SMALL SAMPLE VARIANCES FOR THE U - TERMS IN LOGLINEAR MODELS

I. Elaine Allen, University of Pennsylvania

1. Introduction

The analysis of survey data using loglinear models is now in widespread usage by statisticians, sociologists and demographers. In loglinear analyses the examination of the standardized parameters (or u-terms) of these models involves the estimation of the variances of these parameters. Lee (1977) compared his derivation of closed form δ-method estimates with other asymptotic methods and found them to give smaller variance estimates using a well-known, well-analyzed data set. These δ-method variances, however, being asymptotic approximations, may not give reasonable variance estimates with survey data. Because of the nature of real data, the sample surveys used for loglinear models may involve extremely small cell sizes, triangular of band matrices, or large differences between the counts in the respective cells of the table.

This paper develops more precise variance approximations by adding successive terms of the Taylor Series. These variances are then compared with the variances derived by the δ-method with respect to the special problems of survey data listed above.

2. Two-way Tables

The two-way table under a Poisson sampling scheme is examined because of its simplicity. For a two-way table the model of interest is the saturated model

$$\log m_{ij} = u + u_{1(i)} - u_{2(j)} + u_{12(ij)}) \quad (1)$$

where the u_{ψ} terms are the additive parameters of the model. First, look at the two by two table where

$$u = 1/4(\log x_{11} + \log x_{12} + \log x_{21} + \log x_{22}),$$

$$u_{1(1)} = 1/4(\log x_{11} + \log x_{12} - \log x_{21} - \log x_{22}),$$

$$u_{12(1)} = 1/4(\log x_{11} - \log x_{12} + \log x_{21} - \log x_{22}), \quad (2)$$

where x_{ij} is the observed cell frequency. Then, under Poisson sampling,

$$E(u) = 1/4(\sum_{i=1}^{2}\sum_{j=1}^{2} \log x_{ij})$$

$$1/4 \sum_{i=1}^{2}\sum_{j=1}^{2} E \log x_{ij},$$

$$E(u_{1(1)}) = 1/4[\sum_{j=1}^{2} E \log x_{ij} - \sum_{j=1}^{2} \log x_{2j}], \text{ and} \quad (3)$$

$$E(u_{12(11)}) = 1/4[E \log x_{11} + E \log x_{22} - E \log x_{12} - E \log x_{21}].$$

The other u-terms for this model are derived in a similar fashion. In order to examine the E log x_{ij}, we use the Taylor Series expansion for log x_{ij} about m_{ij}, the expected value of x_{ij}. The first five term of the Taylor Series expansion of log x_{ij} about m_{ij} are:

$$\log x_{ij} \log m_{ij} + (x_{ij}-m_{ij})'(m_{ij}) + \frac{(x_{ij}-m_{ij})^2 f''(m_{ij})}{2!} + \frac{(x_{ij}-m_{ij})^3 f'''(m_{ij})}{} + \frac{(x_{ij}-m_{ij})4f'v(m_{ij})}{} \quad (4)$$

$$+R;$$

where R is a remainder term. Substituting the values of the derivatives assuming a Poisson sampling scheme gives:

$$\log x_{ij} = \log m_{ij} + \frac{(x_{ij}-m_{ij})}{m_{ij}} - \frac{(x_{ij}-m_{ij})^2}{2m_{ij}^2} + \frac{(x_{ij}-m_{ij})^3}{3m_{ij}^3} \quad \frac{(x_{ij}-m_{ij})^4}{8m_{ij}^5} + R. \quad (5)$$

In order to evaluate the variance of any particular u-term $E[u - E\,u]^2$ is needed. Under Poisson sampling the cell means are independently distributed so Var u is simply the sum of the variances of the log cell means

$$\text{Var } u_{1(1)} = \text{Var}[1/4(\sum_{j=1}^{2} \log s_{1j} - \sum_{j=1}^{2} \log s_{2j})] = 1/16\,(\text{Var log } x_{11} + \text{Var log } x_{12} + \text{Var log } x_{21} + \text{Var log } x_{22}). \quad (6)$$

For each cell, the term $E(\log x_{ij} - E \log x_{ij})^2$ is needed. First calculating E log x_{ij}.

$$E \log x_{ij} = E(\log m_{ij} + \frac{(x_{ij}-m_{ij})}{m_{ij}} - \frac{(x_{ij}-m_{ij})^2}{2m_{ij}^2} - \frac{(x_{ij}-m_{ij})^4}{8m_{ij}^5} + R) \quad (7)$$

$$= \log m_{ij} + \frac{\text{Var } x_{ij}}{2m_{ij}^2} + \frac{E(x_{ij}-m_{ij})^3}{3m_{ij}^3}$$

$$- \frac{E(x_{ij}-m_{ij})^4}{8m_{ij}^5} + E(R).$$

Then, subtracting E log x_{ij} from log x_{ij} term by term:

$$\log x_{ij} - E \log x_{ij} = \frac{x_{ij}-m_{ij}}{m_{ij}} - \frac{((x_{ij}-m_{ij})^2 - E(x_{ij}-m_{ij})^2)}{2m_{ij}^2} + \frac{((x_{ij}-m_{ij})^3 - E(x_{ij}-m_{ij})^3)}{3m_{ij}^3} - \frac{((x_{ij}-m_{ij})^4 - E(x_{ij}-m_{ij}))}{8m_{ij}^5} + (R-E(R)). \quad (8)$$

To calculate the variance of log x_{ij} this result is squared and its expected values is taken. Dropping the subscript notations yeilds:

$$\text{Var log } x = \frac{\text{Var } x}{m^2} - \frac{E(x-m)^3}{m^3} - \frac{(3\text{Var}(x-m)^2 - 4E(x-m)^4)}{12m^4} - \frac{(E(x-m)^5 - \text{Var } x\, E(x-m)^3)}{6m^5} - \frac{(9E(x-m)^5 - 8\text{Var}(x-m)^3)}{72m^6} - \frac{(\text{Var } x\, E(x-m)^4 - E(x-m)^6)}{16m^7} - \frac{(E(x-m)^7 - E(x-m)^3 E(x-m)^4)}{24m^8}, \quad (9)$$

after collecting terms.

The first term is equivalent to the variances Lee obtained using the δ-method.

$$\frac{\text{Var } x}{x^2} = \text{Var } x\, [f''(x)]^2 = \sigma^2(\theta)\, [f'(\theta)\, [f'(\theta)]^2, \quad (10)$$

With large enough sample size the remaining terms go to zero. For example, with minimum expected cell counts ranging from 5 down to 1, the values of the denominators of the remaining terms are already very large and are approaching zero very quickly. Even with expected cell counts as low as 2, the terms beyond those involving m^3 are negligible and the remainder term is zero.

The variance of u_{ij} for the two by two table is the sum of the variances of the four log x_{ij}'s. If only one of the cells has an expected value of five or less, we are interested in how this effects the variance of the u_{ij}'s, and if the δ-method approximation is close to the more precise Taylor Series approximations.

To examine this question, five two by two tables are constructed, identical with the exception of one cell. One cell has an expected values, under the saturated model, which varies from five to one. The other cells are all greater than five.

In order to calculate the variances of $u_{1(1)}$ the moments of the Poisson distribution are substituted yielding:

$$\text{Var log } x = \frac{1}{m} - \frac{1}{m} + \frac{(1+10m)}{12m^3} - \frac{(1+9m)}{6m^4} - \frac{(1+10m)}{72m^5} - \frac{(25m+12m^2)}{16m^6} - \frac{(1-57m-108m^2)}{24m^7}. \quad (11)$$

and the corresponding δ-method variance estimate is:

$$AV\ (u_{1(1)} || C_{12}; 2) = \frac{1}{4}\left(\frac{1}{m_{11}} + \frac{1}{m_{12}} + \frac{1}{m_{21}} + \frac{1}{m_{22}}\right), \quad (12)$$

using the notation of Lee (1977).

The values of the variance of $u_{1(1)}$ using the Taylor Series and δ-method are given in Table 1. As expected, when the size of the smallest cell decreases, the difference between the δ-method and the more exact variance increases. If a model other than the saturated model is fit, the expected cell values will be greater than one. The smalllest expected cell count is 3 in the table with the observed cell count 1.

For two-way tables, then, the size of the smallest cell in the table has an effect on the variance calculated. As the number of cells increase from four (two by two table) to a larger r x c table, the effect of the smallest cell diminishes. This can be seen directly from the number of terms in the variance calculation.

There will always be the same number of var log x_{ij} terms as there are cells in the table of the fitted model. In the saturated model this is equal to the number of cells in the original table but, for example, in the model of independence the number of terms is equal to the number of marginal (x_{it} and x_{tj}) terms. So the effect of a small cell on the variance of u changes with the size of the table, which then implies that the effect changes with the model is less than in an unsaturated model.

The size of a small cell need not be less than or equal to some fixed number. If all the cells but one are of one order of magnitude, and the last cell is of an order of magnitude smaller, the δ-method variance will be affected. An illustration of this is given in the next section.

3. Higher than Two-Way Tables

For tables of greater than two dimensions, under Poisson sampling, the variances derived from the Taylor Series expansion will extend directly. Since the cells are independent under Poisson sampling, the variance of $u_{1(1)}$ is again the sum of the variances of the respective log x_{ijk}'s and log x_{1jk}'s.

$$\mathrm{Var}(u_{1(1)}) = \mathrm{Var}[\tfrac{2}{7} \sum_j \sum_k \log x_{ijk} + (\tfrac{1}{27})^2 \sum_{\alpha \neq i} \sum_j \sum_k \mathrm{Var} \log x_{\alpha jk}. \quad (13)$$

Unlike the 2 x 2 table, which reduces to the sum of the reciprocals of expected values, here we have different multipliers for i and $\alpha \neq i$. The Taylor Series expansion for log x_{ijk}, however, will be exactly the same as that for log x_{ij}. However, instead of four terms for each variance, there are now 27. This will reduce the effect of a cell with unit expected value, but only if the other cells are relatively small. If the other cells are large, their reciprocals will be much smaller than one. For example, if we were to fit a table where all the cell values were 10 except for one 1, the variance of $u_{1(1)}$ using the δ-method is .00493. If the table were filled completely with 10's, the corresponding variance is .00412. The difference between these two is not great enough to cause alarm. But if, however, the table is filled with 100's and one 1 the variance is .00127 as compared with a table of 100's where the variance is .00034. This difference is now almost an order of magnitude. From this it is concluded that, even for higher-way tables, the presence of an expected cell value of unity will distort the δ-method variance if the other cells in the table are uniformly much larger than 1.

With high dimensional tables it is very likely that the sampling scheme is not independent Poisson but multinomial instead. Deriving approximate small sample variances using a Taylor Series expansion is more complicated under multinomial sampling because the cells are no longer independent. If a variable has r levels, (r-1) of these levels are independent but the r-th is not. The covariance terms are included for this level and the other (r-1) levels. This variance is calculated for $u_{1(i)}$ of the saturated model of a three-way table:

$$\begin{aligned}\mathrm{Var}(u_{1(i)}) &= \mathrm{Var}(\tfrac{I-1}{IJK} \sum_j \sum_k \log x_{ijk}) \\ &+ \mathrm{Var}(\tfrac{1}{IJK} \sum_{\alpha \neq i} \sum_j \sum_k \log x_{\alpha kj} \\ &+ 2\,\mathrm{Cov}(\tfrac{I-1}{IJK} \sum_J \sum_K \log x_{ijk}, \tfrac{1}{IJK} \sum_{\alpha \neq i} \sum_j \sum_k \log x_{\alpha jk}).\end{aligned} \quad (14)$$

Expanding these and taking appropriate expectations, the result is:

$$\begin{aligned}\mathrm{Var}(u_{1(i)}) &= (\tfrac{I-1}{IJK})^2 \sum_j \sum_k \mathrm{Var} \log x_{ijk} \\ &+ (\tfrac{1}{IJK})^2 \sum_{\alpha \neq i} \sum_j \sum_k \mathrm{Var} \log x_{ijk} \\ &+ 2[\tfrac{I-1}{IJK})^2 \sum_j \sum_k (E(\sum_{\alpha \neq i} \log x_{\alpha jk})(\log x_{ijk}) \\ &- (\sum_{\alpha \neq i} E \log x_{jk})(E \log x_{ijk}))].\end{aligned} \quad (15)$$

The first two terms of this expression reduce to the Taylor Series expansion given earlier with different multipliers for each term. The last term reduces to some extent because only the i-th level of variable one is dependent on the other (i-1) levels. Taking the parts of this term separately, the remaining terms are:

$$E(\sum_{\alpha \neq j} \log x_{ajk})(\log x_{ijk}) \quad (16)$$

$$\begin{aligned}&= E[\log x_{1jk} \log x_{ijk} + \log x_{2jk} \log x_{ijk} + \ldots \\ &+ \log x_{i-1)jk} \log x_{ijk}] = E(\log x_{1jk} \log x_{ijk}) \\ &+ E(\log x_{ijk} \log x_{ijk}) + \ldots + E(\log x_{(i-1)jk} \log x_{ijk}) \\ &E(\sum_{\alpha \neq i} \log x_{jk})(\log x_{ijk}) = E \log x_{1jk} \log x_{ijk} \\ &(E \log x_{ijk}) + (E \log x_{2jk})(E \log x_{ijk}) \\ &+ \ldots + (E \log x_{(i-1)jk})(E \log x_{ijk}).\end{aligned} \quad (17)$$

Each of these terms can be expanded using the

Taylor Series as before.

Comparing these variances to the δ-method variances it is clear that the first two terms again reduce to δ-method variances. The δ-method does not consider the covariance terms, assuming its application to only (r-1) of the r levels of a variable. If the cell sizes are not too small the contribution from the convariance term is extremely small. This can be shown by example. In an extreme case, if the three smallest cells are 1, 2, and 3, and all the other cells of a 3 x 3 x 3 table are equal to 10, the covariance contribution to the entire variance is .00041 for an entire variance of .1649. If the smallest cell size was 2, the contribution would be less than half that much, .00019, for a total variance of .14587.

4. Simulation

To better asses the effect of small sample sizes and small cell sizes on estimated asymptotic variances of u-terms using the δ-method, a Monte Carlo simulation was performed. A simulation allows for a systematic assesment of the δ-method variances because certain factors in the table can be varied and evaluated individually.

a. Design

Four factors of multidimensional contingency tables are controlled: (a) Total sample size; (b) The number of dimensions of the table; (c) The number of margins varied in each table; (d) The probability configuration of the margins. For all tables, the number of levels of each variable is fixed at three.

b. Data Generation

Fortran programs, written by the author, were used to generate the tables and calculate the variance estimates. An IMSL (1976) supplied uniform random number generator (GGUB), whose properties are well-known, was used to generate cell margins fixed as above. All data generation and analysis were done on an IBM 370/168 under VM 340 OS/MVT Release 216 under HASP and CMS. The table were produced in machine readable form producing variance estimates in hard copy and machine readable form.

c. Results

Because of limited space, only the conclusions drawn from the simulation are given. A full description of the results and the computer programs used may be obtained from the author.

From the simulation it is concluded that the δ-method estimates can be used in some small sample situations. If our sample size is at least 4.5 times the number of cells in the table the δ-method estimates are close to the more exact variances. These two estimates are close even if the margins are skewed and one of the cells has expected value less than one. If the sample size is as small as 1.5 times the number of cells in the table, and a maximum of one margin is skewed, the δ-method estimates are also close to the more exact variances. For four-way tables, a maximum of two of the margins may be skewed. In all other cases of more than one skewed margin or smaller sample size, the δ-method variance estimates are no longer reliable. These results give new rules of thumb to the analyst of contingency tables. Instead of being guided only by sample size alone or cell size alone, these results show how the two interact. Also, the skewness of the margins has also been shown to be important.

5. References

Bishop, Y., S. Fienberg, and P. Holland (1975) Discrete Multivariate Analysis, M.I.T. Press, Cambridge, MA.

Haberman, S. (1974) The Analysis of Frequency Data, U. of Chicago Press, Chicago, ILL.

Haberman, S. (1977) "Loglinear Models and Frequency Tables with Small Expected Cell Counts," Ann. Statist. 5,6, 1148-1169.

Lee, S.K. (1977) "On the Asymptotic Variances of f-Terms in Loglinear Models of Multidimensional Contingency Tables," J. Amer. Statist. Assoc., 72, 358-412-419

6. Acknowledgements

This research was supported by NSF Contract OCR75-13373 awarded to Professor Ivor Francis, Department of Economics and Social Statistics, Cornell University. The author wishes to thank Dr. Francis and Dr. Thomas Santner for comments and suggestions in this research.

TABLE 1

Values of the Variance of $u_{1(1)}$ Using the Exact Method and the δ-method

	EXACT METHOD		δ-METHOD	
$E(x_{11})$	Var $(u_{1(1)})$	Var $(\log x_{11})$	AV$(u_{1(1)})$	Var$(\log x_{11})$
5	.03754	.01175	.03879	.0125
4	.03923	.01344	.4192	.01563
3	.04436	.01856	.4712	.2083
2	.05146	.02836	.05754	.03125
1	.25704	.22764	8879	625