The Korean Communications
in Statistics Vol. 6, No. 1, 1999
pp. 327-336

A Note on Estimation of Multinomial Probabilities when Some Frequency Counts are Merged

Sang Eun Lee[1] and C. J. Park[2]

Abstract

In a multinomial sampling scheme, some categories may be observed as partially classified because of technical or economic reasons. In this paper the maximum likelihood estimators(M.L.E) of multinomial probabilities are obtained when some frequencies are merged. We obtained the M.L.E, and a method to evaluate the information gained by including merged frequencies in M.L.E. we obtained an estimator of the covariance matrix, and it is used to examine the information gained by including the merged frequency counts in estimating the cell probabilities. When certain individual frequency counts are missing, a method is proposed for estimating the cell probabilities using EM algorithms.

1. Introduction

In a multinomial sampling scheme, some categories may be observed as partially classified because of technological or economic reasons. In this paper the maximum likelihood estimators of the multinomial probabilities are obtained when some frequencies are merged. The maximum likelihood estimators with incomplete multinomial data are obtained in general case in [1]. The case considered in this paper are special case of [1] and [2], however, we obtained the closed forms of the maximum likelihood estimators, and a method of obtained to evaluate the information gained by including merged frequencies in maximum likelihood estimators. More specially, consider the following multinomial sampling scheme: a random sample of size n_1 is drawn from a multinomial population with cell probability p_1, p_2, p_3, p_4 and p_5, where $p_5 = 1 - p_1 - p_2 - p_3 - p_4$; second random sample of size n_2 is drawn from multinomial population, and the merged frequency counts of cell 1 and cell 2 is observed; a third sample of size n_3 is drawn and merged frequency counts cell 1, cell 2 and cell 3 is observed; a forth sample of size n_4 is drawn and the merged frequency counts cell 1 through cell 4 are observed. The first sample of size n_1 consisting of the individual cell frequency

1) Kyonggi University, Suwon, Kyonggi-Do
2) San Diego State University, San Diego, California

counts of cell 1 through cell 4, denoted by x_{11}, x_{12}, x_{13} and x_{14} respectively; a second sample of size n_2 consisting of the merged frequency counts cell 1 and cell 2 denoted by $x_{2(12)}$ and the cell frequency counts for the third and fourth cell denoted by x_{23} and x_{24} respectively; a third sample of size n_3 consisting of the merged frequency counts of cell 1 through cell 3 denoted by $x_{3(13)}$, and the frequency counts fourth frequency counts cell x_{34} ; finally a fourth sample size of n_4 consisting of the merged frequency counts of cell 1 through cell 4 denoted by $x_{4(14)}$. The merged frequency counts $x_{2(12)}, x_{3(13)}$, and $x_{4(14)}$ are used together with the individual cell frequency counts $x_{11}, x_{12}, x_{13}, x_{14}, x_{23}, x_{24}$, and x_{34} to obtain the maximum likelihood estimators of the cell probabilities.

An estimator of the variance matrix is obtained, and it is used to examine the information gained by including the merged frequency counts in estimating the cell probabilities. When certain individual frequency counts are missing, a method is proposed for estimating the cell probabilities using E_M algorithms.

To facilite the presentation of the maximum likelihood estimators of the cell probabilities, the following notations are adopted by amalgamating the cell independent samples.

cell probabilities	Observed frequency counts
p_1	$x_1 = x_{11}$
p_2	$x_2 = x_{12}$
p_3	$x_3 = x_{13} + x_{23}$
p_4	$x_4 = x_{14} + x_{24} + x_{34}$
$p_5 = 1 - p_1 - p_2 - p_3 - p_4$	$x_5 = (n_1 + n_2 + n_3 + n_4 - x_1 - x_2 - x_3 - x_4 - y_2 - y_3 - y_4)$
$(p_1 + p_2)$	$y_2 = x_{2(12)}$
$(p_1 + p_2 + p_3)$	$y_3 = x_{3(13)}$
$(p_1 + p_2 + p_3 + p_4)$	$y_4 = x_{4(14)}$

The natural logarithm of the likelihood function of p_1, p_2, p_3 and p_4 can be written as

$$\ln(p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} p_5^{x_5} (p_1 + p_2)^{y_2} (p_1 + p_2 + p_3)^{y_3} (p_1 + p_2 + p_3 + p_4)^{y_4})$$

$$= \sum_{i=1}^{5} x_i \ln p_i + y_2 \ln(p_1 + p_2) + y_3 \ln(p_1 + p_2 + p_3) + y_4 \ln(p_1 + p_2 + p_3 + p_4)$$

Differentiating the logarithm of the likelihood function with respect to p_1, p_2, p_3 and p_4 and setting them equal to zero, the following system of four nonlinear equations are obtained :

$$\frac{x_1}{p_1}+\frac{y_2}{(p_{1+p_2})}+\frac{y_3}{(p_1+p_2+p_3)}+\frac{y_4}{(p_1+p_2+p_3+p_4)}-\frac{x_5}{p_5}=0$$

$$\frac{x_2}{p_2}+\frac{y_2}{(p_1+p_2)}+\frac{y_3}{(p_1+p_2+p_3)}+\frac{y_4}{(p_1+p_2+p_3+p_4)}-\frac{x_5}{p_5}=0$$

$$\frac{x_3}{p_3}+\frac{y_3}{p_1+p_2+p_3)}+\frac{y_4}{(p_1+p_2+p_3+p_4)}-\frac{x_5}{p_5}=0$$

$$\frac{x_4}{p_4}+\frac{y_4}{(p_1+p_2+p_3+p_4)}-\frac{x_5}{p_5}=0$$

By solving the system equations, the following closed forms of the maximum likelihood estimators are obtained

$$\hat{p}_5=1-\hat{p}_1-\hat{p}_2-\hat{p}_3-\hat{p}_4$$

$$\hat{p}_4=[x_4+\frac{x_4y_4}{(x_1+x_2+x_3+x_4+y_2+y_3)}]/n$$

$$\hat{p}_3=[\frac{x_3}{(x_1+x_2+x_3+y_2)}][\frac{(x_1+x_2+x_3+y_2+y_3)+(x_1+x_2+x_3+y_2+y_3)y_4}{(x_1+x_2+x_3+x_4+y_2+y_3)}]/n$$

$$\hat{p}_2=[\frac{x_2}{(x_1+x_2)}][\frac{(x_1+x_2+y_2)}{(x_1+x_2+x_3+y_2)}][\frac{(x_1+x_2+x_3+y_2+y_3)+(x_1+x_2+x_3+y_2+y_3)y_4}{(x_1+x_2+x_3+x_4+y_2+y_3)}]/n$$

$$\hat{p}_1=[\frac{x_1}{(x_1+x_2)}][\frac{(x_1+x_2+y_2)}{(x_1+x_2+x_3+y_2)}][\frac{(x_1+x_2+x_3+y_2+y_3)+(x_1+x_2+x_3+y_2+y_3)y_4}{(x_1+x_2+x_3+x_4+y_2+y_3)}]/n$$

where $n=n_1+n_2+n_3+n_4$

Suppose that the merged frequency counts are not used for estimating the cell probabilities, then the independent samples can be summarized as follows.

cell probabilities	Observed frequency counts
p_1	$x_1=x_{11}$
p_2	$x_2=x_{12}$
p_3	$x_3=x_{13}+x_{23}$
p_4	$x_4=x_{14}+x_{24}+x_{34}$
$p_5=1-p_1-p_2-p_3-p_4$	$x_5=(n*-x_1-x_2-x_3-x_4)$
	where $n*=n_1+n_2+n_3+n_4-y_2-y_3-y_4$

Using these frequency counts, the conditional maximum likelihood estimators of $p_i's$ are given by :

$$\hat{p}_5 = 1 - \hat{p}_1 - \hat{p}_2 - \hat{p}_3 - \hat{p}_4$$

$$\hat{p}_1 = x_1 / n*$$

$$\hat{p}_2 = x_2 / n*$$

$$\hat{p}_3 = x_3 / n*$$

$$\hat{p}_4 = x_4 / n*$$

For the case of sampling from a multinomial population with 4 cells and are considered. Suppose a random samples are drawn from a multinomial population with 4 cells, and the independent samples can be summarized as follows. Then the maximum likelihood estimators of $p_i's$ can be similarly obtained

cell probabilities	Observed frequency counts
p_1	$x_1 = x_{11}$
p_2	$x_2 = x_{12}$
p_3	$x_3 = x_{13} + x_{23}$
$p_4 = 1 - p_1 - p_2 - p_3$	$x_4 = (n_1 + n_2 + n_3 - x_1 - x_2 - x_3 - y_2 - y_3)$
$(p_1 + p_2)$	$y_2 = x_{2(12)}$
$(p_1 + p_2 + p_3)$	$y_3 = x_{3(13)}$

The maximum likelihood estimators of $p_i's$ are given by :

$$\hat{p}_4 = 1 - \hat{p}_1 - \hat{p}_2 - \hat{p}_3$$

$$\hat{p}_3 = [x_3 + \frac{x_3 y_3}{(x_1 + x_2 + x_3 + y_2 + y_3)}]/n$$

$$\hat{p}_2 = [\frac{x_2}{(x_1 + x_2)}][\frac{(x_1 + x_2 + y_2) + (x_1 + x_2 + y_2)y_3}{(x_1 + x_2 + x_3 + y_2 + y_3)}]/n$$

$$\hat{p}_1 = [\frac{x_1}{(x_1 + x_2)}][\frac{(x_1 + x_2 + y_2) + (x_1 + x_2 + y_2)y_3}{(x_1 + x_2 + x_3 + y_2 + y_3)}]/n$$

$$where \quad n = n_1 + n_2 + n_3$$

Suppose that independent samples are drawn from a trinomial population, a multinomial population with 3 cells, and the independent samples can be summarized as follows. Then the maximum likelihood estimators of $p_i's$ similarly derived

cell probabilities	Observed frequency counts
p_1	$x_1 = x_{11}$
p_2	$x_2 = x_{12}$
$p_3 = 1 - p_1 - p_2$	$x_3 = (n_1 + n_2 - x_1 - x_2 - y_2)$

The maximum likelihood estimators of $p_i's$ are given by :

$$\hat{p}_3 = 1 - \hat{p}_1 - \hat{p}_2$$

$$\hat{p}_2 = [x_2 + \frac{x_2 y_2}{(x_1 + x_2)}]/n$$

$$\hat{p}_1 = [x_1 + \frac{x_1 y_2}{(x_1 + x_2)}]/n$$

where $n = n_1 + n_2$

2. Covariance matrix

The effect of using the merged frequency counts in estimating the cell probabilities can be examined by considering the covariance matrix of the maximum likelihood estimators. More specifically, let H denote the Hessian matrix when independent individual cell frequency counts and the merged cell frequency counts are used, and H_0 denote the Hessian matrix when only the independents individual cell frequency counts are used. Suppose there is a matrix A such that $H = H_0 + A$, then it can be verified that $H^{-1} = H_0^{-1} - H_0^{-1}AH^{-1}$. Hence, it follows that if the matrix $H_0^{-1}AH^{-1}$ is a semi-positive matrix, then the variances of the maximum likelihood estimators obtained by using only the merged individual cell frequency counts. But it can be shown that if the matrix A is semi-positive then so is the matrix $H_0^{-1}AH^{-1}$. Thus the effect of using the merged cell frequency counts in addition to the individual cell frequency counts on the variance can be examined by the semi-positiveness of the matrix A.

3. Numerical Examples

It is unthinkable to derive a closed form of the covariance matrix of the estimators of $p_i's$ given by in section 1. However, the covariance matrix can be evaluated from the Hessian Matrix, which can be easily obtained by differentiating the system of equations used for the maximum likelihood estimators are summarized in Table 1. For the four independent samples

of sizes n_1, n_2, n_3, and n_4, the individual frequency counts and merged frequency counts are selected such that the estimates of p_i's are the same for the each run. The estimates of p_i's are $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.15$, $\hat{p}_3 = 0.2$, $\hat{p}_4 = 0.25$

For the run number 1, the covariance matrix is calculated.

$$H^{-1} = \begin{pmatrix} .00310662 & -.0028401 & -.0001504 & -.0000067 \\ & .0032398 & -.0002256 & -.0001016 \\ & & .0006082 & -.0001355 \\ & & & .0004258 \end{pmatrix}$$

Table 1. sample sizes and variance

Runs	n_1	n_2	n_3	n_4	$var(\hat{p}_1)$	$var(\hat{p}_2)$	$var(\hat{p}_3)$	$var(\hat{p}_4)$
1	20	200	200	200	0.00311	0.00324	0.00060	0.00042
2	20	200	200	20	0.00311	0.00325	0.00062	0.00044
3	20	200	20	200	0.00312	0.00328	0.00068	0.00073
4	20	200	20	20	0.00313	0.00329	0.00070	0.00077
5	20	20	200	200	0.00349	0.00410	0.00295	0.00073
6	20	20	200	20	0.00349	0.00411	0.00298	0.00077
7	20	20	20	200	0.00359	0.00433	0.00337	0.00278
8	20	20	20	20	0.00363	0.00422	0.00352	0.00301

From table 1, the effective sample size $\hat{\hat{n}}_i$ can be calculated by $\hat{\hat{n}}_i = \hat{p}_i(1-\hat{p}_i) \,/\, var(\hat{p}_i)$ for i= 1, 2, 3, 4 where the $var(\hat{p}_i)$ is the variance of the estimate of the $\hat{p}_i$ calculated from the Hessian matrix. The effective sample sizes for run number 1 are : $n_1 = 28.97$, $n_2 = 39.35$, $n_3 = 263.03$, $n_4 = 440.29$. The effective sample size can be interpreted as the sample size in a sample from a multinomial population when the individual cell probabilities p_1, p_2, p_3 and p_4 are $n_1 = n_2 = 20$, $n_3 = 240$, and $n_4 = 420$, respectively. It can be seen that there are gains among sample sizes when the merged cell frequency counts are used together with the individual cell frequency counts.

4. Newton-Raphson Method

When there are no closed forms of the maximum likelihood estimators of the cell probabilities or to solve the set of the derivatives of the likelihood equations, the following Newton Raphson iteration method can be used in calculating the maximum likelihood estimates. For j=1,2,3 and 4,

$$p_j^{(i+1)} = p_j^{(i)} - H^{-1}\frac{d}{dp_j}\ln(p_1^{x_1}p_2^{x_2}p_3^{x_3}p_4^{x_4}p_5^{x_5}$$
$$(p_1+p_2)^{y_2}(p_1+p_2+p_3)^{y_3}(p_1+p_2+p_3+p_4)^{y_4})$$

where the Hessian matrix H and the derivatives of the logarithm of likelihood function are evaluated with the value $p_j^{(i)}$ of the i^{th} iteration.

5. Missing observations and E-M Algorithm

When there are missing individual cell frequency counts a method is proposed for estimating the cell probabilities using the E-M Algorithm along with the maximum likelihood estimates. An example is presented when the individual frequency counts x_{13} and x_{24} are missing. For the example the following iterative steps based on the E-M Algorithm is used for obtaining the maximum likelihood estimates of the individual cell probabilities. It is simple and easy to apply and several cases have shown that one needs a few steps of iterations to obtain the estimates with a desired accuracy.

The following examples of iterative steps can be easily implemented on a computer for a quick calculation :

$$\hat{p}_4^{(i+1)} = [x_4 + \frac{x_4 y_4}{(x_1+x_2+n_1 p_3^{(i)}+n_1 p_4^{(i)}+y_2+y_3)}]/n$$

$$\hat{p}_3^{(i+1)} = [\frac{n_1\,\hat{p}_3^{(i)}}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2)}]$$
$$[\frac{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3+(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3)y_4}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+n_1\,\hat{p}_4^{(i)}+y_2+y_3)}]/n$$

$$\hat{p}_2^{(i+1)} = [\frac{x_2}{(x_1+x_2)}][\frac{(x_1+x_2+y_2)}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2)}]$$
$$[\frac{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3)+(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3)y_4}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+n_1\,\hat{p}_4^{(i)}+y_2+y_3)}]/n$$

$$\hat{p}_1^{(i+1)} = [\frac{x_1}{(x_1+x_2)}][\frac{(x_1+x_2+y_2)}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2)}]$$
$$[\frac{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3)+(x_1+x_2+n_1\,\hat{p}_3^{(i)}+y_2+y_3)y_4}{(x_1+x_2+n_1\,\hat{p}_3^{(i)}+n_1\,\hat{p}_4^{(i)}+y_2+y_3)}]/n$$

$$\text{where } x_3^{(i)} = (n_1 p_3^{(1)} + x_{23}),$$
$$x_4^{(i)} = (x_{14} + n_2 p_4^{(i)} + x_{34})$$
$$n = n_1 + n_2 + n_3 + n_4$$

when there are missing individual cell frequency counts, the Newton-Raphson iterative method can be used in calculating the individual cell probabilities, instead of using the closed form of the maximum likelihood estimators as shown in the example above. The missing value can be estimated by its estimated expected values in the iterative steps of Newton-Raphson method.

6. Simulations and Conclusion

Three sets of simulations are carried out in order to examine the performance of the method presented in this paper. The simulation results are used to calculate the means, the standard deviations, and the effective sample size, and they are summarized as follows :

The first simulation consists of ninety four replications where the values of cell probabilities are $p_1 = 0.1$, $p_2 = 0.15$ $p_3 = 0.2$ and $p_4 = 0.25$ and the sample sizes are $n_1 = 20$, $n_2 = n_3 = n_4 = 200$.

Cell probabilities	$p_1 = 0.1$	$p_2 = 0.15$	$p_3 = 0.2$	$p_4 = 0.25$
Means	0.1105	0.1426	0.2018	0.2465
Standard deviations	0.06697	0.06440	0.02578	0.02020
Effective sample sizes	21.92	29.48	242.36	455.38

The second stimulation consists of one hundred replications where the values of cell probabilities are $p_1 = 0.1$, $p_2 = 0.15$, $p_3 = 2$ and $p_4 = 0.25$, and the sample sizes are $n_1 = 20$, $n_2 = n_3 = n_4 = 200$.

Cell probabilities	$p_1 = 0.1$	$p_2 = 0.15$	$p_3 = 0.2$	$p_4 = 0.25$
Means	0.1086	0.1437	0.2018	0.2481
Standard deviations	0.06727	0.06362	0.02518	0.02131
Effective sample sizes	21.39	30.40	254.05	410.79

The third simulation consists of one hundred replications where the values of cell probabilities are $p_1 = 0.2$, $p_2 = 0.2$, $p_3 = 0.2$ and $p_4 = 0.2$, and the sample sizes are $n_1 = 20$, $n_2 = n_3 = n_4 = 200$.

Cell probabilities	$p_1 = 0.2$	$p_2 = 0.2$	$p_3 = 0.2$	$p_4 = 0.2$
Means	0.2109	0.1894	0.2006	0.2006
Standard deviations	0.07960	0.07756	0.02388	0.02024
Effective sample sizes	26.27	25.52	281.21	391.45

Even though the estimates are biased, the percentages of bias look reasonably small not to be concerned for the practical applications. The methods are simple and even the excel software can be used. In estimating the cell probabilities, there should be many situations where the methods presented in this paper can be cost effectively, specifically in obtaining the preliminary estimates.

The covariance matrix are calculated for the second simulation run to compare it with the inverse of the Hessian matrix.

The estimates of covariance matrix =

0.00452520	-0.00397107	-0.00009269	0.00031891
	0.00404805	-0.00025341	-0.00012629
		0.00063414	-0.00014143
			0.00045397

Note that the estimates of covariance matrix and the inverse of Hessian matrix given in the previous section are in fair agreement over all elements. In each of above simulations the actual sample sizes that can be used in estimating the individual cell frequency counts are :

$n_1^* = 20$, $n_2^* = 20$, $n_3^* = 220$, and $n_4^* = 440$. The effective sample sizes are calculated by $\hat{\hat{n}}_i = \hat{p}_i(1-\hat{p}_i)/Var(\hat{p}_i)$, where the mean is used for $\hat{p}_i$ and the $Var(\hat{p}_i)$ are calculated from the covariance matrix for i = 1,2,3,4. The effective sample sizes are larger than the actual sample sizes, and the method presented in the paper suggests that there will be an improvement in terms of the variance when the merged frequency counts are combined with the individual cell frequency counts of the estimators of the cell probabilities.

References

[1] Blumenthal, Saul(1968). Multinomial Sampling with Partially Categorized Data, *Journal of the American Statistical Association,* June, 63, 542-551

[2] Hocking, R. R. and Oxspring H. H(1971). Maximum Likelihood Estimation with Incomplete Multinomial Data, *Journal of the American Statistical Association,* March, 66, 65-70.

[3] Park, Taesung and Brown, Morton B(1994). Models for Categorical Data with Bonignorable nonresponse *Journal of the American Statistical Association,* March, 89, 44-53.