Journal of Statistical
Theory & Methods
1996, Vol. 7, pp 105~112

On The Generation of Multivariate Multinomial Random Numbers[1]

Daehak Kim[2]

Abstract Softwares including random number generation are abundant in modern informative society. But it's hard to get directly multivariate multinomial random numbers from existing softwares. Multivariate multinomial random numbers are greatly used in social and medical sciences. In this paper, we show that desired multivariate multinomial random numbers can be easily generated by the aids of existing random number generating software. Some characteristics of multivariate multinomial distribution are surveyd. Measures of association for the generated random numbers were computed and compared with population ones via simulation study

Keywords : multivariate multinomial distribution, random numbers, measures of association, simulation.

1. Introduction

Many researches, particularly in the social sciences, deal with populations of individual which are thought of as cross-classified by two or more polytomies. The double polytomy is the most common one due to the simplicity and easy tabulation. A double polytomy may be represented by the following Table 1 where classification A divides the population into a classes $A_1, A_2, \cdots, A_a$ and classification B divides the population into b classes $B_1, B_2, \cdots, B_b$. The population proportion that is classified as both A_i and B_j will be denoted by p_{ij} and the marginal proportions will be denoted by p_{i+} and p_{+j}, respectively.

Most of results related to cross-classified data suppose the population completely known in regard to classifications. After some measures for a known

[1] This research was supported by general grant, Catholic University of Taegu-Hyosung, 1996.

[2] Department of statistics, Catholic University of Taegu-Hyosung, Kyungsan, 712-702, Korea

population are selected, one should consider the sampling problems associated with estimation and tests about this population parameters.

Table 1. Double polytomy

$A \setminus B$	B_1	B_2	$\cdots$	B_p	Total
A_1	p_{11}	p_{12}	$\cdots$	p_{1b}	p_{1+}
A_2	p_{21}	p_{22}	$\cdots$	p_{2b}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_a	p_{a1}	p_{a2}	$\cdots$	p_{ab}	p_{a+}
Total	p_{+1}	p_{+2}	$\cdots$	p_{+b}	1

Variety measures of association for the cross-classified data were developed by Kendall(1948), Yule and Kendall(1950), Goodman and Kruskal(1959). Traditional measure of association is the chi-square statistic

$$\chi^2 = v(\sum_i \sum_j \frac{p_{ij}^2}{p_{i+} \cdot p_{+j}} - 1), \tag{1}$$

where v is the total number of population elements. Another measures of association suggested by Goodman and Kruskal(1963) include

$$\lambda_b = (\sum_i p_{im} - p_{+m}) / (1 - p_{+m}) \tag{2}$$

and

$$\lambda_a = (\sum_j p_{mj} - p_{m+}) / (1 - p_{m+}), \tag{3}$$

where

$$p_{+m} = \max_j p_{+j}, \quad p_{m+} = \max_i p_{i+}, \quad p_{im} = \max_{1 \le j \le n} p_{ij}, \quad p_{mj} = \max_{1 \le i \le n} p_{ij} \tag{4}$$

For more details of measures of association, see Goodman and Kruskal (1979).

But in many practical situations, it is necessary to get random samples which follow certain covariance or correlation structure instead of complete knowledge (exact values of all cell probabilities) for population.

In this paper, we consider a random number generation problem for these cases when the distribution of the mariginal sums of random variables is multivariate multinomial. Also, we consider some measures of association for these generated random numbers and compare with that of known population via Monte Carlo simulation.

2. Multivariate Multinomial distribution

Continuous multivariate distributions are popular in many areas of statistics. As well as continuous distribution, discrete multivariate distributions are widely applied to many areas, such as sociology, biology and medical sicences.

Among discrete multivariate distributions, we will consider p-th multivariate multinomial distribution with q-th categories for each dimension. Of course, different numbers of categories for each dimension are also possible. But for the simple notations, we assume the same number of categories for each dimension.

Let $N_{i_1 i_2 \cdots i_p} (1 \le i_j \le q,\ 1 \le j \le p)$ be random variables from multinomial distribution $M(n, p_{i_1 i_2 \cdots i_p})$ where

$$\sum_{i_1}\sum_{i_2}\cdots\sum_{i_p} p_{i_1 i_2 \cdots i_p} = 1 \text{ and } \sum_{i_1}\sum_{i_2}\cdots\sum_{i_p} N_{i_1 i_2 \cdots i_p} = n \tag{5}$$

and let

$$\begin{cases} X^{(1)} = (N_{1+\cdots+}, N_{2+\cdots+}, \cdots, N_{q+\cdots+}) \\ X^{(2)} = (N_{+1+\cdots+}, N_{+2+\cdots+}, \cdots, N_{+q+\cdots+}) \\ \quad \vdots \\ X^{(p)} = (N_{+\cdots+1}, N_{+\cdots+2}, \cdots, N_{+\cdots+q}) \end{cases} \tag{6}$$

be a set of random variables of marginal sums for each dimension. Then, each $X^{(i)}$ has multinomial distribution with corresponding marginal sums of probabilties $p^{(i)}$, that is, $X^{(i)} \sim M(n, p^{(i)})$ for $i = 1, \cdots, p$,(see Bishop, Fienberg and Paul, p445).

Now we consider p-dimensional random vector

$$R_{(p)} = (X^{(1)}, X^{(2)}, \cdots, X^{(p)}).$$

Variance-covariance matrix Σ_p of $R_{(p)}$ can be written as follows.

$$\Sigma_p = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{p11} & \Sigma_{p2} & \cdots & \Sigma_{pp} \end{bmatrix}_{pq \times pq}$$

where

$$\sigma_{kl}^{(i,j)} = \mathrm{Cov}(N_{\underbrace{+\cdots+k+\cdots+}_{i}}, N_{\underbrace{+\cdots+l+\cdots+}_{j}}) = (p_{\underbrace{+\cdots+k+\cdots+}_{i-1}\underbrace{k}\underbrace{+\cdots+l+\cdots+}_{j-i}} - p_{\underbrace{+\cdots+k+\cdots+}_{i}} \cdot p_{\underbrace{+\cdots+l+\cdots+}_{j}}) \cdot n,$$

$$\Sigma_{ij} = \mathrm{Cov}(X^{(i)}, X^{(j)}) = \begin{bmatrix} \sigma_{11}^{(i,j)} & \sigma_{12}^{(i,j)} & \cdots & \sigma_{1q}^{(i,j)} \\ \sigma_{21}^{(i,j)} & \sigma_{22}^{(i,j)} & \cdots & \sigma_{2q}^{(i,j)} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{q1}^{(i,j)} & \sigma_{q2}^{(i,j)} & \cdots & \sigma_{qq}^{(i,j)} \end{bmatrix}_{q\times q}$$

For the simplicity, we will explain the bivariate trinomial distribution. In this case, p=2 and q=3. Adopting the above notations, we can summarize the bivariate trinomial random vectors and corresponding probabilities as follows.

Table 2. Trinomial Variables

N_{11}	N_{12}	N_{13}	N_{1+}
N_{21}	N_{22}	N_{23}	N_{2+}
N_{31}	N_{32}	N_{33}	N_{3+}
N_{+1}	N_{+2}	N_{+3}	n

Table 3. Trinomial Probability

p_{11}	p_{12}	p_{13}	p_{1+}
p_{21}	p_{22}	p_{23}	p_{2+}
p_{31}	p_{32}	p_{33}	p_{3+}
p_{+1}	p_{+2}	p_{+3}	1

N_{ij}'s in Table 2 are multinomial random variables with p_{ij} $(1 \le i, j \le 3)$. Then $X^{(1)}$ and $X^{(2)}$ reduce to the marginal $X^{(1)} = (N_{1+}, N_{2+}, N_{3+})$, $X^{(2)} = (N_{+1}, N_{+2}, N_{+3})$ which follows trinomial distribution $M(n, p^{(1)})$ and $M(n, p^{(2)})$ respectively, where $p^{(1)} = (p_{1+}, p_{2+}, p_{3+})$ and $p^{(2)} = (p_{+1}, p_{+2}, p_{+3})$. These mariginals consititute a random vector $R_{(2)} = (X^{(1)}, X^{(2)})$ which has bivariate trinomial distribution.

Lemma 1. Any two variables N_{i+}, N_{+j} form a bivariate binomial random vector with $\mathrm{Cov}(N_{i+}, N_{+j}) = (p_{ij} - p_{i+} \cdot p_{+j}) \cdot n$. That means variance-covariance matrix of $R_{(2)}$ can be written as

$$\Sigma_2 = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ where } \Sigma_{12} = \mathrm{Cov}(X^{(1)}, X^{(2)}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

and $\sigma_{kl} = \mathrm{Cov}(N_{k+}, N_{+l}) = (p_{kl} - p_{k+} \cdot p_{+l}) \cdot n.$

Lemma 2. The correlation matrix Ψ_2 of $R_{(2)} = (X^{(1)}, X^{(2)})$ is also as follows

$$\Psi_2 = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix} \text{ where } \Psi_{12} = \mathrm{Corr}(X^{(1)}, X^{(2)}) = \begin{bmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \\ \rho_{31} & \rho_{32} & \rho_{33} \end{bmatrix}$$

and

$$\rho_{kl} = \mathrm{Corr}(N_{k+}, N_{+l}) = (p_{kl} - p_{k+} \cdot p_{+l}) / \sqrt{p_{k+}(1-p_{k+})}\sqrt{p_{+l}(1-p_{+l})}.$$

Remark. We are greatly concerned for the upper right corner Σ_{12} of the covariance matrix Σ_2 since it reveals the covariance relationships between $X^{(1)}$ and $X^{(2)}$. The diagonal parts Σ_{11}, Σ_{22} of covariance matrix Σ_2 represent inter-dependence of multinomial random variables themselves.

3. Bivariate Trinomial Random Number generation

We want to generate bivariate trinomial random vectors with a specific given correlation matrix Ψ_2 or covariance matrix Σ_2. These procedures can be extended to multivariate multinomial cases without loss of generality. For the completely known probabilities p_{ij}, multinomial random vectors $(X^{(1)}, X^{(2)})$ can be easily obtained by summing corresponding multinomial random numbers. When random vectors are to be generated based on certain correlation or covariance matrix, not each probability, we should solve a sets of nonlinear equations.

The correlation matrix Ψ_2 can be written as a function of $p^{(1)} = (p_{1+}, p_{2+}, p_{3+})$ and $p^{(2)} = (p_{+1}, p_{+2}, p_{+3})$ by Lemma 2. We are greatly concerned to the upper right coner Ψ_{12} of Ψ_2 because it reveals the correlation structure between $X^{(1)}$ and $X^{(2)}$. Exact expression for Ψ_{12} is as follows.

$$\Psi_{12} = \begin{bmatrix} \frac{p_{11} - p_{1+} \cdot p_{+1}}{\sqrt{p_{1+}(1-p_{1+})}\sqrt{p_{+1}(1-p_{+1})}} & \frac{p_{12} - p_{1+} \cdot p_{+2}}{\sqrt{p_{1+}(1-p_{1+})}\sqrt{p_{+2}(1-p_{+2})}} & \frac{p_{13} - p_{1+} \cdot p_{+3}}{\sqrt{p_{1+}(1-p_{1+})}\sqrt{p_{+3}(1-p_{+3})}} \\ \frac{p_{21} - p_{2+} \cdot p_{+1}}{\sqrt{p_{2+}(1-p_{2+})}\sqrt{p_{+1}(1-p_{+1})}} & \frac{p_{22} - p_{2+} \cdot p_{+2}}{\sqrt{p_{2+}(1-p_{2+})}\sqrt{p_{+2}(1-p_{+2})}} & \frac{p_{23} - p_{2+} \cdot p_{+3}}{\sqrt{p_{2+}(1-p_{2+})}\sqrt{p_{+3}(1-p_{+3})}} \\ \frac{p_{31} - p_{1+} \cdot p_{+1}}{\sqrt{p_{3+}(1-p_{3+})}\sqrt{p_{+1}(1-p_{+1})}} & \frac{p_{32} - p_{3+} \cdot p_{+2}}{\sqrt{p_{3+}(1-p_{3+})}\sqrt{p_{+2}(1-p_{+2})}} & \frac{p_{33} - p_{3+} \cdot p_{+3}}{\sqrt{p_{3+}(1-p_{3+})}\sqrt{p_{+3}(1-p_{+3})}} \end{bmatrix}$$

Without loss of generality, we can assume the probabilities p_{ij} are symmetric i.e., $p_{ij} = p_{ji}$. Then correlation matrix Ψ_{12} is symmetrized and 9 elements of Ψ_{12} are reduced to 6 elements. In order to generate bivariate trinomial random numbers for a given specific Ψ_{12} we should solve the following 6 nonlinear equations

$$\frac{p_{ij} - p_{i+} \cdot p_{+j}}{\sqrt{p_{i+}(1-p_{i+})}\sqrt{p_{+j}(1-p_{+j})}} = \rho_{ij} \quad (1 \le i, j \le 3). \tag{7}$$

These sets of nonlinear equations can be solved by the software, **IMSL** (International Mathmatical and Statistical Library), subroutine, **neqnf**. We can get each probability p_{ij} by using neqnf. And then by using these p_{ij} and subroutine **rnmtn** in IMSL, we can generate multinomial random numbers $N_{ij} \sim M(n, p_{ij})$.

Finally we can construct N_{i+} and N_{+j} by summing corresponding cells.

4. Numerical studies

We consider the following correlation matrix as an example.

$$\Psi_{12} = \mathrm{Corr}((N_{1+}, N_{2+}, N_{3+}), (N_{+1}, N_{+2}, N_{+3})) = \begin{bmatrix} 1/4 & -1/8 & -1/8 \\ -1/8 & 1/4 & -1/8 \\ -1/8 & -1/8 & 1/4 \end{bmatrix}$$

By using the subrouitne **neqnf**, we get the following solutions for the set of equations (7).

$$p_{11} = p_{22} = p_{33} = 1/6,\ p_{12} = p_{13} = p_{23} = 1/12$$

Also, by using subroutine **rnmtn,** we can get a set of bivariate trinomial random numbers. In Table 4, we show 3 sets of bivariate trinomial random vectors when $n = 100$.

Table 4. Bivariate Trinomial Random Samples

replication	$X^{(1)} = (N_{1+}, N_{2+}, N_{3+})$	$X^{(2)} = (N_{+1}, N_{+2}, N_{+3})$
1	(36,31,33)	(37,30,33)
2	(34,34,32)	(24,36,40)
3	(37,31,32)	(37,35,28)
⋮	⋮	⋮

In reality, the bivariate trinomial sample 1, (36,31,33) and (37,30,33) are column and row marginals of Table 5, respectively.

Table 5. Random Numbers

24	3	9	36
8	18	5	31
5	9	19	33
37	30	33	100

The measures of association for these generated random vectors are calculated with maximum likelihood estimator of (2)

$$L_b = \frac{\Sigma R_{im} - R_{+m}}{1 - R_{+m}} = \frac{\Sigma N_{im} - N_{+m}}{n - N_{+m}}, \tag{8}$$

of (2) where $R_{ij} = N_{ij}/n$ is sample proportion and so on. Maximum likelihood

estimater L_a of λ_a can be defined similarly,(Goodman and Kruskal, 1954). The results are tabulated in Table 6 and Table 7. All computations are carried on the workstation SS-10. In Table 6, we give the average values of measure of association (2) and (3), respectively. It is based on 1000 replications for each sample size. Standard errors are denoted by s.e. As we can see, the values of measures of association come closer to the population values as n increases in both cases, and the standard errors become smaller as n increases.

Table 6. Measures of Association

criterion	λ_b	λ_a
population	0.25	0.25
n	Average(s.e)	Average(s.e)
100	0.2094(0.0614)	0.2100(0.0608)
200	0.2184(0.0505)	0.2184(0.0505)
300	0.2248(0.0429)	0.2243(0.0431)
400	0.2252(0.0387)	0.2252(0.0370)
500	0.2317(0.0267)	0.2320(0.0259)

Finally, we consider coverage probabilities for measures of association (2) and (3) with the generated random vectors, respectively. The asymptotic standard normality of

$$\sqrt{n}(L_b - \lambda_b)\sqrt{\frac{(1-R_{+m})^3}{(1-\Sigma R_{im})(\Sigma R_{im} + R_{+m} - 2\Sigma^r R_{im})}} \tag{9}$$

was discussed in Goodman and Kruskal(1972) where L_b is maximum likelihood estimator of λ_b and $\Sigma^r R_{im}$ represents the sum of R_{im} over those values of i such that R_{im} is taken on in that column in which R_{+m} is taken. For λ_a, the same result holds if the roles of columns and rows are interchanged.

Table 7. Coverage probabilities

criterion	λ_b			λ_a		
n \ level	90%	95%	99%	90%	95%	99%
100	0.941	0.990	0.999	0.939	0.987	1.000
200	0.893	0.961	0.994	0.905	0.962	0.996
300	0.890	0.945	0.990	0.896	0.943	0.987
400	0.895	0.957	0.993	0.894	0.947	0.994
500	0.898	0.947	0.990	0.897	0.948	0.989

Based on the asymptotic standard normality of (9), the confidence interval for λ_b can be easily obtained,

$$P(L_b - z_{\alpha/2} \cdot Q^{-1} \le \lambda_b \le L_b + z_{\alpha/2} \cdot Q^{-1}) \cong 1-\alpha$$

where $Q = \sqrt{(1-R_{+m})^3 / (1-\Sigma R_{im})(\Sigma R_{im} + R_{+m} - 2\Sigma^r R_{im})}$. Similarly the confidence interval for λ_a can be obtained.

We considered three confidence levels 90%, 95% and 99%. All computations were also based on 1000 replications. Table 7 represents empirical coverage probability for each sample size. It looks like that coverage probabilitites become closer to the confidence levels, respectively, as sample size n increases.

References

Bishop, Y. M. M ., Fienberg, S. E. and Paul, W. H. (1975), *Discrete Multivariate analysis, Theory and Practice*, MIT Press Cambridge, Massachusetts, and London, England.

Goodman, L. A. and Kruskal, W. H. (1954), Measures of association for cross-classification, *Journal of the Americal Statisical Association, 49. 732-764*

Goodman, L. A. and Kruskal, W. H. (1963), Measures of association for cross-classification Ⅲ : Approximate sampling theory., *Journal of the American Statisical Association, 58. 123-163*

Goodman, L. A. and Kruskal, W. H. (1972), Measures of association for cross-classification Ⅳ : Simplification of asymptotic variance, *Journal of the American Statisical Association, 67. 415-421*

Goodman, L. A. and Kruskal, W. H. (1979), *Measures of Association for cross classifications,* Springer-verlag, New York Heidelberg, Berlin.

IMSL(1991), User's Manual : *Stat/Library and Math/Library*

Kendall, M. G.(1948), *The advanced Theory of Statistics,* London, Charles Griffin and co. ltd.

Yule, G. U. and Kendall, M. G. (1950), *An introduction to the theory of Statistics,*