中國統計學報
第31卷第2期
八十二年九月
253-270頁

種類數估計量之一致性
—紀念陳文成教授去世12年

林惠賢　　趙蓮菊　　李燊銘
清華大學　　清華大學　　逢甲大學
統計所　　統計所　　統計系

摘　要

本文利用陳文成教授著作中的一個定理之推廣以證明Chao et al. (1993)對估計種類數之鞅估計函數(martingale estimating function)所得之估計量的一致性。

關鍵詞：多項式模式，差異性，樣本涵蓋率，零均值鞅(ZMM)。
美國數學會分類索引：主要62P10，次要62F10。

1. 前言

1.1 問題描述

母體種類數的估計在生物學的應用上是一個古典的問題。Bunge and Fitzpatrick (1993)回顧關於這個問題的不同模式和估計方法，他們也收集了與這個題目相關的參考論文超過550篇。

本文主要討論最常用的多項式模式(multinomial model)：考慮一個母體，其組成的個體可分類爲N個未知的不同種類。我們每次選取一個個體，記下它的種類別，再把它放回母體。n次個體選取完成後，稱此選取爲n階段選取(n-stage selection)。假設這些種類以任意次序標上1, 2, $\cdots$, N。令p_i爲被選取者屬於第i種類的機率，且令X_{ik}爲k階段選取屬於第i種類的個體數，則對所有的$k = 1, 2, \cdots, n$, $(X_{1k}, X_{2k}, \cdots, X_{Nk})$服從多項式分配。我們的目的就是在$n$次選取完成後估計種類數$N$。

1.2 文獻回顧及研究動機

Bunge et al. (1992)在其論文中提及過去處理本問題在頻率論(frequentist)觀點下有三個主要的估計法：

(1) 假定多項式分配中的種類機率有某一函數形式(functional form)（例如：McNeil, 1973），亦即存在某一函數f使$p_i = f(i), i = 1, \cdots, N$。
(2) 以一有母數(parametric)機率密度函數(p.d.f)描述或近似種類機率的分配（例如：Sichel, 1986）。
(3) 無母數樣本涵蓋率(sample coverage)方法：藉由樣本涵蓋率的估計來估計種類數（例如：Chao and Lee, 1992）。在樣本涵蓋率法中種類機率的變異係數(coefficient of variation)於測度種類機率之差異性(heterogeneity)時扮演重要的角色。

最近Chao et al. (1993)提出另一個無母數方法－使用鞅估計函數(martingale estimating function)。這個方法提供了另一類估計量－鞅估計函數計量（或簡稱鞅估計量）。此類估計量包括種類機率相等時的最大概似估計量(maximum likelihood estimator)(Darroch, 1958)和種類機率不等時的無母數樣本涵蓋率估計量(Chao and Lee, 1992)。Chao and Lee (1992)並列有樣本涵蓋率應用在種類數和母體個體總數估計的歷史可供參考；亦可參看Becker (1984)， Becker and Heyde (1990), Yip (1989,1991)和Yip et al. (1993)關於重複捕取(capture-recapture)模式下鞅估計函數的應用。然而，前述的鞅估計函數法只處理種類機率均等的情況。而Chao, Yip and Lin (1993)可同時處理種類機率不等的情況。爲了整篇文章之完整性，我們將在第二節簡略複習樣本涵蓋率估計量及鞅估計函數估計量

。而本文主要將在第三節證明在某些條件下，鞅估計函數估計量具一致性(consistency)。

2.樣本涵蓋率估計量及鞅估計函數估計量

2.1 樣本涵蓋率估計量及其變異數之估計(Chao and Lee, 1992)

令 f_{ik} 爲在 k 階段選取恰好出現 i 次的種類數，亦即 $f_{ik}=\sum_{j=1}^{N} I[X_{jk}=i]$，其中 $I(\cdot)$ 是常用的指示函數(indicator funtion)。令 D_k 爲 k 階段選取的不同種類數，$D_k=\sum_{i=1}^{k} f_{ik}=\sum_{j=1}^{N} I[X_{jk}>0]$。定義 k 階段選取的樣本涵蓋率 C_k 爲

$$C_k=\sum_{i=1}^{N} p_i I\ [\text{第 } i \text{ 類在 } k \text{ 階段選取曾出現過}]。\tag{2.1}$$

則估計種類數之樣本涵蓋率估計量(Chao and Lee, 1992)爲

$$\hat{N}_s=\frac{D_n}{\hat{C}_n}+\frac{f_{1n}}{\hat{C}_n}\hat{\gamma}_n^2\ ,\tag{2.2}$$

其中爲 $\hat{C}_n=1-f_{1n}/n$ 爲 C_n 之估計量(Good, 1953；或Robbins, 1968)，$\hat{\gamma}_n$ 爲種類機率的變異係數之估計量（定義及估計量見下節及(2.20)式）。而此估計量近似變異數的估計值爲

$$\hat{var}\ (\hat{N}_s)=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial \hat{N}_s}{\partial f_{in}}\frac{\partial \hat{N}_s}{\partial f_{jn}}\ \hat{cov}(f_{in},f_{jn})\ ,$$

其中

$$\hat{cov}\ (f_{in},\ f_{jn})=\begin{cases} f_{in}(1-f_{in}/\hat{N}_s) & \text{if } i=j \\ -f_{in}f_{jn}/\hat{N}_s & \text{if } i\neq j\ 。\end{cases}$$

2.2 鞅估計函數估計量及其變異數之估計(Chao et al. 1993)

令 F_k 表由 k 階段選取的選取過程所生成的 σ-域 (σ- field)，亦即 $F_k=\sigma\{X_{1i},X_{2i},\cdots,X_{Ni},i=1,2,\cdots,k\}$。令 $m_k=I$〔第 k 次選取爲一曾出現過的種類〕及 $u_k=I$〔第 k 次選取爲一不曾出現過的種類〕，則對所有 $k,u_k+m_k=1$，且

$$E(u_k\mid F_{k-1})=1-C_{k-1}\ 。\tag{2.3}$$

同理，我們可得

$$E(m_k \mid F_{k-1}) = C_{k-1} \circ \tag{2.4}$$

我們進一步可得到變異數

$$var(u_k \mid F_{k-1}) = C_{k-1}(1 - C_{k-1}) , \tag{2.5}$$

$$var(m_k \mid F_{k-1}) = C_{k-1}(1 - C_{k-1}) , \tag{2.6}$$

及共變異數

$$cov(u_k, m_k \mid F_{k-1}) = -C_{k-1}(1 - C_{k-1}) \circ \tag{2.7}$$

根據(2.3)和(2.4)我們考慮下列缺差：

$$\begin{aligned} \mathcal{D}_k &= NC_{k-1}[u_k - (1 - C_{k-1})] - N(1 - C_{k-1})[m_k - C_{k-1}] \\ &= NC_{k-1}u_k - N(1 - C_{k-1})m_k \circ \end{aligned} \tag{2.8}$$

若NC_{k-1}和$N(1 - C_{k-1})$是F_{k-1}可測(F_{k-1}-measurable)，則過程 $M = \{M_k : k = 1, 2, \cdots\}$，其中$M_k = \sum_1^k \mathcal{D}_i$，是一個 ZMM (zero mean martingale)。為了得到一個N的一般化的缺估計函數，我們可取一個有界且相對於M是可預測(predictable)的加權函數w_{k-1}，使估計函數成為

$$M_n^* = \sum_{k=1}^{n} w_{k-1}\mathcal{D}_k$$

$$= \sum_{k=1}^{n} w_{k-1} \left[NC_{k-1}u_k - N(1 - C_{k-1})m_k\right] \tag{2.9}$$

$$= \sum_{k=1}^{n} w_{k-1} \left[(NC_{k-1}) - N\, m_k\right] \circ \tag{2.10}$$

在本文中，我們討論兩種可能的加權函數：對所有的$k = 1, 2, \cdots, N, w_{k-1} = 1$，和 Godambe (1985)所建議的最佳加權函數。他證明對一已知缺差函數$\mathcal{D}_k$之最佳加權函數是

$$w_{k-1} = \frac{E[\partial \mathcal{D}_k / \partial N \mid F_{k-1}]}{E[\mathcal{D}_k^2 \mid F_{k-1}]} \circ \tag{2.11}$$

在此"最佳"意指可造一個N的最短信賴區間。

對所有p_i都相等的特別情況，亦即$p_1 = p_2 = \cdots = p_N = 1/N$，我們可得$NC_{k-1} = D_{k-1}(D_0 \equiv 0)$和$N(1 - C_{k-1}) = N - D_{k-1}$。所以估計函數(2.9)式變成

$$\sum_{k=1}^{n} w_{k-1}[D_{k-1}u_k - (N - D_{k-1})m_k] \circ \tag{2.12}$$

這個估計函數類似於Yip (1989, 1991)，Yip et al. (1993)，Becker (1984)和Becker and Heyde (1990)對重複捕取模式所考慮的。令(2.12)等於其均值零，求解得N的估計量爲

$$\sum_{k=1}^{n} w_{k-1}\, D_{k-1} / \sum_{k=1}^{n} w_{k-1}\, m_k \circ \tag{2.13}$$

假如$w_{k-1} = 1$，上式簡化成

$$\hat{N}_0 = \sum_{k=1}^{n} D_{k-1} / \sum_{k=1}^{n} m_k \circ \tag{2.14}$$

這和重複捕取模式中的Schnabel估計量完全相同（假設將每次選取看成一個陷阱(trapping)樣本）（Schnabel, 1938；或Seber, 1982）。注意Schnabel估計量和選取的次序有關。換言之，由於我們的估計量是根據前面的歷史而定，所以Schnabel估計量當選取的次序不同時會有不同的結果。

當所有p_i都相等時，最大概似估計量(MLE) $\hat{N}_{0,mle}$是下列方程式(Darroch, 1958)的解

$$\sum_{i=0}^{D_n - 1} (N - i)^{-1} = n/N \,, \tag{2.15}$$

其漸近變異數(asymptotic variance)爲

$$var(\hat{N}_{0,mle}) = N/[\exp(n/N) - (n/N) - 1] \circ$$

注意此最大概似估計量和選取的次序無關。

由(2.6)，(2.11)和(2.12)可導出此情況下的最佳加權函數爲$w^*_{k-1} = 1/(N - D_{k-1})$。則對應於此最佳加權數的最佳估計量爲下列方程式之解

$$\sum_{k=1}^{n} [D_{k-1} - Nm_k]/(N - D_{k-1}) = 0 \circ \tag{2.16}$$

我們可證明上方程式和(2.15)是完全相同的。類似Yip (1991)的導法，$\hat{N}_0$和$\hat{N}_{0,mle}$的漸近標準分別爲下式中的w_{k-1}以1和$1/(N-D_{k-1})$替代而得：

$$\left\{\sum_{k=1}^{n} w_{k-1}^2 D_{k-1}(N-D_{k-1})\right\}^{1/2} / \sum_{k=1}^{n} w_{k-1}m_k \circ$$

在大部分的實際應用上，種類機率均等的假設是不成立的。以前處理有差異情況的方法多數都對種類機率的分配做有參數的假設。例如：有些作者假設$p_1, p_2, \cdots, p_N$是一組來自某一分配的隨機取樣(random sample) (Fisher et al. 1943是其中一例) 。但在我們的方法中，我們把$p_1, p_2, \cdots, p_N$當做固定的參數(fixed parameters)，但在導變異數估計量時例外。注意此地雖然允許N (N可能很大) 個不同之參數，但並無辨認性(identifiability)之問題，因爲在本論文中將展示種類之估計只與最重要之參數平均值$\bar{p} = \sum_{i=1}^{N} p_i/N = 1/N$和異係數$(CV)\gamma = [\sum_{i=1}^{N}(p_i - \bar{p})^2/N]^{1/2}/\bar{p}$有關。只要平均值和變異係數相同，實際上$p_i$各自的大小並不影響估計結果。故在整個討論中只看成二個參數。此處基本的動機如下：我們不直接估計N，而藉由(NC_{k-1})的估計來估計N。從(2.10)式，假如可找到NC_{k-1}的估計量$\widehat{NC}_{k-1}$，則我們有下列N的估計量

$$\sum_{k=1}^{n} w_{k-1}(\widehat{NC}_{k-1}) / \sum_{k=1}^{n} w_{k-1}m_k \circ \tag{2.17}$$

當估計量$\widehat{NC}_{k-1}$是F_{k-1}可測時，(2.17)之漸近標近標準差可以證明爲

$$N\left\{\sum_{k=1}^{n} w_{k-1}^2 [C_{k-1}(1-C_{k-1})]\right\}^{1/2} / \sum_{k=1}^{n} w_{k-1}m_k \circ \tag{2.18}$$

到此我們的過程就是要找出$E(NC_{k-1})$的估計量，既而用它當做(NC_{k-1})的估計量。我們可導出

$$E(NC_{k-1}) \approx E(D_{k-1}) + \gamma^2 \cdot E(f_{1,k-1}) , \tag{2.19}$$

其中$f_{10} \equiv 0$。若每得一觀測值，即作一次CV之修正估計，即對每一m，求一$(CV)^2$估計如下：

$$\hat{\gamma}_m^2 = max\left\{\frac{D_m \sum_{i=1}^{m} i(i-1)f_{im}}{m(m-1)[1-f_{1m}/m]} - 1, 0\right\} \circ \tag{2.20}$$

從(2.17)，(2.19)和(2.20)，我們提出的估計量在$w_{k-1}=1$時爲

$$\hat{N}=\sum_{k=k_0}^{n}[D_{k-1}+\hat{\gamma}_{k-1}^2 f_{1,k-1}]/\sum_{k=k_0}^{n} m_k \circ \tag{2.21}$$

其中所有累加從起始值k_0開始是因爲我們需要足夠的觀測值去得到一個CV的穩定(stable)估計量。在Chao et al. (1993)中，我們選擇$k_0=n/2$。CV之估計實爲此過程中最困難之處。另一個替代的方法是我們等所有選取都做完後再估計此參數。換言之，我們考慮下列含修正的CV估計之估計量

$$\begin{aligned}\hat{N} &= \sum_{k=1}^{n}[D_{k-1}+\hat{\gamma}_n^2 f_{1,k-1}]/\sum_{k=1}^{n} m_k \\ &= \hat{N}_0+\hat{\gamma}_n^2\cdot\sum_{k=1}^{n} f_{1,k-1}/\sum_{k=1}^{n} m_k\end{aligned} \tag{2.22}$$

此時最佳加權函數可以證明爲

$$w_{k-1}^*=1/(1-C_{k-1}) \circ \tag{2.23}$$

所以藉由使用C_{k-1}之估計量$\hat{C}_{k-1}=1-f_{1,k-1}/(k-1)$可以得到$w_{k-1}^*$之估計量$\hat{w}_{k-1}^*$。因此對應於此最佳加權函數估計量，我們有兩個估計量

$$\hat{N}_w=\sum_{k=k_0}^{n}\hat{w}_{k-1}^*[D_{k-1}+\hat{\gamma}_{k-1}^2\cdot f_{1,k-1}]/\sum_{k=k_0}^{n}\hat{w}_{k-1}^* m_k \tag{2.24}$$

和

$$\tilde{N}_w=\sum_{k=1}^{n}\hat{w}_{k-1}^*[D_{k-1}+\hat{\gamma}_n^2\cdot f_{1,k-1}]/\sum_{k=1}^{n}\hat{w}_{k-1}^* m_k \circ \tag{2.25}$$

有關各種估計量之比較可參Chao et al. (1993)。

在Chao et al. (1993)中採用重抽法(bootstrap resampling method)來估計變異數。作法如下：首先假設$p_1,p_2,\cdots,p_N$是一組來自分配$F(p)$的隨機取樣。然後無條件地(unconditionally)，$(f_{0n},f_{1n},\cdots,f_{nn})$服從含參數$N$和種類機率$\binom{n}{k}\int p^k(1-p)^{n-k}dF(p)$, $k=0,1,\cdots,n$之多項式分配。由含參數$\hat{N}$和估計的種類機率$f_{kn}/\hat{N}, k=1,2,\cdots,n$; $f_{0n}=\hat{N}-\sum_{k=1}^{n} f_{kn}$之多項式分配重複生成$(f_{0n}^*,f_{1n}^*,\cdots,f_{nn}^*)$。然而我們仍需要製造每一種類的出現歷史，以便計算缺估計量。對出現i次的每一種類，我們從$1,2,\cdots,n^*$中，

$n^* = \sum_{i=1}^{n} i f_{in}^*$，隨機選取$i$個觀測值（但被選取過的時間除外），令此種類出現在所選取的時間且不出現在其他時間。

對每一組資料，我們重複生成B次，因此得到B個重抽法估計值$\hat{N}_i^*, i = 1, 2, \cdots B$。$\hat{N}$的重抽法變異數就是此$B$個$\hat{N}_i^*$的樣本變異數，即

$$\hat{var}(\hat{N}) = \left[\sum_{i=1}^{B} (\hat{N}_i^*)^2 - \left(\sum_{i=1}^{B} \hat{N}_i^*\right)^2 / B\right]^2 / (B-1) \text{。}$$

3.估計量之一致性

3.1 樣本涵蓋率估計量之一致性

Lee and Chao (1993)根據Chen (1980, 1981a, 1981b)的結果已證明樣本涵蓋率估計量的一致性。現將它轉述如下：

定理3.1：在多項式模式下，假如種類機率$(p_1, p_2, \cdots, p_N)$服從一含參數α且對稱的Dirichlet分配，則其$(CV^2)\gamma^2 \approx 1/\alpha$。當$\lim_{N\to\infty} \frac{n}{N} = \lambda > 0$時，假如$CV$已知，則樣本涵蓋率估計量具一致性，即我們有

$$N^{-1}\left[\frac{D_n}{\hat{C}_n} + \frac{f_{1n}}{\hat{C}_n}\gamma^2\right] \xrightarrow{P} 1 \text{。}$$

在證明此定理時，主要用到Chen (1980, 1981a, 1981b)的下列定理：即在定理3.1之條件下，對$i = 0, 1, \cdots$

$$\frac{f_{in}}{N} \xrightarrow{P} \frac{\Gamma(i+\alpha)}{i!\Gamma(\alpha)} \left(\frac{\lambda}{\lambda+\alpha}\right)^i \left(\frac{\alpha}{\lambda+\alpha}\right)^\alpha \text{。} \tag{3.1}$$

3.2 鞅估計量之一致性

本文主要利用(3.1)之推廣以證明鞅估計量亦具有一致性。

定理3.2：在多項式模式下，假如種類機率$(p_1, p_2, \cdots, p_N)$服從一含參數α且對稱的Dirichlet分配，則其$(CV^2)\gamma^2 \approx 1/\alpha$。當$\lim_{N\to\infty} \frac{n}{N} = \lambda > 0$時，假如$CV$已知，則鞅估計量具一致性。換言之，對任意起始值$k_0$，我們可以證明

$$N^{-1}\left\{\frac{\sum_{k=k_0}^{n}[D_{k-1} + \gamma^2 f_{1,k-1}]}{\sum_{k=k_0}^{n} m_k}\right\} \xrightarrow{P} 1 \text{。}$$

假如CV已知，則缺估計量具一致性。換言之，對任意起始值k_0，我們可以證明

$$N^{-1}\left\{\frac{\sum_{k=k_0}^{n}[D_{k-1}+\gamma^2 f_{1,k-1}]}{\sum_{k=k_0}^{n} m_k}\right\} \xrightarrow{P} 1 \ 。$$

在證明定理3.2之前，先證明幾個需要的引理。在引理3.1, 3.3和3.4中，我們令

$$f_N(x)=\left(1-\frac{1}{N}\cdot\frac{\alpha+1}{\alpha+\lambda x}\right)^{N(\alpha+\lambda x)-\alpha-\frac{3}{2}} \ 。$$

引理3.1：當$N\to\infty$時，在$[0,1]$上$f_N(x)\xrightarrow{u}e^{-(\alpha+1)}$，其中$\xrightarrow{u}$表均勻收斂。

引理3.2：若φ_N和φ在$[0,1]$上可積，且當$N\to\infty$時，在$[0,1]$上$\varphi_N\xrightarrow{u}\varphi$，則對任意極限值爲無窮大之序列$\{a_N\}_{N=1}^{\infty}$，我們有：當$N\to\infty$時，

$$\frac{1}{a_N}\sum_{k=1}^{a_N}\varphi_N\left(\frac{k}{a_N}\right)\to\int_0^1\varphi(x)dx \ 。$$

引理3.3：當$\lim_{N\to\infty}\frac{n}{N}=\lambda$時，

$$\begin{aligned}T(N) \ = \ & \frac{1}{n}\Bigg|\sum_{k=1}^{n}\frac{k}{N}\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\left(1-\frac{1}{N}\cdot\frac{\alpha+1}{\alpha+\frac{k}{N}}\right)^{N(\alpha+\frac{k}{N})-\alpha-\frac{3}{2}} \\ & -\sum_{k=1}^{n}\frac{k}{n}\lambda\left(\frac{\alpha}{\alpha+\frac{k}{n}\lambda}\right)^{\alpha+1} f_N\left(\frac{k}{n}\right)\Bigg|\to 0 \ 。\end{aligned}$$

證明：

令$T_0(k)=\frac{k}{N}\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\left(1-\frac{1}{N}\frac{\alpha+1}{\alpha+\frac{k}{N}}\right)^{N(\alpha+\frac{k}{N})-\alpha-\frac{3}{2}}$，且令$T_i$爲將$T_{i-1}$中第一個$\frac{k}{N}$以$\frac{k}{n}\lambda$替代而成之函數，則

$$T(N)\le\frac{1}{n}\sum_{k=1}^{n}\sum_{i=0}^{3}|T_i(k)-T_{i+1}(k)| \ 。 \tag{3.2}$$

因爲對$k\ge 1$，我們都有$\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\le 1$，$\left(1-\frac{\alpha+1}{N\alpha+k}\right)^{N\alpha+k}\le e^{-(\alpha+1)}$和$\left(1-\frac{\alpha+1}{N\alpha+k}\right)^{-\alpha-\frac{3}{2}}\le\left(1-\frac{\alpha+1}{N\alpha}\right)^{-\alpha-\frac{3}{2}}$，所以

$$\begin{aligned}
&\frac{1}{n}\sum_{k-1}^{n}\ |T_0(k)-T_1(k)|\\
\leq\ & 1\cdot e^{-(\alpha+1)}\left(1-\frac{\alpha+1}{N\alpha}\right)^{\alpha\ \frac{3}{2}}\cdot\frac{1}{n}\sum_{k=1}^{n}\left|\frac{k}{N}-\frac{k}{n}\lambda\right|\\
=\ & e^{-(\alpha+1)}\left(1-\frac{\alpha+1}{N\alpha}\right)^{\alpha\ \frac{3}{2}}\left(\frac{1}{n^2}\sum_{k=1}^{n}k\right)\cdot\left|\frac{n}{N}-\lambda\right|\\
&\to 0
\end{aligned}$$

同理，當$1\leq i\leq 3$時，我們有

$$\frac{1}{n}\sum_{k=1}^{n}\ |T_i(k)-T_{i+1}(k)|\to 0\ \circ$$

故由(3.2)，可得證此引理。 ■

引理3.4：當$\lim_{N\to\infty}\frac{n}{N}=\lambda$時，

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\frac{k}{N}\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\left(1-\frac{\alpha+1}{N\alpha+k}\right)^{(N-1)\alpha+k\ \frac{3}{2}}\\
\to\ & \int_0^1\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}e^{-(\alpha+1)}dx\ \circ
\end{aligned}$$

證明：因爲當$x\in[0,1]$時，我們有

$$\begin{aligned}
&\left|\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}f_N(x)-\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}e^{-(\alpha+1)}\right|\\
\leq\ & \lambda\,|\,f_N(x)-e^{-(\alpha+1)}\,|\,,
\end{aligned}$$

且由引理3.1可得當$N\to\infty$時，在$[0,1]$上

$$\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}f_N(x)\xrightarrow{u}\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}e^{-(\alpha+1)}\ \circ$$

再由引理3.2我們可得：當$\lim_{N\to\infty}\frac{n}{N}=\lambda$時，

$$\frac{1}{n}\sum_{k=1}^{n}\lambda\frac{k}{n}\left(\frac{\alpha}{\alpha+\lambda\frac{k}{n}}\right)^{\alpha+1} f_N\left(\frac{k}{n}\right)$$
$$\to \int_0^1 \lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1} e^{-(\alpha+1)}dx \circ$$

最後由引理3.3即得證此引理。 ■

現在開始證明定理3.2如下：不失一般性，可假設$k_0=1$。我們首先推廣(3.1)至以下形式：

(a) $\frac{1}{n}\sum_{k=1}^{n}\frac{f_{1,k-1}}{N}\xrightarrow{P}\int_0^1\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}dx$，
(b) $\frac{1}{n}\sum_{k=1}^{n}\frac{D_{k-1}}{N}\xrightarrow{P}1-\int_0^1\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha}dx$ 及
(c) $\frac{1}{n}\sum_{k=1}^{n}m_k\xrightarrow{P}1-\int_0^1\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}dx$。

首先讓k固定。因爲

$$\begin{aligned}
E\left(\frac{f_{1k}}{N}\right) &= Pr(X_{1k}=1)\\
&= \int_0^1 kp(1-p)^{k-1}p^{\alpha-1}(1-p)^{(N-1)\alpha-1}\frac{\Gamma(N\alpha)}{\Gamma(\alpha)\Gamma((N-1)\alpha)}\,dp\\
&= \frac{k\cdot\Gamma(N\alpha)}{\Gamma(\alpha)\Gamma((N-1)\alpha)}\int_0^1 p^{\alpha}(1-p)^{(N-1)\alpha+k-2}\,dp\\
&= \frac{k\cdot\Gamma(N\alpha)}{\Gamma(\alpha)\Gamma((N-1)\alpha)}\cdot\frac{\Gamma(\alpha+1)\Gamma((N-1)\alpha+k-1)}{\Gamma(N\alpha+k)}\\
&= k\alpha\cdot\frac{\Gamma(N\alpha)\Gamma[(N-1)\alpha+k-1]}{\Gamma[(N-1)\alpha]\Gamma(N\alpha+k)},
\end{aligned}$$

所以

$$E\left(\frac{f_{1k}}{N}\right)=\frac{k}{N}\left\{N\alpha\cdot\frac{\Gamma(N\alpha)\Gamma[(N-1)\alpha+k-1]}{\Gamma[(N-1)\alpha]\Gamma(N\alpha+k)}\right\}\circ$$

令

$$h(N,k)=N\alpha\,\frac{\Gamma(N\alpha)\Gamma[(N-1)\alpha+k-1]}{\Gamma[(N-1)\alpha]\Gamma(N\alpha+k)}\circ$$

利用$\log\Gamma(z)=(z-1/2)\log(z)-z+\frac{1}{2}\log(2\pi)+O(\frac{1}{z})$（參看Whittaker and Watson, 1927），我們得到

$$\begin{aligned}\log h(N,k) &= (\alpha+1)\log(N\alpha)-(\alpha+1)\log(N\alpha+k)\\ &\quad +1-[(N-1)\alpha-1/2]\log(1-\frac{1}{N})\\ &\quad +[(N-1)\alpha+k-1-1/2]\log\left(1-\frac{\alpha+1}{N\alpha+k}\right)\\ &\quad +O(\frac{1}{N})+O(\frac{1}{N\alpha+k})\ \circ\end{aligned}$$

既而令

$$h(N,k)=\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\cdot R(N,k)\ ,$$

其中

$$\begin{aligned}R(N,k) &= e\cdot(1-\frac{1}{N})^{-[(N-1)\alpha-1/2]}\cdot\left(1-\frac{\alpha+1}{N\alpha+k}\right)^{(N-1)\alpha+k-3/2}\\ &\quad \cdot e^{O(\frac{1}{N})}\ e^{O(\frac{1}{N\alpha+k})}\ \circ\end{aligned}$$

故

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}E\left(\frac{f_{1k}}{N}\right)\\ =\ &\frac{1}{n}\sum_{k=1}^{n}\frac{k}{N}\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\cdot R(N,k)\\ =\ &e\cdot(1-1/N)^{-[(N-1)\alpha-1/2]}\cdot e^{O(1/N)}\cdot\frac{1}{n}\\ &\cdot\sum_{k=1}^{n}\left\{\frac{k}{N}\left(\frac{\alpha}{\alpha+\frac{k}{N}}\right)^{\alpha+1}\left(1-\frac{\alpha+1}{N\alpha+k}\right)^{(N-1)\alpha+k-3/2}\right\}\ \circ\end{aligned}$$

再由引理3.4，我們可得

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}E\left(\frac{f_{1,k-1}}{N}\right)\\ =\ &\frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{k=1}^{n-1}E\left(\frac{f_{1k}}{N}\right)\\ \to\ &\int_0^1\lambda x\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}dx\ \circ\end{aligned}$$

又

$$var\left(\frac{1}{n}\sum_{k=1}^{n}\frac{f_{1k}}{N}\right)=\frac{1}{n^2}\sum_{k=1}^{n}\frac{var(f_{1k})}{N^2}+\frac{1}{n^2}\sum_{k\neq 1}\frac{cov(f_{1k},f_{1l})}{N^2},$$

由繁瑣的計算可得

$$var(f_{1k})=O(N),cov(f_{1k},f_{1l})=O(N),$$

所以

$$var\left(\frac{1}{n}\sum_{k=1}^{n}\frac{f_{1,k-1}}{N}\right)=\frac{n\cdot O(N)}{n^2\,N^2}+\frac{(n^2-n)\cdot O(N)}{n^2\,N^2}=O(1/N),$$

即 $var\left(\frac{1}{n}\sum_{k-1}^{n}\frac{f_{1,k-1}}{N}\right)\to 0$，故得證(a)。
同理，因

$$\begin{aligned}\frac{1}{n}\sum_{k-1}^{n}E\left(\frac{D_{k-1}}{N}\right) &= \frac{1}{n}\sum_{k=1}^{n}E\left(\frac{N-f_{0,k-1}}{N}\right)\\ &= 1-\frac{1}{n}\sum_{k=1}^{n}E\left(\frac{f_{0,k-1}}{N}\right)\to 1-\int_0^1\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha}dx,\end{aligned}$$

且

$$var\left(\frac{1}{n}\sum_{k=1}^{n}\frac{D_{k-1}}{N}\right)=var\left(\frac{1}{n}\sum_{k=1}^{n}\frac{f_{0,k-1}}{N}\right)\to 0,$$

故得證(b)。而(c)的證明如下：
因

$$\begin{aligned}\frac{1}{n}\sum_{k-1}^{n}E(m_k) &= \frac{1}{n}\sum_{k=1}^{n}E\left(1-\frac{f_{1k}}{k}\right)\\ &= 1-\frac{1}{n}\sum_{k=1}^{n}\frac{1}{k}E(f_{1k})\\ &\to 1-\int_0^1\left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1}dx,\end{aligned}$$

且

$$
\begin{aligned}
var\left(\frac{1}{n}\sum_{k=1}^{n} m_k\right) &= \frac{1}{n^2} var(n - D_n) \\
&= \frac{1}{n^2} var(D_n) \\
&= \frac{1}{n^2} var(N - f_{0n}) \\
&\to 0 ,
\end{aligned}
$$

故得證(c)。最後由(a)，(b)，(c)且因

$$
\gamma^2 = \frac{1}{\alpha}\left(1 - \frac{\alpha + 1}{N\alpha + 1}\right) ,
$$

故

$$
\begin{aligned}
& N^{-1}\left\{\sum_{k=1}^{n} [D_{k-1} + \gamma^2 f_{1,k-1}] / \sum_{k=1}^{n} m_k\right\} \\
= \; & \frac{\frac{1}{Nn}\sum_{k=1}^{n} [D_{k-1} + \gamma^2 f_{1,k-1}]}{\frac{1}{n}\sum_{k=1}^{n} m_k} \\
\xrightarrow{P} \; & \frac{1 - \int_0^1 \left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha} dx + \frac{1}{\alpha}\int_0^1 \lambda x \left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1} dx}{1 - \int_0^1 \left(\frac{\alpha}{\alpha+\lambda x}\right)^{\alpha+1} dx} \\
= \; & 1 。 \quad \blacksquare
\end{aligned}
$$

因此，很明顯地CV在證明一致性時扮演最重要的角色。假如 CV已知，我們可以一致地(consistently)估計看不到的種類數。否則，一致性就變成一個達不到的理想而已。誠如Bunge and Fitzpatrick (1993)中引述I. J. Good之語"總是會有很多極稀少而看不到的種類"。

誌謝詞：阮感謝台美文化基金會kap陳文憲教授（Miami大學）提供陳文成教授生前所有e著作。阮beh用本文來紀念陳文成教授去世12年，並衰心願望伊生前e理想卡早實現。（1993年7月1日）

參考文獻

Becker, N. G. (1984). Estimating Population Size from Capture-Recapture Experiments in Continous Time. Australian Journal of Statistics. 26, pp1-7.

Becker, N. G. and Heyde, C. C. (1990). Estimating Population Size from Capture-Recapture Experiments. Stochastic Processes and Their Applications. 36, pp77-83.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: Recent Developments. Journal of American Statistical Association. 88, pp364-373.

Bunge, J., Fitzpatrick, M. and Handley, J. (1992). Simulation Comparison of Three Estimators of the Number of Species. Cornell University Statistics Center, Technical Report #92.25.

Chao, A. and Lee, S.-M. (1992). Estimating the Number of Classes via Sample Coverage. Journal of American Statistical Association. 87, pp210-217.

Chao, A., Yip, P. and Lin, H-S. (1993). Estimating the Number of Species via a Martingale Estimating Function. Technical Report, Institute of Statistics, Tsing-Hua University, Taiwan. Feb. 1993. (Submitted).

Chen, W-C (1980). On the Weak Form of Zipf's Law. Journal of Applied Probability. 18, pp611-622.

Chen, W-C (1981a). Limit Theorems for General Size Distributions. Journal of Applied Probability. 18, pp139-147.

Chen, W-C (1981b). Some Local Limit Theorems in the Symmetric Dirichlet-Multinomial Urn Models. Annals of Institute Statistical Mathematics. 33A, pp405-415.

Darroch, J. N. (1958). The Multiple-Recapture Census I: Estimation of a Closed Population. Biometrika. 45, pp343-359.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The Relation Between the Number of Species and Species and the Number of Individuals in a Random Sample of an Animal Population. Journal of Animal Ecology. 12, pp42-58.

Godambe, V. P. (1985). The Foundations of Finite-Sample Estimation in Stochastic Processes. Biometrika. 72, pp419-428.

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika. 40, pp237-264.

Lee, S.-M. and Chao, A. (1993). Consistency of Sample Coverage Estimator of the Number of Classes Under a Dirichlet-Multinomial Model. Technical Report, Institute of Statistics, Tsing-Hua University, Taiwan, Feb. 1993.

McNeil, D. (1973). Estimating an Author's Vocabulary. Journal of American Statistical Association. 68, pp92-96.

Robbins, H. (1968). Estimating the Total Probability of the Unobserved Outcomes of an Experiment. Annals of Mathematical Statistics. 39, pp256-257.

Sicel, H. (1986). Parameter Estimation for a Word Frequency Distribution Based on Occupancy Theory. Communications in Statistics. Part A-Theory and Methods. 15, pp935-949.

Schnabel, Z. E. (1938). The Estimation of the Total Fish Population of a Lake. American Mathematical Monthly. 45, pp348-352.

Seber, G. A. F. (1982). The Estimation of Animal Abundance. 2nd edition. London: Griffin.

Whittaker, E. T. and Waton, G. N. (1927). A Course of Modern Analysis. Cambridge University Press.

Yip, P. (1989). An Inference Procedure for a Capture and Recapture Experiment with Time-Dependent Capture Probabilities. Biometrics. 45, pp471-479.

Yip, P. (1991). A Martingale Estimating Equation for a Capture-Recapture Experiment. Biometrics. 47, pp1081-1088.

Yip, P., Fong, D.Y.T. and Wilson, K. (1993). Estimating Population Size by Recapturing Sampling via Estimating Function. Stochastic Models. 9, pp179-194.

[民國 82 年 6 月 26 日收稿]

 JCSA, September 1993, Vol. 31, No. 2, page 253–270

Consistency of an Estimator of the Number of Species ... In Memory of Professor Wen-Chen Chen

H-S Lin
Institute of Statistics
National Tsing Hua University

Anne Chao
Institute of Statistics
National Tsing Hua University

S.-M. Lee
Department of Statistics
Feng-Chia University

ABSTRACT

A theorem in Chen (1980,1981a,1981b) is extended to prove the consistency of an estimator of the number of species derived from a martingale estimating function (Chao et al. 1993).

Key words and phrases. number of classes, multinomial, heterogeneity, sample coverage, zero mean martingale.

AMS 1991 subject classifications. Primary 62P10, Secondary 62F10.