QÜESTIIÓ, vol. 22, 3, p. 417-441, 1998

UNA DISTRIBUCIÓN ASINTÓTICA PARA UN ESTIMADOR NATURAL DEL NÚMERO DE CLUSTERS EN UNA POBLACIÓN

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Un estimador natural, $\hat{K}$, es propuesto para estimar el número de clusters, K, existentes en una población heterogénea. Una ley límite normal es rigurosamente probada para dicho estimador. La demostración utiliza un método de Holst (1979). Un ejemplo para un conjunto de datos reales y un estudio realizado por simulación es presentado para el estimador propuesto.

A little law for a natural estimator of number of clusters in a population.

Palabras clave: Clusters, población heterogénea, ley límite normal.

Clasificación AMS: 1162G05

* Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/Madrid, 126. 28903 Getafe (Madrid)
– Recibido en octubre de 1997.
– Aceptado en abril de 1998.

1. INTRODUCCIÓN

Sea una población constituida por un número desconocido K de clusters. Existe una gran cantidad de trabajos en la literatura estadística sobre los métodos de estimación del número de clusters, pero la mayoría han sido desarrollados en torno a la idea de que las probabilidades de observación de los diferentes clusters son iguales. Ver, por ejemplo, Lewontin y Prout (1956), Darroch (1958), Harris (1968), Johnson y Kotz (1977), Marchand y Schroeck (1982) Darroch y Ratcliff (1980), Holst (1981) y Esty (1985).

Existe un concepto que está muy ligado con el de números de clusters de una población, que es el cubrimiento muestral. Se define como la suma de las probabilidades de los clusters observados en una muestra. En el caso de clusters igualmente probables, el cubrimiento viene dado por el número de clusters observados en una muestra, D, dividido por el número de clusters que constituyen la población, K. Darroch y Ratcliff (1980) utilizaron exactamente la idea del cubrimiento muestral para estimar K.

Ahora bien, considerar la hipótesis de que las probabilidades de los distintos clusters son iguales es, en principio, un caso muy particular y poco frecuente, ya que poblaciones con clusters constituidos por una misma cantidad de elementos es prácticamente imposible. Por ejemplo, no existe una misma cantidad de animales para cada especie en un ecosistema; no se repite con la misma frecuencia cada una de las diferentes palabras que constituyen un texto; no se acuña la misma cantidad de las distintas monedas utilizadas en un país durante un centenario, etc.

La mayoría de los trabajos realizados para poblaciones heterogéneas (es decir, constituidos por clusters no equiprobables) adoptan un enfoque paramétrico. Por ejemplo, Fisher, Corbet y Williams (1943) asumen que para cada cluster, el número de observaciones en la muestra se distribuye según una distribución de Poisson, y el parámetro de dicha distribución se asume que sigue una distribución Gamma. Muchos otros artículos sobre modelos de abundancia de especies en un ecosistema también hacen consideraciones paramétricas. Ver, por ejemplo, McNeil (1973), Engen (1978), Efron y Thisted (1976). Esty (1985) estima el número de clusters en una población heterogénea mediante el concepto de cubrimiento muestral, aunque bajo un modelo paramétrico. Chao y Shen-Ming Lee (1992) propone una técnica de estimación no paramétrica, utilizando también la idea del cubrimiento muestral. Pero hay que subrayar que ninguno de los autores mencionados, como sí hacen algunos autores en el caso equiprobable, estudian cuál es la distribución asintótica del estimador que proponen.

La propuesta de este artículo es justamente el estudio de la distribución asintótica de un estimador para K. Aunque el estimador que aquí se propone es sesgado, lo importante es subrayar la técnica empleada para llegar a dicha distribución, la cual puede ser utilizada próximamente para otros estimadores.

Por tanto, considérese una población cerrada en la cual las observaciones están agrupadas en K clusters. El significado de cerrada hace alusión a que durante el estudio no se producen entradas o salidas de los clusters existentes. A partir de la información obtenida de una muestra aleatoria de tamaño n se propone en el apartado 2 un estimador natural-sesgado para K, $\hat{K}$. El cálculo de su esperanza matemática va a ser importante para cálculos posteriores. Ver el apartado 2.1. Justamente es el apartado 3 el de gran interés. Se estudia la distribución asintótica del estimador propuesto, aplicando un método de Holst (1979). Se prueba que el estimador se distribuye asintóticamente como una normal. En el último apartado se presenta un estudio realizado por simulación para el estimador propuesto. Además se da un ejemplo para un conjunto de datos reales, el cual han sido aplicado por otros autores. A la vista de los resultados se proponen técnicas de reducción del sesgo del estimador.

2. UN ESTIMADOR NATURAL SESGADO

Asúmase que una muestra aleatoria de tamaño n con reemplazamiento ha sido extraída de la población, la cual está formada por K clusters. La probabilidad de observar el cluster j es $p_j \geqslant 0$, con $j = 1, \ldots, K$ y $\sum_{j=1}^{K} p_j = 1$.

Un estimador natural, sesgado y que bajo-estima K cuando éste es grande con respecto a n es: $\hat{K} = \sum_{j=1}^{K} I_j$, donde

$$I_j = \begin{cases} 1 & \text{si el cluster } j \text{ es observado en la muestra.} \\ 0 & \text{en otro caso.} \end{cases}$$

2.1. Momentos del estimador natural

Son presentados a continuación los operadores esperanza y varianza del estimador $\hat{K}$. El segundo no tiene más interés que saber cuál es la varianza del estimador propuesto. En cambio el primero es de gran importancia por su utilización en el cálculo de la distribución asintótica de $\hat{K}$.

2.1.1. La esperanza de $\hat{K}$

Teorema. *La esperanza del estimador natural* $\hat{K}$ *viene expresada por*

$$E(\hat{K}) = K - \sum_{j=1}^{K} (1-p_j)^n = K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

siendo $F(x)$ *una función de distribución.*

Demostración. Se tiene que:

(1)

$$E(\hat{K}) = E\left(\sum_{j=1}^{K} I_j\right) = \sum_{j=1}^{K} p\,(I_j = 1) = \sum_{j=1}^{K} [1 - p(I_j = 0)] = K - \sum_{j=1}^{K} (1-p_j)^n.$$

A continuación se demuestra que (1) se puede expresar como:

$$K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

siendo $F(x)$ una función de distribución, dada en la demostración.

■

El interés que tiene dicha expresión es su utilización en el cálculo de la distribución asintótica.

Considérese la expresión:

$$\frac{\sum_{j=1}^{K} \left[1 - (1-p_j)^n\right] - \sum_{j=1}^{K} \left[1 - e^{-np_j}\right]}{\sum_{j=1}^{K} \left[1 - e^{-np_j}\right]},$$

donde $0 \leqslant p_j \leqslant 1$ y $\sum_{j=1}^{K} p_j = 1$. Ver el artículo de Harris (1968) donde se utiliza esta expresión, y demostrando que $E(\hat{K}) \cong \sum_{j=1}^{K} \left(1 - e^{-np_j}\right)$. Para ello se aplica el siguiente lema.

Lema. *Si* $a_i,\ b_i > 0, \quad i = 1, 2, \ldots,$ *y* $\dfrac{1}{b} = \sup_i \dfrac{a_i}{b_i}$, *entonces* $\dfrac{a}{b} \geqslant \dfrac{\sum_i a_i}{\sum_i b_i}$. *Entonces se*

tiene que:

$$\frac{\sum_{j=1}^{K}\left[1-(1-p_j)^n\right]-\sum_{j=1}^{K}\left[1-e^{-np_j}\right]}{\sum_{j=1}^{K}\left[1-e^{-np_j}\right]} \leqslant \sup_{p_j}\frac{e^{-np_j}-(1-p_j)^n}{1-e^{-np_j}}=\frac{e^{-np}-(1-p)^n}{1-e^{-np}},$$

donde p es justamente una de las p_j's donde se alcanza dicho supremo.

Como $(1-p)^n = e^{n\log(1-p)} = e^{-np-\frac{np^2}{2}}$, entonces

$$\frac{e^{-np}-(1-p)^n}{1-e^{-np}} \leqslant \frac{e^{-np}-e^{-np}\dfrac{np^2}{2}}{1-e^{-np}} \leqslant \frac{e^{-np}\left(1-e^{-\frac{np^2}{2}}\right)}{1-e^{-np}}.$$

Considérese dos casos posibles para p: cuando $p < 1/\sqrt{n}$ y cuando $p \geqslant 1/\sqrt{n}$ ($n \longrightarrow \infty$ en ambos casos).

Si $p \geqslant \dfrac{1}{\sqrt{n}}$, entonces

$$\frac{e^{-np}-(1-p)^n}{1-e^{np}} \leqslant \frac{e^{-np}}{1-e^{-np}} \leqslant \frac{e^{-\sqrt{n}}}{1-e^{-\sqrt{n}}},$$

que tiende a cero cuando $n \longrightarrow \infty$. Por consiguiente, y teniendo en cuenta la expresión de partida, se tiene que:

$$\sum_{j=1}^{K}\left[1-(1-p_j)^n\right] \cong \sum_{j=1}^{K}\left[1-e^{-np_j}\right]. \tag{2}$$

Si $p < \dfrac{1}{\sqrt{n}}$, considérese

$$(1-p)^n = e^{n\log(1-p)} = e^{-np-\frac{np^2}{2}} \cong e^{-np-\frac{np^2}{2(1-x)^2}} \qquad (0 \leqslant x \leqslant p).$$

Entonces:

$$\begin{aligned}\frac{e^{-np}-(1-p)^n}{1-e^{-np}} &\leqslant \frac{e^{-np}-e^{-np-\frac{np^2}{2(1-(1/\sqrt{n}))^2}}}{1-e^{-np}} = \\ &= \frac{e^{-np}\left(1-e^{-np-\frac{-n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{1-e^{-np}}\end{aligned}$$

Llamando $h_n(p)$ a esta última expresión, calculando la función derivada e igualando a cero, se obtiene:

$$\begin{aligned}
h_n'(p) &= \left(\frac{-ne^{-np}\left(1-e^{\frac{-n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{(1-e^{-np})^2} + \frac{e^{-np}\left(\frac{n^2 2p}{2(\sqrt{n}-1)^2}e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{(1-e^{-np})^2}(1-e^{-np}) - \right. \\
&- \left. \frac{\left(e^{-np}\left(1-e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)n(e^{-np})\right)}{(1-e^{-np})^2} \right) = \frac{-ne^{-np}\left(1-e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)(1-e^{-np})}{(1-e^{-np})^2} + \\
&+ \frac{e^{-np}(1-e^{-np})\frac{n^2 2p}{2(\sqrt{n}-1)^2}e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}}{(1-e^{-np})^2} - \frac{e^{-np}\left(1-e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)ne^{-np}}{(1-e^{-np})^2} = \\
&= \frac{-ne^{-np}\left(1-e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)+e^{-np}(1-e^{-np})\frac{n^2 2p}{2(\sqrt{n}-1)^2}e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}}{(1-e^{-np})^2} = 0.
\end{aligned}$$

Dividiendo por ne^{-np} y sacando factor común a $e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}$,

$$-1+e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\left(1+p\frac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right)\right)=0.$$

Tomando logaritmos,

$$\frac{-n^2p^2}{2(\sqrt{n}-1)^2}+\log\left[1+p\frac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right)\right]=0.$$

Como el segundo sumando se puede aproximar por

$$p\frac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right),$$

haciendo $\log(1+p)=p-(p^2/2)+(p^3/3)-\cdots\cong p+0(p)$, entonces

$$\frac{-n^2p^2}{2(\sqrt{n}-1)^2}+p\frac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right)\cong 0.$$

Así, $p\dfrac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right) \cong \dfrac{n^2 p^2}{2(\sqrt{n}-1)^2}$, que es equivalente a decir

$$1-e^{-np} \cong \frac{np}{2}.$$

Al resolver dicha ecuación computacionalmente se obtiene que el máximo de $h_n(p)$ es $p \cong (1,\ 6/n)$. Por consiguiente,

$$h_n\left(\frac{1,6}{n}\right) \cong \frac{e^{-1,6}\left(1-e^{\frac{-2,56}{2(\sqrt{n}-1)^2}}\right)}{1-e^{-1,6}} \longrightarrow 0, \quad \text{cuando} \quad n \longrightarrow \infty.$$

De esta manera se llega a la misma conclusión que (2), es decir,

$$\sum_{j=1}^{K}\left(1-(1-p_j)^n\right) \cong \sum_{j=1}^{K}\left(1-e^{-np_j}\right).$$

Con esto,

$$E(\hat{K} \cong \sum_{j=1}^{K}\left(1-e^{-np_j}\right).$$

Supóngase ahora que cuando K y n tiende a infinito, con las p_j's distintas, la distribución empírica de $np_1,\ np_2,\ \ldots,\ np_k$, definida como

$$F_n(x) = \frac{1}{K}\sum_{j=1}^{K} I\left(n\,p_j \leqslant x\right)$$

converge en probabilidad a $F(x)$ sobre $(0,\ \infty)$. $I(A)$ es la conocida función indicadora. Entonces, se tiene que

$$E(\hat{K}) \cong \sum_{j=1}^{K}\int_0^{\infty}\left(1-e^{-x}\right)\,dF(x) = K\int_0^{\infty}\left(1-e^{-x}\right)\,dF(x).$$

Se define X_j como una variable aleatoria que indica el número de veces que se ha observado el cluster j en la muestra, con $j=1,\ \ldots,\ K$; y considérese el siguiente lema de Holst (1979).

Lema. *$P(x_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_k = x_k) = p\left(Y_1 = x_1, Y_2 = x_2, \ldots, Y_k = \right.$*
$\left. = x_k \Big/ \sum_{j=1}^{K} Y_j = n\right)$, donde $\{Y_n\}$ son variables aleatorias independientes de Poisson con media np_j.

Entonces:

$$\begin{aligned} E(\hat{K}) &= E\left(\sum_{j=1}^{K} I(X_j > 0)\right) = E\left(\sum_{j=1}^{K} I(Y_j > 0)\right) = \sum_{j=1}^{K} Prob\,(Y_j > 0) = \\ &= \sum_{j=1}^{K} \left[1 - ProbI(Y_j = 0)\right] = \\ &= \sum_{j=1}^{K} \left[1 - e^{-np_j}\right] \cong \sum_{j=1}^{K} \int_0^{\infty} \left[1 - e^{-np_j}\right] dF(x) = \\ &= K \int_0^{\infty} \left[1 - e^{-np_j}\right] dF(x), \end{aligned}$$

justamente lo que se quería demostrar.

2.1.2. La varianza de $\hat{K}$

Se tiene que:

$$\mathrm{var}\,(\hat{K}) = E\,(\hat{K}^2) - E^2(\hat{K}).$$

La esperanza de $\hat{K}$ ha sido calculada anteriormente. Ahora queda por determinar quién es $E(\hat{K}^2)$.

$$E(\hat{K}^2) = \sum_{j=1}^{K}\sum_{l=1}^{K} p(I_j = I_l = 1) = \sum_{j=1}^{K}\sum_{l=1}^{K} \left\{p(I_j = 1) + p(I_l = 1) - p(I_j \text{ o } I_l = 1)\right\}.$$

Como $p(I_j = 1) = 1 - p(I_j = 0) = 1 - (1 - p_j)^n$, y la probabilidad de elegir el cluster j o el cluster l es $p_j + p_l$ $(j \neq l)$, entonces

$$p\,((I_j = 1) \quad \text{o} \quad (I_l = 1)) = 1 - (1 - (p_j + p_l)^n), \quad j \neq 1.$$

Pero si $l = j$, entonces, $(I_1 = 1) \cup (I_l = 1) = (I_l = 1)$, y

$$p\,((I_l = 1) \quad \text{o} \quad (I_l = 1)) = 1 - (1 - p_j)^n.$$

Por consiguiente:

$$E(\hat{K}^2) = 2K\sum_{l=1}^{K} p(I_l = 1) - \sum_{\substack{j=1 \\ j \neq l}}^{K}\sum_{l=1}^{K} \left\{p(I_j = 1) + p(I_l = 1)\right\} - \sum_{l=1}^{K} p(I_l = 1) =$$

$$= (2K-1)\left(K-\sum_{l=1}^{K}(1-p_l)^n\right)-K(K-1)+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n =$$

$$= K^2(2K-1)\sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n.$$

Por tanto,

$$\begin{aligned}
\operatorname{var}(\hat{K}) &= K^2-(2K-1)\sum_{l=1}^{K}(1-p_l)^n+ \\
&+ \sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n-\left[K-\sum_{j=1}^{K}(1-p_j)^n\right]^2 = \\
&= K^2-(2K-1)\sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n- \\
&- \left[K^2+\left(\sum_{j=1}^{K}(1-p_j)^n\right)^2-2K\sum_{j=1}^{K}(1-p_j)^n\right] = \\
&= \sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n-\left\{\sum_{j=1}^{K}(1-p_j)^n\right\}^2.
\end{aligned}$$

Hay que notar que dicha expresión coincide por la dada por McNeil (1973).

3. DISTRIBUCIÓN ASINTÓTICA

A continuación se prueba la normalidad asintótica de $\hat{K}$ mediante el método de Holst (1979) (ver también Esty (1985)).

Teorema. *La distribución asintótica de la expresión*

$$K^{-1/2}\left(\hat{K}-E\left(\hat{K}\right)\right)$$

converge a una distribución normal de media cero y varianza σ_1^2, la cual está dada en la demostración.

Demostración. Nótese que $\hat{K} = K - N_0$, donde N_0 es una variable aleatoria que indica el número de clusters no observados en la muestra que se define como $N_0 = \sum_{j=1}^{K} I(X_j = 0)$ y $E(N_0)$, utilizando la variable aleatoria indicatriz

$$(3) \qquad Z_j^{(i,n)} = \begin{cases} 1 & \text{si el cluster } j \text{ ocurre } i \text{ veces en la muestra.} \\ 0 & \text{en otro caso.} \end{cases}$$

es igual a $\sum_{j=1}^{K} (1-p_j)^n$. Entonces:

$$\hat{K} - E(\hat{K}) = \hat{K} - \left[K - \sum_{j=1}^{K} (1-p_j)^n\right] = -N_0 + \sum_{j=1}^{K} (1-p_j).$$

Sea:

$$f(X_j) = [I(X_j = 0) - (1-p_j)^n].$$

Se define

$$Z_M = \sum_{j=1}^{K} f(x_j), \quad M < K.$$

Obsérvese que si M es todo K,

$$Z = Z_M = \sum_{j=1}^{K} f(X_j) = \sum_{j=1}^{K} [I(X_j) - (1-p_j)^n] = N_0 - \sum_{j=1}^{K} (1-p_j)^n.$$

Ahora, el problema consiste en encontrar la distribución asintótica de $Z = \sum_{j=1}^{K} f(X_j)$. Para ello se va a seguir el método de Holst (1979), demostrando que la función característica de

$$K^{-1/2}\left(N_0 - \sum_{j=1}^{K} (1-p_j)^n\right)$$

converge a una distribución normal de media cero y varianza σ_1^2, dada en la demostración. Para ello se prueba primero a continuación cuál es la distribución asintótica de $K^{-1/2} Z_M$.

Considérese de nuevo el lema enunciado en el apartado anterior:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k \Big/ \sum_{j=1}^{K} Y_j = n\right),$$

donde $\{Y_j\}$ son variables aleatorias independientes de Poisson con media np_j. Entonces:

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

■

Considérese ahora el siguiente lema de Holst (1979).

Lema. *Si Si* (U, V) *es un vector bidimensional con* U *entero, entonces*

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

Entonces,

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

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

Ahora bien, como $E\left(\sum_{j=1}^{K} Y_j\right) = \sum_{j=1}^{K} n\,p_j = n$ y $n! = e^{-n}\sqrt{2\pi n}\,n^n$, entonces

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

Haciendo el cambio de variable $t = u\sqrt{n}$,

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

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

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

Sea

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

donde

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

y

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

Ahora,

$$\begin{aligned}
h_{2n}(t) &= \prod_{j=M+1}^{K}\sum_{m=0}^{\infty} e^{itn^{-1/2}(m-np_j)}\,e^{-np_j}\frac{(n\,p_j)^m}{m!} = \\
&= \prod_{j=M+1}^{K} e^{-itn^{-1/2}np_j}\sum_{m=0}^{\infty} e^{itn^{-1/2}m}\,e^{-np_j}\frac{(n\,p_j)^m}{m!} = \\
&= \prod_{j=M+1}^{K} e^{-itn^{-1/2}\,p_j}\,e^{-np_j}\sum_{m=0}^{\infty}\frac{\left(e^{itn^{-1/2}}\,np_j\right)^m}{m!} = \\
&= \prod_{j=M+1}^{K} e^{-itn^{-1/2}\,p_j}\,e^{-np_j}\,e^{np_j e^{itn^{-1/2}}} = \\
&= \prod_{j=M+1}^{K} e^{-itn^{-1/2}\,p_j}\,e^{-np_j\left(e^{itn^{-1/2}}-1\right)} \\
&= e^{-itn^{-1/2}\sum_{j=M+1}^{K}p_j}\,e^{n\sum_{j=M+1}^{K}p_j\left(e^{itn^{-1/2}}-1\right)}
\end{aligned}$$

Como $e^{itn^{-1/2}} = 1 + (it/\sqrt{n}) - \left(t^2/2n\right) + 0(n)$, entonces:

$$\begin{aligned} h_{2n}(t) &= e^{-itn^{-1/2}\sum_{j=M+1}^{K} p_j} \; e^{n\sum_{j=M+1}^{K} p_j \left(1 + \left(it/\sqrt{n}\right) - \left(t^2/2n\right) - 1\right)} = \\ &= e^{-itn^{-1/2}\sum_{j=M+1}^{K} p_j} \; e^{\sum_{j=M+1}^{K} p_j\, itn^{-1/2} - n\sum_{j=M+1}^{K} p_j \left(t^2/2n\right)} = \\ &= e^{-\sum_{j=M+1}^{K} p_j \left(t^2/2n\right)} \end{aligned}$$

Considerando $h_{1n}(s,\, t)$, se tiene:

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

donde

$$\begin{aligned} g_j(s,\, t) &= E\left(e^{itn^{-1/2}\left(Y_j - np_j\right) + isK^{-1/2} f(Y_j)}\right) = \\ &= E\left(e^{itn^{-1/2}\left(Y_j - np_j\right) + isK^{-1/2} I(Y_j = 0)}\right)\left(e^{-isK^{-1/2}\left(1 - p_j\right)^n}\right). \end{aligned}$$

Ahora bien, el primer factor es igual a:

$$\begin{aligned} & E\left(e^{itn^{-1/2}\left(Y_j - np_j\right) + isK^{-1/2} I(Y_j = 0)}\right) = \\ & = e^{-itn^{-1/2} p_j + isK^{-1/2}} e^{-np_j} + \\ & + e^{-itn^{-1/2} p_j} e^{-np_j} \left\{ \sum_{R=0}^{\infty} \frac{(np_j)^R \left(e^{itn^{-1/2}}\right)^R}{R!} - 1 \right\} = \\ & = e^{-np_j} e^{-itn^{-1/2} p_j} \left(e^{isK^{-1/2}} - 1\right) + e^{-itn^{-1/2} p_j} e^{-np_j} e^{np_j e^{itn^{-1/2}}} \end{aligned}$$

El primer sumando se puede poner de la forma:

$$\begin{aligned} & e^{-np_j}\left(1 - itn^{-1/2} p_j - \frac{nt^2 p_j^2}{2}\right)\left(isK^{-1/2} - \frac{s^2}{2K}\right) = \\ & = e^{-np_j}\left\{isK^{-1/2} - \frac{s^2}{2K} + \frac{ts}{n^{1/2} K^{1/2}}(np_j)\right\} + 0\left(K^{-1}\right). \end{aligned}$$

Y el segundo sumando se puede escribir recordando el desarrollo de $h_{2n}(t)$, como $e^{(-p_j/2)t^2}$.

Por consiguiente,

$$E\left(e^{itn^{-1/2}\left(Y_j-np_j\right)+isK^{-1/2}I(Y_j=0)}\right) =$$

$$= e^{-np_j}\left\{isK^{-1/2}-\frac{s^2}{2K}+\frac{ts}{n^{1/2}K^{1/2}}(np_j)\right\}+e^{(-p_j/2)t^2}+0\left(K^{-1}\right).$$

El segundo factor, teniendo en cuenta que

$$\begin{aligned}(1-p_j)^n &= e^{\log\left(1-p_j\right)^n}\cong e^{-np_j},\\ e^{-isK^{-1/2}\left(1-p_j\right)^n} &= 1-isK^{-1/2}e^{-np_j}-\frac{s^2}{2K}e^{-2np_j}+0\left(K^{-1}\right).\end{aligned}$$

De esta forma,

$$\begin{aligned}g_j(s,t) &= e^{-\left(p_j/2\right)t^2}\times\\ &\times \left\{1+e^{(p_j/2)t^2}e^{-np_j}\left(isK^{-1/2}-\frac{s^2}{2K}+\frac{ts}{n^{1/2}K^{1/2}}(np_j)\right\}\times\right.\\ &\times \left\{1-isK^{-1/2}e^{-np_j}-\frac{s^2}{2K}e^{-2np_j}+0\left(K^{-1}\right)\right\}\end{aligned}$$

Haciendo $e^{(p_j/2)t^2}=1+0(p_jt^2)$,

$$\begin{aligned}g_j(s,t) &= e^{-(p_j/2)t^2}\left\{1-isK^{-1/2}e^{-np_j}-\frac{s^2}{2K}e^{-2np_j}+e^{-np_j}iKsK^{-1/2}-\right.\\ &- \left.\frac{s^2}{2K}e^{-np_j}+\frac{ts}{n^{1/2}K^{1/2}}e^{-np_j}(np_j)+e^{-2np_j}s^2K^{-1}\right\}=\\ &= e^{-(p_j/2)t^2}\left\{1+\frac{s^2}{2K}e^{-2np_j}-\frac{s^2}{2K}e^{-np_j}+\frac{ts(np_j)}{n^{1/2}K^{1/2}}e^{-np_j}\right\}\end{aligned}$$

Por tanto:

$$\begin{aligned}\prod_{j=1}^{M}g_j(s,t) &= e^{-t^2/2\sum_{j=1}^{M}p_j}\times\\ &\times \prod_{j=1}^{M}\left\{1-\frac{s^2}{2K}e^{-np_j}+\frac{s^2}{2K}e^{-2np_j}+\frac{ts(np_j)}{n^{1/2}K^{1/2}}e^{-np_j}\right\}\cong e^{-t^2/2\sum_{j=1}^{M}p_j}\times\end{aligned}$$

$$\times \quad e^{-\sum_{j=1}^{M}\left[\left(s^2/2K\right)e^{-np_j}-\left(s^2/2K\right)e^{-2np_j}\right]+\sum_{j=1}^{M}\left(tsnp_j/n^{1/2}K^{1/2}\right)e^{-np_j}}$$

Entonces:

$$H_n(s) = \frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} e^{-t^2/2\left(\sum_{j=1}^{M}p_j+\sum_{j=M+1}^{K}p_j\right)} e^{\sum_{j}^{M}\left(tsnp_j/n^{1/2}K^{1/2}\right)e^{-np_j}}\, dt \times$$

$$\times \quad e^{-s^2/2\left\{(1/K)\sum_{j=1}^{M}e^{-np_j}-(1/K)\sum_{j=1}^{M}e^{-2np_j}\right\}} =$$

$$= \int_{-\pi n^{1/2}}^{+\pi n^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-1/2\left\{t-\frac{\sum_{j}^{M}n\,p_j e^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2} dt \times$$

$$\times \quad e^{-\left(s^2/2\right)\left\{\sum_{j=1}^{M}(np_j)e^{-np_j}\right\}^2\Big/nK} \quad e^{(-s^2/2)\sum_{j=1}^{M}\left\{(1/K)e^{-np_j}-(1/K)e^{-2np_j}\right\}}$$

Si el número de clusters en la población y el tamaño de la muestra son sumamente grandes, tomando el límite de $H_n(s)$ cuando n y K tienden a infinito, se tiene:

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

$$\times \quad \lim_{n,K\longrightarrow\infty}\left\{ e^{-\left(s^2/2\right)\left\{\sum_{j=1}^{M}(np_j)e^{-np_j}\right\}^2\Big/nK} \times \right.$$

$$\left. \times \quad e^{-\left(s^2/2\right)\sum_{j=1}^{M}\left\{(1/K)e^{-np_j}-(1/K)e^{-2np_j}\right\}} \right\}$$

Aplicando el teorema de convergencia dominada:

$$\lim_{n,K\longrightarrow\infty}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}\frac{1}{\sqrt{2\pi}}e^{-1/2\left\{t-\frac{\sum_j^M np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2}dt=$$

$$=\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{\sqrt{2\pi}}e^{-1/2\left\{t-\frac{\sum_j^M np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2}\right\}dt=$$

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

Por consiguiente:

$$\lim_{n,K\longrightarrow\infty}H_n(s)=$$

$$=\lim_{n,K\longrightarrow\infty}e^{(-s^2/2)\left\{\sum_{j=1}^M(1/K)e^{-np_j}\sum_{j=1}^M(1/K)e^{-2np_j}-\sum_{j=1}^M(1/nK)\left((np_j)e^{-np_j}\right)^2\right\}}=$$

$$=e^{-(s^2/2)\lim_{n,K\longrightarrow\infty}\left\{\sum_{j=1}^M(1/K)e^{-np_j}\sum_{j=1}^M(1/K)e^{-2np_j}-\sum_{j=1}^M(1/nK)\left((np_j)e^{-np_j}\right)^2\right\}}$$

Así, $K^{-1/2}Z_M$ se distribuye asintóticamente como una $N(0,\sigma_M^2)$, con

$$\sigma_M^2=\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{K}\sum_{j=1}^M e^{-np_j}-\frac{1}{K}\sum_{j=1}^M e^{-2np_j}-\frac{1}{nK}\left\{\sum_{j=1}^M np_je^{-np_j}\right\}^2\right\}.$$

Ahora bien, $Z=Z_M+Z_{MC}$, siendo $Z_{MC}=\sum_{j=M+1}^K f(X_j)$. Se puede probar exactamente igual que $K^{-1/2}Z_{MC}$ se distribuye asintóticamente según una $N(0,\sigma_{MC}^2)$, donde

$$\sigma_{MC}^2=\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{K}\sum_{j=M+1}^K e^{-np_j}-\frac{1}{K_j}\sum_{j=M+1}^K e^{-2np_j}-\frac{1}{nK}\left\{\sum_{j=M+1}^K np_je^{-np_j}\right\}^2\right\}.$$

Así,

$$K^{-1/2}\left(N_0 - \sum_{j=1}^{K}(1-p_j)^n\right) \longrightarrow N(0,\,\sigma_1^2),$$

donde $\sigma_1^2 = \sigma_M^2 + \sigma_{MC}^2$, tal que

$$\sigma_1^2 = \lim_{n,K \longrightarrow \infty}\left\{\frac{1}{K}\sum_{j=1}^{K} e^{-np_j} - \frac{1}{K}\sum_{j=1}^{K} e^{-2np_j} - \frac{1}{nK}\left\{\sum_{j=1}^{K} np_j e^{-np_j}\right\}^2\right\}$$

y, por consiguiente,

$$\left(\sum_{j=1}^{K}(1-p_j)^n - N_0\right) \longrightarrow N(0,\,\sigma_1^2(\hat{K})),$$

donde $\sigma_1^2(\hat{K}) = K\sigma_1^2$.

Supóngase ahora, como se hizo en el apartado anterior, que la distribución empírica de $np_1,\ np_2,\ \ldots,\ np_k$, definida como

$$F_n(x) = \frac{1}{K}\sum_{j=1}^{K} I(np_j \leqslant x)$$

converge débilmente a $F(x)$ sobre $(0,\,\infty)$. Entonces,

$$\begin{aligned}
\sigma_1^2 &= \frac{1}{K}\sum_{j=1}^{K}\int_0^\infty e^{-x}\,dF(x) - \frac{1}{K}\sum_{j=1}^{K}\int_0^\infty e^{-2x}\,dF(x) - \\
&- \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(\sum_{j=1}^{K}\int_0^\infty (xe^{-x})\,dF(x)\right)^2 = \\
&= \int_0^\infty \left(e^{-x}\left(1-e^{-x}\right)\right)\,dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(\sum_{j=1}^{K}\int_0^\infty (xe^{-x})\,dF(x)\right)^2,
\end{aligned}$$

ya que:

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

Una cota superior aparente para la varianza asintótica de $\hat{K}$ es:

$$\begin{aligned}
\sigma_1^2(\hat{K}) &= K\left(\int_0^\infty \left(1-e^{-x}\right)\,dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(K\int_0^\infty \left(xe^{-x}\right)\,dF(x)\right)^2\right) = \\
&= K\int_0^\infty \left(1-e^{-x}\right)\,dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1} K\left(K\int_0^\infty \left(xe^{-x}\right)\,dF(x)\right)^2.
\end{aligned}$$

Sea la variable N_i que se define como el número de clusters que se observan exactamente i veces en la muestra. A partir de la variable aleatoria indicatriz definida en (3),

$$E(N_i) = \sum_{j=1}^{K} \left(\begin{array}{c} n \\ i \end{array} \right) p_j^i (1-p_j)^{n-i} \cong \sum_{j=1}^{K} e^{-np_j} \frac{(np_j)^i}{i!} \cong \sum_{j=1}^{K} \int_0^{\infty} \left(e^{-x} \frac{x}{i!} \right) dF(x),$$

tal que

$$i!\,E(N_i) \cong K \int_0^{\infty} \left(e^{-x} \frac{x}{i!} \right) dF(x),$$

y como $E(\hat{K}) \cong K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x)$, se obtiene que,

$$\sigma_1^2(\hat{K}) = E(\hat{K}) - \frac{E^2(N_i)}{E\left(\sum_{j=1}^{K} X_j \right)}.$$

Por consiguiente, un estimador de $\sigma_1^2(\hat{K})$ es

$$\hat{\sigma}_1^2(\hat{K}) = \hat{k} - \left(n_1^2/n\right),$$

al reemplazar las esperanzas por los valores observados.

4. RESULTADOS NUMÉRICOS

Antes de presentar los resultados numéricos es necesario indicar que, aunque el objetivo principal de este artículo es la propuesta de presentar una técnica de obtención de una distribución asintótica para un estimador del número de clusters en una población, es necesario y conveniente resaltar como corregir $\hat{K}$. Existen técnicas jackknife y bootstrap que corrigen y ajustan el estimador según el sesgo cometido. En Prieto (1998, a y b) se presentan justamente estas técnicas como reducción y corrección del sesgo. También ha sido utilizadas por Burnham y Overton (1979) para estimar el número de individuos en una población, o por Heltshe y Forrester (1983) para estimar el número de especies en un ecosistema.

Ejemplo 1

Este ejemplo es propuesto por Fisher, Corbet y Williams (1943). 1421 especies fueron cogidas en una trampa - muestra en la localidad de Rothamsted y clasificadas

por especies. Los datos se resumen en la siguiente tabla.

i	1	2	3	4	5	6	7	≥ 8
n_i	35	11	15	14	10	11	5	139

Fisher, Corbet y Williams estimaron el número de especies en 261.9. $\hat{K} = 240$, y la varianza es

$$\hat{\sigma}_1^2(\hat{K}) = \hat{k} - \left(n_1^2/n\right) = 240 - \frac{35^2}{1421} \cong 240.$$

Obsérvese que n debe tender a infinito cuando K es sumamente grande. A continuación se presenta un ejemplo por simulación para ver la eficiencia de $\hat{K}$ según las probabilidades de observacion de cada cluster.

Ejemplo 2

Entonces para comprobar la eficacia del estimador propuesto $\hat{K}$ se ha evaluado mediante métodos computacionales por simulación. La evaluación de $\hat{K}$ ha sido llevada a cabo simulando una muestra aleatoria o bien de tamaño 50 o bien de tamaño 100 de una población de 200 clusters.

Las probabilidades de observar los diferentes clusters han sido consideradas pertenecientes al intervalo $[0.0020; 0.01]$. Se han considerado 6 casos posibles. En el primer caso se ha considerado las probabilidades iguales. En el segundo los primeros 100 clusters tienen probabilidades 0.004 de ser observados y los 100 siguientes 0.006. Los siguientes casos se van considerando poblaciones más heterogéneas. Cada caso se ha simulado 50 veces y se han tomado el promedio de los resultados.

Los resultados obtenidos indican que:

- $\hat{K}$ siempre bajo estima el valor de K, siendo muy sesgado. Técnicas para corregir el estimador han sido ya mencionadas.
- Para cualquier población, el sesgo cometido por $\hat{K}$ cuando $n = 50$ es siempre mayor que cuando $n = 100$.
- El sesgo de cada estimador aumenta a medida que la población es más heterogénea.
- La varianza de $\hat{K}$ cuando $n = 50$ es má pequeña que cuando $n = 100$. A medida que la población es más heterogénea, la varianza es ligeramente mayor.

Tabla 1

Casos	n	p_j	$\hat{K}$	$\hat{\sigma}_1^2$	$E(\hat{K}-K)$	E.C.M.	$V(\hat{K})$
1	50	$p_j = 0.005$ $j = 1-200$	119	48.91	-81	6593.14	32.14
	100		132	53.41	-68	4662.09	38.09
2	50	$p_j = 0.004$ $j = 1-100$ $p_j = 0.006$ $j = 101-200$	103	44.12	-97	9447.01	38.01
	100		118	48.43	-82	6768.80	44.80
3	50	$p_j = 0.0035$ $j = 1-90$ $p_j = 0.0045$ $j = 91-180$ $p_j = 0.014$ $j = 181-200$	94	39.10	-106	11274.10	38.14
	100		107	43.41	-93	8694.04	45.04
4	50	$p_j = 0.01$ $j = 1-10$ $p_j = 0.004$ $j = 11-100$	82	41.27	-118	13967.20	43.24
	100	$p_j = 0.003$ $j = 101-190$ $p_j = 0.023$ $j = 191-200$	97	46.92	-103	10657.10	48.12

Continuación

Tabla 1 (cont.)

Casos	n	p_j	$\hat{K}$	$\hat{\sigma}_1^2$	$E(\hat{K}-K)$	E.C.M.	$V(\hat{K})$
5	50	$p_j = 0.0035$	67	44.12	-133	17735.20	46.28
		$j = 1-50$					
		$p_j = 0.006$					
		$j = 151-100$					
		$p_j = 0.002$					
		$j = 101-125$					
	100	$p_j = 0.009$	85	48.43	-115	13274.20	49.21
		$j = 126-150$					
		$p_j = 0.005$					
		$j = 151-200$					
6	50	$p_j = 0.006$	56	48.91	-144	20785.20	49.24
		$j = 1-25$					
		$p_j = 0.0025$					
		$j = 26-50$					
		$p_j = 0.009$					
		$j = 51-75$					
	100	$p_j = 0.008$	69	53.41	-131	17216.30	55.39
		$j = 76-100$					
		$p_j = 0.001$					
		$j = 101-125$					
		$p_j = 0.002$					
		$j = 126-150$					
		$p_j = 0.005$					
		$J = 151-175$					
		$p_j = 0.004$					
		$j = 176-200$					

AGRADECIMIENTOS

Quiero agradecer a Dr. Anne Chao (Institute of Statistics in National Tsing Hua University, Hsin-Chu, Taiwan) su importante colaboración en este artículo.

BIBLIOGRAFÍA

[1] **Burnham, K.P.** y **Overton, W.S.** (1978). «Estimation of the size of a closed population when capture probabilities vary among animals». *Biometrika*, **63, 3**, 625-633.

[2] **Chao, A.** y **Shen-Ming Lee.** (1992). «Estimating the number of classes via sample coverage». *Journal of the American Statistical Association*, **87, N-417**, 211-217.

[3] **Darroch, J.N.** (1958). «The multiple recapture census I: Estimation of a closed population». *Biometrika*, **45**, 343-359.

[4] **Darroch, J.N.** y **Ratcliff, D.** (1980). «A note on capture-recapture estimation». *Biometrika*, **40**, 343-359.

[5] **Efron, B.** y **Thisted, R.** (1976). «Estimating the number of unseen species: How many words did Shakespeare know?». *Biometrika*, **63**, 435-447.

[6] **Engen, S.** (1978). *Stochastic Abundance Models*. London: Chapman-Hall.

[7] **Esty, W.W.** (1983). «A normal limit law for a nonparametric estimator of the coverage of a random sample». *The Annals of Statistic*, **11**, 905-912.

[8] **Esty, W.W.** (1985). «Estimation of the number of classes in a population and the coverage of a sample». *Mathematical Scientist*, **10**, 41-50.

[9] **Fisher, R.A., Corbet, A.S.** y **Williams, C.B.** (1943). «The relation between the number of species and the number of individuals in a random sample of a animal population». *Journal of Animal Ecology*, **12**, 42-58.

[10] **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.

[11] **Heltshe, J.F.** y **Forrester, N.E.** (1983). «Estimating species richess using the jackknife procedure». *Biometrics*, **39**, 1-11.

[12] **Holst, L.** (1979). «A unified approach to limit theorems for urn models». *Journal of Applied Probability*, **16**, 154-162.

[13] **Holst, L.** (1981). «Some asymptotic results for incomplete multinomial or poisson samples». *Scandinavian Journal of Statistic*, **8**, 243-246.

[14] **Johnson, N.L.** y **Kotz, S.** (1977). *Urn models and their applications: an approach to modern discrete probability theory*. New York: Yohn Wiley.

[15] **Lewontin, R.C.** y **Prout, T.** (1956). «Estimation of the number of different classes in a population». *Biometrics*, **12**, 211-223.

[16] **Marchand, J.P.** y **Schroeck, P.E.** (1982). «On the estimation of the number of equally likely classes in a population». *Communications in a Statistic, Part A-Theory and Methods*, **11**, 1139-1146.

[17] **McNeil, D.R.** (1973). «Estimating an author's vocabulary». *Journal of the American Statistical Association*, **68, N-341**, 92-96.

[18] **Prieto M., J.J.** (1998, a). «¿Cuántos clusters hay en una población?». *Qüestiió*, **1.**

[19] **Prieto M.. J.J.** (1998, b). «Estimación del número de clusters en una población aplicando el jackknife generalizado». *Qüestiió*, **2.**

[20] **Robbins, H.** (1968). «Estimating the total probability of the unobserved outcomes of an experiment». *Annals of mathematical Statistics*, **39**, 256-257.

ENGLISH SUMMARY

A LITTLE LAW FOR A NATURAL ESTIMATOR OF NUMBER OF CLUSTERS IN A POPULATION

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

A natural estimator, $\hat{K}$, is propousted to estimate the number of clusters, K, that there are in a heterogeneous population. A limit law normal is rigorously proved for $\hat{K}$. The proof utilizes 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 examples.

Keywords: Clusters, heterogeneous population, limit law normal.

AMS Classification: 1162G05

* Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/Madrid, 126. 28903 Getafe (Madrid)
– Received October 1997.
– Accepted April 1998.

Assume that a random sample is drawn from a population with unknown number K o clusters. Denote p_j the probability that any observation belong to the jth cluster, $j = 1, \ldots, K$, $\sum_{j=1}^{K} p_j = 1$.

A natural estimator, with bias, and its value belonging to $(0, K)$ is

$$\hat{K} = \sum_{j=1}^{K} I_j,$$

where

$$I_j = \begin{cases} 1 & \text{if the cluster } j \text{ is observed in the sample.} \\ 0 & \text{otherwise.} \end{cases}$$

The expectation of $\hat{K}$ is

$$E(\hat{K}) = K - \sum_{j=1}^{K} (1 - p_j)^n = K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

where $F(x)$ is a distribution function. This result is applied in the proof to obtain the asymptotic distribution. A lemma of Holst (1979) is very important to proof it. The variance of $\hat{K}$ is:

$$\text{var}(\hat{K}) = \sum_{j=1}^{K} (1 - p_j) + \sum_{\substack{j=1 l=1 \\ j \neq l}}^{K} \sum_{}^{K} (1 - p_j - p_l)^n - \left(\sum_{j=1}^{K} (1 - p_j)^n \right)^2,$$

which is similar obtained by McNeil (1973).

The principal goal is the limiting normality of the estimator biased $\hat{K}$, which is derived using the method of Holst (1979). The result is important because this method can be used to obtain the asymptotic distribution of a estimator.

If assume that $n \longrightarrow \infty$, then

$$K^{-1/2} \left(\hat{K} - E(\hat{K})\right) \longrightarrow N(0, \sigma_1^2),$$

where σ_1^2 is given in the proof.

The performance of the proposed estimator is investigated by means of Monte Carlo simulations. Some alternatives are showed to correcting and adjusting $\hat{K}$ for its estimated bias.