Journal of Data Science **5**(2007), 413-423

Dirichlet-multinomial Model with Varying Response Rates over Time

Jeffrey R. Wilson and Grace S. C. Chen
Arizona State University

Abstract: It is believed that overdispersion or extravariation as often referred is present more in survey data due to the existence of heterogeneity among and between the units. One approach to address such a phenomenon is to use a generalized Dirichlet-multinomial model. In its application the generalized Dirichlet-multinomial model assumes that the clusters are of equal sizes and the number of clusters remains the same from time to time. In practice this may rarely ever be the case when clusters are observed over time. In this paper the random variability and the varying response rates are accounted for in the model. This requires modeling another level of variation. In effect, this can be considered a hierarchical model that allows varying response rates in the presence of overdispersed multinomial data. The model and its applicability are demonstrated through an illustrative application to a subset of the well known High School and Beyond survey data.

Key words: Dirichlet-multinomial, extravariation, overdispersion.

1. Introduction

The research related to analyzing cross-classified data in the presence of overdispersion has received a great deal of attention. Recent research includes (Goodman 1991a; Andersen 1991; Benzecri 1991; Baccini *et al.* 1991; Clogg and Rao 1991; Cox 1991; Haberman 1991; Goodman 1991b). The broad applications of cross-classified data (Wilson and Koehler, 1991; Bedrick, 1983; Ennis and Bai, 1999) give rise to the need for continued research. A test of relationships among variables through a generalized linear model, with overdispersion arising naturally through complex survey data, is the focus of this paper. A model approach based on mixture distributions similar to what was utilized in the generalized Dirichlet-multinomial distribution is presented. It provides, in part, an expansion of the use of the distribution used in the generalized Dirichlet-multinomial model as developed by Wilson and Koehler (1991) for surveys with equal and

fixed response rates. Within a subpopulation there is a guarantee that the clusters will always be of equal sizes. As such it is important to model the unknown cluster sizes.

Some researchers believe that this "variance discrepancy" or the fact that the variance of sample data estimates often exceeds those mandated by a predetermined sampling distribution is usually the norm rather than the exception. The sample data estimates are often referred to as being overdispersed. Efron (1986) suggested that such a phenomenon may be caused by clumped sampling. Cox (1983) pointed out that overdispersion in general has two effects. One is that summary statistics have a larger variance than anticipated under the simple model. The second effect is a possible loss of efficiency in using statistics appropriate for the single-parameter family.

Wilson and Koehler (1991) considered the variation among proportions similar to the approach to randomized block design with random components for interval-level data. They presented a model that allowed for the analysis of variation among replicates and among units for a given replicate. Ignoring either level of variation leads to underestimation of the true standard errors of estimated proportions. In particular, they used the Dirichlet-multinomial distribution to incorporate the two types of variation.

Wilson and Koehler (1991) extended the Dirichlet-multinomial model by using a second Dirichlet prior to account for a second random component. A model with similar properties was obtained by applying a Dirichlet prior to the generalized multinomial model as proposed by Tallis (1962) for dependent multinomial random vectors. It is their use of the nested model that is further examined. Nested factors are frequently encountered in longitudinal and repeated-measures studies in which each subject or experimental unit provides responses at several time periods, possibly under different sets of circumstances. The model considered in this paper can be viewed as multivariate extensions of the beta-binomial and beta-correlated binomial models considered by Kupper and Haseman (1978) and Crowder (1978) for binary data. Paul (1987) considered a modification of the beta-correlated binomial model as a means of analyzing affected fetuses in litters of live fetuses. Anderson (1988) reviewed some of the models found useful in the analysis of overdispersed binomial data, including (a) models with a heterogeneity factor, (b) random-effects models, (c) binomial-mixture models, and (d) models for longitudinal data. The model considered in this paper allows flexibility of the generalization of some of the random-effects models considered by Anderson and are suitable for longitudinal data with a possible correlation induced by a repeated-measures design.

A generalized Dirichlet-multinomial model within varying response rates is presented in Section 2. Models as considered in this paper are also useful in

analyzing economic data, social-survey data, and some types of business data. Other applicable kinds of data include tracking of election results over time (the focus groups are certainly of different sizes), the buying habits of consumers (the family of consumers are not necessarily of the same size), and television ratings during prime time weeks (the number of households in the cluster may be of different size). In section 3, data obtained from the well known *High School and Beyond* survey were analyzed using the Dirichlet model in this paper. The data, though useful for illustrative purposes, are sufficient to accurately use the asymptotic properties of some of the test statistics. These illustrative data are used not for educational policy but for the purpose of demonstrating the usefulness and applicability of the model.

2. The Dirichlet-multinomial Model

Consider the following information for a number of clusters observed over period T. The rows represents the responses over I categories and the columns represents the J clusters. Suppose that, for a given period t, J clusters of unequal sizes r_j are randomly selected from a population with the vectors of proportions for cluster j, $\mathbf{p}_j = (p_{1j}, \ldots, p_{Ij})$.

The Dirichlet model specifies that the j-th cluster $(J = 1, \ldots, J)$ of units in a period of time t are such that over a given period of time the J clusters are independent and random. Assume for each of the j-th cluster, the probability of certain responses over I categories has probability vector, $\mathbf{p}_j = (p_{1j}, \ldots, p_{Ij})'$, $(j = 1, \ldots, J)$ is fixed over time period and provides a varying sample of r_j responses resulting in a multinomial vector of counts X_{jt}, of dimension I. The total number of responses over the I categories, r_j made by the j-th cluster follows multinomial distribution with parameters r_j and vector $(p_{1j}, \ldots, p_{Ij})'$. Further assume the vector of probabilities p_j varies among clusters and follow a Dirichlet distribution with parameter β and vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_I)'$. Thus these two assumptions lead to a mixture distribution, for responses among clusters, is given by a mixture of multinomial with a Dirichlet distribution. Thus the model assumes that the response for all clusters, conditional on the response probabilities have a Dirichlet multinomial distribution,

$$f_{\beta_1,\ldots,\beta_I}(r_1, \ldots, r_J) = \frac{\Gamma(\sum_{j=1}^{J} \beta_j) k!}{\Gamma(\sum_{j=1}^{J} \beta_j) + k} \prod_{j=1}^{J} \frac{\Gamma(\beta_j + r_j)}{r_j! \Gamma(\beta_j)}$$

Across period t, assume that the conditional distribution of the probability vector for the t-th period $\boldsymbol{\pi}_t$ is distributed with Dirichlet distribution such that over time, π_t is $Dir(\alpha, \theta)$ where α is an unknown constant and θ is an unknown vector. The model may be displayed as $X_{jt}|r_j, p_{jt}$ multinomial (r_j, p_{jt}) for $j =$

$1, \ldots, J; t = 1, \ldots, T$ and $p_{jt}|\beta, \boldsymbol{\pi}_t$ are approximately independent $Dir(\beta, \boldsymbol{\pi}_t)$ such that $\boldsymbol{\pi}_t|\alpha, \theta$ are approximately independent $Dir(\alpha, \theta)$ for $j = 1, \ldots, J; t = 1, \ldots, T$.

Distribution of Cluster Sizes

Further we assume that the number of responses from a cluster (or the cluster sizes) is such that successive responses among clusters are assumed to be independent with a constant mean μ_j, and behave as if random in chosen time-period. The number of responses r_j made in each of succession of equal non-overlapping periods of relative length then follows a Poisson distribution with mean $\mu_j T$. Also the mean response rates vary among clusters and follow a gamma distribution with density function $e^{-\mu\gamma/\beta}/\{(\beta/\gamma)^\gamma \Gamma(\gamma)\}$.

This model accounts for the extra variation due to time and due to the sampled units. Also it allows the cluster sizes r_j, the number of units per cluster to be modeled as it varies. Thus both the number of units per cluster denoted by r_j, and the underlying probability vectors π_j are allowed to vary. In particular the Dirichlet model is used to incorporate two sets of variability, and the Poisson distribution is used to account for variation among the cluster sizes within a given time period. Ignoring either level of variation might lead to underestimation of the true standard errors of estimated proportions.

For period t, let the cluster sizes r_j have a distribution that of the Poisson, which arises when the units within the clusters are subject to certain probabilities. Thus both the cluster sample sizes and the probability vector associated with each cluster are allowed to vary resulting in a mixture of distributions.

Moments

The first and second moments of the mixture distribution can be easily found. Under the Dirichlet model the covariance matrix for the conditional distribution of the vector of proportions across time $\hat{\boldsymbol{\pi}} = T^{-1}R^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J} w_j X_{jt}$ where $w_j = r_j R^{-1}$, $X_{jt} = (X_{1jt}, \ldots, X_{Ijt})'$ for fixed $R = \sum_{j=1}^{J} r_j$ has mean vector $E(\hat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$ and covariance matrix

$$V(\hat{\boldsymbol{\pi}}) = [(nR)c\{1 + \rho(J-1)\}]^{-1} \sum_{j=1}^{J} w_j^2 V_{\pi_j}$$

where c is part of an overdispersion factor, T is the number of time periods, ρ is a measure of the correlation between units within clusters, $V_{\pi_j} = diag(\pi_{jt}) - \pi_{jt}\pi_{jt}'$, where $diag(\pi_{jt})$ is the diagonal matrix with elements that of the vector, $\boldsymbol{\pi}_{jt} = (\pi_{1jt}, \ldots, \pi_{Ijt})'$. Then the unconditional first and second moments for the

distribution are

$$E(X_t) = nR\sum_{j=1}^{J} r_j \boldsymbol{\pi}_j \quad \text{and}$$

$$V(X_t) = nr\left[b\sum_j \{w_j V_{\pi_j} + n(r_j - J)(J-1)J^{-1}\boldsymbol{\pi}_j\boldsymbol{\pi}_j'\}\right],$$

where $b = c\{diag(\boldsymbol{\pi}_j) - [1 - n(r_j - J)(J-1)b^{-1}J^{-1}]\boldsymbol{\pi}_j\boldsymbol{\pi}_j'\}$. For instance, when $r_j = 1$ for all j, we have $R = \sum_j r_j = J$, and therefore, $V(X_t) = nJ[bV_\pi]$. The variance is a linear combination of the variation due to the Dirichlet-multinomial and the variation due to the unequal sample sizes. Thus, when the variance due to the unequal sample sizes is small there is little difference between the Dirichlet-multinomial and the proposed modeling of the random cluster sizes. Certainly there will be no difference between the model variances when there is one unit per cluster.

Test Statistic

Consider comparing the vector of proportions for two subpopulations. Similar to the assumptions made for the general Dirichlet-multinomial model, Wilson and Koehler (1991) using the appropriate covariance matrix, and given a consistent estimator for c and for the expression $\{1 + \rho(J-1)\}$, asymptotic chi-square tests involving sufficiently smooth functions of π (this can be a linear combination of the vectors of proportions for the different subpopulations) can be obtained as Wald statistics,

$$X^2_{EGD} = \{g(\hat{\pi}) - g(\pi)\}'(D\hat{V}D')^{-}\{g(\hat{\pi}) - g(\pi)\}, \tag{2.1}$$

where $(D\hat{V}D')^{-}$ is a generalized inverse of $(D\hat{V}D')$, $\hat{V}$ is a consistent estimate of V, and D is the matrix of first derivatives of the function g. The degrees of freedom for the asymptotic chi-square distribution of X^2_{EGD}, is the rank of $D\hat{C}D'$.

Model Checking

When using the generalized Dirichlet model with varying and unequal cluster sizes there are three basic assumptions: a) the correlations between units within clusters, are constant for any two units; b) the X_j's, $j = 1, \ldots, J$, are identically multinomial distributed; and c) the cluster sizes are distributed as a Poisson distribution. Test statistics were presented by Wilson and Koehler (1991) to assess the validity of the first two assumptions. Wilson (1986) and Koehler and Wilson (1986) gave large sample tests for the covariance structure associated with

the Dirichlet-multinomial model. Methods of estimating c and the expression $\{1+\rho(J-1)\}$, were given by Wilson and Koehler (1991). One set of estimators can be obtained by constructing an I by J table (where I denotes the dimension of π and J denotes the number of clusters) and a I by n table, where n denotes the number of time periods. For each table, obtain the Pearson chi-square test statistic, $X^2_{(IJ)}$ and $X^2_{(In)}$ respectively, for testing independence in a two-way contingency table. Then estimates of $\{1+\rho(J-1)\}$ and c are $X^2_{(IJ)}/(I-1)(J-1)$ and $X^2_{(In)}/(I-1)(n-1)$ respectively.

A subset of the High School and Beyond data based on three time periods $T=3$ (the 1980, 1982, and 1984); clusters of students from one population, public school and others from a second population, Catholic schools with $I=3$ possible responses (no TV on a weekday, one to five hours on a weekday, more than 5 hours on a weekday) was used. The data can be arranged in a $3\times 2\times 3$ contingency table. For the analysis of these data, we let H_{21} denote the hypothesis for the comparison of vectors of proportions between two subpopulations of schools.

3. Illustrative Example

A numerical example demonstrates the usefulness and applicability of the generalized Dirichlet-multinomial with varying cluster sizes is the High School and Beyond (HSB) longitudinal survey study in the United States, 1980 to 1986. It was designed for studying the transition of high school students from secondary school system to the postsecondary system. The High School and Beyond survey of adolescents provides data on the daily television viewing habits of adolescents for 1980, and follow up for years 1982 and 1984. These data were collected from a complex survey-sampling scheme. The information about television viewing habits was obtained from the original sample of students in 1980, 1982 and 1984. High School and Beyond is part of a larger program of national longitudinal surveys of American youth initiated by the National Center of Education Statistics with the cohort of 1972 high school seniors. In this demonstration the aim is to test the hypothesis that the daily viewing habits of students from the Catholic schools differ from the public schools. It is essential that the use of any test statistic to examine such a hypothesis incorporates the extra variation present in the data. Some educators believe that the amount of homework and the greater demands placed on students from Catholic schools will reduce the time available for students to view television. This in turns may reduce the amount of unpopular programs they are exposed to. The use of these data portrays the applications of modeling data in the presence of overdispersion.

Data taken from the High School and Beyond Survey were obtained primarily by means of questionnaires and tests. The students were selected through a two-

stage probability sample with schools as the first stage units and students within schools as the second stage units (Frankel, Kohnke, Buonanno and Tourangeau, 1981). Two types of schools (clusters) were used for analysis in this paper; the regular public schools and the regular Catholic schools. The combined data for 1980, 1982 and 1984 are summarized in Table 1. The histogram of the data by type of schools is in Figure 1.

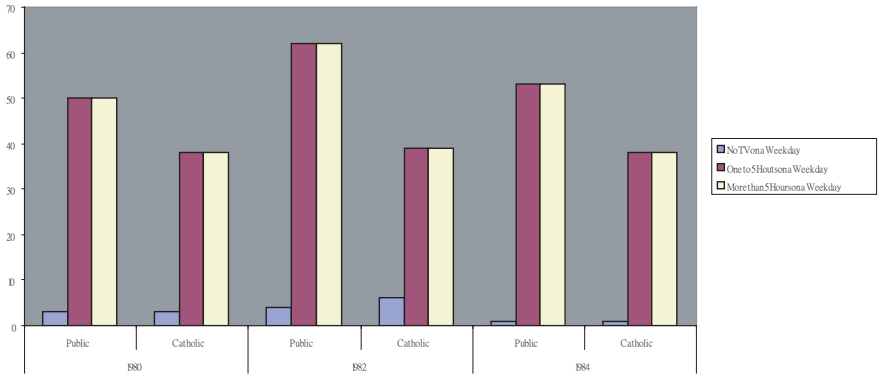

Figure 1: Adolescents' weekly TV viewing habit for 1980, 1982 and 1984 by type of school in "High School and Beyond"

Table 1: Adolescents' weekly TV viewing habit for 1980, 1982, and 1984 by type of school in "high school and beyond

Year	School	No TV on a weekday	One to 5 Hours on a Weekday	More than 5 Hours on a Weekday
1980	Public	3	50	32
	Catholic	3	38	9
1982	Public	4	62	19
	Catholic	6	39	5
1984	Public	1	53	31
	Catholic	1	38	11

3.1 Results

The possible responses to the television viewing question are: less than one hour per day, one to five hours per day, and more than five hours per day. The estimated vector of proportions for the subpopulation of cluster for the public ($h = 1$) and for the Catholic schools ($h = 2$) over the three periods are $\hat{\boldsymbol{\pi}}_1 =$

$(0.0310, 0.647, 0.322)'$ and $\hat{\boldsymbol{\pi}}_2 = (0.067, 0.767, 0.166)'$ respectively. There is some indication that students from the regular public schools viewed more television than their counterparts from the Catholic schools. However, for this comparison to be tested since the extra variation present may be enough to negate such apparent differences.

The usual Pearson chi-square test for checking homogeneity across school has a value of 13.18 on 2 degrees of freedom with an observed significance level of .001. The value of the test statistic is substantially reduced when the effects of cluster sampling, repeated measurement, and sample size variation are taken into consideration. The proposed Dirichlet model with unequal cluster sizes is employed to take into account effects due to cluster sampling, repeated measurement, and sample size variation.

There are 30 clusters with varying sizes from the regular public schools and 21 clusters also with varying sizes from the Catholic schools. The size of the clusters varies between 1 and 20 for public schools and between 1 and 8 for the Catholic schools. There was no reason to believe that the r_j's are not distributed as a Poisson. The model assumptions were checked using the methods of Wilson and Koehler (1991). These assumptions were sufficiently satisfied. The details of these assumptions were omitted here for sake of brevity.

The covariance matrices for subpopulation of public and Catholic schools $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ under the Dirichlet model with unequal cluster sizes are

$$V(\hat{\boldsymbol{\pi}}_1) = \begin{bmatrix} 0.86 & 12.54 & 6.23 \\ 12.54 & 263.70 & 128.50 \\ 6.23 & 128.50 & 66.41 \end{bmatrix} \times 10^{-3}$$

and

$$V(\hat{\boldsymbol{\pi}}_2) = \begin{bmatrix} 3.17 & 27.65 & 6.01 \\ 27.65 & 326.73 & 69.11 \\ 6.01 & 69.11 & 16.91 \end{bmatrix} \times 10^{-3}.$$

The overall probability vector $\hat{\boldsymbol{\pi}}_0 = (0.046, 0.696, 0.258)'$ based on weights of $\alpha_1 = 0.59$ and $\alpha_2 = 0.41$ in the linear combination of $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$. The weights were chosen to be inversely proportional to the relative size of the multiplier in $V(\hat{\boldsymbol{\pi}}_h)$ thus $\alpha_h, h = 1, 2$ is proportional to $n^{-1}m_h^{-1}c_h\{1 + \rho_k(J_h - 1)\}$. The parameter estimates $\hat{c}_1 = 1.156, \hat{c}_2 = 1.130, \{1 + \hat{\rho}(J_1 - 1)\} = 1.749$ and $\{1 + \hat{\rho}_2(J_2 - 1)\} = 1.524$.

To test for homogeneity among the subpopulations, define $g(\hat{\boldsymbol{\pi}}) = \hat{\boldsymbol{\pi}}_h - \hat{\boldsymbol{\pi}}_0$ with covariance matrix $V(\hat{\boldsymbol{\pi}}_h - \hat{\boldsymbol{\pi}}_0) = V(\hat{\boldsymbol{\pi}}_h) - 2\alpha_h V(\hat{\boldsymbol{\pi}}_h) + \sum_{k=1}^{2} \alpha_k^2 V(\hat{\boldsymbol{\pi}}_k)$ (Koehler and Wilson, 1986). Thus, for 30 clusters in the public schools and 21 clusters in the private schools the test statistic X^2_{EGD} from equation (2.1) for

homogeneity has a value of 2.38 with 2 degrees of freedom. Therefore, the apparent differences in viewing habits between public and Catholic school students are not significant. From the interval coverage there seems to be no difference in the television viewing habits based on type of school.

4. Conclusions

A modification to the Dirichlet-multinomial model with unequal cluster sizes can be very useful for analyzing survey data. It takes into account a measure of any clustering effect present in the data. In addition, it incorporates any variation across time and variation between units for a given time period. It provides a means of addressing the problem of extra variation caused by the random cluster sizes. Some previous works suggest an approximate technique for unequal clusters but the Dirichlet model can be expanded to unequal cluster sizes uses the Poisson distribution to address the unequal cluster sizes directly. The proposed model-based method brings together three strategies for modeling complex designs, namely, i) Dirichlet mixing for clustering effects, ii) Dirichlet mixing for time dependencies as well, and iii) Poisson distribution for varying cluster sizes.

References

Andersen, E. B. (1991). Measures, models, and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1111-1112.

Baccini, A, Caussunus, H, de (1991). Measures, models and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1115-1117.

Bedrick, E. J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika* **70**, 591-595.

Benzecri, J. P. (1991). Measures, models, and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1112-1114.

Clogg, C. C. and Rao, C. R. (1991). Measures, models, and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1118-1121.

Cox, D. R. (1991). Measures, models, and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1121-1211.

Ennis, D. M. and Bi, J. (1999). The Dirichlet-multinomial model: Accounting for intertrial variation in replicated ratings. *Journal of Sensory Studies* **14**, 321-345.

Frankel, M., Kohnke, L., Buonanno, D. and Tourangeau, R. (1981). *HS&B Base Sample Design Year Report*, NORC.

Gange, S.J., Jacobson, L. P. and Munoz, A. (1996). Parametric models to assess tracking of categorical high-risk sexual behaviour for HIV infection. *Statistics in Medicine* **15**, 2323-2336.

Goodman, L. A. (1991a). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association* **86**, 1085-1110.

Goodman, L. A. (1991b). Measures, models, and graphical displays in the analysis of cross-classified data (with comment and a rejoinder by the author). *Journal of the American Statistical Association* **86**, 1124-1132.

Haberman, S. J. (1991). Measures, models, and graphical displays in the analysis of cross-classified data: comment. *Journal of the American Statistical Association* **86**, 1121-1123.

Koehler, K. J. and Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communication in Statistics* **A15**, 2977-2990.

Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effects in the Dirichlet-multinomial model. *Communications in Statistics* **A15**, 1235-1249.

Wilson, J. R. and Koehler, J. K. (1991). Hierarchical models for cross-classified data overdispersed multinomial data. *Journal of Business and Economic Statistics* **9**, 103-110.

Received April 1, 2005; accepted February 17, 2006.

Jeffrey R. Wilson
Department of Economics
Arizona State University
Tempe, AZ 85287-4506, USA
jeffrey.wilson@asu.edu

Grace S. C. Chen
Department of Mathematics and Statistics, and
Department of Biomedical Informatics
Arizona State University
Tempe AZ 85287-1804, USA
scchen@math.asu.edu

Also:
Department of Statistics
National Cheng Kung Universsity
Tainan, Taiwan