Far East Journal of Theoretical Statistics

Published Online: October 2013
Available online at http://pphmj.com/journals/fjts.htm
Volume 44, Number 2, 2013, Pages 121-133

MUTUAL INFORMATION FOR THE MULTINOMIAL DISTRIBUTION

Walfredo R. Javier and Arjun K. Gupta

Mathematics Department
Southern University-BR
LA70813, U. S. A.
e-mail: walfredo_javier@subr.edu

Mathematics and Statistics Department
Bowling Green State University
Bowling Green, OH 43403, U. S. A.
e-mail: gupta@math.bgsu.edu

Abstract

The expression for the mutual information measure for the multinomial distribution is derived; the resulting information measure $T(X)$ is the difference of two terms: a constant term, denoted by $K(N, q)$ which is an expression solely in terms of N and q, where N is the trial size and q is the dimension of the random vector, and the other term is the square of the magnitude of the probability vector $(p_1, p_2, ..., p_q)^T$ of the multinomial distribution. Also, an unbiased estimator for the information measure and the derivation of its asymptotic distribution are presented.

Received: June 2, 2013; Accepted: June 27, 2013

2010 Mathematics Subject Classification: 62-XX.

Keywords and phrases: Kullback-Leibler measure of association, multinomial distribution, asymptotic sampling distribution, distribution of quadratic forms derived from a singular multivariate normal distribution, non-central chi-square distribution.

1. Introduction

For a $q \times 1$ random vector $X = (X_1, X_2, ..., X_q)^T$ with probability density function or probability mass function $f(X)$ a measure of dependence, denoted by $T(X)$, among the q component variates $X_1, X_2, ..., X_q$ has been introduced in the statistical research literature, see for example, Guerrero [3]. It is defined as

$$T(X) = \int \cdots \int f(x) \ln \left[f(x) \Big/ \prod_{i=1}^{q} f_i(x_i) \right]_{i=1}^{q} dx_1 dx_2 \cdots dx_q, \tag{1.1}$$

the q-fold integration is over the entire R^q Euclidean space and is replaced by summation if X has a discrete distribution; $f_i(X_i)$ is the marginal density function of the ith component X_i; ln is the natural logarithm base e. $T(X)$, is non-negative [Kullback (1959, p. 14)] for any distribution; it assumes the value 0 when the component vectors are independent; it is a special case of the mean information for discrimination between two competing distributions $f(X)$, and $g(X)$, defined in information theory [Kullback (1959, p. 6)] as

$$I(f : g) = \int \cdots \int f(x) \ln(f(x)/g(x)) dx \tag{1.2}$$

and is known as the Kullback-Leibler number.

$T(X)$ may be considered as a function of the probability vector parameter $(p_1, p_2, ..., p_q)^T$ of the distribution. In practical applications, this parameter vector will be estimated from the available data set, and there is a need to study its sampling distribution to make it useful in statistical inference. In this paper, we shall present the maximum likelihood estimator $\hat{T}(X)$ for the multinomial distribution, and then present results on its sampling distribution for large sample size. This paper is organized into 6 sections.

2. The Multinomial Distribution

An experiment may consist of performing N identical and independent trials; each trial has q number of exhaustive and mutually exclusive outcomes $O_1, O_2, ..., O_q$ with probability of occurrence $p_1, p_2, ..., p_q : 0 \le p_i \le 1$, $\sum_{i=1}^{q} p_i = 1$; these probabilities remain constant from trial to trial. If X_i denotes the frequency count of occurrences of outcome O_i, $i = 1, 2, ..., q$, then the vector $X = (X_1, X_2, ..., X_q)^T$ with $\sum_{i=1}^{q} x_i = N$, has the multinomial distribution with probability mass function

$$f(X) = P[X_1 = x_1, X_2 = x_2, ..., X_q = x_q] = \left(N! \Big/ \prod_{i=1}^{q} x_i! \right) \prod_{i=1}^{q} p_i^{x_i}. \quad (2.1)$$

It is well known that the marginal distribution of each X_i is Binomial with parameters N and p_i; and the joint marginal distribution of any subset of component variates is again multinomial.

When N is large, the distribution of the random vector X is (Cramer [2]) approximately multivariate singular normal $N_q(\vec{\mu}, \Sigma)$, where

$$\Sigma = [\sigma_{ij}], \quad \sigma_{ij} = -Np_i p_j, \quad i \ne j, \ \ \sigma_{ii} = Np_i(1 - p_i),$$

$$rank(\Sigma) = q - 1; \quad \vec{\mu} = (\mu_1, \mu_2, ..., \mu_q)', \quad \mu_i = Np_i \ (i = 1, 2, ..., q).$$

This information will be used to derive the asymptotic distribution of an unbiased estimator for $T(X)$.

3. Mutual Information Measure for the Multinomial Distribution

The mutual information $T(X)$ for the multinomial distribution will now be presented. The proof is provided in Appendix (6.1).

Theorem 1. *The mutual information* $T(X)$, *for the multinomial distribution*, *is given by*

$$T(X) = K(N, q) - \sum_{i=1}^{q} p_i^2, \tag{3.1}$$

where

$$K(N, q) = (q-1)\ln\left(\frac{N^N}{N!}\right) + \frac{q}{2}\ln(2\pi)$$

$$+ \frac{q\ln(N) - 1}{2} + N(1-q) + 1. \tag{3.2}$$

Remark (R3.1). As can be seen in the derivation of the expressions in (3.1) and (3.2) in Appendix (6.1), the equality in (3.1) is derived from the application of two approximations: Stirling's factorial approximation and the linearization of the natural logarithm function near the origin.

We now consider the two terms in (3.1):

(3.1) The term $K(N, q)$

It is to be noted that the term $K(N, q)$ defined in (3.2) is independent of parameter probability vector $(p_1, p_2, \ldots, p_q)^T$. It is linear in q with a verifiable negative slope $\frac{\partial}{\partial q} K(N, q)$; hence for a fixed sample size N, at $q = 2$ (a Binomial distribution), $T(X)$ is largest, and decreases with larger q. For a fixed q, $K(N, q)$ increases with larger sample size N. These statements are supported by an empirical evidence provided by numerical values summarized by the following table:

Table 1. Values of $K(N, q)$ for $q = 2, 3, 5, 10$

For $q = 2/q = 3$

N	$K(N, q)$
2	1.724/1.682
3	1.940/1.913
5	2.207/2.190
10	2.562/2.554
15	2.767/2.762
20	2.913/2.908
25	3.025/3.022
30	3.117/3.114
45	3.320/3.319

For $q = 5/q = 10$

N	$K(N, q)$
5	2.157/2.074
10	2.537/2.495
15	2.751/2.723
20	2.900/2.879
25	3.015/2.988
30	3.108/3.094

The numerical values for $K(N, q)$ in Table 1 were computed by implementing (3.2) in a simple program in TI-83; this program is documented in Appendix (6.2).

(3.2) Now the term $\sum_{i=1}^{q} p_i^2$

By elementary vector algebra arguments, it can be shown that

$$\max\left\{\sum_{i=1}^{q} p_i^2 : p_i \geq 0,\ i = 1, 2, \ldots, q;\ \sum_{i=1}^{q} p_i = 1\right\} = 1$$

and

$$\min\left\{\sum_{i=1}^{q} p_i^2 : p_i \geq 0,\ i = 1, 2, \ldots, q;\ \sum_{i=1}^{q} p_i = 1\right\} = 1/q$$

so that $1/q \leq \sum_{i=1}^{q} p_i^2 \leq 1$, and from (3.1),

$$K(N, q) - 1 \leq T(X) \leq K(N, q) - 1/q. \tag{3.2.1}$$

We require that $K(N, q) \geq 1$ so that $T(X) \geq 0$. For a fixed N, this will be achieved if $2 \leq q \leq q^*$, where $q^* = q^*(N)$ is the solution of the linear equation $K(N, q^*) = 1$ in the unknown q^*. This requirement appears to be the restriction on the validity of the formula for $T(X)$ as given in (3.1).

Setting the right hand side of (3.2) to 1, we can easily solve for $q^* = q^*(N)$ as

$$q^* = q^*(N) = [\ln(N^N/N!) + 1/2 - N][\ln(N^N/N!) + \ln\sqrt{2\pi N} - N]^{-1}. \tag{3.2.2}$$

Not much of a restriction as the next table of numerical values suggests.

Table 2. Ceiling values $q^*(N)$

N	5	10	15	20	25	30	35	40	45
$q^*(N)$	74	189	320	461	609	764	923	1087	1255

4. Unbiased Maximum Likelihood Estimator for $T(X)$

In some situations the true proportions p_i, $i = 1, 2, ..., q$ are not known; they are usually estimated by $\hat{p}_i = X_i/N$, their maximum likelihood estimates. Replacing p_i with $\hat{p}_i$ in the expression for $T(X)$ gives its maximum likelihood estimator:

$$\hat{T}(X) = K(N, q) - \sum_{i=1}^{q} (\hat{p}_i)^2,$$

or equivalently,

$$\hat{T}(X) = K(N, q) - (1/N^2)\sum_{i=1}^{q} (X_i)^2, \tag{4.1}$$

where the X_i 's are the observed frequency counts.

Taking the expectation of (4.1)', with $X_i \sim Bino(N, p_i)$, we have

$$E(\hat{T}(X)) = K(N, q) - 1/N^2 \sum_{i=1}^{q} E(X_i)^2$$

$$= K(N, q) - \frac{1}{N^2}\sum_{i=1}^{q}[Np_i(1 - p_i) + (Np_i)^2]$$

$$= K(N, q) - \left(1 - \frac{1}{N}\right)\sum_{i=1}^{q} p_i^2 - \frac{1}{N}.$$

This plainly shows that $\hat{T}(X)$ is asymptotically unbiased estimator for $T(X)$; furthermore, the last equation suggests the unbiased estimator

$$\widetilde{T}(X) = K(N, q) + \frac{1}{N} - \frac{N}{N-1}\sum_{i=1}^{q} X_i^2. \tag{4.2}$$

5. Asymptotic Sampling Distribution for the Unbiased Estimator $\widetilde{T}(X)$

From Cramer [2], for large N, the multinomial random vector X follows approximate multivariate singular normal distribution $X \sim N_q(\vec{\mu}, \Sigma)$; where

$$\vec{\mu}' = N(p_1, p_2, \ldots, p_q); \quad rank(\Sigma) = q - 1;$$

$$\Sigma = [\lambda_{ij}], \quad \lambda_{ij} = -Np_i p_j \text{ for } i \neq j, \lambda_{ii} = Np_i(1 - p_i). \tag{5.1}$$

We shall follow Rao (1965, p. 528) in identifying the subspace where the random vector X has a non-zero density function; and we shall write its density function as adopted from the same source.

The linear transformation $L : R^q \to R^q$ with matrix Σ has a null space of dimension one and has its corresponding orthogonal subspace of dimension $(q-1)$. Let B be the $(q \times (q-1))$ matrix whose columns are the unit basis vectors for the orthogonal subspace of $L : R^q \to R^q$. Then the asymptotic density function of X is

$$f(X) = \frac{(2\pi)^{-k/2}}{\sqrt{\lambda_1\lambda_2\cdots\lambda_k}} \exp\{(B'X - B'\vec{\mu})'(B'\Sigma B)^{-}(B'X - B'\vec{\mu})/2\}, \tag{5.2}$$

whenever X lies on the subspace orthogonal to the hyperplane $A'X = A'\vec{\mu}$; zero otherwise $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the non-zero eigenvalues of Σ; $k = q - 1$.

Note that $(B'\Sigma B)^{-}$ is a $(q-1)\times(q-1)$ non-singular matrix, and therefore the above density function is a non-singular multivariate density function restricted on the indicated subspace which is orthogonal to the hyperplane $A'X = A'\bar{\mu}$; the density function is zero on that hyperplane.

We now can focus on the random term $\sum_{i=1}^{q}(X_i)^2$ of the expression for $\widetilde{T}(X)$ in (4.2) and indicate the steps needed to write its density function. We recognize this as a quadratic form of a singular multivariate normal distribution. The relevant known result for a non-singular multivariate normal distribution is from Anderson [1, p. 77]. We quote and list as:

Theorem 2. *If V is a random vector of p components and is distributed as* $N(\Lambda, I)$, *then* $V'V$ *has density function*

$$f(v) = \frac{1}{2^{p/2}} \exp\left(-\frac{\tau^2 + \nu}{2}\right) \nu^{p/2-1} \sum_{\beta=0}^{\infty} \left(\frac{\tau^2}{4}\right)^{\beta} \frac{\nu^{\beta}}{\beta!\Gamma(p/2 + \beta)}, \qquad (5.3)$$

where $\tau^2 = \Lambda'\Lambda$ *is the non-centrality parameter*.

(5.3) gives the density of a non-central chi-square distribution with non-centrality parameter $\tau^2 = \Lambda'\Lambda$.

An appropriate non-singular linear transformation of the multinomial random vector *X*, of the form $CB'X$, would achieve the transformation of the variance-covariance matrix Σ in (5.2) into the identity matrix *I*. The matrix *B* is as defined in (5.2); it maps the random vector into the subspace orthogonal to the linear subspace $A'X = A'\bar{\mu}$. This would enable us to apply the above quoted result from Anderson; then, employing the back transformation $X = (CB')^{-1}CB'(X)$ should give us the desired density function for the quadratic form $X'X = \sum_{i=1}^{q-1} X_i^2$ and thus that of $\widetilde{T}(X)$ of equation (4.2). This program of derivation can be carried out, if desired, as outlined in this paragraph.

We would thus be able to write the sampling distribution of $\widetilde{T}(X)$ for large N, correct to within a constant factor, as the density of a non-central chi-square distribution.

For statistical inference applications, the estimate for unknown mean vector $\vec{\mu}^T = N(\widetilde{p}_1, \widetilde{p}_2, ..., \widetilde{p}_q)^T$ will be used to compute the non-centrality parameter $\tau^2 = \Lambda'\Lambda$ in equation (5.3).

6. Appendix

(6.1)

The validity of Theorem 1 will now be shown.

Re-write the right side of equation (1.1) as an expectation

$$T(X) = E\left\{\left[\ln f(X) \Big/ \prod_{i=1}^{q} f_i(x_i)\right]\right\}. \tag{6.1.1}$$

From the multinomial distribution with probability mass function (2.1),

$$\begin{aligned}
\ln\left[f(X) \Big/ \prod_{i=1}^{q} f_i(x_i)\right] &= \ln(f(X)) - \sum_{i=1}^{q} \ln(f_i(X_i)) \\
&= \left[\ln(N!) + \sum_{i=1}^{q} X_i \ln(p_i) - \sum_{i=1}^{q} \ln(X_i!)\right] \\
&\quad - \left[q \ln(N!) + \sum_{i=1}^{q} X_i \ln(p_i)\right. \\
&\quad + \sum_{i=1}^{q} (N - X_i) \ln(1 - p_i) - \sum_{i=1}^{q} \ln(X_i!) \\
&\quad \left. - \sum_{i=1}^{q} \ln((N - X_i)!)\right].
\end{aligned} \tag{6.1.2}$$

Simplifying (6.1.2), we have

$$\ln\left[f(X) \Big/ \prod_{i=1}^{q} f_i(x_i) \right]$$

$$= (1-q)\ln(N!) - \sum_{i=1}^{q} (N - X_i)\ln(1-p_i) + \sum_{i=1}^{q} \ln(N - X_i)!. \quad (6.1.3)$$

Using Stirling's factorial approximation $n! \approx \sqrt{2\pi n}(n)^n e^{-n}$ and $\ln(1-u) \approx -u$ for $0 < u < 1$; then summing up with respect to i, we have for the last term on the right hand side of (6.1.3),

$$\sum_{i=1}^{q} \ln(N - X_i)! = (q/2)\ln(2\pi) + (1/2)\left[\sum_{i=1}^{q} (\ln N - X_i/N)\right]$$

$$+ \sum_{i=1}^{q} (N - X_i)(\ln N - X_i/N) - \sum_{i=1}^{q} (N - X_i). \quad (6.1.4)$$

Using (6.1.4) in (6.1.3) and taking expectation gives

$$T(X) = (1-q)\ln N! + (q/2)\ln(2\pi) + (q\ln N - 1)/2 + (q-1)N\ln N - Nq$$

$$- \sum_{i=1}^{q} (N - Np_i)\ln(1-p_i) + \sum_{i=1}^{q} E(X_i^2)/N$$

$$= (1-q)\ln N! + (q/2)\ln(2\pi) + (q\ln N - 1)/2 + (q-1)N\ln N - Nq$$

$$+ \sum_{i=1}^{q} (N - Np_i)\, p_i + (1/N)\sum_{i=1}^{q} (Np_i(1-p_i) - N^2 p_i^2). \quad (6.1.5)$$

Simplifying, we have

$$T(X) = (1-q)\ln N! + (q/2)\ln(2\pi) + ((q-1)N\ln N - 1)/2$$

$$+ (q-1)N\ln N - Nq + N + 1 - \sum_{i=1}^{q} p_i^2$$

$$= [(q-1)\ln(N^N/N!) + (q/2)\ln(2\pi) - Nq + N + 1] - \sum_{i=1}^{q} p_i^2. \quad (6.1.6)$$

This is the expression (3.1) in Theorem 1.

(6.2)

Program *MULTI*

: prompt N, Q

: prod $(seq(1 + x/(N - x),\ x,\ 0,\ N - 1,\ 1)) \xrightarrow{store} R$

: $(q/2)\ln(2\pi) + (Q\ln(N) - 1)/2 \xrightarrow{store} S$

: $N(1 - Q) + 1 \xrightarrow{store} T$

: $(Q - 1)\ln R + S + T \xrightarrow{store} U$

: *display U*

:

Remarks

(R6.2.1) The second line in this program can compute $N^N/N!$ for any inputted positive integer up to $N = 230$ in the hand held scientific calculator TI-83.

(R6.2.2) This program can be modified to include the computation of q^* of equation (3.2.2); this was done to generate Table 2.

References

[1] T. W. Anderson, An Introduction to Multivariate Analysis, 3rd ed., Wiley, 1984.

[2] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, 1945.

[3] J. L. Guerrero, Multivariate mutual information, Comm. Statist. Theory Methods 23(5) (1994), 1319-1339.

[4] W. R. Javier and A. K. Gupta, Mutual information for the mixture of two multivariate normal distributions, Far East J. Theor. Stat. 26(1) (2008), 47-58.

[5] W. R. Javier and A. K. Gupta, Mutual information for certain multivariate distributions, Far East J. Theor. Stat. 29(1) (2009), 39-51.

[6] A. Mood, F. Graybill and D. Boes, Introduction to the Theory of Statistics, 3rd ed., McGraw-Hill, 1974.

[7] C. R. Rao, Linear Statistical Inference and its Applications, 2nd ed., Wiley, 2002.

[8] K. Solomon, Information Theory and Statistics, Dover, 1959.

[9] M. S. Srivastava and C. G. Khatri, An Introduction to Multivariate Statistics, Elsevier, North Holland, 1979.