Bayesian Fixed Sample Size Procedure for Selecting the Least Probable Event in Multinomal Distribution

By

Saad A.Madhi and Kawther F .Hamza

Dept .of Mathematics College of Education.

ABSTRACT

In this paper, a fixed sample size prosedure for selecting the least probable event (i.e ,the cell with smallest probability) in multinomial distribution is given .Bayesian decision –theoretic approach is used to construct this procedure.Bayes risks of taking decisions under linear losses and Dirichelet priors are derived .

1. Introduction

Consider the multinomial distribution with k cells and unknown probabilities of an observation in the ith cell p_i ,(i=1,2,….,k),where $\sum_{i=1}^{k} p_i = 1$. It is required to find the cell with the smallest (least) probability (best cell in this sense).There are many practical situations where a solution to this problem is required . For example ,we have a sample of blood from each of a large number of populations in a certain city ,and each person is classifed as type A, type B, type AB, or type O,since we wish to identify the blood type that is more rare ,the goal is to determine which blood type occurs least frequently among these persons .[5]

The problem of selecting the smallest cell probability has been considered by Alam and Thompson (1972) using indefference zone approach . According to this approach ,the

cell with smallest count is selected as the least probable event ,with ties broken by randomization .

Let p^* (1/k < p^* <1) and c ,(0<c<1/k-1) by given .the smallest count should be determined for probability of correct select selection $p\{CS\} \geq p^*$ whenever $p_{[i]} - p_{[1]} \geq c$, (i=1, 2, …, k).

This paper deals with Bayesian fixed sample size procedure for selecting the least cell probability in multinomial distribution whose parameters are distributed a prior according to a Dirichlet distribution. Section 2 contains the formulation of the problem .Section 3 ,presents the prior and posterior probabilities .In section 4 ,we develop a procedure for selecting the smallest cell probability in multinomial distribution using a Bayesian Decision – theoretic approach . Section 5 contains some concluding remarks and future works .

2. Formulation of the Problem

The problem of selecting the smallest cell probability is formulated as follows :

Let $\underline{n}$ has the multinomial distribution with probability mass function

$$P(\underline{n} \mid \underline{p}) = \frac{m!}{n_1! n_2! .. n_k!} \prod_{i=1}^{k} p_i^{n_i} \quad , \sum_{i=1}^{k} n_i = m$$

$$\underline{n} = (n_1, ..., n_k) \quad , \quad \sum_{i=1}^{k} p_i = 1$$

and $\underline{p} = (p_1, p_2, ..., p_k)$ with p_i is the probability of an observation in the cell i .

Let $p_{[1]} \leq p_{[2]} \leq ... \leq p_{[k]}$ denote the ordered values of the $p_i (1 \leq i \leq k)$

the goal of the experimenter is to select the least cell probability ,that is the cell . associated with $p_{[1]}$.

3. Prior and Posterior Probabilities

In general ,the vector $\underline{p}$ may be unknown ,and the decision _maker is assumed to assess a prior distribution $\pi(\underline{p})$ on the unknown parameters . The revision of the

prior distribution in light of sample information by Bayes rule is simplified if the prior a member of a family that is conjugate to the multinomial distribution .To show this simplificaton ; consider a prior density of the Dirichlet (vector beta) family [wilks 1962]

$$\pi(\underline{p}) = \frac{\Gamma\left(\sum_{i=1}^{k} n_i'\right)}{\prod_{i=1}^{k}\Gamma(n_i')} \prod_{i=1}^{k} p_i^{n_i'-1} \quad .$$

The distribution may be denoted by $Dir(n_1', n_2', ..., n_k', m')$ $m' = \sum_{i=1}^{k} n_i'$ and its marginal distribution for p_i is the Beta

$$f(p_i) = \frac{(m'-1)!}{(n_i'-1)!(m'-n_i'-1)!} p_i^{n_i'-1}(1-p_i)^{m'-n_i'-1} \quad .$$

Since

$P(\underline{n} \mid \underline{p}) \propto p_1^{n_1} .... \ p_k^{n_k}$ and

$\pi(\underline{p}) \propto \ p_1^{n_1'-1} .... \ p_k^{n_k'-1}$

hence the posterior probability $\pi(\underline{p} \mid \underline{n}) \propto \ p_1^{n_1+n_1'-1} .... \ p_k^{n_k+n_k'-1}$

which is a member of the Dirichlet family with parameters

$n_i'' = n_i' + n_i$ and $m'' = m' + m$ (1,2, ., k) with mean $\hat{p}_i = \frac{n_i''}{m''}$

will be termed the posterior frequency in the cell i .The normalizing constant is :

$$\frac{(m''-1)!}{(n_1''-1)!(n_2''-1)!....(n_k''-1)!} p_1^{n_1''-1} ....p_k^{n_k''-1}.$$

The posterior distribution is denoted by $Dir(n_1', n_2', ..., n_k', m')$ and its marginal distribution for $p_i (1 \le i \le k)$ is the Beta .

$$f(p_i \mid n_i'') = \frac{\Gamma(m'')}{\Gamma(n_i'')\Gamma(m''-n_i'')} p_i^{n_i''-1}(1-p_i)^{m''-n_i''-1}$$

4. Construction of the Bayes Procedures

In this section a fixed sample size procedure for selecting the least cell probability in multinomial distribution using Bayesian decision theoritic approach is developed .

Before we introduce the Bayesian procedures, certain definitions and notations are needed . Let $\Omega_k : \{\underline{p} = (p_1, p_2, ..., p_k) : \sum_{i=1}^{k} p_i = 1 \ ; p_i \geq 0\}$ be the parameter space and $D = \{d_1, d_2, ..., d_k\}$ be the decision space where in the following terminal *k*-decision rule:

$d_i : p_i$ is the sallest cell probability (*i*=1, 2, …, *k*).

That is, d_i denote the decision to select the event associated with the i^{th} cell as the least probable event, after the sampling is terminated.

Suppose the loss function in making decisions d_i, defined on $\Omega_k \times D$, is given as follows.

$$L(d_i, \underline{p}^*) = \begin{Bmatrix} k^*(p_i - p_{[1]}) & \text{if } (p_{[1]} \neq p_i) \\ 0 & \text{if } (p_{[1]} = p_i) \end{Bmatrix} \quad \ldots \quad (4.1)$$

That is the loss if decision d_i is made when the true value of $\underline{p} = \underline{p}^*$. Where k^* is the loss constant, giving losses in terms of cost.

The Bayesian approach requires that we specify a prior probability density function $\pi(\underline{p})$ expressing our beliefs about $\underline{p}$ before we obtain the data

From a mathematical point of view, it would be convenient if $\underline{p}$ is assigned a prior distribution which is a member of a family of distributions closed under multinomial sampling or as a member of the conjugate family.

let $\underline{p}$ is assigned dirichlet prior distribution with parameters $m', n_1', n_2', ..., n_k'$

After m observations have been taken ,the total loss is given by

$$L(d_i, \underline{p}^*) = mc + k^*(p_i - p_{[1]})$$

The stopping risk (the posterior expected loss) of the terminal decision d_i when the posterior distribution for $\underline{p}$ has parameters $(n_1'', n_2'', ..., n_k''; m'')$, that is when the sample path has reached $(n_1'', n_2'', ..., n_k''; m'')$ from the origin $(n_1', n_2', ..., n_k'; m')$, denoted by $S_i(n_1'', n_2'', ..., n_k''; m'')$, can be found as follows.

$$S_i(n_1'', n_2'', ..., n_k''; m'') = \underset{\pi(\underline{p}|\underline{n})}{E}[L(d_i, \underline{p}^*)]$$

$$= mc + k^* \left[\frac{n_i''}{m''} - \underset{\pi(\underline{p}|\underline{n})}{E}(\text{þ}_{[1]}) \right] \quad \text{. ..(4,2)}$$

the value of $\underset{\pi(\underline{p}|\underline{n})}{E}[\text{þ}_{[1]}]$ is derived as follows.

$$\underset{\pi(\underline{p}|\underline{n})}{E}[\text{þ}_{[1]}] = \int_0^1 p_{[1]}.g(p_{[1]})dp_{[1]} ,$$

where $g(p_{[1]}) = k\, f(p_{[1]}).\left[1 - F(p_{[1]})\right]^{k-1}$ be the probability density function of the largest order statistics $p_{[1]}$. Let the ordered values of $n_1'', n_2'', ..., n_k''$ is $n_{[1]}'' \le n_{[2]}'' \le, ..., \le n_{[k]}''$. The marginal posterior probability density function of p_i if $p_i = p_{[1]}$ is

$$f(p_{[1]}) = \frac{(m''-1)!}{(n_{[1]}''-1)!(m''-n_{[1]}''-1)!} p_{[1]}^{n_{[1]}''-1}(1-p_{[1]})^{m''-n_{[1]}''-1} \quad \text{.....(4,3)}$$

and the cumulative density function is

$$F(p_{[1]}) = \sum_{j=n_{[1]}''}^{m''-1} \frac{(m''-1)!}{j!(m''-1-j)!}.p_{[1]}{}^{j}(1-p_{[1]})^{m''-1-j} .$$

Then,

$$\underset{\pi(\underline{p}|\underline{n})}{E}(\text{þ}_{[1]}) = \frac{k\left[(m''-1)!\right]^k}{(n_{[1]}''-1)!(m''-n_{[1]}''-1)!} \left\{ \int_0^1 \left[1 - \sum_{j_l=n_{[1]}''}^{m''-1} \frac{\left[\frac{p_{[1]}}{(1-p_{[1]})}\right]^j}{j!(m''-1-j)!} \right]^{k-1} \right.$$

$$\left. p_{[1]}^{n_{[1]}''}(1-p_{[1]})^{km''-n_{[1]}''-1} \right\}.dp_{[1]}$$

$$= \frac{k[(m''-1)!]^k}{(n_{[1]}''-1)!(m''-n_{[1]}''-1)!} \sum_{l=0}^{k-1} \left[\left(\frac{k-1}{l}\right)(-1)^l((m''-1)!)^l \sum_{j_1=n_{[1]}''}^{m''-1} \sum_{j_2=n_{[1]}''}^{m''-1} ..... \sum_{jl=n_{[1]}''}^{m''-1} \int_0^1 \left(\frac{p_{[1]}}{1-p_{[1]}}\right)^{j_1+j_2+...+j_l+n_{[1]}''} \right.$$

$$\left. \frac{p_{[1]}^{n_{[1]}''}(1-p_{[1]})^{m''-n_{[1]}''-1}}{j_1!(m''-j_1-1)!...j_l!(m''-j_l-1)!} \right] dp_{[1]}$$

.

$$= \frac{k[(m''-1)!]^k}{(n''_{[1]}-1)!(m''-n''_{[1]}-1)!}\sum_{l=0}^{k-1}\left[\binom{k-1}{l}(-1)^l((m''-1)!)^l \sum_{j_1=n''_{[1]}}^{m''-1}\sum_{j_2=n''_{[1]}}^{m''-1}\ldots\ldots\sum_{jl=n''_{[1]}}^{m''-1}\right.$$

$$\left.\frac{(j_1+j_2+\ldots+j_l+n''_{[1]})!(1+l)m''-n''_{[1]}-l-j_1-j_2-\ldots-j_l)!}{j_1!j_2!\ldots j_l!((1+l)m''-j_1-1)!((1+l)m''-j_2-1)!\ldots.((1+l)m''-j_l-1)!}\right] \qquad \ldots(4.4)$$

5 Conclusion and Future Work

1. Conclution

It is quite clear that the problem of selecting the category with largest probability is not equivalent (and not reducible) to that of selecting the category with the smallest probability .the former problem is treated in Bechhofer ,Elmaghraby,and Morse (1959),and the latter is by Alam and Thompson (1972) .

Although these two papers both use a standard type requirement for the probability of correct selection ,they actually require different measures of distance to obtain a solution . in the former paper the measure of distance is the ratio $p_{[k]}/p_{[k-1]}=\delta$, $\delta \geq 1$ and $p\{CS\} \geq p^*$ wherever $\delta \geq \delta^*$, $\{\delta^*, p^*\}$ requirement is determined in advance .in the latter paper the measure of distance is $p_{[k]}-p_{[k-1]}=\delta^*$ where $\delta \geq \delta^*$.

2. Future Work

- Our plan in future is to produce some numerical results.
- Fully Bayesisn sequential scheme to selecting the smallest multinomial selection problem can be developed and comparisons with the Bayesian fixed sample size procedure can be tried.
- An upper bound for risks may be found using functional analysis.
- General loss functions may be used, where linear loss is considered as a special case.
- To simplify the formula (4.4) we can use stirling's approximation for large factorials and hence we will get an approximate formula to (4.4) .

References

1. Alam, K. (1971). On selecting the most probable category Technometrics, vol.13, No.4, pp. 843-850.
2. Alam, K. Seo, K. and Thompson, J. R. (1971). A sequential sampling rule for selecting the most Probable Multinomial event. Annals of the INST. Of statist. Math. 33, No.3, pp. 365- 374.
3. Alam, K. Kenzo, S. and James, R. (1970). A sequential sampling rule for selecting the most probable Multinomial event.
4. Alam, K. and Thampson, J. R. (1972). On selecting the least probable Multinomial Event. Ann. Math. Statist. 43. pp. 1981-1990.
5. Bechhofer, R. E., Elmaghraby, S. and Morse, N. (1959). A single- sample multiple-decision procedure for selecting the Multinomial event which has the largest probability. Ann. Math. Statist., 30, pp. 102-119.
6. Bechhofer, R. E. and Kulkarni, R.V. (1984). Closed sequential which have the largest probabilities. Communications in statistics-Theory and Methods. A13 (24). pp. 2997- 3031.

الملخص

في هذا البحث ، تم عرض أجراء ذو حجم عينة ثابت لأختيار الحدث الاقل احتمال ،أي الخلية ذات الاحتمال الأقل في توزيع متعدد الحدود .

أستخدم منهج القرار البيزيني لبناء هذا الاجراء .وتضمن البحث أشتقاق خطورة بيز الناتجة من اتخاذ القرارات باستخدام الخسارات الخطية وأستخدام توزيع درشلت كاحتمال قبلي .