Rev. Acad. Canar. Cienc., XIV (Núms. 1-2), 47-64 (2002)

ESTIMACIÓN NO PARAMÉTRICA DEL NÚMERO DE ESPECIES EN UN ECOSISTEMA DE TAMAÑO DESCONOCIDO VÍA ESTIMADORES JACKKNIFE GENERALIZADOS EN POBLACIONES FINITAS

Juan José Prieto Martínez
Dpto. de Estadística e Investigación Operativa.
Fac. de C.C. Matemáticas. Universidad Complutense de Madrid.
e-mail: jjprieto@mat.ucm.es

Resumen

En este artículo se presenta una colección de estimadores no paramétricos para el número S de especies en un ecosistema de tamaño N conocido, al aplicar una extensión del jackknife generalizado de Gray y Shucany (1972). Cuando N es desconocido, se presenta una cota basada en la información dada por la variable aleatoria número de especies observadas y las variables número de especies observadas una y dos veces. Su distribución asintótica sirve de vía para la obtención de un estimador, $\hat{N}$, para N. El citado estimador es utilizado para estimar posteriormente S. La realización de los estimadores S y $\hat{N}$ son investigados por simulación de Monte Carlo.

Palabras claves: Número de especies, jackknife generalizado, distribución asintótica.

Clasificación AMS: 1162G05

Abstract

We use an extension of the generalized jackknifed approach of Gray and Shucany (1972) to obtain new nonparametric estimators for the number S of species in a ecosytem of size N known. If N is unknown it shows an estimator. A limit law is rigurously proved for it and an estimator is proposed via the asymptotic distribution. This estimator is just used to estimate N and can be applied later to the estimators for S.

The performance of the estimators is investigated by means of Monte Carlo experiments.

Key words: Number of species, jackknife generalized, asymptotic distribution.

1. *Introducción*

La estimación del número de clusters de una población presenta un interés relevante en el campo de la ecología. En este último, las distintas especies que comparten un cierto hábitat conformarían los clusters y su número se decanta en el índice más significativo a la hora de evaluar la *diversidad* del ecosistema o hábitat citado. Por lo general, la estimación del número de especies se lleva a cabo mediante métodos de captura-recaptura, siendo la referencia clásica esencial sobre el tema Seber (1982). Una revisión histórica de los principales trabajos al respecto indica que, al menos en sus orígenes, la mayoría asumen la hipótesis de que las probabilidades de captura son iguales para todas las especies de la población. La citada hipótesis es, sin duda, muy cuestionable, buscando tan solo una cierta manejabilidad en la búsqueda de resultados. Muchos trabajos pioneros utilizaron la citada hipótesis (Crowcroft y Jeffers (1961), Huber (1962), Edwards & Eberhardt (1967), Bailey (1969), Carothers (1973), Esty (1983) y Chao (1992)) y probaron empíricamente que no se sostenía en la realidad. Posteriormente, otros autores mostraron que los sesgos derivados de la heterogeneidad en las probabilidades de captura podrían ser muy elevados, comprobándose así la infactibilidad de los resultados obtenidos a través de la utilización de los estimadores citados. Por ejemplo, Manly (1971), Carothers (1973) y Burnham y Overton (1979) llegan a evaluar la magnitud del citado sesgo en determinados casos prácticos, llegando a proponer, tras observar su elevada cuantía, estimadores alternativos mediante la ampliación de técnicas como la del jackknife generalizado o la de Jolly-Seber. Hoy día, la utilización de este último tipo de estimadores que relajan la condición de equiprobabilidad de captura se han extendido enormemente. Pese a ello, una buena parte de los estimadores más populares exigen el conocimiento del tamaño poblacional global N que influye a todas las especies y han sido desarrollados mediante las técnicas del jackknife generalizado.

En el epígrafe 2 se proponen una colección de estimadores para el número de especies en poblaciones finitas de tamaño N (conocido) aplicando la técnica de remuestreo jackknife generalizado. El fin de la citada colección de estimadores es utilizarlos también cuando el tamaño poblacional es desconocido tras la utilización del estimador propuesto en Prieto (2001), presentado aquí en el epígrafe 3, para estimar N. Es por consiguiente una estimación en dos etapas. Nótese que, por ejemplo, en entornos ecológicos parece natural que no se conozca cuál es el tamaño total de la población si el objetivo es estimar el número de especies que constituye dicha población.

2. Una colección de estimadores para el número de especies en un ecosistema de tamaño conocido

Supongamos que una población de tamaño conocido N está constituida por S especies o clusters $C_1, C_2, \ldots C_S$. Definamos T_j el tamaño de la especie C_j donde $N=\sum_j T_j$. Es claro que

$$S = \sum_{i=1}^{N} F_i \qquad \text{y} \qquad N = \sum_{i=1}^{N} iF_i,$$

donde F_i es el número de especies de tamaño i en la población. Finalmente N_i denota el número de especies representadas exactamente i veces en una muestra de tamaño n y D_n el número total de especies observadas en ella, se tiene que

$$D_n = \sum_{i=1}^{n} N_i \qquad \text{y} \qquad n = \sum_{i=1}^{n} iN_i.$$

En entornos ecológicos la mayoría de los estimadores empleados utilizan la técnica del jackknife generalizado (j.g.) para construir estimadores (cuasi)insesgados a partir de otros que no lo son. Los estimadores finales j.g. adoptan la forma (véase Gray y Shucany (1972))

$$G(\hat{S}_1, \hat{S}_2) = \frac{\hat{S}_1 - R\hat{S}_2}{1-R},$$

donde $\hat{S}_1$ y $\hat{S}_2$ son dos estimadores sesgados para S y $R \neq 1$ es un valor real. Este último valor de R queda abierto con el objetivo de que pueda elegirse convenientemente de modo que se minimice el sesgo de G. La idea es pues obtener un estimador insesgado (sofisticado) a partir de otros dos sencillos fácilmente obtenibles por experimentos de captura-recaptura. Los más habituales por sencillez son $\hat{S}_1 = D_n$ (número de especies observadas en la muestra) y

$$\hat{S}_2 = \frac{1}{n}\sum_{j=1}^{n} D_{n-1}(j),$$

donde $D_{n-1}(j)$ representa el número de especies observadas en la muestra tras haber eliminado la *j-ésima* observación en la misma. Es claro que $D_{n-1}(j)=D_n-1$ si la especie asociada a la *j-ésima* observación ha sido observada una sola vez. En caso contrario, $D_{n-1}(j)=D_n$. Por consiguiente,

$$D_{n-1}(j) = D_n - \frac{N_1}{n},$$

donde N_1 es el número de especies observadas exactamente una vez. Es fácil comprobar que eligiendo

$$R = \frac{E(D_n) - S}{E(D_{(n-1)}) - S}$$

conseguimos que $G(D_n, D_{(n-1)})$ sea insesgado para S. En lo que sigue se utilizará la notación $\hat{S} = G(D_n, D_{(n-1)})$. Claramente

$$\hat{S} = \frac{D_n - RD_{(n-1)}}{1-R} = D_n + Z\frac{N_1}{n}, \qquad [1]$$

siendo

$$Z = \frac{R}{1-R} = \frac{E(D_n) - S}{E(D_{(n-1)}) - E(D_n)}.$$

Puesto que X_j representa el número de observaciones de C_j, se tiene que

$$E(D_n) = E\left(\sum_{j=1}^{S} I(X_j > 0)\right) = \sum_{j=1}^{S} P(X_j > 0) = S - \sum_{j=1}^{S} P(X_j = 0),$$

donde $I(A)$ es la función indicadora, y

$$E(N_1) = \sum_{j=1}^{S} P(X_j = 1),$$

de donde

$$Z = n\frac{\sum_{j=1}^{S} P(X_j = 0)}{\sum_{j=1}^{S} P(X_j = 1)}. \qquad [2]$$

Basándonos en una aproximación propuesta por Shlosser (1981) (igualmente utilizada y justificada en Haas y Stokes (1996)) según la cual

$$P(X_j = k) \approx \binom{T_j}{k} q^k (1-q)^{T_j - k},\ 0 \le k \le n,\ 1 \le j \le S, \qquad [3]$$

donde $q=n/N$, y sustituyendo la expresión anterior en [2] se llega a que

$$Z \approx n \frac{\sum_{j=1}^{S} (1-q)^{T_j}}{\sum_{j=1}^{S} T_j q (1-q)^{T_j - 1}}. \qquad [4]$$

Los valores T_j son desconocidos como es lógico. Aproximando cada uno de ellos por el promedio

$$\bar{T} = \frac{1}{S} \sum_{j=1}^{S} T_j = \frac{N}{S}$$

se llega a $Z \approx \frac{nS(1-q)}{Nq}$.

De este modo, [1] se puede expresar como

$$\hat{S} = D_n + \frac{nS(1-q)}{Nq} \frac{N_1}{n} = D_n + \frac{S(1-q)N_1}{n}.$$

Finalmente, reemplazando S por $\hat{S}$ en el segundo miembro de la igualdad y despejando se llega al estimador:

$$\hat{S}_1 = \left(1 - \frac{(1-q)N_1}{n}\right)^{-1} D_n \qquad [5]$$

Un razonamiento similar permite llegar a justificar los distintos estimadores del número de especies utilizados hoy en día con mayor asiduidad en ecología (aparte del anterior [5]). A continuación describiremos tres de ellos que tienen una relevancia especial. El primero parte del valor óptimo de Z dado en [2] y reescribe [1] mediante

$$\hat{S} = D_n + \frac{\sum_{j=1}^{S} P(X_j = 0)}{E(N_1 / n)} \frac{N_1}{n}.$$

Identificando $E(N_1)$ como el valor observado de esta variable y haciendo uso de [4] se llega a

$$S = D_n + \sum_{j=1}^{S} (1-q)^{T_j}$$

en donde, de nuevo, el sumatorio de la derecha no puede ser evaluado. Finalmente, sustituyendo T_j por su valor promedio $\overline{T} = N/S$ se llega a

$$S = D_n + S(1-q)^{N/S}$$

que conduce directamente al M-estimador

$$\hat{S}_2 = \underset{\hat{S}}{cero}\left\{S(1-(1-q)^{N/\hat{S}}) - D_n\right. \qquad [6]$$

tras razonar como en $\hat{S}_1$.

El siguiente estimador que comentamos es del tipo jackknife de segundo orden. Utilizando en [5] las expansiones de Taylor siguientes (pivotando en $\overline{T}$)

$$(1-q)^{T_j} \approx (1-q)^{\overline{T}} + (1-q)^{\overline{T}} \ln(1-q)(T_j - \overline{T}), \qquad [7]$$

$$T_j q(1-q)^{T_j-1} \approx T_j q\{(1-q)^{\overline{T}-1} + (1-q)^{\overline{T}-1} \ln(1-q)(T_j - \overline{T})\}, \forall j \in \{1,...,S\}, \qquad [8]$$

se llega a

$$Z \approx S(1-q)\frac{1}{1+\ln(1-q)\overline{T}\gamma^2} \approx S(1-q)(1-\ln(1-q)\overline{T}\gamma^2), \qquad [9]$$

siendo γ el coeficiente de variación de los tamaños de las distintas especies que componen la población. Sustituyendo [9] en [1] se llega a la expresión

$$\hat{S} = D_n + S(1-q)(1-\ln(1-q)\overline{T}\gamma^2)\frac{N_1}{n} =$$

$$= D_n + \frac{SN_1(1-q)}{n} - \frac{N_1(1-q)\ln(1-q)\gamma^2}{q}, \qquad [10]$$

de la cual, actuando como en $\hat{S}_1$ y $\hat{S}_2$, se deriva el nuevo M-estimador

$$\hat{S}_3 = \underset{\hat{S}}{cero}\left\{\left(1-\frac{N_1(1-q)}{n}\right)\hat{S} - D_n + \frac{N_1(1-q)\ln(1-q)\gamma^2}{q}\right\}. \qquad [11]$$

Un enfoque alternativo en la búsqueda de estimadores del número de especies cuando N es conocido consiste en estimar Z en [1] mediante la técnica de Horvitz-Thompson (véase Sarndal, Swensson y Wretman (1992)). Para ello obsérvese que Z adopta la forma $Z \approx A/B$, siendo A y B sumas no observables pero que pueden ser evaluadas ambas (o al menos una de ellas) mediante estimadores Horvitz-Thompson. Esto conduce a proponer estimadores del tipo

$$\hat{S} = D_n + \frac{\hat{A}}{\hat{B}} \frac{N_1}{n}. \qquad [12]$$

Denotando por $f(z)=(1-q)^z$ y $g(z)=n^{-1}zq(1-q)^{z-1}$, es claro que $A = \sum_{j=1}^{S} f(T_j)$ y $B = \sum_{j=1}^{S} g(T_j)$. Tan solo nos referiremos a la variante más utilizada como botón de muestra. Estimando B por Horvitz-Thompson se llega a que

$$B_{HT} = \sum_{\{j:X_j>0\}} \frac{g(T_j)}{1-P(X_j=0)} = \sum_{\{j:X_j>0\}} \frac{g(T_j)}{1-(1-q)^{T_j}},$$

que puede ser aproximado por

$$\hat{B}_{HT} = \sum_{\{j:X_j>0\}} \frac{g(\overline{T}_j)}{1-(1-q)^{\overline{T}_j}}.$$

Aunque A podría ser estimada por esta misma técnica, la citada variante, hoy ampliamente usada, simplemente estima A mediante la expresión ya usada anteriormente $\hat{A} = S(1-q)^{N/S}$. La sustitución de los anteriores valores $\hat{A}$ y $\hat{B}$ en [12] nos lleva a $\hat{S} = D_n + \frac{S(1-q)^{N/S}}{\hat{B}_{HT}} \frac{N_1}{n}$, de donde se deriva el M-estimador

$$\hat{S}_4 = \underset{\hat{S}}{cero}\left\{\hat{S}\left(1 - \frac{N_1(1-q)^{N/\hat{S}}}{n\hat{B}_{HT}}\right) - D_n\right\}. \qquad [13]$$

Todos los estimadores presentados en este apartado han sido propuestos para estimar el número de clusters en una población finita de tamaño N conocido. Nuestro interés se centrará en cómo poder adecuar aquéllos cuando el tamaño poblacional N es desconocido. En entornos ecológicos es difícil pensar en situaciones en las que desconociendo S se conozca N. Por consiguiente la adecuación antes citada presenta un interés innegable. Nuestra propuesta consiste en incorporar los tamaños estimados, utilizando los resultados asintóticos presentados en Prieto (1998,2001), en lugar del tamaño poblacional real que figura en todos los estimadores citados en este epígrafe. Consiste por tanto en una técnica de estimación en dos etapas.

3. Estimación del tamaño total de un ecosistema.

El estimador citado $\hat{N}$ (ver Prieto (2001)) no solamente es candidato para estimar el número de especies S sino que también lo es para estimar el tamaño poblacional N. A continuación se propone una técnica de estimación para S con N desconocido en dos etapas: la primera etapa consiste en estimar N a partir del estimador citado $\hat{N}$; en la segunda etapa se aplican directamente los estimadores propuestos del epígrafe anterior. Es claro que la estimación de N en la primera etapa simplemente requiere considerar cada individuo de la población como un clase.

Se asume que una muestra aleatoria de tamaño n con reemplazamiento ha sido extraída de la población constituida por K individuos. En cada extracción, la probabilidad de observar el individuo j será $p_j \geq 0$, con $\sum_{j=1}^{K} p_j = 1$. Sea X_j la variable aleatoria número de veces que es observado el individuo j. Por consiguiente $(X_1,...,X_K)$ se distribuye como una multinomial $M(n; p_1,..., p_K)$. Para $i=0,1,...,n$, denotemos por $N_i = \sum_{j=1}^{K} I(X_j = i)$ el número de individuos que han sido observados exactamente i veces en la muestra. Un estimador natural-sesgado y que infraestima N es el número $K_{obs} = \sum_{j=1}^{K} I(X_j > 0)$. El número N_i de clusters observados exactamente i veces cumple que

$$E(N_i) = \sum_{j=1}^{K} \binom{n}{i} p_j^i (1-p_j)^{n-i}. \qquad [14]$$

La desigualdad de Schwartz permite escribir

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

que es equivalente a $\left(\sum_{j=1}^{K} p_j (1-p_j)^{n-1}\right)^2 \left(\sum_{j=1}^{K} p_j^2 (1-p_j)^{n-2}\right)^{-1} \leq \sum_{j=1}^{K} (1-p_j)^n.$

Utilizando [14], esta última desigualdad se convierte en

$$E(N_0) \geq \frac{(E(N_1))^2 (n-1)}{2nE(N_2)}. \qquad [15]$$

Por otra parte, es claro que $N=K_{obs}+N_0$ de donde, tomando esperanzas y teniendo en cuenta la expresión [15], se deriva que una cota inferior para N es

$$C_S = E(K_{obs}) + \frac{n-1}{2n}\frac{E(N_1)^2}{E(N_2)}.$$

Esta cota será tanto más precisa (pudiendo llegar a convertirse la desigualdad en igualdad) cuanto más accesible sea la citada cota de Schwartz, por ejemplo, para el caso de probabilidades iguales o moderadamente desiguales. Supongamos, por un momento, que $E(N_2)$ sea conocida. Si así fuera, sustituyendo las restantes esperanzas por sus respectivos valores observados se obtendría un estimador para C_S dado por

$$\hat{C}_S = K_{obs} + \frac{n-1}{2n}\frac{n_1^2}{E(N_2)}.$$

Por lo general, $E(N_2)$ no será conocida. En consecuencia, la expresión anterior no sería un estimador en sentido estricto porque su valor quedará indeterminado tras observar la muestra. En cualquier caso, mantendremos la condición de que la citada esperanza es conocida a la espera de relajar esta limitación en resultados numéricos.

El siguiente teorema indica cuál es la distribución asintótica del valor $\hat{C}_S$, la cual sirve como vía para la obtención del estimador citado $\hat{N}$.

Teorema. *Bajo políticas de muestreo $n=n(K)\to\infty$ cuando $K\to\infty$, si las dos siguientes condiciones técnicas se cumplen: i) $n=o(K)$ y $n^{-1}=o(K^{-1/2})$; ii) $n\cdot max(p_j)=o(1)$, entonces $\frac{\sqrt{n}(\hat{C}_S - C_S)}{\sqrt{K}\sigma^2}$ converge en distribución a una normal estándar, donde σ^2 viene dado por*

$$\sigma^2 \approx \frac{1}{K}\sum_{j=1}^{K}\left\{e^{-np_j}(A_1^2 + A_2^2 np_j) - e^{-2np_j}(A_1 + A_2 np_j)^2\right\} - \frac{1}{nK}\left\{\sum_{j=1}^{K} e^{-np_j}(A_1 np_j + A_2 np_j(np_j - 1))\right\}^2,$$

con $A_1 = -1, A_2 = \frac{n-1}{n}\frac{E(N_1)}{E(N_2)}$. ■

De aquí se deduce que un estimador para K es:

$$\hat{N} = \frac{n\hat{V}ar(\hat{C}_s)}{s^2}, \qquad [16]$$

donde

$$s^2 = \int_0^n e^{-x}\left(1 - e^{-x}\right)dG(x) + \left(\frac{n-1}{n}\right)^2\left(\frac{n_1}{n_2}\right)^2\int_0^n (xe^{-x})(1 - xe^{-x})dG(x) +$$

$$+2\left(\frac{n-1}{n}\right)\left(\frac{n_1}{n_2}\right)^2\int_0^{t} xe^{-2x}dG(x)-$$

$$-\left\{\int_0^{t} xdG(x)\right\}^{-1}\left\{\int_0^{t} xe^{-x}dG(x)+\left(\frac{n-1}{n}\right)\left(\frac{n_1}{n_2}\right)\int_0^{\infty} x(x-1)e^{-x}dG(x)\right\}^2$$

En la fórmula [16] la estimación de $\hat{V}ar(\hat{C}_S)$ queda libre. Existen para su elección distintas posibilidades. Por ejemplo, a partir de una muestra observada, se podría estimar mediante técnicas habituales de partición en grupos aleatorios, semimuestras reiteradas,...(véase, por ejemplo, Efron (1982) y Wolter (1990)). En entornos ecológicos (e.g., estimación del número de especies en una población), su mera observación reiterada conduciría a obtener varias muestras (digamos m) a partir de las cuales la varianza empírica de los valores $\hat{C}_S$ observados en la *j-ésima* muestra conduciría a la estimación empírica habitual mediante

$$\hat{V}ar(\hat{C}_S)=\frac{1}{m}\sum_{h=1}^{m}\hat{C}_{Sh}^2-\left(\frac{1}{m}\sum_{h=1}^{m}\hat{C}_{Sh}\right)^2. \qquad [17]$$

Ver Prieto (2001). Esta expresión será utilizada en los resultados numéricos mostrados en el siguiente epígrafe concernientes al numerador del estimador propuesto en [16]. Queda por determinar el cálculo integral del denominador, que puede siempre ser aproximado por métodos de cuadraturas. $\hat{N}$ resulta ser un estimador basado en m muestras. Cada muestra j proporciona un valor observado $\hat{C}_S < K$ y el conjunto de los m valores intervienen en la obtención de su varianza estimada. Una característica importante a subrayar del estimador $\hat{N}$ es que el resultado asintótico sirve de vía para obtener el estimador propuesto y exige que n sea una $o(K^{\beta})$ con $\frac{1}{2}<\beta<1$. Por otra parte, la necesidad de no tener que imponer la asunción de equiprobabilidad y considerar que las p_j's toman valores cercanos a cero en todos los casos puede modelizar una situación inicial de investigación bastante frecuente en contextos de sobreestimación. Por consiguiente, el estimador propuesto puede ser aplicado a distintas áreas bajo este contexto citado.

Una característica que diferencia [16] de otros estimadores es que no involucra parámetro alguno. Se trata, por tanto, de un estimador de tipo no paramétrico. De facto, resulta ser también libre de distribución puesto que no exige el conocimiento distribucional de ninguna de las variables aleatorias que intervienen en el desarrollo. A

continuación se presentan los resultados numéricos obtenidos por simulación de Monte Carlo para comprobar la eficiencia del estimador propuesto bajo distintos escenarios de trabajo.

Nótese que para evaluar la varianza es necesario asumir que K es suficientemente grande y que las probabilidades $p_1,...,p_K$ han sido generadas de la forma siguiente. A partir de una distribución dada positiva y admitiendo esperanzas, se extrae una muestra aleatoria simple $(\xi_1,...,\xi_K)$, aceptando que las p_j´s se corresponden con las partes fraccionales de las ξ_j´s es decir, $p_j = \frac{\xi_j}{\sum_{j=1}^{K} \xi_j}$. La distribución de Dirichlet es un claro candidato a modelizar el comportamiento estocástico en la generación conjunta de $p_1,...,p_K$ como proceso. Claramente $p_j \to 0$ *c.s.* cuando $K \to \infty$, siendo estas últimas convergencias compatibles con las condiciones del teorema. Ello indica que las p_j´s tienden a ser incorreladas cuando $K \to \infty$. Adicionalmente, las p_j´s pueden considerarse igualmente distribuidas por simetría. Puesto que en el teorema se exige que $np_j \to 0$, puede comprobarse que esta condición se satisface *c.s.* por el proceso de generación citado. Siendo así, aceptamos que F (en realidad $F=F_n$) es la distribución (cuasi)subyacente de las np_j, de donde, las siguientes aproximaciones directamente se desprenden del teorema de Kolmogorov-Smirnov:

$$\text{a)}\ \frac{1}{K}\sum_{j=1}^{K} e^{-np_j}(1-e^{-np_j}) \approx \int_0^n e^{-x}(1-e^{-x})dF(x),$$

$$\text{b)}\ \frac{1}{nK}\left(\sum_{j=1}^{K} np_j e^{-np_j}\right)^2 \approx \frac{\left(\int_0^n xe^{-x}dF(x)\right)^2}{\int_0^n xdF(x)}.$$

Si las p_j´s se rigen por un proceso de Dirichlet, es obvio que F debería igualarse con una Beta reescalada por el tamaño muestral.

4. Resultados numéricos.

La Tabla 4 presentada a continuación muestran los resultados computacionales que se han obtenido por simulación de Monte Carlo para los M-estimadores $S_i, i = 1,2,3,4$. Se han considerado poblaciones de tamaños N=10.000 y N=26.000 con un número real de

especies son desiguales, las cuales están reflejadas en la Tabla 1. Las fracciones de muestreo consideradas han sido del 10% y 20%.

Tabla 1. Estructura de un ecosistema en el caso no equiprobable.

Clusters	Probabilidades p_j	N° de individuos Por especie.
Del cluster 1 al 50	0,004	40
Del cluster 51 al 100.	0,002	20
Del cluster 101 al 150.	0,008	80
Del cluster 151 al 200.	0,006	60

El estimador $\hat{N}$ se ha utilizado ahora para estimar N y los valores estimados sustituyen a N en las fórmulas de los M-estimadores citados. Puesto que el primero exige la observación reiterada de varias muestras para llevar a cabo las estimaciones empíricas de la varianza, hemos fijado el número de las muestras reiteradas en 100. Las Tablas 2 y 3 muestran sucintamente los sesgos y errores cuadráticos medios de $\hat{N}$ en los casos equiprobable y no equiprobable, respectivamente.

Tabla 2. Estimación de *N*. Caso equiprobable.

N	n	$\hat{E}(\hat{K}_{Prieto})$	$\hat{S}esgo(\hat{K}_{Prieto})$	$\hat{E}CM$
10.000	1000	9680	-320	103625
10.000	2000	9760	-240	59449
26.000	2600	25060	-940	884825
26.000	5200	26780	780	609129

Tabla 3. Estimación del número *N*. Caso no equiprobable.

N	n	$\hat{E}(\hat{K}_{Prieto})$	$\hat{S}esgo(\hat{K}_{Prieto})$	$\hat{E}CM$
10.000	1000	9286	-714	510796
10.000	2000	10344	344	119231
26.000	2600	27122	1122	1260253
26.000	5200	25060	-940	885121

La Tabla 4 incorpora la eficiencia empírica de los M-estimadores $\hat{S}_i$, con $i=1,2,3,4,$ en términos de los valores de los sesgos y errores cuadráticos medios. Se intenta comparar su precisión bajo situaciones análogas. Adicionalmente se intenta contrastar la sensibilidad de los citados estimadores en los distintos escenarios contemplados en las tablas. En particular, ante los elementos claves siguientes:

(1) Tamaño de la población *N;*

(2) Fracción de muestreo;

(3) Estimador $\hat{N}$ utilizado como sustituto el valor real *N* en las fórmulas [3], [6], [11] y [13].

A raíz de los resultados se contempla, en el mejor de los casos para $N=10.000,$ que el error relativo en término de sesgo del $\hat{S}_3$ es del 7,5% frente a un 12,5% para $\hat{S}_4$. Para $N=26.000,$ los sesgos relativos de todos los $\hat{S}_i$ tienden a aumentar. $\hat{S}_3$ es el que mantiene el sesgo relativo más bajo. La influencia de $\hat{N}$ es pequeña debido a que el número de especies en la población es muy elevado, estando acorde con las hipótesis de los resultados asintóticos en los que el estimador está basado. Con respecto a la fracción de muestreo las precisiones aumentan a medida que el tamaño muestral aumenta tal como cabría esperar. En el caso no equiprobable se han obtenido valores de sesgo similares para todos los M-estimadores. Para $n=2000$ y $N=10.000$ el error relativo en cuanto a sesgo y perteneciente al mejor de los casos es de un 7%, el cual corresponde a $\hat{S}_3$, frente a un 13% que pertenece a $\hat{S}_2$. Puede observarse al respecto que los sesgos relativos aumentan con respecto al caso equiprobable. En ambos, la influencia del estimador de la primera etapa $\hat{N}$ es mínima.

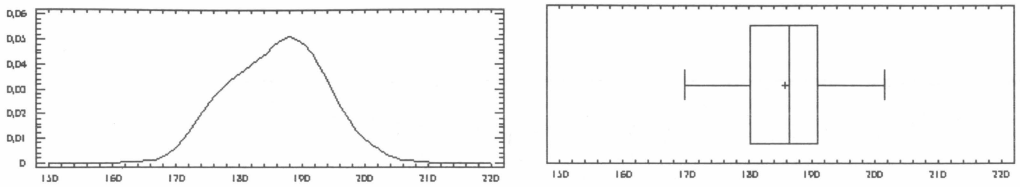
Tabla 4. Estimación del número de especies en una población.

Estimador $\hat{S}_1$							
		Caso equiprobable			*Caso no equiprobable*		
N	*n*	$\hat{E}(\hat{S}_1)$	$\hat{S}esgo(\hat{S}_1)$	$\hat{E}CM$	$\hat{E}(\hat{S}_1)$	$\hat{S}esgo(\hat{S}_1)$	$\hat{E}CM$
10.000	1000	170	-30	1504	159	-41	2112
10.000	2000	184	-16	658	172	-28	1212
26.000	2600	149	-51	2859	135	-65	5039
26.000	5200	165	-35	1692	150	-50	3121
Estimador $\hat{S}_2$							
		Caso equiprobable			*Caso no equiprobable*		
N	*n*	$\hat{E}(\hat{S}_2)$	$\hat{S}esgo(\hat{S}_2)$	$\hat{E}CM$	$\hat{E}(\hat{S}_2)$	$\hat{S}esgo(\hat{S}_2)$	$\hat{E}CM$
10.000	1000	168	-32	1568	161	-39	1955
10.000	2000	179	-21	901	170	-30	1304
26.000	2600	152	-48	2917	139	-61	4077
26.000	5200	162	-38	1881	149	-51	3013
Estimador $\hat{S}_3$							
		Caso equiprobable			*Caso no equiprobable*		
N	*n*	$\hat{E}(\hat{S}_3)$	$\hat{S}esgo(\hat{S}_3)$	$\hat{E}CM$	$\hat{E}(\hat{S}_3)$	$\hat{S}esgo(\hat{S}_3)$	$\hat{E}CM$
10.000	1000	172	-28	1703	164	-36	1698
10.000	2000	185	-15	1097	177	-23	1705
26.000	2600	158	-42	2723	145	-55	3717
26.000	5200	167	-33	1989	165	-35	1796
Estimador $\hat{S}_4$							
		Caso equiprobable			*Caso no equiprobable*		
N	*n*	$\hat{E}(\hat{S}_4)$	$\hat{S}esgo(\hat{S}_4)$	$\hat{E}CM$	$\hat{E}(\hat{S}_4)$	$\hat{S}esgo(\hat{S}_4)$	$\hat{E}CM$
10.000	1000	167	-33	2003	159	-41	2199
10.000	2000	175	-25	1520	170	-30	1319
26.000	2600	150	-50	3581	157	-50	5371
26.000	5200	162	-38	1821	150	-43	2411

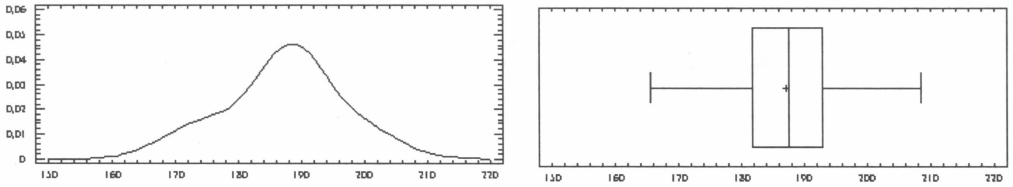
Las Figuras 1 y 2 muestran gráficamente las densidades empíricas y los diagrama de caja asociados de los estimadores $\hat{S}_i$ (i=1,2,3,4) en los casos equiprobables y no equiprobables cuando N es estimado previamente mediante $\hat{N}$. Puede observarse que son asimétricas y centradas en valores inferiores a 200 aunque con sesgos reducidos.

Es importante subrayar que los resultados empíricos obtenidos han sido comparados con otros estimadores propuestos por Goodman (1949), Korwar (1988) y Chao y Lee (1992). Los resultados aquí presentados son muy satisfactorios, sobre todo en el caso no equiprobable, frente a los obtenidos por los citados autores.

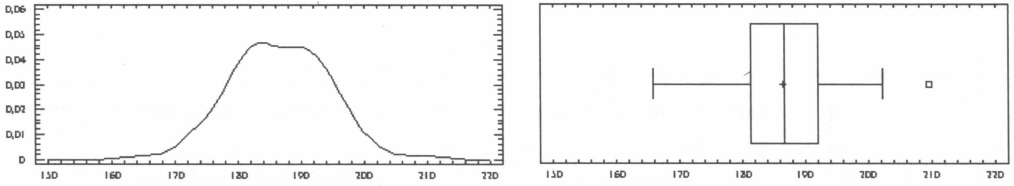
Figura 1. Densidades empíricas y diagramas de caja de los M-estimadores. Caso equiprobable.



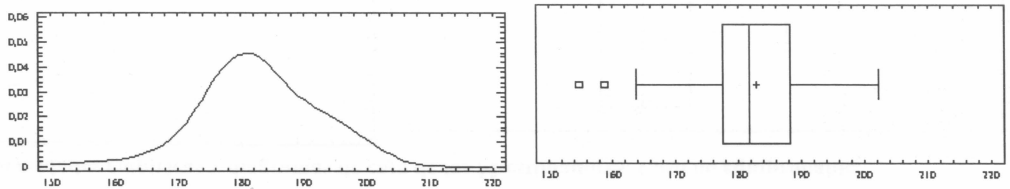
a) Caso equiprobable. Estimación de *S1. S=200. N=10.000. n=2.000.*



b) Caso equiprobable. Estimación de *S2. S=200. N=10.000. n=2.000.*

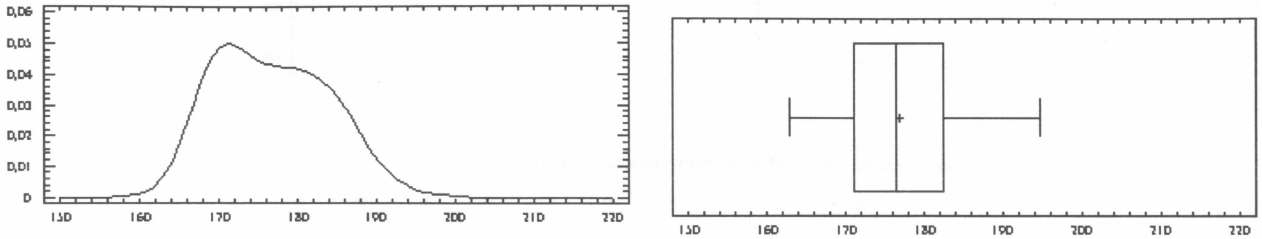


c) Caso equiprobable. Estimación de *S3. S=200. N=10.000. n=2.000.*

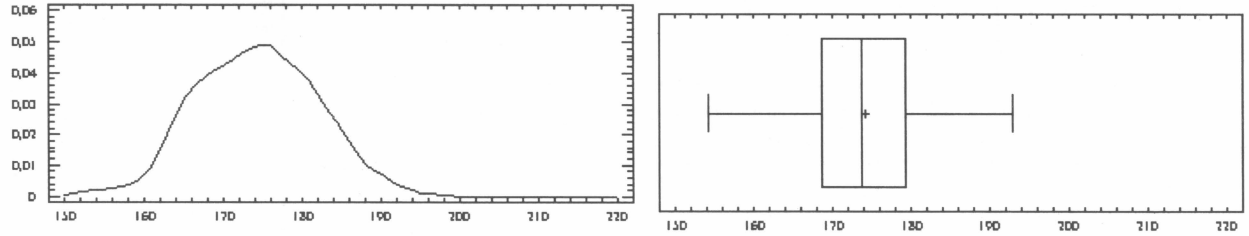


d) Caso equiprobable. Estimación de *S4. S=200.N=10.000. n=2.000*

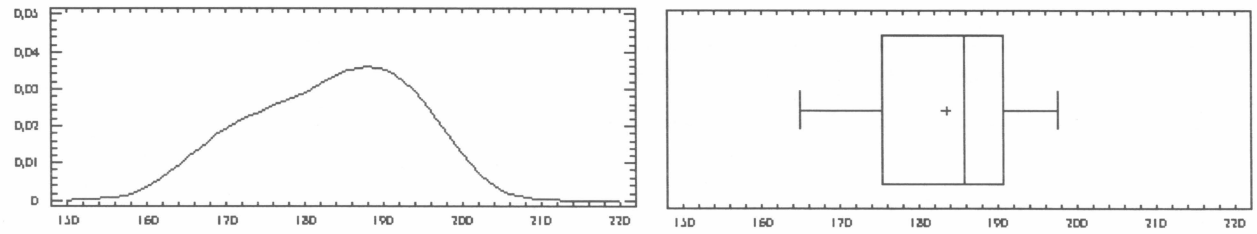
Figura 2. Densidades empíricas y diagramas de caja de los M-estimadores. Caso no equiprobable.



a) Caso no equiprobable. Estimación de *S1* con $\hat{N}$. *S=200. N=10.000. n=2.000.*



b) Caso no equiprobable. Estimación de *S2* con $\hat{N}$. *S=200. N=10.000. n=2.000.*



c) Caso no equiprobable. Estimación de *S3* con $\hat{N}$. *S=200. N=10.000. n=2.000.*

4. Bibliografía.

[1] **BAILEY, J. A.** 1969. "Trap responses of wild cottontails". *Journal Wildlife Management*. 33, 48-58.

[2] **BURNHAM, K.P. & OVERTON, W.S.** 1979. "Robust estimation of population size when capture probabilities vary among animals". *Ecology*, 60, 927-936.

[3] **CAROTHERS, A.D.** 1973. "Capture-recapture methods applied to a population with known parameters". *Journal of Animal Ecology*, 42, 125-146.

[4] **CHAO, A. & LEE, S.M.** 1992. "Estimating the number of classes via sample coverage". *Journal of the American Statistical Association*, 87, 210-217.

[5] **CROWCROFT, P. & JEFFERS, J.N.R.** 1961. "Variability in the behaviour of wild house nice toward live traps". *Proceeding of the Zoological Society*, 137, 573-582.

[6] **EDWARDS, W.R. & EBERHARDT, L.L.** 1967. "Estimating cottontail abundance from live trapping data". *Journal of Wildlife Management,* 31, 87-96.

[7] **EFRON, B.** 1982. *The jackknife, the bootstrap and the others resampling plans*. SIAM. Monographs, Nº 38, Philadelpia. SIAM.

[8] **GRAY, H. L & SHUCANY, W.R.** 1972. *The generalized jackknife statistic*, New York: Marcel Dekker.

[9] **GOODMAN, L.A.** 1949. "On the estimation of the number of classes in a population". *Annals of Mathematical Statistics*, 20, 572-579.

[10] **HAAS, P.J. & STOKES, L.** 1996. *Estimating the number of classes in a finite population*. IBM Research. Report RJ 10025, IBM Almaden Research Centre, San José, CA.

[11] **HUBER, J.J.** 1962. "Trap response of confined cottontail populations". *Journal of Wildlife Management,* 26, 177-185.

[12] **KORWAR, R.M.** 1988. "On the observed number of classes from multivariante power series and hipergeometric distributions". *Sankhya: The Indian Journal of Statistics*, 50, 39-59.

[13] **MANLY, B.F.J.** 1971. "Estimates of a marking effect with captures-recapture sampling". *Journal Applied Ecology*, 8, 181-189.

[14] **PRIETO MARTÍNEZ, J.J.** 1998. "Una distribución asintótica para un estimador natural del número de clusters en una población". *Qüestiió,* 22, 3, 417-441

[15] **PRIETO MARTÍNEZ, J.J.** 2001. *The number of dies to strike the coinage: A statistical problem in ancient numismatic.* XXVI Congreso Nacional de Estadística e Investigación Operativa. (6-9 de Noviembre de 2001). Pag de Acta 158.

[16] **SARNDAL, C.E., SWENSSON, B. & WRETMAN, J.** 1992. *Model-assisted survey sampling*. New York: Springer-Verlag.

[17] **SEBER, G.A.F.** 1982. *The Estimation of Animal Abundance.* 2nd edition. London: Griffin.

[18] **SHLOSSER, A.** 1981. "On estimation of the size of the Dictionary of a long text on the basis of a sample". *Engineering Cybernetics*, 19, 97-102.