ON A BAYESIAN APPROACH TO ESTIMATING CLASS NUMBER

STEPHEN MONTGOMERY-SMITH

Abstract. We examine a Bayesian approach to estimating the number of classes in a population, in the situation that we are able to take many independent samples from an infinite population.

Introduction. Suppose we capture 1000 birds from an effectively infinite population, and observe 15 different species. Estimate how many species have not been observed. While it is clear that one can only get a lower estimate (because there may be a very large number of extremely rare species which we have essentially no hope of observing), nevertheless if, say, 4 of the species appear only once, we would assume that most likely that there are species not observed, because if 15 were the exact number we would likely have not observed every one of them.

More generally, suppose that we have an observation $\mathcal{O}$ of n independent samples from an infinite population. Each sample belongs to one of M classes $C_1, C_2, \dots, C_M$, and the probability that a sample belongs to class C_i is p_i, where $0 \leq p_i \leq 1$ and $\sum_{i=1}^{M} p_i = 1$. The goal of this paper is to present a Bayesian method to estimate the class number, M. The difficulty is to estimate the number of classes that have not been observed.

A survey of various approaches is given in [2]. In particular, we point out a rather effective method due to A. Chao [1]. If m_k is the number of classes observed k times, then a lower estimator for the number of classes not observed is $m_1^2/2m_2$.

The Bayesian approach will be to treat M and $p = (p_1, \dots, p_M)$ as random variables with prior distributions. After computing the posteriors, we will integrate out p. This has already been explored, for example, in [3], where the final distribution is obtained using Monte-Carlo techniques. In this paper we present some prior distributions for which we are able to give explicit formulae for the final distributions.

The purpose of this paper is not to advocate for the Bayesian approach, but rather to identify that in this particular situation that the Bayesian approach seems to give rather poor quality answers. We will present a computer generated simulation in which our methods will significantly overestimate. In particular, the beta prior distributions will be shown to give significantly bland answers, because the final formulae do not depend upon the numbers m_k, but only upon $m = \sum_{k=1}^{n} m_k$, the total number of classes observed, and $n = \sum_{k=1}^{n} k m_k$, the sample size.

The posterior. Let $n_1, n_2, \dots, n_m$ denote the number of samples from each observed class, that is, a sequence in which each number k appears exactly m_k times.

Now we do not know which class is which, and so

$$\Pr(\mathcal{O}|(M,p)) = \sum_{\sigma \in B_{m,M}} \prod_{i=1}^{m} p_{\sigma(i)}^{n_i},$$

where $B_{m,M}$ denotes the set of one to one mappings from $\{1,2,\dots,m\}$ to $\{1,2,\dots,M\}$. (We mention that this formula assumes that we are also told the order in which the samples are picked. If we do not know the order in which they are picked, and we are simply given the numbers $n_1, n_2, \dots, n_m$, then there is an extra factor $n!/n_1! \cdots n_m! m_1! \cdots m_n!$. However this factor is independent of M and so it will disappear in the final normalization that takes place in Bayes' Theorem.)

The prior for p given M is a distribution μ_M on the simplex

$$\Delta_{M-1} = \left\{ p = (p_1, \dots, p_M) : p_i \geq 0, \sum_{i=1}^{M} p_i = 1 \right\},$$

which, since we do not know *a priori* which class is which, must be invariant under rearrangements of $p_1, \dots, p_M$. Thus

$$\Pr(\mathcal{O}|M) = \int_{\Delta_{M-1}} \Pr(\mathcal{O}|(M,p)\, d\mu_M(p) = \frac{M!}{(M-m)!} \int_{\Delta_{M-1}} \prod_{i=1}^{m} p_i^{n_i}\, d\mu_M(p).$$

Prior distributions from populations. The prior distributions of p that we will consider will be derived from the size of the population of each class, X_i, which are very large (that is, effectively infinite) random variables, but independent and identically distributed. We will assume that each X_i has a continuous probability distribution $\lambda\phi(\lambda x)\, dx$, where $\lambda > 0$ is very small. In the appendix we will show that

$$\Pr(\mathcal{O}|M) = \frac{M!}{(M-m)!(n-1)!} \int_{s=0}^{\infty} s^{-M-1} h_0(s)^{M-m} h_1(s)^{m_1} h_2(s)^{m_2} \cdots h_n(s)^{m_n}\, ds,$$

where

$$h_k(s) = \int_0^{\infty} x^k e^{-x} \phi(x/s)\, dx.$$

Beta prior distributions. This distribution arises when the population size of each class is given by the Gamma distribution, namely the distribution is $\phi(x) = x^a e^{-x}/\Gamma(a+1)$, where $a > -1$. Then $d\mu_M(p)$ is a beta distribution, that is, it is proportional to $(p_1 \cdots p_M)^a$ times Lebesgue measure on Δ_{M-1}. The case $a = 0$ is the uniform distribution on Δ_{M-1}. A direct calculation shows

$$\int_{\Delta_{M-1}} \prod_{i=1}^{M} p_i^{n_i}\, d\mu_M(p) = \frac{\Gamma(M + Ma + 1)}{\Gamma(M + Ma + n + 1)} \prod_{i=1}^{m} \frac{\Gamma(n_i + a + 1)}{\Gamma(a+1)}.$$

After applying Bayes' Theorem, and remembering the normalization that takes place, we find that

$$\Pr(M|\mathcal{O}) \propto \frac{M!\Gamma(M + Ma + 1)}{(M-m)!\Gamma(M + n + Ma + 1)} \Pr(M),$$

where the constant of proportionality is so that the probabilities on the left hand side add up to 1. We notice that the answer does not depend upon the values of $n_1, n_2, \dots, n_m$, but only upon the sample size n, and the number of classes observed m.

Uniform distribution on the sphere. Here we suppose that the population sizes of each class have a semi-Gaussian distribution, that is $\phi(x) = \sqrt{2/\pi}\, e^{-x^2/2}$. This gives rise to a distribution which is formed by projecting the normalized uniform distribution on the positive sphere

$$S_{M-1}^{+} = \left\{ q = (q_1, q_2, \ldots, q_M) : q_i \geq 0, \sum_{i=1}^{M} q_i^2 = 1 \right\}$$

onto Δ_{M-1} via the map $q \mapsto q/(q_1 + q_2 + \cdots + q_M)$. In this case we obtain

$$h_k(s) = \sqrt{\frac{2}{\pi}} \int_0^\infty x^k e^{-x} \exp(-x^2/2s^2)\, dx.$$

Although $h_k(s)$ can be computed explicitly as $(-1)^k s^{k+1}$ times the kth derivative of $\exp(s^2/2)\operatorname{erfc}(s/\sqrt{2})$, the resulting formula is numerically unstable for large s. Instead it is rather easily computed by a standard numerical integration package.

Uniform distribution on the rectangle. This next possibility is if the population of each class is uniformly distributed. In this case $h_k(s)$ is a lower incomplete Gamma function, which can easily be numerically computed.

A numerical simulation. A computer simulated a sample of $n = 500$ taken from a population of 100 classes, of which 50 of the classes occurred with a probability of 0.05, and the other 50 classes occurred with a probability of 0.15. The data obtained had $m_1 = 12$, $m_2 = 15$, $m_3 = 12$, $m_4 = 8$, $m_5 = 6$, $m_6 = 15$, $m_7 = 8$, $m_8 = 4$, $m_9 = 4$, $m_{10} = 8$, $m_{11} = 2$, $m_{14} = 1$, $m_{15} = 2$, and all other $m_k = 0$. Thus the number of observed classes is $m = 97$. The Chao estimate gives 3.2 for the number of unobserved classes, which is very close to the correct answer.

Using the improper prior distribution where $\Pr(M)$ is constant, then the beta with $a = 0$, uniform spherical, and uniform rectangular priors for p respectively give a final probability of $M \leq 100$ equal to 3.46×10^{-6}, 0.00028, and 0.0048, and a mid-90th percentile range for M of 121–131, 106–121, and 103–115. It can be seen that the uniform rectangular prior seems to work the best, but all of them significantly overestimate the class number.

Appendix: derivation of prior distribution from populations. We have that

$$\int_{\Delta_{M-1}} \prod_{i=1}^{M} p_i^{n_i}\, d\mu_M(p) = \int_{\mathbb{R}_+^M} \frac{x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m}}{(x_1 + \cdots + x_M)^n} \lambda^M \phi(\lambda x_1) \cdots \phi(\lambda x_M)\, dx.$$

Since

$$\int_{s=0}^\infty s^{n-1} e^{-sy}\, ds = \frac{(n-1)!}{y^n},$$

the right hand side may be seen to equal

$$\frac{\lambda^{M+n}}{(n-1)!} \int_{s=0}^\infty s^{n-1} \int_{\mathbb{R}_+^M} x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m} e^{-\lambda(x_1+\cdots+x_M)s} \phi(\lambda x_1) \cdots \phi(\lambda x_M)\, dx\, ds,$$

which after making the substitution $x_i \mapsto x_i/\lambda s$ can be written as

$$\frac{1}{(n-1)!} \int_{s=0}^\infty s^{-M-1} h_{n_1}(s) \cdots h_{n_m}(s) h_0(s)^{M-m}\, ds.$$

Acknowledgments. The author would like to extend his sincere thanks to Anthony Garrett introducing him to the Bayesian approach to this particular problem. He also extends his gratitude to Frank Schmidt for bringing the class number problem and reference [1] to his attention.

References

[1] Chao, A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11, 265-270.
[2] Chao, A. 2005. Species richness estimation, Pages 7909-7916 in N. Balakrishnan, C. B. Read, and B. Vidakovic, eds. Encyclopedia of Statistical Sciences. New York, Wiley.
[3] Rodrigues J., Milan L.A., Leite J.G. 2001. Hierarchical Bayesian estimation for the number of species. Biometrical J., 43, 737-746.

Math Dept, University of Missouri, Columbia, MO 65211, U.S.A.
E-mail address: stephen@math.missouri.edu