COMPARE THE NUMBER OF SPECIES IN TWO POPULATIONS: GROUP SAMPLING CASE

Jack C. Yue

Department of Statistics
National Chengchi University
Taipei, Taiwan, R.O.C.

Murray K. Clayton

Department of Statistics
University of Wisconsin–Madison
Madison, WI 53706

ABSTRACT

In Yue and Clayton (1996), a sequential Bayesian approach was used to compare the number of species in two populations based on a cost structure, which consists of a sampling cost and a misclassification loss. We extend this study to the case when the observations are taken in groups. Based on the Bayes risk, decision rules with different number of observations in a group are compared.

Key Words and Phrases: Sampling species; Species diversity; Number of species; Sequential sampling; Bayesian decision making; Group sampling

1. INTRODUCTION

The number of species in a population is a frequently used measure in comparing populations. The estimation of Shakespeare's vocabulary (Efron and Thisted, 1976) is one of the famous applications in Linguistics, in which case the vocabulary of an author is treated as the number of species in a population. However, most of the past work (see Bunge and Fitzpatrick, 1993 or Yue, 1994 for detailed discussion)

only deals with the estimation of the number of species, instead of the comparison of populations in terms of the number of species.

Yue and Clayton(1996) considered a decision theory approach, combining Bayesian and sequential frameworks, to compare two populations. They assume one observation is collected at a time from each population with a sampling cost c. When sampling is terminated, a decision of whether two populations has the same number of species must be made. No loss is added if the correct decision is made; otherwise, there will be a decision loss 1 (i.e. 0–1 loss). They showed that the optimal rule exists and is truncated when the prior distribution of the number of species is bounded, or the prior distribution is unbounded and satisfies certain conditions. Note that the prior structure used in Yue and Clayton is proposed by Lewins and Joanes(1984):

$$\pi(s) \times f(p_1, \ldots, p_s | s, K) = \pi(s) \times \frac{\Gamma(Ks)}{[\Gamma(K)]^s} \prod_{i=1}^{s} p_i^{K-1}, \tag{1}$$

where s is the number of species, p_i is the species proportion of species i, and $K > 0$.

The assumption in Yue and Clayton that observations are sampled in pairs is questionable. In reality, this is not cost efficient and sometimes not possible. For example, when setting a trap to catch insects, it is difficult to catch exactly one insect. It is appropriate to consider a version of this problem that incorporates group sampling. In this paper, we will study the group sampling case and compare the Bayes risk of decision rules with different number of observations in a group (or different group sizes). The known group size case is discussed in the following section, and the unknown group size case is in Section 3.

2. KNOWN GROUP SIZE

We will consider first the case when the number of observations taken from each population is known and the same, and leave the unknown case to Section 3. Since the capacity of traps usually has a limit, it is reasonable to assume that the number of observations taken in a group is bounded. We shall assume that this limit is M_0 and that the group size is at most M_0 in this study. Also, we assume that, when taking one group of observations from each population, the group sizes shall be the same and the sampling cost is c for one group of observations from each population.

We first define some notation. Let k denote the group size and $r^k(\pi, \delta)$ be the Bayes risk of decision rule δ when the prior distribution of the number of species is π and the group size is k. Also, let $r^k(\pi)$ be the Bayes risk of the optimal rule given the prior π and group size k, if the optimal rule exists and is truncated. We shall

consider as well the m–truncated rule that at most m groups will be taken from each population, and let $r_m^k(\pi)$ be the Bayes risk of the m–truncated rule. Similar to Yue and Clayton (1996), we can use Doob's process to show that in the case of 0–1 loss, more groups of observations lead to smaller expected posterior loss and Bayes risk. Similarly, we can also show that the Bayes rule in the group sampling case will still be truncated when the prior distribution of the number of species is bounded, or the prior distribution is unbounded and satisfies certain conditions (Theorems 5 and 6 in Yue and Clayton). The convergence of the m–truncated rule to the optimal stopping rule also follows.

Theorem 1 The Bayes risk of the m–truncated rule converges to the Bayes risk of the optimal rule, i.e.

$$\lim_{m\to\infty} r_m^k(\pi) = r^k(\pi),$$

provided that the optimal rule exists and is bounded.

Because the group size is not restricted to one, we can compare the efficiency of different sampling methods. For instance, we are interested in knowing whether sampling more observations at every stage would lower the expected posterior loss if the sampling cost is the same per pair of groups.

Theorem 2 The Bayes risk of the m–truncated rule is a non-increasing function of the group size k, i.e.

$$r_m^{k+1}(\pi) \le r_m^k(\pi).$$

Proof: Let τ_m^k be the decision rule which uses the m–truncated procedure as the stopping rule when the group size is k for every draw. Define a decision rule $\hat{\tau}_m^k$, which also uses the m–truncated procedure as the stopping rule but $\hat{\tau}_m^k$ takes j extra observations when τ_m^k stops sampling after taking j groups of observations from each population. From Theorem 2 of Yue and Clayton(1996), since $\hat{\tau}_m^k$ has a smaller expected posterior loss than τ_m^k. As a result, the Bayes risk of $\hat{\tau}_m^k$ is smaller than that of τ_m^k. Note that although $\hat{\tau}_m^k$ follows the same stopping rule as τ_m^k, it is a decision rule that has $k+1$ observations in each group and allows a maximum number of m draws from each population. Because τ_m^{k+1} has the smallest Bayes risk among the decision rules which take $k+1$ observations in a group and take at most m groups of samples, it is obvious that

$$r_m^{k+1}(\pi) \le r_m^k(\pi).\square$$

Corollary 1 The Bayes risk of the optimal rule is a non-increasing function of the group size k, provided that the optimal rule is truncated.

The idea of Theorem 2 can be generalized to the case that the number of observations in a group is not fixed. In particular, suppose there are two sampling methods, say A and B. Let A_i and B_i be the group sizes at stage i for using sampling method A and B, respectively. Assume that

$$\sum_{i=1}^{n} A_i \leq \sum_{i=1}^{n} B_i, \qquad \text{for all } n > 0,$$

which means that the accumulated number of observations by using method B is greater than or equal to that of using method A. Then the Bayes risk of the m–truncated rule using sampling method B is not larger than that of the m-truncated rule using sampling method A, or $r_m^A(\pi) \geq r_m^B(\pi)$ for all m. The proof is similar to the proof in Theorem 2. Simply use the difference of $\sum_{i=1}^{j} A_i$ and $\sum_{i=1}^{j} B_i$ to replace j in the proof of Theorem 2, when the m–truncated rule using sampling method A stops after j groups of observations are taken from each population.

Note that when setting traps to catch insects and butterflies, the sampling cost is often of the form: $g + kc$, where g is the cost of setting a trap, k is the number of observations in the trap, and c is the cost of counting and recording the observations in the trap. Under this cost structure, most results in the previous group sampling case are still valid, for example, the truncation of the Bayes risk and the convergence of the m–truncated rule to the optimal stopping rule. However, more observations in a group do not guarantee smaller Bayes risk (Theorem 2), since there are extra costs for more observations.

Example 1. Let s_1 and s_2 denote the numbers of species in populations 1 and 2, respectively. Assume that s_1 and s_2 have the same prior: $\pi(s) = 0.5\, I\{s = 3\} + 0.5\, I\{s = 4\}$. Then $r_0(\pi) = 0.5$ and $r_1^k(\pi)$ for $k = 1, 2, 3, 4$ is:

	$k=1$	$k=2$	$k=3$	$k=4$
$r_1^k(\pi)$	$\min\{0.5,$ $0.5 + g + c\}$	$\min\{0.5,$ $0.495 + g + 2c\}$	$\min\{0.5,$ $0.495 + g + 3c\}$	$\min\{0.5,$ $0.470 + g + 4c\}$

With $g = 0.002$ and $c = 0.001$, $r_1^k(\pi) = 0.5, 0.499, 0.5$, and 0.476 for $k = 1, 2, 3, 4$, respectively. In other words,

$$r_1^1(\pi) = r_1^3(\pi) > r_1^2(\pi) > r_1^4(\pi),$$

which means that r_m^k is not a non-increasing function of k, and r_m^k can also have two or more than two local minimums. In the following discussion, we will focus on the original group sampling cost structure, and consider the unknown group size.

3. UNKNOWN GROUP SIZE

The result of Theorem 2 can be further extended to the case that the group size is random, under certain conditions. Let Y be stochastically larger than X, or $X \leq_{st} Y$, if

$$P(X \leq z) \geq P(Y \leq z), \qquad \text{for } z \geq 0,$$

where X and Y are random variables. For discussion of stochastic ordering, see, for example, Marshall and Olkin (1979). Then $A_i \leq_{st} B_i$ for all $i = 1, 2, \ldots$, where $P(A_i \leq M_0) = P(B_i \leq M_0) = 1$ and M_0 is a positive number, also implies that $r_m^A(\pi) \geq r_m^B(\pi)$ for all m. Note that $A_i \leq_{st} B_i$ for all i means that using method B always has a better chance to take more observations than using method A at each stage.

Theorem 3 Suppose that at every stage the number of observations taken under sampling method B is stochastically larger than that under method A, and the sampling cost is the same under these two methods. Then the Bayes risk of truncated rules under method B is smaller than or equal to that under method A.

Proof: Note that if $A_i \leq_{st} B_i$ then we can write:

$$P(A_i = A_{ij}) = q_{ij} = P(B_i = B_{ij}), \sum_j q_{ij} = 1 \qquad \text{for all } i \text{ and } q_{ij} \geq 0, \tag{2}$$

where the A_{ij}'s and B_{ij}'s are positive integers and are bounded increasing functions of i for every j, and $A_{ij} \leq B_{ij}$ for all i, j. To see this, suppose for every i,

$$P(A_i = k_j) = r_{ij} > 0 \text{ and } P(B_i = k_j^*) = r_{ij}^* > 0,$$

where k_j and k_j^* are positive integers and are increasing functions of j, $r_{ij} > 0$ and $r_{ij}^* > 0$ with $\sum_j r_{ij} = \sum_j r_{ij}^* = 1$ for each i. In other words, r_{ij}'s and r_{ij}^*'s form two partitions of [0,1], with

$$0 < r_{i1} < r_{i1} + r_{i2} < \cdots < \sum_{j=1}^{M_0} r_{ij} = 1$$

and

$$0 < r_{i1}^* < r_{i1}^* + r_{i2}^* < \cdots < \sum_{j=1}^{M_0} r_{ij}^* = 1.$$

We can combine the above two partitions and define a new, refined partition on [0,1] by defining $q_{ij} > 0$, where

$$q_{ij} = \min_n \{ \sum_{m=1}^{n} r_{im} - \sum_{m=1}^{j-1} q_{im}, \sum_{m=1}^{n} r_{im}^* - \sum_{m=1}^{j-1} q_{im} \},$$

$q_{i1} = \min\{r_{i1}, r^*_{i1}\}$, and

$$0 < q_{i1} < q_{i1} + q_{i2} < \cdots < \sum_{j=1}^{2M_0} q_{ij} = 1.$$

If $A_i \leq_{st} B_i$, then we can rearrange the k_j's and k^*_j's to come out with expression (2).

On the other hand, expression (2) also leads to $A_i \leq_{st} B_i$ for all i since

$$\begin{aligned} P(A_i \leq z) &= \sum_j P(A_i = A_{ij}) I\{A_{ij} \leq z\} \\ &= \sum_j P(B_i = B_{ij}) I\{A_{ij} \leq z\} \\ &\geq \sum_j P(B_i = B_{ij}) I\{B_{ij} \leq z\} \\ &= P(B_i \leq z). \end{aligned}$$

Since $A_{ij} \leq B_{ij}$, which implies that $I\{A_{ij} \leq z\} \geq I\{B_{ij} \leq z\}$ for all z, the second line to the third line of the above equation follows. Therefore, it is sufficient to use (2) to show that $r^A_m(\pi) \geq r^B_m(\pi)$ for $A_i \leq_{st} B_i$, if the sampling costs of using method A and B are the same.

Suppose τ^A_m and τ^B_m are the m–truncated procedures on which method A and B are used. Define $\hat{\tau}^B_m$ similar to τ^A_m, except that when τ^A_m stops, $\hat{\tau}^B_m$ takes x_i more pairs of observations at stage i, where $P(x_i = B_{ij} - A_{ij}) = q_{ij}$ for $j = 1, 2, \ldots, 2M_0$. Therefore, on average, $\hat{\tau}^B_m$ takes

$$\sum_{i=1}^{N_0} \sum_j q_{ij} I\{k = A_{ij}\}(B_{ij} - A_{ij}).$$

more pairs of observations if the m–truncated rule using method A stops sampling after N_0 stages of observations.

Note that the sampling distribution of the number of observations taken at every stage in $\hat{\tau}^B_m$ is the same as under method B. Thus, $\hat{\tau}^B_m$ is also a truncated decision rule using sampling method B with truncation bound m and so $r^B_m \leq r^A_m$.□

Example 2. Suppose that using method A, $P(A_i = 1) = 0.5 = P(A_i = 2)$ for all i and that using method B, $P(B_i = 1) = 0.4 = 1 - P(B_i = 2)$ for all i. Then $q_{i1} = 0.4, q_{i2} = 0.1, q_{i3} = 0.5$ for all i and $A_{i1} = 1, A_{i2} = 1, A_{i3} = 2$, while $B_{i1} = 1, B_{i2} = 2, B_{i3} = 2$. Suppose τ^A_m stops at $N_0 = 5$ and

the numbers of observations taken at every stage from each population are $(1,1,2,1,2)$, i.e. one observation each in the first, second, and fourth groups, and two observations each in the third and fifth groups. Then with probability $0.4\times(1-1)+0.1\times(2-1)=0.1$, $\hat{\tau}_m^B$ takes an extra $0.4\times(1-1)+0.1\times(2-1)=0.1$ pair of observations from each population at stage 1. Stage 2 to 5 are similar and thus $\hat{\tau}_m^B$ takes on average

$$0.1\times(2-1)+0+0.1\times(2-1)+0.1\times(2-1)+0=0.3$$

more pairs of observations and then stops sampling. From Theorem 2, since $\rho_0(\pi^n)\geq E^*[\rho_0(\pi^n)\mid y_{n+1},\ldots,y_{n+j}]$, it follows that the expected posterior loss of $\hat{\tau}_m^B$ is smaller than that of τ_m^A.□

4. DISCUSSION

Since $A_i\leq_{st}B_i$ for every i can guarantee that $r_m^A\geq r_m^B$ for every m, we might expect that the result in Theorem 3 can be extended to the case that $\sum_{i=1}^n A_i\leq_{st}\sum_{i=1}^n B_i$ for all n, i.e.

$$P(\sum_{i=1}^n A_i\leq z)\geq P(\sum_{i=1}^n B_i\leq z). \tag{3}$$

However, this is not true, as is shown in the following example.

Example 3. Suppose that $P(A_1=1)=0.872$ and $P(A_1=2)=0.128$. Also $A_2=2$, $B_1=2$, and $B_2=1,2$ with probabilities 0.872 and 0.128, respectively. Let the sampling cost $c=0.0015$. Assume that s_1 and s_2 have the same prior: $\pi(s)=0.5I\{s=3\}+0.5I\{s=4\}$. It is obvious that $P(A_1\leq 1)=0.872>0=P(B_1\leq 1)$ and $P(A_1\leq 2)=1=P(B_1\leq 2)$. Similarly, $P(A_1+A_2\leq 3)=0.872=P(B_1+B_2\leq 3)$ and $P(A_1+A_2\leq 4)=1=P(B_1+B_2\leq 4)$ and so $\sum_{i=1}^n A_i\leq_{st}\sum_{i=1}^n B_i$ for $n=1$ and 2.

It can be shown that if method B and the 2–truncated rule are used, then the sampling is terminated at $n=2$, and $r_2^B=0.49650$. On the other hand, if method A and the 2–truncated rule are used, then the sampling is terminated at $n=3$ or 4, and $r_2^A=0.49648$. This implies that $r_2^A<r_2^B$.□

Although we assume that both populations have the same group size in the proceeding discussion, the results are also valid when populations have different group size. For example, one sampling plan with 4 and 5 observations per group from populations 1 and 2, has smaller Bayes risk than the other with 2 and 4 observations per group from populations 1 and 2. And in general, if for each population using

sampling method B always leads to more observations in a group than using method A, i.e. $A_i \leq_{st} B_i$, then using method B would give us a smaller Bayes risk. The proof is similar and is thus omitted.

BIBLIOGRAPHY

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer–Verlag, New York.

Bunge, J. and Fitzpatrick, M. (1993) "Estimating the Number of Species: A Review", *Journal of the American Statistical Association*, 88, 364-373.

Efron, B. and Thisted, R. (1976) "Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?", *Biometrika*, 63, 435-447.

Lee, J. (1989) *On Asymptotics for the NPMLE of the Probability of Discovering a New species and an Adaptive Stopping rule in Two-Stage Searches*, Ph.D. Thesis, Department of Statistics, University of Wisconsin–Madison.

Lewins, W. A. and Joanes, D. N. (1984) "Bayesian Estimation of the Number of Species", *Biometrics*, 40, 323-328.

Marshall, A. W. and Olkin, I. (1979) *Inequalities: Theory of Majorization and Its Applications*, Academic Press.

Rasmussen, S. L. and Starr, N. (1979) Optimal and Adaptive Stopping in the Search for New Species, *Journal of the American Statistical Association*, 74, 661-667.

Yue, C. J. (1994) *Bayesian Sequential Tests for Comparing the Species Richness of Two Populations*, Ph.D. Thesis, Department of Statistics, University of Wisconsin–Madison.

Yue, C. J. and Clayton, M. K. (1996) "Bayesian Sequential Tests for Comparing the Number of Species in Two Population", *Sequential Analysis*, 15, 185-210.