DEPARTMENT OF MATHEMATICAL SCIENCES
CLEMSON UNIVERSITY
Clemson, South Carolina
AN EMPIRICAL BAYES ESTIMATE
OF MULTINOMIAL PROBABILITIES

*Khursheed Alam
Clemson University

&

Amitava Mitra
Auburn University

N-137

Technical Report #381
Department of Mathematical Sciences
Clemson University

February, 1982

*This work was supported by the U. S. Office of Naval Research under Contract N00014-75-C-0451.
An Empirical Bayes Estimate of Multinomial Probabilities

Khursheed Alam* & Amitava Mitra

Clemson University & Auburn University

*This work was supported by the U.S. Office of Naval Research under contract N00014-75-C-0451.
An Empirical Bayes Estimate of Multinomial Probabilities

Khursheed Alam & Amitava Mitra
Clemson University & Auburn University

ABSTRACT

This paper deals with the problem of estimating simultaneously the parameters (cell probabilities) of \( m > 2 \) independent multinomial distributions, with respect to a quadratic loss function. An empirical Bayes estimator is given which is shown to have smaller risk than the maximum likelihood estimator for sufficiently large values of \( m \), at all points of the parameter space outside its boundary.

Key words: Multinomial distribution; Maximum likelihood; Empirical Bayes; Quadratic loss

AMS Classification: 62F10

This work was supported by the U.S. Office of Naval Research under contract N00014-75-C-0451.
1. INTRODUCTION

The multinomial distribution arises in various statistical analyses. In this paper we consider the problem of estimating simultaneously the parameters (cell probabilities) of \( m \geq 2 \) independent multinomial distributions. A new estimator is obtained from empirical Bayes considerations, where the estimate of the parameters of one distribution depends on the observations taken from the other distributions. We compare the new estimator with the maximum likelihood estimator (MLE) and show that for large values of \( m \) the risk of the new estimator is smaller than the risk of the MLE at all points excluding a small subset of the parameter space.

The problem of estimating the parameters of a multinomial distribution has been considered by various authors. Alam (1979) and Johnson (1971) have shown that for \( m = 1 \) the MLE is admissible with respect to the sum of squared errors as the loss function. The estimation of multinomial probabilities has been considered from a decision theoretic point of view by Steinhäus (1957), Trybula (1958) and Rutkowska (1977). In a recent paper Olkin and Sobel (1979) have discussed the problem of estimating the parameters of \( m \geq 2 \) binomial distributions.

In Section 2 we introduce the new empirical Bayes estimator. In Section 3 we compare the risk of the empirical Bayes estimator with the MLE.

2. EMPIRICAL BAYES ESTIMATOR

Let \( \mathbf{x}_i = (x_{i1}, \ldots, x_{iK}) \), \( i = 1, \ldots, m \) denote independent observations from \( m \) multinomial distributions, each with \( K \geq 2 \) cells, where \( \sum_{j=1}^{K} x_{ij} = n \). Let \( \mathbf{p}_i = (p_{i1}, \ldots, p_{iK}) \) denote the vector of cell probabilities associated with the \( i \)th distribution, where \( p_{ij} > 0 \), \( j = 1, \ldots, K \) and \( \sum_{j=1}^{K} p_{ij} = 1 \).
For estimating the cell probabilities simultaneously, let the loss function be given by

$$L(\delta, P) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{K} (\delta_{ij} - p_{ij})^2$$

where $\delta_{ij}$ denotes the estimate of $p_{ij}$, $\delta = (\delta_{ij})$ and $P = (p_{ij})$.

First we obtain a Bayes estimator of $P$, as follows. Suppose that the parameters of the $m$ distributions are distributed a priori, independently and identically according to the Dirichlet distribution, given by the density function

$$f(\theta_1, \ldots, \theta_K) = \frac{\Gamma(K)}{\Gamma(\lambda)} \theta_1^{\alpha-1} \cdots \theta_K^{\alpha-1}$$

where $\lambda$ is a positive number, $0 < \theta_j < 1$, $j = 1, \ldots, K$ and $\sum_{j=1}^{K} \theta_j = 1$.

A Bayes estimator of $P$ with respect to the Dirichlet prior distribution is given by $\delta^* = (\delta_{ij}^*)$, where

$$\delta_{ij}^* = \frac{x_{ij} + \lambda}{(m + K\lambda)}.$$  \hspace{1cm} (2.1)

Let $q_i = 1 - \sum_{j=1}^{K} q_{ij}^2$, $q = \frac{1}{m} \sum_{i=1}^{m} q_i$ and $y = \sum_{i=1}^{m} \sum_{j=1}^{K} x_{ij}^2$. The risk of $\delta^*$ is given by

$$R(\delta^*) = \frac{[(n-\lambda^2K^2)q + K(K-1)\lambda^2]}{(n+K\lambda)^2}.$$  \hspace{1cm} (2.2)

The value of $\lambda$ minimizing $R(\delta^*)$ is equal to $\lambda_0$, given by

$$\lambda_0 = q(K-1-Kq)^{-1}.$$  \hspace{1cm} (2.3)

Let $R^*$ denote the value of $R(\delta^*)$ for $\lambda = \lambda_0$. We have

$$R^* = \frac{q}{n+K\lambda_0} < \frac{q}{n} = R(\delta_0^*)$$  \hspace{1cm} (2.4)
where \( \hat{\delta} \) denotes the MLE, given by \( \hat{\delta} = \frac{x_{ij}}{n} \). An unbiased estimator of \( q \) is given by \( \hat{q} = \frac{(mn^2 - y)}{mn(n-1)} \). Substituting \( \hat{q} \) for \( q \) in (2.3) we get an estimated value of \( \lambda \), given by

\[
\hat{\lambda} = \hat{q}(K-1-K\hat{q})^{-1}
\]

(2.5)

= \infty \text{ for } \hat{q} = (K-1)/K.

Substituting \( \hat{\lambda} \) for \( \lambda \) in (2.1), we obtain an empirical Bayes estimator \( \hat{\delta} \) of \( \Psi \), given by

\[
\hat{\delta}_{ij} = \frac{(x_{ij} + \hat{\lambda})}{(n+K\hat{\lambda})}
\]

(2.6)

= 1/K \text{ for } \hat{\lambda} = \infty.

The quantity \( q \) is a measure of diversity of the \( i \)th multinomial distribution. It is equal to zero when the probability mass of the distribution is concentrated in a single cell. It achieves its maximum value equal to \( (K-1)/K \) when the probability mass is equally distributed among the \( K \) cells of the distribution. The value of \( q \) represents the average diversity of the \( m \) multinomial distributions. In the following section we show that the empirical Bayes estimator \( \hat{\delta} \) has smaller risk than the MLE for sufficiently large values of \( mq^2 \). Therefore, in practice when the parameters of a large number of multinomial distributions are simultaneously estimated for which the average diversity is unlikely to be very small, the empirical Bayes estimator should be preferred to the MLE.

The preceding observation is supported by the numerical results shown in Table 1 below. The table gives the empirical values of \( R(\hat{\delta}) \), \( R(\hat{\delta}^0) \) and \( mq^2 \) for certain values of \( m, n, K \) and randomly chosen values of \( \Psi \). The value of \( R(\hat{\delta}) \) and \( R(\hat{\delta}^0) \) for any fixed value of \( \Psi \) was obtained from simulation by the Monte Carlo method. It is seen from the table that \( R(\hat{\delta}) < R(\hat{\delta}^0) \) in all
the observed cases except one when \( m = 5, n = 10, K = 2 \) for which \( R(\hat{\delta}) = 0.0247 \) and \( R(\delta^0) = 0.0241 \). This case corresponds to the smallest value of \( mq^2 = 0.2856 \) given in the table.

3. RISK OF \( \hat{\delta} \)

From (2.2) it is seen that \( R(\delta^*) \) first decreases then increases as \( \lambda \) varies from 0 to \( \infty \). It is equal to \( \frac{a}{n} \) and \( \frac{K-1}{K} - q \) for \( \lambda = 0 \) and \( \infty \), respectively. Hence, \( R(\delta^*) < R(\delta^0) \) for all positive values of \( \lambda \) if \( \frac{a}{n} > \frac{K-1}{K} - q \). Therefore, we shall assume that

\[
q \leq \frac{n(K-1)}{K(n+1)} . \tag{3.1}
\]

The value of \( R(\hat{\delta}) \) for large \( m \) is given as follows: From (2.3) and (2.5) we get

\[
\hat{\lambda} - \lambda_0 = \frac{(K-1)(\hat{q}-q)}{(K-1-Kq)(K-1-Kq)} . \tag{3.2}
\]

Now, \( \hat{E}q = q \) and \( \text{Var}(\hat{q}) \leq \frac{mq}{m(n-1)} \) by Lemma 3.1 below. Therefore, for large \( m \)

\[
\hat{E}\hat{\lambda} = \lambda_0 + \frac{K(K-1)}{(K-1-Kq)^2} \text{Var}(\hat{q}) + o\left(\frac{1}{m}\right) \quad \text{and} \tag{3.3}
\]

\[
\text{Var}(\hat{\lambda} - \lambda_0)^2 = \left[\frac{(K-1)}{(K-1-Kq)^4}\right] \text{Var}(\hat{q}) + o\left(\frac{1}{m}\right). \tag{3.4}
\]

The value of \( R(\hat{\delta}) \) as obtained from (2.6) by direct computation, is given after simplification by

\[
R(\hat{\delta}) = R^* + nk(n+K\lambda_0)^{-2}E(n+K\hat{\lambda})^{-2}(n(K-1)-K(n-1)\hat{q})(\hat{\lambda} - \lambda_0)^2 \tag{3.5}
\]

\[
+ 2nk(n+K\lambda_0)^{-1}E(n+K\hat{\lambda})^{-1}(\hat{\lambda} - \lambda_0)\left\{ \frac{(K-1)\lambda_0+(n-1)\hat{q}-(n+K\lambda_0)q}{n+K\lambda_0} + Z \right\}
\]

where \( Z = \frac{1}{m} \sum_{j=1}^{m} \sum_{j=1}^{K} \pi_{ij} (x_{ij} - np_{ij}) \). Now \( EZ = 0 \) and
\[
\text{Var}(Z) = nm^{-2} \left( \sum_{i=1}^{m} \sum_{j=1}^{K} P_{ij}^3 - \sum_{i=1}^{m} (1-q_i)^2 \right) \tag{3.6}
\]

\[
\leq nm^{-2} \sum_{i=1}^{m} q_i (1-q_i)
\]

\[
\leq \frac{n}{m} q.
\]

Using (3.3), (3.4), and (3.6) in (3.5) we get

\[
R(\hat{\delta}) = R^* + \frac{cq}{m} + o\left(\frac{1}{m}\right)
\]

\[
= q \left( \frac{1}{n+K\lambda} \right) + \frac{c}{m} + o\left(\frac{1}{m}\right)
\]

where \(c = c(P)\) is a number less than \(K^2n^2\) in absolute value. Then

\[
R(\delta^0) - R(\hat{\delta}) = q \left( \frac{1}{n} - \frac{1}{n+K\lambda} \right) - \frac{c}{m} + o\left(\frac{1}{m}\right)
\]

\[
= q \left( \frac{K\lambda}{n(n+K\lambda)} \right) - \frac{c}{m} + o\left(\frac{1}{m}\right)
\]

\[
= q^2 \left[ \frac{K(n+K\lambda)}{n} \right]^{-1} (K-1-Kq)^{-1} - \frac{c}{mq} + o\left(\frac{1}{m}\right).
\]

By virtue of the inequality (3.1) the quantity inside the square bracket is minorized by

\[
\frac{K}{2n^2(K-1)} - \frac{c}{mq}.
\]

Therefore, we have the following result.

**Theorem 3.1.** The empirical Bayes estimator \(\hat{\delta}\) has smaller risk than the maximum likelihood estimator \(\delta^0\) for sufficiently large values of \(mq^2\).

**Remark.** It has been mentioned in the introduction that the MLE is admissible with respect to the quadratic loss function when \(m = 1\). It is also noted by Johnson (1971) that the admissibility of the MLE is essentially for the reason that its risk is small near the boundary of the parameter.
space, given by \( q = 0 \). Therefore, the condition of Theorem 3.1 that the value of \( q \) should be sufficiently large for the risk of \( \hat{\delta} \) to be smaller than the risk of \( \delta^0 \) is consistent with that observation. Johnson has also noted that there is no Stein-effect for \( m > 1 \), that is, the MLE cannot be improved upon by letting the estimate of parameters of one distribution depend on the observations taken from the other distributions. Therefore, we do not expect that \( \hat{\delta} \) has a uniformly smaller risk than \( \delta^0 \).

Finally, we give the lemma which was used in the proof of Theorem 3.1.

**Lemma 3.1.** \( \text{Var}(q) \leq \frac{pq}{m(n-1)} \)

**Proof:** We have

\[
\text{Var}\left( \sum_{j=1}^{K} x_{ij}^2 \right) = E\left( \sum_{j=1}^{K} x_{ij}^2 \right)^2 - (E \sum_{j=1}^{K} x_{ij}^2)^2 \\
\leq (E \sum_{j=1}^{K} x_{ij}^2)(n^2 - E \sum_{j=1}^{K} x_{ij}^2) \\
\leq n^2(n^2 - E \sum_{j=1}^{K} x_{ij}^2) \\
= n^3(n-1)q_i.
\]

Summing the above inequality for \( i = 1, \ldots, m \) we have

\[
\text{Var}(y) \leq mn^3(n-1)q.
\]

Therefore

\[
\text{Var}(\hat{q}) \leq \frac{pq}{m(n-1)}.
\]
References


### Table 1. Values of $R(\delta)$, $R(\delta^0)$ and $mq^2$

<table>
<thead>
<tr>
<th>m</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>$R(\delta)$</td>
<td>$R(\delta^0)$</td>
<td>$mq^2$</td>
<td>$R(\delta)$</td>
</tr>
<tr>
<td>5</td>
<td>0.0661</td>
<td>0.0669</td>
<td>0.5599</td>
<td>0.0582</td>
</tr>
<tr>
<td>10</td>
<td>0.0247</td>
<td>0.0241</td>
<td>0.2856</td>
<td>0.0293</td>
</tr>
<tr>
<td>15</td>
<td>0.0208</td>
<td>0.0217</td>
<td>0.5381</td>
<td>0.0229</td>
</tr>
<tr>
<td>K = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0797</td>
<td>0.0883</td>
<td>0.8782</td>
<td>0.0694</td>
</tr>
<tr>
<td>10</td>
<td>0.0323</td>
<td>0.0338</td>
<td>0.5828</td>
<td>0.0339</td>
</tr>
<tr>
<td>15</td>
<td>0.0272</td>
<td>0.0289</td>
<td>0.9254</td>
<td>0.0237</td>
</tr>
<tr>
<td>K = 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0849</td>
<td>0.1111</td>
<td>1.5426</td>
<td>0.0773</td>
</tr>
<tr>
<td>10</td>
<td>0.0480</td>
<td>0.0572</td>
<td>1.6411</td>
<td>0.0429</td>
</tr>
<tr>
<td>15</td>
<td>0.0325</td>
<td>0.0369</td>
<td>1.5235</td>
<td>0.0304</td>
</tr>
<tr>
<td>n</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>R(δ)</td>
<td>R(δ^0)</td>
<td>m^2</td>
<td>R(δ)</td>
</tr>
<tr>
<td>5</td>
<td>0.0502</td>
<td>0.0757</td>
<td>4.3331</td>
<td>0.0492</td>
</tr>
<tr>
<td>10</td>
<td>0.0293</td>
<td>0.0346</td>
<td>3.5217</td>
<td>0.0289</td>
</tr>
<tr>
<td>15</td>
<td>0.0180</td>
<td>0.0195</td>
<td>2.6324</td>
<td>0.0190</td>
</tr>
<tr>
<td>5</td>
<td>0.0626</td>
<td>0.0781</td>
<td>4.4657</td>
<td>0.0644</td>
</tr>
<tr>
<td>10</td>
<td>0.0369</td>
<td>0.0431</td>
<td>5.4888</td>
<td>0.0385</td>
</tr>
<tr>
<td>15</td>
<td>0.0281</td>
<td>0.0331</td>
<td>7.3489</td>
<td>0.0277</td>
</tr>
<tr>
<td>5</td>
<td>0.0721</td>
<td>0.0929</td>
<td>6.3544</td>
<td>0.0735</td>
</tr>
<tr>
<td>10</td>
<td>0.0431</td>
<td>0.0530</td>
<td>8.4576</td>
<td>0.0426</td>
</tr>
<tr>
<td>15</td>
<td>0.0273</td>
<td>0.0298</td>
<td>6.1387</td>
<td>0.0304</td>
</tr>
</tbody>
</table>
This paper deals with the problem of estimating simultaneously the parameters (cell probabilities) of $m > 2$ independent multinomial distributions, with respect to a quadratic loss function. An empirical Bayes estimator is given which is shown to have smaller risk than the maximum likelihood estimator for sufficiently large value of $m$, at all points of the parameter space outside its boundary.