Multivariate Generalizations of the Multiplicative Binomial Distribution: Introducing the MM Package

P. M. E. Altham
University of Cambridge

Robin K. S. Hankin
Auckland University of Technology

Abstract

We present two natural generalizations of the multinomial and multivariate binomial distributions, which arise from the multiplicative binomial distribution of Altham (1978). The resulting two distributions are discussed and we introduce an \textit{R} package, \texttt{MM}, which includes associated functionality.

This vignette is based on Altham and Hankin (2012).

\textbf{Keywords:} Generalized Linear Models, Multiplicative Binomial, Overdispersion, Overdispersed Binomial, Categorical Exponential Family, Multiplicative Multinomial Distribution, \textit{R}.

1. Introduction

The uses of the binomial and multinomial distributions in statistical modelling are very well understood, with a huge variety of applications and appropriate software, but there are plenty of real-life examples where these simple models are inadequate. In the current paper, we first remind the reader of the two-parameter exponential family generalization of the binomial distribution first introduced by Altham (1978) to allow for over- or under-dispersion:

\[ P(Z = j) = \binom{n}{j} p^j q^{n-j} \theta^{j(n-j)} \left/ f(p, \theta, n) \right. \quad j = 0, \ldots, n \]  

(1)

where

\[ f(p, \theta, n) = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \theta^{j(n-j)}. \]  

(2)

Here, \( 0 \leq p \leq 1 \) is a probability, \( p + q = 1 \), and \( \theta > 0 \) is the new parameter which controls the shape of the distribution; the standard binomial \( \text{Bi}(n, p) \) is recovered if \( \theta = 1 \). Altham points out that this distribution is more sharply peaked than the binomial if \( \theta > 1 \), and more diffuse if \( \theta < 1 \). As far as we are aware, no other generalization has this type of flexibility; for example, the beta-binomial distribution (Johnson, Kemp, and Kotz 2005) only allows for \textit{over}-dispersion relative to the corresponding binomial distribution.

We then introduce two different generalizations, both of which are of exponential family form. We call these:
The multivariate multinomial distribution

Take non-negative integers $y_1, \ldots, y_k \geq 0$ with $\sum y_i = y$, a fixed integer. Then suppose that the probability mass function of $Y_1, \ldots, Y_k$ is

$$P(y_1, \ldots, y_k) = C^{-1} \left( \frac{y}{y_1 \ldots y_k} \right) \prod_{i=1}^{k} p_i^{y_i} \prod_{1 \leq i < j \leq k} \theta_{ij}^{y_i y_j}$$

where the free parameters are $p = (p_1, \ldots, p_k)$ and $\theta = \theta_{ij}$, with $p_i \geq 0$ for $1 \leq i \leq k$, $\sum p_i = 1$, and $\theta_{ij} > 0$ for $1 \leq i < j \leq k$ [these restrictions on $i, j$ understood henceforth]. Here $C = C(y, p, \theta)$ is a normalization constant. Thus the standard multinomial is recovered if $\theta_{ij} = 1$, and in this case $C = 1$.

The multivariate multiplicative binomial distribution

For simplicity of notation, we restrict attention to the bivariate case. The proposed frequency function $P(X_1 = x_1, X_2 = x_2) = f(x_1, x_2)$ is

$$f(x_1, x_2) = C^{-1} \left( \frac{m_1}{x_1 z_1} \right) p_1^{x_1} q_1^{z_1} \cdot \left( \frac{m_2}{x_2 z_2} \right) p_2^{x_2} q_2^{z_2} \cdot \phi^{x_1 x_2}$$

where $p_i + q_i = 1$ and $x_i + z_i = m_i$, $i = 1, 2$; all parameters are strictly positive and $C$ is again a normalization constant. Thus $X_1, X_2$ are independent iff $\phi = 1$. Furthermore, if $\phi = 1$, then $\theta_1 = \theta_2 = 1$ corresponds to $X_1, X_2$ independent Binomial: $X_i \sim Bi(m_i, p_i)$.

We then introduce an R (R Development Core Team 2011) package MM which implements some functionality for these distributions. Because of their simple exponential family forms, both these distributions may be fitted to appropriate count data by using the glm() R function with the Poisson distribution and log link function. This follows from the ingenious result of Lindsey and Mersch (1992), and has the very important consequence that the computationally expensive normalizing constants in Equations 3 and 4 above need never be evaluated.

Both these distributions are clearly exponential family-type distributions (Cox and Hinkley 1974), and may be seen as discrete analogues of the multivariate normal distribution in the following two respects, which we illustrate only for the multiplicative multinomial distribution. Firstly, suppose we have observed frequencies $n(y_1, \ldots, y_k)$ with $\sum y_i = y$, then the log likelihood of this dataset may be written as

$$\sum n(y_1, \ldots, y_k) \log P(y_1, \ldots, y_k).$$

where the summation is over $y_1, \ldots, y_k \geq 0$ with $\sum y_i = y$. This gives rather simple expressions, essentially the sample mean and sample covariance matrix of the vector $(y_1, \ldots, y_k)$, for the minimal sufficient statistics of this exponential family. Hence, by standard theory for exponential families, at the maximum likelihood value of the parameters, the observed and fitted values of the minimal sufficient statistics will agree exactly.

Secondly, as we show later, each of the distributions given in Equations 3 and 4 is reproductive under conditioning on components.
1.1. A probabilistic derivation of the Multiplicative Multinomial

It is possible to consider the MM distribution from the perspective of contingency tables, which for simplicity we will carry out for \( k = 3, \ y = 4 \): The general case is notationally challenging.

Our preferred interpretation is drawn from the field of psephology: Consider a household of 4 indistinguishable voters, each of whom votes for exactly one of 3 political parties, say \( \wp_1, \wp_2, \wp_3 \). Let \( y_1, y_2, y_3 \) be the total number of votes obtained from this household for \( \wp_1, \wp_2, \wp_3 \) respectively, and so \( y_1 + y_2 + y_3 = 4 \).

The 4 voters in the household may be considered as corresponding to the rows, columns, layers and the 4th dimension of a \( 3 \times 3 \times 3 \times 3 \) contingency table, with cell probabilities \( P_{ijhl} \) for \( 1 \leq i, j, h, l \leq 3 \), which we will assume have the following symmetric form

\[
P_{ijhl} = \frac{1}{C'} \cdot p_ip_jp_hp_l \cdot \theta_{ij} \theta_{ih} \theta_{ji} \theta_{jl} \theta_{hl}(6)
\]

where \( \theta_{rs} = \theta_{sr} \) for \( s, r \in \{i, j, h, l\} \) [notation is analogous to that used in Equation 8 of Altham (1978), with \( \theta \) written for \( \phi \)], and without loss of generality \( \theta_{rr} = 1 \).

The parameters \( \theta_{ij} \) may be interpreted in terms of conditional cross-ratios. Recalling that \( P_{ijhl} = P_{ijlh} = \ldots = P_{lhji} \) we have, for example:

\[
P_{12hl} P_{21hl} = \theta_{12}^2 \quad \text{for each } h, l. \tag{7}
\]

By enumerating the possible voting results for a given family of size 4, we may find the resulting joint distribution of \((Y_1, Y_2, Y_3)\), where random variable \( Y_i \) is the household total of votes for party \( \wp_i, \ i = 1, 2, 3 \). For example, \( P(Y_1 = 4, Y_2 = 0, Y_3 = 0) = \frac{1}{C'} \cdot p_1^4 \) is the probability that all 4 members of the household vote for \( \wp_1 \). Similarly, \( P(Y_1 = 3, Y_2 = 1, Y_3 = 0) = \frac{1}{C'} \cdot 4p_1^3p_2^1\theta_{12}^3 \) is the probability that 3 members of the household vote for \( \wp_1 \) and the remaining 1 member votes for \( \wp_2 \).

This clearly corresponds to the given Multiplicative Multinomial distribution, so \( C = C' \). We return to this example with a synthetic dataset in Section 3.1 below.

1.2. Marginal and conditional distributions

There does not appear to be an elegant expression for the marginal distribution of \((Y_1, \ldots, Y_r)\) where \( r < k \). However, the multiplicative multinomial behaves ‘elegantly’ under conditioning on a subset of the variables \((Y_1, \ldots, Y_k)\). For example,

\[
P(Y_1, Y_2 \mid y_3, \ldots, y_k) \propto \phi_1^{y_1} \phi_2^{y_2} \theta_{12}^{y_1+y_2} y_1! y_2! \tag{8}
\]

where

\[
\phi_1 = p_1 \theta_{13} \ldots \theta_{1k} \quad \phi_2 = p_2 \theta_{23} \ldots \theta_{2k}. \tag{9}
\]

Hence the distribution of \( Y_1 \), conditional on the values of \((y_3, \ldots, y_k)\) is multiplicative binomial in the sense of Altham (1978). Similarly, the distribution of \((Y_1, \ldots, Y_\nu)\), conditional on the values of \((y_{\nu+1}, \ldots, y_k)\) is multiplicative multinomial in the sense defined above.
1.3. The normalization constant

The constant $C$ in Equation 3 must be determined by numerical means:

$$C = \sum_{y_1+y_2+y_3=y} \left( \frac{y}{y_1 \cdots y_k} \right) \prod_{i=1}^{k} p_i^{y_i} \prod_{1 \leq i < j \leq k} \theta_{ij}^{y_iy_j}. \quad (10)$$

Although this is provided in the MM package [function NormC()], it is computationally expensive and difficult to evaluate for all but small $k$ and $y$.

1.4. The Poisson method for fitting distributions with an intractable normalizing constant

The parameters $(p, \theta)$ may be estimated without determining the normalizing constant $C$ by transforming the problem into a Generalized Linear Model. The method presented here follows Lindsey and Mersch (1992); for simplicity of notation we take $k = 3$. Equation 3 is equivalent to

$$\log P(y_1, y_2, y_3) = \mu + \sum y_i \log p_i + \sum_{i<j} y_i \cdot y_j \log \theta_{ij} + \text{offset}[y_1, y_2, y_3] \quad (11)$$

where $\text{offset}[y_1, y_2, y_3] = -\sum \log (y_i!)$ accounts for the multinomial term on the right hand side. The log-likelihood $\mathcal{L}$ of the dataset $n(y_1, y_2, y_3)$ is given by

$$\mathcal{L} = \sum \log P(y_1, y_2, y_3). \quad (12)$$

Thus, treating $n(y_1, y_2, y_3)$ as independent Poisson variables with parameters$^1$ given by Equation 11, we may fit the parameters of Equation 3 using $\text{glm}(\ldots, \text{family}=\text{poisson})$, using the canonical log link function, and regressing $n(y_1, y_2, y_3)$ on the variables

$$y_1, y_2, y_3, y_1y2, y1y3, y2y3.$$

With obvious notation, the R idiom is

$$\text{glm}(n \sim -1+\text{offset(Off)} + y1 + y2 + y3 + y1:y2 + y1:y3 + y2:y3, \text{family} = \text{poisson})$$

(recall that $y_1 + y_2 + y_3 = y$, fixed), which is given by function Lindsey() in the package.

$^1$The distribution of independent Poisson random variables conditional on their total is multinomial with probabilities equal to the scaled Poisson parameters. If $X_i \sim \text{Po} (\lambda_i)$, then elementary considerations show

$$P \left( X_1 = x_1, \ldots, X_k = x_k \bigg| \sum x_i = N \right) = \binom{N}{x_1, \ldots, x_k} \prod \left( \frac{\lambda_i^{x_i}}{\sum \lambda_i} \right)^{x_i},$$

the right hand side being recognisable as a multinomial distribution. Given that the distribution is of the exponential family, it is the case that $\sum n = \sum \lambda_i$, the normalizing constant is not needed.
1.5. Multivariate multiplicative binomial

Considering the bivariate case for simplicity, suppose \((X_1, X_2)\) to be non-negative integers not exceeding known fixed maxima \(m_1, m_2\) respectively.

We introduce a 5-parameter distribution of exponential family form. In common with the multiplicative binomial, it has the property that at the maximum likelihood values of these parameters, the observed and fitted values of the means of \(X_i, i = 1, 2\) will agree exactly, and similarly for the observed and fitted values of the covariance matrix of \(X_1, X_2\). This distribution is easy to fit to frequency data (again using the Lindsey and Mersch Poisson device). The distribution has some nice properties, but there do not appear to be simple formulae for its moments.

The proposed frequency function \(P(X_1 = x_1, X_2 = x_2) = f(x_1, x_2)\) is

\[
f(x_1, x_2) = C^{-1} \left( \frac{m_1}{x_1 z_1} p_1^{x_1 z_1} \theta_1^{z_1} \cdot \frac{m_2}{x_2 z_2} p_2^{x_2 z_2} \theta_2^{z_2} \right) \cdot \phi^{x_1 x_2}
\]

where \(p_i + q_i = 1\) and \(x_i + z_i = m_i\); all parameters are strictly positive.

Here, \(C\) is the normalization constant:

\[
C = \sum_{x_1 + z_1 = m_1} \sum_{x_2 + z_2 = m_2} \left( \frac{m_1}{x_1 z_1} p_1^{x_1 z_1} \theta_1^{z_1} \cdot \frac{m_2}{x_2 z_2} p_2^{x_2 z_2} \theta_2^{z_2} \right) \cdot \phi^{x_1 x_2}
\]

Thus \(X_1, X_2\) are independent if \(\phi = 1\). As already noted, if \(\phi = 1\), then \(\theta_1 = \theta_2 = 1\) corresponds to \(X_1, X_2\) independent Binomial: \(X_i \sim Bi(m_i, p_i)\).

Although there does not seem to be a simple expression for the correlation between \(X_1\) and \(X_2\), it is easily seen that \(\phi\) controls their independence in a likelihood ratio fashion, with

\[
\frac{f(x_1, x_2) f(x_1 + 1, x_2 + 1)}{f(x_1 + 1, x_2) f(x_1, x_2 + 1)} = \phi.
\]

Indeed, following Lehmann (1966) we can prove a much stronger statement: The variables \(X_1, X_2\) are positive monotone likelihood ratio dependent\(^2\) for \(\phi > 1\), negative if \(\phi < 1\).

The conditional distribution is again of multiplicative binomial form, since we can write

\[
P(X_1 = x_1 | X_2 = x_2) \propto \left( \frac{m_1}{x_1 z_1} p_1^{x_1 z_1} \theta_1^{z_1} \right) \phi^{x_1 z_1}.
\]

\[
(16)
\]

2. The MM package

The MM package associated with this article provides R functionality for assessing the multiplicative multinomial and multivariate binomial. We have provided user-friendly wrappers to expedite use of the distributions in a data analysis setting.

\(^2\)Random variables \(X_1, X_2\) are positive monotone likelihood ratio dependent if \(f(x_1, x_2) f(x_1', x_2') \leq f(x_1, x_2) f(x_1', x_2')\) for all \(x_1 < x_1', x_2 < x_2'\), and negative monotone likelihood ratio dependent if the inequality is reversed.
The **MM** package uses an object-oriented approach: The set of free parameters (one vector, one upper-diagonal matrix) is not a standard R object, and is defined to be an object of S4 class **paras**. The objects thus defined are user-transparent and a number of manipulation methods are provided in the package.

For example, consider Equation 3 with \( k = 5 \) and \( p_i = \frac{1}{5}, \ 1 \leq i \leq 5 \) and \( \theta_{ij} = 2 \) for \( 1 \leq i < j \leq 5 \). This distribution would be underdispersed compared with the corresponding multinomial. It is straightforward to create an object corresponding to the parameters for this distribution using the package:

```r
R> library("MM")
R> pm1 <- paras(5,pnames=letters[1:5])
R> theta(pm1) <- 2
R> pm1

$p
   a b c d e
0.2 0.2 0.2 0.2 0.2

$theta
   a b c d e
 a NA 2 2 2 2
 b NA NA 2 2 2
 c NA NA NA 2 2
 d NA NA NA NA 2
 e NA NA NA NA NA
```

Now we may sample repeatedly from the distribution (sampling is quick because it does not require evaluation of the normalization constant). Consider \( y = 20 \):

```r
R> set.seed(0)
R> (sample1 <- rMM(n=10, Y=20, paras=pm1))

   a b c d e
[1,] 3 4 3 5 5
[2,] 3 3 4 5 5
[3,] 4 3 5 5 3
[4,] 4 3 6 4 3
[5,] 5 4 3 3 5
[6,] 3 4 4 5 4
[7,] 5 4 4 4 3
[8,] 4 5 4 3 4
[9,] 5 4 4 3 4
[10,] 2 4 4 5 5
```

See how closely clustered the sample is around its mean of \((4, 4, 4, 4, 4)\); compare the wider dispersion of the multinomial:
Thus sample2 is drawn from the classical multinomial. It is then straightforward to perform a likelihood ratio test on, say, sample1:

R> support1 <- MM_allsameasum(sample1, paras=pm1)
R> support2 <- MM_allsameasum(sample1, paras=pm2)
R> support1-support2
[1] 14.562

Function MM_allsameasum() calculates the log likelihood for a specific parameter object (in this case, pm1 and pm2 respectively) and we see that, for sample1, hypothesis pm1 is preferable to pm2 on the grounds of a likelihood ratio of about \( \Lambda = 0.47 \times 10^{-6} \), corresponding to 14.56 units of support. This would exceed the two units of support criterion suggested by Edwards (1992) and we could reject pm2. Alternatively, we could observe that \(-2 \log \Lambda\) is in the tail region of its asymptotic distribution, \(\chi^2_1\).

The package includes a comprehensive suite of functionality which is documented through the R help system and accessible by typing `library(help="MM")` at the command prompt.

### 3. The package in use

The package comes with a number of datasets, four of which are illustrated here. We begin with a small synthetic dataset, then consider data taken from the social sciences, previously analyzed by Wilson (1989); analyze some pollen counts considered by Mosimann (1962) in the context of palaeoclimatology; and finally assess a marketing science dataset.

#### 3.1. Synthetic voting dataset

We begin with a small synthetic dataset which is simple enough to illustrate the salient aspects of the multiplicative multinomial distribution, and the MM package.
This dataset arises from 96 households each of size 4, in which each member of the household is noted as voting Lib, Con or Lab respectively. We take \( n(\cdot,\cdot,\cdot) \) as the voting tally; thus \( n(0,0,4) = 5 \) [the first line] means that there are exactly 5 households in which all 4 members vote Labour; similarly \( n(0,1,3) = 8 \) means that there are exactly 8 households in which 1 member votes Conservative and the remaining 3 vote Labour.

\[
\begin{array}{ccc|c}
\text{Lib} & \text{Con} & \text{Lab} & \text{voting_tally} \\
[1,] & 0 & 0 & 4 & 5 \\
[2,] & 0 & 1 & 3 & 8 \\
[3,] & 0 & 2 & 2 & 7 \\
[4,] & 0 & 3 & 1 & 4 \\
[5,] & 0 & 4 & 0 & 6 \\
[6,] & 1 & 0 & 3 & 1 \\
[7,] & 1 & 1 & 2 & 7 \\
[8,] & 1 & 2 & 1 & 4 \\
[9,] & 1 & 3 & 0 & 9 \\
[10,] & 2 & 0 & 2 & 5 \\
[11,] & 2 & 1 & 1 & 7 \\
[12,] & 2 & 2 & 0 & 12 \\
[13,] & 3 & 0 & 1 & 2 \\
[14,] & 3 & 1 & 0 & 7 \\
[15,] & 4 & 0 & 0 & 12 \\
\end{array}
\]

One natural hypothesis is that the data are drawn from a multinomial distribution (alternative hypotheses might recognize that individuals within a given household may be non-independent of each other in their voting).

The multinomial hypothesis may be assessed using \texttt{glm()} following \textit{Lindsey and Mersch (1992)} but without the interaction terms:

\[
\text{R> Off } \leftarrow -\text{rowSums(lfactorial(voting))} \\
\text{R> summary(glm(voting_tally } \sim -1 + (.) + \text{offset(Off)}, \\
\text{ data } = \text{as.data.frame(voting), family } = \text{poisson})
\]

Call:
\[
\text{glm(formula } = \text{voting_tally } \sim -1 + (.) + \text{offset(Off), family } = \text{poisson,} \\
\text{ data } = \text{as.data.frame(voting))}
\]

Deviance Residuals:
\[
\begin{array}{cccc}
\text{Min} & 1Q & \text{Median} & 3Q & \text{Max} \\
-3.341 & -1.234 & -0.103 & 1.753 & 4.904 \\
\end{array}
\]

Coefficients:
\[
\text{Estimate Std. Error z value Pr(}>|z|<) }
\]
Thus the model fails to fit (77.315 being much larger than the corresponding degrees of freedom, 12). This is because the observed frequencies of the cells in which all members of the household vote for the same party (namely for rows 1, 5 and 15 of the data) greatly exceed the corresponding expected numbers under the simple multinomial model.

The next step is to take account of the fact that individuals within a given household may be non-independent of each other in their voting intentions (and may indeed tend to disagree with each other rather than all vote the same way). Positive dependence between individuals in a household could be modelled by the Dirichlet-multinomial distribution (Mosimann 1962), but by using the Multiplicative Multinomial introduced here, we are allowing dependence between individuals in a household to be positive or negative.

The MM parameters may be estimated, again following Lindsey and Mersch (1992) but this time admitting first-order interaction, using bespoke function Lindsey():

R> Lindsey(voting, voting_tally, give_fit = TRUE)

$MLE
$p
   Lib  Con  Lab
0.36695 0.31515 0.31790

$theta
   Lib  Con  Lab
Lib  NA 0.67351 0.48259
Con  NA  NA 0.65153
Lab  NA  NA  NA

$fit

Call:
glm(formula = jj$d ~ -1 + offset(Off) + (.)^2, family = poisson, data = data.frame(jj$tbl))

Deviance Residuals:
The MM distribution

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.330</td>
<td>-1.014</td>
<td>0.293</td>
<td>0.606</td>
<td>1.400</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate       | Std. Error | z value | Pr(>|z|) |
|----------------|------------|---------|----------|
| Lib            | 1.3745     | 0.0712  | 19.31    | < 2e-16 |
| Con            | 1.2224     | 0.0898  | 13.61    | < 2e-16 |
| Lab            | 1.2311     | 0.0925  | 13.31    | < 2e-16 |
| Lib:Con        | -0.3952    | 0.0844  | -4.68    | 2.8e-06 |
| Lib:Lab        | -0.7286    | 0.1023  | -7.13    | 1.0e-12 |
| Con:Lab        | -0.4284    | 0.0965  | -4.44    | 8.9e-06 |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 543.236 on 15 degrees of freedom
Residual deviance: 11.501 on 9 degrees of freedom
AIC: 77.16

Number of Fisher Scoring iterations: 5

Observe that the MLEs of $p$ [viz $0.367, 0.315, 0.318$] are obtained as proportional to the exponential of the estimated regression coefficients: $e^{1.375}, e^{1.222}, e^{1.231}$, normalized to add to 1.

This model is quite a good fit in the sense that the null deviance 11.5 is not in the tail region of $\chi^2_9$, its null distribution; it can be seen that all 3 interaction parameters are significant and, for example, $\hat{\theta}_{23} = 0.652 = \exp(-0.428)$.

The corresponding conditional cross-ratios are all significantly greater than 1; for example

$$\frac{\hat{P}_{11hl}\hat{P}_{22hl}}{\hat{P}_{12hl}\hat{P}_{21hl}} = \frac{1}{\theta_{12}^2} = \frac{1}{0.6735116^2} = 2.2045. \quad (17)$$

3.2. Housing satisfaction data

We now consider a small dataset taken from Table 1 of Wilson (1989), who analyzed the dataset in the context of overdispersion (Wilson himself took the dataset from Table 1 of Brier (1980)).

In a non-metropolitan area, there were 18 independent neighbourhoods each of 5 households, and each household gave its response concerning its personal satisfaction with their home. The allowable responses were ‘unsatisfied’ (US), ‘satisfied’ (S), and ‘very satisfied’ (VS).

```r
R> data("wilson")
R> head(non_met)

   US  S  VS
[1,] 3  2  0
```
Thus the first neighbourhood had three households responding US, two reporting S, and zero reporting VS; the second neighbourhood had the same reporting pattern.

Observe that the 5 households within a neighbourhood may not be independent in their responses. The first step is to recast the dataset into a table format; the package provides a function `gunter()` [named for the R-lister who suggested the elegant and fast computational method]:

R> wilson <- gunter(non_met)
R> wilson

$table

US S VS
1 5 0 0
2 4 1 0
3 3 2 0
4 2 3 0
5 1 4 0
6 0 5 0
7 4 0 1
8 3 1 1
9 2 2 1
10 1 3 1
11 0 4 1
12 3 0 2
13 2 1 2
14 1 2 2
15 0 3 2
16 2 0 3
17 1 1 3
18 0 2 3
19 1 0 4
20 0 1 4
21 0 0 5

d

[1] 1 5 4 2 0 2 1 0 0 1 1 0 0 1 0 0 0 0 0 0

Thus 1 neighbourhood reported c(5,0,0), and 5 neighbourhoods reported c(4,1,0) [because d[1]=1 and tbl[1,]=c(5,0,0); and d[2]=5 and tbl[2,]=c(4,1,0) respectively].

The hypothesis that the data are drawn from a multinomial distribution may again be assessed by using Lindsey and Mersch’s technique:
The MM distribution

R> attach(wilson)
R> Off <- -rowSums(lfactorial(tbl))
R> summary(glm(d~1+(.)+offset(Off),data=tbl,family=poisson))

Call:
glm(formula = d ~ -1 + (.) + offset(Off), family = poisson, data = tbl)

Deviance Residuals:
     Min      1Q  Median      3Q     Max
-1.728   -0.398  -0.105   0.338   2.222

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
US          0.88600   0.11092   7.967  1.7e-15
S           0.67289   0.13259   5.101  3.4e-07
VS         -1.35512   0.43748  -3.103  0.0019

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 110.17 on 21 degrees of freedom
Residual deviance: 21.22 on 18 degrees of freedom
AIC: 49.19

Number of Fisher Scoring iterations: 6

Thus the multinomial model is a reasonable fit, in the sense that the residual deviance of 21.22
is consistent with the null distribution, $\chi^2_{18}$. The slightly increased value would be because
the observed frequencies for neighbourhoods in agreement (that is, either perfect agreement—
c(5,0,0) or c(0,5,0) or c(0,0,5)—or near-perfect, as in c(4,1,0)) exceed the corresponding
expected numbers under the simple multinomial model.

The MM parameters may be estimated, again following Lindsey and Mersch (1992) but this
time admitting first-order interaction:

R> Lindsey(wilson,give_fit=TRUE)

$MLE
$p
  US   S   VS
0.494469 0.411254 0.094277

$theta
  US   S   VS
US NA  0.74424 0.59647
S  NA   NA  0.88449
VS NA   NA   NA
$fit$

Call:
\[
glm(formula = jj$d \sim -1 + offset(Off) + (.)^2, family = poisson, 
data = data.frame(jj$tbl))
\]

Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.913</td>
<td>-0.498</td>
<td>-0.130</td>
<td>0.342</td>
<td>1.192</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| US       | 1.140      | 0.112   | 10.20   | < 2e-16 |
| S        | 0.956      | 0.161   | 5.95    | 2.6e-09 |
| VS       | -0.517     | 1.652   | -0.31   | 0.7543 |
| US:S     | -0.295     | 0.112   | -2.64   | 0.0084 |
| US:VS    | -0.517     | 0.508   | -1.02   | 0.3086 |
| S:VS     | -0.123     | 0.545   | -0.23   | 0.8219 |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 110.173 on 21 degrees of freedom
Residual deviance: 13.425 on 15 degrees of freedom
AIC: 47.4

Number of Fisher Scoring iterations: 6

Thus in this dataset, only the first interaction parameter US:S is significant. This might be interpreted as an absence of VS responses coupled with a broader than expected spread split between US and S. Note that the residual deviance is now less than the corresponding degrees of freedom.

In this case, the three categories US, S, and VS are ordered, a feature which is not used in the present approach. It is not clear at this stage how we could best include information about such ordering into our analysis.

We now check agreement of the observed and expected sufficient statistics:

R> summary(suffstats(wilson))

$row_sums

<table>
<thead>
<tr>
<th>US</th>
<th>S</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.61111</td>
<td>2.11111</td>
<td>0.27778</td>
</tr>
</tbody>
</table>

$cross_prods

<table>
<thead>
<tr>
<th>US</th>
<th>S</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 9.27778</td>
<td>3.38889</td>
<td>0.38889</td>
</tr>
</tbody>
</table>
The MM distribution

S  3.38889  6.55556  0.61111
VS 0.38889  0.61111  0.38889

The summary() method gives normalized statistics so that, for example, the row_sums total y = 5. This may be compared with the expectation of the maximum likelihood MM distribution:

R> L <- Lindsey(wilson)
R> expected_suffstats(L,5)

$row_sums
   US   S   VS
2.61111 2.11111 0.27778

cross_prods
   US   S   VS
US 9.27778 3.38889 0.38889
S  3.38889 6.55556 0.61111
VS 0.38889 0.61111 0.38889

showing agreement to within numerical precision.

3.3. Mosimann’s forest pollen dataset

Palynology offers a unique perspective on palaeoclimate; pollen is durable, easily identified, and informative about climate (Faegri and Iversen 1992). We now consider a dataset collected in the context of palaeoclimate (Sears and Clisby 1955; Clisby and Sears 1955), and further analyzed by Mosimann (1962).

We consider a dataset taken from the Bellas Artes core from the Valley of Mexico (Clisby and Sears 1955, Table 2); details of the site are given by Foreman (1955). The dataset comprises a matrix with N = 73 observations, each representing a depth in the core, and k = 4 columns, each representing a different type of pollen. We follow Mosimann in assuming that the 73 observations are independent, and in restricting the analysis to depths at which a full complement of 100 grains were identified.

R> data("pollen")
R> pollen <- as.data.frame(pollen)
R> head(pollen)

Pinus Abies Quercus Alnus
1  94  0  5  1
2  75  2 14  9
3  81  2 13  4
4  95  2  3  0
5  89  3  1  7
6  84  5  7  4
Thus each row is constructed to sum to 100, and there are 4 distinct types of pollen; hence in our notation $y = 100$ and $k = 4$.

Observe that this dataset, in common with the housing satisfaction data considered above, has to be coerced to histogram form; but this time the numbers are larger. The partitions package (Hankin 2006) uses generating functions to determine that there are exactly

\[ R > S(rep(100,4),100) \]

[1] 176851

possible non-negative integer solutions to $y_1 + y_2 + y_3 + y_4 = 100$ (most of these have zero observed count). Each of these solutions must be generated and this is achieved using the compositions() function of the partitions package (Hankin 2007).

First we repeat some of Mosimann’s calculations as a check. Using the ordinary multinomial distribution $Mn(y, p)$, we find $\hat{p}$ to be

\[ R > p.hat <- colSums(pollen)/sum(pollen) \]

Pinus  Abies  Quercus  Alnus
0.8627  0.0141  0.0907  0.0325

The observed sample variances for the counts are

\[ R > apply(pollen,2,var) \]

Pinus  Abies  Quercus  Alnus
48.51  2.08  25.87  8.19

but if the ordinary multinomial model held, we would expect these variances to be

\[ R > 100*p.hat*(1-p.hat) \]

Pinus  Abies  Quercus  Alnus
11.84  1.39  8.25  3.14

respectively. This shows that the dataset has pronounced over-dispersion compared to the ordinary multinomial. Furthermore, the sample correlation matrix is not what we would expect from the ordinary multinomial.

As Mosimann points out, the sample correlation matrix is

\[ R > cor(pollen) \]

```
Pollen         Pinus       Abies      Quercus      Alnus
Pollen       1.0000 -0.3018 -0.8960 -0.6892
Abies         -0.3020  1.0000  0.0870  0.0761
Quercus       -0.8960  0.0870  1.0000  0.3595
Alnus          -0.6892  0.0761  0.3600  1.0000
```
while the correlation matrix for the multinomial corresponding to $\hat{p}$ is actually

$$
\begin{array}{cccc}
\text{Pinus} & \text{Abies} & \text{Quercus} & \text{Alnus} \\
\text{Pinus} & 1.000 & -0.2999 & -0.7917 & -0.4592 \\
\text{Abies} & -0.300 & 1.0000 & -0.0378 & -0.0219 \\
\text{Quercus} & -0.792 & -0.0378 & 1.0000 & -0.0578 \\
\text{Alnus} & -0.459 & -0.0219 & -0.0578 & 1.0000 \\
\end{array}
$$

(We have corrected what seems to be a small typo in the 4th column of this matrix in Mosimann’s Table 2). It is particularly striking that the data show positive correlations for 3 entries. Such positive correlations could never arise from the Dirichlet-multinomial distribution, but they will be exactly matched by our new multiplicative multinomial model. The full sample covariance matrix for the dataset is

$$
R> \text{var(pollen)}
$$

$$
\begin{array}{cccc}
\text{Pinus} & \text{Abies} & \text{Quercus} & \text{Alnus} \\
\text{Abies} & -3.03 & 2.079 & 0.638 & 0.314 \\
\text{Quercus} & -31.74 & 0.638 & 25.870 & 5.233 \\
\text{Alnus} & -13.74 & 0.314 & 5.233 & 8.188 \\
\end{array}
$$

which is precisely the covariance of the multiplicative multinomial distribution at the maximum likelihood (ML) parameters.

Calculating the Normalizing constant for the MM distribution is computationally expensive; \text{NormC()} takes over 60 seconds to execute on a 2.66 GHz Intel PC running linux. For direct maximization of the log-likelihood function, for example by \text{MM} function \text{optimizer()}, one would have to call \text{NormC()} many times. Thus function \text{Lindsey()} represents, in this case, a considerable saving of time in maximizing the log-likelihood (the call below took under 15 seconds elapsed time):

$$
\text{Lindsey(pollen, give_fit=TRUE)}
$$
Figure 1: Marginal frequency distributions for numbers of each of four pollen types based on the multinomial distribution (black) and the multiplicative multinomial (red).
### 3.4. Marketing science: An example of the multivariate multiplicative binomial

We now illustrate the multivariate multiplicative binomial with an example drawn from the field of economics. Danaher and Hardie (2005) considered a dataset obtained from a sample of \( N = 548 \) households over four consecutive store trips. For each household, they counted the total number of egg purchases in their four eligible shopping trips, and the total number of bacon purchases for the same trips; the hypothesis was that egg consumption was correlated with bacon consumption.
The dataset is provided with the MM package:

```r
R> data("danaher")
R> danaher
     eggs bacon 0 1 2 3 4
0   254 115 42 13 6
1    34  29 16  6  1
2     8   8  3  3  1
3     0   0  4  1  1
4     1   1  1  0  0
```

Thus 16 households purchased eggs twice and bacon once (this is Danaher and Hardie’s Table 1). The purchases of eggs and bacon are not independent\(^3\) and we suggest fitting this data to the distribution given in Equation 4; here \(m_1 = m_2 = 4\). The Poisson device of Lindsey and Mersch is again applicable:

```r
R> fit <- Lindsey_MB(danaher)
R> summary(fit)
```

Call:
```
 glm(formula = d ~ ., family = poisson, data = x, offset = Off)
```

Deviance Residuals:
```
  Min      1Q  Median      3Q     Max
-1.8786  -0.5751   0.0381   0.5032   1.5891
```

Coefficients:
```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.5117     0.0606 90.922   <2e-16 
bacon       -1.6528    0.1667 -9.920   <2e-16 
eggs       -1.0487     0.0810 -12.940   <2e-16 
`bacon:nbacon` -0.5149    0.0627  -8.220   <2e-16 
`eggs:neggs` -0.3555     0.0406  -8.760   <2e-16 
`bacon:eggs`  0.3006     0.0598   5.030  5.0e-07 
```

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3046.440  on 24  degrees of freedom
Residual deviance: 18.666  on 19  degrees of freedom
AIC: 108.3

Number of Fisher Scoring iterations: 5

---

\(^3\)Fisher’s exact test gives a \(p\) value of \(11 \times 10^{-6}\) with \(10^6\) replicates.
The MM distribution

and `glm()` gives a good fit in the sense that the residual deviance of 18.666 is compatible with its asymptotic null distribution $\chi^2_{19}$.

The `bacon:eggs` coefficient (ie $\log \hat{\phi} = 0.3006$) gives $\hat{\phi} = e^{0.3006} = 1.3507$, showing strong positive association.

We can now verify that the expected (marginal) number of egg purchases and bacon purchases under the ML distribution match the observed. The first step is to create $M$, the expected contingency matrix:

```r
R> M <- danaher
R> M[] <- fitted.values(fit)
R> M
```

<table>
<thead>
<tr>
<th></th>
<th>eggs 0</th>
<th>eggs 1</th>
<th>eggs 2</th>
<th>eggs 3</th>
<th>eggs 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>bacon 0</td>
<td>247.566</td>
<td>119.448</td>
<td>43.999</td>
<td>14.665</td>
<td>3.732</td>
</tr>
<tr>
<td>bacon 1</td>
<td>40.470</td>
<td>26.374</td>
<td>13.122</td>
<td>5.907</td>
<td>2.030</td>
</tr>
<tr>
<td>bacon 3</td>
<td>1.484</td>
<td>1.765</td>
<td>1.602</td>
<td>1.315</td>
<td>0.825</td>
</tr>
<tr>
<td>bacon 4</td>
<td>0.333</td>
<td>0.535</td>
<td>0.656</td>
<td>0.727</td>
<td>0.616</td>
</tr>
</tbody>
</table>

Then we may verify, for example, that the fitted sum of bacon purchases matches its observed value:

```r
R> bacon <- slice.index(danaher,1)
R> eggs   <- slice.index(danaher,2)
R> sum(bacon*eggs)  # Observed number of bacon purchases
[1] 710
R> sum(bacon*M)     # Expectation; matches
[1] 710
```

As a final check, we can verify that the sample covariance matches the distribution’s covariance at the MLE:

```r
R> sum(bacon*eggs*eggs) /N - sum(bacon*eggs)*sum(eggs*eggs) /N^2
[1] 0.144
R> sum(bacon*eggs*M)   /N - sum(bacon*M   )*sum(eggs*M   ) /N^2
[1] 0.144
```
4. Suggestions for further work

The multiplicative multinomial is readily generalized to a distribution with $2^k - 1$ parameters:

$$P(y_1, \ldots, y_k) = \left( \frac{y}{y_1 \ldots y_k} \right) \prod_{S \subseteq [k]} (\Theta_S)^{\Pi_{i \in S} y_i}$$  \hspace{1cm} (18)

where $[k] = \{1, 2, \ldots, k\}$ is the set of all strictly positive integers not exceeding $k$. Here, the parameters are indexed by a subset of $[k]$; it is interesting to note that $\Theta_S$ formally represents the normalization constant $C$. In this notation, Equation 3 becomes

$$\prod_{S \subseteq [k]} \left( \frac{y}{y_1 \ldots y_k} \right) (\Theta_S)^{\Pi_{i \in S} y_i}.$$  \hspace{1cm} (19)

Equation 18 leads to a distribution of the exponential family type; but interpretation of the parameters is difficult, and further work would be needed to establish the usefulness of this extension.

Further, Equation 4 generalizes to

$$\prod_{i=1}^{t} \left( \frac{m_i}{x_i z_i} \right) p_i^{x_i} q_i^{z_i} \phi^{x_i z_i}.$$  \hspace{1cm} (20)

It is possible to generalize the equations in a slightly different way. Consider an $r \times c$ matrix $n$ with entries $n_{ij}$ and fixed marginal totals. Now suppose that each row of $n$ comprises independent observations from a multinomial distribution with probabilities $p_1, \ldots, p_r$, and likewise the columns are multinomial $q_1, \ldots, q_c$. This is the null of Fisher’s exact test. Then one natural probability measure would be

$$\mathcal{P}(n) = \frac{1}{C} \cdot \frac{\prod_{1 \leq i_1 < i_2 \leq r} \theta_{i_1 i_2}^{n_{i_1, i_2}} \prod_{1 \leq j_1 < j_2 \leq c} \phi_{j_1 j_2}^{n_{j_1, j_2}}}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}$$  \hspace{1cm} (21)

(the fixed known row- and column- sums mean that $p_i$ and $q_j$, and the marginal multinomial terms, are absorbed into the normalizing constant $C$). With a slight abuse of notation this can be written

$$\mathcal{P}(n) = \frac{1}{C} \cdot \frac{\prod \Theta^{n^T} \prod \Phi^{n^T} n!}{\prod n^!}$$  \hspace{1cm} (22)

where $\Theta$ governs row-wise departures from multinomial and $\Phi$ governs column-wise departures; there are a total of $r(r - 1)/2 + c(c - 1)/2$ free parameters.

The Poisson device of Lindsey and Mersch is again applicable, with the difference that the enumeration carried out by compositions() is replaced by enumeration of contingency tables with the correct marginal totals: Function allboards() of the aylmer package (West and Hankin 2008). A simple example is given under help(sweets).
As pointed out by an anonymous referee, it might be possible to extend either or both of the new distributions to the context of regression on covariates.

5. Conclusions

In this paper, we considered natural generalizations of the multiplicative binomial distribution to the multivariate case. The resulting distributions have a number of desirable features, including a more precise control over the variance than the multinomial, and a straightforward interpretation in terms of contingency tables.

The distributions belong to the exponential family; this makes fast calculation of maximum likelihood estimates possible using generalized linear model techniques; in R idiom, the \texttt{glm()} function is used.

Novel analyses are presented on data drawn from the fields of social science and palaeoclimatology.

Acknowledgements

We thank Professor Gianfranco Lovison for helpful discussions; the AE and referees of the JSS review process provided constructive suggestions which improved the quality of the manuscript and associated software.

References


**Affiliation:**
P. M. E. Altham
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road Cambridge CB3 0WB
United Kingdom
E-mail: P.M.E.Altham@statslab.cam.ac.uk

Robin K. S. Hankin
Auckland University of Technology
14 Wakefield Street
Auckland 1010
New Zealand
E-mail: hankin.robin@gmail.com