응용통계연구 제 7권 1호
1994년 3월 pp. 159 - 172

Empirical Bayes Estimation of the Probability of Discovering a New Species[1]

Jooho Lee[2]

Abstract

An empirical Bayes estimator of the probability of discovering a new species is proposed when some prior information is available on the number of species. The new estimator is shown via simulations to have only a moderate bias and a smaller RMSE than Good's estimator when the species population follows a truncated geometric distribution.

1. Introduction

The estimation of the probability of discovering a new species in a population is an inferential problem that some statisticians have long struggled with since Good(1953) proposed a nonparametric estimator. To be more specific, let s be the number of species in an infinite population and let $\boldsymbol{p} = (p_1, \cdots, p_s)$ denote the species relative abundances in the population. Assume that a sample of size n contains t distinct species. Then the probability of discovering a new species at the $(n+1)$th observation can be expressed as

$$U(n) = \sum_i p_i I(Y_i = 0),$$

where Y_i, $i = 1, \cdots, s$, denotes the abundance of the ith species in the sample. The estimator of $U(n)$ that Good derived using a nonparametric Bayesian argument is

$$V_0 = (1/n) \sum_i I(Y_i = 1).$$

Much of later work was devoted to studying statistical properties of Good's estimator and its generalized version (Robbins, 1968, Starr, 1979, and Clayton and Frees, 1987). A few other estimators were proposed as alternatives, but Clayton and Frees' nonparametric maximum likelihood estimator(NPMLE) seems to be the only all-round competitor aside from its underbiasedness. Although a considerable negative bias that the NPMLE has was addressed and partly corrected by Lee(1989, 1993), more reduction in bias would be necessary for the NPMLE to become a superior alternative to Good's estimator.

However, if some prior information is available on the number of species before sampling

[1] This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1992.

[2] Department of Statistics, Chungnam National University, Taejeon, Korea, 305-764.

is done, then Good's estimator has no way of using this prior information. In such a situation, which is not so unusual in reality, a parametric Bayesian approach would be more appropriate. Hill(1968, 1979) is the first statistician that used a parametric Bayesian approach to derive the posterior distribution of the probability of discovering a new species. Hill's posterior mean and mode, however, are not useful as an estimator due to a heavy negative bias and a large MSE. Later, Lewins and Joanes(1984) briefly discussed a generalized version of Hill's estimator, but they did not present it in an explicit form and failed to provide a formal method for estimating the prior parameters.

The purpose of this paper is to propose an empirical Bayes estimator using the same approach as in Lewins and Joanes. Unlike in Lewins and Joanes, however, the new estimator is given in an explicit form and the estimation of the prior parameters is also discussed in detail.

The posterior probability of discovering a new species is derived in the next section. Section 3 describes how to estimate prior parameters using actual data. Section 4 presents the results of small-sample simulation study under several hypothetical population distributions.

2. Posterior Probability of Discovering a New Species

Kempton and Wedderburn(1978) argued that the frequency distribution of the species abundances can be reasonably well approximated by a gamma distribution. If the abundances of species in the population constitute independent observations from a gamma distribution, then it can be seen that the relative abundances jointly follow a Dirichlet distribution. Therefore, the conditional prior distribution for $\boldsymbol{p}$ given s is assumed to be

$$\pi_2(p \mid s) = \frac{\Gamma(ks)}{[\Gamma(k)]^s} \prod_{i=1}^{s} p_i^{k-1}. \tag{2.1}$$

As in Lewins and Joanes, the prior distribution for s is assumed to be the zero-truncated negative binomial distribution with the density

$$\pi_1(s) = \binom{s+r-1}{s}\left[\frac{\theta^r(1-\theta)^s}{1-\theta^r}\right],\ 0<\theta\leq 1,\ r\geq 1,\ s\geq 1. \tag{2.2}$$

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_s)$, $x_1 \geq x_2 \geq \cdots \geq x_s$, denote the species abundances in the sample. Since no unique labeling is available to the species appearing in the sample, the likelihood function is given by

$$f(x \mid p, s) = \sum_{R} \frac{n!}{x_{r_1}! \cdots x_{r_s}!} \prod_{i=1}^{s} p_i^{x_{r_i}},\ x_i \geq 0,\ i = 1, \cdots, s,\ \sum_{i=1}^{s} x_i = n, \tag{2.3}$$

where R is the set of permutations of $\{1, \cdots, s\}$ which have distinct $\prod_{i=1}^{s} p_i^{x_{r_i}}$ and $\{r_1, \cdots,$

$r_s\} \in R$.

We first derive the conditional marginal posterior distribution of p_1 given s to find the posterior probability of discovering a new species. The conditional marginal posterior density of p_1 given s is obtained as

$$\begin{aligned}
\pi(p_1 \mid x, s) &= \int \cdots \int \pi(p \mid x, s)\, dp_2 \cdots dp_{s-1} \\
&\propto \int \cdots \int \pi_2(p \mid s) f(x \mid p, s)\, dp_2 \cdots dp_{s-1} \\
&\propto \sum_R \frac{n!}{x_{r_1}! \cdots x_{r_s}!} \int \cdots \int \prod_{i=1}^{s} p_i^{k+x_{r_i}-1}\, dp_2 \cdots dp_{s-1} \\
&\propto \sum_R \frac{n!}{x_{r_1}! \cdots x_{r_s}!} \cdot \frac{\Gamma(k+x_{r_2}) \cdots \Gamma(k+x_{r_s})}{\Gamma((s-1)k+n-x_{r_1})}\, p_1^{k+x_{r_1}-1} \\
&\qquad \cdot (1-p_1)^{(s-1)k+n-x_{r_1}-1} \\
&\propto \frac{n!}{x_1! \cdots x_s!}\, t \binom{s}{t} \sum_{i=1}^{t} [\Gamma((s-1)k+n-x_i)]^{-1} p_1^{k+x_i-1} \\
&\qquad \cdot (1-p_1)^{(s-1)k+n-x_i-1} \sum_{R_i} \Gamma(k+x_{r_2}) \cdots \Gamma(k+x_{r_s}) \\
&\propto \sum_{i=1}^{t} \frac{\prod_{j \neq i} \Gamma(k+x_j)}{\Gamma((s-1)k+n-x_i)}\, p_1^{k+x_i-1}(1-p_1)^{(s-1)k+n-x_i-1},
\end{aligned} \tag{2.4}$$

where R_i is the set of permutations of $\{1, \cdots, s\} - \{i\}$ which have distinct $\prod_{i=2}^{s} p_i^{x_{r_i}}$ and $\{r_2, \cdots, r_s\} \in R_i$. Thus the conditional posterior mean of p_1 given s can be expressed as

$$\begin{aligned}
E(p_1 \mid x, s) &= \frac{\sum_{i=1}^{t} [\Gamma((s-1)k+n-x_i)]^{-1} \prod_{j \neq i} \Gamma(k+x_j) \int p_1^{k+x_i}(1-p_1)^{(s-1)k+n-x_i-1} dp_1}{\sum_{i=1}^{t} [\Gamma((s-1)k+n-x_i)]^{-1} \prod_{j \neq i} \Gamma(k+x_j) \int p_1^{k+x_i-1}(1-p_1)^{(s-1)k+n-x_i-1} dp_1} \\
&= \frac{[\Gamma(ks+n+1)]^{-1} \sum_{i=1}^{t} (k+x_i) \prod_{j=1}^{t} \Gamma(k+x_j)}{[\Gamma(ks+n)]^{-1} \sum_{i=1}^{t} \prod_{j=1}^{t} \Gamma(k+x_j)} \\
&= \frac{1}{t} \cdot \frac{kt+n}{ks+n}.
\end{aligned} \tag{2.5}$$

Noting that $E(p_i \mid \mathbf{x}, s)$ is the same for all i by symmetry, the posterior probability of discovering a new species is expressed as

$$Q = 1 - E[\,t\,E(p_1 \mid x, s) \mid x]$$
$$\quad = 1 - (kt + n)\,E[\,(ks + n)^{-1} \mid x],$$

which is exactly the form that Lewins and Joanes provided without any detail. Now we will proceed one step further to express Q in terms of integrals. The new expression for Q would significantly reduce the computing time for simulation study in the next section. From Lewins and Joanes the posterior distribution of s is proportional to

$$\pi^*(s \mid x) = (1-\theta)^s \binom{s+r-1}{s} \binom{s}{t} \binom{ks+n-1}{n}^{-1}, \; s \geq t,$$

so that

$$E[\,(ks+n)^{-1} \mid x]$$
$$= \frac{\sum_{s=t}^{\infty}(ks+n)^{-1}\pi^*(s \mid x)}{\sum_{s=t}^{\infty}\pi^*(s \mid x)}$$
$$= \frac{\sum_{s=0}^{\infty}[k(s+t)+n]^{-1}(1-\theta)^{s+t}\binom{s+t+r-1}{s+t}\binom{s+t}{t}\binom{k(s+t)+n-1}{n}^{-1}}{\sum_{s=0}^{\infty}(1-\theta)^{s+t}\binom{s+t+r-1}{s+t}\binom{s+t}{t}\binom{k(s+t)+n-1}{n}^{-1}}$$
$$= \frac{\sum_{s=0}^{\infty}\{\,[k(s+t)+n][k(s+t)+n-1]\cdots[k(s+t)]\}^{-1}(1-\theta)^s\binom{s+t+r-1}{s}}{\sum_{s=0}^{\infty}\{\,[k(s+t)+n-1][k(s+t)+n-2]\cdots[k(s+t)]\}^{-1}(1-\theta)^s\binom{s+t+r-1}{s}}$$
$$= \frac{E\{\,[k(S+t)][k(S+t)+1]\cdots[k(S+t)+n]\}^{-1}}{E\{\,[k(S+t)][k(S+t)+1]\cdots[k(S+t)+n-1]\}^{-1}}, \tag{2.6}$$

where S is a random variable that follows the negative binomial distribution with parameters θ and $t + r$. We need the following lemma developed in Hill(1979) to evaluate the expectations in (2.6).

Lemma: Let Y be a nonnegative random variable with probability-generating function $M(u) = E(u^Y)$, $0 < u \leq 1$. Let $a > 0$, and let b be a nonnegative integer. Then

$$E[\,(Y+a)(Y+a+1)\cdots(Y+a+b)]^{-1}$$
$$= [\Gamma(b+1)]^{-1}\int_0^1 u^{a-1}(1-u)^b M(u)\,du.$$

Since the probability-generating function for S is $M(u) = \{\theta/[1-(1-\theta)u]\}^{t+r}$, it follows from the lemma that

$$
\begin{aligned}
Q &= 1 - (kt+n) \cdot \frac{[\Gamma(n+1)]^{-1}\int_0^1 u^{kt-1}(1-u)^n M(u^k)du}{[\Gamma(n)]^{-1}\int_0^1 u^{kt-1}(1-u)^{n-1} M(u^k)du} \\
&= 1 - \frac{kt+n}{n} \cdot \frac{\int_0^1 u^{kt-1}(1-u)^n [1-(1-\theta)u^k]^{-(t+r)}du}{\int_0^1 u^{kt-1}(1-u)^{n-1} [1-(1-\theta)u^k]^{-(t+r)}du} \qquad (2.7)
\end{aligned}
$$

In particular, if $k = 1$, then (2.7) reduces to Hill(1979)'s (3.3).

3. Estimation of Prior Parameters

In order to use Q given in (2.7) as an estimator of the probability of discovering a new species, we need estimate prior parameters, k, θ, and r. A full-fledged empirical Bayes approach may be used to estimate these hyperparameters from the data. It is, however, not unusual that an experimenter has some prior information on s in the form of most likely value and confidence interval, which can be used to subjectively determine θ and r. The parameter k reflects the species relative abundances, with larger k corresponding to more uniform relative abundances. Gill and Joanes(1979) recommended values between 0 and 1 for k, but it is not intuitively easy to subjectively determine the value of k. In this paper we propose a partial empirical Bayes approach in the sense that only k is estimated from the data and θ and r are determined subjectively. Note that there may be more than one pair of values for θ and r which have the same prior mode. Among those pairs of values for θ and r, larger pairs reflect more accurate information on s.

A most commonly used empirical Bayes approach to prior determination is the ML-II approach, which finds the prior that maximizes the marginal density of $\boldsymbol{x}$. Using the lemma, the marginal density of $\boldsymbol{x}$ is seen to be proportional to

$$
\begin{aligned}
&m^*(x \mid \pi_1, \pi_2) \\
&= \left(\frac{\theta^r}{1-\theta^r}\right)\left[\frac{\prod_{i=1}^{t}\Gamma(k+x_i)}{\{\Gamma(k)\}^t}\right]\sum_{s=t}^{\infty}\frac{(s+r-1)!}{(s-t)!\,(r-1)!}(1-\theta)^s\frac{\Gamma(ks)}{\Gamma(ks+n)} \\
&\propto \left[\frac{\prod_{i=1}^{t}\Gamma(k+x_i)}{\{\Gamma(k)\}^t}\right] E\{[k(S+t)][k(S+t)+1]\cdots[k(S+t)+n-1]\}^{-1} \qquad (3.1) \\
&\propto \left[\frac{\prod_{i=1}^{t}\Gamma(k+x_i)}{\{\Gamma(k)\}^t}\right] \int_0^1 u^{kt-1}(1-u)^{n-1}[1-(1-\theta)u^k]^{-(t+r)}du.
\end{aligned}
$$

Although we can numerically find the value of k which maximizes (3.1) by taking the derivative and setting to zero, the procedure seems to be too time-consuming to be used for Monte Carlo simulations with large number of replicates. We instead consider a simpler approach that is based on the repeat rate. The repeat rate, which is defined as $\lambda = \sum_i p_i^2$, measures how uniform the population is. It takes a value between 0 and 1, with larger values implying more uneven relative abundances (see Good, 1965, for details). Good suggested estimating k by equating the expected value of the repeat rate with an estimate of it. Noting that each p_i follows the beta distribution with parameters k and $k(s-1)$ for given s, we have

$$\begin{aligned} E^{\pi_2}(\lambda \mid s) &= \sum_{i=1}^{s} E^{\pi_2}(p_i^2 \mid s) = \sum_{i=1}^{s} \frac{\Gamma(ks)\Gamma(k+2)}{\Gamma(k)\Gamma(ks+2)} \\ &= \frac{\Gamma(ks+1)\Gamma(k+2)}{\Gamma(k+1)\Gamma(ks+2)} = \frac{k+1}{ks+1}. \end{aligned} \tag{3.2}$$

Equating (3.2) with the sample repeat rate $\hat{\lambda} = \sum_i (x_i/n)^2$, we obtain

$$\hat{k} = \frac{1-\hat{\lambda}}{\hat{\lambda}s-1}. \tag{3.3}$$

Because the resulting estimate of k depends on the unknown parameter s, it could be determined recursively by substituting for s the prior mode initially and a tentative posterior mode thereafter in (3.3) until the value of s satisfying (3.3) coincides with the posterior mode of s. However, as Lewins and Joanes pointed out, this procedure always yields a posterior mode of s larger than its prior mode and hence too small a value for k. Therefore, instead of relying on this procedure, we will remove the dependence of k on s by taking expectation of (3.2) with respect to the prior distribution of s;

$$\begin{aligned} &E^{\pi_1}[E^{\pi_2}(\lambda \mid s)] \\ &\quad = E^{\pi_1}\left(\frac{k+1}{ks+1}\right) = (k+1)\int_0^1 M_s(u^k)\,du, \end{aligned}$$

where $M_s(u)$ is the probability generating function for s. Note that the second equality in the above follows from Hill's lemma. Since s has the zero-truncated negative binomial distribution with parameters θ and r, $M_s(u)$ is given by

$$\begin{aligned} M_s(u) &= \sum_{s=0}^{\infty} u^s \binom{s+r-1}{s} \frac{\theta^r(1-\theta^s)}{1-\theta^r} - \frac{\theta^r}{1-\theta^r} \\ &= \theta^r(1-\theta^r)^{-1}\{[1-(1-\theta)u]^{-r}-1\}, \end{aligned}$$

and it follows that

$$E^{\pi_1}[E^{\pi_2}(\lambda \mid s)] = \frac{(k+1)\theta^r}{1-\theta^r}\left\{\int_0^1 [1-(1-\theta)u^k]^{-r}du - 1\right\}.$$

Thus, for given θ, r, and x, the estimate of k can be obtained by numerically solving the equation

$$\frac{(k+1)\theta^r}{1-\theta^r}\left\{\int_0^1 [1-(1-\theta)u^k]^{-r}du - 1\right\} = \hat{\lambda}. \tag{3.4}$$

Here note that

$$\begin{aligned}\frac{\partial}{\partial k}E^{\pi_1}[E^{\pi_2}(\lambda \mid s)] &= E^{\pi_1}\left[\frac{\partial}{\partial k}\left(\frac{k+1}{ks+1}\right)\right] \\ &= E^{\pi_1}\left[\frac{1-s}{(ks+1)^2}\right] \le 0,\end{aligned}$$

where the equality holds if and only if $s = 1$ a.s. Hence, unless the prior distribution for s is degenerate at 1, (3.4) has a unique solution in k.

We next examine how sensitive the posterior probability is to changes in prior parameters using the Mount Kenya data from Lewins and Joanes. The data was obtained by sampling n = 1043 units from the insect population residing in Mount Kenya area. A total of 32 species were discovered and their respective abundances are shown in Table 1, where f_r denotes the number of species that have r representatives in the sample.

Table 1. Sample Abandances for the Insect Population in Mount Kenya

r	f_r	r	f_r	r	f_r	r	f_r	r	f_r
1	8	5	1	12	1	46	1	109	1
2	3	6	3	18	1	56	1	157	1
3	2	7	2	21	1	95	1	335	1
4	1	10	1	25	1	98	1		

Assuming that the prior mode of s is 45 with an 80% Bayesian interval of (35, 55), θ and r are determined as 0.7 and 107, respectively. Figure 1 shows the posterior means of probabilities of discovering a new species for various values of k when θ = 0.7 and r = 107. It can be seen that a change in the value of k causes a significant change in the posterior mean.

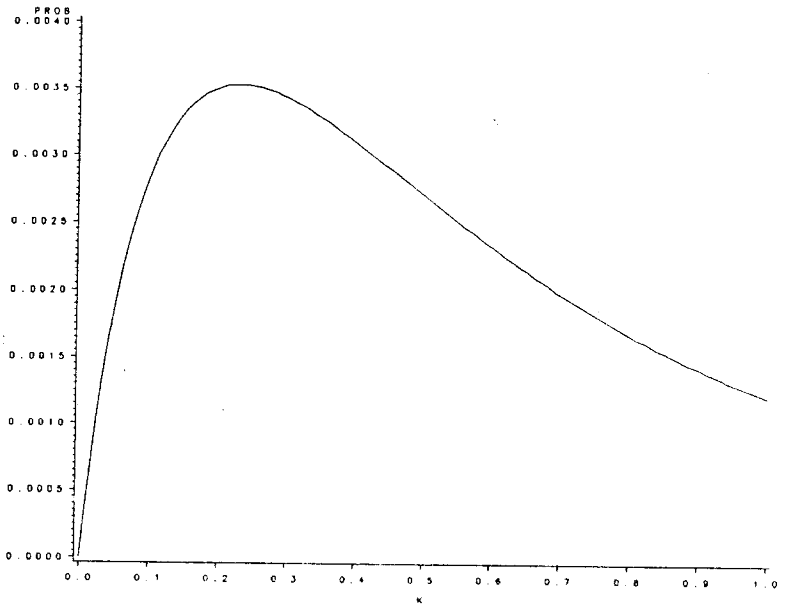

Figure 1. Posterior Means for Different Values of k (θ = 0.7 and r = 107)

If the prior mode of s is fixed at 45 but an 80% Bayesian interval is assumed otherwise, then the values of θ and r will also be different. For example, 80% interval of (37, 54) yields θ = 0.9 and r = 410, while 80% interval of (34, 58) yields θ = 0.5 and r = 47. For these three pairs of θ and r values, the values of k were estimated using (3.4) and the resulting posterior means were computed. The results in Table 2 suggest that the value of k and the resulting posterior mean do not change markedly even with severe changes in θ and r as long as the prior mode of s is maintained at the same level. This fact will be further confirmed by the simulation results in the next section. Incidentally, note that for θ = 0.7 and r = 107 our estimate 0.00324 of the probability of discovering a new species is almost three times as large as Hill's estimate of 0.00121 resulting from k = 1. This difference might be due to the fact that Hill's estimate effectively assumes the uniform conditional prior distribution for $\boldsymbol{p}$ given s, and that for relatively large samples the chance of discovering a new species is likely to become lower as the population gets more even.

Table 2. Posterior Means for Different Values of θ and r

θ	r	k	post.mean
.7	107	.1364	.00324
.9	410	.1365	.00306
.5	47	.1342	.00350

4. Simulation Results

In this section the finite sample performance of the proposed empirical Bayes estimator is studied by simulation. Good's estimator, which is the most commonly used one, is taken as a reference. In order to represent a wide spectrum of unevenness in relative abundances of a species population, truncated geometric distributions having densities $p_i = (1-p)p^{i-1}/(1-p^s)$, $i = 1, \cdots, s$, are used with p = 0.9, 0.8, and 0.7, and s = 100. Note that the truncated geometric distribution converges to the discrete uniform distribution on $\{1, \cdots, s\}$ as $p \rightarrow 1$ and the degenerate distribution at 1 as $p \rightarrow 0$, and that p closer to 1 reflects more even relative abundances. To represent different levels of accuracy in prior information, four sets of values for θ and r were considered; θ = 0.6 and r = 152, θ = 0.2 and r = 26, θ= 0.5 and r = 52, and θ = 0.7 and r = 354. They give a prior mode of 100, 99, 50, and 151, and an 80% interval of (84, 117), (74, 131), (39, 64), and (133, 170), respectively. Note that both the first and the second set estimate the number of species almost correctly but the first set represents more accurate prior information, and that the third set much underestimates the number of species, whereas the fourth set much overestimates it. For each combination of species distribution and prior distribution for the number of species, pseudo-random samples of size 50, 100, and 200 were drawn 1000 times, respectively, to approximate the means and the root mean square errors(RMSE) of Good's estimator and the empirical Bayes estimator. Uniform pseudo-random numbers were drawn using the IMSL subroutine rnun (multiplicative congruential generators with shuffling). Table 3 shows the summary of simulation results.

For all four prior distributions considered, the empirical Bayes estimator Q has smaller RMSE than Good's estimator V_0, with the ratio of two RMSE's ranging from 1.01 to 1.49 depending on the species distribution and the sample size. It is against our intuition that the ratio is not highest when prior information on the number of species is most accurate (θ = 0.6 and r = 152). In fact, the ratio is almost consistently highest when the number of species is understated in a prior distribution (θ = 0.2 and r = 52), and lowest when the number of species is overstated (θ = 0.7 and r = 354). This phenomenon could even be capitalized on if an experimenter gives a prior distribution for the number of species rather conservatively. Incidentally, note that there does not exist much difference in the ratio between the first two prior distributions, which suggests that the empirical Bayes estimator is robust with respect to change of θ and r as long as the prior mode is fixed. There appears to be no definite pattern in the ratios as the species distribution or the sample size changes.

Unlike the NPMLE that is underbiased, the empirical Bayes estimator seems to be overbiased except when θ = 0.5, r = 52, and p = 0.9, with the percentage bias lying between -17.2% and 48.0%. Q is least biased when θ = 0.5 and r = 52, which can be again rather advantageous in application if one takes a conservative prior distribution for the number of species. The percentage bias tends to increase slightly either as the sample size increases or as the species distribution gets more uneven. Considering both RMSE and bias, the empirical Bayes estimator performs best when θ = 0.5 and r = 52, while it performs worst when θ = 0.7 and r = 354.

Table 3. Comparison of V_0 and Q in RMSE

	p	n	$E[U(n)]$	$E(V_0)$ $RMSE(V_0)$	$E(Q)$ $RMSE(Q)$	$RMSE(V_0)/$ $RMSE(Q)$
θ = .6 r = 152	.9	50	.18347	.18920 .07915	.19059 .05752	1.37591
		100	.09288	.09521 .04121	.10975 .03370	1.22301
		200	.04718	.04756 .02137	.05513 .01670	1.27988
	.8	50	.08681	.08992 .05798	.11234 .04968	1.16698
		100	.04438	.04486 .02822	.06025 .02551	1.10611
		200	.02221	.02300 .01504	.03022 .01282	1.17258
	.7	50	.05413	.05560 .04677	.07379 .03987	1.17321
		100	.02740	.02848 .02209	.03856 .01854	1.19181
		200	.01442	.01391 .01153	.01932 .00937	1.23030
θ = .2 r = 26	.9	50	.18544	.18680 .08174	.19688 .06473	1.26272
		100	.09408	.09412 .04174	.10806 .03410	1.22425
		200	.04753	.04694 .02130	.05211 .01610	1.32316
	.8	50	.08810	.08946 .05608	.11116 .05011	1.11927
		100	.04491	.04477 .03027	.05727 .02542	1.19089
		200	.02223	.02215 .01472	.02804 .01160	1.26877
	.7	50	.05572	.05488 .04624	.07213 .03803	1.21600
		100	.02831	.02801 .02477	.03719 .01978	1.25248
		200	.01370	.01423 .01165	.01826 .00899	1.29637

Table 3. (continued)

	p	n	$E[U(n)]$	$E(V_0)$ $RMSE(V_0)$	$E(Q)$ $RMSE(Q)$	$RMSE(V_0)/$ $RMSE(Q)$
	.9	50	.18521	.18958 .08411	.17512 .05941	1.41562
		100	.09455	.09456 .04316	.08928 .02994	1.44162
		200	.04674	.04720 .02109	.03869 .01634	1.29123
θ = .5	.8	50	.08849	.08880 .05770	.10295 .04505	1.28078
		100	.04377	.04490 .02854	.05166 .02099	1.35988
r = 52		200	.02214	.02273 .01496	.02405 .01005	1.48814
	.7	50	.05557	.05476 .04656	.06843 .03500	1.33015
		100	.02783	.02795 .02348	.03450 .01765	1.32987
		200	.01390	.01434 .01187	.01665 .00812	1.46215
	.9	50	.18686	.18976 .08119	.19491 .06041	1.34409
		100	.09418	.09465 .04049	.11705 .03849	1.05202
		200	.04723	.04735 .02134	.06202 .02108	1.01256
θ = .7	.8	50	.08853	.08972 .05961	.11478 .05176	1.15161
		100	.04407	.04418 .03005	.06351 .02887	1.04065
r = 354		200	.02218	.02231 .01500	.03255 .01421	1.05572
	.7	50	.05383	.05684 .04397	.07555 .03909	1.12466
		100	.02753	.02790 .02160	.03977 .01968	1.09722
		200	.01379	.01413 .01198	.02041 .01035	1.15784

5. Conclusion

In this paper we proposed an empirical Bayes estimator of the probability of discovering a new species when some prior information is available on the number of species. Irrespective of the accuracy in prior specification, the new estimator appears to have smaller RMSE but to be more biased than Good's estimator that is most commonly used. It seems that RMSE increases and the amount of overbias decreases as the number of species becomes understated in the prior distribution. Therefore, by specifying the prior distribution of the number of species on the conservative side, one could expect to reduce both bias and RMSE. Although not intended to be used in ignorance situations, the proposed estimator may possibly be used in such situations by conservatively taking the prior mode of the number of species to be the number of species found in the sample. A Bayesian interval need also be specified to uniquely determine the hyperparameters θ and r, but it may not be so critical due to the robustness of the proposed estimator with respect to the prior confidence. However, it would be more desirable in these situations to use a full-fledged empirical Bayes approach where not only k but also θ and r are estimated from the data. The research in this direction is currently in progress by the auther.

References

[1] Abramowitz, M. and Stegun, I.(1965), *Handbook of Mathematical Functions*, New York: Dover Publications.

[2] Clayton, M.K. and Frees, E.W.(1987), Nonparametric Estimation of the Probability of Discovering a New Species, *Journal of the American Statistical Association*, Vol. 82, 305-311.

[3] Conte, S.D. and de Boor, C.(1972), *Elementary Numerical Anaysis*, New York: McGraw-Hill.

[4] Good, I.J.(1953), On the Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, Vol. 40, 237-264.

[5] Good, I.J.(1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge, Massachusetts: MIT Press.

[6] Gill, C.A. and Joanes, D.N.(1979), Bayesian Estimation of Shannon's Index of Diversity, *Biometrika*, Vol. 66, 81-85.

[7] Hill, B.M.(1968), Posterior Distribution of Percentiles: Bayes' Theorem for Sampling from a Population, *Journal of the American Statistical Association*, Vol. 63, 677 - 691.

[8] Hill, B.M.(1979), Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding a New Species, *Journal of the American Statistical Association*, Vol. 74, 668-673.

[9] IMSL STAT/LIBRARY User's Manual(1992), Sugar Land Texas: IMSL.

[10] Kempton, R.A. and Wedderburn, R.W.M.(1978), A Comparison of Three Measures of

Species Diversity, *Biometrics*, Vol. 34, 25-37.

[11] 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, *Unpublished PhD. Thesis*, Department of Statistics, University of Wisconsin, Madison.

[12] Lee, J.(1993), On Asymptotics for a Bias-Corrected Version of the NPMLE of the Probability of Discovering a New Species, *The Korean Journal of Applied Statistics*, Vol. 6, 341-353.

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

[14] Robbins, H.(1968), Estimating the Total Probability of the Unobserved Outcomes of an Experiment, *Annals of Mathematical Statistics*, Vol. 39, 256-257.

[15] Starr, N.(1979), Linear Estimation of the Probability of Discovering a New Species, *Annals of Statistics*, Vol. 7, 644-652.

신종발견확률의 경험적 베이지안 추정에 관한 연구[1]

이주호[2]

요약

여러개의 종으로 구성된 모집단으로부터 일정 크기의 표본을 추출한 경우 다음에 관측된 종이 신종일 확률에 대한 추정량으로 가장 널리 사용되어 온 것은 Good의 추정량이다. 본 논문에서는 종의 총 수효에 관한 사전정보가 존재할 경우 Good의 추정량에 대한 대안으로서 새로운 경험적 베이지안 추정량을 제안하였다. 모집단이 절단기하분포를 따를 경우의 소표본 시뮬레이션 결과는 새로운 추정량의 편의가 별로 크지 않으며 RMSE가 Good의 추정량보다 작음을 보여 주었다.

[1] 본 연구는 1992년도 학술진흥재단 자유공모과제 연구비에 의해 수행되었음.

[2] (305-764) 대전직할시 유성구 궁동 220 충남대학교 통계학과 조교수