The Annals of Statistics
2004, Vol. 32, No. 1, 340–383

MULTINOMIAL–POISSON HOMOGENEOUS MODELS FOR CONTINGENCY TABLES

BY JOSEPH B. LANG

University of Iowa

A unified approach to maximum likelihood inference for a broad, new class of contingency table models is presented. The model class comprises multinomial–Poisson homogeneous (MPH) models, which can be characterized by an independent sampling plan and a system of homogeneous constraints, $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, where $\mathbf{m}$ is the vector of expected table counts. Maximum likelihood (ML) fitting and large-sample inference for MPH models are described. The MPH models are partitioned into well-defined equivalence classes and explicit comparisons of the large-sample behaviors of ML estimators of equivalent models are given. The equivalence theory not only unifies a large collection of previously known results, it also leads to useful generalizations and many new results. The practical, computational implication is that ML fit results for any particular MPH model can be obtained directly from the ML fit results for any conveniently chosen equivalent model. Issues of hypothesis testability and parameter estimability are also addressed. To illustrate, an example based on statistics journal citation patterns is given for which the data can be used to test the hypothesis that a certain model holds, but they cannot be used to estimate any of that model's parameters.

1. Introduction. We present a unified theory of maximum likelihood (ML) inference for a broad, new class of contingency table models. This class comprises multinomial–Poisson homogeneous (MPH) models, which can be characterized by an independent sampling plan and a system of homogeneous constraints, $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, where $\mathbf{m}$ is the vector of expected table counts. This article considers a wide variety of sampling plans that lead to sufficient counts composed of independent blocks of Poisson and multinomial variables. The constraint function $\mathbf{h}$ is sufficiently smooth and homogeneous, relative to the sampling plan, in a sense akin to Euler's homogeneous functions [see "Euler's theorem for homogeneous functions" as presented in Apostol (1974), pages 364 and 365]. Maximum likelihood fitting and large-sample inference for the broad class of MPH models is described. The theoretical development is based largely on the approach of Aitchison and Silvey (1958), who viewed a model as a system of constraints.

The current article shows that both ML estimation theory and ML fitting are straightforward for a much broader class of models than previously assumed.

Received May 2001; revised May 2002.

AMS 2000 subject classifications. Primary 62H17; secondary 62E20, 62H12, 62H15.

Key words and phrases. Approximate normality, categorical data, equivalent models, estimability, homogeneous constraint, homogeneous statistic, large-sample inference, restricted maximum likelihood, sampling plan, testability.

The collection of models that are useful for analyzing contingency, or cross-classification, tables is much richer than the class of log-linear/logit models, a class that is most directly useful for describing the association among the classification variables. Often questions of interest concern marginal distributions or other many-to-one functions of the contingency table probabilities. As simple examples, consider models of marginal homogeneity [see Kullback (1971), Agresti (1990), page 390], mean and marginal mean response models [see Agresti (1990), page 333] and linear predictor models such as those considered in Grizzle, Starmer and Koch (1969), Lang and Agresti (1994) or Glonek and McCullagh (1995). In general, these models are not members of the log-linear/logit family, but they are members of the broader class of MPH models. Similarly, models for tables with given marginal distributions [Ireland and Kullback (1968)], models of pairwise independence [Haber (1986)] and models for tables based on both completely and partially cross-classified responses [Chen and Fienberg (1974)] generally are not members of the log-linear/logit family, but they are typically MPH models. This article shows that ML estimation is straightforward and an attractive alternative to weighted least squares for the aforementioned "nonstandard" (i.e., nonlog-linear/logit) models.

In contingency table analyses, it is common practice to exploit equivalences between certain models. As a simple example, when faced with fitting a product-multinomial log-linear model, one might choose to fit the "equivalent" Poisson log-linear model for convenience. The literature is full of examples where the relationship, or "equivalence," between multinomial and Poisson models is exploited [e.g., Palmgren (1981), Lyons and Hutcheson (1986), Cormack and Jupp (1991), Chambers and Welsh (1993), Baker (1994), Matthews and Morris (1995), Lang (1996a), Lipsitz, Parzen and Molenberghs (1998), Fienberg (2000) and Bergsma and Rudas (2002)]. This article extends and formalizes the notion of model equivalence. We show that MPH models can be partitioned into well-defined equivalence classes. We give explicit comparisons of the large-sample behaviors of ML estimators of equivalent models. The practical implication is that ML fit results for any particular MPH model can be directly obtained from the ML fit results for any conveniently chosen equivalent model. This has obvious computational utility.

We address questions as to whether collected data can be used to test a hypothesis and/or estimate an estimand of interest. Section 10 gives a simple example where the collected data can be used to test the hypothesis that a certain model holds, but they cannot be used to estimate any of that model's parameters. By way of example, Section 10 also addresses estimability and testability for the logit model under retrospective, or case-control, sampling.

This article highlights the fact that the choice of model constraints need not dictate the choice of inferential distribution. For example, logit models typically have been used in conjunction with a product-multinomial inferential distribution, even when full-multinomial or full-Poisson sampling was actually

used. Unfortunately, this conditional-inference approach precludes estimation of the expected counts or the underlying joint distribution subject to the logit constraints on the conditional distributions. This article shows that this problem can be avoided. For example, the ML fit results for the actual (full-multinomial or full-Poisson) data model with constraints specified in terms of logit constraints on the conditional distributions can be obtained directly from the fit results of the artificial, but equivalent, product-multinomial logit model.

The theory herein serves not only to formally unify a large collection of well-established contingency table model results, it also generalizes these results and presents many new ones. As examples of existing results, Birch (1963) gave general conditions under which two contingency table models give numerically the same expected count estimates. Haberman (1974) presented a rich collection of related numerical and asymptotic results in the special log-linear model setting. Using different approaches, Andersen (1974), Palmgren (1981) and Christensen [(1990), Chapter XV] also gave related log-linear model results. Baker (1994) compared inferences for Poisson and product-multinomial models that lend themselves to the multinomial-to-Poisson transformation. Bergsma (1997) used homogeneity properties of a special class of multinomial and Poisson marginal models to derive certain numerical and approximation results. Lang, McDonald and Smith (1999) gave related results for a special class of generalized log-linear models. With the exception of Haberman's results for certain conditional-Poisson (e.g., hypergeometric) log-linear models, all of these results are obtained as special cases of the general theory outlined herein.

This article is organized as follows. Section 2 introduces notation and gives preliminary definitions. Section 3 describes the sampling plans and multinomial–Poisson data models. Section 4 gives an example, using statistics journal citation data, that serves to illustrate the notation of the first three sections and provides motivation for the subsequent sections. Homogeneous constraint functions are described and their properties are explored in Section 5. Multinomial–Poisson homogeneous models are introduced in Section 6. Section 7 gives numerical, asymptotic and approximation maximum likelihood results for MPH models. Section 8 introduces a formal definition of model equivalence and explicitly compares maximum likelihood fit results for equivalent models. The useful class of $\mathbf{Z}$-homogeneous statistics is introduced in Section 9. Section 10 addresses issues of estimability and testability. Using the journal citation data of Section 4 as an example, Section 11 describes the numerical computation of ML fit results. Section 12 gives a brief discussion. Selected proofs are given in the Appendix.

2. Introductory notation and definitions. Conventional notation will be used. The symbol $\mathbf{D}^{\alpha}(\mathbf{m})$ (or $\operatorname{diag}^{\alpha}\{m_i,\ i = 1, \ldots, c\}$) represents the αth power, where α is any real number, of the diagonal matrix with the components in $\mathbf{m}$ on the diagonal. Functions that typically operate on scalars, like powers and logarithms, act on vectors in a componentwise fashion. For example, $\log \mathbf{m}$ is

defined as $(\log m_1, \ldots, \log m_c)^T$, where the T represents the transpose. With this componentwise convention, the representation $\mathbf{D}^\alpha(\mathbf{m}) = \mathbf{D}(\mathbf{m}^\alpha)$ is well defined. Componentwise multiplication and division of two compatible vectors $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ will be denoted $\boldsymbol{\delta} \cdot \boldsymbol{\gamma}$ and $\boldsymbol{\delta}/\boldsymbol{\gamma}$.

Projection notation will be used; it is simplest to define this notation via an example. Let $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ and let $\phi = (1, 3, 4)$ be an ordered subset of $\{1, 2, 3, 4\}$. Then the subvector $\mathbf{x}_\phi = (x_1, x_3, x_4)^T$. Moreover, the value $\mathbf{x}_{\phi, j}$ is the jth component in the vector $\mathbf{x}_\phi$. For example, $\mathbf{x}_{\phi,2} = x_3$.

To denote a sum over a certain dimension of an array, a $+$ sign will be used. For example, a matrix $\mathbf{Z}$ with components Z_{ik} has kth column sum equal to Z_{+k}. The direct sum of matrices $B_1, \ldots, B_b$ will be denoted by $\bigoplus_{i=1}^b B_i$. The symbol $A \otimes B$ will represent the Kronecker (right direct) product of matrices A and B. Note that $B \oplus B = \mathbf{I}_2 \otimes B$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Finally, the indicator functional $I(\cdot)$ is defined as $I(E) = 1$ or 0 as the condition E is true or false.

The general results in this article are most conveniently stated using a coordinate-free representation of the categorical variables and distributions of interest. For example, a coordinate-based description of the joint distribution of the bivariate categorical random vector (A, B) might be expressed as $P_{ij} = P(A = i, B = j), i = 1, \ldots, I,\ j = 1, \ldots, J$. Using the coordinate-free approach, we would identify C with (A, B) and the events $(C = 1), (C = 2), \ldots, (C = c \equiv IJ)$ with $(A = 1, B = 1), (A = 1, B = 2), \ldots, (A = I, B = J)$. The probabilities $P(C = 1) = P_1, P(C = 2) = P_2, \ldots, P(C = c) = P_c$ are identified with $P_{11}, P_{12}, \ldots, P_{IJ}$.

Let C represent the (composite) categorical variable of interest. The unrestricted model for C can be written as

$$C \sim \mathbf{P} \qquad \text{where } \mathbf{1}_c^T \mathbf{P} = 1, \mathbf{P} > 0.$$

The probability vector $\mathbf{P} = (P_1, \ldots, P_c)^T$ has components defined as $P_i = P(C = i)$. Herein, we also consider restricted models for C of the form $C \sim \mathbf{P} \in \Theta$, where $\Theta = \{\mathbf{P} : \mathbf{1}_c^T \mathbf{P} = 1, \mathbf{P} > 0, \mathbf{h}(\mathbf{P}) = \mathbf{0}\}$. Because we can define the model $C \sim \mathbf{P} \in \Theta$ before data are collected or a sampling plan conceived, we call it a *predata model.* One of the primary objectives is to conduct inferences about the predata probabilities in $\mathbf{P}$. For example, we may wish to test the hypothesis that $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ or we may wish to estimate the function $\mathbf{S}(\mathbf{P})$.

3. Sampling plans and multinomial–Poisson data models. Inferences about predata probabilities $\mathbf{P}$ are to be based on data obtained from $K \geq 1$ independent random samples from conditional distributions of $\mathbf{P}$. The independent sampling plans considered in this article can be characterized by so-called population matrices and sampling constraints. Because of their importance, we carefully define both population and sampling constraint matrices, and give several useful properties.

3.1. *Population and sampling constraint matrices.*

DEFINITION 1. The matrix $\mathbf{Z}$, with components Z_{ik}, is a *population matrix* if (i) $Z_{ik} \in \{0, 1\}$, (ii) $Z_{i+} = 1$ and (iii) $Z_{+k} \geq 1$. Let $\mathcal{P}$ be the collection of population matrices.

For example, consider

$$\mathbf{Z}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \mathbf{Z}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{Z}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\mathbf{Z}_4 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{Z}_5 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{3.1}$$

The matrices $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ are all in $\mathcal{P}$; the matrices $\mathbf{Z}_4$ and $\mathbf{Z}_5$ are not.

DEFINITION 2. The matrix $\mathbf{Z}_F$ is a *sampling constraint matrix* with respect to population matrix $\mathbf{Z}$ if $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F$, where $\mathbf{Q}_F \in \mathbf{\Gamma} \equiv \{\mathbf{Q}: Q_{ij} \in \{0, 1\}, Q_{+j} = 1, Q_{i+} \leq 1\} \cup \{\mathbf{0}\}$. Let $\mathcal{S}(\mathbf{Z})$ be the collection of sampling constraint matrices with respect to $\mathbf{Z}$.

The requirement that $\mathbf{Q}_F$ fall in $\mathbf{\Gamma}$ implies that $\mathbf{Z}_F \in \mathcal{S}(\mathbf{Z})$ comprises a subset of columns of $\mathbf{Z}$ or it is the zero matrix. Note that $\mathbf{Z} \in \mathcal{S}(\mathbf{Z})$.

It will be convenient to define the complement of $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F$ as $\mathbf{Z}_R = \mathbf{Z}\mathbf{Q}_R$, where $\mathbf{Q}_R \in \mathbf{\Gamma}$ is the orthogonal complement of $\mathbf{Q}_F$. This implies that (i) $\mathbf{Z}_R$ is the collection of columns of $\mathbf{Z}$ not included in $\mathbf{Z}_F$ (or it is the zero matrix), (ii) the range space of $[\mathbf{Q}_F, \mathbf{Q}_R]$ is R^K, (iii) $\mathbf{Q}_F\mathbf{Q}_F^T + \mathbf{Q}_R\mathbf{Q}_R^T = \mathbf{I}$, (iv) $\mathbf{Q}_x\mathbf{Q}_x^T$ is diagonal, $x = F, R$, and (v) provided $\mathbf{Q}_x$ is not a zero matrix, $\mathbf{Q}_x^T\mathbf{Q}_x = \mathbf{I}$, $x = F, R$.

As an example, consider the population matrix $\mathbf{Z}_2$ of (3.1). The matrix $\mathbf{Z}_{2F} = [1, 1, 0, 0]^T$ is a member of $\mathcal{S}(\mathbf{Z}_2)$, because $\mathbf{Z}_{2F} = \mathbf{Z}\mathbf{Q}_F$ where $\mathbf{Q}_F = [1, 0]^T \in \mathbf{\Gamma}$. Here, the complement of $\mathbf{Z}_{2F}$ is $\mathbf{Z}_{2R} = \mathbf{Z}\mathbf{Q}_R = \mathbf{Z}[0, 1]^T = [0, 0, 1, 1]^T$.

The balance of the article makes use of the following useful properties of population and sampling constraint matrices. The proofs are straightforward and are omitted. The arbitrary vectors of positive numbers, $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ and $\mathbf{m}$, are assumed to be of compatible dimension for matrix multiplication. Let $\mathbf{Z} \in \mathcal{P}$, let $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F \in \mathcal{S}(\mathbf{Z})$ and let $\mathbf{Z}_R = \mathbf{Z}\mathbf{Q}_R \in \mathcal{S}(\mathbf{Z})$ be the complement of $\mathbf{Z}_F$.

S1. $\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{Z}_F = \mathbf{Z}\mathbf{D}(\boldsymbol{\delta})\mathbf{Q}_F$.

S2. $\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D}(\mathbf{Z}\boldsymbol{\delta}) = \mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{Z}_F\mathbf{Z}_F^T$.

S3. $\mathbf{D}^{\alpha}(\mathbf{Z}_F\boldsymbol{\delta}) = \mathbf{D}(\mathbf{Z}_F\boldsymbol{\delta}^{\alpha})$.

S4. $\mathbf{Z}_F^T\mathbf{D}(\mathbf{m})\mathbf{Z}_F = \mathbf{D}(\mathbf{Z}_F^T\mathbf{m})$.
S5. $\log(\mathbf{Z}\boldsymbol{\delta}) = \mathbf{Z}\log\boldsymbol{\delta}$.
S6. $\mathbf{D}(\mathbf{Z}_F\boldsymbol{\delta})\mathbf{D}(\mathbf{Z}_F\boldsymbol{\gamma}) = \mathbf{D}(\mathbf{Z}_F(\boldsymbol{\delta}\cdot\boldsymbol{\gamma}))$.
S7. $\mathbf{Z}_F\mathbf{Z}_F^T = \mathbf{Z}\mathbf{Z}^T - \mathbf{Z}_R\mathbf{Z}_R^T$.
S8. $\mathbf{Z}_F^T\mathbf{D}(\mathbf{m})\mathbf{Z}_R = \mathbf{0}$.
S9. $\mathbf{Z}_F\mathbf{1} + \mathbf{Z}_R\mathbf{1} = \mathbf{1}$.
S10. The set of all population matrices, $\mathcal{P}$, is closed under the operation of compatible multiplication. That is, if $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are population matrices of compatible dimensions, the product $\mathbf{Z}_1\mathbf{Z}_2$ is also a population matrix.

3.2. *The sampling plan triple and data model parameters.* Let $\mathbf{Z} \in \mathcal{P}$ be a $c \times K$ population matrix and let $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F$ be a sampling constraint matrix in $\mathcal{S}(\mathbf{Z})$. A sampling plan is characterized by the triple $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ in the following sense. Let $\phi_k = (i : Z_{ik} = 1), k = 1, \ldots, K$. Consider K independent random samples

$$C_h(k) \text{ indep} \sim C | C \in \phi_k, \qquad h = 1, \ldots, N_k, k = 1, \ldots K.$$

Assume that (i) $\{\{C_h(k)\}, \{N_k\}\}$ are mutually independent and (ii) $N_k = n_k$ or $N_k \sim \text{Po}(\delta_k)$ according to whether $\mathbf{Z}_F$ includes the kth column of $\mathbf{Z}$ or not. In words, $\mathbf{Z}$ determines the strata from which samples are drawn, $\mathbf{Z}_F$ indicates which samples have a priori fixed sample sizes (the nonfixed sample sizes have Poisson distributions) and $\mathbf{n}$ gives the fixed sample sizes.

Define the *data model probabilities* as $\pi_i \equiv P(C = i | C \in \phi_{k_i})$, where k_i is the column in $\mathbf{Z}$ that has a 1 in the ith row (i.e., $\phi_{k_i} \ni i$). The data model expected sample sizes are in $\boldsymbol{\gamma} \equiv E(N_1, \ldots, N_K)^T$, which will be assumed positive throughout this article. The data model parameters $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ satisfy:

(i) $\boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta} > 0$,
(ii) $\boldsymbol{\pi} = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{P})\mathbf{P} > 0$ and
(iii) $\mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K$.

Note that the dimensions of $\mathbf{n}$ and $\boldsymbol{\delta}$ equal the numbers of columns in $\mathbf{Z}_F$ and $\mathbf{Z}_R$, respectively. We also point out that the a priori fixed sample sizes $\mathbf{n}$, corresponding to a nonzero $\mathbf{Z}_F$, are assumed to be positive throughout this work.

3.3. *Multinomial–Poisson distribution.* Using standard distribution theory arguments, it follows that the counts $Y_i \equiv \#(C_h(k) = i), i = 1, \ldots, c$, are sufficient for $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ and $\mathbf{Y}_{\phi_1}, \ldots, \mathbf{Y}_{\phi_K}$ are mutually independent. Moreover, standard arguments [e.g., Ross (1993), pages 216 and 217] lead to

$$\mathbf{Y}_{\phi_k} \sim \text{mult}(n_k, \boldsymbol{\pi}_{\phi_k}) \qquad \text{if } \mathbf{Z}_F \text{ includes the } k\text{th column of } \mathbf{Z},$$

$$\mathbf{Y}_{\phi_k, j} \text{ indep} \sim \text{Po}(\delta_k \boldsymbol{\pi}_{\phi_k, j}), \qquad j = 1, \ldots, Z_{+k},$$

$$\text{if } \mathbf{Z}_F \text{ does not include the } k\text{th column of } \mathbf{Z}.$$

It also follows that the sufficient counts $\mathbf{Y} \equiv (Y_1, \ldots, Y_c)^T$ have the following properties:

(i) $\mathbf{Z}^T\mathbf{Y} = (N_1, N_2, \ldots, N_K)^T$,
(ii) $P(\mathbf{Z}_F^T\mathbf{Y} = \mathbf{n}) = 1$,
(iii) $E(\mathbf{Y}) = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}$,
(iv) $\operatorname{var}(\mathbf{Y}) = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})[\mathbf{D}(\boldsymbol{\pi}) - \mathbf{D}(\boldsymbol{\pi})\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D}(\boldsymbol{\pi})]$.

The probability density function of the sufficient count vector $\mathbf{Y}$ can be parameterized in terms of $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ or, owing to the one-to-one result of Proposition 1 below, the expected count vector $\mathbf{m} \equiv E(\mathbf{Y})$. It will prove useful to give the general form of the probability density for both parameterizations—the $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameterization is convenient for the study of asymptotic behavior of model estimators and the $\mathbf{m}$ parameterization is convenient for model fitting and specification.

The random vector $\mathbf{Y}$, with corresponding sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, will be called a *multinomial–Poisson* (*MP*) *random vector*. When the density is parameterized in terms of $(\boldsymbol{\gamma}, \boldsymbol{\pi})$, we will write $\mathbf{Y} \sim \mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi} | \mathbf{Z}_F, \mathbf{n})$, and when the density is parameterized in terms of $\mathbf{m}$, we will write $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{m} | \mathbf{Z}_F, \mathbf{n})$.

3.3.1. *The* $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ *parameterization of the MP density.* It can be shown that the probability density function of $\mathbf{Y} \sim \mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi} | \mathbf{Z}_F, \mathbf{n})$ has the general form

$$P(\mathbf{Y} = \mathbf{y}) = c^*(\mathbf{y}) \exp\{\mathbf{y}^T \log \boldsymbol{\pi} + \mathbf{y}^T\mathbf{Z}_R \log(\mathbf{Q}_R^T\boldsymbol{\gamma}) - \mathbf{1}^T\mathbf{Q}_R^T\boldsymbol{\gamma}\} I(\mathbf{y} \in S), \tag{3.2}$$

where $c^*(\mathbf{y}) \equiv \mathbf{n}!/\mathbf{y}!$ if $\mathbf{Z}_F \neq 0$ and $c^*(\mathbf{y}) \equiv 1/\mathbf{y}!$ if $\mathbf{Z}_F = 0$. Here, $(x_1, \ldots, x_m)! \equiv x_1!x_2! \cdots x_m!$ and S is the support set as described in Section 3.3.2.

The collection of admissible $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameter values will be written as

$$\omega_Z^*(0|\mathbf{Z}_F, \mathbf{n}) \equiv \{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta}, \boldsymbol{\delta} > 0, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K\}.$$

By convention, $\omega_Z^*(0|0) = \{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \boldsymbol{\delta}, \boldsymbol{\delta} > 0, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K\}$ and $\omega_Z^*(0|\mathbf{Z}, \mathbf{n}) = \{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \mathbf{n}, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K\}$.

3.3.2. *The* $\mathbf{m}$ *parameterization of the MP density.* The following proposition gives the one-to-one correspondence between the $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameters and the expected count or mean parameter $\mathbf{m} \equiv E(\mathbf{Y}) = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}$.

PROPOSITION 1. *Let* $\mathbf{R}(\boldsymbol{\gamma}, \boldsymbol{\pi}) \equiv \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi} = \mathbf{m}$. (1) *The function* $\mathbf{R} : \omega_Z^*(0|\mathbf{Z}_F, \mathbf{n}) \mapsto \mathbf{R}(\omega_Z^*(0|\mathbf{Z}_F, \mathbf{n}))$ *is one-to-one*, (2) *the inverse function* $\mathbf{R}^{-1} : \mathbf{R}(\omega_Z^*(0|\mathbf{Z}_F, \mathbf{n})) \mapsto \omega_Z^*(0|\mathbf{Z}_F, \mathbf{n})$ *is defined as* $\mathbf{R}^{-1}(\mathbf{m}) = (\mathbf{Z}^T\mathbf{m}, \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{m})\mathbf{m}) \equiv (\boldsymbol{\gamma}, \boldsymbol{\pi})$ *and* (3) *the range set* $\mathbf{R}(\omega_Z^*(0|\mathbf{Z}_F, \mathbf{n}))$ *can be reexpressed as* $\mathbf{R}(\omega_Z^*(0|\mathbf{Z}_F, \mathbf{n})) = \{\mathbf{m} : \mathbf{m} > 0, \mathbf{Z}_F^T\mathbf{m} = \mathbf{n}\} \equiv \omega(0|\mathbf{Z}_F, \mathbf{n})$.

By convention, if $\mathbf{Z}_F = 0$ (i.e., $\mathbf{Q}_F = 0$), we have $\omega(0|0) = \{\mathbf{m} : \mathbf{m} > 0\}$. The proof of Proposition 1 is relatively straightforward and is omitted.

The probability density function of $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{m}|\mathbf{Z}_F, \mathbf{n})$ has the general form

$$P(\mathbf{Y} = \mathbf{y}) = c(\mathbf{y}) \exp\{\mathbf{y}^T \log \mathbf{m} - \mathbf{1}^T \mathbf{Z}_R^T \mathbf{m}\} I(\mathbf{y} \in S), \tag{3.3}$$

where $c(\mathbf{y}) = \mathbf{n}! \exp\{-\mathbf{n}^T \log \mathbf{n}\}/\mathbf{y}!$ if $\mathbf{Z}_F \neq 0$ and $c(\mathbf{y}) = 1/\mathbf{y}!$ if $\mathbf{Z}_F = 0$.

The collection of admissible $\mathbf{m}$ parameter values is

$$\omega(0|\mathbf{Z}_F, \mathbf{n}) \equiv \{\mathbf{m} : \mathbf{m} > 0, \mathbf{Z}_F^T \mathbf{m} = \mathbf{n}\}.$$

The support set S, which satisfies $P(\mathbf{Y} \in S) = 1$, is related to the $\mathbf{m}$ parameter space $\omega(0|\mathbf{Z}_F, \mathbf{n})$. Specifically, $S = \{\mathbf{y} \in \mathcal{Z}^c : \mathbf{y} \geq 0, \mathbf{Z}_F^T \mathbf{y} = \mathbf{n}\}$, where $\mathcal{Z}$ is the set of integers.

REMARK. Multinomial–Poisson distributions have been used before for the purpose of comparing expected count estimates for two different contingency table models. For example, Birch (1963) and Bishop, Fienberg and Holland [(1975), pages 446 and 447] described data models with corresponding densities that can be shown to have the MP density form (3.3). The current article not only uses a derivation that motivates the appropriateness of this density, but it also introduces the alternative $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameterization (3.2), which will prove very useful for model interpretation and for describing the large-sample behavior of MP estimators.

REMARK. Two common, special-case MP distributions are $\mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi}|0) = \mathrm{MP}_Z(\mathbf{m}|0)$ (i.e., product Poisson) and $\mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi}|\mathbf{Z}, \mathbf{n}) = \mathrm{MP}_Z(\mathbf{m}|\mathbf{Z}, \mathbf{n})$ (i.e., product multinomial). The well-known relationship between the two distributions [see Bishop, Fienberg and Holland (1975), page 440] can be stated here as follows. If $\mathbf{Y} \sim \mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi}|0)$, then $\mathbf{Y}|\mathbf{Z}^T\mathbf{Y} = \mathbf{n} \sim \mathrm{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi}|\mathbf{Z}, \mathbf{n})$.

3.4. *Multinomial–Poisson data models.* A primary goal in contingency table analysis is to use data $\mathbf{y}$, a realization of some MP random vector $\mathbf{Y}$ (denoted $\mathbf{y} \leftarrow \mathbf{Y}$), to model and conduct inferences about the predata probability vector $\mathbf{P}$ and, at times, the unknown expected sample size, or rate parameters in $\boldsymbol{\gamma}$. For now, suppose that it is possible to write a model of interest for $\mathbf{P}$ in terms of constraints on the expected counts, say $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. Section 10 gives sufficient conditions under which this is possible.

The "unrestricted" MP data model states that $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{m}|\mathbf{Z}_F, \mathbf{n})$, where $\mathbf{m}$ is some value in $\omega(0|\mathbf{Z}_F, \mathbf{n})$. That is, any $\mathbf{m}$ that satisfies the sample size constraint $\mathbf{Z}_F^T \mathbf{m} = \mathbf{n}$ and the positivity constraint $\mathbf{m} > 0$ is a candidate value.

More generally, a MP data model imposes the additional constraints $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ on $\mathbf{m}$. That is, in general, a MP data model constrains $\mathbf{m}$ to fall in the model space

$$\omega(\mathbf{h}|\mathbf{Z}_F, \mathbf{n}) \equiv \{\mathbf{m} : \mathbf{m} > 0, \mathbf{Z}_F^T \mathbf{m} = \mathbf{n}, \mathbf{h}(\mathbf{m}) = \mathbf{0}\}, \tag{3.4}$$

where $\mathbf{h}$ is some model constraint function. The notation $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ will mean that $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{m}|\mathbf{Z}_F, \mathbf{n})$, where $\mathbf{m} \in \omega(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$.

The MP data model $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ is equivalent to the reparameterized version $\mathbf{Y} \sim \mathrm{MP}^*_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$, which means $\mathbf{Y} \sim \mathrm{MP}^*_Z(\boldsymbol{\gamma}, \boldsymbol{\pi}|\mathbf{Z}_F, \mathbf{n})$, where $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ is some value in

$$\begin{aligned}\omega^*_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n}) \equiv \big\{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta},\\ \boldsymbol{\delta} > 0, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K, \mathbf{h}\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}\big) = \mathbf{0}\big\}.\end{aligned} \tag{3.5}$$

4. Motivating example: citation patterns in statistics journals. Stigler (1994) investigated citation patterns in several journals of statistics and probability. Herein, we restrict attention to the *Journal of the American Statistical Association* (JASA), *Biometrics* (BMCS) and *The Annals of Statistics* (ANNS). Let $C = (A, B)$ be a cross-citation, where A is the journal of the citing article (1 = JASA, 2 = BMCS, 3 = ANNS) and B is the journal of the cited article (1 = JASA, 2 = BMCS, 3 = ANNS). Let $P_{ij} = P(A = i, B = j)$ be the probability that a randomly selected cross-citation refers to an article published in journal i and a cited reference published in journal j. One of the primary objects of inference is $\mathbf{P} = (P_{11}, P_{12}, \ldots, P_{33})^T$ or some function thereof.

4.1. *Unrestricted MP data models.* Stigler (1994) used citation data on journals published between 1987 and 1989. Let (a_h, b_h) be the citing journal and cited journal, respectively, for the hth selected cross-citation from the 1987–1989 journal issues. Using similar arguments to those given in Stigler (1994), a tenable 1987–1989 data model is

$$(a_h, b_h) \leftarrow (A_h, B_h) \text{ i.i.d. } \sim (A, B), \qquad h = 1, \ldots, N \sim \mathrm{Po}(\delta).$$

The data parameters are $\pi_{ij} = P(A = i, B = j) = P_{ij}$ and δ, which is the expected number of cross-citations involving JASA, BMCS and ANNS during 1987–1989. The counts $Y_{ij} \equiv \#(A_h = i, B_h = j)$ are sufficient for this model. Moreover, the resulting sufficient data model is $y_{ij} \leftarrow Y_{ij}$ indep $\sim \mathrm{Po}(\delta\pi_{ij}), i, j = 1, 2, 3$. Using MP model notation, $\mathbf{y} \leftarrow \mathbf{Y} \sim \mathrm{MP}_{\mathbf{1}_9}(\mathbf{m}|0) = \mathrm{MP}^*_{\mathbf{1}_9}(\boldsymbol{\gamma}, \boldsymbol{\pi}|0)$, where $\boldsymbol{\gamma} = \delta$ and $\mathbf{m} = \mathbf{D}(\mathbf{1}_9\boldsymbol{\gamma})\boldsymbol{\pi} = \delta\boldsymbol{\pi}$. The observed counts $\mathbf{y}$, which are taken from Table 4 of Stigler (1994), are reproduced in Table 1.

TABLE 1
1987–1989 *statistics journals citation pattern counts*

	Cited		
Citing	**JASA**	**BMCS**	**ANNS**
JASA	1072	264	739
BMCS	348	770	155
ANNS	340	42	1623

We conducted a small-scale study of citation patterns for more recently published statistics articles appearing in JASA, BMCS and ANNS. Specifically, the cited journals in the bibliographies of articles appearing in the December 1999 issues of JASA and BMCS were recorded. Because the ANNS includes fewer articles and, hence, fewer references per issue than the other two, it was decided to start with the December 1999 issue and go back in time to sample until 225 JASA, BMCS or ANNS articles were referenced. With this stopping rule, it was necessary to go to the fourth article in the August 1999 issue of ANNS.

For the current 1999 cross-citation study, let b_{ih} be the cited journal for the hth reference in journal i. A tenable data model is

$$b_{ih} \leftarrow B_h(i) \,\text{indep} \sim B|A = i, \qquad h = 1, \ldots, N_i,$$

where $N_i \sim \text{Po}(\delta_i), i = 1, 2$, and $N_3 = n_3 = 225$. The $B_h(i)$'s and N_i's are mutually independent. Note that the parameters of this model are $\pi_{ij} \equiv P(B = j|A = i) = P_{ij}/P_{i+}$, δ_1 and δ_2. The rate parameters δ_i give the expected number of JASA, BMCS and ANNS references per issue of journal i.

To illustrate the coordinate-free notation of the previous section, identify events $(C = 1), (C = 2), \ldots, (C = 9)$ with $(A = 1, B = 1), (A = 1, B = 2), \ldots, (A = 3, B = 3)$. It follows that the sampling plan is characterized by $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, where $\mathbf{Z} = \bigoplus_{k=1}^{3} \mathbf{1}_3$, $\mathbf{Z}_F$ is equal to the third column of $\mathbf{Z}$ and $\mathbf{n} = n_3 = 225$. Notice that $B|A = k \sim C|C \in \phi_k$, where $\phi_k = (i : Z_{ik} = 1), k = 1, 2, 3$.

The 1999 data model above implies that the counts $Y_{ij} = \#(B_h(i) = j)$ are sufficient. Using standard arguments it can be shown that

$$\begin{aligned} y_{ij} \leftarrow Y_{ij} \,\text{indep} &\sim P(\delta_i \pi_{ij}), \qquad i = 1, 2, j = 1, 2, 3, \\ (y_{31}, y_{32}, y_{33}) \leftarrow (Y_{31}, Y_{32}, Y_{33}) &\sim \text{mult}(n_3 = 225, \pi_{31}, \pi_{32}, \pi_{33}). \end{aligned} \tag{4.1}$$

The multinomial and the six Poisson random variables are mutually independent. More succinctly, we can write $\mathbf{y} \leftarrow \mathbf{Y} \sim \text{MP}_Z(\mathbf{m}|\mathbf{Z}_F, \mathbf{n}) = \text{MP}_Z^*(\boldsymbol{\gamma}, \boldsymbol{\pi}|\mathbf{Z}_F, \mathbf{n})$, where $\boldsymbol{\gamma} = [\delta_1, \delta_2, n_3]^T$, $\mathbf{m} = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}$ and $\mathbf{n} = n_3 = 225$. The observed counts $\mathbf{y}$ are shown in Table 2.

TABLE 2
1999 *statistics journals citation pattern counts* ($n_3 = 225$)

	Cited		
Citing	**JASA**	**BMCS**	**ANNS**
JASA	104	24	65
BMCS	76	146	30
ANNS	50	9	166

4.2. *Some estimands, hypotheses and models of interest.* Some estimands of interest include the Gini concentrations of citations for each of the journals, $S_i(\mathbf{P}) = G_i = \sum_{j=1}^{3}(P_{ij}/P_{i+})^2, i = 1, 2, 3$; the odds that a JASA article cites a JASA rather than a BMCS or ANNS paper, $S_4(\mathbf{P}) = P_{11}/(P_{12} + P_{13})$; the probability that a BMCS article cites an ANNS paper rather than BMCS or JASA, $S_5(\mathbf{P}) = P_{23}/P_{2+}$; and, for a randomly selected cross-citation that involves BMCS and JASA, the odds that a BMCS article cites a JASA paper rather than the other way around, $S_6(\mathbf{P}) = P_{21}/P_{12}$. It is also of interest to estimate the expected number of ANNS references per issue of JASA and the expected number of JASA references per issue of ANNS.

One hypothesis of interest is the hypothesis of no change in Gini concentrations from the 1987–1989 values. Using the counts in Table 1, we have that the observed 1987–1989 Gini concentrations are 0.410, 0.455 and 0.684 for JASA, BMCS and ANNS, respectively. Treating these as population values, the no-change hypothesis can be written as $\mathbf{h}_1(\mathbf{P}) = (G_1 - 0.410, G_2 - 0.455, G_3 - 0.684)^T = \mathbf{0}$. Other candidate hypotheses include $\mathbf{h}_2(\mathbf{P}) = (G_1 - G_2, G_1 - G_3) = \mathbf{0}$, $h_3(\mathbf{P}) = P_{12} - P_{22} = 0$ and $h_4(\mathbf{P}) = \log(P_{21}/P_{12}) - \log(P_{31}/P_{13}) + \log(P_{32}/P_{23}) = 0$. The hypotheses corresponding to $\mathbf{h}_2$ and h_3 have straightforward interpretations, but h_4 requires some elaboration.

Stigler (1994) used the 1987–1989 citation data to model the probabilities P_{ij} through the so-called *exchange score model*. Specifically, the exchange score model in our setting has the form

$$\log \frac{P_{ij}}{P_{ji}} = \alpha_i - \alpha_j, \qquad i > j,\ i, j = 1, 2, 3,$$

and is equivalent to the Bradley–Terry paired-comparison model [see Agresti (1990), page 370]. Without loss of generality, α_1 can be set to zero for identifiability. The values $\alpha_1 \equiv 0, \alpha_2$ and α_3 are the *exchange scores*, which measure the level of information exchange among the three journals. This exchange score model can be restated in terms of the single constraint, $h_4(\mathbf{P}) = \log(P_{21}/P_{12}) - \log(P_{31}/P_{13}) + \log(P_{32}/P_{23}) = 0$.

Section 10 argues that the constraints specified using $\mathbf{h}_1, \mathbf{h}_2$ and h_4 can be reexpressed in terms of the 1999 data model expected count parameters $\mathbf{m}$. Specifically, $\mathbf{h}_i(\mathbf{P}) = \mathbf{0}$ if and only if $\mathbf{h}_i(\mathbf{m}) = \mathbf{0}$ for $i = 1, 2, 4$. This implies that tests of these hypotheses using the 1999 data are equivalent to tests of goodness of fit of the 1999 restricted data models $\mathrm{MP}_Z(\mathbf{h}_i|\mathbf{Z}_F, \mathbf{n}), i = 1, 2, 4$. In contrast, Section 10 argues that $h_3(\mathbf{P}) = 0$ is not equivalent to constraints on the 1999 expected counts $\mathbf{m}$.

5. Z-homogeneous functions. Most contingency table models commonly used in practice can be directly specified using a constraint function $\mathbf{h}$ that has several convenient properties. For example, $\mathbf{h}$ often satisfies (i) $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ if and only if $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$ if and only if $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ and (ii) the collection of constraints

$\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$ and $\mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}$ is nonredundant. Provided a MP model has constraint function $\mathbf{h}$ that is sufficiently smooth, nonredundant and satisfies properties (i) and (ii), model fitting and inference are simplified, as are comparisons of inferences for different sampling plans. This section introduces a broad class of functions $\mathbf{h}$ that have these useful properties.

DEFINITION 3. Let $\Omega = \{\mathbf{x} \in R^c : \mathbf{x} > 0\}$. A function $\mathbf{h}: \Omega \to R^u$ is $\mathbf{Z}$-*homogeneous* [*of order* $\mathbf{p} = (p(1), \ldots, p(u))^T$] if

$$\mathbf{h}\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}\big) = \mathbf{G}(\boldsymbol{\delta})\mathbf{h}(\mathbf{x}) \qquad \forall\, \boldsymbol{\delta} > 0, \forall\, \mathbf{x} \in \Omega,$$

where $\mathbf{G}(\boldsymbol{\delta}) = \operatorname{diag}\{\delta_{\nu(j)}^{p(j)} : j = 1, \ldots, u\}$. Here, $\mathbf{Z}$ is a $c \times K$ population matrix and $\nu(j) \in \{1, \ldots, K\}$. When the orders are not important the phrase in square brackets is omitted. The function $\mathbf{h}$ is $\mathbf{Z}$-*homogeneous of order* 0 if $\mathbf{p} = \mathbf{0}$. In this special case,

$$\mathbf{h}\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}\big) = \mathbf{h}(\mathbf{x}) \qquad \forall\, \boldsymbol{\delta} > 0, \forall\, \mathbf{x} \in \Omega.$$

The zero function defined as $h(\mathbf{x}) \equiv 0$ is a zero-order $\mathbf{Z}$-homogeneous function for any $\mathbf{Z}$.

EXAMPLE 5.1. Consider the population matrices $\mathbf{Z}_1 = \mathbf{1}_4$, $\mathbf{Z}_2 = \bigoplus_1^2 \mathbf{1}_2$ and $\mathbf{Z}_3 = \mathbf{1}_2 \otimes \mathbf{I}_2$. The function defined as $\mathbf{h}(\mathbf{x}) = [x_1 - x_3,\ x_2^2 - x_2x_4]^T$ is $\mathbf{Z}_3$-homogeneous of order $\mathbf{p} = (1, 2)^T$. To see this note that

$$\mathbf{h}\big(\mathbf{D}(\mathbf{Z}_3\boldsymbol{\delta})\mathbf{x}\big) = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2^2 \end{bmatrix} \mathbf{h}(\mathbf{x}).$$

This function $\mathbf{h}$ is also $\mathbf{Z}_1$-homogeneous of order $\mathbf{p} = (1, 2)^T$, but it is not $\mathbf{Z}_2$-homogeneous. The function defined as $h(\mathbf{x}) = x_1/(x_1 + x_2) - x_3/(x_3 + x_4)$ is $\mathbf{Z}_1$- and $\mathbf{Z}_2$-homogeneous of order 0, but it is not $\mathbf{Z}_3$-homogeneous.

EXAMPLE 5.2. Consider the functions S_i, $i = 1, \ldots, 6$, of Section 4.2. Define population matrices $\mathbf{Z}_1 = \mathbf{1}_9$, $\mathbf{Z}_2 = \bigoplus_{k=1}^3 \mathbf{1}_3 = \mathbf{I}_3 \otimes \mathbf{1}_3$ and $\mathbf{Z}_3 = \mathbf{1}_3 \otimes \mathbf{I}_3$. It is straightforward to see that, for example, the first Gini concentration function S_1, defined as $S_1(\mathbf{P}) = \sum_{j=1}^3 (P_{1j}/P_{1+})^2$, is $\mathbf{Z}_1$- and $\mathbf{Z}_2$-homogeneous of order 0, but it is not $\mathbf{Z}_3$-homogeneous. The function S_6, defined as $S_6(\mathbf{P}) = P_{21}/P_{12}$, is $\mathbf{Z}_1$-homogeneous of order 0, but it is not $\mathbf{Z}_2$- or $\mathbf{Z}_3$-homogeneous.

Consider the population matrices of the previous paragraph and the constraint functions h_3 and h_4 of Section 4.2. The function h_3 is $\mathbf{Z}_1$- and $\mathbf{Z}_3$-homogeneous of order 1, but it is not $\mathbf{Z}_2$-homogeneous. The function h_4 is $\mathbf{Z}_1$-, $\mathbf{Z}_2$- and $\mathbf{Z}_3$-homogeneous of order 0.

For notational convenience, we let $\mathcal{H}(\mathbf{Z})$ be the set of all $\mathbf{Z}$-homogeneous functions. The subset $\mathcal{H}_{\mathbf{p}}(\mathbf{Z})$ contains only $\mathbf{Z}$-homogeneous functions of order $\mathbf{p}$. The following definition gives other useful subsets and supersets.

DEFINITION 4. The set $\mathcal{H}''(\mathbf{Z})$ contains all functions $\mathbf{h}: \Omega \mapsto R^u$ that satisfy the following four conditions:

H_0. $\omega(\mathbf{h}|\mathbf{0}) \equiv \{\mathbf{x}: \mathbf{x} > 0, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} \neq \varnothing$.
H_1. $\mathbf{h}$ has continuous second-order derivatives on Ω.
H_2. $\mathbf{H}(\mathbf{x}) \equiv \partial \mathbf{h}^T(\mathbf{x})/\partial \mathbf{x}$ is full column rank u on Ω.
H_3. $\mathbf{h} \in \mathcal{H}(\mathbf{Z})$.

The subset $\mathcal{H}''_{\mathbf{p}}(\mathbf{Z})$ includes only $\mathcal{H}''(\mathbf{Z})$ functions of order $\mathbf{p}$. The superset $\mathcal{H}'' \equiv \bigcup_{\mathbf{Z} \in \mathcal{P}} \mathcal{H}''(\mathbf{Z})$ and $\mathcal{H}''_{\mathbf{p}} \equiv \bigcup_{\mathbf{Z} \in \mathcal{P}} \mathcal{H}''_{\mathbf{p}}(\mathbf{Z})$.

To conveniently state results that accommodate the singular case, unrestricted model (with zero constraint function $h \equiv 0$), it will prove convenient to include the zero function in $\mathcal{H}''(\mathbf{Z})$. To accomplish this, we consider the following conventions: (i) the zero function will be considered a mapping from Ω to $R^0 \equiv \{0\}$ and (ii) the derivative of the zero function, $\mathbf{H} \equiv \mathbf{0}$, will be considered to be of *full column rank* $u = 0$.

Propositions 2–7 give useful properties of $\mathbf{Z}$-homogeneous functions. With the exception of Proposition 5, the proofs are relatively straightforward and are omitted.

PROPOSITION 2. *Let* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *be compatible population matrices. If* $\mathbf{h} \in \mathcal{H}_{\mathbf{p}}(\mathbf{Z}_1)$, *then* $\mathbf{h} \in \mathcal{H}_{\mathbf{p}}(\mathbf{Z}_1\mathbf{Z}_2)$.

PROPOSITION 3. *If* $\mathbf{h} \in \mathcal{H}_{\mathbf{p}}(\mathbf{Z})$, *then there exists a matrix-valued function* $\mathbf{B}(\cdot)$ *such that* $\mathbf{B}(\mathbf{x})$ *is diagonal and positive definite on* Ω *and* $\mathbf{h}_0 \equiv \mathbf{B}\mathbf{h} \in \mathcal{H}_0(\mathbf{Z})$.

As an example, $\mathbf{h}(\mathbf{x}) = [x_1 - x_2, x_1^3 - x_1x_2x_3]^T$ is $\mathbf{1}_3$-homogeneous of orders $(1, 3)^T$. Defining $\mathbf{B}(\mathbf{x}) = \operatorname{diag}\{1/x_1, 1/x_1^3\}$, it follows that $\mathbf{h}_0 \equiv \mathbf{B}\mathbf{h}$ is $\mathbf{1}_3$-homogeneous of order 0.

For modeling purposes, the constraint function $\mathbf{h}_0$ could be called *zero-order version of* $\mathbf{h}$. This language is reasonable because the $\mathbf{m}$ parameter space (3.4) satisfies $\omega(\mathbf{h}|\mathbf{Z}_F, \mathbf{n}) = \omega(\mathbf{h}_0|\mathbf{Z}_F, \mathbf{n})$. Thus, without loss of generality, homogeneous function model spaces could be specified in terms of zero-order homogeneous constraint functions. We point out, however, that in practice it may be simpler to work with homogeneous constraints of nonzero order.

PROPOSITION 4. *Provided first-order derivatives exist, if* $\mathbf{h} \in \mathcal{H}(\mathbf{Z})$, *with* $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}) = \mathbf{G}(\boldsymbol{\delta})\mathbf{h}(\mathbf{x})$, *then*

$$\mathbf{H}\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}\big) = \mathbf{D}^{-1}(\mathbf{Z}\boldsymbol{\delta})\mathbf{H}(\mathbf{x})\mathbf{G}(\boldsymbol{\delta}) \qquad \forall\, \boldsymbol{\delta} > 0, \forall\, \mathbf{x} \in \Omega.$$

PROPOSITION 5 (Generalized Euler's homogeneous function theorem). *Provided first-order derivatives exist*,

$$\mathbf{h} \in \mathcal{H}_{\mathbf{p}}(\mathbf{Z}) \quad \textit{if and only if} \quad \mathbf{Z}^T\mathbf{D}(\mathbf{x})\mathbf{H}(\mathbf{x}) = \mathbf{A}\mathbf{D}(\mathbf{p})\mathbf{D}(\mathbf{h}(\mathbf{x})) \qquad \forall\, \mathbf{x} \in \Omega,$$

where the matrix $\mathbf{A}$ *has components that satisfy* $A_{ij} \in \{0, 1\}$, $A_{+j} = 1$.

Proposition 5 is very important for the subsequent results in this article; its proof is given in the Appendix. An inspection of the proof of the necessity part of this theorem indicates that $\mathbf{A}$ has components of the form $A_{ij} = I(\nu(j) = i)$, where the $\nu(j)$'s are subscripts in $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}) = \mathbf{G}(\boldsymbol{\delta})\mathbf{h}(\mathbf{x})$, where $\mathbf{G}(\boldsymbol{\delta}) = \operatorname{diag}\{\delta_{\nu(j)}^{p(j)} : j = 1, \ldots, c\}$. This in turn leads to the useful identity

$$\mathbf{A}\mathbf{D}(\mathbf{p}) = \frac{\partial \mathbf{d}_{\mathbf{G}}(\mathbf{1})^T}{\partial \boldsymbol{\delta}}, \tag{5.1}$$

where $\mathbf{d}_{\mathbf{G}}(\boldsymbol{\delta})$ is the diagonal vector of $\mathbf{G}(\boldsymbol{\delta})$.

Proposition 5 and its direct corollaries, Propositions 6 and 7, lead to simplifications in model fitting and in derivations of the model equivalence results of Section 8.

PROPOSITION 6. *If* $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, *then* $\mathbf{Z}^T\mathbf{D}(\mathbf{x})\mathbf{H}(\mathbf{x}) = \mathbf{0}\ \forall\, \mathbf{x} \in \omega(\mathbf{h}|0)$. *In case* $\mathbf{h} \in \mathcal{H}_0''(\mathbf{Z})$, *the identity holds for all* $\mathbf{x} \in \Omega$.

PROPOSITION 7. *If* $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, *then the matrix* $[\mathbf{H}(\mathbf{x})\vdots\mathbf{Z}]$ *is full column rank on* $\omega(\mathbf{h}|0)$. *In case* $\mathbf{h} \in \mathcal{H}_0''(\mathbf{Z})$, *the full rank condition holds throughout* Ω.

Notice that condition H_2 of Definition 4 implies that $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ contains no redundant constraints. Proposition 7, which follows from the orthogonality property of Proposition 6, implies that the entire collection of model and identifiability constraints, $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{Z}^T\mathbf{x} = \mathbf{1}$, is nonredundant as well. This further implies that $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{Z}_F^T\mathbf{x} = \mathbf{n}$ are nonredundant also.

6. MP homogeneous data models.

DEFINITION 5. The data model $\mathcal{M}$ is said to be an *MP homogeneous model* if there exist a sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ and a homogeneous constraint function $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$ such that $\mathcal{M} = \mathrm{MP}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$. This MPH model will be denoted $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$.

Multinomial–Poisson homogeneous data models of the form $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ have constraint functions $\mathbf{h}$ that satisfy the two useful properties discussed at the beginning of Section 5. That is, when $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, (i) $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ if and only if $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$ if and only if $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ and (ii) the collection of constraints $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$

and $\mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}$ is nonredundant. Property (i) follows directly from the definition of a $\mathbf{Z}$-homogeneous function and property (ii) follows from Proposition 7.

Multinomial–Poisson homogeneous models have $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameter spaces that are well structured. In particular the $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameter space (3.5) of Section 3.4 can be written as a product space, namely

$$
\begin{aligned}
&\omega_Z^*(\mathbf{h}|\mathbf{Z}_F, \mathbf{n}) \\
&\quad = \{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta}, \boldsymbol{\delta} > 0, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K, \mathbf{h}\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}\big) = \mathbf{0}\} \\
&\quad = \{(\boldsymbol{\gamma}, \boldsymbol{\pi}) : \boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta}, \boldsymbol{\delta} > 0, \boldsymbol{\pi} > 0, \mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}_K, \mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}\} \\
&\quad \equiv \mathcal{D}(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}) \times \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}),
\end{aligned} \tag{6.1}
$$

where $\mathcal{D}(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}) \equiv \{\boldsymbol{\gamma} : \boldsymbol{\gamma} = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\boldsymbol{\delta}, \boldsymbol{\delta} > 0\}$. Moreover, because Proposition 7 implies that the $u + K$ constraints, $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$ and $\mathbf{Z}^T\boldsymbol{\pi} = \mathbf{1}$, are nonredundant and because $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$ is well behaved, the space $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$, as defined in (3.4), is a $(c - u - K)$-dimensional manifold [see Fleming (1977), Section 4.7] and is topologically well behaved. It is precisely this product-space manifold representation that leads to many of the subsequent results in this article.

EXAMPLE 6.1 (Log-linear models). Suppose that $\mathbf{Y} \sim \mathrm{MP}_Z(\mathbf{m}|0)$, where $\log\mathbf{m} = \mathbf{X}\boldsymbol{\beta}$. This is the Poisson log-linear model, which can be written as the MP model $\mathrm{MP}_Z(\mathbf{h}|0)$, where $\mathbf{h}(\mathbf{m}) = \mathbf{U}^T\log\mathbf{m} = \mathbf{0}$ and the matrix $\mathbf{U}$ is a full-column-rank orthogonal complement of $\mathbf{X}$. The $\mathbf{m}$ parameter space is $\omega(\mathbf{h}|0) = \{\mathbf{m} : \mathbf{m} > 0, \mathbf{h}(\mathbf{m}) = \mathbf{0}\}$. If the range (or column space) of $\mathbf{X}$, denoted $R(\mathbf{X})$, contains $R(\mathbf{Z})$ so that $\mathbf{U}^T\mathbf{Z} = \mathbf{0}$, then $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}) = \mathbf{U}^T\log(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}) = \mathbf{U}^T\mathbf{Z}\log\boldsymbol{\gamma} + \mathbf{U}^T\log\boldsymbol{\pi} = \mathbf{U}^T\log\boldsymbol{\pi} = \mathbf{h}(\boldsymbol{\pi})$. Thus, provided that $R(\mathbf{X})$ contains $R(\mathbf{Z})$, the function $\mathbf{h}$ is in $\mathcal{H}''(\mathbf{Z})$ and $\mathrm{MP}_Z(\mathbf{h}|0) = \mathrm{MPH}_Z(\mathbf{h}|0)$; that is, it is an MPH model. Reparameterizing in terms of $\boldsymbol{\gamma} = \mathbf{Z}^T\mathbf{m}$ and $\boldsymbol{\pi} = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{m})\mathbf{m}$, the $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameter space simplifies to $\omega_Z^*(\mathbf{h}|0) = \mathcal{D}(\mathbf{Z}, \mathbf{0}) \times \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) = \{\boldsymbol{\gamma} : \boldsymbol{\gamma} > 0\} \times \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$.

The multinomial analogue of the previous Poisson log-linear model is $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}, \mathbf{n})$. The $\mathbf{m}$ parameter space is $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{n}) = \{\mathbf{m} : \mathbf{m} > 0, \mathbf{Z}^T\mathbf{m} = \mathbf{n}, \mathbf{h}(\mathbf{m}) = \mathbf{0}\}$. The $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameter space simplifies to $\omega_Z^*(\mathbf{h}|\mathbf{Z}, \mathbf{n}) = \mathcal{D}(\mathbf{Z}, \mathbf{Z}, \mathbf{n}) \times \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) = \{\mathbf{n}\} \times \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$. The similarity between $\omega_Z^*(\mathbf{h}|0)$ and $\omega_Z^*(\mathbf{h}|\mathbf{Z}, \mathbf{n})$ hints at an "equivalence" between the Poisson and multinomial log-linear models. This equivalence is formally addressed in Section 8.

7. Numerical, asymptotic and approximation results.

7.1. *Numerical results for maximum likelihood estimates.* Consider the MPH data model $\mathbf{y} \leftarrow \mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$. The ML estimate $\hat{\mathbf{m}}$ of $\mathbf{m}$ is the maximizer of the log likelihood $\mathbf{y}^T\log\mathbf{m} - \mathbf{1}^T\mathbf{Z}_R^T\mathbf{m}$ [see density (3.3)], subject to $\mathbf{m} \in \omega(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$). Assume that $\hat{\mathbf{m}}$ exists and uniquely solves the restricted (or

Lagrangian) likelihood equations

$$\begin{bmatrix} \mathbf{y} - \mathbf{D}(\mathbf{m})\mathbf{Z}_R\mathbf{1} + \mathbf{D}(\mathbf{m})\mathbf{H}(\mathbf{m})\boldsymbol{\lambda} + \mathbf{D}(\mathbf{m})\mathbf{Z}_F\boldsymbol{\tau} \\ \mathbf{h}(\mathbf{m}) \\ \mathbf{Z}_F^T\mathbf{m} - \mathbf{n} \end{bmatrix} = \mathbf{0} \tag{7.1}$$

for some Lagrange multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$. The classic articles by Aitchison and Silvey (1958, 1960) and Silvey (1959) give a general discussion of Lagrangian or restricted likelihood equations, and Lang (1996b) considered restricted likelihood equations for a special class of categorical data models.

Theorem 1, which is proven in the Appendix, implies that for MPH models the ML estimate $\hat{\mathbf{m}}$ does not depend on the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$. This, of course, does not mean that the sampling distribution of $\hat{\mathbf{m}}$ does not depend on the sampling plan; indeed, Theorem 4 shows that it does.

THEOREM 1. *Suppose that the maximum likelihood estimate* $\hat{\mathbf{m}}$ *under model* $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *uniquely solves the restricted likelihood equations* (7.1). *Then* $\hat{\mathbf{m}}$ *arises as the solution to the reduced set of equations*

$$\begin{bmatrix} \mathbf{y} - \mathbf{m} + \mathbf{D}(\mathbf{m})\mathbf{H}(\mathbf{m})\boldsymbol{\lambda} \\ \mathbf{h}(\mathbf{m}) \end{bmatrix} = \mathbf{0}. \tag{7.2}$$

Consider the MPH model reparameterized in terms of $(\boldsymbol{\gamma}, \boldsymbol{\pi})$. By the one-to-one result of Proposition 1 and by invariance, $\hat{\mathbf{m}} = \mathbf{D}(\mathbf{Z}\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\pi}}$, so $\hat{\mathbf{m}}$ exists if and only if $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\pi}})$ exists. Theorem 2 shows that the ML estimates have a simple form. The proof is given in the Appendix.

THEOREM 2. *Under the same conditions as in Theorem* 1, *the ML estimate* $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\pi}})$ *under* $\mathrm{MPH}_Z^*(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *can be computed as follows*: $\hat{\boldsymbol{\gamma}} = \mathbf{Z}^T\mathbf{y}$ *and* $\hat{\boldsymbol{\pi}}$ *is the maximizer of* $\mathbf{y}^T \log \boldsymbol{\pi}$ *over* $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$, *which can be written as* $\hat{\boldsymbol{\pi}} = \mathbf{D}^{-1}(\mathbf{Z}\hat{\boldsymbol{\gamma}})\hat{\mathbf{m}}$.

7.2. *Asymptotic results.* Consider the following collection of MPH random vectors indexed by $\nu > 0$: $\mathbf{Y}_\nu \sim \mathrm{MP}_Z(\mathbf{m}_\nu|\mathbf{Z}_F, \mathbf{n}_\nu)$, where $\mathbf{m}_\nu \in \omega(\mathbf{h}|\mathbf{Z}_F, \mathbf{n}_\nu)$ and $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$. For each $\nu > 0$, Proposition 1 and equation (6.1) imply that the mean vector can be written as $\mathbf{m}_\nu = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma}_\nu)\boldsymbol{\pi}_\nu$, where $\boldsymbol{\gamma}_\nu = \mathbf{Q}_F\mathbf{n}_\nu + \mathbf{Q}_R\boldsymbol{\delta}_\nu = \mathbf{Z}^T\mathbf{m}_\nu$ and $\boldsymbol{\pi}_\nu = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{m}_\nu)\mathbf{m}_\nu \in \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$. For asymptotic purposes, we use the sequence $\{\mathbf{Y}_\nu\}$ defined by assuming that $\boldsymbol{\pi}_\nu = \boldsymbol{\pi}$ is fixed with respect to ν and that $\boldsymbol{\gamma}_\nu/\nu \to \mathbf{w} > 0$, as $\nu \to \infty$. Notice that as $\nu \to \infty$, $\mathbf{n}_\nu/\nu \to \mathbf{Q}_F^T\mathbf{w} > 0$, $\boldsymbol{\delta}_\nu/\nu \to \mathbf{Q}_R^T\mathbf{w} > 0$ and $\mathbf{m}_\nu/\nu \to \mathbf{D}(\mathbf{Z}\mathbf{w})\boldsymbol{\pi} > 0$. In words, all the expected sample sizes and all the expected cell counts go to infinity at the same rate.

For convenience, the index ν will be dropped and $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ will denote the sequence $\{\mathbf{Y}_\nu\}$ defined above. Set $\mathbf{N} = \mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{Y})$ and note that, depending on $\mathbf{Z}_F$, $\mathbf{N}$ may or may not be a random matrix. Also, for convenience, let $\mathbf{D} \equiv \mathbf{D}(\boldsymbol{\pi})$, $\mathbf{W} \equiv \mathbf{D}(\mathbf{Z}\mathbf{w})$ and $\mathbf{H} \equiv \mathbf{H}(\boldsymbol{\pi})$.

The next four lemmas give limiting results for the sequence of MPH random vectors $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$. These lemmas will be used to prove subsequent limiting results for MPH model estimators.

LEMMA 1. $\nu^{-1/2}(\mathbf{Y} - \mathbf{m}) \xrightarrow{d} N(0, \mathbf{WD} - \mathbf{WDZ}_F\mathbf{Z}_F^T\mathbf{D})$.

LEMMA 2. $\nu^{-1/2}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}) \xrightarrow{d} N(0, \mathbf{WD} - \mathbf{WDZZ}^T\mathbf{D})$.

LEMMA 3. $\nu^{-1/2}(\mathbf{Z}^T\mathbf{Y} - \boldsymbol{\gamma}) \xrightarrow{d} N(0, \mathbf{Q}_R\mathbf{Q}_R^T\mathbf{D}(\mathbf{w}))$.

LEMMA 4. $\nu^{1/2}(\mathbf{N}^{-1}\mathbf{Y} - \boldsymbol{\pi}) \xrightarrow{d} N(0, \mathbf{W}^{-1}\mathbf{D} - \mathbf{W}^{-1}\mathbf{DZZ}^T\mathbf{D})$.

Haberman [(1974), Theorem 1.1] proved Lemma 1 in the more general conditional–Poisson sampling setting. In the independent sampling setting of this article, the proof can be simplified. For completeness, the Appendix outlines the simplified proof. Lemmas 2–4 follow sequentially from Lemma 1. The proofs are given in the Appendix.

Not surprisingly, the limiting distribution of $\nu^{-1/2}(\mathbf{Y} - \mathbf{m})$ depends on which sample sizes are fixed a priori, that is, it depends on $\mathbf{Z}_F$. Interestingly, the limiting distributions of $\nu^{-1/2}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi})$ and $\nu^{1/2}(\mathbf{N}^{-1}\mathbf{Y} - \boldsymbol{\pi})$ do *not* depend on $\mathbf{Z}_F$. As a special case, the sample proportions $\mathbf{N}^{-1}\mathbf{Y}$, when properly normalized, have the same limiting distributions whether $\mathbf{Y}$ is product-multinomial or Poisson. For example, whether (Y_1, Y_2) is multinomial (with $Y_1 + Y_2$ fixed) or comprises independent Poisson components, the limiting distribution of $\nu^{1/2}(Y_1/(Y_1 + Y_2) - m_1/(m_1 + m_2))$ is unchanged.

The following limiting results for MPH model estimators are most easily derived using the $(\boldsymbol{\gamma}, \boldsymbol{\pi})$ parameterization as described in Section 3.3.1. By Theorem 2, $\hat{\boldsymbol{\pi}}$ is the maximizer over $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ of $\mathbf{Y}^T \log \boldsymbol{\pi}$ and $\hat{\boldsymbol{\gamma}} = \mathbf{Z}^T\mathbf{Y}$. By invariance of ML estimators, $\hat{\mathbf{m}} = \mathbf{D}(\mathbf{Z}\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\pi}} = \mathbf{N}\hat{\boldsymbol{\pi}}$.

The next lemma states that there exists a sequence of local maximizers of $\mathbf{Y}^T \log \boldsymbol{\pi}$ that is strongly consistent for the true $\boldsymbol{\pi}$ value. This implies that if the ML estimators exist and are unique, the ML estimator sequence is strongly consistent. The Appendix gives an outline of the proof, which uses results from Silvey (1959) and Wald (1949).

LEMMA 5. *Suppose that* $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$. *There exists a sequence* $\hat{\boldsymbol{\pi}}$ *of* (*local*) *maximizers in* $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ *of* $\mathbf{Y}^T \log \boldsymbol{\pi}$ *that is strongly consistent. Moreover, for sufficiently large* ν, *with probability going to* 1, *these maximizers emerge through the solution* $\hat{\mathbf{m}} = \mathbf{N}\hat{\boldsymbol{\pi}}$ *to the restricted likelihood equations* (7.2).

This strong consistency result does not come as a surprise because model spaces specified in terms of $\mathcal{H}''(\mathbf{Z})$ functions are well behaved topologically. The

regularity conditions of Definition 4 and Proposition 7 imply that if $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, then $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ is a $(c - u - K)$-dimensional manifold, which is topologically well behaved [see Fleming (1977), page 153]. For example, the implicit function theorem states that, provided $u + K < c$, $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ can be anywhere locally reparameterized in terms of $c - u - K$ freedom parameters. That is, for any $\boldsymbol{\pi}$ in an open neighborhood of a point $\boldsymbol{\pi}_0 \in \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$, there exists a function $\mathbf{f}$ with open domain $D \subseteq R^{c-u-K}$ such that $\boldsymbol{\pi} \in \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ if and only if there exists a $\boldsymbol{\theta} \in D$ such that $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta})$. Moreover, the function $\mathbf{f}$ is locally well behaved (e.g., differentiable) and satisfies regularity conditions like those of Birch (1964). Rather than take the more standard "freedom" approach of Birch (1964), we follow the lead of Aitchison and Silvey (1958) and use the constraint specification of the model-space manifold in the derivation of results. As will become evident, this constraint approach has several advantages over the freedom-specification approach.

The next theorem gives the joint limiting behavior of the sequence of $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ maximum likelihood estimators. The proof, which is outlined in the Appendix, is based on the approach of Aitchison and Silvey (1958). Specifically, a linear approximation to a set of maximum likelihood estimating equations leads to the result.

THEOREM 3. *Suppose that the sequence of models* $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds, with* $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{h}(\mathbf{x})$. *Let the ML estimator* $(\hat{\mathbf{m}} = \mathbf{N}\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ *be the unique solution to the restricted likelihood equations* (7.2). *Then the limiting results*

$$\nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

$$\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}} \xrightarrow{d} N\left(\mathbf{0}, \left[\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}\mathbf{H}\right]^{-1}\right)$$

hold, where $\boldsymbol{\Sigma} = \mathbf{W}^{-1}\mathbf{D} - \mathbf{W}^{-1}\mathbf{D}\mathbf{H}[\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{D}\mathbf{Z}\mathbf{Z}^T\mathbf{D}$. *Moreover, the estimators* $\hat{\boldsymbol{\pi}}$ *and* $\hat{\boldsymbol{\lambda}}$ *are asymptotically independent.*

By Theorems 2 and 3 and Lemma 3, we know the limiting distribution of $\hat{\boldsymbol{\gamma}} = \mathbf{Z}^T\mathbf{Y}$ and the joint limiting distribution of $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$. The next lemma gives an independence result that leads to the joint limiting distribution of the entire vector of estimators $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$. The Appendix gives an outline of the proof.

LEMMA 6. *Suppose that* $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds. The ML estimators* $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\pi}}$ *and* $\hat{\boldsymbol{\lambda}}$ *are mutually asymptotically independent.*

The independence result of Lemma 6 is exploited in the derivation of many of the subsequent theoretical results. As an example, it is used to prove the next theorem, which gives the limiting distribution of the expected count ML estimators $\hat{\mathbf{m}} = \mathbf{N}\hat{\boldsymbol{\pi}}$.

THEOREM 4. *Suppose that* $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds. Then* $\hat{\mathbf{m}}$ *and* $\hat{\boldsymbol{\lambda}}$ *are asymptotically independent and*

$$\nu^{-1/2}(\hat{\mathbf{m}} - \mathbf{m}) \xrightarrow{d} N\big(\mathbf{0}, \mathbf{WD} - \mathbf{DH}\big[\mathbf{H}^T\mathbf{DW}^{-1}\mathbf{H}\big]^{-1}\mathbf{H}^T\mathbf{D} - \mathbf{DWZ}_F\mathbf{Z}_F^T\mathbf{D}\big).$$

PROOF. Note that $\nu^{-1/2}(\hat{\mathbf{m}} - \mathbf{m}) = \mathbf{W}\nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + \mathbf{DZ}\nu^{-1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_P(1)$. However, by Lemma 6, the two summands are asymptotically independent. Their normal limiting distributions are given in Theorem 3 and Lemma 3, respectively. Some algebra leads to the simplified form of the limiting variance. □

REMARK. Haberman [(1974), pages 78 and 87] proved, under certain restrictions, that for log-linear models, which impose constraints of the special form $\mathbf{h}(\mathbf{m}) = \mathbf{U}^T \log \mathbf{m} = \mathbf{0}$, the asymptotic distributions of $\hat{\mathbf{m}}$ and $\hat{\boldsymbol{\pi}}$, respectively, do and do not depend on the sampling constraint. Theorems 3 and 4 generalize these log-linear model results, in the case of independent sampling, to the broader class of MPH models. For MPH models, the limiting distribution of $\hat{\boldsymbol{\pi}}$ depends only on the population matrix $\mathbf{Z}$; it does not depend on the sampling constraint matrix $\mathbf{Z}_F$. The limiting distribution of $\hat{\mathbf{m}}$ depends only on the sampling constraint matrix $\mathbf{Z}_F$; it does not depend on the population matrix $\mathbf{Z}$.

For example, by Theorem 3, whether $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|0)$ (i.e., $\mathbf{Y}$ is Poisson) or $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}, \mathbf{n})$ (i.e., $\mathbf{Y}$ is product multinomial), the ML estimator $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ will have the same limiting distribution. By Theorem 4, whether $\mathbf{Y} \sim \mathrm{MPH}_{Z_1}(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ or $\mathbf{Y} \sim \mathrm{MPH}_{Z_2}(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$, the ML estimator $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ will have the same limiting distribution.

This section closes with a discussion of three of the more common goodness-of-fit statistics (Wald's W^2, Pearson's X^2 and Wilks' likelihood ratio G^2) for testing whether or not $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. They have the following forms [see Silvey (1959)]:

$$\begin{aligned}
W^2(\mathbf{Y}) &= \mathbf{h}(\mathbf{Y})^T\big[\mathbf{H}(\mathbf{Y})^T\mathbf{D}(\mathbf{Y})\mathbf{H}(\mathbf{Y})\big]^{-1}\mathbf{h}(\mathbf{Y}) = \mathbf{h}(\mathbf{Y})^T[\mathrm{avar}(\mathbf{h}(\mathbf{Y}))]^{-1}\mathbf{h}(\mathbf{Y}),\\
X^2(\mathbf{Y}) &= (\mathbf{Y} - \hat{\mathbf{m}})^T\mathbf{D}^{-1}(\hat{\mathbf{m}})(\mathbf{Y} - \hat{\mathbf{m}})\\
&= \hat{\boldsymbol{\lambda}}^T\mathbf{H}(\hat{\mathbf{m}})^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}^T[\mathrm{avar}(\hat{\boldsymbol{\lambda}})]^{-1}\hat{\boldsymbol{\lambda}},\\
G^2(\mathbf{Y}) &= 2\mathbf{Y}^T\log(\mathbf{Y}/\hat{\mathbf{m}}) - 2\big[\mathbf{1}^T\mathbf{Z}_R^T(\mathbf{Y} - \hat{\mathbf{m}})\big] = 2\mathbf{Y}^T\log(\mathbf{Y}/\hat{\mathbf{m}}).
\end{aligned}$$

That Pearson's form of X^2 can be written as a quadratic form in the Lagrange multipliers is a consequence of the form of the restricted likelihood equations (7.2); for example, $\mathbf{Y} - \hat{\mathbf{m}} = -\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}}$. The simplification of the form of G^2 follows because Proposition 6 along with the form of the restricted likelihood equations (7.2) implies that $\mathbf{1}^T\mathbf{Z}_R^T(\mathbf{Y} - \hat{\mathbf{m}}) = \mathbf{1}^T\mathbf{Q}_R^T\mathbf{Z}^T(\mathbf{Y} - \hat{\mathbf{m}}) = 0$.

The next theorem gives the null limiting distribution of all three goodness-of-fit statistics. The proof exploits the asymptotic equivalence of these statistics. Proofs

of the asymptotic equivalences for certain classes of models exist in the literature [see Silvey (1959), Haberman (1974), page 99, and Agresti (1990), page 434]. For completeness, the proofs of these general MPH limiting results are outlined in the Appendix, using notation presented herein.

THEOREM 5. *Suppose that the sequence of models* $\mathbf{Y} \sim \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds. Then* $W^2(\mathbf{Y}) = X^2(\mathbf{Y}) + o_P(1) = G^2(\mathbf{Y}) + o_P(1) \to_d \chi^2(u)$ *as* $\nu \to \infty$, *where* $\chi^2(u)$ *is a central chi-squared random variable with degrees of freedom* $u = \dim(\mathbf{h})$, *the dimension of the model constraint function* $\mathbf{h}$.

Interestingly, the common limiting distribution does not depend on the true parameter values $\mathbf{m}$ or the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$.

7.3. *Asymptotic-based approximation results.* Most of the asymptotic results above are not directly applicable in practice, because (i) $\mathbf{w}$ and ν are not identifiable parameters and (ii) the limiting distributions have variances that depend on the unknown parameter $\boldsymbol{\pi}$. This section uses a generalization of "asymptotic normality" as described in Serfling (1980) to formally state more directly applicable approximation results based on asymptotic arguments.

The next definition shows how one can use an asymptotic result to obtain an approximation result when orders of convergence are allowed to vary, as they will for homogeneous models of nonzero orders.

DEFINITION 6. The sequence $\mathbf{U}_\alpha$ is said to have an *approximate normal distribution* with mean $\boldsymbol{\mu}_\alpha$ and variance $\mathbf{V}_\alpha$, denoted $\mathbf{U}_\alpha \sim \widehat{\mathrm{AN}}(\boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha)$, if, as $0 < \alpha \to \infty$, (i) $\alpha^s \mathbf{A}^{\mathbf{p}}(\mathbf{U}_\alpha - \boldsymbol{\mu}_\alpha) \to_d N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_\alpha$ is a constant sequence, $\boldsymbol{\Sigma}$ has positive diagonal terms, $\mathbf{A}^{\mathbf{p}} = \mathrm{diag}\{a_i^{p_i}\}$ and $a_i/\alpha \to b_i > 0$, and (ii) the sequence of deterministic or stochastic matrices $\mathbf{V}_\alpha$ satisfies $\alpha^{2s}\mathbf{A}^{\mathbf{p}}\mathbf{V}_\alpha\mathbf{A}^{\mathbf{p}} \to_P \boldsymbol{\Sigma}$. The matrix $\mathbf{V}_\alpha$ is called an *approximating variance* of $\mathbf{U}_\alpha$ and is denoted $\mathrm{avar}(\mathbf{U}_\alpha) = \mathbf{V}_\alpha$.

The $\widehat{\mathrm{AN}}$ (approximate normal) notation is meant to resemble Serfling's (1980) AN (asymptotic normal) notation; the circumflex over the AN indicates that the sequence of approximating variances $\mathbf{V}_\alpha$ is allowed to be stochastic. (Serfling's AN definition does not allow for stochastic $\mathbf{V}_\alpha$.) Sometimes the orders of convergence are noted as well. The sequence $\mathbf{U}_\alpha - \boldsymbol{\mu}_\alpha$ of Definition 6 is of order $O_P(\alpha^{-(s+\mathbf{p})})$ because the ith component $U_{\alpha,i} - \mu_{\alpha,i}$, when multiplied by α^{s+p_i}, converges in distribution to a normal random variable with nondegenerate variance.

The justification for Definition 6 is based on the following observations. (i) It can be shown using standard limiting results (e.g., Slutsky's theorem) analogous

to the asymptotic normal arguments used in Serfling (1980) that

$$\frac{\boldsymbol{\theta}^T(\mathbf{U}_\alpha - \boldsymbol{\mu}_\alpha)}{\sqrt{\boldsymbol{\theta}^T\mathbf{V}_\alpha\boldsymbol{\theta}}} \xrightarrow{d} N(0,1) \qquad \forall\,\boldsymbol{\theta} \in \Lambda(\mathbf{B}^{-\mathbf{p}}\boldsymbol{\Sigma}\mathbf{B}^{-\mathbf{p}}, \mathbf{p}),$$

where $\mathbf{B}^{-\mathbf{p}} = \operatorname{diag}\{b_i^{-p_i}\}$, $\Lambda(\boldsymbol{\Gamma}, \mathbf{p}) = \{\boldsymbol{\theta} \neq \mathbf{0} : \boldsymbol{\tau}(\boldsymbol{\theta})^T\boldsymbol{\Gamma}\boldsymbol{\tau}(\boldsymbol{\theta}) > 0\}$, $\boldsymbol{\tau}(\boldsymbol{\theta}) = \operatorname{diag}\{I(p_i = p(\boldsymbol{\theta}))\}\boldsymbol{\theta}$ and $p(\boldsymbol{\theta}) = \min\{p_i : \theta_i \neq 0\}$. (ii) Because $\boldsymbol{\Sigma}$ and $\mathbf{B}^{-\mathbf{p}}$ have nonzero diagonal terms, it follows that the elementary vectors of the form $\boldsymbol{\theta} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ fall in $\Lambda(\mathbf{B}^{-\mathbf{p}}\boldsymbol{\Sigma}\mathbf{B}^{-\mathbf{p}}, \mathbf{p})$. Thus, $(U_{\alpha i} - \mu_{\alpha i})/\sqrt{V_{\alpha,ii}} \to_d N(0,1)$, $i = 1, \ldots, c$. Note also that if $\boldsymbol{\Sigma}$ is positive definite, then $\Lambda(\mathbf{B}^{-\mathbf{p}}\boldsymbol{\Sigma}\mathbf{B}^{-\mathbf{p}}, \mathbf{p}) = R^c - \{\mathbf{0}\}$, that is, every nonzero vector $\boldsymbol{\theta}$ belongs to the set. (iii) Finally, note that $\alpha^{s+p(\boldsymbol{\theta})}\boldsymbol{\theta}^T(\mathbf{U}_\alpha - \boldsymbol{\mu}_\alpha) \to_d N(0, \sigma_{\boldsymbol{\theta}^2})$, where $\sigma_{\boldsymbol{\theta}^2} \equiv \boldsymbol{\tau}(\boldsymbol{\theta})^T\mathbf{B}^{-\mathbf{p}}\boldsymbol{\Sigma}\mathbf{B}^{-\mathbf{p}}\boldsymbol{\tau}(\boldsymbol{\theta})$. It can be shown that, for any deterministic sequence $\{q_\alpha\}$, it follows that

$$\left|P(\boldsymbol{\theta}^T\mathbf{U}_\alpha \leq q_\alpha) - P(N_\alpha \leq q_\alpha|\mathbf{V}_\alpha)\right| \xrightarrow{P} 0 \qquad \text{where } N_\alpha|\mathbf{V}_\alpha \sim N(\boldsymbol{\theta}^T\boldsymbol{\mu}_\alpha, \boldsymbol{\theta}^T\mathbf{V}_\alpha\boldsymbol{\theta}). \tag{7.3}$$

Serfling's definition of asymptotic normality follows as a special case of approximate normality, when $\mathbf{p} = \mathbf{0}$ and $\mathbf{V}_\alpha$ is nonstochastic. We note that the stochastic convergence of the $\widehat{\text{AN}}$ result (7.3) can be replaced by deterministic convergence in the AN case of Serfling.

In practice, the approximating variance $\mathbf{V}_\alpha = \operatorname{avar}(\mathbf{U}_\alpha)$ in $\mathbf{U}_\alpha \sim \widehat{\text{AN}}(\boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha)$ is chosen so that it is an identifiable estimator. To illustrate, consider two independent sequences of Poisson random variables $Y_{i\nu} \sim \text{Po}(\nu)$, $i = 1, 2$. It is straightforward to see that $\nu^{-1/2}(Y_{1\nu} - Y_{2\nu}) \to_d N(0, 2)$. By definition, either of 2ν or $Y_{1\nu} + Y_{2\nu}$ could serve as an approximating variance of $(Y_{1\nu} - Y_{2\nu})$, because $\nu^{-1}2\nu$ and $\nu^{-1}(Y_{1\nu} + Y_{2\nu})$ both converge in probability to 2. We would choose $\operatorname{avar}(Y_{1\nu} - Y_{2\nu}) = (Y_{1\nu} + Y_{2\nu})$ because, unlike 2ν, $(Y_{1\nu} + Y_{2\nu})$ is an estimator. We write $Y_{1\nu} - Y_{2\nu} \sim \widehat{\text{AN}}(0, Y_{1\nu} + Y_{2\nu}) = O_P(\nu^{1/2})$. An estimator of $P(Y_{1\nu} - Y_{2\nu} \leq 10)$ is $P(N_\nu \leq 10|Y_{1\nu}, Y_{2\nu})$, where $N_\nu|Y_{1\nu}, Y_{2\nu} \sim N(0, Y_{1\nu} + Y_{2\nu})$. As an example, when $\nu = 100$, $P(Y_{1\nu} - Y_{2\nu} \leq 10) = 0.771$. Five realizations of the estimator $P(N_\nu \leq 10|Y_{1\nu}, Y_{2\nu})$ were 0.766, 0.753, 0.770, 0.761 and 0.765.

Several useful approximation results are given in the next theorem. Here, $\mathbf{Y}$ is viewed as a member of the sequence $\{\mathbf{Y}_\nu\}$ defined in the previous section. The Appendix outlines the proofs of Ax3 and Ax7. The other proofs follow analogously.

THEOREM 6 (MPH approximation results). *Suppose that* $\mathbf{Y} \sim \text{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds, with* $\mathbf{h} \in \mathcal{H}''_{\mathbf{p}}(\mathbf{Z})$. *Then the following approximation results* Ax1 *through* Ax8 *are valid, with the approximations improving as the components in* $\mathbf{m}$ *increase. For convenience, define* $\hat{\mathbf{D}} \equiv \mathbf{D}(\hat{\mathbf{m}})$ *and* $\hat{\mathbf{H}} \equiv \mathbf{H}(\hat{\mathbf{m}})$.

Ax1. $\mathbf{N}^{-1}\mathbf{Y} - \boldsymbol{\pi} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \mathbf{N}^{-1}[\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}\mathbf{Z}^T\hat{\mathbf{D}}]\mathbf{N}^{-1}) = O_P(\nu^{-1/2})$.

Ax2. $\hat{\boldsymbol{\pi}} - \boldsymbol{\pi} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \mathbf{N}^{-1}[\hat{\mathbf{D}} - \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}\mathbf{Z}^T\hat{\mathbf{D}}]\mathbf{N}^{-1}) = O_P(\nu^{-1/2})$.

Ax3. $\mathbf{Y} - \mathbf{m} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}) = O_P(\nu^{1/2})$.

Ax4. $\hat{\mathbf{m}} - \mathbf{m} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{D}} - \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}) = O_P(\nu^{1/2})$.

Ax5. $\mathbf{Y} - \hat{\mathbf{m}} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}}) = O_P(\nu^{1/2})$.

Ax6. $\hat{\boldsymbol{\lambda}} \sim \widehat{\mathrm{AN}}(\mathbf{0}, [\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}}]^{-1}) = O_P(\nu^{1/2-\mathbf{p}})$.

Ax7. $\mathbf{h}(\mathbf{Y}) - \mathbf{h}(\mathbf{m}) \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}}) = O_P(\nu^{\mathbf{p}-1/2})$.

Ax8. $\log\hat{\mathbf{m}} - \log\mathbf{m} \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{D}}^{-1} - \hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T - \mathbf{N}^{-1}\mathbf{Z}_F\mathbf{Z}_F^T) = O_P(\nu^{-1/2})$.

Several remarks are in order.

1. Ax1 and Ax2 imply that the probability estimators have approximating variances that depend only on the population matrix $\mathbf{Z}$; they do not depend on the sampling constraints.
2. Ax2 and Ax4 taken together indicate how one can use the approximating variance of $\hat{\mathbf{m}}$ to directly compute the approximating variance of $\hat{\boldsymbol{\pi}}$ without resorting to the delta method.
3. Because $\mathbf{N}^{-1}\mathbf{Z}_F\mathbf{Z}_F^T$ can be shown to equal $\mathbf{Z}_F\mathbf{D}^{-1}(\mathbf{Z}_F^T\mathbf{Y})\mathbf{Z}_F^T$, Ax3 and Ax4 imply that the expected count estimators have approximating variances that depend on the the sampling constraint matrix $\mathbf{Z}_F$, but *not* on the population matrix $\mathbf{Z}$. This means, for example, that we can always use $\mathbf{Z} = \mathbf{1}$ for convenience when considering the approximating distributions of expected count estimators $\hat{\mathbf{m}}$ for Poisson sampling, that is, when $\mathbf{Z}_F = \mathbf{0}$.
4. Ax3–Ax5 taken together imply that the residuals $\mathbf{Y} - \hat{\mathbf{m}}$ and fitted values $\hat{\mathbf{m}}$ for MPH models are asymptotically independent, because their approximating variances add up to the approximating variance of $\mathbf{Y} - \mathbf{m}$.
5. Ax5 implies that the approximating variance of the residuals does not depend on the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$.
6. Ax6 and Ax7 imply that the approximating distributions of $\hat{\boldsymbol{\lambda}}$ and $\mathbf{h}(\mathbf{Y})$ do not depend on the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$.

8. Equivalence results for MPH models. In contingency table analyses, it is common practice to exploit equivalences between certain models. As a simple example, when faced with fitting a product-multinomial log-linear model, one might choose to fit the "equivalent" Poisson log-linear model for convenience.

This section formalizes the notion of model equivalence. Specifically, we introduce a formal definition of model equivalence and explicitly compare the ML fit results for two equivalent models. This has both theoretical and computational utility.

8.1. *Equivalence classes of MPH models.* Let $U(\mathbf{y}) \equiv \{\mathcal{M} : \mathcal{M}$ is a MPH model for data $\mathbf{y}\}$. Notice that any model $\mathcal{M}$ in $U(\mathbf{y})$ can be written as $\mathcal{M} = \mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ for some sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ and some constraint function $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$; the sample sizes $\mathbf{n}$ are determined by $\mathbf{Z}_F$ and $\mathbf{y}$ through $\mathbf{n} = \mathbf{Z}_F^T\mathbf{y}$.

DEFINITION 7. Two models $\mathcal{M}_1, \mathcal{M}_2 \in U(\mathbf{y})$ are *equivalent*, denoted $\mathcal{M}_1 \approx \mathcal{M}_2$, if there exist sampling plans $(\mathbf{Z}_1, \mathbf{Z}_{1F}, \mathbf{n}_1)$ and $(\mathbf{Z}_2, \mathbf{Z}_{2F}, \mathbf{n}_2)$ and a constraint function $\mathbf{h} \in \mathcal{H}''(\mathbf{Z}_1) \cap \mathcal{H}''(\mathbf{Z}_2)$ such that $\mathcal{M}_1 = \mathrm{MPH}_{Z_1}(\mathbf{h}|\mathbf{Z}_{1F}, \mathbf{n}_1)$ and $\mathcal{M}_2 = \mathrm{MPH}_{Z_2}(\mathbf{h}|\mathbf{Z}_{2F}, \mathbf{n}_2)$. If, in addition, the population matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are identical (the sampling constraint matrices $\mathbf{Z}_{1F}$ and $\mathbf{Z}_{2F}$ need not be identical), then the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are *population equivalent*, denoted $\mathcal{M}_1 \overset{P}{\approx} \mathcal{M}_2$.

Loosely, two MPH models that can be specified in terms of the same constraints on the expected table counts are equivalent. Two equivalent MPH models that are based on the same population matrix are population equivalent. Obviously, population equivalence is a stronger version of equivalence. That is, if two models are population equivalent, then they are equivalent; the converse is not true.

It will be useful to introduce a notation for the equivalence classes that corresponds to the equivalence relationships "$\approx$" and "$\approx_P$." Let $\mathcal{E}(\mathbf{h}, \mathbf{y}) \equiv \{$MPH models for $\mathbf{y}$ that can be specified using constraint function $\mathbf{h}\}$ and let subset $\mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y}) \equiv \{$MPH models for $\mathbf{y}$, with population matrix $\mathbf{Z}$, that can be specified using constraint function $\mathbf{h}\}$. The set $\mathcal{E}(\mathbf{h}, \mathbf{y})$ is an equivalence class induced by "$\approx$" in that any two models in $\mathcal{E}(\mathbf{h}, \mathbf{y})$ are equivalent. Similarly, the set $\mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y})$ is an equivalence class induced by "$\approx_P$" in that any two models in $\mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y})$ are population equivalent.

Note that although $U(\mathbf{y}) = \bigcup_{h \in \mathcal{H}''} \mathcal{E}(\mathbf{h}, \mathbf{y}) = \bigcup_{h \in \mathcal{H}''} \bigcup_{Z \in \mathcal{P}} \mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y})$, the sets $\mathcal{E}(\mathbf{h}, \mathbf{y})$ are not disjoint; neither are the sets $\mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y})$. For example, $\mathcal{E}(3\mathbf{h}, \mathbf{y})$ and $\mathcal{E}(\mathbf{h}, \mathbf{y})$ are identical. Herein, we have no need for a disjoint partition of $U(\mathbf{y})$; the equivalence classes $\mathcal{E}(\mathbf{h}, \mathbf{y})$ and $\mathcal{E}(\mathbf{h}, \mathbf{Z}, \mathbf{y})$ will suffice.

The choice of equivalence relationship definition is reasonable because, when models are equivalent, we will show that, among other things, (i) expected count ML estimates are identical, (ii) goodness-of-fit statistics are identical and (iii) adjusted residuals [see Haberman (1973)] are identical.

8.2. *Numerical equivalence and comparison results.* Theorem 7 gives a collection of useful comparisons between ML fit results for two equivalent data models, $\mathbf{y} \leftarrow \mathbf{Y}_1 \sim \mathcal{M}_1$ and $\mathbf{y} \leftarrow \mathbf{Y}_2 \sim \mathcal{M}_2$. In the statement of the theorem, maximum likelihood estimates and goodness-of-fit statistics for model $\mathcal{M}_i$ are subscripted with an i, $i = 1, 2$. The symbol $\mathbf{N}_i \equiv \mathbf{D}(\mathbf{Z}_i\mathbf{Z}_i^T\mathbf{Y}_i) = \mathbf{D}(\mathbf{Z}_i\hat{\boldsymbol{\gamma}}_i)$ for $i = 1, 2$ and $\hat{\mathbf{D}} = \mathbf{D}(\hat{\mathbf{m}}_1) = \mathbf{D}(\hat{\mathbf{m}}_2)$. The approximating variances, as derived in Section 7.3 and denoted avar, are viewed as nonrandom estimates, evaluated using

the observed data $\mathbf{y}$. All of these results are direct consequences of the numerical results of Section 7.1 and the approximation results of Section 7.3; hence, proofs are omitted.

THEOREM 7. *Consider two candidate data models,* $\mathbf{y} \leftarrow \mathbf{Y}_1 \sim \mathcal{M}_1$ *and* $\mathbf{y} \leftarrow \mathbf{Y}_2 \sim \mathcal{M}_2$. *Assume that* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are equivalent in that they both are members of the same equivalence class* $\mathcal{E}(\mathbf{h}, \mathbf{y})$. *For specificity, assume that* $\mathcal{M}_i = \mathrm{MPH}_{Z_i}(\mathbf{h}|\mathbf{Z}_{iF}, \mathbf{n}_i)$, $i = 1, 2$. *Then the following numerical equivalences and comparisons hold.*

NE1. $\hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2$.

NE2. $\mathrm{avar}(\hat{\mathbf{m}}_1) - \mathrm{avar}(\hat{\mathbf{m}}_2) = \mathbf{N}_2^{-1}\hat{\mathbf{D}}\mathbf{Z}_{2F}\mathbf{Z}_{2F}^T\hat{\mathbf{D}} - \mathbf{N}_1^{-1}\hat{\mathbf{D}}\mathbf{Z}_{1F}\mathbf{Z}_{1F}^T\hat{\mathbf{D}}$.

NE3. $\mathrm{avar}(\log \hat{\mathbf{m}}_1) - \mathrm{avar}(\log \hat{\mathbf{m}}_2) = \mathbf{N}_2^{-1}\mathbf{Z}_{2F}\mathbf{Z}_{2F}^T - \mathbf{N}_1^{-1}\mathbf{Z}_{1F}\mathbf{Z}_{1F}^T$.

NE4. $\hat{\boldsymbol{\pi}}_1 = \mathbf{N}_1^{-1}\mathbf{N}_2\hat{\boldsymbol{\pi}}_2$. *If, in addition,* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are* population *equivalent, then* $\hat{\boldsymbol{\pi}}_1 = \hat{\boldsymbol{\pi}}_2$.

NE5. $\mathrm{avar}(\hat{\boldsymbol{\pi}}_1) - \mathrm{avar}(\hat{\boldsymbol{\pi}}_2)$ *depends only on* $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\hat{\mathbf{m}} = \hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2$ *and* $\mathbf{H}(\hat{\mathbf{m}})$. *The difference can be computed using* Ax2. *If, in addition,* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are* population *equivalent, then* $\mathrm{avar}(\hat{\boldsymbol{\pi}}_1) = \mathrm{avar}(\hat{\boldsymbol{\pi}}_2)$.

NE6. $\hat{\boldsymbol{\lambda}}_1 = \hat{\boldsymbol{\lambda}}_2$.

NE7. $\mathrm{avar}(\hat{\boldsymbol{\lambda}}_1) = \mathrm{avar}(\hat{\boldsymbol{\lambda}}_2)$.

NE8. $\mathrm{avar}(\mathbf{h}(\mathbf{Y}_1)) = \mathrm{avar}(\mathbf{h}(\mathbf{Y}_2))$.

NE9. $\mathrm{avar}(\mathbf{Y}_1 - \hat{\mathbf{m}}_1) = \mathrm{avar}(\mathbf{Y}_2 - \hat{\mathbf{m}}_2)$.

NE10. *The adjusted residuals, defined as* $\hat{r}_{it} = (y_t - \hat{m}_{it})/\sqrt{\mathrm{avar}(Y_{it} - \hat{m}_{it})}$, $i = 1, 2$, *are numerically identical* (*i.e.*, $\hat{r}_{1t} = \hat{r}_{2t}, t = 1, \ldots, c$). *Here,* $\hat{m}_{it}$ *is the* t*th fitted value for model* $\mathcal{M}_i$.

NE11. $W_1^2(\mathbf{y}) = W_2^2(\mathbf{y})$, $X_1^2(\mathbf{y}) = X_2^2(\mathbf{y})$ *and* $G_1^2(\mathbf{y}) = G_2^2(\mathbf{y})$.

Inspection of the form of the differences between point estimates or approximate variance estimates for two equivalent models indicates the practical utility of these numerical equivalence results. Specifically, any numerical difference can be explicitly computed using the ML fit results of either equivalent model. This means the ML fit results for any particular MPH model can be obtained directly from the ML fit results of any conveniently chosen equivalent model.

Birch (1963) gave a necessary and sufficient condition for numerical equivalence NE1 of Theorem 7 when a Poisson model is compared to a MP model. Restated in terms of constraint models, Birch's result implies that for the two models $\mathcal{M} = \mathrm{MP}_Z(\mathbf{h}|0)$ and $\mathcal{M}_2 = \mathrm{MP}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n} = \mathbf{Z}_F^T\mathbf{y})$, the fitted values are numerically equal (i.e., $\hat{\mathbf{m}} = \hat{\mathbf{m}}_2$) if and only if $\mathbf{Z}_F^T\hat{\mathbf{m}} = \mathbf{n}$. In practice, with the exception of special cases like log-linear models [see Haberman (1974)], it is not always easy to know what models will generally lead to $\mathbf{Z}_F^T\hat{\mathbf{m}} = \mathbf{n}$.

Here, we give a very general and simple to verify sufficient condition for when the identity $\mathbf{Z}_F^T\hat{\mathbf{m}} = \mathbf{n}$ holds. Namely, it holds when $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$. This can be argued

as follows: by Proposition 6, $\mathbf{Z}_F^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}}) = \mathbf{0}$. This implies that when the first set of equations in (7.2) is multiplied by $\mathbf{Z}_F^T$, the identity $\mathbf{Z}_F^T\hat{\mathbf{m}} = \mathbf{Z}_F^T\mathbf{y} = \mathbf{n}$ is obtained. Bergsma (1997) implicitly used a result similar to Proposition 6 to assert that Poisson and multinomial point estimates are identical in the special case of zero-order homogeneous constraint models.

Numerical equivalence NE3 affords a straightforward comparison of the approximating variances of the linear predictors in any two equivalent log-linear models for data $\mathbf{y}$. For example, the difference between the approximating variance estimate of $\log\hat{\mathbf{m}}$ for a Poisson log-linear model and a multinomial–Poisson log-linear model with sampling constraint matrix $\mathbf{Z}_F$ is simply $\mathbf{N}^{-1}\mathbf{Z}_F\mathbf{Z}_F^T = \mathbf{Z}_F\mathbf{D}^{-1}(\mathbf{Z}_F^T\mathbf{y})\mathbf{Z}_F^T$.

EXAMPLE 8.1 (Log-linear models). Suppose that the counts in $\mathbf{y}$ are observed and the log-linear model of the form $\log\mathbf{m} = \mathbf{X}\boldsymbol{\beta}$, or equivalently, $\mathbf{h}(\mathbf{m}) = \mathbf{U}^T\log\mathbf{m} = \mathbf{0}$, is considered. Provided $R(\mathbf{X})$ contains both $R(\mathbf{Z}_1)$ and $R(\mathbf{Z}_2)$, the Poisson log-linear model $\mathcal{M}_1 = \mathrm{MPH}_{Z_1}(\mathbf{h}|0)$ and the product-multinomial log-linear model $\mathcal{M}_2 = \mathrm{MPH}_{Z_2}(\mathbf{h}|\mathbf{Z}_2, \mathbf{n})$ are equivalent [i.e., $\mathcal{M}_1, \mathcal{M}_2 \in \mathcal{E}(\mathbf{h}, \mathbf{y})$]. It follows that the many numerical equivalences outlined above hold. For example, we obtain the well-known result [see Bishop, Fienberg and Holland (1975), Agresti (1990) and Fienberg (2000)] that the equivalent log-linear model fitted values are identical (i.e., $\hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2$). Furthermore, by the numerical comparison result NE3 of this section, $\mathrm{avar}(\log\hat{\mathbf{m}}_1) - \mathrm{avar}(\log\hat{\mathbf{m}}_2) = \mathbf{Z}_2\mathbf{D}^{-1}(\mathbf{Z}_2^T\mathbf{y})\mathbf{Z}_2^T$. This result leads directly to many useful equivalence results for the log-linear coefficient estimators in $\hat{\boldsymbol{\beta}}$, because $\mathbf{X}\hat{\boldsymbol{\beta}}_i = \log\hat{\mathbf{m}}_i$. As an example, if the jth component of $\boldsymbol{\beta}$, say β_j, corresponds to a column in $\mathbf{X}$ that is not needed to span the space of $\mathbf{Z}_2$, then $\mathrm{avar}(\hat{\beta}_{j1}) = \mathrm{avar}(\hat{\beta}_{j2})$. This special case example is well known [see Haberman (1974), Palmgren (1981), Christensen (1990), page 215, and Lang (1996a)].

EXAMPLE 8.2. Consider the hypothesis $h(\mathbf{m}) = \sum_{i=1}^R m_{i+}^2 - \sum_{i=1}^R m_{+i}^2 = 0$ for an $R \times R$ contingency table. The function h is in $\mathcal{H}_2''(\mathbf{1})$, so the Poisson and multinomial data models $\mathcal{M}_1 = \mathrm{MPH}_{\mathbf{1}}(h|0)$ and $\mathcal{M}_2 = \mathrm{MPH}_{\mathbf{1}}(h|\mathbf{1}, n)$ are population equivalent. It follows that $\mathcal{M}_1$ and $\mathcal{M}_2$ have identical fitted values ($\hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2 = \hat{\mathbf{m}}$); the difference $\mathrm{avar}(\hat{\mathbf{m}}_1) - \mathrm{avar}(\hat{\mathbf{m}}_2) = \hat{\mathbf{m}}\hat{\mathbf{m}}^T/n$; the estimated cell probabilities and their approximating variance estimates are identical; the adjusted residuals are identical; and the goodness-of-fit statistic values are identical. We point out that $h(\mathbf{m}) = 0$ if and only if $h(\mathbf{P}) = 0$, so $\mathcal{M}_1$ and $\mathcal{M}_2$ are models of equal marginal Gini concentration.

EXAMPLE 8.3. Consider the hypothesis of a fixed odds ratio, $h(\mathbf{m}) = (m_{11}m_{22})/(m_{12}m_{21}) - 3 = 0$ for a 2×2 table. Let $\mathbf{Z}_1 = \mathbf{1}_4$, $\mathbf{Z}_2 = \mathbf{I}_2 \otimes \mathbf{1}_2$ and $\mathbf{Z}_3 = \mathbf{1}_2 \otimes \mathbf{I}_2$. The function h is in $\mathcal{H}''(\mathbf{Z}_i)$, $i = 1, 2, 3$, so the six data models $\mathrm{MPH}_{Z_i}(h|\mathbf{Z}_i, \mathbf{n}_i)$, $\mathrm{MPH}_{Z_i}(h|0)$, $i = 1, 2, 3$, are all equivalent. By the equivalence

results of this section, the ML fit results for one of the models, say the Poisson model $\mathrm{MPH}_{Z_1}(h|0)$, can be explicitly adjusted to give the ML fit results for the other five equivalent models.

9. Z-homogeneous statistics. Analysis of contingency tables often involves estimation of functions of the expected counts $\mathbf{m}$. Previous sections described estimation of certain functions, such as $\mathbf{S}(\mathbf{m}) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{m})\mathbf{m} = \boldsymbol{\pi}$ and $\mathbf{S}(\mathbf{m}) = \mathbf{m}$. This section explores the broader class of smooth $\mathbf{Z}$-homogeneous functions and describes the limiting behavior of the corresponding ML estimators. These ML estimators, which have the form $\mathbf{S}(\hat{\mathbf{m}})$, are called $\mathbf{Z}$-homogeneous statistics. In practice, many estimators of interest can be written as $\mathbf{Z}$-homogeneous statistics.

DEFINITION 8. The estimator $\mathbf{S}(\hat{\mathbf{m}})$ is a *$\mathbf{Z}$-homogeneous statistic of order* $\mathbf{p}$ if (i) $\hat{\mathbf{m}}$ is the ML estimator under $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$, (ii) $\mathbf{S} \in \mathcal{H}_{\mathbf{p}}(\mathbf{Z})$ and (iii) $\mathbf{S}$ has continuous first-order derivatives at the true $\boldsymbol{\pi}$.

The asymptotic distribution of properly normalized $\mathbf{S}(\hat{\mathbf{m}})$ can be found using Theorem 3 and Lemma 6 in conjunction with the delta method. The proof can be found in the Appendix. In the following $\mathbf{D} \equiv \mathbf{D}(\boldsymbol{\pi})$ and $\mathbf{H} \equiv \mathbf{H}(\boldsymbol{\pi})$.

THEOREM 8. *Suppose that the sequence of MPH models* $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$ *holds. Let* $\mathbf{S}(\hat{\mathbf{m}})$ *be a* $\mathbf{Z}$*-homogeneous statistic of order* $\mathbf{p}$, *where* $\mathbf{S}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{S}(\mathbf{x})$. *Then*

$$\nu^{1/2}\mathbf{G}^{-1}(\boldsymbol{\gamma})\big(\mathbf{S}(\hat{\mathbf{m}}) - \mathbf{S}(\mathbf{m})\big) \xrightarrow{d} N\Big(0, \frac{\partial \mathbf{S}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\big[\boldsymbol{\Sigma}^* - \mathbf{W}^{-1}\mathbf{D}\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D}\big]\frac{\partial \mathbf{S}(\boldsymbol{\pi})^T}{\partial \boldsymbol{\pi}}\Big),$$

where $\boldsymbol{\Sigma}^* = \mathbf{W}^{-1}\mathbf{D} - \mathbf{W}^{-1}\mathbf{D}\mathbf{H}[\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}$.

As in Section 7.3, this asymptotic result can be used to obtain a practically useful approximation result. As before, let $\mathbf{N} \equiv \mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{Y})$, $\hat{\mathbf{D}} \equiv \mathbf{D}(\hat{\mathbf{m}})$ and $\hat{\mathbf{H}} \equiv \mathbf{H}(\hat{\mathbf{m}})$.

COROLLARY 1. *Under the conditions of Theorem* 8, *the following approximation result is obtained*:

$$\begin{aligned}&\mathbf{S}(\hat{\mathbf{m}}) - \mathbf{S}(\mathbf{m})\\ &\quad\sim \widehat{\mathrm{AN}}\Big(0, \frac{\partial \mathbf{S}(\hat{\mathbf{m}})}{\partial \mathbf{m}^T}\big[\hat{\mathbf{D}} - \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}\big]\frac{\partial \mathbf{S}(\hat{\mathbf{m}})^T}{\partial \mathbf{m}}\Big)\\ &\quad= O_P(\nu^{\mathbf{p}-1/2}).\end{aligned}$$

REMARK. In view of Ax4, the approximating variance has the form

$$\frac{\partial \mathbf{S}(\hat{\mathbf{m}})}{\partial \mathbf{m}^T}\operatorname{avar}(\hat{\mathbf{m}})\frac{\partial \mathbf{S}(\hat{\mathbf{m}})^T}{\partial \mathbf{m}}.$$

Because of this result, it is often mistakenly assumed that one can find the approximating distribution of $\mathbf{S}(\hat{\mathbf{m}})$, for *any sufficiently smooth function* $\mathbf{S}$, by formally applying the delta method directly to $\mathbf{S}(\hat{\mathbf{m}})$. Corollary 1 shows that when $\mathbf{S}$ is $\mathbf{Z}$-homogeneous this is true. However, it is *not* true in general. Consider an example from Bishop, Fienberg and Holland [(1975), pages 489 and 490]. When $Y \sim \text{Po}(m)$, the approximating distribution of $S_1(Y) = Y^r$ can be found by formally applying the delta method because S_1 is homogeneous of order r; thus, $Y^r \sim \text{AN}(m^r, r^2 m^{2r-1})$ or $Y^r \sim \widehat{\text{AN}}(m^r, r^2 Y^{2r-1})$. In contrast, the approximating distribution of the nonhomogeneous function $S_2(Y) = \exp(Y)$ cannot be found using a formal application of the delta method; that is, $\exp(Y)$ is *not* $\text{AN}(\exp(m), m\exp(2m))$. We point out that there are classes of $\mathbf{S}$ functions other than $\mathbf{Z}$-homogeneous functions for which a formal application of the delta method works. For example, it can be shown that the formal method works if $\mathbf{S}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{a}(\boldsymbol{\gamma}) + \mathbf{S}(\mathbf{x})$, where $\mathbf{a}(\boldsymbol{\gamma}_1) - \mathbf{a}(\boldsymbol{\gamma}_2) = \mathbf{a}(\boldsymbol{\gamma}_1/\boldsymbol{\gamma}_2) - \mathbf{a}(\mathbf{1})$. As a special case example, the function $\mathbf{S}(\mathbf{m}) = \log(\mathbf{m})$ satisfies this condition.

If $\mathbf{S}$ is $\mathbf{Z}$-homogeneous of order 0 [i.e., $\mathbf{S}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{S}(\mathbf{x})$], the approximation result of Corollary 1 simplifies because $\mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}(\partial \mathbf{S}(\hat{\mathbf{m}})^T)/\partial \mathbf{m} = \mathbf{0}$ by Proposition 5. The result can be stated as follows.

COROLLARY 2. *Under the conditions of Theorem* 8, *if* $\mathbf{S}$ *is* $\mathbf{Z}$*-homogeneous of order* 0, *then*

$$\begin{aligned}\mathbf{S}(\hat{\mathbf{m}}) - \mathbf{S}(\mathbf{m}) &\sim \widehat{\text{AN}}\Big(0, \frac{\partial \mathbf{S}(\hat{\mathbf{m}})}{\partial \mathbf{m}^T}[\hat{\mathbf{D}} - \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}}]\frac{\partial \mathbf{S}(\hat{\mathbf{m}})^T}{\partial \mathbf{m}}\Big)\\ &= O_P(\nu^{-1/2}).\end{aligned}$$

The previous two corollaries can be used to compare the approximating distributions of $\mathbf{Z}$-homogeneous statistics for equivalent models. We state the result in the form of a corollary.

COROLLARY 3. *Suppose that* $\mathbf{y}$ *is observed and that* $\mathcal{M}_1 = \text{MPH}_{Z_1}(\mathbf{h}|\mathbf{Z}_{1F}, \mathbf{n}_1)$ *and* $\mathcal{M}_2 = \text{MPH}_{Z_2}(\mathbf{h}|\mathbf{Z}_{2F}, \mathbf{n}_2)$ *are any two equivalent models in* $\mathcal{E}(\mathbf{h}, \mathbf{y})$. *If* $\mathbf{S}(\hat{\mathbf{m}}_i)$ *is a* $\mathbf{Z}_i$*-homogeneous statistic for* $i = 1, 2$, *then* $\mathbf{S}(\hat{\mathbf{m}}_1)$ *and* $\mathbf{S}(\hat{\mathbf{m}}_2)$ *are numerically identical and*

$$\begin{aligned}&\operatorname{avar}\big(\mathbf{S}(\hat{\mathbf{m}}_1)\big) - \operatorname{avar}\big(\mathbf{S}(\hat{\mathbf{m}}_2)\big)\\ &\quad= \frac{\partial \mathbf{S}(\hat{\mathbf{m}})}{\partial \mathbf{m}^T}[\mathbf{N}_2^{-1}\hat{\mathbf{D}}\mathbf{Z}_{2F}\mathbf{Z}_{2F}^T\hat{\mathbf{D}} - \mathbf{N}_1^{-1}\hat{\mathbf{D}}\mathbf{Z}_{1F}\mathbf{Z}_{1F}^T\hat{\mathbf{D}}]\frac{\partial \mathbf{S}(\hat{\mathbf{m}})^T}{\partial \mathbf{m}},\end{aligned}$$

where $\hat{\mathbf{m}} = \hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2$. *If, in addition,* $\mathbf{S}$ *is of order* 0, *then* $\operatorname{avar}(\mathbf{S}(\hat{\mathbf{m}}_1)) = \operatorname{avar}(\mathbf{S}(\hat{\mathbf{m}}_2))$.

EXAMPLE 9.1 (Log odds ratio estimator). Suppose that outcomes are classified according to dichotomous variables $(A, B) \sim \mathbf{P}$ and that the vector of counts $\mathbf{y} = (y_{11}, y_{12}, y_{21}, y_{22})^T = (32, 18, 12, 38)^T$ is observed. Here, y_{ij} = number of $(A = i, B = j)$ events. It is of interest to estimate the log odds ratio $S(\mathbf{P}) = \log(P_{11}P_{22}/P_{12}P_{21})$.

Consider the following three equivalent—they all fall in $\mathcal{E}(0, \mathbf{y})$—unrestricted data models: $\mathbf{y} \leftarrow \mathbf{Y} \sim \mathrm{MPH}_{Z_i}(0|\mathbf{Z}_{iF}, \mathbf{n}_i), i = 1, 2, 3$, where $\mathbf{Z}_1 = \mathbf{1}_4, \mathbf{Z}_2 = \mathbf{I}_2 \otimes \mathbf{1}_2, \mathbf{Z}_3 = \mathbf{1}_2 \otimes \mathbf{I}_2$ and $\mathbf{Z}_{iF}^T\mathbf{y} = \mathbf{n}_i$. Because the log odds ratio function S is in $\mathcal{H}_0(\mathbf{Z}_i)$ for $i = 1, 2, 3$, $S(\mathbf{P}) = S(\mathbf{m})$ for any of the three data models. This implies that the unrestricted ML estimators of $S(\mathbf{P})$ have the form $S(\hat{\mathbf{m}}_i) = S(\mathbf{Y})$. Now, the estimator $S(\hat{\mathbf{m}}_i) = S(\mathbf{Y})$ is a $\mathbf{Z}_i$-homogeneous statistic of order 0, $i = 1, 2, 3$. Thus, not only are the $S(\hat{\mathbf{m}}_i)$ numerically identical [and equal to $S(\mathbf{y}) = 1.728044$], but Corollary 3 states that they have the same approximating variance estimate, namely from Corollary 2,

$$\operatorname{avar}\big(S(\hat{\mathbf{m}}_i)\big) = \frac{\partial S(\hat{\mathbf{m}}_i)}{\partial \mathbf{m}^T}\hat{\mathbf{D}}\frac{\partial S(\hat{\mathbf{m}}_i)^T}{\partial \mathbf{m}} = \frac{\partial S(\mathbf{y})}{\partial \mathbf{m}^T}\mathbf{D}(\mathbf{y})\frac{\partial S(\mathbf{y})^T}{\partial \mathbf{m}} = \sum\sum y_{ij}^{-1} = 0.1965.$$

These results hold regardless of the choice of sampling constraint matrices $\mathbf{Z}_{iF}, i = 1, 2, 3$.

EXAMPLE 9.2 (Conditional probability estimator). Haberman [(1974), pages 88–90] described the asymptotic distributions of conditional probability ML estimators for conditional–Poisson sampling and log-linear model constraints. Using Theorem 8 and its corollaries, we can generalize these log-linear model results for the independent sampling schemes used in this article to the broader class of MPH models.

Suppose it is desired to estimate the conditional probabilities in $\mathbf{S}(\mathbf{P}) = \mathbf{D}^{-1}(\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{P})\mathbf{P}$. Consider two candidate data models, $\mathbf{y} \leftarrow \mathbf{Y} \sim \mathrm{MPH}_{Z_1}(\mathbf{h}|\mathbf{Z}_{1F}, \mathbf{n}_1)$ and $\mathbf{y} \leftarrow \mathbf{Y} \sim \mathrm{MPH}_{Z_2}(\mathbf{h}|\mathbf{Z}_{2F}, \mathbf{n}_2)$, where $\mathbf{Z}_2 = \mathbf{Z}_1\mathbf{Z}$ for some population matrix $\mathbf{Z}$. Because $\mathbf{S}$ is in $\mathcal{H}_0(\mathbf{Z}_1)$, Proposition 2 implies that it is also in $\mathcal{H}_0(\mathbf{Z}_2)$. It follows that $\mathbf{S}(\mathbf{P}) = \mathbf{S}(\mathbf{m})$ under either data model. Therefore the ML estimators of $\mathbf{S}(\mathbf{P})$ that correspond to the two data models are $\mathbf{S}(\hat{\mathbf{m}}_1)$ and $\mathbf{S}(\hat{\mathbf{m}}_2)$. Because both models are equivalent [i.e., both belong to $\mathcal{E}(\mathbf{h}, \mathbf{y})$], the ML estimates are numerically identical. Also, for $i = 1, 2$, the estimator $\mathbf{S}(\hat{\mathbf{m}}_i)$ is a $\mathbf{Z}_i$ homogeneous statistic of order 0, so by Corollary 3 the approximating variance estimates are identical as well. These results hold regardless of the choice of sampling constraint matrices $\mathbf{Z}_{iF} \in \mathcal{S}(\mathbf{Z}_i)$.

Note that $\mathbf{S}(\hat{\mathbf{m}}_1) = \mathbf{D}^{-1}(\mathbf{Z}_1\mathbf{Z}_1^T\hat{\mathbf{m}}_1)\hat{\mathbf{m}}_1 = \mathbf{N}_1^{-1}\hat{\mathbf{m}}_1 = \hat{\boldsymbol{\pi}}_1$ and that $\mathbf{S}(\hat{\mathbf{m}}_2) = \mathbf{D}^{-1}(\mathbf{Z}_1\mathbf{Z}_1^T\hat{\mathbf{m}}_2)\hat{\mathbf{m}}_2 = \mathbf{N}_1^{-1}\hat{\mathbf{m}}_2$ (because $\mathbf{Z}_1^T\hat{\mathbf{m}}_2 = \mathbf{Z}_1^T\hat{\mathbf{m}}_1 = \mathbf{Z}_1^T\mathbf{Y}$). Theorem 6 gives the common approximating variance estimate as $\operatorname{avar}(\mathbf{S}(\hat{\mathbf{m}}_1)) = \operatorname{avar}(\mathbf{S}(\hat{\mathbf{m}}_2)) = \mathbf{N}_1^{-1}[\hat{\mathbf{D}} - \hat{\mathbf{D}}\hat{\mathbf{H}}(\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{D}} - \mathbf{N}_1^{-1}\hat{\mathbf{D}}\mathbf{Z}_1\mathbf{Z}_1^T\hat{\mathbf{D}}]\mathbf{N}_1^{-1}$.

10. On estimability and testability. A primary goal of contingency table inference is to estimate functions of, or test hypotheses about, the predata probabilities $\mathbf{P}$ and perhaps certain rate parameters. Of course, we are restricted by the fact that our model for data $\mathbf{y}$, which depends on the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, only affords inferences about the data model parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\pi} = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{P})\mathbf{P}$ or, equivalently, $\mathbf{m} = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}$. This section formally defines estimability and testability, and gives sufficient conditions for when an estimand $\mathbf{S}(\mathbf{P})$ is estimable and/or a hypothesis $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ is testable.

10.1. *Estimability of functions of* $\mathbf{P}$.

DEFINITION 9. A function $\mathbf{S}(\mathbf{P})$ is said to be $\mathbf{Z}$-*estimable* if there exists a function $\mathbf{S}^*$ satisfying $\mathbf{S}(\mathbf{P}) = \mathbf{S}^*(\mathbf{m})$.

Note that every estimand $\mathbf{S}(\mathbf{P})$ is $\mathbf{1}$-estimable, because in this case $\mathbf{P} = \mathbf{m}/\mathbf{1}^T\mathbf{m}$. More generally, a simple to verify sufficient condition for $\mathbf{Z}$-estimability of $\mathbf{S}(\mathbf{P})$ is that $\mathbf{S} \in \mathcal{H}_0(\mathbf{Z})$; that is, $\mathbf{S}$ is zero-order $\mathbf{Z}$-homogeneous. This follows because $\mathbf{S}(\mathbf{m}) = \mathbf{S}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}) = \mathbf{S}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{P})\mathbf{P}) = \mathbf{S}(\mathbf{P})$, where the last equality follows when $\mathbf{S}$ is in $\mathcal{H}_0(\mathbf{Z})$.

That this zero-order homogeneity is not a necessary condition follows, for example, from $S(\mathbf{P}) = P_1 - P_2$ being $\mathbf{1}_2$-estimable. Similarly, there are simple examples that show that the more general condition $\mathbf{S} \in \mathcal{H}(\mathbf{Z})$ is not sufficient for $\mathbf{Z}$-estimability of $\mathbf{S}(\mathbf{P})$.

REMARK. Consider data model $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$. If $\mathbf{S} \in \mathcal{H}_0(\mathbf{Z})$ and first-order derivatives exist, then $\mathbf{S}(\mathbf{P})$ is $\mathbf{Z}$-estimable *and* $\widehat{\mathbf{S}(\mathbf{P})} = \mathbf{S}(\hat{\mathbf{m}})$ is a zero-order $\mathbf{Z}$-homogeneous statistic.

EXAMPLE 10.1. Consider the estimands of Section 4.2. The 1987–1989 data and the 1999 data resulted from sampling plans with population matrices $\mathbf{1}_9$ and $\mathbf{Z} = \bigoplus_1^3 \mathbf{1}_3$, respectively. It follows that all of the estimands can be estimated using the 1987–1989 data. The sufficient condition for $\mathbf{Z}$-estimability can be used to show that the Gini concentrations $S_i(\mathbf{P}) = G_i = \sum_{j=1}^3 (P_{ij}/P_{i+})^2$, $i = 1, 2, 3$, as well as $S_4(\mathbf{P}) = P_{11}/(P_{12} + P_{13})$ and $S_5(\mathbf{P}) = P_{23}/P_{2+}$, can be estimated using the 1999 data. In contrast, the estimand $S_6(\mathbf{P}) = P_{21}/P_{12}$ does not satisfy the sufficient condition for $\mathbf{Z}$-estimability. In fact, it is straightforward to see that $S_6(\mathbf{P})$ is *not* $\mathbf{Z}$-estimable. Note that, for $i = 1, \ldots, 5$, $\widehat{S_i(\mathbf{P})} = S_i(\hat{\mathbf{m}})$ is a zero-order $\mathbf{Z}$-homogeneous statistic under the 1999 data model.

EXAMPLE 10.2. Consider the exchange score model of Section 4.2. The exchange scores have the form $\alpha_i = \log(P_{i1}/P_{1i}) \equiv \alpha_i(\mathbf{P})$. It follows that, for $i = 2, 3$, the function $\alpha_i(\cdot)$ is not in $\mathcal{H}_0(\mathbf{Z})$; in fact, it is straightforward to see that $\alpha_i(\mathbf{P})$ is not $\mathbf{Z}$-estimable. That is, the exchange scores are not estimable using the 1999 data.

EXAMPLE 10.3. Consider the linear logit model $\log(P_{i2}/P_{i1}) = \alpha + \beta * i$, $i = 1, \ldots, 4$, for the probabilities in a 4×2 table. Suppose that the observed MP counts are the result of a sampling plan with $\mathbf{Z} = \mathbf{1}_4 \otimes \mathbf{I}_2$. If the dichotomous column variable is viewed as the response of interest, we would call this a retrospective or case-control sampling plan. It is straightforward to see that the data *cannot* be used to estimate $\log(P_{i2}/P_{i1})$ or α, but they can be used to estimate the slope coefficient β. For example, the $\mathbf{Z}$-estimability of β follows because $\beta = \log(P_{i1}P_{i+1,2}/(P_{i+1,1}P_{i,2}))$ for any i, and these log odds ratios are $\mathbf{Z}$-estimable.

10.2. *Testability of* $\mathbf{h}(\mathbf{P}) = \mathbf{0}$.

DEFINITION 10. A hypothesis $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ is said to be $\mathbf{Z}$-*testable* if there exists a function $\mathbf{h}^*$ satisfying $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ if and only if $\mathbf{h}^*(\mathbf{m}) = \mathbf{0}$.

Note that every hypothesis is $\mathbf{1}$-testable. More generally, a simple to verify sufficient condition for $\mathbf{Z}$-testability of $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ is that $\mathbf{h} \in \mathcal{H}(\mathbf{Z})$; that is, $\mathbf{h}$ is $\mathbf{Z}$-homogeneous of any order. This follows because $\mathbf{h}(\mathbf{m}) = \mathbf{0}$ iff $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{P})\mathbf{P}) = \mathbf{0}$ iff $\mathbf{h}(\mathbf{P}) = \mathbf{0}$, where the last equivalence follows when $\mathbf{h}$ is in $\mathcal{H}(\mathbf{Z})$.

That this homogeneity condition is not necessary follows from a simple example: Let $\mathbf{P} = (P_{11}, P_{12}, P_{21}, P_{22})^T$ and consider the data model with population matrix $\bigoplus_1^2 \mathbf{1}_2$. The function h defined as $h(\mathbf{P}) = \log(P_{11}/P_{+1}) - \log(P_{12}/P_{+2})$ is not in $\mathcal{H}(\bigoplus_1^2 \mathbf{1}_2)$, but $h(\mathbf{P}) = 0$ if and only if $h^*(\mathbf{m}) \equiv \log(m_{11}m_{22}) - \log(m_{12}m_{21}) = 0$. This equivalence follows because, in 2×2 tables, the relative risk equals 1 if and only if the odds ratio equals 1.

REMARK. When $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, the hypothesis $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ is $\mathbf{Z}$-testable and is equivalent to the test of goodness of fit of the MPH data model $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$, which imposes the constraint $\mathbf{h}(\mathbf{m}) = \mathbf{0}$.

EXAMPLE 10.4. Consider the four hypotheses of Section 4.2. Using the sufficient condition for testability, it is straightforward to see that the 1999 data as modeled in Section 4.1 can be used to test $\mathbf{h}_1(\mathbf{P}) = \mathbf{0}$, $\mathbf{h}_2(\mathbf{P}) = \mathbf{0}$ and $h_4(\mathbf{P}) = 0$. In contrast, the sufficient condition does not hold for h_3; in fact, $h_3(\mathbf{P}) = 0$ is not testable using the 1999 data. We point out, however, that it is testable using the 1987–1989 data.

The fact that $h_4(\mathbf{P}) = 0$, the hypothesis that the exchange score model of Section 4.1 holds, is testable using the 1999 data leads to what at first glance appears to be a paradox. Example 10.2 argued that the parameters in the exchange score model are not estimable using the 1999 data. Thus, the 1999 data can be used to test whether the exchange score model holds, but they cannot be used to estimate any of that model's parameters.

EXAMPLE 10.5. Consider the setting of Example 10.3 and recall that $\mathbf{Z} = \mathbf{1}_4 \otimes \mathbf{I}_2$. Denote the local odds ratios by $\theta_i = (P_{i1}P_{i+1,2})/(P_{i2}P_{i+1,1}), i = 1, 2, 3$. The constraint form of the linear logit model, $\log(P_{i2}/P_{i1}) = \alpha + \beta * i$, $i = 1, \ldots, 4$, can be written as $\mathbf{h}(\mathbf{P}) = (\log(\theta_1/\theta_2), \log(\theta_1/\theta_3)) = \mathbf{0}$. It follows that $\mathbf{h}$ is in $\mathcal{H}''(\mathbf{Z})$, so by the sufficiency condition $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ is $\mathbf{Z}$-testable. The test is equivalent to the goodness-of-fit test of the model $\mathrm{MPH}_Z(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$.

10.3. *Estimability of rate parameters.* The estimability of rate quantities is dependent on the sample size assumptions. If a sample size is fixed a priori, then we cannot estimate the sampling rate parameter without additional information. When the sample sizes are random and events occur according to a Poisson process, we can estimate the rate parameters in a natural way. Note that estimation must proceed unconditionally, that is, it is not appropriate to conduct inferences conditional on sample sizes in this setting.

EXAMPLE 10.6. The 1999 citation data can be used to estimate the expected number of ANNS references per issue of JASA—this rate is equal to $E(Y_{13}) = \delta_1\pi_{13} = m_{13}$. In contrast, the 1999 data cannot be used to estimate the expected number of JASA references per issue of the ANNS, because the expected number of ANNS references $E(Y_{3+}) = 225$ is fixed by sampling design.

11. Numerical computation of ML fit results: statistics journal citation data analysis. Maximum likelihood fitting for MPH models is relatively straightforward. I have written a software program in R [discussed in Ihaka and Gentleman (1996)] that produces maximum-likelihood fit results for any MPH model that has a constraint function that can be explicitly specified. The program uses a modified Newton–Raphson algorithm to solve the restricted likelihood equations (7.2) of Theorem 1. The algorithm is related to the algorithms of Aitchison and Silvey (1958) and Lang and Agresti (1994). By-products of the fitting algorithm, in conjunction with the MPH model approximation results presented herein, are used to compute a wide variety of ML fit results including goodness-of-fit statistics, residuals and approximating variance estimates.

Section 10.2 argued that the hypothesis $\mathbf{h}_1(\mathbf{P}) = (G_1 - 0.410, G_2 - 0.455, G_3 - 0.684)^T = \mathbf{0}$ that the Gini concentrations of cited journals have not changed from the 1987–1989 values (see Section 4.2) is testable using the 1999 data model, denoted $\mathrm{MPH}_Z(0|\mathbf{Z}_F, 225)$. In fact, this test is equivalent to the test of goodness of fit of the model $\mathrm{MPH}_Z(\mathbf{h}_1|\mathbf{Z}_F, 225)$. The observed likelihood ratio and Pearson score statistics are $G^2 = 7.773$ and $X^2 = 8.421$, respectively. Comparing these values to the approximate null distribution $\chi^2(3)$ gives p-values of 0.051 and 0.038. To assess the local fit of the no-change model, the nine adjusted residuals are computed: their values are $(0.455, -0.455, -0.455, 0.642, -0.642, 0.642, 2.793, 2.793, -2.793)$. Evidently, the no-change hypothesis is questionable, especially regarding the citation patterns in ANNS. The observed 1999 Gini concentrations are 0.419 (ase = 0.021), 0.441 (ase = 0.022) and 0.595 (ase = 0.032); the

JASA and BMCS concentrations are nearly unchanged from 1987–1989, but there appears to be less concentration in 1999 than in 1987–1989 for ANNS.

Section 10.2 argued that the 1999 data could be used to test the hypothesis $h_4(\mathbf{P}) = 0$ that Stigler's (1994) exchange score model $\log(P_{ij}/P_{ji}) = \alpha_i - \alpha_j$ holds. This model holds if and only if the data model parameters $\mathbf{m}$ satisfy $\log(m_{ij}/m_{ji}) = \beta_i - \beta_j$ or, in generic matrix notation, $\mathbf{L}_1(\mathbf{m}) = \mathbf{X}_1\boldsymbol{\beta}_1$. Alternatively, the exchange score model holds if and only if the data model parameters $\mathbf{m}$ satisfy the quasisymmetry model $\log m_{ij} = \beta + \beta_i^A + \beta_j^B + \beta_{ij}$, where $\beta_{ij} = \beta_{ji}$ [see Fienberg and Larntz (1976) and Stigler (1994)], or in generic matrix notation, $\mathbf{L}_2(\mathbf{m}) = \mathbf{X}_2\boldsymbol{\beta}_2$.

The models $\mathbf{L}_1(\mathbf{m}) = \mathbf{X}_1\boldsymbol{\beta}_1$ and $\mathbf{L}_2(\mathbf{m}) = \mathbf{X}_2\boldsymbol{\beta}_2$ are qualitatively different in that $\mathbf{L}_1$ is a many-to-one link and $\mathbf{L}_2$ is a one-to-one link. Standard approaches to ML fitting require the link to be one-to-one because the likelihood is to be reparameterized in terms of the $\boldsymbol{\beta}$ "freedom" parameters. For this reason, the many-to-one link models are typically fitted using non-ML methods, such as weighted least squares [see Stokes, Davis and Koch (2000), Chapter 13]. This need not be the case. Following the approach of Aitchison and Silvey (1958), either model can be easily fitted using ML because they both can be respecified in terms of constraints: $\mathbf{L}_1(\mathbf{m}) = \mathbf{X}_1\boldsymbol{\beta}_1$ is equivalent to $h_{4,1}(\mathbf{m}) \equiv \mathbf{U}_1^T\mathbf{L}_1(\mathbf{m}) = \mathbf{0}$ and $\mathbf{L}_2(\mathbf{m}) = \mathbf{X}_2\boldsymbol{\beta}_2$ is equivalent to $h_{4,2}(\mathbf{m}) \equiv \mathbf{U}_2^T\mathbf{L}_2(\mathbf{m}) = \mathbf{0}$; here $\mathbf{U}_i$ is the orthogonal complement of $\mathbf{X}_i$, $i = 1, 2$. It can be shown that h_4, $h_{4,1}$ and $h_{4,2}$ lie in $\mathcal{H}''(\mathbf{Z})$, so the exchange score data model $\mathrm{MPH}_Z(h_4|\mathbf{Z}_F, \mathbf{n})$ can be equivalently respecified as $\mathrm{MPH}_Z(h_{4,1}|\mathbf{Z}_F, \mathbf{n})$ or $\mathrm{MPH}_Z(h_{4,2}|\mathbf{Z}_F, \mathbf{n})$. The goodness-of-fit statistic values for the exchange score data model are $G^2 = 0.187$, $X^2 = 0.190$ and $\mathrm{df} = 1$. Furthermore, all the adjusted residuals are less than 0.4361 in absolute value. Thus, it appears that the exchange score model fits these data very well. Recall from Section 10.1 that the exchange score parameters α_i cannot be estimated using these 1999 data.

We estimate the expected number of ANNS references per JASA issue, which in terms of the data model parameters is $E(Y_{13}) = \delta_1\pi_{13} = m_{13}$, using the exchange score model. To this end, it is important that we do not conduct inference conditional on the sample sizes. Instead, we use the restricted data model $\mathrm{MPH}_Z(h_4|\mathbf{Z}_F, \mathbf{n})$. The ML estimate of m_{13} under the exchange score model is $\hat{m}_{13} = 64.12$ and the approximating standard error is 7.75. Interestingly, the corresponding approximate 95% confidence interval [48.90, 79.34] contains the average number of ANNS references per 1987–1989 JASA issue, which was $739/12 = 61.58$, so we do not observe a statistically significant change in the rate from 1987–1989. If we inappropriately condition on the sample sizes and use the equivalent product-multinomial model $\mathrm{MPH}_Z(h_4|\mathbf{Z}, \mathbf{n}_1 = (193, 252, 225))$, the ML estimate of m_{13} is the same, but the approximating standard error 6.22 is smaller than 7.75, the correct standard error. The difference in standard errors can be computed using NE2.

The estimated Gini concentration for JASA under the 1999 exchange score data model $\mathrm{MPH}_Z(h_4|\mathbf{Z}_F, \mathbf{n})$ is 0.417 (ase = 0.020). If two references in JASA are randomly selected, given they refer to JASA, BMCS or ANNS, the probability they refer to the same journal is estimated to be 41.7%. Because this Gini concentration estimator is a $\mathbf{Z}$ homogeneous statistic of order 0, the ML estimate and approximate standard error estimate would be unchanged if, for example, we instead used the population equivalent product-multinomial data model $\mathrm{MPH}_Z(h_4|\mathbf{Z}, \mathbf{n}_1 = (193, 252, 225))$ or the equivalent Poisson model $\mathrm{MPH}_{\mathbf{1}_9}(\mathbf{h}_1|0)$.

Section 10.2 argued that the hypothesis of common Gini concentration $\mathbf{h}_2(\mathbf{P}) = \mathbf{0}$ (see Section 4.2 also) is testable using the 1999 data. The common Gini model is equivalent to $\mathbf{L}(\mathbf{m}) = \alpha\mathbf{1}_3$, where $\mathbf{L}(\mathbf{m}) \equiv (\sum_{j=1}^3 (m_{1j}/m_{1+})^2, \sum_{j=1}^3 (m_{2j}/m_{2+})^2, \sum_{j=1}^3 (m_{3j}/m_{3+})^2)^T = \mathbf{L}(\mathbf{P})$. The link $\mathbf{L}$ is a many-to-one link, but as argued above, this causes no problems with ML fitting because $\mathbf{L}(\mathbf{m}) = \alpha\mathbf{1}_3$ if and only if $\mathbf{h}_{2,1}(\mathbf{m}) \equiv \mathbf{U}^T\mathbf{L}(\mathbf{m}) = \mathbf{0}$, where $\mathbf{U}$ is the orthogonal complement of $\mathbf{1}_3$. Note that $\mathbf{h}_{2,1} \in \mathcal{H}''(\mathbf{Z})$. The fit of the restricted 1999 data model $\mathrm{MPH}_Z(\mathbf{h}_2|\mathbf{Z}_F, \mathbf{n})$ gives $\hat{\alpha} = 0.478$ (ase = 0.015), $G^2 = 24.63$, and df = 2 ($p < 0.0001$). Evidently, the model of common Gini concentrations is untenable—in particular, it appears that there is more concentration in ANNS than the other two journals.

12. Discussion. This article makes a clear distinction between the predata probabilities $\mathbf{P}$ and the data model parameters $\boldsymbol{\pi} = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{P})\mathbf{P}$ and $\boldsymbol{\gamma} = E(\mathbf{Z}^T\mathbf{Y})$. It is generally more informative to begin a contingency table analysis by specifying a model and/or estimands in terms of the predata probabilities $\mathbf{P}$, rather than the data model parameters. It can then be determined whether the model and/or estimands can be restated in terms of the data model parameters that correspond to the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$. As an example, consider a 2×2 table, where $(A, B) \sim \mathbf{P} = (P_{11}, P_{12}, P_{21}, P_{22})$. Suppose the estimand of interest is the relative risk $S(\mathbf{P}) = (P_{11}/P_{1+})/(P_{21}/P_{2+})$. Under sampling plan $(\mathbf{Z}_1, \mathbf{Z}_{1F}, \mathbf{n}_1)$, where $\mathbf{Z}_1 = \bigoplus_1^2 \mathbf{1}_2$, the data model probabilities are defined as $\pi_{ij} = P_{ij}/P_{i+}$ and the relative risk $S(\mathbf{P}) = \pi_{11}/\pi_{21}$ is estimable. In contrast, under sampling plan $(\mathbf{Z}_2, \mathbf{Z}_{2F}, \mathbf{n}_2)$, where $\mathbf{Z}_2 = \mathbf{1}_2 \otimes \mathbf{I}_2$, the data model probabilities are $\pi_{ij} = P_{ij}/P_{+j}$ and $S(\mathbf{P})$ is not estimable.

I have written a MPH model ML fitting program in R [Ihaka and Gentleman (1996); see also http://cran.r-project.org]. This program (available from the author upon request) can be used to compute ML fit statistics for many less standard contingency table models. An important example is the class of linear predictor models of the form $\mathbf{L}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{L}$ is generally many-to-one. Currently available statistical packages typically fit these many-to-one link models using nonlikelihood methods, such as weighted least squares [see Stokes, Davis and Koch (2000), Chapter 13]. The current article and the companion fitting program illustrate that models like these can be fitted easily using maximum likelihood.

Aside from the examples, this article focused on results for the very general class of MPH models. Lang (2000) exploited the special structure of and gave more in-depth results for several important subclasses of MPH models. Included among these subclasses are *generalized log-linear models* [see Lang and Agresti (1994) and Lang, McDonald and Smith (1999)] and the broad class of *probability freedom models* that lend themselves to the multinomial-to-Poisson transformation of Baker (1994).

APPENDIX: SELECTED PROOFS

PROOF OF PROPOSITION 5 (Generalized Euler's homogeneous function theorem). A version of Euler's theorem can be stated as follows: *Suppose that first-order partial derivatives of the scalar-valued function* $h : \Omega \mapsto R$ *exist. Then*

$$\delta^p h(\mathbf{x}) = h(\delta\mathbf{x}) \qquad \forall\, \delta > 0, \forall\, \mathbf{x} \in \Omega$$

if and only if

$$ph(\mathbf{x}) = \mathbf{x}^T \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} \qquad \forall\, \mathbf{x} \in \Omega.$$

The proof is straightforward [see Fleming (1977), page 89]. This result will be used repeatedly in the proof of Proposition 5.

Let $\phi_k = (i : Z_{ik} = 1)$ and for $\mathbf{x} = (x_1, \ldots, x_c)^T$ define $\mathbf{y}_k = \mathbf{x}_{\phi_k}$, that is, $\mathbf{y}_k$ comprises those components in $\mathbf{x}$ that correspond to the kth column in $\mathbf{Z}$. Define the argument-permuting function $\mathbf{E}_Z$ as $\mathbf{E}_Z(\mathbf{y}_1, \ldots, \mathbf{y}_K) = \mathbf{x}$. For example, if $\mathbf{Z}$ is a 4×2 population matrix that gives $\phi_1 = (1, 3)$ and $\phi_2 = (2, 4)$, then $\mathbf{y}_1 = (x_1, x_3)$ and $\mathbf{y}_2 = (x_2, x_4)$. The argument-permuting function would be defined as $\mathbf{E}_Z(a_1, a_2, a_3, a_4) = (a_1, a_3, a_2, a_4)$, so that $\mathbf{E}_Z(\mathbf{y}_1, \mathbf{y}_2) = (x_1, x_2, x_3, x_4) = \mathbf{x}$.

Necessity. Assume that $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}) = \operatorname{diag}\{\delta_{\nu(j)}^{p(j)},\ j = 1, \ldots, u\}\mathbf{h}(\mathbf{x})\ \forall\, \boldsymbol{\delta} > 0$, $\forall\, \mathbf{x} \in \Omega$. Let $\mathbf{x} \in \Omega$. The matrix $\mathbf{Z}^T\mathbf{D}(\mathbf{x})\mathbf{H}(\mathbf{x})$ can be written as $\mathbf{Z}^T\mathbf{D}(\mathbf{x})[(\partial h_1(\mathbf{x}))/\partial\mathbf{x}, \ldots, (\partial h_u(\mathbf{x}))/\partial\mathbf{x}]$. The jth column can be written as

$$\mathbf{Z}^T\mathbf{D}(\mathbf{x})\frac{\partial h_j(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \mathbf{y}_1^T \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{y}_1} \\ \vdots \\ \mathbf{y}_K^T \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{y}_K} \end{bmatrix}.$$

Define $h_j^{(k)}(\mathbf{y}_k) = h_j(\mathbf{x})$, where $\mathbf{y}_j$, $j \neq k$, are viewed as fixed.

By assumption (i.e., $\mathbf{h}$ is $\mathbf{Z}$-homogeneous), for every $\delta > 0$,

$$h_j^{(k)}(\delta\mathbf{y}_k) = \begin{cases} \delta^{p(j)} h_j^{(k)}(\mathbf{y}_k), & \text{if } k = \nu(j), \\ h_j^{(k)}(\mathbf{y}_k), & \text{if } k \neq \nu(j), k = 1, \ldots, K. \end{cases}$$

By Euler's theorem, as stated at the beginning of this proof,

$$\mathbf{y}_k^T \frac{\partial h_j^{(k)}(\mathbf{y}_k)}{\partial \mathbf{y}_k} = \begin{cases} p(j)h_j^{(k)}(\mathbf{y}_k), & \text{if } k = \nu(j), \\ 0, & \text{if } k \neq \nu(j), k = 1, \ldots, K. \end{cases}$$

Equivalently,

$$\mathbf{y}_k^T \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{y}_k} = \begin{cases} p(j)h_j(\mathbf{x}), & \text{if } k = \nu(j), \\ 0, & \text{if } k \neq \nu(j), k = 1, \ldots, K. \end{cases}$$

Therefore,

$$\mathbf{Z}^T \mathbf{D}(\mathbf{x}) \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} I(\nu(j) = 1) \\ \vdots \\ I(\nu(j) = K) \end{bmatrix} p(j)h_j(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega$$

or

$$\mathbf{Z}^T \mathbf{D}(\mathbf{x})\mathbf{H}(\mathbf{x}) = \mathbf{A}\mathbf{D}(\mathbf{p})\mathbf{D}(\mathbf{h}(\mathbf{x})) \qquad \forall \mathbf{x} \in \Omega,$$

where $\mathbf{A}$ is $K \times u$ with (i, j)th component $A_{ij} = I(\nu(j) = i)$. Note that $A_{ij} \in \{0, 1\}$ and $\sum_{i=1}^K A_{ij} = 1$. This proves the necessity of the condition.

Sufficiency. Assume that $\mathbf{Z}^T \mathbf{D}(\mathbf{x})\mathbf{H}(\mathbf{x}) = \mathbf{A}\mathbf{D}(\mathbf{p})\mathbf{D}(\mathbf{h}(\mathbf{x}))\ \forall \mathbf{x} \in \Omega$, where $A_{ij} \in \{0, 1\}$ and $\sum_{i=1}^K A_{ij} = 1$. By assumption, for any $\mathbf{x} \in \Omega$,

$$\mathbf{Z}^T \mathbf{D}(\mathbf{x}) \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{x}} = p(j)h_j(\mathbf{x}) \begin{bmatrix} A_{1j} \\ \vdots \\ A_{Kj} \end{bmatrix}$$

or, equivalently,

$$\mathbf{y}_k^T \frac{\partial h_j(\mathbf{x})}{\partial \mathbf{y}_k} = \begin{cases} 0, & \text{if } A_{kj} = 0, \\ p(j)h_j(\mathbf{x}), & \text{if } A_{kj} = 1, k = 1, \ldots, K. \end{cases}$$

This implies, using Euler's theorem as stated above, that for $k = 1, \ldots, K$,

$$\begin{aligned} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \ldots, \delta_k \mathbf{y}_k, \ldots, \mathbf{y}_K)\big) &= \begin{cases} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \ldots, \mathbf{y}_K)\big), & \text{if } A_{kj} = 0, \\ \delta_k^{p(j)} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \ldots, \mathbf{y}_K)\big), & \text{if } A_{kj} = 1, \end{cases} \\ &= \delta_k^{p(j)A_{kj}} h_j(\mathbf{x}). \end{aligned}$$

Letting $\mathbf{y}_{0k} \equiv \mathbf{x}_{0\phi_k}$ be the components in $\mathbf{x}_0$ that correspond to the kth column in $\mathbf{Z}$, we can summarize by defining property k, for $k = 1, \ldots, K$:

PROPERTY k. $h_j(\mathbf{E}_Z(\mathbf{y}_{01}, \ldots, \delta_k \mathbf{y}_{0k}, \ldots, \mathbf{y}_{0K})) = \delta_k^{p(j)A_{kj}} h_j(\mathbf{x}_0)\ \forall \delta_k > 0$ $\forall \mathbf{x}_0 \in \Omega$.

The result now follows by sequential applications of properties 1 through K. Let $\boldsymbol{\delta} > 0, \mathbf{x} \in \Omega$. Then

$$
\begin{aligned}
h_j\big(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}\big) &= h_j\big(\mathbf{E}_Z(\delta_1\mathbf{y}_1, \delta_2\mathbf{y}_2, \delta_3\mathbf{y}_3, \ldots, \delta_K\mathbf{y}_K)\big) \\
&= \delta_1^{p(j)A_{1j}} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \delta_2\mathbf{y}_2, \delta_3\mathbf{y}_3, \ldots, \delta_K\mathbf{y}_K)\big) \\
&\qquad\qquad \text{[apply property 1 with} \\
&\qquad\qquad \mathbf{x}_0 = \mathbf{E}_Z(\mathbf{y}_1, \delta_2\mathbf{y}_2, \delta_3\mathbf{y}_3, \ldots, \delta_K\mathbf{y}_K) \in \Omega\text{]} \\
&= \delta_1^{p(j)A_{1j}} \delta_2^{p(j)A_{2j}} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \mathbf{y}_2, \delta_3\mathbf{y}_3, \ldots, \delta_K\mathbf{y}_K)\big) \\
&\qquad\qquad \text{[apply property 2 with} \\
&\qquad\qquad \mathbf{x}_0 = \mathbf{E}_Z(\mathbf{y}_1, \mathbf{y}_2, \delta_3\mathbf{y}_3, \ldots, \delta_K\mathbf{y}_K) \in \Omega\text{]} \\
&\quad\ \vdots \\
&= \big(\delta_1^{A_{1j}} \cdots \delta_K^{A_{Kj}}\big)^{p(j)} h_j\big(\mathbf{E}_Z(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_K)\big) \\
&\qquad\qquad \text{[apply property } K \text{ with} \\
&\qquad\qquad \mathbf{x}_0 = \mathbf{E}_Z(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_K) = \mathbf{x} \in \Omega\text{]} \\
&= \delta_{\nu(j)}^{p(j)} h_j(\mathbf{x}),
\end{aligned}
$$

where $\nu(j) \in \{1, \ldots, K\}$. The last equality follows because $A_{ij} \in \{0, 1\}$ and $\sum_{i=1}^{K} A_{ij} = 1$. Therefore, in matrix notation $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\delta})\mathbf{x}) = \mathbf{G}(\boldsymbol{\delta})\mathbf{h}(\mathbf{x})\ \forall\, \boldsymbol{\delta} > 0$, $\forall\, \mathbf{x} \in \Omega$, where $\mathbf{G}(\boldsymbol{\delta}) = \operatorname{diag}\{\delta_{\nu(j)}^{p(j)},\ j = 1, \ldots, u\}$. This proves the sufficiency of the condition. □

PROOF OF THEOREM 1. By convention, if $\mathbf{Z}_F = \mathbf{0}$ (so $\mathbf{Z}_R = \mathbf{Z}$), the equations (7.1) reduce to (7.2). Consider $\mathbf{Z}_F \neq \mathbf{0}$. Using sampling constraint matrix property S8 and Proposition 6, the first set of equations in (7.1), when premultiplied by $\mathbf{Z}_F^T$, reduces to $\mathbf{Z}_F^T\mathbf{y} + \mathbf{Z}_F^T\mathbf{D}(\mathbf{m})\mathbf{Z}_F\boldsymbol{\tau} = \mathbf{0}$. Alternatively using S4 and the fact that $\mathbf{Z}_F^T\mathbf{m} = \mathbf{n} = \mathbf{Z}_F^T\mathbf{y}$, the equations reduce to $\mathbf{n} + \mathbf{D}(\mathbf{n})\boldsymbol{\tau} = \mathbf{0}$. Thus, $\boldsymbol{\tau} = -\mathbf{1}$ and the solution $\mathbf{m}$ must satisfy the reduced set of equations

$$
\begin{bmatrix} \mathbf{y} - \mathbf{D}(\mathbf{m})\big[\mathbf{Z}_R\mathbf{1} + \mathbf{Z}_F\mathbf{1}\big] + \mathbf{D}(\mathbf{m})\mathbf{H}(\mathbf{m})\boldsymbol{\lambda} \\ \mathbf{h}(\mathbf{m}) \end{bmatrix} = \mathbf{0}.
$$

However, by sampling constraint matrix property S9 these equations are identical to those in (7.2). Therefore, the restricted likelihood equations can be reduced to (7.2). □

PROOF OF THEOREM 2. That $\hat{\boldsymbol{\pi}} = \mathbf{D}^{-1}(\mathbf{Z}\hat{\boldsymbol{\gamma}})\hat{\mathbf{m}}$ follows from the invariance of ML estimates. By the form of the log likelihood [see density (3.2)] and the product-space form (6.1) of $\omega_Z^*(\mathbf{h}|\mathbf{Z}_F, \mathbf{n})$, it follows that $\hat{\boldsymbol{\pi}}$ is the maximizer over $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$

of $\mathbf{y}^T \log \boldsymbol{\pi}$ and $\hat{\boldsymbol{\gamma}}$ is the maximizer over $\mathcal{D}(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ of $\mathbf{y}^T \mathbf{Z}_R \log \mathbf{Q}_R^T \boldsymbol{\gamma} - \mathbf{1}^T \mathbf{Q}_R^T \boldsymbol{\gamma}$. If $\mathbf{Z}_R = \mathbf{0}$, then $\mathcal{D}(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}) = \{\mathbf{n}\}$, so that $\hat{\boldsymbol{\gamma}} = \mathbf{n} = \mathbf{Z}^T \mathbf{y}$. Otherwise, to find $\hat{\boldsymbol{\gamma}}$ we note that $\hat{\boldsymbol{\gamma}} = \mathbf{Q}_F \mathbf{n} + \mathbf{Q}_R \hat{\boldsymbol{\delta}}$, where $\hat{\boldsymbol{\delta}}$ is the unrestricted ($\boldsymbol{\delta} > 0$) maximizer of $\mathbf{y}^T \mathbf{Z}_R \log[\mathbf{Q}_R^T(\mathbf{Q}_F \mathbf{n} + \mathbf{Q}_R \boldsymbol{\delta})] - \mathbf{1}^T \mathbf{Q}_R^T(\mathbf{Q}_F \mathbf{n} + \mathbf{Q}_R \boldsymbol{\delta}) = \mathbf{y}^T \mathbf{Z}_R \log \boldsymbol{\delta} - \mathbf{1}^T \boldsymbol{\delta}$. It follows that $\hat{\boldsymbol{\delta}} = \mathbf{Z}_R^T \mathbf{y}$, so that $\hat{\boldsymbol{\gamma}} = \mathbf{Q}_F \mathbf{n} + \mathbf{Q}_R \hat{\boldsymbol{\delta}} = \mathbf{Q}_F \mathbf{Z}_F^T \mathbf{y} + \mathbf{Q}_R \mathbf{Z}_R^T \mathbf{y} = [\mathbf{Q}_F \mathbf{Q}_F^T + \mathbf{Q}_R \mathbf{Q}_R^T] \mathbf{Z}^T \mathbf{y}$. Therefore, using the fact that $\mathbf{Q}_F \mathbf{Q}_F^T + \mathbf{Q}_R \mathbf{Q}_R^T = \mathbf{I}$, we have that $\hat{\boldsymbol{\gamma}} = \mathbf{Z}^T \mathbf{y}$. □

PROOF OF LEMMA 1. For appropriately chosen permutation matrix $\mathbf{E}_Z$, the vector $\mathbf{Y} - \mathbf{m}$ can be written as

$$\mathbf{E}_Z \begin{bmatrix} \mathbf{Y}_{\phi_1} - \mathbf{m}_{\phi_1} \\ \mathbf{Y}_{\phi_2} - \mathbf{m}_{\phi_2} \\ \vdots \\ \mathbf{Y}_{\phi_K} - \mathbf{m}_{\phi_K} \end{bmatrix},$$

where $\phi_k = (i : Z_{ik} = 1)$. Because $\mathbf{Y}$ is a MP random vector, $\mathbf{Y}_{\phi_k}$'s are independent and either $\mathbf{Y}_{\phi_k}$ is multinomial or it comprises independent Poisson random variables. By standard results, the limiting distribution as $\nu \to \infty$ of $\mathbf{D}^{-1/2}(\mathbf{m}_{\phi_k})(\mathbf{Y}_{\phi_k} - \mathbf{m}_{\phi_k})$ is normal with mean vector zero and variance equal to $\mathbf{I} - \boldsymbol{\pi}_{\phi_k}^{1/2} \boldsymbol{\pi}_{\phi_k}^{T/2}$ or $\mathbf{I}$ depending on whether $\mathbf{Y}_{\phi_k}$ is multinomial or has independent Poisson components. Using properties of permutation matrices and the properties of ϕ, $\mathbf{Z}$ and $\mathbf{Z}_F$, the limiting distribution of unpermuted $\mathbf{D}^{-1/2}(\mathbf{m})(\mathbf{Y} - \mathbf{m})$ can be shown to be normal with mean vector zero and variance equal to $\mathbf{I} - \mathbf{D}\mathbf{Z}_F \mathbf{Z}_F^T \mathbf{D}$, where $\mathbf{D} \equiv \mathbf{D}(\boldsymbol{\pi})$. Finally, because $\mathbf{m}/\nu = \mathbf{D}(\mathbf{Z}\mathbf{w})\boldsymbol{\pi} + o(1) = \mathbf{W}\boldsymbol{\pi} + o(1)$, it follows that $\nu^{-1/2}(\mathbf{Y} - \mathbf{m}) = \nu^{-1/2}\mathbf{D}^{1/2}(\mathbf{m})\mathbf{D}^{-1/2}(\mathbf{m})(\mathbf{Y} - \mathbf{m})$ has a normal limiting distribution with mean zero and variance $\mathbf{W}\mathbf{D} - \mathbf{W}\mathbf{D}\mathbf{Z}_F \mathbf{Z}_F^T \mathbf{D}$. □

PROOFS OF LEMMAS 2–4. Lemma 2 is proven as follows. Recall that the mean vector can be written as $\mathbf{m} = \mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\pi}$. It follows by properties of population matrices that $(\mathbf{I} - \mathbf{D}\mathbf{Z}\mathbf{Z}^T)\mathbf{m} = \mathbf{0}$. Therefore, it can be shown that $\mathbf{Y} - \mathbf{N}\boldsymbol{\pi} = (\mathbf{I} - \mathbf{D}\mathbf{Z}\mathbf{Z}^T)(\mathbf{Y} - \mathbf{m})$. The result of Lemma 2 now follows from Lemma 1 and properties of population and sampling constraint matrices. Lemma 3 follows directly from Lemma 1 because $\boldsymbol{\gamma} = \mathbf{Z}^T \mathbf{m}$; the simplified form for the limiting variance follows by properties of population and sampling constraint matrices. The result of Lemma 4 follows from Lemmas 2 and 3. Specifically, by Lemma 3, $\mathbf{N}/\nu = \mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{Y}/\nu) = \mathbf{D}(\mathbf{Z}\mathbf{w}) + o_P(1) = \mathbf{W} + o_P(1)$. Therefore, using Lemma 2, $\nu^{1/2}(\mathbf{N}^{-1}\mathbf{Y} - \boldsymbol{\pi}) = \nu\mathbf{N}^{-1}\nu^{-1/2}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}) = \mathbf{W}^{-1}\nu^{-1/2}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}) + o_P(1)$. □

PROOF OF LEMMA 5. The proof is similar in spirit to that of Silvey (1959). Let $\boldsymbol{\pi}_0 \in \omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ be the true value. Define $\ell^*(\boldsymbol{\pi}) \equiv \mathbf{Y}^T \log \boldsymbol{\pi} - \mathbf{Y}^T \mathbf{Z}\mathbf{Z}^T \boldsymbol{\pi}$, $\mathbf{U}(\boldsymbol{\pi}) \equiv \nu^{-1}[\ell^*(\boldsymbol{\pi}) - \ell^*(\boldsymbol{\pi}_0)]$ and $\boldsymbol{\mu}(\boldsymbol{\pi}) = \lim_{\nu \to \infty} E(\mathbf{U}(\boldsymbol{\pi}))$. It can be shown

that as $\nu \to \infty$, $\sup_{\boldsymbol{\pi} \in \mathcal{C}} |\mathbf{U}(\boldsymbol{\pi}) - \mu(\boldsymbol{\pi})| \xrightarrow{\text{a.s.}} 0$ for any compact set $\mathcal{C} \subset \Omega$ and that $\mu(\boldsymbol{\pi}) < \mu(\boldsymbol{\pi}_0) = 0$ for all $\boldsymbol{\pi} \in \Omega$, $\boldsymbol{\pi} \neq \boldsymbol{\pi}_0$. Let $B \subset \Omega$ be a compact neighborhood of $\boldsymbol{\pi}_0$. By similar arguments to those of Wald (1949) [see, e.g., Ferguson (1996), pages 107–115], it follows that there exists $\hat{\boldsymbol{\pi}}$, a sequence of local maximizers in $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) \cap B$ of $\mathbf{U}$ that converges almost surely to $\boldsymbol{\pi}_0$. Now $\nu\mathbf{U}(\boldsymbol{\pi}) = \mathbf{Y}^T \log \boldsymbol{\pi} + \text{const}$ on the set $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) \cap B$. Thus, the strongly consistent sequence $\hat{\boldsymbol{\pi}}$ comprises local [on $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) \cap B$] maximizers of $\mathbf{Y}^T \log \boldsymbol{\pi}$.

We now show that for sufficiently large ν, with probability going to 1, $\hat{\boldsymbol{\pi}}$ emerges as a solution to the restricted likelihood equations (7.2) through $\hat{\mathbf{m}} = \mathbf{N}\hat{\boldsymbol{\pi}}$. Because $\mathbf{h}$ lies in $\mathcal{H}''(\mathbf{Z})$, by Proposition 7 $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1})$ is a $(c - u - K)$-dimensional manifold. Assume for the moment that $u + K < c$ so the manifold has dimension of at least 1. By strong consistency of $\hat{\boldsymbol{\pi}}$, for large enough ν, with probability going to 1, $\hat{\boldsymbol{\pi}}$ will be in $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) \cap \text{int}(B)$, where $\text{int}(B)$ is the open version of the neighborhood of $\boldsymbol{\pi}_0$. In this case, by Lagrange's result [see Fleming (1977), page 161] the maximizer will satisfy the equations

$$\begin{bmatrix} \mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{Y} + \mathbf{H}(\hat{\boldsymbol{\pi}})\hat{\boldsymbol{\lambda}}_\pi + \mathbf{Z}\hat{\boldsymbol{\tau}} \\ \mathbf{h}(\hat{\boldsymbol{\pi}}) \\ \mathbf{Z}^T\hat{\boldsymbol{\pi}} - \mathbf{1} \end{bmatrix} = \mathbf{0} \tag{A.1}$$

for some multipliers $\hat{\boldsymbol{\lambda}}_\pi$ and $\hat{\boldsymbol{\tau}}$. If $u + K = c$, the set $\omega(\mathbf{h}|\mathbf{Z}, \mathbf{1}) = \{\boldsymbol{\pi}_0\}$, and these equations produce the single desired solution $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}_0$.

Premultiply the first set of equations in (A.1) by $\mathbf{Z}^T\mathbf{D}(\hat{\boldsymbol{\pi}})$ and set the result equal to 0. Using Proposition 6, this leads to the solution $\hat{\boldsymbol{\tau}} = -\mathbf{Z}^T\mathbf{Y}$. Premultiplying the first set of equations by $\mathbf{D}(\hat{\boldsymbol{\pi}})$, replacing $\hat{\boldsymbol{\tau}}$ by $-\mathbf{Z}^T\mathbf{Y}$ and omitting the now-redundant constraint $\mathbf{Z}^T\hat{\boldsymbol{\pi}} = \mathbf{1}$, the equations in (A.1) can be reduced to

$$\begin{bmatrix} \mathbf{Y} - \mathbf{N}\hat{\boldsymbol{\pi}} + \mathbf{D}(\hat{\boldsymbol{\pi}})\mathbf{H}(\hat{\boldsymbol{\pi}})\hat{\boldsymbol{\lambda}}_\pi \\ \mathbf{h}(\hat{\boldsymbol{\pi}}) \end{bmatrix} = \mathbf{0},$$

where $\mathbf{N} = \mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{Y})$. Set $\hat{\mathbf{m}} = \mathbf{N}\hat{\boldsymbol{\pi}}$ and note that because $\mathbf{h} \in \mathcal{H}''(\mathbf{Z})$, $\mathbf{h}(\hat{\boldsymbol{\pi}}) = \mathbf{0}$ if and only if $\mathbf{h}(\hat{\mathbf{m}}) = \mathbf{0}$. Using Proposition 4, the previous reduced set of equations can be written as

$$\begin{bmatrix} \mathbf{Y} - \hat{\mathbf{m}} + \mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}} \\ \mathbf{h}(\hat{\mathbf{m}}) \end{bmatrix} = \mathbf{0},$$

where $\hat{\boldsymbol{\lambda}} = \mathbf{G}^{-1}(\mathbf{Z}^T\mathbf{Y})\hat{\boldsymbol{\lambda}}_\pi$. These are the restricted likelihood equations (7.2) of Theorem 1. □

PROOF OF THEOREM 3. The proof of Lemma 5 implies that $\hat{\boldsymbol{\pi}}$ is the solution to

$$\begin{bmatrix} \mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\mathbf{Y} - \mathbf{N}\hat{\boldsymbol{\pi}}) + \mathbf{H}(\hat{\boldsymbol{\pi}})\hat{\boldsymbol{\lambda}}_\pi \\ \mathbf{h}(\hat{\boldsymbol{\pi}}) \end{bmatrix} = \mathbf{0}, \tag{A.2}$$

where $\hat{\boldsymbol{\lambda}}_\pi = \mathbf{G}(\mathbf{Z}^T\mathbf{Y})\hat{\boldsymbol{\lambda}}$. Owing to the strong consistency of $\hat{\boldsymbol{\pi}}$ and the smoothness of $\mathbf{h}$ and $\mathbf{H}$, these equations can be linearly approximated near the true $\boldsymbol{\pi}$. In particular, note that $\mathbf{H}(\hat{\boldsymbol{\pi}}) = \mathbf{H}(\boldsymbol{\pi}) + o_P(1)$, $\mathbf{h}(\hat{\boldsymbol{\pi}}) = [\mathbf{H}(\boldsymbol{\pi})^T + o_P(1)](\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ and, by Lemma 3 $\nu^{-1}\mathbf{N} = \mathbf{W} + o_P(1)$. By Lemma 2 $\nu^{-1/2}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi})$ is bounded in probability, so the equations (A.2) can be written as

$$\text{(A.3)}\qquad \begin{bmatrix} \nu^{-1/2}\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{D}^{-1}\mathbf{W} & -\mathbf{H} \\ -\mathbf{H}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ \nu^{-1/2}\hat{\boldsymbol{\lambda}}_\pi \end{bmatrix} + o_P(1),$$

where $\mathbf{D} \equiv \mathbf{D}(\boldsymbol{\pi})$ and $\mathbf{H} \equiv \mathbf{H}(\boldsymbol{\pi})$. Write these equations in the generic form $\mathbf{A}_\nu = \mathbf{V}\mathbf{B}_\nu + o_P(1)$. Lemma 2 implies that $\mathbf{A}_\nu \to_d N(\mathbf{0}, \mathbf{V}_0)$, where

$$\mathbf{V}_0 = \begin{bmatrix} \mathbf{W}\mathbf{D}^{-1} - \mathbf{W}\mathbf{Z}\mathbf{Z}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Because $\mathbf{H}$ is full column rank, it can be shown that $\mathbf{V}$ is nonsingular. It follows that $\mathbf{B}_\nu \to_d N(\mathbf{0}, \mathbf{V}^{-1}\mathbf{V}_0\mathbf{V}^{-1})$. Going through some tedious algebra, the covariance matrix simplifies and we have that

$$\text{(A.4)}\qquad \begin{bmatrix} \nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ \nu^{-1/2}\hat{\boldsymbol{\lambda}}_\pi \end{bmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & (\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}\mathbf{H})^{-1} \end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}$ is as defined in the statement of the theorem. Now Lemma 3 implies that $\nu^{-1/2}\hat{\boldsymbol{\lambda}}_\pi = \nu^{-1/2}\mathbf{G}(\mathbf{Z}^T\mathbf{Y})\hat{\boldsymbol{\lambda}} = \nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}} + o_P(1)$. Thus, the limiting result of (A.4) is unchanged if we replace $\hat{\boldsymbol{\lambda}}_\pi$ by $\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}$. Finally, by properties of the normal distribution, the block diagonal structure of the variance matrix in (A.4) implies that $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ are asymptotically independent. □

PROOF OF LEMMA 6. The approximating equations (A.3) in the proof of Theorem 3, with $\hat{\boldsymbol{\lambda}}_\pi$ replaced by $\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}$, can be augmented as

$$\text{(A.5)}\qquad \begin{bmatrix} \nu^{-1/2}(\mathbf{Z}^T\mathbf{Y} - \boldsymbol{\gamma}) \\ \nu^{-1/2}\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\mathbf{D}^{-1} & -\mathbf{H} \\ \mathbf{0} & -\mathbf{H}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \nu^{-1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ \nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ \nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}} \end{bmatrix} + o_P(1).$$

Using properties of population matrices and a result cited in the proof of Lemma 2, the left-hand side can be written as

$$\begin{bmatrix} \mathbf{Z}^T \\ \mathbf{D}^{-1} - \mathbf{Z}\mathbf{Z}^T \\ \mathbf{0} \end{bmatrix} \nu^{-1/2}(\mathbf{Y} - \mathbf{m}).$$

Lemma 1 states that $\nu^{-1/2}(\mathbf{Y} - \mathbf{m}) \to_d N(\mathbf{0}, \mathbf{W}\mathbf{D} - \mathbf{W}\mathbf{D}\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D})$. Therefore, the left-hand side of (A.5) has a normal limiting distribution with variance matrix

$$\begin{bmatrix} \mathbf{Z}^T \\ \mathbf{D}^{-1} - \mathbf{Z}\mathbf{Z}^T \\ \mathbf{0} \end{bmatrix} (\mathbf{W}\mathbf{D} - \mathbf{W}\mathbf{D}\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D})[\mathbf{Z},\ \mathbf{D}^{-1} - \mathbf{Z}\mathbf{Z}^T,\ \mathbf{0}].$$

By properties of population and sampling constraint matrices, and after some tedious algebra, this variance matrix can be shown to have the block-diagonal form

$$\begin{bmatrix} \mathbf{Q}_R\mathbf{Q}_R^T\mathbf{D}(\mathbf{w}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\mathbf{D}^{-1} - \mathbf{W}\mathbf{Z}\mathbf{Z}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

It follows that

$$\begin{bmatrix} \nu^{-1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ \nu^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ \nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}} \end{bmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}),$$

where the variance matrix $\boldsymbol{\Gamma}$ simplifies as

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\mathbf{D}^{-1} & -\mathbf{H} \\ \mathbf{0} & -\mathbf{H}^T & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Q}_R\mathbf{Q}_R^T\mathbf{D}(\mathbf{w}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\mathbf{D}^{-1} - \mathbf{W}\mathbf{Z}\mathbf{Z}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\times \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}\mathbf{D}^{-1} & -\mathbf{H} \\ \mathbf{0} & -\mathbf{H}^T & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{Q}_R\mathbf{Q}_R^T\mathbf{D}(\mathbf{w}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\mathbf{H}^T\mathbf{D}\mathbf{W}^{-1}\mathbf{H})^{-1} \end{bmatrix},$$

where $\boldsymbol{\Sigma}$ is as defined in the statement of Theorem 3. By properties of the normal distribution, the block diagonal variance matrix implies that $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ are mutually asymptotically independent. □

PROOF OF THEOREM 5. Define $\bar{\boldsymbol{\pi}} = \mathbf{N}^{-1}\mathbf{Y}$. Show that $G^2 = X^2 + o_P(1)$,

$$\begin{aligned} G^2 &= 2\mathbf{Y}^T \log(\mathbf{Y}/\hat{\mathbf{m}}) \\ &= 2\bar{\boldsymbol{\pi}}^T\mathbf{N}\log(\bar{\boldsymbol{\pi}}/\hat{\boldsymbol{\pi}}) \\ &= 2\bar{\boldsymbol{\pi}}^T\mathbf{N}\log\bigl(\mathbf{1} + \mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})\bigr) \\ &= 2\bar{\boldsymbol{\pi}}^T\mathbf{N}\bigl[\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}) \\ &\qquad - \tfrac{1}{2}\mathbf{D}\bigl(\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})\bigr)\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}) + O_P(\nu^{-3/2})\bigr] \\ &= (\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T\mathbf{N}\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}) + o_P(1) \\ &= (\mathbf{Y} - \hat{\mathbf{m}})^T\mathbf{D}^{-1}(\hat{\mathbf{m}})(\mathbf{Y} - \hat{\mathbf{m}}) + o_P(1) \\ &= X^2 + o_P(1), \end{aligned}$$

where the fourth equality follows using the expansion $\log(\mathbf{1} + \mathbf{x}) = \mathbf{x} - \frac{1}{2}\mathbf{D}(\mathbf{x})\mathbf{x} + O(|\mathbf{x}|^3, \|\mathbf{x}\| \to 0)$ and the fifth equality uses the fact that, owing to the form of the restricted likelihood equations (7.2), $\mathbf{1}^T(\mathbf{Y} - \hat{\mathbf{m}}) = 0$.

To show that $W^2 = X^2 + o_P(1)$,

$$\begin{aligned} W^2 &= \mathbf{h}(\mathbf{Y})^T\big[\mathbf{H}(\mathbf{Y})^T\mathbf{D}(\mathbf{Y})\mathbf{H}(\mathbf{Y})\big]^{-1}\mathbf{h}(\mathbf{Y}) \\ &= \nu^{1/2}\mathbf{h}(\bar{\boldsymbol{\pi}})^T\big[\mathbf{H}(\bar{\boldsymbol{\pi}})^T\mathbf{D}(\bar{\boldsymbol{\pi}})\nu\mathbf{N}^{-1}\mathbf{H}(\bar{\boldsymbol{\pi}})\big]^{-1}\nu^{1/2}\mathbf{h}(\bar{\boldsymbol{\pi}}) \\ &= \nu^{1/2}\mathbf{h}(\bar{\boldsymbol{\pi}})^T\big[\nu\mathbf{G}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{H}(\hat{\mathbf{m}})^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\mathbf{G}^{-1}(\hat{\boldsymbol{\gamma}})\big]^{-1}\nu^{1/2}\mathbf{h}(\bar{\boldsymbol{\pi}}) + o_P(1) \\ &= \hat{\boldsymbol{\lambda}}^T\mathbf{H}(\hat{\mathbf{m}})^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}} + o_P(1) \\ &= X^2 + o_P(1), \end{aligned}$$

where the second equality follows upon writing $\mathbf{h}(\mathbf{Y}) = \mathbf{G}(\hat{\boldsymbol{\gamma}})\mathbf{h}(\bar{\boldsymbol{\pi}})$ and using Proposition 4 to reexpress the inner matrix; the third equality follows because the inner matrix can be shown to converge in probability to the same constant matrix as the inner matrix of the previous expression and $\nu^{1/2}\mathbf{h}(\bar{\boldsymbol{\pi}}) = O_P(1)$; the fourth equality uses (i) the relationship $\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}} = -\mathbf{N}^{-1}\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}}$, which follows from the restricted likelihood equations (7.2), and (ii) the approximation $\mathbf{h}(\bar{\boldsymbol{\pi}}) = \mathbf{H}(\hat{\boldsymbol{\pi}})^T(\bar{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}) + o_P(\nu^{-1/2}) = -\mathbf{G}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{H}(\hat{\mathbf{m}})^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}} + o_P(\nu^{-1/2})$.

To show that $X^2 \to_d \chi^2(u)$,

$$\begin{aligned} X^2 &= \hat{\boldsymbol{\lambda}}^T\mathbf{H}(\hat{\mathbf{m}})^T\mathbf{D}(\hat{\mathbf{m}})\mathbf{H}(\hat{\mathbf{m}})\hat{\boldsymbol{\lambda}} \\ &= \hat{\boldsymbol{\lambda}}^T\mathbf{G}(\hat{\boldsymbol{\gamma}})\mathbf{H}(\hat{\boldsymbol{\pi}})^T\mathbf{D}(\hat{\boldsymbol{\pi}})\mathbf{N}^{-1}\mathbf{H}(\hat{\boldsymbol{\pi}})\mathbf{G}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\lambda}} \\ &= \big[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}\big]^T \\ &\quad \times \big\{\big[\mathbf{G}^{-1}(\boldsymbol{\gamma})\mathbf{G}(\hat{\boldsymbol{\gamma}})\big]\mathbf{H}(\hat{\boldsymbol{\pi}})^T\mathbf{D}(\hat{\boldsymbol{\pi}})\nu\mathbf{N}^{-1}\mathbf{H}(\hat{\boldsymbol{\pi}})\big[\mathbf{G}(\hat{\boldsymbol{\gamma}})\mathbf{G}^{-1}(\boldsymbol{\gamma})\big]\big\}\big[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}\big] \\ &= \big[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}\big]^T\big[\mathbf{H}(\boldsymbol{\pi})^T\mathbf{D}(\boldsymbol{\pi})\mathbf{W}^{-1}\mathbf{H}(\boldsymbol{\pi})\big]\big[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}\big] + o_P(1) \\ &\xrightarrow{d} \chi^2(u), \end{aligned}$$

where the second equality uses Proposition 4; the fourth equality uses facts (i) $[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}] = O_P(1)$ and (ii) the inner matrix is the probability limit of the inner matrix in braces of the previous expression; and the limiting result follows because $[\nu^{-1/2}\mathbf{G}(\boldsymbol{\gamma})\hat{\boldsymbol{\lambda}}]$ converges to a normal distribution (of dimension u) with mean zero and variance $[\mathbf{H}(\boldsymbol{\pi})^T\mathbf{D}(\boldsymbol{\pi})\mathbf{W}^{-1}\mathbf{H}(\boldsymbol{\pi})]^{-1}$. □

PROOF OF AX3. By Lemma 1, $\nu^{-1/2}(\mathbf{Y} - \mathbf{m})$ has a normal limiting distribution with variance matrix $\mathbf{WD} - \mathbf{DZ}_F\mathbf{Z}_F^T\mathbf{WD}$. Now

$$\begin{aligned} \nu^{-1}\big[\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}\big] &= \nu^{-1}\mathbf{ND}(\hat{\boldsymbol{\pi}}) - \mathbf{D}(\hat{\boldsymbol{\pi}})\mathbf{Z}_F\mathbf{Z}_F^T\nu^{-1}\mathbf{ND}(\hat{\boldsymbol{\pi}}) \\ &\xrightarrow{P} \mathbf{WD} - \mathbf{DZ}_F\mathbf{Z}_F^T\mathbf{WD}. \end{aligned}$$

Therefore, by Definition 6 $[\hat{\mathbf{D}} - \mathbf{N}^{-1}\hat{\mathbf{D}}\mathbf{Z}_F\mathbf{Z}_F^T\hat{\mathbf{D}}]$ is an approximating variance and the approximation result follows. □

PROOF OF AX7. Recalling that $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$ because the model is assumed to hold, we can write

$$\begin{aligned}
&\nu^{1/2}\mathbf{G}^{-1}(\boldsymbol{\gamma})\big(\mathbf{h}(\mathbf{Y}) - \mathbf{h}(\mathbf{m})\big) \\
&\quad = \nu^{1/2}\mathbf{G}(\hat{\boldsymbol{\gamma}}/\boldsymbol{\gamma})\big[\mathbf{h}(\mathbf{N}^{-1}\mathbf{Y}) - \mathbf{h}(\boldsymbol{\pi})\big] + \nu^{1/2}[\mathbf{G}(\hat{\boldsymbol{\gamma}}/\boldsymbol{\gamma}) - \mathbf{G}(\mathbf{1})]\mathbf{h}(\boldsymbol{\pi}) \\
&\quad = \nu^{1/2}\big[\mathbf{h}(\mathbf{N}^{-1}\mathbf{Y}) - \mathbf{h}(\boldsymbol{\pi})\big] + o_P(1) \\
&\quad \xrightarrow{d} N\big(\mathbf{0}, \mathbf{H}^T(\mathbf{W}^{-1}\mathbf{D} - \mathbf{W}^{-1}\mathbf{D}\mathbf{Z}\mathbf{Z}^T\mathbf{D})\mathbf{H}\big) \sim N\big(\mathbf{0}, \mathbf{H}^T\mathbf{W}^{-1}\mathbf{D}\mathbf{H}\big),
\end{aligned}$$

where the limiting result follows from Lemma 4 and the simplification of the variance matrix follows from Proposition 6. Also, by Proposition 4 $\hat{\mathbf{H}} = \mathbf{H}(\hat{\mathbf{m}}) = \mathbf{N}^{-1}\mathbf{H}(\hat{\boldsymbol{\pi}})\mathbf{G}(\hat{\boldsymbol{\gamma}})$ and

$$\begin{aligned}
\nu\mathbf{G}^{-1}(\boldsymbol{\gamma})\big[\hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}}\big]\mathbf{G}^{-1}(\boldsymbol{\gamma}) \;&=\; \mathbf{G}(\hat{\boldsymbol{\gamma}}/\boldsymbol{\gamma})\mathbf{H}(\hat{\boldsymbol{\pi}})^T\nu\mathbf{N}^{-1}\mathbf{D}(\hat{\boldsymbol{\pi}})\mathbf{H}(\hat{\boldsymbol{\pi}})\mathbf{G}(\hat{\boldsymbol{\gamma}}/\boldsymbol{\gamma}) \\
&\xrightarrow{P} \mathbf{H}^T\mathbf{W}^{-1}\mathbf{D}\mathbf{H}.
\end{aligned}$$

Finally, because $\mathbf{G}^{-1}(\boldsymbol{\gamma}) = \operatorname{diag}\{\gamma_{u(j)}^{-p(j)} : j = 1, \ldots, c\}$ and $\gamma_{u(j)}/\nu$ converges to a positive constant, Definition 6 implies that $\mathbf{h}(\mathbf{Y}) - \mathbf{h}(\mathbf{m}) \sim \widehat{\mathrm{AN}}(\mathbf{0}, \hat{\mathbf{H}}^T\hat{\mathbf{D}}\hat{\mathbf{H}}) = O_P(\nu^{\mathbf{p}-1/2})$. □

PROOF OF THEOREM 8. Because $\mathbf{S} \in \mathcal{H}(\mathbf{Z})$, we can write

$$\begin{aligned}
&\nu^{1/2}\mathbf{G}^{-1}(\boldsymbol{\gamma})\big(\mathbf{S}(\hat{\mathbf{m}}) - \mathbf{S}(\mathbf{m})\big) \\
&\quad = \mathbf{G}^{-1}(\boldsymbol{\gamma})\mathbf{G}(\hat{\boldsymbol{\gamma}})\nu^{1/2}\big(\mathbf{S}(\hat{\boldsymbol{\pi}}) - \mathbf{S}(\boldsymbol{\pi})\big) + \nu^{1/2}\mathbf{D}(\mathbf{S}(\boldsymbol{\pi}))\big(\mathbf{d}_{\mathbf{G}}(\hat{\boldsymbol{\gamma}}/\boldsymbol{\gamma}) - \mathbf{d}_{\mathbf{G}}(\mathbf{1})\big).
\end{aligned}$$

Now, by Lemma 6, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\gamma}}$ are asymptotically independent. The limiting distribution follows from Lemma 3 and Theorem 3. The simplified form for the limiting variance follows using the fact that $\mathbf{D}(\mathbf{S}(\boldsymbol{\pi}))(\partial\mathbf{d}_{\mathbf{G}}(\mathbf{1})/\partial\boldsymbol{\gamma}^T) = (\partial\mathbf{S}(\boldsymbol{\pi})/\partial\boldsymbol{\pi}^T)\mathbf{D}(\boldsymbol{\pi})\mathbf{Z}$ by identity (5.1) following Proposition 5. □

Acknowledgments. I would like to thank the referees, an Associate Editor and the Editor for their helpful comments and suggestions.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

AITCHISON, J. and SILVEY, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29** 813–828.

AITCHISON, J. and SILVEY, S. D. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *J. Roy. Statist. Soc. Ser. B* **22** 154–171.

ANDERSEN, A. H. (1974). Multidimensional contingency tables. *Scand. J. Statist.* **1** 115–127.

APOSTOL, T. M. (1974). *Mathematical Analysis*, 2nd ed. Addison–Wesley, Reading, MA.

BAKER, S. G. (1994). The multinomial–Poisson transformation. *The Statistician* **43** 495–504.

BERGSMA, W. P. (1997). *Marginal Models for Categorical Data*. Tilburg Univ. Press.

BERGSMA, W. P. and RUDAS, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159.

BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **25** 220–233.

BIRCH, M. W. (1964). A new proof of the Pearson–Fisher theorem. *Ann. Math. Statist.* **35** 817–824.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.

CHAMBERS, R. L. and WELSH, A. H. (1993). Log-linear models for survey data with nonignorable nonresponse. *J. Roy. Statist. Soc. Ser. B* **55** 157–170.

CHEN, T. and FIENBERG, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30** 629–642.

CHRISTENSEN, R. R. (1990). *Log-Linear Models*. Springer, New York.

CORMACK, R. M. and JUPP, P. E. (1991). Inference for Poisson and multinomial models for capture–recapture experiments. *Biometrika* **78** 911–916.

FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, London.

FIENBERG, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *J. Amer. Statist. Assoc.* **95** 643–647.

FIENBERG, S. E. and LARNTZ, K. (1976). Log linear representation for paired and multiple comparisons models. *Biometrika* **63** 245–254.

FLEMING, W. (1977). *Functions of Several Variables*, 2nd ed. Springer, New York.

GLONEK, G. F. V. and MCCULLAGH, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B* **57** 533–546.

GRIZZLE, J. E., STARMER, C. F. and KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25** 489–504.

HABER, M. (1986). Testing for pairwise independence. *Biometrics* **42** 429–435.

HABERMAN, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics* **29** 205–220.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.

IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5** 299–314.

IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

KULLBACK, S. (1971). Marginal homogeneity of multidimensional contingency tables. *Ann. Math. Statist.* **42** 594–606.

LANG, J. B. (1996a). On the comparison of multinomial and Poisson loglinear models. *J. Roy. Statist. Soc. Ser. B* **58** 253–266.

LANG, J. B. (1996b). Maximum likelihood methods for a generalized class of loglinear models. *Ann. Statist.* **24** 726–752.

LANG, J. B. (2000). Maximum-likelihood analysis for useful classes of multinomial–Poisson homogeneous-function models. Technical Report 298, Dept. Statistics and Actuarial Science, Univ. Iowa.

LANG, J. B. and AGRESTI, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89** 625–632.

LANG, J. B., MCDONALD, J. W. and SMITH, P. W. F. (1999). Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *J. Amer. Statist. Assoc.* **94** 1161–1171.

LIPSITZ, S. R., PARZEN, M. and MOLENBERGHS, G. (1998). Obtaining the maximum likelihood estimates in incomplete $R \times C$ contingency tables using a Poisson generalized linear model. *J. Comput. Graph. Statist.* **7** 356–376.

LYONS, N. I. and HUTCHESON, K. (1986). Estimation of Simpson's diversity when counts follow a Poisson distribution. *Biometrics* **42** 171–176.

MATTHEWS, J. N. S and MORRIS, K. P. (1995). An application of Bradley–Terry-type models to the measurement of pain. *Appl. Statist.* **44** 243–255.

PALMGREN, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68** 563–566.

ROSS, S. M. (1993). *Introduction to Probability Models*, 5th ed. Academic Press, San Diego, CA.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SILVEY, S. D. (1959). The Lagrangian multiplier test. *Ann. Math. Statist.* **30** 389–407.

STIGLER, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statist. Sci.* **9** 94–108.

STOKES, M. E., DAVIS, C. S. and KOCH, G. G. (2000). *Categorical Data Analysis Using the SAS System*, 2nd ed. SAS Institute, Cary, NC.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242
USA
E-MAIL: jblang@stat.uiowa.edu