

Nonparametric Estimation of the Number of Classes in a Population

ANNE CHAO

National Tsing Hua University

ABSTRACT. Assume that a random sample is drawn from a population with an unknown number of classes. This work proposes a nonparametric method to estimate the number of classes when most of the information is concentrated on the low order occupancy numbers. The percentile method (Efron, 1981, 1982) is applied to construct confidence intervals based on bootstrap distributions. Using real data sets, we also compare the proposed point and interval estimates with previously published results.

Key words: number of classes, population size, occupancy number, jackknife estimator, percentile method

1. Introduction

Assume that there is an unknown number θ of different classes in a population. We search this population by selecting one member at a time, noting its class identity and returning it to the population. We also assume that the classes are indexed by $1, 2, \dots, \theta$. In general applications, the classes may be species of insects or different dies by which coins were produced in minting. For practical examples, see Goodman (1949), Efron & Thisted (1976), Burnham & Overton (1978, 1979) and Holst (1981). Suppose N selections have been made and p_j denotes the probability that a randomly selected member belongs to the j th class, $j=1, 2, \dots, \theta$, $\sum p_j=1$. Our interest is to estimate θ based on the occupancy numbers n_r , $r=1, 2, \dots, N$, where n_r denotes the number of classes observed exactly r times in the sample. This is a familiar problem in ecological studies.

If all θ classes are equally likely ($p_i=1/\theta$ for all i), the problem reduces to an inference problem involving only one parameter. In this case, traditional estimation procedures (e.g. maximum likelihood, minimum variance unbiased and Bayes) have been investigated by many authors including Lewontin & Prout (1956), Harris (1968), Samuel (1968), Johnson & Kotz (1977, pp. 136–139), and Marchand & Schroeck (1982). Holst (1981) further constructed a confidence interval of θ and provided a test for the equiprobability hypothesis. Esty (1982, 1983) obtained nonparametric confidence intervals for the sample coverage. Since the sample coverage in the equiprobable case is the number of observed classes divided by θ , his results will automatically produce confidence intervals of θ under the equiprobability assumption.

When the hypothesis of equiprobability is false or in doubt, most previous approaches were to adopt specific parametric models, see Fisher et al. (1943), McNeil (1973), Engen (1974, 1978), Efron & Thisted (1976) and many others.

Let d be the total number of classes seen in the sample. Applying the generalized jackknife technique to the naive estimator d , Burnham & Overton (1978, 1979) have developed nonparametric estimators under the assumption that the bias of d is expressible in a power series in N^{-1} . Based on the subsamples which at most k observations are deleted, one can compute the so called k th order jackknife estimator, which eliminates the

