MAXIMUM LIKELIHOOD ESTIMATION FOR THE MULTINOMIAL DISTRIBUTION WITH INFINITE NUMBER OF CELLS

By C. RADHAKRISHNA RAO

Indian Statistical Institute, Calcutta

SUMMARY. Maximum likelihood (m.l.) estimate of the infinite multinomial distribution exists with probability 1 and is consistent under a simple condition on the cell probabilities. To prove the consistency of an m.l. estimate of a parameter it is necessary to assume that the parameter is a continuous function of the distribution. The existence of the m.l. estimates of parameters and their consistency is established under conditions slightly weaker than those assumed by earlier writers.

1. Introduction

The object of this paper is to extend the results obtained earlier (Rao, 1957) in the problem of maximum likelihood estimation for the finite multinomial distribution (f.m.d.) to the case of a multinomial distribution with infinite number of cells, which may be denoted by (i.m.d.).

As in the case of the f.m.d., the consistency of the maximum likelihood (m.l.) estimate of the i.m.d. is established first, and then the properties of m.l. equation estimates are studied.

In the parametric case, it has been assumed for simplicity in presentation that there is only one unknown parameter. But the method of proof given here holds good for any number of unknown parameters without any modification or without any extra results being proved.

2. Definitions and preliminary lemmas

The hypothetical frequencies in the infinite number of classes are represented by $(\pi_1, \pi_2, \ldots)$ and the observed relative frequencies by $(p_1, p_2, \ldots)$. All summations in the following definitions and in the rest of the paper are taken from 1 to ∞ unless otherwise stated.

Definition 1 : The m.l. estimate of the i.m.d. is a distribution $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots)$, if it exists, belonging to a given admissible class A of distributions π, for which the expression

$$\Sigma\, p_i \log \pi_i \text{ is a maximum} \qquad \ldots \quad (2.1)$$

when π is restricted to A.

Definition 1.2 : n^* is said to be an approximate or near m.l. estimate if

$$\Sigma\, p_i \log n_i^* \geq \frac{\log c}{n} + \sup_{\pi \in \Delta} \Sigma\, p_i \log \pi_i \qquad \ldots \quad (2.2)$$

where n is the sample size and $0 < c < 1$.

Definition 2 : The m.l. estimate of θ occurring in the specification of the hypothetical frequencies is a value of θ, if one exists in the admissible set of values of θ, for which

$$\Sigma\, p_i \log \pi_i(\theta) \text{ is a maximum.} \qquad \ldots \quad (2.3)$$

Near m.l. estimate of θ can be defined in the same way as in definition (2.2).

Definition 3 : The m.l. equation is

$$\Sigma\, \frac{p_i}{\pi_i} \frac{d\pi_i}{d\theta} = 0. \qquad \ldots \quad (2.4)$$

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 the likelihood when θ is restricted to the roots.

The m.l.e. estimate is not defined if the equation (2.4) has no roots.

In what follows, we shall denote a sequence of independent observations from the given multinomial population by $X_\infty = (x_1, x_2, \ldots)$.

Definition 5 : A sequence of functions $t_n = t_n(x_1, \ldots, x_n)$, $n = 1, 2, \ldots$ ad inf, is said to be a *consistent root of the likelihood equation* if, except for a set of sequences $X_\infty = (x_1, x_2, \ldots$ ad inf) of probability zero,

(i) t_n is well-defined for all sufficiently large n,

(ii) t_n is a root of the likelihood equation for all sufficiently large n, and

(iii) $t_n \to \theta_0$ as $n \to \infty$, where θ_0 is the actual parameter value.

In Definition 5, the phrase 'all sufficiently large n' means that for any given sequence X_∞ there is an m, depending on the sequence, such that the statement is true for all $n > m$.

Definition 6 : For each β in some index set let $\{t_{n\beta}\}$ be a sequence of estimates of θ. This family is said to be uniformly consistent if

$$u_n = \inf_\beta \{t_{n\beta}\} \text{ and } v_n = \sup_\beta \{t_{n\beta}\}$$

are themselves consistent estimates of θ.

Lemma 1 : *If* $\Sigma\, a_i$ *and* $\Sigma\, b_i$ *are two convergent sequences of non-negative numbers such that* $\Sigma\, a_i \geqslant \Sigma\, b_i$, *then*

$$\Sigma' a_i \log \frac{b_i}{a_i} \leqslant 0, \qquad \ldots \quad (2.5)$$

where the summation Σ' *is extended over non-zero values of* a_i. *The equality is attained when and only when* $a_i = b_i$ *for all* i.

The proof of Lemma 1 is similar to that given by Rao (1957) when the number of elements is finite, following a method used by Kullback and Liebler (1951).

Lemma 2 : *If the hypotheses of Lemma* 1 *are satisfied and if* $a_i \leqslant 1, b_i \leqslant 1$ *for all* i, *then*

$$2\,\Sigma' a_i \log \frac{a_i}{b_i} \geqslant \Sigma' a_i (a_i - b_i)^2. \qquad \ldots \quad (2.6)$$

To establish this, we note that for $x > 0$,

$$\log x = (x-1) - \frac{(x-1)^2}{2y^2} \text{ with } y \in (1, x).$$

It follows hence that

$$\Sigma' a_i \log \frac{b_i}{a_i} = \Sigma' b_i - \Sigma' a_i - \Sigma' a_i \frac{(b_i - a_i)^2}{2(z_i^2)} \qquad \ldots \quad (2.7)$$

where $z_i^2 \in (a_i^2, b_i^2)$ and hence $\leqslant 1$ for each i. Changing the signs on both sides of (2.7) and observing that $\Sigma' a_i - \Sigma' b_i$ is not less than zero, we obtain

$$2\,\Sigma' a_i \log \frac{a_i}{b_i} \geqslant \Sigma' a_i \frac{(b_i - a_i)^2}{z_i^2} \geqslant \Sigma' a_i (b_i - a_i)^2$$

as desired.

3. Consistency of m.l. estimates

Let A be the given admissible set of i.m. distributions, a typical member of which is denoted by $\pi = (\pi_1, \pi_2, \ldots)$. We shall make the following assumption about the true distribution $\pi^0 = (\pi_1^0, \pi_2^0, \ldots)$:

Assumption A_1 : $\Sigma\, \pi_i^0 \log \pi_i^0 > -\infty$.

The consistency of the m.l. estimate of the i.m.d. is established under this sole assumption. The consistency of m.l. estimates under the same assumption can be also proved by using the arguments of Wald (1949), as pointed out by Kiefer and Wolfowitz (1956). The proof given here, however, seems simpler and more direct.

It may be worthwhile to note here that Assumption A_1 is not necessary for consistency. For, if we allow A to be the set of all possible i.m. distributions, the assumption A_1 is clearly not satisfied but the m.l. estimate, which is then the observed distribution, is consistent. On the other hand, it is known that m.l. estimates can be inconsistent when A_1 is not true (Bahadur, 1958).

First let us suppose that the m.l. estimate exists and let us represent it by $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots)$. If $p = (p_1, p_2, \ldots)$ represents the observed cell proportions, then

$$\Sigma\, p_i \log\, \pi_i^0 \leqslant \Sigma\, p_i \log \hat{\pi}_i \leqslant \Sigma\, p_i \log\, p_i. \qquad \ldots \quad (3.1)$$

By the strong law of large numbers

$$\Sigma\, p_i \log \pi_i^0 \rightarrow \Sigma\, \pi_i^0 \log\, \pi_i^0, \qquad \ldots \quad (3.2)$$

where in what follows the symbol $\rightarrow$. when applied to random variables, indicates convergence with probability 1. For any given k and any sample sequence $X_\infty = (x_1, x_2, \ldots, x_n, \ldots)$ we have

$$\Sigma\, p_i \log\, p_i \leqslant \sum_1^k p_i \log\, p_i$$

for every n, and therefore

$$\limsup_{n\rightarrow\infty} \Sigma\, p_i \log\, p_i \leqslant \limsup_{n\rightarrow\infty} \sum_1^k p_i \log\, p_i.$$

Consequently, $\limsup_{n\rightarrow\infty} \Sigma\, p_i \log\, p_i \leqslant \sum_1^k \pi_i^0 \log\, \pi_i^0$ with probability 1. Since k is arbitrary, it follows that

$$\limsup_{n\rightarrow\infty} \Sigma\, p_i \log\, p_i \leqslant \Sigma\, \pi_i^0 \log\, \pi_i^0 \qquad \ldots \quad (3.3)$$

with probability 1.

On the other hand, $\Sigma\, p_i \log p_i \geqslant \Sigma\, p_i \log\, \pi_i^0$ by (3.1). Hence

$$\liminf_{n\rightarrow\infty} \Sigma\, p_i \log\, p_i \geqslant \liminf_{n\rightarrow\infty} \Sigma\, p_i \log\, \pi_i^0 = \Sigma \pi_i^0 \log\, \pi_i^0 \text{ with probability 1.} \quad \ldots (3.4)$$

From (3.3) and (3.4) it follows that

$$\Sigma\, p_i \log\, p_i \rightarrow \Sigma\, \pi_i^0 \log\, \pi_i^0$$

and using (3.2) we see that all the three terms in (3.1) tend to the same limit. Since this common limit is finite by assumption A_1, we have

$$\Sigma p_i \log \frac{\hat{\pi}}{p_i} \to 0.$$

Hence, by an application of Lemma 2,

$$\Sigma p_i (\hat{\pi}_i - p_i)^2 \to 0,$$

hence $p_i (\hat{\pi}_i - p_i)^2 \to 0$ for each i. This implies that $(\hat{\pi}_i - \pi_i^0) \to 0$ for each i for which $\pi_i^0 \neq 0$. Consequently, $\hat{\pi}_i \to \pi_i^0$ for every i. Thus the m.l. estimate of each of the individual cell proportions tends to the true value. According to a theorem of Scheffe (1947), this is equivalent to

$$\Sigma \mid \pi_i - \pi_i^0 \mid \to 0 \qquad \ldots \quad (3.5)$$

which exhibits the consistency of the m.l. estimate of the entire distribution in a strong form.

When the m.l. estimate does not exist but near m.l. estimates are considered, their consistency can be proved by a modification of the preceding argument (cf. Rao, 1957). The modified argument is presented in Theorem 1 below for the parametric case.

As observed earlier (Rao, 1957), the estimation of a parameter presents some difficulty unless it is suitably related to the admissible distributions. If $A = \{\pi(\theta)\}$ is a given parametric representation of the set A of admissible distributions, with θ a real parameter, and θ_0 is the true value of the parameter, and assumption A_1 holds, then the consistency of the m.l. estimate of the parameter follows immediately, provided only that the following continuity assumption holds:

Assumption A_2: If $\{\theta_m\}$ is a sequence of parameter values for $m = 0, 1, 2, \ldots$ such that for each $i = 1, 2, \ldots, \pi_i(\theta_m) \to \pi_i(\theta_0)$ as $m \to \infty$, then $\theta_m \to \theta_0$ as $m \to \infty$.

We shall now prove the following theorem.

Theorem 1: If A_1 *and* A_2 *hold, then the class of all approximate m.l. estimates (definitions* 1.2 *and* 2) *corresponding to a fixed* c, $0 < c \leqslant 1$, *is uniformly consistent (definition* 6).

To prove this result, we observe first that if $\theta^* = \theta_n(x_1, x_2, \ldots, x_n)$ is *any* approximate m.l. estimate we have

$$\Sigma p_i \log \pi_i(\theta_0) \leqslant \sup_{\theta} \Sigma p_i \log \pi_i(\theta)$$

$$\leqslant -\frac{\log c}{n} + \Sigma p_i \log \pi_i(\theta^*)$$

$$\leqslant -\frac{\log c}{n} + \Sigma p_i \log p_i.$$

Hence

$$\Sigma p_i \log \frac{\pi_i(\theta_0)}{p_i} + \frac{\log c}{n} \leqslant \Sigma p_i \log \frac{\pi_i(\theta^*)}{p_i} \leqslant 0. \qquad \ldots (3.6)$$

It has already been shown above that under A_1, $\Sigma p_i \log \frac{\pi_i(\theta_0)}{p_i} \to 0$. It follows hence by (3.6) that

$$\inf_{\theta^*} \Sigma p_i \log \frac{\pi_i(\theta^*)}{p_i} \to 0 \qquad \ldots (3.7)$$

An application of Lemma 2 to (3.7) shows that

$$\sup_{\theta^*} \Sigma p_i [\pi_i(\theta^*) - p_i]^2 \to 0. \qquad \ldots (3.8)$$

Since $\Sigma |p_i - \pi_i(\theta_0)| \to 0$, it follows easily from (3.8) that

$$\sup_{\theta^*} \Sigma \pi_i(\theta_0)[\pi_i(\theta^*) - \pi_i(\theta_0)]^2 \to 0. \qquad \ldots (3.9)$$

We observe next, in consequence of A_2, that for any $\delta > 0$,

$$\inf_{|\theta - \theta_0| \geqslant \delta} \Sigma \pi_i(\theta_0)[\pi_i(\theta) - \pi_i(\theta_0)]^2 > 0. \qquad \ldots (3.10)$$

It follows from (3.9) and (3.10) that with probability 1, for any $\delta > 0$, there exists an m (depending on c, δ, and the sequence $x_1, x_2 \ldots$) such that for all $n > m$ every approximate m.l. estimate θ^* is in the interval $(\theta_0 - \delta, \theta_0 + \delta)$. This establishes Theorem 1.

4. PROPERTIES OF M. L. EQUATION ESTIMATES

In this section, we state and prove some properties of the m.l. equation estimates and also establish the existence of m.l. estimates in the parametric case under certain additional assumptions.

Assumption A_3 : There exists a neighbourhood of θ_0 in which each $\pi_i(\theta)$ is differentiable with respect to θ.

Theorem 1. *Under assumptions* A_1, A_2 *and* A_3, *for all sufficiently large* n *with probability* 1,

(i) *m.l. estimates exist, and*

(ii) *every m.l. estimate is an m.l.e. estimate, and conversely.*

In other words, the class of m.l. estimates is non-empty and coincides with the class of m.l.e. estimates (definitions 2 and 4). It follows from Theorem 1 that this class is uniformly consistent.

To establish Theorem 2, choose and fix a $\delta > 0$ such that each $\pi_i(\theta)$ is differentiable (and therefore continuous) in $[\theta_0-\delta, \theta_0+\delta]$. Next, choose and fix a c, $0 < c < 1$, and consider the class of approximate m. l. estimates corresponding to c. It follows from Theorem 1 that with probability 1, for all sufficiently large n, every approximate m.l. estimate θ is in $(\theta_0-\delta, \theta_0+\delta)$. Consider a particular sequence $x_1, x_2, \ldots$ for which this is true, and a particular sufficiently large n. Let $L_n(\theta)$ denote the log likelihood at θ. If $\theta_1, \theta_2, \ldots$ is any sequence such that $L_n(\theta_k) \to \sup_\theta \{L_n(\theta)\}$, we shall have $L_n(\theta_k) \geqslant \frac{\log c}{n} + \sup_\theta \{L_n(\theta)\}$ for all sufficiently large k, since $0 < c < 1$, and hence $\theta_0-\delta < \theta_k < \theta_0+\delta$ for all sufficiently large k. It follows that the supremum of $L_n(\theta)$ equals the supremum with θ restricted to $[\theta_0-\delta, \theta_0+\delta]$, and is therefore attained in the latter interval. Thus m.l. estimates exist.

Now let $\hat{\theta}$ be any value of θ at which $L_n(\theta)$ is a maximum. Since then $\hat{\theta}$ is also an approximate m.l. estimate, we must have $\theta_0-\delta < \hat{\theta} < \theta_0+\delta$. Hence, as pointed out by Le Cam and Kraft (1956), the derivative of $L_n(\theta)$ must vanish at $\hat{\theta}$. Thus every m.l. estimate is a root of the likelihood equation, and therefore is an m.l.e. estimate. Conversely, every m.l.e. estimate must be an m.l. estimate, since the maximum of the likelihood with θ restricted to the roots of the likelihood equation has just been shown to equal the maximum likelihood with θ unrestricted. This completes the proof of Theorem 2.

The assumptions A_1, A_2 and A_3 under which the existence of consistent roots, and their identification as m.l. estimates has been established here seem to be considerably weaker than those assumed by earlier authors. Also the argument does not depend on the number of unknown parameters so that the additional difficulty in establishing the existence of the roots of the m.l. equation present in some of the earlier proofs when the number of unknown paramoters is more than one is avoided. However, proofs of other properties such as uniqueness of the m.l. estimate, and non-existence of two distinct consistent roots (Huzurbazar's theorem, 1948) seem to require additional assumptions. For example, if certain conditions on the second derivatives of $\pi_i(\theta)$ are satisfied, it is possible to deduce that, for some $r > 0$, with probability 1,

$$\max_{|\theta-\theta_0|<r} \{L_n''(\theta)\} < 0 \qquad \ldots \quad (4.1)$$

for all sufficiently large n, where dashes denote differentiation with respect to θ.

Let $\{t_{1n}\}$ and $\{t_{2n}\}$ be two consistent roots. If for some sequence and some n, $t_{1n} \neq t_{2n}$, it follows from $L_n'(t_{1n}) = L_n'(t_{2n}) = 0$ that there exists a

$$t_{3n} \in (t_{1n}, t_{2n}) \text{ with } L_n''(t_{3n}) = 0. \qquad \ldots \quad (4.2)$$

Since, with probability 1, the interval (t_{1n}, t_{2n}) is contained in (θ_0-r, θ_0+r) for all sufficiently large n, it follows from (4.1) and (4.2) that we must have $t_{1n} = t_{2n}$ for all sufficiently large n. Thus the consistent root of the likelihood equation is unique in the sense that any two consistent roots coincide for all sufficiently large n. It follows in particular that the m.l.e. estimate is unique, and hence so is the m.l. estimate.

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

REFERENCES

BAHADUR, R. R. (1958): Examples of inconsistency of maximum likelihood estimates. *Sankhyā*, 20, 207.

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

KIEFER, J. AND WOLFOWITZ, J. (1956): Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.*, 27, 887.

KNAPT, C. AND LE CAM, 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. (1957): Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139.

SCHEFFÉ, H. (1947): A useful convergence theorem for probability distributions. *Ann. Math. Stat.*, 18, 434.

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

Paper received : June, 1958.