LOGARITHMIC SERIES DISTRIBUTION AND ITS USE IN ANALYZING DISCRETE DATA

Jeffrey R. Wilson, Arizona State University
Tempe, Arizona 85287

1. Introduction

In a previous paper, Wilson and Koehler (1988) used the generalized-Dirichlet multinomial model to account for extra variation. The model allows for a second order of pairwise correlation among units, a type of assumption found reasonable in some biological data. In that paper, the two-way crossed generalized Dirichlet Multinomial model was used to analyze repeated measure on the categorical preferences of insurance customers. The number of respondents was assumed to be fixed and known.

In this paper a generalization of the model is made allowing the number of respondents m, to be random. Thus both the number of units m, and the underlying probability vector are allowed to vary. The model presented here uses the logarithmic series distribution to account for the variation among number of units and the Dirichlet distribution to model the probabilities. In particular the Dirichlet-Multinomial distribution is used to incorporate the two types of variation, and the logarithmic series distribution is used to account for variation among the number of units within a given time period. Ignoring either level of variation leads to underestimation of the true standard errors of estimated proportions.

Tallis (1962) proposed the use of the generalized-multinomial model for dependent multinomials. Wilson and Koehler (1988) extended the model to allow for a second random component. The extended model considered can be viewed as multivariate extensions of the beta-binomial and correlated binomial models considered by Kupper and Haseman (1978) and Crowder (1978) for binary data. Paul (1987) considered a modification of the beta-correlated binomial as, a means of analyzing affected foetuses in litters of live foetuses. Section 2 outlines the generalized-multinomial model. Section 3 discusses the Dirichlet-Multinomial model. Section 4 presents the generalized Dirichlet-Multinomial model. The logarithmic distribution is presented in Section 5. An extended generalized Dirichlet-Multinomial model is developed in Section 6. Tests for certain hypotheses and fit of the model are developed in Section 7. Parameter estimates are obtained in Section 8.

2. Generalized-Multinomial Model

One way to view the generalized-multinomial model is to consider mJ vectors of outcomes for a set of J units that are simultaneously subjected to a series of m trials. At each trial, each unit is classified as being in exactly one of I mutually exclusive states. Let the random variable X_{ijk} take the value 1 if the k-th trial of the j-th unit is observed to be in the i-th state, and zero otherwise.

The probability that X_{ijk} takes the value 1 is assumed to be π_i for any unit and any trial. Furthermore, for each unit the m trials are identical and independent so upon summing across trials $\underset{\sim}{X}_j = (X_{1j}, X_{2j}, \ldots, X_{Ij})'$, the vector of counts for the j-th unit has a multinomial distribution with probability vector $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_I)'$ and sample size m. However, responses given by the J units at a particular trial may be correlated, producing a set of J correlated multinomial random vectors, $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_J$.

Tallis (1962) developed a model for this situation, which he called the generalized-multinomial distribution in which a single parameter ρ, is used to reflect the common dependency between any two of the dependent multinomial random vectors. The distribution of the category total $X_{ij} = \sum_{k=1}^{m} X_{ijk}$ is binomial with sample size m and parameter π_i, for each unit. Tallis formalized the dependencies among unit totals by specifying the joint moment generating function as

$$G_J(\underset{\sim}{\mu}) = \rho \left\{ \sum_{i=0}^{m} p_i \left(\prod_{j=1}^{J} e^{u_j} \right)^i \right\} + (1-\rho) \left\{ \prod_{j=1}^{J} P(e^{u_j}) \right\}, \quad 1,2.....I, \tag{2.1}$$

where $P(e^{u_r}) = \sum_{i=1}^{I} p_i e^{iu_r}$ and $\underset{\sim}{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)'$. The parameter ρ appearing in (2.1) is the correlation coefficient between X_{ij} and $X_{ij'}$, for any $j \neq j'$. When $\rho \neq 0$, $G_J(\underset{\sim}{u})$ is a weighted mean of moment generating function for a distribution with perfect correlation and one with complete independence, the weights being ρ and $1-\rho$ respectively. Altham (1978) proposed a similar model for a joint moment generating function for correlated binary variables.

Consider the overall vector of category totals $\underset{\sim}{X} = \sum_{j=1}^{J} \underset{\sim}{X}_j$. From the moment generating function in (2.1) it can be shown that $E(\underset{\sim}{X}) = Jm\underset{\sim}{\pi}$ and $V(\underset{\sim}{X}) = Jm\{1+(J-1)\rho\}M_{\underset{\sim}{\pi}}$ for the generalized-multinomial model, where $M_{\underset{\sim}{\pi}} = \text{diag}(\underset{\sim}{\pi}) - \underset{\sim}{\pi}\underset{\sim}{\pi}'$ and $\text{diag}(\underset{\sim}{\pi})$ is a diagonal matrix with diagonal elements provided by the vector $\underset{\sim}{\pi}$. Consequently, $\underset{\sim}{\pi} = (Jm)^{-1}\underset{\sim}{X}$ is an unbiased estimator for $\underset{\sim}{\pi}$. Tallis (1962) proposed estimators for ρ, but he did not discuss techniques for making inferences about $\underset{\sim}{\pi}$. We consider here a technique for making such inferences.

One approach is to use the limiting normal distribution of $\underset{\sim}{X}$ as $m \to \infty$. At trial k consider an IJ-dimensional response vector $\underset{\sim}{X}_{k(J)} = (\underset{\sim}{X}'_{1k},$

$X'_{2k}, \ldots, X'_{Jk})'$ where $X_{jk} = (X_{1jk}, X_{2jk}, \ldots, X_{Ijk})'$ is the response vector for the j-th unit at the k-th trial. Define $X_{(J)} = \sum_{k=1}^{m} X_{k(J)}$

Since the $X_{k(J)}$ vectors are independent and the first and second moments of $X_{k(J)}$ are finite, the multivariate central limit theorem implies that $m^{-\frac{1}{2}}(X_{(J)} - m\mu) \to N_{IJ}(0, \Sigma)$, as $m \to \infty$, where $\mu = 1_J \otimes \pi$, $\Sigma = M_\pi \otimes Q$, 1_J is a J-dimensional vector of ones, $\otimes$ denotes direct product between and Q is a square matrix of dimension J with ones on the diagonal and ρ as each off diagonal element. Now $X = GX_{(J)}$ where $G = 1'_J \otimes E_I$, and E_I is the identity matrix of dimension I. Then, by Rao (1973 page 124) the limiting distribution of π is specified by

$$m^{\frac{1}{2}}(\hat{\pi} - \pi) \to N_I(0, (mJ)^{-1}\{1+(J-1)\rho\}M_\pi). \quad (2.2)$$

Given a consistent estimator for ρ, asymptotic chi-square tests involving sufficiently smooth functions of π can be obtained as Wald statistics,

$$X^2_W = mJ\{1+(J-1)\rho\}^{-1}[g(\hat{\pi})- g(\pi)]'[DM\hat{}D']^-[g(\hat{\pi})-g(\pi)] \quad (2.3)$$

where D is the matrix of first partial derivatives of g evaluated at π, and $[DM\hat{}D]^-$ is a generalized inverse of $DM\hat{}D$. The degrees of freedom correspond to the rank of $DM_\pi D'$. In some applications it may be necessary to replace replace D with D (i.e. D evaluated at π).

3. Dirichlet-Multinomial Model

For each of N units observe a multinomial vector of responses, with parameters $p = (p_1, p_2, \ldots, p_I)'$ and sample size S. Furthermore assume the probability vector p has a Dirichlet distribution with mean vector $\pi = (\pi_1, \pi_2, \ldots, \pi_I)'$ and scaling parameter α.

For this model the sum of the vector of counts has a Dirichlet-Multinomial distribution and the vector of proportions has first moment π and covariance matrix $N^{-1}(S+\alpha)(1+\alpha)^{-1}M_\pi$. The Dirichlet distribution provides a convenient model for describing variation among vectors of proportions since it has relatively simple mathematical properties. The Dirichlet-Multinomial model has been studied by Mosimann (1962) and Good (1965). Brier (1980) used the model to analyze sample proportions obtained from a single two-stage cluster sample. Koehler and Wilson (1986) extended some of Brier's results to analyze vectors of proportions obtained from several two-stage cluster samples.

4. Generalized Dirichlet-Multinomial Model

In this section a generalized Dirichlet-Multinomial model, Wilson and Koehler (1988), is reviewed for which the observed vectors of counts may be correlated as in the generalized-multinomial model. Suppose for a given time t, J units are randomly selected from a population for which the vectors of proportions are distributed with respect to a Dirichlet distribution with parameter σ and $\pi_t = (\pi_{1t}, \pi_{2t}, \ldots, \pi_{It})'$.

As in the generalized-multinomial model, the X_{jt} vectors (j=1,2.....J) are identically distributed and are not independent. The observations taken at time t on the J individuals are equally pairwise correlated as measured by the parameter ρ. The vector of total counts $X = \sum_{t=1}^{n} X_t$, where $X_t = \sum_{j=1}^{J} X_{jt}$ for the generalized Dirichlet-Multinomial model, has mean vector $E(X) = N\pi$ and covariance matrix $V(X) = NC\{1+\rho(J-1)\}M_\pi$, where N=nS is the total number of observations, S is the total number of units at time t and $C = (S+\sigma)(1+\sigma)^{-1}$. Using an argument similar to the one in section 2, it can be shown that $n^{-\frac{1}{2}}(X-N\pi) \to N_I(0, SC\{1+(J-1)\rho\}M_\pi)$ and tests of hypotheses about π or vector functions $g(\pi)$, where g is a continuous function with second partial derivatives, can be obtained using the large sample chi-square distributions for the Wald statistic

$$N\{C\{1+(J-1)\rho\}\}^{-1}(g(\hat{\pi}) - g(\pi))'(DM_\pi D')^-(g(\hat{\pi}) - g(\pi)), \quad (4.1)$$

where $[DM_\pi D']^-$ denotes the generalized inverse of $DM_\pi D'$, with degrees of freedom equal to rank of $DM_\pi D'$. The greater imprecision in the estimation for π due to variation in vectors of proportion among individuals is accounted for by the factor C which cannot be less than one. The consequence of ignoring this extra variation is an inflation of the type I error levels for such tests.

5. The Univariate Logarithmic Series Distribution

The logarithmic series distribution was introduced by Fisher, Corbett and Williams (1943) to investigate the distribution of butterflies in the Malayan Peninsula. It has been used in the sampling of quadrants for plant species, the distribution of animal species, population growth and in economic applications. Chatfield et al (1966) used the logarithmic series distribution to represent the distribution of numbers of items of a product purchased by a buyer in a specified time period. They point out that the logarithmic series has the advantage of dependency on only one parameter θ.

The random variable M has a logarithmic series distribution if the probability function

$$P(M=k) = a\theta^k/k \quad (k=1,2,\ldots; 0<\theta<1)$$

where $a=-[\log(1-\theta)]^{-1}$. The probabilities are equivalent to the terms in the series expansion of $-a\log(1-\theta)$. Thus it is a power series distribution. Johnson and Kotz (1969) and Patil and Wani (1965), have given the moment generating function of M as

$$E(e^{tM}) = [\log(1-\theta e^t)]/[\log(1-\theta)],$$

so the variance is

$$\text{var}(M) = a\theta(1-a\theta)(1-\theta)^{-2},$$

with mean

$E(M) = a\theta/(1-\theta)$. Since $\theta<1$, then it follows that the ratio

$$(k+1)P(M=k+1) = k\theta P(M=k) < 1.$$

Hence the maximum value of P(M=k) is at the initial value k=1 and the value of P(M=k) decreases as k increases.

In the model presented in section 6 the number of clusters J, and the total number of observation S from the J clusters may be expected to increase proportionately. The logarithmic series distribution is used here to explain the variation in the number of units in the cluster. Supposing the 'index of diversity', δ remains constant then S and J would be expected to be related by the formula $e^{J/\delta} = 1+S/\delta$. If S/δ is large then $e^{J/\delta} = S/\delta$.

The idea to use the logarithmic series distribution in conjunction with the Dirichlet distribution is a result of work done by Engen (1975). He demonstrated the use of the limit of the Dirichlet distribution in deriving the logarithmic series distributions.

Consider the joint conditional distribution of $m_1, m_2 \ldots m_J$ for each fixed sum $\Sigma m_j = S$ i.e.

$$P(M_j=m_j;j=1,2\ldots J \mid \Sigma m_j=S) = S!/J! \prod_{j=1}^{J} (1/m_j)/F(s,J),$$

for $1<m_j<S-(J-1)$, and for any integer $S>J$ and where $F(s,J)$ is the absolute value of the stirling number of the first kind. The sum ΣM_j, is a sufficient statistic since the conditional distribution do not depend upon the parameter θ. The sum ΣM_j, is a complete sufficient statistic for θ and $P(M_j=m_j \mid \Sigma M_j=S)$ is minimum variance unbiased estimator of the probability function of LSD. Shanmugam and Singh (1984) noted that $E(m_j \mid \Sigma M_j=s)=Sn^{-1}$ regardless of the underlying probability distribution for the random sample.

6. An Extended Generalized Dirichlet-Multinomial Model

The logarithmic series distribution discussed in the previous section is used to extend the generalized Dirichlet-Multinomial model. The number of units per cluster is assumed to vary according to a logarithmic series distribution. Thus both m and the probability vector associated with each cluster are allowed to be random variables and we have

$$\underset{\sim}{h(t)} = \int \Sigma\, g_m f_{\underset{\sim}{m}}(P,t)\phi(P)dP$$

where $f_{\underset{\sim}{m}}(P,t)\phi(P)$ represents the conditional distribution for given m, and represented here by the generalized Dirichlet-Multinomial. The term, $g_m=P(M=m)$ represents the marginal distribution of the sample sizes. Here the conditional distribution given a sum of random samples from logarithmic distribution is used to represent such a marginal distribution.

The problem of obtaining expressions for $\underset{\sim}{h(t)}$ is now considerably magnified by the nature of the expression for the conditional distribution given a sum of logarithmic series distribution variables. However, the first and second moments of the distribution h(t) can be found.

Under the generalized Dirichlet-Multinomial model the covariance matrix for the conditional distribution of $\underset{\sim}{X}$ for given m is

$$V_{\underset{\sim}{m}}(X) = mBM_{\underset{\sim}{\pi}}$$

where $B = RC\{1+\rho(J-1)\}$ and $R=nJ$. Thus the moments for the distribution given by h(t) are

$$V(\underset{\sim}{X}) = n^{-1}SBM_{\underset{\sim}{\pi}} + SJ^{-2}(S-J)(J-1)B^2\underset{\sim\sim}{\pi\pi'}$$

$$= n^{-1}SB\{M_{\underset{\sim}{\pi}} + nBJ^{-2}(S-J)(J-1)\underset{\sim\sim}{\pi\pi'}\}$$

and

$$E(\underset{\sim}{X}) = SR\underset{\sim}{\pi}.$$

It follows that the covariance matrix, $V(\underset{\sim}{X})$ can be written as

$$V(\underset{\sim}{X}) = SBM_{\underset{\sim}{\pi}} + Sn^2C^2\{1+\rho(J-1)\}^2(S-J)(J-1)\underset{\sim\sim}{\pi\pi'}$$

$$= SB\Delta_{\underset{\sim}{\pi}} - [SB\, SJ^{-2}(S-J)(J-1)B^2]\underset{\sim\sim}{\pi\pi'}$$

$$= SB\{\Delta_{\underset{\sim}{\pi}} - [1-J^{-2}B(S-J)(J-1)]\underset{\sim\sim}{\pi\pi'}\},$$

where

$$t=(1-a\theta)(1-\theta)^{-1}R^2B^{-1}.$$

From ( ) $a = [-\log(1-\theta)]^{-1}$ and $0<\theta<1$. Also $R^2B^{-1} = RC^{-1}\{1+\rho(J-1)\}^{-1}$

$$V(X)=SB\{[1-J^{-2}B(S-J)(J-1)]M_{\underset{\sim}{\pi}} + J^{-2}B(S-J)(J-1)\underset{\sim\sim}{\pi\pi'}.$$

The covariance matrix has some simularities with the covariance matrix under the general Dirichlet-Multinomial. In the extended model case the variance is a sum of the variation due to the generalized Dirichlet Multinomial and the variation due to the variation among the samples sizes. Thus when the variance among the sample sizes is small there is little difference between the two models, generalized Dirichlet Multinomial and the extended generalized Dirichlet Multinomial. Certainly there is no difference between the models when there is one unit per cluster.

Similar to the assumption in (2.2) with the appropriate covariance matrix and given consistent estimator for C and $\{1+\rho(J-1)\}$ asymptotic chi-square tests involving sufficiently smooth functions of $\underset{\sim}{\pi}$ can be obtained as Wald statistics,

$$X^2_{GLD} = [g(\hat{\underset{\sim}{\pi}})-g(\underset{\sim}{\pi})]' \, [\hat{DVD'}]^{-}[g(\hat{\underset{\sim}{\pi}})-g(\underset{\sim}{\pi})]$$

where $[\hat{DVD'}]^{-}$ is a generalized inverse of $\hat{DVD'}$ and V is a consistent estimate of V(X) in ( ). The degrees of freedom correspond to the rank of DVD'.

7. Test of the Model Assumptions

In using the extended generalized Dirichlet-Multinomial model there arethree basic assumptions:

a) the correlations between the units $\underset{\sim}{X}_j$, and $\underset{\sim}{X}_{j'}$, are constant for any $j \neq j'$ b) the $\underset{\sim}{X}_j$'s j=1, 2,...,J; are identically multinomially distributed and c) the sample sizes are distributed as logarithmic series distribution. Test statistics were presented to assess the validity of the first two assumptions by Wilson & Koehler (1988). Large sample tests for the covariance structure associated with the Dirichlet-Multinomial model were given by Wilson (1986) and by Koehler and Wilson (1986). Here we make mention of some procedures for testing

that $m_1, m_2 \ldots m_J$ belong to a logarithm series distribution. One method of testing that $m_1, m_2 \ldots m_J$ is a random sample from a logarithmic series distribution is to consider the characterization of the distribution, Shanmugen and Singh (1984). For any fixed s, let

$$Q=(\underset{\sim}{m}-\underset{\sim}{\mu})'\Sigma^{-}(\underset{\sim}{m}-\underset{\sim}{\mu})$$

be a test statistic where $\underset{\sim}{m}=(m_1,m_2\ldots,m_J)'$ is the observed vector, and $\underset{\sim}{\mu}=(u_1,\mu_2\ldots,\mu_J)'$ with $u_i=E(m_i|\Sigma m_i=s)$ is the vector of expected values and $\Sigma=\{\text{cov}(m_j m_i)|\Sigma m_i=S\}$ is the vector of weights. The rank of Σ is J-1. It can be shown that asymptotically

$$\text{cov}(m_j m_j)|\Sigma m_i=S) \sim S(S-J)(J-1)/J^2 \quad j=j'$$
$$\cong -S(S-J)/J^2 \quad j\neq j'.$$

The structure of the dispersion matrix Σ is of the intraclass correlation matrix type. Thus Q simplifies to

$$Q \sim (S-J)^{-1}[(S/J)\sum_{j=1}^{J} n_j^2 - S] .$$

Through the asymptotic distribution of $\sum_{j=1}^{J} m_j^2$, it can be shown (Shanmugam and Singh 1984) that for a given level of significance α, we would reject the null hypothesis that a random sample $m_1, m_2 \ldots m_J$ is from a logarithmic series distribution if

$$\left| \frac{\Sigma m_j^2 - S(S-J+1)}{\sqrt{S(J-1)\binom{S-J+1}{3}}} \right| \geq Z_{\alpha/2}$$

where $Z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution.

8. Parameter Estimates

The problem of estimating θ given values of J random variables $m_1, m_2, \ldots, m_J$; each having a logarithmic series distribution has been considered by Johnson and Kotz (1969). The maximum likelihood estimator θ satisfies the equation for $\bar{m}$ (the mean of the m's) where

$$\bar{m} = J^{-1}\sum_{j=1}^{J} m_j = -\hat{\theta}\{(1-\hat{\theta})\log(1-\hat{\theta})\}^{-1}. \quad (8.1)$$

Since the logarithmic distribution is a generalized power series distribution equation (8.1) can be solved by equating the sample and population means. Other estimators of θ are presented in Johnson and Kotz.

When using the logarithmic series distribution to obtain an extension of the generalized Dirichlet distribution to test hypotheses concerning $g(\underset{\sim}{\pi})$ there is no need to obtain estimates of θ. However, estimates of C and $\{1+\rho(J-1)\}$ must be obtained.

Methods of estimating C and $\{1+\rho(J-1)\}$ are presented by Wilson and Koehler (1988). One set of estimators can be obtained by constructing an IxJ and an Ixn table. From each table obtain the Pearson chi-square statistic $X^2_{(IJ)}$ and $X^2_{(In)}$ for testing independence in a two-way contingency table.

Then $\hat{C}=X^2_{(IJ)}/(I-1)(J-1)$ and $\{1+\rho(J-1)\}=X^2_{(In)}/(I-1)(n-1)$.

References

Altham, P. M. E. (1978). Two generalizations of the binomial distribution Applied Statistics 27, 162-167.

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. John Wiley and Sons, New York.

Anscombe, F. J. (1950). Sampling Theory of the Negative Binomial and Logarithmic Series Distributions. Biometrika 37, 358-82.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. Biometrika 67, 591-596.

Chatfield, C., Ehrenberg, A. S. C. and Goodhardt, G. J. (1966). Progress on a simplified model of stationary purchasing behavior (with discussion), Journal of The Royal Statistical Society, Series A, 129, 317-367.

Cochran, W. G. (1943). Analysis of variance for percentages based on unequal numbers. Journal of the American Statistical Association 38, 287-301.

Choi, J. W. (1987). A direct estimate of Intracluster correlation. Section on Survey. Research Methods. Proceedings of American Statistical Association.

Crosby, L.A. and Stephens, N. (1987). Effects of Relationship Marketing on the Satisfaction, Retention and Prices in the Life Insurance Industry. Journal of Marketing Research. (Forthcoming).

Crowder, M. J. (1978). Beta-binomial Anova for proportions. Applied Statistics 27, 34-37.

Engen, S. (1975). A note on the geometric series as a species frequency model. Miscellanea.

Fisher, R. A., Corbett, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals on a random sample from an animal population. Journal of Animal Ecology, 10, 446-56.

Good, I. J. (1965). Estimation of Probabilities. MIT Press, Cambridge, Massachusetts.

Healy, M. J. R. (1972). Animal litters as experimental units. Applied Statistics, 21, 155-159.

Johnson, N. L. and Kotz, S. (1969). Discrete Distributions. Houghton Mifflin Company, New York.

Koehler, K. J. and Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. Communication in Statistics Vol. A15, No. 10, Theory and Methods.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated bionomial model for the analysis of certain toxicological experiments. Biometrics 34, 69-76.

Landis, J. R. and Koch, G. G. (1977). A one-way component of variance model for categorical data. Biometrics, 33, 671-79.

Lawley, D. N. (1963). On testing a set of correlation coefficients for equality. Annals of mathematical statistics 34, 149-151.

Moore, D. S. (1977). Generalized inverses, Wald's method and construction of chi-squared tests of fit. Journal of the American Statistical Association 72: 131-137.

Moseman, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlation among proportion. Biometrika 49: 65-82.

Patil, G. P., and Bildekar, S., (1967). Multivariate Logarithmic Series Distribution as a Probability Model in Population and Community Ecology and Some of its Statistical Properties. Journal of American Statistical Association.

Patil, G. P. and Wani, J. K. (1965). On certain structural properties of the logarithmic series distribution and the first type Sterling distribution, Sankhya Series A, 27, 271-280.

Paul, S. R. (1987) On the Beta - Correlated Binomial (BCB) Distribution A Three Parameter Generalization of the Binomial Distribution Communication in Statistic, Vol. 6, (5), 1473-1478.

Shanmugam, R., and Singh, J., (1984). A characterization of the Logarithmic Series Distribution and Its Application. Communication in Statistics, 13(7), 865-875.

Tallis, G. M. (1962). The use of a generalized multinomial distribution in the estimation of correlation in discrete data. J. R. Statistical Soc., Series B, 24, 530-534.

Tallis, G. M. (1964). Further Models for estimating correlation in discrete data. J. R. Statist. Soc., Series B, 26, 82-85.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math., Soc., 54, 426-482.

Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effects in the dirichlet multinomial model. Communication in Statistics, Vol. A15, No. 4, Theory and Methods.

Wilson, J. R. and Turner, D. (1987). A simulation Study to compare different Test Statistics for Complex Sampling Data. Technical Report DIS, Arizona State University, Tempe, Arizona.