MAXIMUM LIKELIHOOD ESTIMATION FOR THE MULTINOMIAL DISTRIBUTION

By C. RADHAKRISHNA RAO

Indian Statistical Institute, Calcutta

1. Introduction

In an earlier paper (Rao 1953), the author considered minimum Chi-Square [m.c.] estimates and derived some of their large sample properties. Using the same techniques, it is possible to derive similar properties of maximum likelihood [m.l.] estimates, and these properties seem to be worth recording. For instance, it is not explicitly made clear in literature what analytical properties are possessed by m.l. estimates, under the usual regularity conditions (see section 4 of this paper) on the hypothetical cell frequencies as functions of unknown parameters. In fact, many probability statements concerning the m.l. estimates are direct consequences of the continuity and differentiability of the m.l. estimate as a function of observed relative frequencies.

Wald (1949) proved, under some regularity conditions, that the m.l. estimate is consistent while Huzurbazar (1949) showed, under Cramer's regularity conditions (Cramer, 1946), that with probability tending to unity a consistent root of the m.l. equation provides a local maximum of the likelihood. It was not known whether at such a root the likelihood attains an absolute maximum. It is shown in this paper that in the case of the multinomial distribution, a m.l. estimate is, with probability 1, a root of the likelihood equation, and provides a maximum of the likelihood when the parameter is restricted to the roots. We thus have a complete theory of the method of m.l. for the multinomial distribution. No assumption is made about the existence of the m.l. estimate, but the existence and uniqueness are deduced as a consequence of the regularity conditions.

To make the arguments free from unnecessary complications only the one parameter case is considered. The additional complication in the proof for the multiparameter case is in establishing the existence of the roots of the likelihood equation. Once this is done, the rest of the argument is the same.

In the development of this paper, the following scheme is adopted. First, the consistency of the m.l. estimates of the hypothetical frequencies of the multinomial distribution is established without any regularity conditions whatsoever. Second, the consistency of the m.l. estimate of a parameter occurring in the specification of the hypothetical frequencies is established under a very natural restriction on the parameter,

Third, the m.l. estimate is identified (with probability 1 as the sample size tends to infinity) as *the* root of the likelihood equation providing the maximum of the likelihood with the parameter restricted to the roots, and its analytical properties discussed.

2. DEFINITIONS AND PRELIMINARY LEMMAS

The hypothetical frequencies in k classes are represented by $\pi_1(\theta), \ldots, \pi_k(\theta)$ and the observed relative frequencies by $p_1, \ldots, p_k$.

Definition 1: The likelihood of the hypothetical frequencies $\pi_1, \ldots, \pi_k$ given the observed relative frequencies $p_1, \ldots, p_k$ in a sample of size n is proportional to

$$L(\pi) = \pi_1^{np_1} \ldots \pi_k^{np_k} \qquad \ldots (2.1)$$

Definition 2.1: The m.l. estimate of the multinomial distribution is a set of values $\pi_1, \ldots, \pi_k$ (if it exists) belonging to a given admissible class A of distributions and for which the likelihood $L(\pi)$ or the equivalent expression

$$p_1 \log \pi_1 + \ldots + p_k \log \pi_k \qquad \ldots (2.2)$$

is a maximum when π is restricted to A.

Definition 2.2: The m.l. estimate of a parameter θ occurring in the specification of the hypothetical frequencies is a value θ, if one exists in the admissible set of values of θ, for which

$$p_1 \log \pi_1(\theta) + \ldots + p_k \log \pi_k(\theta) \qquad \ldots (2.3)$$

is a maximum.

Definition 3: The m.l. equation is obtained by equating the derivative of (2.1) to zero, i.e.,

$$\frac{p_1}{\pi_1}\frac{d\pi_1}{d\theta} + \ldots + \frac{p_k}{\pi_k}\frac{d\pi_k}{d\theta} = 0 \qquad \ldots (2.4)$$

Of course, the differentiability of $\pi_1, \ldots, \pi_k$ is assumed in this definition.

Definition 4: The maximum likelihood equation [or m.l.e.] estimate is that root (or a root) of the likelihood equation which provides the maximum of (2.3) when θ is restricted to the roots of (2.4).

Definition 5: A statistic $T(p_1, \ldots, p_k)$ which is a function of relative frequencies only is Fisher consistent [FC] for θ if $T\{\pi_1(\theta), \ldots, \pi_k(\theta)\} \equiv \theta$.

MAXIMUM LIKELIHOOD ESTIMATION FOR MULTINOMIAL DISTRIBUTION

In what follows, a neighbourhood of the point $\pi = (\pi_1, \ldots, \pi_k)$ in the space of *all* multinomial distributions is denoted by $N(\pi)$. The notation $p \in N(\pi)$ means that the point $p = (p_1, \ldots, p_k)$ is in a neighbourhood of $\pi = (\pi_1, \ldots, \pi_k)$.

Lemma 1: *Let θ be the parameter to be estimated. If*

(a) *$T(p_1, \ldots, p_k)$ is FC for θ,*

(b) *$T(p_1, \ldots, p_k)$ admits continuous first partial derivatives,*

then (i) *$T(p_1, \ldots, p_k)$ is asymptotically normally distributed with mean θ and variance*

$$V(T) = \left\{ \sum \pi_i \left(\frac{dT}{d\pi_i} \right)^2 - \left(\sum \pi_i \frac{dT}{d\pi_i} \right)^2 \right\} \div n \qquad \ldots (2.5)$$

and (ii)

$$nV(T) \geqslant 1/i(\theta) \qquad \ldots (2.6)$$

where $i(\theta) = \sum \frac{1}{\pi_i} \left(\frac{d\pi_i}{d\theta} \right)^2$ is information as defined by Fisher.

The conditions (a) and (b) need be satisfied only in the neighbourhood of the true value of θ. For a detailed proof of these results see Rao (1955) or Kallianpur and Rao (1956).

Lemma 2: *For $a_1, \ldots, a_k$ positive and $b_1, \ldots, b_k$ non-negative such that $\Sigma a_i = \Sigma b_i$,*

$$\sum_1^k a_i \log \frac{b_i}{a_i} \leqslant 0. \qquad \ldots (2.7)$$

The equality is attained when and only when $a_i = b_i$ for all i.

We follow a proof similar to that given by Kullback and Liebler (1951) in the case of continuous distributions. The function $\log x$ has the expansion

$$\log x = \log 1 + (x-1) - (x-1)^2 \frac{1}{2y^2}, \quad y \in (1, x)$$

On substituting this for each term in the expression (2.7) we have

$$\sum a_i \log \frac{b_i}{a_i} = -\sum a_i \left(\frac{b_i}{a_i} - 1 \right)^2 \frac{1}{2y_i^2}, \quad y_i \in \left(1, \frac{b_i}{a_i} \right) \qquad \ldots (2.8)$$

which is always negative and can be zero when and only when $b_i = a_i$ for all i.

3. Consistency of m. l. estimates

3.1. *Estimation of the multinomial distribution under no regularity conditions.* Let A be the admissible set of hypothetical cell frequencies. No assumption is made about this set, which is arbitrary but fixed. The object of estimation is to choose one element of the set as an estimate, given an observed set of relative frequencies $p = (p_1, \ldots, p_k)$. We denote a typical member of A by $\pi = (\pi_1, \ldots, \pi_k)$ and the true value by $\pi^0 = (\pi_1^0, \ldots, \pi_k^0)$. The m.l. estimate is that value of π belonging to A such that $L(\pi)$ is a maximum, with π restricted to A.

Let us suppose first that the m.l. estimate exists at least when $p \in N(\pi^0)$, and represent it by $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$. Then

$$L(\pi^0) \leqslant L(\hat{\pi}) \leqslant L(p) \qquad \ldots \quad (3.1)$$

since $L(p)$ is the unconditional maximum of the likelihood, while $L(\hat{\pi})$ is the maximum when π is restricted to A. By taking logarithms, (3.1) can be written as

$$\Sigma p_i \log \pi_i^0 \leqslant \Sigma p_i \log \hat{\pi}_i \leqslant \Sigma p_i \log p_i \qquad \ldots \quad (3.2)$$

It follows from the law of large numbers that (with probability 1) the first and the last sums in (3.2) tend to the common limit $\Sigma \pi_i^0 \log \pi_i^0$. Hence, with probability 1

$$\Sigma p_i \log \hat{\pi}_i \to \Sigma \pi_i^0 \log \pi_i^0 \qquad \ldots \quad (3.3)$$

Without loss of generality we can suppose $\pi_i^0 \neq 0$ for each i, in which case when p is close to π^0, the $\log \hat{\pi}_i$ are bounded, by (3.3). It now follows from (3.3) and the fact that $p \to \pi^0$ that

$$\Sigma \pi_i^0 (\log \hat{\pi}_i - \log \pi_i^0) \to 0 \qquad \ldots \quad (3.4)$$

with probability 1. This implies that $\hat{\pi} \to \pi^0$ with probability 1, by an application of (2.8).

Let us now consider the case where the m.l. estimate may not exist, however close p is to π^0. In this case we can prove a more general result that any estimate π^* such that

$$L(\pi^*) \geqslant c \operatorname*{Sup}_{\pi \in A} L(\pi),$$

where L stands for the likelihood function defined in section 2, and $0 < c \leqslant 1$, is also consistent. This covers the case where a m.l. estimate may not exist for any p but

a near m.l. could be chosen. Such approximate m.l. estimates π^* were first considered by Wald (1949).

We have the general relationships

$$L(p) \geqslant L(\pi^*) \geqslant c \sup_{\pi \in A} L(\pi) \geqslant c\, L(\pi^0)$$

Taking logarithms and dividing by n the sample size,

$$\Sigma\, p_i \log p_i \geqslant \Sigma\, p_i \log \pi_i^* \geqslant \frac{\log c}{n} + \Sigma\, p_i \log \pi_i^0$$

As $n \to \infty$ and $p \to \pi^0$ we have with probability 1,

$$\Sigma\, p_i \log \pi_i^* \to \Sigma\, \pi_i^0 \log \pi_i^0$$

and hence by the preceding argument $\pi^* \to \pi^0$ with probability 1.

The argument employed here can also be used to establish consistency of estimates obtained by other methods of estimation, such as minimum Chi-square, minimum distance etc.

3.2. *Estimation of a parameter.* In general any problem of estimation can be viewed as that of finding the underlying distribution function as a whole, although this is usually done by first estimating a parameter occurring in the distribution function and substituting the estimate for the unknown value. A parameter is a function (or a functional) of the distribution function and for purposes of identification we shall assume that the functional correspondence is one to one. Stated as such the choice of a parameter (i.e. the functional) is to some extent arbitrary. While it was possible to show in section 3.1. that the m.l. estimate of the distribution function is strongly consistent under no regularity conditions it would be of interest to examine whether such a statement of consistency could be made about the m.l. estimate of θ, where $A = \{\pi(\theta)\}$ is a given parametric representation of the set A of admissible distributions.

The answer is in the affirmative provided the parametric representation satisfies the following natural assumption.

Assumption (A_1): The expression

$$I(\theta_0, \theta) = -\Sigma\, \pi_i(\theta_0) \log \frac{\pi_i(\theta)}{\pi_i(\theta_0)}, \quad (I \geqslant 0)$$

which provides an average amount of discrimination between the multinomial distribution defined by θ and the true one defined by θ_0 is bounded away from zero when $\|\theta - \theta_0\| > \delta$ for each $\delta > 0$.

To prove the assertion made we observe that for any given δ

$$\inf_{|\theta-\theta_0|>\delta} \Sigma\, \pi_i(\theta_0) \log \frac{\pi_i(\theta_0)}{\pi_i(\theta)} = \epsilon > 0$$

Let θ^* be a m.l. or approximate m.l. estimate. Then according to section 3.1. we have (with probability 1)

$$-\Sigma\, \pi_i(\theta_0)\, \log \frac{\pi_i(\theta^*)}{\pi_i(\theta_0)} < \epsilon$$

for all sufficiently large n. Hence $|\theta^*-\theta_0| \leqslant \delta$ for all sufficiently large n. Since δ is arbitrary, $\theta^* \to \theta_0$ with probability 1.

It is interesting to note that no other regularity conditions are needed. The assumption (A_1) seems to be a natural one to make since its violation would imply that two distributions with close values of the parameter could be better discriminated than those with a larger difference in the parameter which would indeed be an unnatural parametric representation.

Recently Kraft and Lecam (1956) gave an example of a multinomial distribution where the cell frequencies are regular functions of a parameter and for which m.l. estimate of the parameter is not consistent. I am indebted to Dr. R. R. Bahadur who pointed out that the assumption (A_1) introduced in this paper is not satisfied in their case. In fact, in their example, the discrimination between two distributions $\to$ zero as the difference in the corresponding parameter values $\to \infty$.

It may be noted that the corresponding assumption in the minimum Chi-square theory (see Rao, 1955) is that the Chi-square separator

$$\Sigma\, \frac{[\pi_i(\theta)-\pi_i(\theta_0)]^2}{\pi_i(\theta_0)}$$

be bounded away from zero whenever $|\theta-\theta_0| > \delta$ for any δ.

4. PROPERTIES OF THE M.L.E. ESTIMATE OF THE PARAMETER

We now make the following additional assumptions concerning $\pi_i(\theta)$.

Assumption (A_2): $\pi_1(\theta), \ldots, \pi_k(\theta)$ have continuous partial derivatives of the second order, at least in the neighbourhood of the true value θ_0.

Assumption (A_3): $\pi_j(\theta_0) \neq 0$ for each j, and $(d\pi_j/d\theta) \neq 0$ for at least one j.

As a consequence of this assumption $i(\theta_0)$, which is Fisher's information at θ_0 is $\neq 0$.

Assumption (A_4): $\pi_i(\theta) = \pi_i(\xi)$ for all i implies that $\theta = \xi$.

This provides one to one correspondence between the values of the parameter and the hypothetical cell frequencies

Theorem: *Under the assumptions* $(A_1), (A_2), (A_3)$ *and* (A_4) *there exists a neighbourhood of the true proportions* π^0, *say* $N(\pi^0)$ *and a positive* δ *such that* $p \in N(\pi^0)$ *implies,*

(i) *There exists one and only one root* $\hat{\theta}$ *of the likelihood equation* (2.4) *which differs from the true value* θ_0 *by less than* δ. *This root, as a function of the relative frequencies, is continuous at* π^0 *where it tends to* θ_0 *and is Fisher consistent.*

(ii) $\hat{\theta}$ *is Frechit. differentiable,*

(iii) $\hat{\theta}$ *is the unique m.l. estimate and is therefore in particular the m.l.e. estimate,*

Remark 1: The results (i) and (ii) are true under assumptions (A_2), (A_3) and (A_4) only. In addition the assumption (A_1) is needed to prove the stronger result (iii) that a m.l. estimate exists and is unique at least when p is close to the true value π^0. It may be noted that the assumption A_1 implies A_4.

Remark 2: It follows from the strong law of large numbers that when $\pi = \pi^0$ (i.e., $\theta = \theta_0$)

$$\text{prob.}\{p \in N(\pi^0) \text{ for all sufficiently large } n\} = 1.$$

Consequently the assertions (i), (ii) and (iii) of the Theorem are valid (with probability 1) for all sufficiently large n. A weaker statement but perhaps worthwhile to make is that the probability that the assertions (i)—(iii) are true $\to 1$ as $n \to \infty$.

Remark 3: As a consequence of results (i)–(iii) of the Theorem it follows that the m.l. estimate has an asymptotic normal distribution with the least asymptotic variance specified by (2.6) (see Rao. 1955).

Proof of (i): As $p \to \pi(\theta_0)$ and $\theta \to \theta_0$, with $\psi_i = d \log \pi_i / d\theta$, we have

$$\Sigma p_i \frac{d\psi_i}{d\theta} \to \Sigma \pi_i(\theta_0)\left\{\frac{1}{\pi_i(\theta_0)}\frac{d^2\pi_i}{d\theta_0^2} - \frac{1}{\pi_i^2(\theta_0)}\left(\frac{d\pi_i}{d\theta_0}\right)^2\right\}$$

$$= -\Sigma \frac{1}{\pi_i(\theta_0)}\left(\frac{d\pi_i}{d\theta_0}\right)^2 = -i(\theta_0) < 0.$$

This implies that there exists a neighbourhood $N(\pi_0)$ for p and positive δ such that $p \in N(\pi_0)$ and $|\theta - \theta_0| < \delta$ implies

$$\Sigma\, p_i \psi_i'(\theta) < 0. \qquad \ldots \quad (4.1)$$

More generally we have

$$\Sigma\, p_i \psi_i'(\theta_i) < 0 \qquad \ldots \quad (4.2)$$

whenever $p \in N(\pi^0)$ and $|\theta_i - \theta_0| < \delta$. The m.l. equation can be expressed as

$$0 = \Sigma\, p_i \psi_i(\theta_0) + (\theta - \theta_0)\, \Sigma p_i \psi_i'(\theta_i) \qquad \ldots \quad (4.3)$$

where θ_i is in (θ, θ_0). As $p \to \pi^0$, the expression on the right hand side $\to (\theta - \theta_0)\Sigma \pi_i (\theta_0)\psi_i'(\theta_i)$, which is < 0 for $(\theta - \theta_0) > 0$ and > 0 for $(\theta - \theta_0) < 0$ provided $|\theta - \theta_0| < \delta$. Hence when $p \in N(\pi^0)$, $\Sigma\, p_i \psi_i(\theta)$ is positive at $\theta_0 - \delta$ and negative at $\theta_0 + \delta$. Hence it vanishes for some $\hat{\theta}$ such that $|\hat{\theta} - \theta_0| < \delta$.

If there are two roots $\hat{\theta}_1$ and $\hat{\theta}_2$ both of which differ from θ_0 by less than δ then it follows, by Rolle's theorem, that

$$\Sigma\, p_i \psi_i'(\hat{\theta}_3) = 0$$

where $\hat{\theta}_3$ is in $(\hat{\theta}_1, \hat{\theta}_2)$ and hence $|\hat{\theta}_3 - \theta_0| < \delta$. This is impossible when $p \in N(\pi^0)$. Therefore, there is only one root close to θ_0 and the others are separated from θ_0 by a distance greater than δ.

It is now obvious that the root closer to θ_0 is a continuous function of the relative frequencies and $\to \theta_0$ as $p \to \pi^0$. Also it is Fisher consistent at least locally.

Proof of (ii): The expression (4.3) with $\hat{\theta}$ the continuous root is

$$0 = \Sigma p_i \psi_i(\hat{\theta}) = \Sigma p_i \psi_i(\theta_0) + (\hat{\theta} - \theta_0) \Sigma\, p_i \psi_i'(\theta_i)$$

where $$\theta_i \to \theta_0 \quad \text{as} \quad p_i \to \pi_i(\theta_0).$$

Hence
$$(\hat{\theta} - \theta_0) = -\Sigma\, p_i \psi(\theta_0) / \Sigma\, p_i \psi_i'(\theta_i)$$

$$= -\Sigma [p_i - \pi_i(\theta_0)] \psi_i(\theta_0) / \Sigma p_i \psi_i'(\theta).$$

Now consider,

$$\frac{1}{\max |p_i - \pi_i(\theta_0)|} \left| (\hat{\theta} - \theta_0) + \frac{1}{i(\theta_0)} \sum \frac{p_i - \pi_i(\theta_0)}{\pi_i(\theta_0)} \frac{d\pi_i}{d\theta_0} \right|$$

$$= \sum \frac{|p_i - \pi_i(\theta_0)|}{\max |p_i - \pi_i(\theta_0)|} \left| - \frac{\psi_i'(\theta_0)}{\Sigma p_i \psi_i'(\theta_i)} + \frac{1}{i(\theta_0)} \frac{1}{\pi_i(\theta_0)} \frac{d\pi_i}{d\theta_0} \right|$$

$$< \sum \left| \frac{\psi_i'(\theta_0)}{\Sigma p_i \psi_i'(\theta_i)} - \frac{1}{i(\theta_0)} \frac{1}{\pi_i(\theta_0)} \frac{d\pi_i}{d\theta_0} \right| \to 0$$

as $\max |p_i - \pi_i(\theta_0)| \to 0$. The function $(\hat{\theta} - \theta_0)$ is thus Frechet differentiable and can be approximated by the linear functional

$$-\frac{1}{i(\theta_0)} \sum \frac{p_i - \pi_i(\theta_0)}{\pi_i(\theta_0)} \frac{d\pi_i}{d\theta_0}$$

and hence $\sqrt{n}(\hat{\theta} - \theta_0)$ has asymptotic normal distribution with zero mean and variance $1/i(\theta_0)$ which is the minimum attainable under the conditions of lemma 1.

Proof of (iii): The proof consists of two parts

(A) For any θ such that $|\theta - \theta_0| < \delta$ and $p \in N(\pi^0)$, in which case $|\hat{\theta} - \theta_0| < \delta$,

$$\Sigma p_i \log \pi_i(\theta) = \Sigma p_i \log \pi_i(\hat{\theta}) + \frac{(\hat{\theta} - \theta)^2}{2} \Sigma p_i \psi_i(\theta') \leqslant \Sigma p_i \log \pi_i(\hat{\theta}) \quad \ldots \quad (4.4)$$

since by (4.1), $\Sigma p_i \psi_i'(\theta')$ is negative.

(B) For θ outside $|\theta - \theta_0| < \delta$, let S be the set of θ such that

$\Sigma(\pi_i(\theta_0) - \epsilon') \log \pi_i(\theta) < \Sigma(\pi_i(\theta_0) + \epsilon') \log \pi_i(\theta_0)$, where ϵ' is a small positive number such that the cube $\pi_i(\theta_0) \pm \epsilon'$ is within $N(\pi^0)$. Then, for θ in S and p in the cube,

$$\begin{aligned} \Sigma p_i \log \pi_i(\theta) &\leqslant \Sigma(\pi_i(\theta_0) - \epsilon') \log \pi_i(\theta) \\ &< \Sigma(\pi_i(\theta_0) + \epsilon') \log \pi_i(\theta_0) \leqslant \Sigma p_i \log \pi_i(\theta_0). \end{aligned}$$

Now consider θ outside $|\theta - \theta_0| < \delta$ and S. If $\epsilon' < \min\{\pi_i(\theta_0)\}$ the quantities $\log \pi_i(\theta)$ are bounded. Also

$$\Sigma p_i \log \frac{\pi_i(\theta)}{\pi_i(\theta_0)} = \Sigma \pi_i(\theta_0) \log \frac{\pi_i(\theta)}{\pi_i(\theta_0)} + \Sigma[p_i - \pi_i(\theta_0)] \log \frac{\pi_i(\theta)}{\pi_i(\theta_0)} \quad \ldots \quad (4.5)$$

$$\leqslant -c + \nu \quad \ldots \quad (4.6)$$

where by assumption (A_1) the first expression on the right side of (4.5) is not greater than say, $-c < 0$, and $\nu \to 0$ as $p \to \pi^0$. Hence (4.6) can be made < 0 by choosing $p \in N(\pi^0)$. Thus when $|\theta - \theta_0| \geq \delta$

$$\Sigma p_i \log \pi_i(\theta) < \Sigma p_i \log \pi_i(\theta_0) \leqslant \Sigma p_i \log \pi_i(\hat{\theta}) \quad \ldots \quad (4.7)$$

the latter inequality being true in virtue of what is proved in (A).

Since the equality in (4.4) is attained when and only when $\theta = \hat{\theta}$ for $|\theta - \theta_0| < \delta$ and outside $|\theta - \theta_0| < \delta$ the inequality (4.7) is strictly true, it follows that $\hat{\theta}$ is the unique m.l. estimate and in particular it is the unique m.l.e. estimate.

In conclusion, I wish to thank my colleague Dr. R. R. Bahadur for the useful discussions I had with him during the preparation of this paper.

REFERENCES

CRAMÉR, H. (1946): *Mathematical methods of statistics*, Princeton University Press.

HUZURBAZAR, V. S. (1948): The likelihood equation, consistency and the maximum of the likelihood function. *Ann. Eugen.*, **14**, 185.

KALLIANPUR, G. and RAO, C. R. (1955): On Fisher's lower bound to the asymptotic variance of an estimate. *Sankhya*, **15**, 331.

KRAFT, C. AND LECAM, L. (1956): A remark on the roots of the maximum likelihood equation. *Ann. Math. Stat.*, **27**, 1174.

KULLBACK, S. AND LEIBLER, R. A. (1951): On information and sufficiency. *Ann. Math. Stat.*, **22**, 79.

RAO, C. R. (1955): The theory of the method of minimum chi-square estimation. *Proc. Int. Stat. Conf.* 1955 (in press).

WALD, A. (1949): Note on the consistency of the maximum likelihood estimation. *Ann. Math. Stat.*, **20**, 595.

Paper received : May, 1957.