ESTIMATION OF MULTINOMIAL PROBABILITIES.
ESTIMATION OF MULTINOMIAL PROBABILITIES

By

Khursheed Alam
Clemson University

REPORT N102
TECHNICAL REPORT #294
NOVEMBER, 1978

Research Supported in part by

THE OFFICE OF NAVAL RESEARCH

Task NR 042-271 Contract N00014-75-C-0451

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited
ESTIMATION OF MULTINOMIAL PROBABILITIES

Khursheed Alam*

Clemson University

ABSTRACT

This paper deals with the estimation of the parameters (cell probabilities) of a multinomial distribution. The maximum likelihood estimator (MLE) is known to be minimax and admissible with respect to a quadratic loss function. It is shown that the MLE is inadmissible with respect to a non-quadratic loss function. For the parameters of \( m \) multinomial distributions being estimated simultaneously and the loss being quadratic, an estimator is given which is shown to have smaller risk than the MLE for all but a small subset of the parameter space, when \( m \) is large.

Key words: Multinomial Distribution; Maximum Likelihood; Admissible Minimax Estimators.

AMS Classification: 62C15

*The author's work was supported by the Office of Naval Research under Contract N00014-75-C-0451.
1. Introduction and main results. Let \( x = (x_1, \ldots, x_k) \) be distributed according to a multinomial distribution \( M(x, p, n) \) with \( k \) cells, where \( p = (p_1, \ldots, p_k)' \), \( 0 \leq p_i \leq 1 \) (\( i = 1, \ldots, k \)), \( \sum_{i=1}^{k} p_i = 1 \) and \( \sum_{i=1}^{k} x_i = n \). For estimating \( p \) we consider below two loss functions, given by

\[
L(\hat{\delta}, p) = n \sum_{i=1}^{k} (\hat{\delta}_i - p_i)^2
\]

\[
L^*(\hat{\delta}, p) = n \sum_{i=1}^{k} (\hat{\delta}_i - p_i)^2 / p_i
\]

where \( \hat{\delta}_i = \hat{\delta}_i(x) \) and \( \hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_k) \) denotes an estimator of \( p \). Let \( \hat{\delta}^* = x/n \) denote the maximum likelihood estimator (MLE) whose risk is given by

\[
R(\hat{\delta}^*, p) = E L(\hat{\delta}^*, p)
\]

\[
= 1 - \sum_{i=1}^{k} p_i^2
\]

\[
R(\hat{\delta}^*, p) = E L^*(\hat{\delta}^*, p)
\]

\[
= k - 1.
\]

First we consider the loss (1.2). The covariance matrix of \( x \) is given by \( \Sigma = (\sigma_{ij}) \), where \( \sigma_{ij} = -np_i p_j \), \( i \neq j \) and \( \sigma_{ii} = np_i (1 - p_i) \). Clearly, \( \Sigma \) is a singular matrix. A generalized inverse of \( \Sigma \) is given by a diagonal matrix \( \Sigma^- \) whose \( i \)th diagonal element is equal to \( (np_i)^{-1} \), as it can be verified that \( \Sigma^{-1} \Sigma = \Sigma \). Hence, the loss function (1.2)
represents the Mahalanobis distance function \( n(\hat{\theta} - \theta)' \Sigma^{-1} (\hat{\theta} - \theta) \). It is also seen that Pearson's chi-square test statistic used for testing goodness of fit, represents the loss due to the MLE.

Olkin and Sobel (1978) have shown that the maximum likelihood estimator \( \hat{\theta}_0 \) is admissible for estimating \( \theta \) with respect to the loss function (1.2), among all estimators \( \hat{\theta} \) for which

\[
\sum_{i=1}^{k} \hat{\theta}_i = 1. \tag{1.5}
\]

Since the risk of \( \hat{\theta}_0 \) is constant, as given by (1.4), the MLE is also minimax among those estimators. If the condition (1.5) is removed then \( \hat{\theta}_0 \) is inadmissible. This is shown, as follows, by finding an estimator \( \hat{\theta}^* \) which dominates \( \hat{\theta}_0 \).

The Dirichlet distribution is a conjugate prior distribution for the parameter of a multinomial distribution. Suppose that \( \theta \) is distributed a priori according to the Dirichlet distribution \( D(\lambda) \), given by the density function

\[
f(\theta) = \frac{\Gamma(\lambda)}{\Gamma(\lambda)} (\theta_1 \cdots \theta_k)^{-1}, \quad 0 \leq \theta_i \leq 1. \tag{1.6}
\]

A Bayes estimator of \( \theta \) with respect to (1.6) and the loss function (1.2) is \( \hat{\theta}^* \), given by

\[
\hat{\theta}_i^* = \begin{cases} 
\frac{x_i - 1}{n + k - 1}, & x_i > 0, \\
0, & x_i = 0.
\end{cases} \tag{1.7}
\]
By direct computation we obtain the risk of \( \hat{\theta}^* \), given by

\[
R^*(\hat{\theta}^*, \mathcal{P}) = n(n+k\nu-1)^{-2} (\nu-1)^2 \sum_{i=1}^{k} \frac{1-(1-p_i)^n}{p_i} + 2nk(\nu-1)
\]

\[
+ n(n+k-1) - 2n(n+k\nu-1)^{-1} (\nu-1) \sum_{i=1}^{k} (1-(1-p_i)^n) + n
\]

\[
+ n.
\]

For \( \nu = 1 \), (1.8) reduces to

\[
R^*(\hat{\theta}^*, \mathcal{P}) = \frac{n(k-1)}{n+k-1}
\]

\[
< R^*(\hat{\theta}^0, \mathcal{P}).
\]

For \( \nu < 1 \) we have from (1.8)

\[
R^*(\hat{\theta}^*, \mathcal{P}) \leq n^2(n+k\nu-1)^{-2} [k(\nu-1)^2 + 2k(\nu-1) + (n+k-1)]
\]

\[
- 2\nu n^2(n+k\nu-1)^{-1} + n
\]

\[
= n^2(n+k\nu-1)^{-2}(k\nu^2+n-1) - 2n^2(n+k\nu-1)^{-1} + n.
\]

The quantity on the right hand side of the equality in (1.9) is equal to \( \frac{n(k-1)}{n+k-1} \) for \( \nu = 1 \). Therefore, \( R^*(\hat{\theta}^*, \mathcal{P}) < R^*(\hat{\theta}^0, \mathcal{P}) \) for \( 1 < \nu \leq 1 \), where \( \gamma \) is a positive number depending on the values of \( n \) and \( k \). Thus \( \hat{\theta}^* \) dominates \( \hat{\theta}^0 \) for certain values of \( \nu \). Note that \( \hat{\theta}^* \) does not satisfy the condition (1.5). Note also that \( \hat{\theta}^* \) is admissible, being a Bayes estimator.
Next, we consider the quadratic loss given by (1.1). Johnson (1971) and Alam (1978) have shown that the maximum likelihood estimator is admissible with respect to the quadratic loss. Steinhaus (1957) and Trybula (1958) have obtained minimax estimators for the more general loss function of the form \( \sum_{i=1}^{k} c_i (\hat{\xi}_i - \xi_i)^2 \), where \( c_i \) are constants and \( \sum_{i=1}^{k} \xi_i = 1 \). It should be observed that an estimator \( \hat{\delta} \) which does not satisfy the condition (1.5) is inadmissible with respect to the loss function given by (1.1), since the projection of \( \hat{\delta} \) on the hyperplane \( \sum_{i=1}^{k} \xi_i = 1 \) gives an estimator satisfying (1.5) for which the loss is smaller.

We consider below the problem of estimating simultaneously the parameters of \( m \geq 2 \) multinomial populations. Let \( \tau_1, \ldots, \tau_m \) denote the \( m \) populations, and let \( (p_{i1}, \ldots, p_{ik}) \) denote the vector of cell probabilities associated with \( \tau_i \), where \( \sum_{j=1}^{k} p_{ij} = 1 \). A sample of \( n \) observations is taken from each population. Let \( x_{ij} \) denote the sample frequency associated with the \( j \)th cell of \( \tau_i \). The loss is given by \( \sum_{i=1}^{m} \sum_{j=1}^{k} (\hat{\delta}_{ij} - p_{ij})^2 \), equal to \( n \) times the sum of squared errors, where \( \hat{\delta}_{ij} \) denotes an estimate of \( p_{ij} \), depending on the entire set of observations, even though the set of observations from \( \tau_i \) alone seems to be relevant. A sort of empirical Bayes estimator for the given problem is obtained, as follows.

Without confusing with the notation used above, we shall denote below the MLE and a Bayes estimator for the problem of simultaneous estimation by \( \hat{\xi}^o \) and \( \hat{\xi}^* \), respectively, and let \( \xi = (p_{11}, \ldots, p_{mk}) \). Let \( y = \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}^2, q_i = 1 - \sum_{j=1}^{k} p_{ij}^2 \).
and \( q = \sum_{i=1}^{m} q_i \). The MLE is given by \( \hat{\delta}_{ij} = x_{ij}/n \) and its risk is given by

\[
R(\hat{\delta}, p) = q.
\]

A Bayes estimator with respect to the Dirichlet prior (1.6) is given by

\[
\delta^*_{ij} = \frac{(x_{ij} + \nu)}{(n+k_1 \nu)}
\]

and by direct computation its risk is given by

\[
R(\delta^*, p) = n(n+k_1 \nu)^{-2} \left[ (n-k_1^2)q + \nu^2 mk(k-1) \right].
\]

A value of \( \nu \) minimizing (1.12) is given by

\[
\nu = q(m(k-1)-kq)^{-1}
\]

for which

\[
R(\delta^*, p) = \frac{nc}{n+k_1 \nu} < R(\hat{\delta}, p).
\]

Since \( \nu \) is unknown, the above inequality suggests that we might use \( \hat{\nu} \) for \( \nu \). But \( \hat{\nu} \) is also unknown, since \( q \) is unknown. But an estimate of \( q \) is given by \( (mn^2-y)/n(n-1) \) since its expected value is equal to \( q \). Substituting the estimate for \( q \) in (1.13) we get after simplification

\[
\hat{\nu} = \frac{mn^2-y}{ky-mn(n+k-1)}.
\]
Since the value of $\hat{\nu}$ given by (1.14) is negative for certain values of $y$, we make a minor modification and finally come up with a value of $\nu = \lambda$, say, given by

$$(1.15) \quad \lambda = \frac{mn^2 - y}{ky - mn^2}.$$ 

The empirical Bayes estimator is obtained from $\hat{\nu}$ by substituting $\lambda$ for $\nu$ in (1.11). We shall denote it by $\hat{\nu}$. It is shown below in Section 2 that $R(\hat{\nu}^*, p) < R(\hat{\nu}^0, p)$ for all values of $p$ for which

$$(1.16) \quad q > \frac{2n^4}{(n-1)^2} m^2, \quad m > n^2/(1-\delta)$$

where $0 < \delta < \frac{1}{2}$. That is, $\hat{\nu}^*$ has smaller risk than $\hat{\nu}^0$ for sufficiently large values of $m$, except for a set of values of $p$ approaching the null set as $m \to \infty$. Johnson (1971) has shown that there is no "Stein effect", that is, there is no estimator which dominates the MLE for a given value of $m$. This is essentially for the reason, as Johnson points out, that the risk of the MLE is small near the boundary of the parameter space, given by $q = 0$. A numerical comparison of the risk of $\hat{\nu}^*$ and $\hat{\nu}^0$ is given in Section 3.

The above results are summarized in the following theorems.
Theorem 1.1. The MLE is admissible with respect to (1.2) among all estimators satisfying the condition (1.5) but inadmissible among all estimators and is dominated by \( \hat{\psi}^* \), given by (1.7).

Theorem 1.2. \( R(\hat{\psi}^*, p) < R(\hat{\psi}^0, p) \) for all values of \( p \) for which (1.16) holds, where \( \hat{\psi}^* \) is given by (1.11) with the value of \( \psi \) given by (1.15).

2. Proof of Theorem 1.2. First we give a preliminary result which will be used in the sequel. Let

\[
(2.1) \quad z = (1 + k \lambda)^{-1} \frac{ky - mn^2}{(k-1)mn^2}.
\]

We have

\[
E(z) = 1 - \left( \frac{n-1}{n} \right) \left( \frac{k}{k-1} \right) \frac{q}{m}.
\]

Note that \( \frac{q}{m} \leq \frac{k-1}{k} \). As

\[
\text{Var} \left( \sum_{j=1}^{k} x_{ij} \right) = E \left( \sum_{j=1}^{k} x_{ij}^2 \right) - (E \sum_{j=1}^{k} x_{ij})^2
\]

\[
\leq n^2 E \left( \sum_{j=1}^{k} x_{ij}^2 \right) - (E \sum_{j=1}^{k} x_{ij})^2
\]

\[
\leq n^2 (n^2 - E \sum_{j=1}^{k} x_{ij}^2)
\]

\[
= n^3 (n-1) q_i
\]
we have

(2.2) \[ \text{Var}(z) \leq \left( \frac{n-1}{n} \right) \left( \frac{k}{k-1} \right)^2 \frac{q}{m^2}. \]

The risk of \( \hat{\psi}^{**} \) is given after simplification by

(2.3) \[ n^{-1} R(\hat{\psi}^{**}, \theta) = E \sum_{i=1}^{m} \sum_{j=1}^{k} (\hat{\psi}_{ij} - \theta_{ij})^2 \]

\[ = E \left( \hat{\psi}_{ij}^2 \right) - 2 \sum_{i=1}^{m} \sum_{j=1}^{k} \hat{\psi}_{ij} \theta_{ij} \]

\[ \leq E \left( \hat{\psi}_{ij}^2 \right) - 2 \sum_{i=1}^{m} \sum_{j=1}^{k} \hat{\psi}_{ij} \theta_{ij} \]

\[ \leq E \left( \hat{\psi}_{ij}^2 \right) - 2 \sum_{i=1}^{m} \sum_{j=1}^{k} \hat{\psi}_{ij} \theta_{ij} \]

\[ = E \left[ y + 2mn \frac{\lambda + mk}{2} (n+k) \right] - 2 \left( \frac{m}{n} \right) \left( \frac{k}{k-1} \right) \frac{q}{m} \]

\[ + (q-m) \frac{(k-1)}{2n} \frac{z^2}{1+(n-1)z} - 1 \]

\[ + 2n (m \frac{(k-1)}{k}) \frac{z}{1+(n-1)z}. \]
\[ \frac{1}{m(n^{-1})} \left( \sum_{i=1}^{k} \frac{1}{F(x)} \right) \]

where \( F \) denotes the cumulative distribution function of \( z \). The inequality in (2.3) is derived from the fact that

\[ E_{j=1}^{k} P_{j} x_{ij} \] is nondecreasing in \( \sum_{j=1}^{k} x_{ij}^{2} \). This is shown by proving it for \( k = 2 \) by direct computation of the expected value, and recursively for \( k > 2 \). The last step in (2.3) is derived through integration by parts. From (1.10) and (2.3) we get

(2.4) \[ (mn)^{-1} \left( R(z^{*}, P) - R(z^{e}, P) \right) \leq 2 \left( \frac{n}{n^{-1}} \right) \left( \frac{k-1}{k} \right) \int_{0}^{1} \frac{x F(x) dx}{(1+(n-1)x)^{3}} \]

\[ + \left( \frac{n}{n^{-1}} \right)^{2} \left( \frac{k-1}{k} \right) + 2n \left( \frac{k-1}{k} - \frac{q}{m} \right) \int_{0}^{1} \frac{F(x) dx}{(1+(n-1)x)^{2}} - \frac{q}{m} \left( \frac{1}{n} + \frac{1}{n^{-1}} \right) \]

\[ \leq 2 \left( \frac{n}{n^{-1}} \right) \left( \frac{k-1}{k} \right) + \left( \frac{n}{n^{-1}} \right)^{2} \left( \frac{k-1}{k} \right) + 2n \left( \frac{k-1}{k} - \frac{q}{m} \right) \int_{0}^{1} \frac{F(x) dx}{(1+(n-1)x)^{2}} \]

\[ - \frac{q}{m} \left( \frac{1}{n} + \frac{1}{n^{-1}} \right) \]
Let \( 0 < \varepsilon < \frac{1}{n} \). We have

\[
\int_0^1 \frac{F(x)dx}{(1+(n-1)x)^2} - \frac{q}{m^2} \frac{1}{n} + \frac{1}{n-1}.
\]

The first integral on the right hand side of (2.5) is majorized by

\[
\int_0^1 \frac{F(x)dx}{(1+(n-1)x)^2} = \int_0^{E(z)-\varepsilon} + \int_{E(z)-\varepsilon}^{E(z)} + \int_{E(z)}^{1} \frac{F(x)dx}{(1+(n-1)x)^2}.
\]

by Chebychev's inequality

\[
\leq \frac{\text{Var}(Z)}{\varepsilon^2} \leq \frac{(n-1)}{n} \frac{k}{(k-1)}^2 \frac{q}{m^2} \frac{1}{n} \text{ by (2.2).}
\]

The second integral is majorized by

\[
(1+(n-1)(E(z)-\varepsilon))^{-2} = \frac{\varepsilon^2}{n^2} (1 - \frac{n-1}{n})^2 \frac{k}{(k-1)}^2 \frac{q}{m^2} - (\frac{n-1}{n} \varepsilon)^{-2}.
\]

The third integral is majorized by

\[
\int_{E(z)}^{1} (1+(n-1)x)^{-2}dx = \frac{1-E(z)}{1+(n-1)E(z)}
\]

\[
= \frac{q}{mn} \frac{(n-1)}{n} \frac{k}{(k-1)} \frac{1}{(1 - \frac{n-1}{n})^2 \frac{k}{(k-1)}^2 \frac{q}{m^2}^{-1}}.
\]
Using (2.5) through (2.8) in (2.4) we get

\[(2.9) \quad (mn)^{-1}(R(\hat{\xi}^*, \rho) - R(\hat{\xi}^0, \rho)) \leq n\left(\frac{n}{n-1}\right)\left(\frac{k-1}{k}\right) \left(\frac{2n-1}{n-1}\right)\]

\[\times 2\left(\frac{n-1}{n}\right)\left(\frac{k}{k-1}\right) \left(\frac{q}{m}\right)^2 \left(\frac{(n-1)}{n} \left(k - \frac{1}{n}\right) \left(\frac{k-1}{k}\right)^2 \frac{q}{m} \epsilon^2 \right) + \frac{\epsilon}{n^2} \left(1 - \left(\frac{n-1}{n}\right)^2\right)\]

\[\left(\frac{k}{k-1}\right) \left(\frac{q}{m}\right)^{-1}\]

\[\leq n\left(\frac{n}{n-1}\right)\left(\frac{k-1}{k}\right) \frac{2n-1}{n-1} \left(\frac{n-1}{n} \left(k-1\right) \left(\frac{k}{k-1}\right)^2 \frac{q}{m} \epsilon^2 \right) + \frac{\epsilon}{n^2}\]

\[\left(1 - \left(\frac{n-1}{n}\right)^2 \left(k - \frac{1}{n}\right)^2 \left(\frac{q}{m}\right)^2\right)\]

since \(\epsilon \leq \frac{1}{n}\) and \(\left(\frac{k}{k-1}\right) \left(\frac{q}{m}\right) \leq 1\).

Let \(Q\) denote the quantity inside the square bracket on the right side of the second inequality in (2.9). Suppose that

\[(2.10) \quad \frac{q}{m} > \gamma m^{-\delta}\]

where \(\gamma = \frac{2n^4}{(n-1)^2}\) and \(0 < \delta \leq \frac{1}{5}\). Putting \(\epsilon = m^{-\frac{1}{2}}\) in \(Q\) we get

\[Q \leq \frac{2n-1}{n-1} \left(\frac{q}{m}\right)^{\frac{1-1}{2}} - \left(\frac{n-1}{n} \left(k - \frac{1}{n}\right) \left(\frac{n-1}{n} \left(\frac{k}{k-1}\right)^2 \frac{q}{m} \epsilon^2 \right) - \frac{2n-1}{n-1}\right) \left(\frac{q}{m}\right)\]

\[\leq \frac{2n-1}{n-1} \left(\frac{q}{m}\right) \left(\frac{n-1}{n} \left(k - \frac{1}{n}\right) \left(\frac{n-1}{n} \left(\frac{k}{k-1}\right)^2 \frac{q}{m} \epsilon^2 \right) - \frac{2n-1}{n-1}\right)\]

\[\leq 0\] for \(n \geq 2\).
Therefore

\[ R(\frac{1}{2}^*, p) - R(\frac{1}{2}^0, p) < 0 \]

for all values of \( p \), satisfying the inequality (1.16).

3. **Numerical comparison of the risk of \( \frac{1}{2}^* \) and \( \frac{1}{2}^0 \).** By Theorem 2.2 if \( m \) is large then \( R(\frac{1}{2}^*, p) < R(\frac{1}{2}^0, p) \) for all but a small subset of the values of \( p \) near the boundary of the parameter space, given by \( q = 0 \). In many practical situations requiring simultaneous estimation of the parameters of several multinomial populations the value of \( q \) is a priori bounded away from zero so that the inequality holds for moderately large values of \( m \). Therefore, \( \frac{1}{2}^* \) should be ordinarily preferred to \( \frac{1}{2}^0 \). Let

\[ \zeta(\frac{1}{2}^0, \frac{1}{2}^*) = \frac{R(\frac{1}{2}^0, p) - R(\frac{1}{2}^*, p)}{R(\frac{1}{2}^0, p)} \]

denote the relative saving in the risk of \( \frac{1}{2}^* \). The following table gives for illustration 8 sets of values of \( \zeta(\frac{1}{2}^0, \frac{1}{2}^*) \), computed by Monte Carlo method for \( m = 10, n = 10, 20 \) and \( k = 2, 3, 4 \) with the values of \( p \) being chosen randomly. It is seen from the table that there is considerable saving in the risk due to \( \frac{1}{2}^* \).
TABLE I - Values of \( p(\tilde{\theta}, \tilde{\delta}) \)

<table>
<thead>
<tr>
<th></th>
<th>n = 10</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 2</td>
<td>.193</td>
<td>.128</td>
<td>.193</td>
<td>.170</td>
<td>.045</td>
<td>.166</td>
<td>.686</td>
</tr>
<tr>
<td>3</td>
<td>.257</td>
<td>.187</td>
<td>.243</td>
<td>.236</td>
<td>.176</td>
<td>.390</td>
<td>.203</td>
</tr>
<tr>
<td>4</td>
<td>.298</td>
<td>.210</td>
<td>.251</td>
<td>.211</td>
<td>.119</td>
<td>.168</td>
<td>.481</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>n = 20</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.000</td>
<td>.099</td>
<td>.232</td>
<td>.042</td>
<td>-.017</td>
<td>.083</td>
<td>-.041</td>
</tr>
<tr>
<td>3</td>
<td>.132</td>
<td>.024</td>
<td>.031</td>
<td>.144</td>
<td>.056</td>
<td>.263</td>
<td>.013</td>
</tr>
<tr>
<td>4</td>
<td>.165</td>
<td>.215</td>
<td>.180</td>
<td>.118</td>
<td>.069</td>
<td>.205</td>
<td>.372</td>
</tr>
</tbody>
</table>
References


Estimation of Multinomial Probabilities

Multinomial Distribution; Maximum Likelihood; Admissible Minimax Estimators.

This paper deals with the estimation of the parameters (cell probabilities) of a multinomial distribution. The maximum likelihood estimator (MLE) is known to be minimax and admissible with respect to a quadratic loss function. It is shown that the MLE is admissible with respect to a non-quadratic loss function. For the parameters of m multinomial distributions being estimated simultaneously and the loss being quadratic, an estimator is given which is shown to have smaller risk than the MLE for all but a small subset of the parameter space, when m is large.