Rev.Acad.Canar.Cienc., IX (Núm. 1), 97-124 (1997)

AN ESTIMATOR FOR THE NUMBER OF CLUSTERS IN A POPULATION

JUAN JOSÉ PRIETO MARTÍNEZ
University Carlos III of Madrid. Department of Statistics and Econometrics. Calle Madrid, 126
28093 - Getafe - MADRID. SPAIN

Abstract:

Assume that a random sample is drawn from a population with an unknown number K of clusters. This work proposes a nonparametric method to estimate the number of clusters when most of the information is concentrated on the low order occupancy numbers. This paper derives an estimator to K and proves the asymptotic distribution using a method of Holst (1979). The performance of the estimator is investigated by means of Monte Carlo experiments and it is applied to one real data example.

Key words: number of clusters,
method of Holst,
asymptotic normality.

1./ Introduction:

Assume that there is an unknown number K of different clusters in a population. We search this population by selecting one member at a time, noting its class identity and returning it to the population. Suppose n selections have been made and p_j denotes the probability that a ramdomly selected member belongs to the jth cluster, $j=1,\ldots,K$, $\sum p_j=1$. If $p_j=1/K$, $\forall$ $j=1,\ldots,K$ (the equally likely or equiprobable assumption), the problem reduces to an inference problem involving only one parameter. See, for example Lewontin & Prout (1956), Darroch (1958), McNeill (1973), Johnson & Kotz (1977), Harris (1968), Host (1981) and Marchand & Schroeck (1982).

Probably, in most practical applications, the equally likely asumption is not valid. For instance, the insects in a forest classified by species, the words in a computer file classified by precise letter sequence, or a archeological artifacts classified by type. Most authors adopted a parametric approach to handle heterogeneous populations (i.e. unequal clusters probabilities). For example, Fisher, Corbet and William (1943), assumed that for each species the number of elements observed in the sample follows a Poisson distribution and the Poisson parameter is assumed to have a gamma-type distribution. Many others papers on stochastic abundance models also make parametric assumption; see, for example Engen (1978) for a review.

The sample coverage of a random sample from a multinomial population is defined to be the sum of the probabilities of the observed clusters. For an equiprobable population, the estimator proposed by Darroch and Ratcliff (1980) exactly used the idea of sample coverage. For heterogeneous populations, Esty (1985) was the first to apply the concept of sample coverage to estimate the number of clusters in a parametric setup. The clusters discussed by Esty are the different dies in minting. The assumed that the number of coins that each die produced follows a negative binomial distribution and obtained an estimator of the number of dies in

terms of the sample coverage and the parameter of the negative binomial distribution. A nonparametric estimation technique is proposed by Chao (1992) to estimate the number of clusters using the idea of sample coverage. She generalizes the result of Esty (1985) to a nonparametric approach and extends Darroch and Ratcliff (1980) to incorporate the heterogeneity of the clusters probabilities.
The previous authors do not proves the asymptotic distribution of the estimator proposed.

This work just porposes a nonparametric method to estimate the number of clusters when most of the information is concentrated on (D, N_1, N_2), where D is the total number of clusters observed in the sample, and N_i is the number of clusters observed exactly i times in the sample. See section 2.

In the section 3, the asymptotic normality of the estimator is proved by applying a result of Holst (1979). In section 4, the results of a simulation study to investigate the performance of the estimator is showed and also the bound is applied to one real data example.

2./ An estimator for K.

Let C_0, C_1, $C_2,\ldots,C_K$ be the clusters in a population. Suppose n selections have been made and p_j denotes the probability that a ramdomly selected member belongs to the jth cluster, $j=0,1,\ldots,K$; $\sum_{j=0}^{K} p_j=1$.

Let $\varepsilon_{ij}=I$(the ith observation belongs to the jth cluster) where $I(A)$ is the indicator function, $i=1,\ldots,n$; $j=0,1,\ldots,K$. Let $X_j=\sum_{i=1}^{n}\varepsilon_{ij}$ be the number of observations belonging to the jth cluster, then $(X_0, X_1,\ldots,X_K)$ is distributed as a multinomial distribution with parameter $(n; p_0, p_1,\ldots,p_K)$. Let $N_i=\sum_{j=1}^{K} I(X_j=i)$, $i=0,1,\ldots,n$, be the number of clusters with i representatives in the sample and $D=\sum_{j=1}^{K} I(X_j>0)$ be the number of observed distinct clusters.

Define

$$Z_{j,i}=\begin{cases}1 \text{ if the cluster j is observed i times in the sample,}\\ 0 \text{ otherwise.}\end{cases}$$

Hence,

$$E(N_i)=E\left(\sum_{j=1}^{K} Z_{j,i}\right)=\sum_{j=1}^{K}\text{Prob}(Z_{j,i}=1)=\sum_{j=1}^{K}\binom{n}{i}p_j^i(1-p_j)^{n-i},$$

$$\forall i=1,2,\ldots,n.$$

In particular,

$$E(N_0)=\sum_{j=1}^{K}(1-p_j)^n, \qquad (2.1)$$

$$E(N_1)=\sum_{j=1}^{K} np_j(1-p_j)^{n-1} \qquad (2.2)$$

and

$$E(N_2)=\sum_{j=1}^{K}\binom{n}{2}p_j^2(1-p_j)^{n-2}. \qquad (2.3)$$

It follows from the Cauchy-Schartz inequality that:

$$\left(\sum_{j=1}^{K}p_j(1-p_j)^{n-1}\right)^2=\left(\sum_{j=1}^{K}[p_j\ (1-p_j)^{n/2-1}(1-p_j)^{n/2}]\right)^2\leq$$

$$\leq\sum_{j=1}^{K}(p_j(1-p_j)^{n/2-1})^2\sum_{j=1}^{K}((1-p_j)^{n/2})^2=\sum_{j=1}^{K}p_j^2(1-p_j)^{n-2}\sum_{=1}^{K}(1-p_j)^n.$$

This is,

$$\left(\sum_{j=1}^{K}p_j(1-p_j)^{n-1}\right)^2\leq\sum_{j=1}^{K}p_j^2(1-p_j)^{n-2}\sum_{j=1}^{K}(1-p_j)^n\ ,$$

that is equivalent to:

$$\left(\sum_{j=1}^{K}p_j(1-p_j)^{n-1}\right)^2\left(\sum_{j=1}^{K}p_j^2(1-p_j)^{n-2}\right)^{-1}\leq\sum_{j=1}^{K}(1-p_j)^n,$$

that is equal to:

$$\sum_{j=1}^{K}(1-p_j)^n\geq\frac{\left(\sum_{j=1}^{K}np_j(1-p_j)^{n-1}\right)^2}{n^2}\ \frac{n(n-1)}{\sum_{j=1}^{K}n(n-1)p_j^2(1-p_j)^{n-2}}\ .$$

Combining (2.1), (2.2) and (2.3):

$$E(N_0)\geq\frac{(E(N_1))^2(n-1)}{nE(N_2)}\ .$$

Note that $K=N_0+D$ (the number of clusters in a population is equal the number of clusters with 0 representatives in the sample and the number of clusters observated). Thus a lower bound of K is:

$$\underline{K} \geq E(D) + \frac{n-1}{n} \frac{E(N_1)^2}{E(N_2)} .$$

Replacing $E(D)$, $E(N_1)$ y $E(N_2)$ by the observed value, an estimator if $n_2>0$ is:

$$\hat{\underline{K}} = d + \frac{n-1}{n} \frac{n_1^2}{n_2} .$$

Note that if $n \longrightarrow \infty$,

$$\hat{\underline{K}} \cong d + \frac{n_1^2}{n_2} .$$

3./ A normal limit law.

A limit distribution is rigorously proved for $\hat{\underline{K}}$ using a method of Holst (1979). It is shown the characteristic function of

$$K^{-1/2}(\hat{\underline{K}} - \underline{K})/\sigma^2 .$$

converges in distribution to a standard normal, where σ^2 is specified in the proof.
Let the hipótesis:

1. $\sum_{j=1}^{K} K^{-1} n^{1/2} p_j \longrightarrow 0.$

$$2.\ \sum_{j=1}^{K} K^{-1/2} n\ p_j^2 = n^{-1}K^{-1/2} \sum_{j=1}^{K} (np_j)^2 \longrightarrow 0.$$

$$3.\ \sum_{j=1}^{K} K^{-1/2} n^{3/2} p_j^3 = n^{-3/2}\ K^{-1/2} \sum_{j=1}^{K} (np_j)^3 \longrightarrow 0. \qquad (3.1)$$

$$4.\ \sum_{j=1}^{K} O(K^{-3/2}) \longrightarrow 0.$$

Let:

$$\hat{\underline{K}} - \underline{K} = \Big(\sum_{j=0}^{K} - (I(X_j=0) + (1-p_j)^n)\Big) + \frac{n-1}{n} \frac{E(N_1)}{E(N_2)} \Big(\sum_{j=0}^{K} (I(X_j=1) - np_j(1-p_j)^{n-1})\Big) + \frac{-(n-1)}{n} \frac{E^2(N_1)}{E^2(N_2)} \Big(\sum_{j=0}^{K} (I(X_j=2) - \binom{n}{2} p_j^2 (1-p_j)^{n-2})\Big),$$

where I(A) is the usual indicator function, using Taylor series expansion. Taylor series:

$$f(x_1, x_2, \ldots, x_n) = f(x_{1_0}, x_{2_0}, \ldots, x_{n_0}) + \left.\frac{\delta f}{\delta x_i}\right|_{(x_{1_0}, x_{2_0}, \ldots, x_{n_0})} + R.$$

Now regard $\hat{\underline{K}}$ as a function of $I(X_j=0)$, $I(X_j=1)$ and $I(X_j=2)$, $j=1,2,\ldots,K$. Expand $\hat{\underline{K}}$ in the point:

$$[(1-p_j)^n,\ np_j(1-p_j)^{n-1},\ \binom{n}{2} p_j^2 (1-p_j)^{n-2}]$$

Now the problem is to find the asymptotic distribution of $\sum_{j=1}^{K} f(X_j)$

and show that the characteristic function $\Psi(s)$ of $K^{-1/2}\sum_{j=1}^{K} f(X_j)$ converges to that of a normal distribution, where

$$f(X_j) = -(I(X_j=0) - (1-p_j)^n) + \frac{n-1}{n}\,\frac{E(N_1)}{E(N_2)}(I(X_j=1) - np_j(1-p_j)^{n-1}) +$$

$$+ \frac{-(n-1)}{n}\,\frac{E^2(N_1)}{E^2(N_2)}\,(I(X_j=2) - \binom{n}{2}p_j^2(1-p_j)^{n-2})$$

First note that:

$$P(X_0=x_0, X_1=x_1, \ldots X_K=x_K) = P(Y_0=y_0, Y_1=y_1, \ldots, Y_K=y_K / \sum_{j=0}^{K} Y_j = n),$$

where $\{Y_j\}$ are independent Poisson random variable with mean np_j.

Its proof is:

$$P(Y_0=y_0, Y_1=y_1, \ldots, Y_K=y_K / \sum_{j=0}^{K} Y_j = n) = \frac{P(Y_1=x_1, \ldots, Y_K=x_K, \sum_{j}^{K} Y_j = n)}{P(\sum_{j=1}^{K} Y_j = n)} =$$

$$= \frac{P(Y_1=x_1, \ldots, Y_K=x_K)}{P(\sum_{j=1}^{K} Y_j = n)} = \frac{[(np_j)^{x_1}/x_1!]e^{-np_1} \ldots [(np_j)^{x_K}x_K!]e^{-np_K}}{[(Knp_j)/n!]e^{-Knp_j}} =$$

$$= \frac{n!}{x_1! \ldots x_n!}\left(\frac{1}{K}\right)^n = P(X_0=x_0, X_1=x_1, \ldots X_K=x_K).$$

Hence,

$$\Psi(s)=E\left\{e^{isK^{-1/2}\sum_{j=1}^{K}f(X_j)}\right\}=E\left\{e^{isK^{-1/2}\sum_{j=1}^{K}f(Y_j)}\Big/\sum_{j=0}^{K}Y_j=n\right\}.$$

It then follows from Lemma 2 of Holst (1979):

"Let (U,V) be a two dimensional random vector with U integer valued. Then

$$E(e^{ivV}/U=n)=\frac{1}{2\pi P(U=n)}\int_{-\pi}^{\pi}E(e^{iu(U-n)+ivV})\,du)\,."$$

that:

$$\psi(s)=\frac{1}{2\pi(\sum_{j=0}^{K}Y_j=n)}\int_{-\pi}^{\pi}E(e^{iu\sum_{j=0}^{K}(Y_j-np_j)+isK^{-1/2}\sum_{j=1}^{K}f(Y_j)})\,du.$$

Since,

$$E(\sum_{j=0}^{K}Y_j)=\sum_{j=0}^{K}E(Y_j)=\sum_{j=0}^{K}np_j=n \text{ and } n!=e^{-n}\sqrt{2\pi n}\,n^n,$$

$$P(\sum_{j=0}^{n}Y_j=n)=e^{-n}\frac{n^n}{e^{-n}\sqrt{2\pi n}\,n^n}=\frac{1}{\sqrt{2\pi n}}\,.$$

Let $t=u\sqrt{n}$ to obtain

$$\Psi(s)=\frac{1}{\frac{1}{2\pi}\sqrt{2\pi n}}\int_{-\pi n^{1/2}}^{\pi n^{1/2}}E(e^{itn^{-1/2}\sum_{j=0}^{K}(Y_j-np_j)+isK^{-1/2}\sum_{j=1}^{K}f(Y_j)})\,n^{-1/2}dt=$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} E\left(e^{itn^{-1/2}\sum_{j=0}^{K}(Y_j - np_j) + isK^{-1/2}\sum_{j=1}^{K} f(Y_j)}\right) dt .$$

Define:

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} h_{1n}(s,t) h_{2n}(t)\, dt ,$$

where

$$h_{1n}(s,t) = \prod_{j=1}^{K} E\left(e^{itn^{-1/2}(Y_j - np_j) + isK^{-1/2} f(Y_j)}\right)$$

and

$$h_{2n}(t) = E\left(e^{itn^{-1/2}(Y_0 - np_0)}\right) .$$

Note that Y_0 for given n is a Poisson random variable,

$$h_{2n}(t) = E\left(e^{itn^{-1/2}(Y_0 - np_0)}\right) = E\left(e^{itn^{-1/2}Y_0} e^{-itn^{-1/2}np_0}\right) =$$

$$= e^{-itn^{-1/2}np_0} \sum_{j=0}^{\infty} e^{itn^{-1/2}j} e^{-np_0} \frac{(np_0)^j}{j!} =$$

$$= e^{-itn^{-1/2}np_0} \sum_{j=0}^{\infty} \frac{(e^{itn^{-1/2}} np_0)^j}{j!} e^{-np_0} =$$

$$e^{-itn^{1/2}p_0} e^{np_0 e^{itn^{-1/2}}} e^{-np_0} = e^{-itn^{1/2}p_0} e^{np_0(e^{itn^{-1/2}} - 1)} =$$

$$= e^{-itn^{1/2}p_0} e^{np_0([1+(it/n^{1/2})-(t^2/2n)+O(t^2)]-1)} =$$

$$= e^{-itn^{1/2}p_0} e^{np_0(it/n^{1/2})-np_0(t^2/2n)+O(t^2)} =$$

$$= e^{-itn^{1/2}p_0+p_0itn^{1/2}-np_0(t^2/2n)+O(t^2)} = e^{(-t^2p_0/2)+O(t^2)}$$

Considerer the factors of $h_{1n}(s,t)$:

$$h_{1n}(s,t) = \prod_{j=1}^{K} g_j(s,t),$$

where $g_j(s,t) = E(e^{it(Y_j-np_j)n^{-1/2}+isK^{-1/2}f(Y_j)})$.

Now,

$$E(e^{it(Y_j-np_j)n^{-1/2}+isK^{-1/2}f(Y_j)}) =$$

$$= E(e^{it(Y_j-np_j)n^{-1/2}+isK^{-1/2}(A_1I(Y_j=o)+A_2I(Y_j=1)+A_3I(Y_j=2))}) \times \quad (A)$$

$$\times e^{-isK^{-1/2}(A_1(1-p_j)^n+A_2np_j(1-p_j)^{n-1}+A_3\binom{n}{2}p_j^2(1-p_j)^{n-2})}, \quad (B)$$

where

$$A_1=-1,\ A_2=\frac{n-1}{n}\ \frac{E(N_1)}{E(N_2)}\ \text{y}\ A_3=-\ \frac{n-1}{n}\ \frac{E^2(N_1)}{E^2(N_2)}\ .$$

Part A involves expectation:

$$e^{-itn^{1/2}p_j+isK^{-1/2}A_1}e^{-np_j}+e^{it(1-np_j)n^{-1/2}+isK^{-1/2}A_2}np_je^{-np_j}+$$

$$+e^{-it(2-np_j)n^{-1/2}+isA_3K^{-1/2}}\frac{(np_j)^2}{2}e^{-np_j}+$$

$$+\sum_{R=3}^{\infty}e^{itn^{-1/2}(R-np_j)}\frac{(np_j)^R}{R!}e^{-np_j}.$$

Since $\{Y_j\}$ are independent Poisson random variable with mean np_j,

$$e^{-np_j}e^{-itn^{1/2}p_j}\times$$

$$\times\left\{[e^{isK^{-1/2}A_1}-1]+np_je^{itn^{-1/2}}[e^{isK^{-1/2}A_2}-1]+\right. \qquad (C)$$

$$\left.+\frac{(np_j)^2}{2}e^{2itn^{-1/2}}[e^{isK^{-1/2}A_3}-1]\right\}+$$

$$+e^{-itn^{1/2}p_j}e^{np_j(e^{itn^{-1/2}}-1)} \qquad (D).$$

Note that (D) is:

$$e^{-itn^{1/2}p_j}e^{np_j(e^{itn^{-1/2}}-1)}=$$

$$=e^{-itn^{1/2}p_j}e^{np_j(itn^{-1/2}-(t^2/2n)+O(n^{-3/2}t^3)}=$$

$$=e^{-p_j(t^2/2)+O(n^{-1/2}p_jt^3)}=e^{-p_jt^2/2}(1+O(n^{-1/2}t^3p_j)).$$

The expression (C) is:

$$e^{-np_j}e^{-itn^{1/2}p_j}\left\{[e^{isK^{-1/2}A_1}-1]+np_je^{itn^{-1/2}}[e^{isK^{-1/2}A_2}-1]+\right.$$
$$\left.+\frac{(np_j)^2}{2}e^{2itn^{-1/2}}[e^{isK^{-1/2}A_3}-1]\right\}.$$

Now, expanding:

$$e^{-itn^{1/2}p_j}\left\{[e^{isK^{-1/2}A_1}-1]+np_je^{itn^{-1/2}}[e^{isK^{-1/2}A_2}-1]+\right.$$
$$\left.+\frac{(np_j)^2}{2}e^{2itn^{-1/2}}[e^{isK^{-1/2}A_3}-1]\right\}=$$

$$=(1-itn^{1/2}p_j+\frac{(-1)t^2np_j^2}{2}+O(n^{3/2}p_j^3))\times$$
$$\times\left\{[isK^{-1/2}A_1-\frac{s^2A_1^2}{2K}+O(K^{-3/2})]+np_j[1+\frac{it}{n^{1/2}}+O(n^{-1})]\right.$$
$$[isK^{-1/2}A_2-\frac{s_2A_2^2}{2K}+O(K^{-3/2})]+$$
$$\left.+\frac{(np_j^2)}{2}[1+\frac{2ti}{n^{1/2}}+O(n^{-1})][isK^{-1/2}A_3-\frac{s^2A_3^2}{2K}+O(K^{-3/2})\right\}=$$

$$=(1-itn^{1/2}p_j-\frac{t^2np_j^2}{2}+O(n^{3/2}p_j^3)\left\{[isK^{-1/2}A_1-\frac{s^2A_1^2}{2K}+O(K^{-3/2})]+\right.$$
$$+np_j[isK^{-1/2}A_2-\frac{s^2A_2^2}{2K}-\frac{stA_2}{n^{1/2}K^{1/2}}+O(K^{-3/2})]+$$
$$\left.+\frac{(np_j)^2}{2}[isK^{-1/2}A_3-\frac{s^2A_3^2}{2K}-\frac{2stA_3}{n^{1/2}K^{1/2}}+O(K^{-3/2})]\right\}=$$

$$= (1 - itn^{1/2}p_j - \frac{np_j t^2}{2} + O(n^{3/2}p_j^3)) \left\{ isK^{-1/2}(A_1 + np_j A_2 + \frac{(np_j)^2}{2} A_3) - \right.$$

$$- \frac{s^2}{2K}(A_1^2 + np_j A_2^2 + \frac{(np_j)^2}{2} A_3^2) -$$

$$\left. - \frac{ts}{n^{1/2}K^{1/2}}(np_j A_2 + \frac{(np_j)^2}{2} A_3) + O(K^{-3/2}) \right\} =$$

$$= isK^{-1/2}(A_1 + np_j A_2 + \frac{(np_j)^2}{2} A_3) - \frac{s^2}{2K}(A_1^2 + np_j A_2^2 + \frac{(np_j)^2}{2} A_3^2) -$$

$$- \frac{2ts}{n^{1/2}K^{1/2}}(np_j A_2 + \frac{(np_j)^2}{2} A_3) + O(K^{-3/2}) +$$

$$+ \frac{ts(np_j)}{n^{1/2}K^{1/2}}(A_1 + np_j A_2 + \frac{(np_j)^2}{2} A_3) +$$

$$+ O(K^{-1}n^{1/2}p_j^2) + O(K^{-1/2}np_j^2) + O(K^{-1/2}n^{3/2}p_j^3) .$$

Under assumption (3.1), this expression is equal to:

$$isK^{-1/2}(A_1 + np_j A_2 + \frac{(np_j)^2}{2} A_3) - \frac{s^2}{2K}(A_1^2 + np_j A_2^2 + \frac{(np_j)^2}{2} A_3^2) +$$

$$+ \frac{st}{n^{1/2}K^{1/2}}(np_j A_1 + np_j A_2(np_j - 1) + \frac{(np_j)^2}{2}(np_j - 2)A_3) + O(K^{-1}) .$$

Hence:

$$\prod_{j=1}^{K} g_j(s,t) = \prod_{j=1}^{K} e^{-p_j t^2/2} \left\{ [1 + O(n^{-1/2}t^3 p_j)][1 + O(p_j t^2)] \times \right.$$

$$\times e^{-np_j}\left[isk^{-1/2}\left(A_1+np_jA_2+\frac{(np_j)^2}{2}A_3\right)-\frac{s^2}{2K}\left(A_1^2+np_jA_2^2+\frac{(np_j)^2}{2}A_3^2\right)+\right.$$

$$\left.+\frac{ts}{n^{1/2}K^{1/2}}\left(np_jA_1+np_jA_2(np_j-1)+-\frac{(np_j)^2}{2}(np_j-2)A_3\right)+O(K^{-1})\right]\Bigg\}\times$$

$$\times e^{-isK^{-1/2}\left(A_1(1-p_j)^n+A_2(1-p_j)^{n-1}np_j+A_3\binom{n}{2}(1-p_j)^{n-2}p_j^2\right)}=$$

$$=\prod_{j=1}^{K} e^{-p_jt^2/2}\Bigg\{\left(1+O(n^{-1/2}t^3p_j)+(1+O(p_jt^2)e^{-np_j}\times\right.$$

$$\times\left[isK^{-1/2}\left(A_1+np_jA_2+\frac{(np_j)^2}{2}A_3\right)-\frac{s^2}{2K}\left(A_1^2+np_jA_2^2+\frac{(np_j)^2}{2}A_3^2\right)+\right.$$

$$\left.\left.+\frac{ts}{n^{1/2}K^{1/2}}\left(np_jA_1+np_jA_2(np_j-1)+\frac{(np_j)^2}{2}(np_j-2)A_3\right)+O(K^{-1})\right]\Bigg\}\times\right.$$

$$\times\left\{1-isK^{-1/2}\left[A_1e^{-np_j}+A_2np_je^{-np_j}+A_3\frac{(np_j)^2}{2}e^{-np_j}\right]-\right.$$

$$\left.-\frac{s^2}{2K}\left[A_1e^{-np_j}+A_2np_je^{-np_j}+A_3\frac{(np_j)^2}{2}e^{-np_j}\right]^2+O(K^{-1})\right\}=$$

$$=\prod_{j=1}^{K} e^{-p_jt^2/2}\Bigg\{1-isK^{-1/2}\left[A_1e^{-np_j}+A_2np_je^{-np_j}+A_3\frac{(np_j)^2}{2}e^{-np_j}\right]-$$

$$-\frac{s^2}{2K}\left[A_1e^{-np_j}+A_2np_je^{-np_j}+A_3\frac{(np_j)^2}{2}e^{-np_j}\right]^2+$$

$$+isK^{-1/2}\left[A_1e^{-np_j}+A_2np_je^{-np_j}+A_3\frac{(np_j)^2}{2}e^{-np_j}\right]-$$

$$-\frac{s^2}{2K}e^{-np_j}\left(A_1^2+np_jA_2^2+\frac{(np_j)^2}{2}A_3^2\right)+$$

$$: \quad + \frac{ts}{n^{1/2}K^{1/2}} e^{-np_j}\left(A_1 np_j + A_2 np_j (np_j - 1) + A_3 \frac{(np_j)^2}{2}(np_j - 2)\right) +$$

$$+ \frac{s^2}{K}\left[A_1 e^{-np_j} + A_2 np_j e^{-np_j} + A_3 \frac{(np_j)^2}{2} e^{-np_j}\right]^2 + O(K^{-1})\Bigg\} =$$

$$= \prod_{j=1}^{K} e^{-p_j t^2/2}\Bigg\{1 - \frac{s^2}{2K}\Bigg\{e^{-np_j}\left(A_1 + np_j A_2^2 + A_3^2 \frac{(np_j)^2}{2}\right) -$$

$$- \left[A_1 e^{-np_j} + A_2 np_j e^{-np_j} + A_3 \frac{(np_j)^2}{2} e^{-np_j}\right]^2\Bigg\} +$$

$$+ \frac{ts}{n^{1/2}K^{1/2}} e^{-np_j}\left(A_1 np_j + A_2 np_j (np_j - 1) + A_3 \frac{(np_j)^2}{2}(np_j - 2)\right) + O(K^{-1}) =$$

$$= \prod_{j=1}^{K} e^{-p_j t^2/2} \prod_{j=1}^{K}\Bigg\{1 - \frac{s^2}{2K}\Bigg\{e^{-np_j}\left(A_1^2 + A_2^2 np_j + A_3^2 \frac{(np_j)^2}{2}\right) -$$

$$- \left[A_1 e^{-np_j} + A_2 e^{-np_j} np_j + A_3 \frac{(np_j)^2}{2} e^{-np_j}\right]^2\Bigg\} +$$

$$+ \frac{ts}{n^{1/2}K^{1/2}} e^{-np_j}\left(A_1 np_j + A_2 p_j (np_j - 1) + A_3 \frac{(np_j)^2}{2}(np_j - 2) + O(K^{-1})\right)\Bigg\}.$$

Then,

$$g_j(s,t) = e^{-(1-p_0)t^2/2} \exp\Bigg\{-\frac{s^2}{2K}\Bigg\{\sum_{j=1}^{K}\Bigg\{e^{-np_j}\left(A_1^2 + A_2^2 np_j + A_3 \frac{(np_j)^2}{2}\right) -$$

$$- \left[A_1 e^{-np_j} + A_2 np_j e^{-np_j} + A_3 \frac{(np_j)^2}{2} e^{-np_j}\right]^2\Bigg\} +$$

$$+ \frac{ts}{n^{1/2}K^{1/2}} e^{-np_j}\left(A_1 np_j + A_2 np_j (np_j - 1) + A_3 \frac{(np_j)^2}{2}(np_j - 2)\right)\Bigg\},$$

and

$$\prod_{j=1}^{K} g_j(s,t)\,h_{2n}(t) = e^{-(1-p_0)t^2/2 - p_0 t^2/2} \times$$

$$\times \exp\left\{ \frac{ts}{n^{1/2}K^{1/2}} \left\{ \sum_{j=1}^{K} e^{-np_j} \left[A_1 np_j + A_2 np_j (np_j - 1) + A_3 \frac{(np_j)^2}{2} (np_j - 2) \right] \right\} \right\}$$

$$\exp\left\{ -\frac{s^2}{2K} \left\{ \sum_{j=1}^{K} \left[A_1^2 e^{-np_j} + A_2^2 np_j e^{-np_j} + A_3^2 \frac{(np_j)^2}{2} e^{-np_j} - \right.\right.\right.$$

$$\left.\left.\left. - e^{-2np_j} \left[A_1 + A_2 np_j + A_3 \frac{(np_j)^2}{2} \right]^2 \right\} \right\} =$$

$$= \exp\left\{ -\frac{t^2}{2} + \frac{ts}{n^{1/2}K^{1/2}} \right.$$

$$\left. \left\{ \sum_{j=1}^{K} e^{-np_j} \left[A_1 np_j + A_2 np_j (1 - np_j) + A_3 \frac{(np_j)^2}{2} (np_j - 2) \right] \right\} \right\} \times$$

$$\times \exp\left\{ -\frac{s^2}{2K} \left\{ \sum_{j=1}^{K} A_1^2 e^{-np_j} + A_2^2 np_j e^{-np_j} + A_3 \frac{(np_j)^2}{2} e^{-np_j} - \right.\right.$$

$$\left.\left. e^{-2np_j} \left[A_1 + A_2 np_j + A_3 \frac{(np_j)^2}{2} \right]^2 \right\} \right\}.$$

Then,

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\pi n^{1/2}}^{+\pi n^{1/2}} e^{(-t^2/2)} \exp\left\{ (ts/(nK)^{1/2}) \sum_{j=1}^{K} \beta(j) - (s^2/2K) \sum_{j=1}^{K} \alpha(j) \right\} dt =$$

$$\vdots$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\pi n^{1/2}}^{+\pi n^{1/2}} \exp\left\{ (-1/2)\left(t - \frac{\sum_{j=1}^{K} \beta(j)}{(nK)^{1/2}} s\right)^2 \right\} dt \times$$

$$\times \exp\left\{ (s^2/2nK) \left(\sum_{j=1}^{K} \beta(j) \right)^2 \right\} \exp\left\{ (-s^2/2K) \sum_{j=1}^{K} \alpha(j) \right\},$$

where

$$\alpha(j)=e^{-np_j}\left(A_1^2+A_2^2np_j+A_3\frac{(np_j)^2}{2}\right)-e^{-2np_j}\left(A_1+A_2np_j+A_3\frac{(np_j)^2}{2}\right)^2$$

and

$$\beta(j)=e^{-np_j}\left(A_1np_j+A_2np_j(np_j-1)+A_3\frac{(np_j)^2}{2}(np-2)\right).$$

If $n, K\longrightarrow\infty$, the limit of $H_n(s)$ is:

$$\lim_{n,K\to\infty}H_n(s)=\lim_{n,K\to\infty}\frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}e^{-t^2/2}\exp\left\{(-1/2)\left(t-\frac{\sum_{j=1}^{K}\beta(j)}{(nK)^{1/2}}\right)^2\right\}dt\times$$

$$\times\lim_{n,K\to\infty}\left\{\exp\left\{(s^2/2nK)\left(\sum_{j=1}^{K}\beta(j)\right)^2\right\}\exp\left\{(-s^2/2K)\sum_{j=1}^{K}\alpha(j)\right\}\right\}.$$

It follows from the dominated convergence theorems that:

$$\lim_{n,K\to\infty}\frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}e^{-t^2/2}\exp\left\{(-1/2)\left(t-\frac{\sum_{j=1}^{K}\beta(j)}{(nK)^{1/2}}s\right)^2\right\}dt=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\lim_{n,K\to\infty}e^{-t^2/2}\exp\left\{(-1/2)\left(t-\frac{\sum_{j=1}^{K}\beta(j)}{(nK)^{1/2}}s\right)^2\right\}dt\cong$$

$$\cong\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-t^2/2}dt=1.$$

Then,

$$\lim_{n,K\to\infty} H_n(s) = \lim_{n,K\to\infty} \exp\left\{ (s^2/2nK) \left(\sum_{j=1}^{K} \beta(j)\right)^2 - (s^2/2K) \sum_{j=1}^{K} \alpha(j) \right\} =$$

$$= \exp\left\{ (-s^2/2) \lim_{n,K\to\infty} \left[\frac{-1}{nK} \left(\sum_{j=1}^{K} \beta(j)\right)^2 + \frac{1}{K} \sum_{j=1}^{K} \alpha(j) \right] \right\}.$$

Hence $K^{-1/2}(\hat{\underline{K}}-\underline{K})$ converges to a normal distribution with mean 0 and variance:

$$\sigma_2^2 = \lim_{K,n\to\infty} \left\{ \frac{1}{K} \sum_{j=1}^{K} \alpha(j) - \frac{1}{nK} \left(\sum_{j=1}^{K} \beta(j)\right)^2 \right\}.$$

and $(\hat{\underline{K}}-\underline{K})$ converges to a normal distribution normal with mean 0 and variance $\sigma^2(\hat{\underline{K}}) = K\sigma_2^2$.

Now, assume that the empirical distribution $G_n(x)$ of $np_1, np_2 \ldots, np_k$ converge to a probability distribution $G(x)$ on $(0,\infty)$, where

$$G_n(x) = \frac{1}{K} \sum_{j=0}^{K} I(np_j \leq x).$$

Then

$$\lim_{K,n\to\infty} \left\{ \frac{1}{K} \sum_{j=1}^{K} \alpha(j) \right\} =$$

$$= \lim_{K,n\to\infty} \left\{ \frac{1}{K} \sum_{j=1}^{K} e^{-np_j} \left(A_1^2 + A_2^2 np_j + A_3 \frac{(np_j)^2}{2}\right) - e^{-2np_j} \left(A_1 + A_2 np_j + A_3^2 \frac{(np_j)^2}{2} \right)^2 \right\} \cong$$

$$\cong \frac{1}{K} K \int_0^{\infty} \left\{ e^{-x} \left(A_1^2 + A_2^2 x + A_3 \frac{x^2}{2}\right) - e^{-2x} \left(A_1 + A_2 x + A_3^2 \frac{x^2}{2}\right)^2 \right\} dG(x),$$

and

$$\lim_{K,\,n\to\infty}\left\{\frac{1}{nK}\left(\sum_{j=1}^{K}\beta(j)\right)^2\right\}=$$

$$\lim_{K,\,n\to\infty}\left\{\frac{1}{nK}\left\{\sum_{j=1}^{K}e^{-np_j}\left(A_1np_j+A_2np_j(np_j-1)+A_3\frac{(np_j)^2}{2}(np_j-2)\right)\right\}^2\right\}\cong$$

$$\cong\left\{K\int_0^\infty[Kx]\,dG(x)\right\}^{-1}\left\{K\int_0^\infty\left[e^x\left(A_1x+A_2x(x-1)+A_3\frac{x^2}{2}(x-2)\right)\right]dG(x)\right\}^2,$$

because

$$n=E\left(\sum_{j=1}^{K}X_j\right)=E\left(\sum_{j=1}^{K}Y_j\right)=\sum_{j=1}^{K}np_j=K\int_0^\infty x\,dG(x)\,.$$

Therefore,

$$\sigma_2^2=\int_0^\infty\left\{e^{-x}\left(A_1^2+A_2^2x+A_3^2\frac{x^2}{2}\right)-e^{-2x}\left(A_1+A_2x+A_3^2\frac{x^2}{2}\right)^2\right\}dG(x)-$$

$$-\left\{K\int_0^\infty[Kx]\,dG(x)\right\}^{-1}\left\{K\int_0^\infty\left[e^x\left(A_1x+A_2x(x-1)+A_3\frac{x^2}{2}(x-2)\right)\right]dG(x)\right\}^2=$$

$$=\int_0^\infty\left\{e^{-x}\left(A_1^2+A_2^2x+A_3\frac{x^2}{2}\right)-e^{-2x}\left(A_1^2+A_2^2x^2+A_3^4\frac{x^4}{4}+\right.\right.$$

$$\left.\left.+2A_1A_2x+A_1A_3^2x^2+A_2xA_3^2x^2\right)\right\}dG(x)-$$

$$-\left\{K\int_0^\infty[Kx]\,dG(x)\right\}^{-1}\left\{K\int_0^\infty\left[e^x\left(A_1x+A_2x(x-1)+A_3\frac{x^2}{2}(x-2)\right)\right]dG(x)\right\}^2=$$

$$= \int_0^\infty [e^{-x}(A_1^2 - A_1^2 e^{-x})]\,dG(x) + A_2^2 \int_0^\infty [e^{-x}x(1-xe^{-x})]\,dG(x) +$$

$$+ \frac{A_3^2}{2} \int_0^\infty [x^2 e^{-x}(1 - A_3^2 x^2 e^{-x}/2)]\,dG(x) -$$

$$\int_0^\infty [xe^{-2x}(2A_1A_2 + A_1A_3^2 x + A_2A_3^2 x^2)]\,dG(x) -$$

$$- \left\{K\int_0^\infty [Kx]\,dG(x)\right\}^{-1} \left\{K\int_0^\infty [xe^{x}(A_1 + A_2(x-1) + A_3\frac{x}{2}(x-2))]\,dG(x)\right\}^2 .$$

An upper bound of asymptotic variance of $\hat{\underline{K}}$ is

$$\bar{\sigma}^2(\hat{\underline{K}}) = K\left\{\int_0^\infty [1-e^{-x}]\,dG(x) + A_2^2\int_0^\infty [e^{-x}x]\,dG(x) + \right.$$

$$+ \frac{A_3^2}{2}\int_0^\infty [x^2e^{-x}]\,G(x) - \int_0^\infty [xe^{-x}(2A_1A_2 + A_1A_3^2x + A_2A_3^2x^2)]\,dG(x) -$$

$$\left. - \left\{K\int_0^\infty [Kx]\,dG(x)\right\}^{-1} \left\{K\int_0^\infty [xe^{x}(A_1 + A_2(x-1) + A_3\frac{x}{2}(x-2))]\,dG(x)\right\}^2\right\} =$$

Note that:

$$E(N_i) = \sum_{j=1}^{K} \binom{n}{i} p_j^i (1-p_j)^{n-i} \cong \sum_{j=1}^{K} e^{-np_j}(np_j)^i/i! \cong \sum_{j=1}^{K}\int_0^\infty [e^{-x}\,x^i/i!]\,dG(x) \cong$$

$$\cong K\int_0^\infty [e^{-x}x^i/i!]\,dG(x),$$

then $i!E(N_i) \cong K\int_0^\infty [e^{-x}x^i]\,dG(x)$.

Also

$$E(D) = E\left(\sum_{j=1}^{K} I(X_j>0)\right) = \sum_{j=1}^{K} \text{Prob}(X_j>0) = \sum_{j=1}^{K} [1 - \text{Prob}(X_j=0)] =$$

$$=\sum_{j=1}^{K}[1-\text{Prob}(Y_j=0)]\cong\sum_{j=1}^{K}[1-e^{-np_j}]\cong E(S)\cong K\int_0^{\infty}(1-e^{-x})\,dG(x)\,.$$

Using this relations, $\bar{\sigma}^2(\hat{\underline{K}})$ is equal to:

$$\bar{\sigma}^2(\hat{\underline{K}})=E(D)+A_2^2E(N_1)+\frac{A_3^2}{2}\,2E(N_2)-2A_1A_2E(N_1)-A_1A_3^2 2E(N_2)-A_2A_3^2 3!E(N_3)-$$

$$-\frac{[A_1E(N_1)+2A_2E(N_2)-A_2E(N_1)+(A_3/2)3!E(N_3)-A_3 2E(N_2)]^2}{E(\sum_{j=1}^{K}Y_j)}\,.$$

Replacing $E(N_i)$ (with i=1, 2, 3) and E(D) by the observed values, $\bar{\sigma}^2(\hat{\underline{K}})$ can be estimated by:

$$\hat{\bar{\sigma}}^2(\hat{\underline{K}})=d+\left(\frac{n-1}{n}\right)^2\frac{n_1^3}{n_2^2}+\frac{(n-1)^2}{4n^2}\frac{n_1^4}{n_2^3}+2\frac{n-1}{n}\frac{n_1^2}{n_2}+$$

$$+\left(\frac{n-1}{n}\right)^2\frac{n_1^4}{2n_2^3}-\left(\frac{n-1}{n}\right)^3\frac{n_1^5}{2n_2^5}\frac{3}{2}n_3-$$

$$-\frac{1}{(\sum_{j=1}^{K}Y_j)}\left(-n_1+\frac{n-1}{n}\frac{n_1}{n_2}2n_2-\frac{n-1}{n}\frac{n_1^2}{n_2}-\right.$$

$$\left.-6\frac{n-1}{4n}\frac{n_1^2}{n_2^2}n_3+\frac{n-1}{n}\frac{n_1^2}{n_2}\right)^2=$$

$$=d+\left(\frac{n-1}{n}\right)^2\left\{\frac{n_1^3}{n_2^2}+\frac{n_1^4}{4n_2^3}+\frac{n}{n-1}\frac{2n_1^2}{n_2}+\frac{n_1^4}{2n_2^3}-\frac{n-1}{n}\frac{n_1^5}{n_2^5}\frac{3}{2}n_3\right\}-$$

$$-\frac{1}{(\sum_{j=1}^{K}Y_j)}\left(-n_1+\frac{n-1}{n}2n_1-6\frac{n-1}{4n}\frac{n_1^2}{n_2^2}n_3\right)^2.$$

4./ Numerical examples.

Example 1.

A simulation study was carried out to investigate the performance of the estimator:

$$\hat{\underline{K}}=d+\frac{n-1}{2n}\frac{n_1^2}{n_2}.$$

The true number of cluster was fixed at 200. Several populations with observation probability ranging from 0.002 to 0.01. For each given population the program produced 100 simulation runs with size sample n=50 and n=100. These 100 values were average to give the results of table 1. Note that the values most important in the sample are N_1, N_2 and $\hat{\underline{K}}$.
Also it was calculated:

$$\text{Bias}(\hat{\underline{K}})=E[(\hat{\underline{K}}_j-K)]=\frac{1}{50}\sum_{j=1}^{50}(\hat{\underline{K}}_j-K).$$

$$\text{ECM}(\hat{\underline{K}})=E[(\hat{\underline{K}}_j-K)^2]=\frac{1}{50}\sum_{j=1}^{50}(\hat{\underline{K}}_j-K)^2.$$

The simulation results indicate that:
- For population with equal observation probability (case 1), the estimator work very well.

- For any fixed popuation, the standard error for the estimator when n=100 is smaller than that of n=50.

- The standard error of estimator increases as the degree of heterogeneity of the population is increased.

- $\text{Var}(\hat{\underline{K}})$ decrease when n increase.

Table 1.

Cases	n	p_j	$\hat{\underline{K}}$	$Var(\hat{\underline{K}})$	$E(\hat{\underline{K}}_j - K)$	ECM
1.	50	p_j=0.005 j=1-200	203.32	11.74	2.96	20.46
	100		201.01	2.98	0.87	3.73
2	50	p_j=0.004 j=1-100 p_j=0.006 j=101-200	199.21	2.30	0.64	2.70
	100		202.13	6.87	1.87	10.2
3	50	p_j=0.0035 j=1-90 p_j=0.0045 j=91-180 p_j=0.014 j=181-200	203.41	11.36	2.14	15.57
	100		201.52	3.23	1.79	6.43
4	50	p_j=0.01 j=1-10 p_j=0.004 j=11-100 p_j=0.003 j=101-191 p_j=0.023 j=191-200	204.49	18.86	3.96	34.54
	100		198.21	6.03	2.07	10.28

Continue table 1.

Cases	n	p_j	$\hat{\underline{K}}$	$Var(\hat{\underline{K}})$	$E(\hat{\underline{K}}_j - K)$	ECM
5	50	p_j=0.0035 j=1-50 p_j=0.006 j=51-100 p_j=0,002 j=101-125 p_j=0.009 j=126-150 p_j=0.005 j=151-200	195.68	28.97	5.11	56.28
	100		203.34	11.31	3.14	20.85
6	50	p_j=0.006 j=1-25 p_j=0.0025 j=26-50 p_j=0.009 j=51-75 p_j=0.008 j=76-100 p_j=0.001 j=101-125 p_j=0.002 j=126-150 p_j=0.005 j=151-175 p_j=0.004 j=176-200	212.23 190.02	154.85 114.48	11.83 9.76	292.24 208.47

- Note that ECM-$[E(\hat{\underline{K}}_j - K)]$ is roughly $Var(\hat{\underline{K}})$.

Eample 2.

This interesting example was describe in Holst (1981). Given the number of dies produced r coins, r=1,2,..., in a hoard, the problem is to estimate the number of dies used in the minting process. I first discuss the reverse side: 204 coins were found in a hoard of ancient coins, 156 appeared once, 19 twice, 2 three times, and 1 four times, no die appeared more than four times. For this frecuency 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 clusters. The estimate proposed in this work is $\hat{K}$=818.

Acknowledgements.

I want to thanks Dr. Anne Chao (Institute of Statistics in National Tsing Hua University, Hsin-Chu, Taiwan) for their important contributions to this work.

Bibliography:

Chao, A. (1992). Estimating the number of classes via sample coverage. Journal of the American Statistical Association, 87, 417, 210-217.

Darroch, J.N. (1958). The multiple recapture census I: Estimation of a closed population. Biometrika, 40, 343-359.

Darroch, J.N and Ratclif, D. (1980). A note on capture-recapture estimation". Biometrika, 45, 343-359.

Engen, S. (1978). Stochastic Abundance Models, London: Chapman and Hall.

Esty, W. W. (1985). The estimation of the number of classes in a population and the coverage of a sample. Mathematical Scientist, 10, 41-50.

Fisher et all. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology, 12, 42-58.

Harris, B. (1968). Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. Journal of the American Statistical Association, 63, 837-847.

Holst, L. (1979). A unified approach to limit theorems for urn models. Journal applied probability, 16, 1, 154-162.

Holst, L. (1981). Some asymptotic result for incomplete multinomial or poisson samples. Scandinavian Journal of Statistic, 8, 243-246.

Johnson, N.L. and Kotz, S. (1977). Urn models and their applications: An approach to modern discrete probability theory, New YorK: John Wiley.

Lewontin R.C. and Prout, T. (1956). Estimation of the number of classes in a population" Biometrics, 12, 211-223.

Marchand, J.P. and Schroech, P.E. (1982). On estimation of the number of equally likely classes in a population. Communications in a Statistics, Part A-Theory and Methods, 11, 1139-1146.

McNeil, D. (1973). Estimating an author's vocabulary. Journal of the American Statistical Association, 68, 341, 92-97.