for two members of the family ($\lambda = 1$ and $\lambda = \frac{2}{3}$) when they are used for goodness-of-fit testing in sparse contingency tables.

Keywords and phrases: Power-divergence statistic,sparse contingency table, goodness-of-fit

AMS Classification:

1 Introduction

Oftentimes when we analyze contingency tables, the sample size is not much larger than the number of cells in the contingency table. This is either because the sample size itself is small or the number of categories classifying the table is too large. The result is what is called a sparse contingency table-one with most of the cells having zero frequencies. In common statistical practice, we regard a table to be sparse if at least twenty percent of the cells have expected frequencies less than 5 [5].

The analysis of sparse tables leads to two types of problems. The first class of problems is associated with goodness-of-fit testing since the asymptotic approximations of the standard chi-squared statistics tend to be poor for sparse tables [5].

Another class of problems is related to the non-existence of the maximum likelihood estimates. Parameter estimates sometimes take on values of plus or minus infinity. In such cases algorithms like Newton-Raphson may fail to converge [5].

A common way around the two problems above is to collapse the categories until expected frequencies are large enough. We obviously loose some information when we do this.

The latter problem is also handled by adding a small constant, say 0.5 to every cell of the table prior to analysis [5]. Bayesian approaches that prevent the estimation problems mentioned above have also been suggested [5].

In relation to the first problem, the use of exact tests has been recommended [1]. Cressie and Read (1984) also introduced the idea of the Power-divergence family of Statistics which links many goodness-of-fit statistics through a single family parameter λ. As a byproduct of their work, a new Goodness-of-fit statistic (when $\lambda = \frac{2}{3}$) emerges that has valuable properties. The most important property is that for the equiprobable hypothesis, "the critical value of this statistic is well approximated by the chi-squared critical value under certain conditions" [7].

Cressie and Read (1988) also postulate that "The accuracy of the chi-squared critical value for the power-divergence statistic based on $\lambda = \frac{2}{3}$... **appears** to carry over to hypothesis with unequal cell probabilities and estimated parameters provided $\min_{1\leq i\leq k} n\pi_i \geq 1$.

This study presents the results of a Monte-Carlo study on the accuracy of the chi-squared approximation for the Cressie and Read Statistic when used as a goodness-of fit statistic in sparse contingency tables. Specifically we are testing the hypothesis of independence under various levels of sparseness. We also compare the Cressie and Read Statistic to the traditional Pearson's Chi-squared statistic.

2 The Power–Divergence Family of Statistics

The power-divergence family of statistics is related through a parameter $\lambda \in \mathbb{R}$. Each member $PDS(\lambda)$ is a sum over all cells of "deviations" between expected and observed counts. The deviation of a single cell a_λ is the scaled distance between the ratio of observed counts to expected counts raised to the power λ and the unit 1.

$$a_\lambda = \frac{2.observed}{\lambda(\lambda+1)}\left[\left(\frac{observed}{expected}\right)^{\lambda} - 1\right]$$

Thus using the usual notation for observed and expected cell frequencies, the formula for the power-divergence statistic is

$$PDS(\lambda) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} y_i\left[\left(\frac{y_i}{n\hat{\theta}_i}\right)^{\lambda} - 1\right], \quad -\infty < \lambda < \infty$$

It can be shown [4] that

$$PDS(1) = \sum_{i=1}^{k} \frac{\left(y_i - n\hat{\theta}_i\right)^2}{n\hat{\theta}_i},$$

which is Pearson's chi-squared statistic. Here y_i is the observed frequency for the i^{th} cell, k is the number of cells and $n = \sum y_i$ is the sample size.

3 Asymptotic Distributional Results

3.1 Introduction

In this section we explain how to calculate an $\alpha\%$ critical value for the power-divergency family of statistics. First we consider the simple hypothesis where there are no parameters to be estimated, then we also give results for the case where there are some parameters to be estimated.

3.2 Simple Hypothesis

We first consider the simple null hypothesis over k cells:

$$H_0 : \Theta = \Theta_0,$$

where $\Theta_0^T = (\theta_{01}, \ldots, \theta_{0k})$ is completely specified and each $\theta_{0i} > 0$. It can be shown [3] that in this case $Pr(PDS(\lambda) \geq \chi^2_{k-1}(\alpha)) \to \alpha$, as $n \to \infty$ for each $\lambda \in (-\infty, \infty)$ and each $\alpha \in (0, 1)$. This result generally holds when expected frequencies are large [3].

3.3 General Hypothesis and Parameter Estimation

The hypothesis are

$$H_0 : \Theta \in \Theta_0$$

Versus

$$H_1 : \Theta \notin \Theta_0$$

where Θ_0 is a set of possible values of Θ. We must estimate one value $\hat{\Theta} \in \Theta_0$ that is most consistent with observed proportions $\mathbf{y}/n$, and test H_0 by calculating $PDS(\lambda)$ for some fixed $-\infty < \lambda < \infty$.

The desired result is given in [3] as $Pr(PDS(\lambda) \geq \chi^2_{k-s-1}(\alpha)) \to \alpha$, as $n \to \infty$ for each $\lambda \in (-\infty, \infty)$. Here s is the number of parameters to be estimated. Again this result is generally known to be true when expected frequencies are large[3].

For an $r \times r$ contingency table, where we are fitting the independence model, the number of parameters to be estimated will be $2(r-1)$ and the total number of cells will be r^2. It follows that

$$\begin{aligned} k-s-1 &= r^2-2(r-1)-1 \\ &= (r-1)^2 \end{aligned}$$

4 Simulation of Square Contingency Tables

We consider only square $r \times r$ contingency tables and following the scheme of Koehler(1986) we consider three cases:

1. Case 1 All marginal probabilities are equal; that is

$$p_{i.} = p_{.i} = \frac{1}{r},\ i = 1, 2, \ldots, r.$$

2. Case 2

$$p_{i.} = p_{.i} = \frac{1}{r}\left[0.1 + 0.9\sum_{j=1}^{r} j^{-1}\right],\ i = 1, 2, \ldots, r.$$

3. Case 3

$$p_{i.} = p_{.i} = \begin{cases} \frac{1.85}{r}, & \text{i=1,2,...,0.5r;} \\ \frac{0.15}{r}, & \text{i=0.5r+1,...,r .} \end{cases}$$

The first case produces tables with equal expected frequencies, while cases two and three give tables with a mixture of large and small expected frequencies. Case 3 gives the most sparse tables. It must be noted that the case 3 tables used in this study are different from those used by Koehler (1986). This deviation was necessitated by the need to reduce computer runtime which was in some cases stretching to several days for Koehler type case three tables.

The probability for each cell of the two-dimensional contingency table was then calculated assuming independence between the rows and columns. The probabilities $(p_1, p_2, \ldots, p_{r\times r})$ were then summed up to give cumulative probability boundaries $(0, q_1, q_2, \ldots, 1) \equiv (0, p_1 + p_2, \ldots, 1)$ with each pair of adjacent values bordering a cell of a two-dimensional contingency table.

To get a two-dimensional table with a sample of n observed frequencies, n uniform random numbers were generated using R commands on the unit interval [0,1]. A uniform number falling in the interval $[q_i, q_{i+1}]$ would result in the number of observations for the corresponding cell in the contingency table increasing by one. Thus the number of observations $n(i, j)$ for each cell was made to be proportional to the corresponding probability for each cell under the model of independence.

Again, following the scheme of Koehler (1986) we chose sample sizes as multiples of the number of categories in each table. Thus for 6×6 tables, sample sizes of 18, 72 and 180

were used. For 10×10 tables, sample sizes used were 50, 200 and 500. The 20×20 table was only considered with a sample size of 200. Detailed R-algorithms for generating tables are given in [2]

For each sample size, 1000 tables were generated for which the rows and columns were independent. Tables for which some of the marginal totals were equal to zero were discarded and in place of them, new tables were generated. This explains why it was taking long to generate and analyze Koehler type case three tables which are very sparse. Members of the Power-divergence family were then used to test for the (correct) hypothesis of independence. Any rejection of the null hypothesis would thus be a type 1 error.

5 Results and Discussion

The tables, as pointed earlier give the number of type 1 errors committed out of a 1000 tests of hypothesis performed at each level of significance. If the chi-squared approximation is correct, we would thus expect about 200 rejections at 20% level of significance, 100 rejections at 10%, 50 rejections at 5% and 10 rejections at 1% level of significance. Fewer rejections indicate that the statistic used takes on smaller values than chi-squared while a large number of rejections point towards a stochastically larger statistic. The last 4 rows of the result tables give the approximate distribution of expected frequencies.

5.1 Case 1 Results

In this case the tables were generated in such a way that all expected frequencies would be equal. The results on the number of type I errors out of a thousand tests of hypothesis are given in table 1. The figures in brackets indicate the percentage absolute deviation of $PDS(\frac{2}{3})$ from $PDS(1)$ over $PDS(1)$.

The table shows that the chi-squared approximation for both the Pearson's chi-squared Statistic and the statistic $PDS(\frac{2}{3})$ is quite accurate in all cases where expected frequencies are at least one (6×6 table with sample size 72, 6×6 table with $n = 180$, 10×10 table with sample sizes of 200 and 500). The result for $PDS(\frac{2}{3})$ is consistent with the aforementioned postulation of Cressie and Read (1988).

For tables with expected frequencies between 0.25 and 1 ($r = 6, n = 18$ and $r = 10, n = 50$), the statistic $PDS(\frac{2}{3})$ is stochastically smaller than a chi-squared random variable with $(r-1)^2$ degrees of freedom. A comparison of the performance of the statistic $PDS(\frac{2}{3})$ and that of the Pearson's chi-squared statistic in the same tables reveals that the Pearson is less sensitive to small expected frequencies as compared to the earlier statistic. This is also reflected in the percentage deviations. The result for the Pearson is in agreement with the findings of Koehler (1986, p489) who concludes that "the chi-squared approximation for the Pearson statistic is quite accurate for case 1,...).

Table 1: Case 1 results-The number of type I errors out of a thousand tests of the independence model.

	Level of Significance	r=6 n=18	r=6 n=72	r=6 n=180	r=10 n=50	r=10 n=200	r=10 n=500	r=20 n=200
χ^2 approx	0.2	179	205	194	195	217	198	207
for $PDS(1)$	0.1	72	100	99	92	112	83	91
With	0.05	33	56	49	38	55	40	41
$(r-1)^2$ d.f	0.01	4	14	7	5	9	15	9
χ^2 approx	0.2	77(57)	212(3)	206(6)	42(78)	225(4)	201(2)	9(96)
for $PDS(\frac{2}{3})$	0.1	16(78)	107(7)	104(5)	12(87)	115(3)	86(4)	1(98)
with	0.05	5(85)	55(2)	50(2)	4(89)	54(2)	42(5)	0(100)
$(r-1)^2$ d.f	0.01	0(100)	13(8)	9(29)	0(100)	8(11)	15(0)	0(100)
number	[0,0.25)	0	0	0	0	0	0	0
of expected	[0.25,1)	36	0	0	100	0	0	100
frequencies in	[1,5)	0	36	0	0	100	0	0
each interval	≥ 5	0	0	36	0	0	100	0

Table 2: The number of type I errors out of a thousand tests of the independence model.

	Level of Significance	r=6 n=18	r=6 n=72	r=6 n=180	r=10 n=50	r=10 n=200	r=10 n=500
χ^2 approx	0.2	262	181	190	248	209	212
for $PDS(1)$	0.1	138	81	90	160	117	108
With	0.05	70	48	55	104	77	61
$(r-1)^2$ d.f	0.01	12	16	21	42	21	16
χ^2 approx	0.2	77	139	189	25	111	181
for $PDS(\frac{2}{3})$	0.1	12	49	83	2	43	86
with	0.05	5	24	44	0	22	44
$(r-1)^2$ d.f	0.01	0	7	13	0	1	6
number	[0,0.25)	15	3	0	45	10	1
of expected	[0.25,1)	16	12	6	43	35	18
frequencies in	[1,5)	5	17	20	12	45	51
each interval	≥ 5	0	4	10	0	10	30

5.2 Case 2 Results

In this case, the tables are more sparse than those for the previous case as reflected by the distribution of expected frequencies. Also, case 2 tables have a mixture of large and small expected frequencies. The results are given in table 2.

The chi-squared approximation is not very good for both statistics in this case. There is however a slight improvement for both statistics in cases where the majority of expected frequencies are at least one ($r = 6, n = 180$ and $r = 10, n = 500$). In these cases, the chi-squared approximation for $PDS(\frac{2}{3})$ looks more accurate in the lower tail than that for the Pearson statistic.

5.3 Case 3 Results

In this case the tables also have a mixture of small and large expected frequencies and there are more sparse than those for the previous case.

Table 3: The number of type I errors out of a thousand tests of the independence model.

	Level of Significance	r=6 n=18	r=6 n=72	r=6 n=180	r=10 n=50	r=10 n=200	r=10 n=500	r=20 n=200
χ^2 approx	0.2	320	230	231	285	276	253	302
for $PDS(1)$	0.1	159	182	142	228	208	168	254
With	0.05	74	49	104	169	162	129	217
$(r-1)^2$ d.f	0.01	27	108	56	94	107	63	173
χ^2 approx	0.2	66	131	144	23	81	121	20
for $PDS(\frac{2}{3})$	0.1	19	72	69	3	38	57	0
with	0.05	4	41	34	2	22	26	0
$(r-1)^2$ d.f	0.01	0	21	14	0	5	10	0
number	[0,0.25)	16	12	9	58	25	25	250
of expected	[0.25,1)	36	0	0	100	0	0	100
frequencies in	[1,5)	0	36	0	0	100	0	0
each interval	≥ 5	0	0	36	0	0	100	0

In this case, the chi-squared approximation for both $PDS(1)$ and $PDS(\frac{2}{3})$ is very poor. Whereas the Pearson's chi-squared resulted too many rejections, the statistic $PDS(\frac{2}{3})$ gives fewer rejections than expected.

6 Conclusions

When expected frequencies are equal and at least one, the chi-square approximation for both the Pearson's chi-squared and $PDS(\frac{2}{3})$ is fairly accurate.

When expected frequencies are equal and between 0.25 and 1, the statistic $PDS(\frac{2}{3})$ is stochastically smaller than a $\chi^2(r-1)$ random variable. The Pearson's chi-squared statistic fairly well approximates a $\chi^2(r-1)$ random variable in this case.

It is evident from this study that for sparse contingency tables, the chi-squared approximation for the statistic $PDS(\frac{2}{3})$ is not a better alternative to that of the Pearson Statistic.

7 Acknowledgments

The authors are highly indebted to Yonas Tesfazghi, for helping with the computer algorithms.

References

[1] Agresti A., Wackerly, D. and Boyet J.M. (1979). *Exact tests for cross-classifications :Approximation of attained significance levels*, Psychometrika.,**44, No 1**, 75-83.

[2] Bere A, A Monte Carlo study of the accuracy of Pearson's chi-squared statistic and the power divergence statistic (parameter= $\frac{2}{3}$) when used as goodness-of-fit statistics in sparse contingency tables (*Unpublished M.Sc, thesis*), University of Zimbabwe, 2002.

[3] Bishop Y.M, Fienberg S.E, and Holland P. W. (1975). Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, Mass.

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

[5] Galindo-Garre F, Vermunt J.k, Ato-Garcia M, Bayesian Approaches to the problem of sparse tables in Loglinear Modelling, http://spitswww.uvt.nl/∼vermunt/clm2000c.pdf.

[6] Koehler K. J. (1986). Goodness-of-fit Tests for Log-linear Models in Sparse Contingency Tables.*Journal of the American Statistical Association,* **81, No 394***, 483-493.*

[7] Read T. R. C. and Cressie, N. A. C. (1988). Goodness-of-fit Statistics for Discrete Multivariate data. New York:Springer.