

# 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  $\theta$  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  $\theta$  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  $\theta$  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  $\theta$  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  $\theta$ , his results will automatically produce confidence intervals of  $\theta$  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

$N^{-1}, N^{-2}, \dots, N^{-k}$  order terms from the bias. For large  $N$ , Burnham & Overton (1979, p. 935) showed that for our problem, the  $k$ th order jackknife estimator is

$$\hat{\theta}_k = d + \sum_{i=1}^k (-1)^{i+1} \binom{k}{i} n_i. \tag{1}$$

Burnham & Overton (1979) also provided a testing procedure to select an appropriate  $k$ .

In Section 2, we present another method to find a nonparametric estimator of  $\theta$ . This method is essentially useful when most of the information is concentrated on  $(d, n_1, n_2)$ . Efron (1981, 1982) introduced the percentile method using bootstrap distributions to set confidence intervals in general nonparametric situations, and justified it from various theoretical points of view. In this work, the percentile method is slightly modified to obtain confidence limits based on the proposed estimator. The performance of the estimator as applied to some real data examples is discussed in the final section.

### 2. A proposed method

The method is similar to that taken by Harris (1959). We first estimate  $En_0$ , the expected value of the number of unobserved classes. Harris (1959) proved that for  $r^2 = o(N)$ ,

$$En_r \sim \sum_{i=1}^{\theta} (Np_i)^r e^{-Np_i/r!}, \tag{2}$$

where the approximation is in the sense that for large  $N$  either both sides are negligible or the ratio of both sides tends to 1. Consider the following distribution function:

$$F(x) = \sum_{Np_i \leq x} (Np_i) e^{-Np_i} / \sum_{i=1}^{\theta} (Np_i) e^{-Np_i}. \tag{3}$$

This distribution was originally used by Harris (1959) and Cobb & Harris (1966) to approach other statistical problems. We find it can easily be employed to obtain estimators of  $En_0$ , as will be described. Note that from (2) and (3) we have

$$\begin{aligned} En_0 &\sim \sum_i e^{-Np_i} \\ &\sim (En_1) \int_0^N x^{-1} dF(x). \end{aligned} \tag{4}$$

The  $r$ th moment  $\mu_r$  of  $F(x)$  is given by

$$\begin{aligned} \mu_r &= \sum_{i=1}^{\theta} (Np_i)^{r+1} e^{-Np_i} / \sum_{i=1}^{\theta} (Np_i) e^{-Np_i} \\ &\sim (r+1)! En_{r+1} / En_1. \end{aligned} \tag{5}$$

Then we can regard

$$m_r = (r+1)! n_{r+1} / n_1$$

as an estimate of  $\mu_r$  whenever  $n_1 \neq 0$ .

Note that if the integrand  $x^{-1}$  in (4) is approximated by a polynomial

$$\sum_{i=1}^k (-1)^{i+1} \binom{k}{i} x^{i-1}/i!,$$

of degree  $k-1$ , it follows from (4) and (5) that

$$En_0 \sim (En_1) \sum_{i=1}^k (-1)^{i+1} \binom{k}{i} \mu_{i-1}/i! \sim \sum_{i=1}^k (-1)^{i+1} \binom{k}{i} En_i.$$

Replacing the expected value  $En_i$  by the observed  $n_i$ ,  $i=1, 2, \dots, k$ , we obtain exactly the jackknife estimator given in (1). Thus this approach also provides a justification of the use of the jackknife estimator.

Instead of approximating the integrand, the proposed procedure is to use the moment estimates to obtain an estimate  $\hat{F}(x)$  of the integrator  $F(x)$  and thus find an estimate of  $\theta$ :

$$d+n_1 \int x^{-1} d\hat{F}(x).$$

We are mainly interested in finding an estimator  $\hat{F}(x)$  of  $F(x)$  such that  $\hat{F}(x)$  has  $m_1, m_2$  as its first two moments. Assume that  $m_1$  and  $m_2$  are legitimate moments, that is,  $m_1$  and  $m_2$  satisfy  $m_2 > m_1^2$  and  $Nm_1 > m_2$ . Let  $C(m_1, m_2)$  denote the class of cumulative distribution functions in  $[0, N]$  with  $m_1$  and  $m_2$  as the first two moments. Following a theorem in Harris (1959), we have

$$\min_{F \in C} \int x^{-1} dF(x) = \int x^{-1} dG(x),$$

where

$$G(x) = \begin{cases} 0 & x < \frac{Nm_1 - m_2}{N - m_1} \\ \frac{(N - m_1)^2}{(N - m_1)^2 + (m_2 - m_1^2)} & \frac{Nm_1 - m_2}{N - m_1} \leq x < N, \\ 1 & N \leq x \end{cases}$$

Hence we obtain a lower bound  $\bar{\theta}_{\min}$  of  $\theta$ :

$$\begin{aligned} \bar{\theta}_{\min} &= d + \int x^{-1} dG(x) \\ &= d + \frac{n_1}{(N - m_1)^2 + m_2 - m_1^2} \left\{ \frac{(N - m_1)^3}{Nm_1 - m_2} + \frac{m_2 - m_1^2}{N} \right\}. \end{aligned}$$

As  $N \rightarrow \infty$ ,

$$\bar{\theta}_{\min} \rightarrow \bar{\theta} = d + n_1^2 / (2n_2). \tag{6}$$

Although  $\bar{\theta}$  is a lower bound, its performance as an estimator of  $\theta$ , especially when  $(d_1, n_1, n_2)$  carries most of the information, is encouraging, as will be shown in the next section.

In order to obtain a confidence interval of  $\theta$ , we first construct a ‘‘pseudo population’’ of

$\hat{\theta}$  cells:  $kn_1$  ( $k=\hat{\theta}/d$ ) cells with cell probabilities  $1/(kN)$ ,  $kn_2$  cells with probabilities  $2/(kN)$ , ... etc. Then Efron's percentile method (Efron 1981, 1982) is applied as follows:

(i) Draw a bootstrap sample of size  $N$  from the "pseudo population" and compute  $\hat{\theta}^*$  based on this sample.

(ii) Do step (i)  $B$  times and obtain  $B$  replications  $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B}$ .

(iii) Let  $H(t)=(\text{number of } \hat{\theta}^{*i} \leq t, i=1, 2, \dots, B)/B$  and define  $H^{-1}(\alpha)=\inf\{t: H(t) \geq \alpha\}$ ,  $0 < \alpha < 1$ .

The percentile method assigns  $[H^{-1}(\alpha/2), H^{-1}(1-\alpha/2)]$  as an approximate  $1-\alpha$  confidence interval of  $\theta$ . Its performance as applied to real data sets is generally satisfactory, as will be discussed in Section 3. Note that Efron also suggested the bias-corrected percentile method. It usually, in our problems, gives a wider interval, though there is some improvements in the coverage probabilities. Hence we will only report the results of the percentile method.

### 3. Numerical examples

*Example 1.* This interesting example was described in Holst (1981). Given the number of dies having produced  $r$  coins,  $r=1, 2, \dots$  in a hoard, the problem is to estimate the number of dies used in the minting process. We first discuss the reverse side: 204 coins were found in a hoard of ancient coins, 156 dies appeared once, 19 twice, 2 three times, and 1 four times, no die appeared more than four times. For this frequency sequence, as explained by Holst (1981), it is plausible to assume that all the classes are equally likely. He further obtained an estimate 731 of the number of classes and constructed a 95% confidence interval (537, 1106). The jackknife estimate is the sixth order value  $\hat{\theta}_6=854$  with a standard error 107.7, which yields an interval (643, 1065) assuming the normality is valid. Esty (1983) obtained a confidence interval for the sample coverage and thus gave an estimate 757 and the interval (559, 1171) under the restrictive equiprobability assumption. The estimate proposed in this work is  $\hat{\theta}=818$ .  $B=1000$  replications were then drawn from the following "pseudo population": 717 cells with probabilities  $1/938$ , 87 cells with probabilities  $2/938$ , 9 cells with probabilities  $3/938$ , and 5 cells with probabilities  $4/938$ . The percentile method gives an interval (522, 1218). All the results are comparable. Although our interval is wider, we will see, in Example 4, that the coverage probability of it is closer to the nominal level than that of the jackknife interval.

We now consider the data for the obverse side of the same coins. The first seven frequencies, in order, are 102, 26, 8, 2, 1, 1, 1, and the other frequencies are 0. Using a variance test, Holst (1981) concluded that the equiprobability assumption is not appropriate. Thus the results of Holst (1981) or Esty (1982, 1983) are insufficient to yield estimates. The resultant jackknife estimate is  $\hat{\theta}_4=423$  with a 95% interval of (316, 530). Our result would correspond to an estimate of 341 and the interval (237, 431), based on 1000 bootstrap replications.

*Example 2.* This is also application to the problem of determining the number of dies. Eddy (1967) (see also Esty, 1978, 1982, who analyzed this data) found among 662 ancient coins that only two pairs were struck from the same dies. Esty (1978) provided an estimate 110000 and a 95% lower confidence bound (Esty, 1982) 35000, assuming that each die produced about the same number of coins. Our estimate is  $\hat{\theta}=108900$  and a 95% lower bound is 43000 based on 1000 replications. Both results agree well. As indicated by Esty (1982), other results based on a normal limit law are not appropriate for this example.

*Example 3.* The problem of estimating the number of classes is equivalent to that of estimating the total number of individuals in capture-recapture studies (see Samuel, 1968).

The basic data in a capture-recapture study are the trapping histories of each individual at  $t$  trapping occasions. In this case,  $n_i$  becomes the number of individuals captured exactly  $i$  times for  $i=1, 2, \dots, t$ . These are the capture frequencies and  $d=\sum n_i$  is the total number of <sup>distinct</sup> captures. The form of the jackknife estimator developed by Burnham & Overton (1978, 1979) for multiple recapture studies is slightly different from that given in (1). The reader is referred to the preceding articles for details. It is clear that the proposed estimator  $\hat{\theta}$  in (6) can also be used as an estimator of population size provided that the total number of captures at each trapping occasion is large enough. Consider the data given in Edwards & Eberhardt (1967). They conducted an 18 livetrapping study on a penned population of 135 wild cottontail rabbits. (Note here that the true value  $\theta=135$  is known, which can be used as a basis for comparison with estimates.) Recorded capture frequencies were ( $n_1$  to  $n_7$ ) 43, 16, 8, 6, 0, 2, 1. Our estimate  $\hat{\theta}=134$  is quite precise in this case and 1000 bootstrap replications give (97, 168) as a 95% confidence interval for  $\theta$ . On the other hand, Burnham & Overton (1978, 1979) suggested to use the jackknife estimate  $\hat{\theta}_3=159$  with the interval of (116, 202). They further recommended an improved interpolated estimate 142 and the interval (112, 172), which is slightly narrower.

*Example 4.* Carothers (1973) conducted a capture-recapture study on the taxicab population of Edinburgh, which numbered 420. The reader may refer to Carothers (1973) for details of the population and of the method of sampling. The jackknife estimates and corresponding 95% confidence intervals are tabulated in Table 1, for part of the collected data subsets only. The sampling schemes and data subsets used in Table 1 are the ones Carothers identifies in his paper. Our point and interval estimates based on 1000 bootstrap replications are also given in Table 1 for comparison. Generally, our estimates are preferable to the jackknife estimates in the sense of being closer to the true value 420. Note that only one out of 14 jackknife confidence intervals covers the true value; in contrast, all except one of our intervals cover the true value, although the intervals are wider.

We finally remark that if  $(n_3, n_4, \dots)$  carries nonnegligible information, the proposed estimates usually underestimate the true parameter (as we discussed before,  $\hat{\theta}$  is actually a lower bound). Other estimators following this approach are still under investigation.

Table 1. Comparison of estimates applied to Carothers' (1973) data (true value  $\theta=420$ )

Sampling scheme	Data subset	Jackknife estimate	95% conf. interval	Proposed estimate	95% conf. interval
A	$\alpha$ a	192	(155, 229)	253	(147, 475)
	b	217	(176, 258)	414	(230, 885)
	c	223	(182, 264)	484	(247, 1207)
	d	325	(274, 376)	384	(251, 540)
	e	332	(281, 383)	366	(250, 513)
	f	350	(297, 403)	430	(275, 616)
	g	407	(350, 464)	404	(283, 495)
B	$\alpha$ a	233	(190, 276)	691	(344, 1808)
	b	199	(160, 238)	325	(183, 726)
	c	213	(172, 254)	439	(226, 1123)
	d	333	(282, 384)	421	(272, 633)
	e	315	(266, 364)	338	(227, 471)
	f	303	(250, 356)	331	(216, 465)
	g	346	(307, 385)	312	(224, 380)