

ON THE ESTIMATION OF THE RELATIVE INCIDENCE PARAMETER OF NEGATIVE MULTINOMIAL DISTRIBUTION.

BY

ADEJUMO, ADEBOWALE OLUSOLA

DEPARTMENT OF STATISTICS

UNIVERSITY OF ILORIN

ILORIN, NIGERIA.

ABSTRACT

In sampling schemes, the easiest approach is to fix the sample size. This approach is sometimes abridged. However, when considering categorical variables and there is a particular sequence of outcomes of the same type that is of interest, this approach is sometimes abridged.

When the variable of interest is polychotomous, the negative multinomial distribution is applicable. In this research we examine properties of the estimator of the relative incidence (the proportion of the total outcome that is of interest) of the outcome upon which our attention is focused under negative multinomial and full multinomial distribution assumption.

KEY WORDS: Negative Multinomial and relative incidence.

1. INTRODUCTION

Sampling from Poisson, Binomial and Multinomial are common sampling plans in categorical data. These sampling models assume asymptotic independent cell counts, negative correlated cell counts and within each of the mutually independent strata, respectively. Bonett (1985).

There are, however some situations that restrict our sampling plan, especially when observing a run of an outcome out of many possible outcomes available. Sampling may have to be terminated when the required run is observed. For a sampling resulting into more than two outcomes, we assume a sampling plan based on the negative multinomial distribution. This is an extension of the negative Binomial distribution. In the case of negative multinomial distribution the sample size is not fixed and this makes the sampling plan fairly restricted.

Steyn (1959) used the multinomial as a sampling distribution for the cells in a two-way contingency table. He considered situations where observations are taken until a fixed number of observations, say five, is observed in a given cell, say the (1,1) cell. Here the count in the (1,1) cell is fixed a priori but the total number of observations in the table is not. In such situation we can no longer use the chi-square test for independence. Steyn (1959) derived an alternate test statistic that does not have an asymptotic chi-square distribution.

Rudolf (1967), extending the work of Steyn, considered a sampling model, which is part multinomial and part negative multinomial. For a two-way contingency table, this model arises fixing one of the row or column marginal totals instead of the total sample size (as for the multinomial case) or one of the cell counts (as in the negative multinomial case) is fixed. Thus the distribution of one set of marginal totals is negative multinomial. Rudolf showed that both the exact and the asymptotic chi-square test statistic for independence is identical to the usual ones under the multinomial model.

In section 3, we consider the estimation of relative frequencies using the methods of moment estimation. We conclude this work giving some examples in section 4.

2. NEGATIVE MULTINOMIAL AND ITS PROPERTIES

The random variable $X = (x_1, x_2, \dots, x_i)$ has a negative multinomial distribution with positive parameters $k, m_2, m_3, \dots, m_i$. Let x_1 be the response of the outcome upon which interest is particularly focused. If

$$P(X = x_1, x_2, \dots, x_i) = f(x_i / k, m_i)$$

$$= \binom{k + \sum_{i=2}^i x_i - 1}{x_1 - 1, x_2, \dots, x_i} \left(\frac{k}{\sum_{j=2}^i m_j + k} \right)^{x_1} \prod_{i=2}^i \left(\frac{m_j}{\sum_{j=2}^i m_j + k} \right)^{x_i}$$

$$f(x_i / k, m_i) = \frac{\left(k + \sum_{i=2}^i x_i - 1 \right)!}{(x_1 - 1)! (x_2! x_3! \dots x_i!)} \left(\frac{k}{\sum_{j=2}^i m_j + k} \right)^{x_1} \prod_{i=2}^i \left(\frac{m_j}{\sum_{j=2}^i m_j + k} \right)^{x_i} \dots\dots\dots 2.1$$

For $t = 2$, equation (2.1) reduces to a negative Binomial distribution. That is

$$P(X = x_1, x_2) = \binom{x_1 + x_2 - 1}{x_1} \left(\frac{k}{m+k} \right)^{x_1} \left(\frac{m}{m+k} \right)^{x_2} \dots \dots \dots 2.2$$

$m' = (k, m_2, \dots, m_t)$ is the mean vector. Clearly $\frac{k}{\sum_{i=2}^t m_i + k}$ is the relative incidence rate.

$E(X_1)$ can be obtained first obtaining $E(X_1 - 1)$. That is,

$$E(X_1 - 1) = \sum_{i=1}^t (x_i - 1) f(x_i / k, m_i)$$

$$E(X_1 - 1) = \sum_{i=1}^t (x_i - 1) \frac{\left(k + \sum_{i=2}^t x_i - 1 \right)!}{(x_1 - 1)! (x_2! x_3! \dots x_t!)} \left(\frac{k}{\sum_{j=2}^t m_j + k} \right)^{x_1} \prod_{i=2}^t \left(\frac{m_i}{\sum_{j=2}^t m_j + k} \right)^{x_i}$$

$$= \sum_{i=1}^t \frac{\left(\sum_{i=2}^t x_i - 1 \right) \left(\sum_{i=2}^t x_i - 2 \right)!}{(x_1 - 2)! (x_2! x_3! \dots x_t!)} \left(\frac{k}{\sum_{j=2}^t m_j + k} \right)^{x_1 - 1} \prod_{i=2}^t \left(\frac{m_i}{\sum_{j=2}^t m_j + k} \right)^{x_i}$$

$$\text{since } \sum_{i=1}^t \frac{\left(\sum_{i=2}^t x_i - 2 \right)!}{(x_1 - 2)! (x_2! x_3! \dots x_t!)} \left(\frac{k}{\sum_{j=2}^t m_j + k} \right)^{x_1 - 1} \prod_{i=2}^t \left(\frac{m_i}{\sum_{j=2}^t m_j + k} \right)^{x_i} = 1$$

$$\therefore E(X_1 - 1) = \frac{\left(\sum_{i=2}^t x_i - 1 \right) k}{\sum_{j=2}^t m_j + k}$$

$$\begin{aligned}
 E(X_1 - 1) &= \frac{\left(\sum_{j=2}^I x_j - 1 \right) k}{\sum_{j=2}^I m_j + k} \\
 \therefore E(X_1) &= \frac{\left(\sum_{j=2}^I x_j - 1 \right) k}{\sum_{j=2}^I m_j + k} + 1 \\
 \Rightarrow E(X_1) &= \left(1 - \frac{(k+1)}{\sum_{j=2}^I m_j + k} \right) k + 1 \dots\dots\dots 2.3
 \end{aligned}$$

But MLE of $\sum x_i = \sum m_i$ (Johnson and Kotz, 1969).

And since $E(X_1) \neq k$, it is a biased estimator. However $E(X_i) = m_i$ for $i \geq 2$

3. ESTIMATION OF RELATIVE INCIDENCE PARAMETER

The maximum likelihood estimate does not exist for k . (Bishop, Fienberg and Holland, 1975), (Neyman, 1965). In estimating (k) we shall therefore use method of moments.

Using method of moment, the estimate of k is obtained as follows: from equation 2.3

$$\begin{aligned}
 E(X_1) &= \left(1 - \frac{(k+1)}{\sum_{i=1}^I m_i + k} \right) k + 1 = x_1 \\
 \Rightarrow k - \frac{(k+1)k}{\sum_{i=1}^I m_i + k} + 1 &= x_1 \\
 \Rightarrow \frac{k \left(\sum_{i=1}^I m_i + k \right) - k(k+1)}{\sum_{i=1}^I m_i + k} &= x_1 - 1
 \end{aligned}$$

$$\begin{aligned}
 &\Rightarrow k \sum_{i=1}^t m_i - k = x_1 \sum_{i=1}^t m_i + kx_1 - \sum_{i=1}^t m_i - k \\
 &\Rightarrow k \sum_{i=1}^t m_i - kx_1 = \sum_{i=1}^t m_i (x_1 - 1) \\
 &\Rightarrow \hat{k} = \frac{\sum_{i=1}^t m_i (x_1 - 1)}{\sum_{i=1}^t m_i - x_1} \dots\dots\dots 3.2
 \end{aligned}$$

We represent the negative multinomial more conventionally as follows.

$$f(\underline{x}) = \binom{x_1 + \dots + x_t - 1}{x_1 - 1, \dots, x_t} \left(\frac{k}{\sum m_i + k} \right)^{x_1} \dots \left(\frac{m_i}{\sum m_i + k} \right)^{x_i} \dots\dots\dots 3.3$$

$$p_1 = \frac{k}{\sum_{i=2}^t m_i + k} \text{ is the relative incidence.}$$

x_i is sufficient and complete for m_i (Johnson and Kotz, 1969).

By Maximum Likelihood Estimation

$$\begin{aligned}
 \hat{p}_1 &= \frac{x_1}{\sum_{i=1}^t x_i} = \frac{k}{\sum_{i=2}^t m_i + k} \\
 \hat{p}_i &= \frac{\hat{m}_i}{\sum_{i=2}^t m_i + k} = \frac{x_i}{\sum_{i=1}^t x_i}
 \end{aligned}$$

For a given set of data, identify x_1 , the outcome in focus. Estimate k using equations 3.2 above, we obtain $p_1 = \frac{k}{\sum_{i=2}^t m_i + k}$ which is the relative incidence rate of negative multinomial distribution.

$p_i = \frac{x_i}{\sum_{i=1}^t x_i}$, $i \geq 2$ and p_1 is proportion of outcome in focus. Based on the underling assumption of full multinomial distribution.

4. EXAMPLES

Example 1 The data in Table 4.1 were obtained from the Federal Ministry of Health, Lagos, on morbidity from some notifiable diseases in 1989 and 1990. This kind of data is best examined by negative multinomial distribution because the sample size is not fixed. We terminated the data the following year (1990) because reported cases of Tuberculosis in 1990 were relatively very high. x_i is the response of the outcome of the disease i , while m_i is mean response of the outcome.

Diseases	1989	1990
Chicken pox	9903	10300
Filariasis	5649	8438
Meaningitis	7800	7804
Tuberculosis	12232	*20122

Table 4.1: Data on some diseases collected from FMH.

Here, the disease in focus is Tuberculosis.

Using equation 3.2 above the estimate of k is 39398.91, so that $p_1=0.324$.

This value if estimated using full multinomial distribution gives $p=0.245$,

This implies that an assumption of full multinomial distribution results in underestimation of the incidence rate.

Example 2. The data in table 4.2 is for another set of diseases. Here we are interested in obtaining the relative incidence for Leprosy because of the large increase in number of cases from 1989 to 1990.

Diseases	1989	1990
Leprosy	9122	*20557
Opthalinia Neonatorum	4906	4176
Infective Hepatitis	2916	5495
Viral influenza	2797	7918
Tetanus	2308	3763

Table 4.2: Reported cases of some diseases in 1989 & 1990.

Using equation 3.2 above the estimate of $k=57552.12$, so that $p_1=0.474$.

This value if estimated using full multinomial distribution gives $p=0.321$

This also implies that using full multinomial distribution gives an underestimate of the incidence rate.

Example 3. Using another set of data obtained from the University of Ilorin Teaching Hospital (UTH), Ilorin, on morbidity from some notifiable diseases collected in 1991 and 1992. This kind of data is best examined by assuming a negative multinomial distribution because the sample size is not fixed. We terminated the data in (1992) because reported cases of Typhoid in 1992 were relatively very high (see Table 4.3). x_i is the response of the outcome of the disease i , while m_i is mean response of the outcome.

Diseases	1991	1992
Typhoid	8737	*19003
Guinea worm	5479	6749
Relapsing fever	5156	8702
Tetanus	1103	896

Table 4.3: Data on diseases from UTH, Ilorin.

Using equation 3.2 above the estimate of $k=59531.21$, so that $p_1=0.516$.

Estimating p using full multinomial distribution gives $p=0.340$

This implies again that full multinomial distribution underestimates the incidence rate.

Example 4. We estimate the relative incidence of Trachoma in Table 4.4

Diseases	1991	1992
STD	18173	19934
Diphtheria	2849	2351
Syphilis	2032	3073
Trachoma	4042	*14704

Table 4.4: When the disease in focus is Trachoma.

Using equation 3.2 above the estimate of $k=26156.93$, so that $p_1=0.280$.

This value if estimated using full multinomial distribution gives $p=0.219$

This also implies full multinomial distribution is under estimating the incidence rate.

5. CONCLUSION

When the sample size is not fixed or predetermined, negative multinomial is preferred to a full multinomial distribution, which is best, when the sample size is fixed.

We observed that the p (relative incidence) obtained directly through the multinomial distribution was smaller than that obtained using negative multinomial distribution. The results rest on the wisdom in the use of the underlying distribution. There would be an under estimation if instead of using negative multinomial the full multinomial distribution is used.

REFERENCE:

1. Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975): "Discrete Multivariate Analysis", MIT press, Cambridge, Mass.
2. Bonett, D. G. (1985a): "A linear negative Multinomial model", *Statistics and Probability letters*, 3, 127-129.
3. Johnson, N. L. and Kot Z, S. (1969): "Distributions in Statistics". Discrete Distributions. Pp 296-297, New York: John Wiley and Sons.
4. McCullagh, P. and Nelder, J. A. (1989). "Generalized Linear Models"; London: Chapman and Hall.
5. Neyman, J. (1965): "Certain chance mechanisms involving discrete distribution", In *Classical and contagious discrete distributions*, pp 4-14. Oxford, Pergamon.
6. Rudolf, G. J. (1967): "A Quasi- Multinomial Type of Contingency Table", *South Africa Journal of Statistic*, 1: 59-65.
7. Steyn, H. S. (1959): "On X^2 test for contingency tables of negative Multinomial types". *Statistica, Neerlandica*.