Rev.Acad.Canar.Cienc., X (Núm. 1), 27-43 (1998)

COMPARACION ENTRE LOS PRINCIPALES ESTIMADORES EQUIPROBABLES Y NO EQUIPROBABLES PARA EL NUMERO DE CLUSTERS EN UNA POBLACION.

Juan José Prieto Martínez.
Depto. de Estadística e Investigación Operativa.
Facultad de C.C. Matemáticas. Universidad Complutense de Madrid.

Abstract.

How many clusters there are in a infinite population?. Let's assume a population formed with K clusters. There are many situations in which, to know the K value is the main interest. For instance, biologist and ecologist many be interested in the assesment of the number of breeds in a plant or animal population; in nummismatic the underlying interest is the different types of strikes, develop during a time period; in the linguistic one their may be concern about the vocabulary size of a particular language known by the autor. This article advances more efficient estimators for the number of clusters tha mahe up an homogeneous or heterogeneous population, dertemining comparisons between each other out of a research done using Monte Carlo method.

Resumen.

¿Cuántos clusters hay en una población infinita?. Supóngase que una población esta constituida por K clusters. En muchas situaciones, el interés principal es conocer su valor. Por ejemplo, biólogos y ecologistas pueden estar interesados en estimar el número de especies en una población de plantas o animales; en el campo de la numismática se puede estar interesado en estimar el número de tipos de acuñación realizadas durante un periodo de tiempo; en el campo de la lingüística se puede estar interesado en conocer el tamaño del vocabulario que en una determinada lengua conoce un autor. En este artículo se presentan los estimadores más eficientes o con menos sesgo para el número de clusters que constituyen una población homogénea o una población heterogénea, realizando oportunas comparaciones entre ellos a partir de un estudio realizado por el método de Monte Carlo.

Palabras claves: Número de clusters, población homogénea, población heterogénea.

1./ Introducción.

Asúmase una población infinita 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) y Holst (1981).

Una aplicación de gran interés es el problema de estimación de especies en un ecosistema. La captura-recaptura de especies ha sido ampliamente utilizada por biólogos y ecologistas para el estudio de la dinámica de los ecosistemas. En poblaciones cerradas (ni nacen ni mueren especies durante el estudio), el problema clásico para experimentos de captura y recaptura es la estimación del número de especies. Pero también la mayoría de los trabajos ecológicos están basados en la hipótesis de que todas las especies tienen la misma probabilidad de captura. Sin embargo, como se indica en algunos de ellos, la heterogeneidad de las probabilidades de captura de especies es una realidad, enfrentándose al problema con técnicas de regresión, modelos de probabilidad o aplicando técnicas de muestreo poco justificados para aplicación en otros campos. Ver, por ejemplo, Young, Ness y Emlen (1952), Tanaka (1956, 1985), Crowcroft y Jeffer (1961), Tanton (1965), Eberhardh (1969), Marten G.G. (1970), Carothers (1973), Wilbur and Landwehr (1974).

Existe un concepto que está muy ligado con el de número 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. Esty (1982) obtuvo un estimador y un intervalo de confianza para K tambien a través de este concepto. Otros autores que utilizan este concepto para estimar K son: Betro y Zielinski (1987), Bickel y Yahav (1986), Chao (1981), Clayton y Frees (1987), Cohen y Sackrowitz (1990), Lo (1992) y Yatracos (1991).

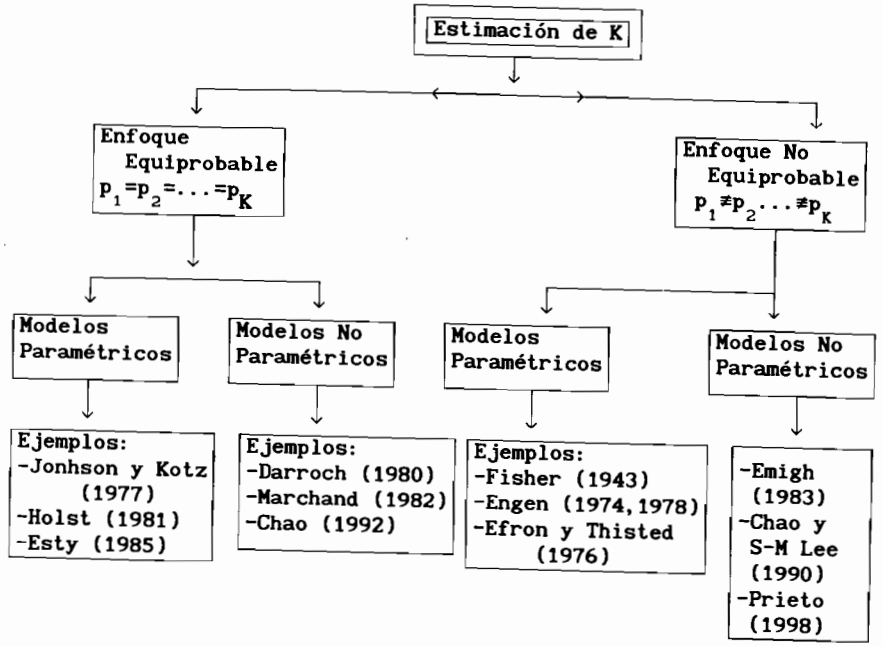
Ahora bien, considerar la hipótesis de que las probabilidades de los distintos clusters son iguales es un caso muy particular y poco frecuente, ya que poblaciones con clusters constituidos por una misma cantidad de elementos es prácticamente imposible. La mayoría de los trabajos realizados para poblaciones heterogéneas (es decir, constituidas por clusters no equiprobables) adoptan un enfoque paramétrico. Por ejemplo, en el artículo clásico de Fisher, Corbet y William (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 que sigue una distribución Gamma.

Muchos otros artículos sobre modelos de abundancia de especies en un ecosistema tambien hacen consideraciones paramétricas. Ver, por ejemplo, McNeil (1973), Engen (1978), Efron y Thisted (1976). Fué Esty (1985) el primero en estimar 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) propuso por primera vez una técnica de estimación no paramétrica, utilizando tambien la idea del cubrimiento muestral. Esta última, aunque asume desde un principio que las probabilidades de los clusters son diferentes, necesita en un momento determinado del desarrollo la hipótesis de equiprobabilidad. Esto justamente hace poner en duda si en verdad se trata de un enfoque verdaderamente no equiprobable de manera que, este trabajo engloba dicho método dentro de un enfoque totalmente equiprobable.

Tres técnicas de estimación no paramétrica alternativas a los estimadores anteriores sin necesidad de plantear un modelo de probabilidad ni de recurrir al concepto de cubrimiento muestral, son los de Emigh (1983), Chao y Shen-Ming Lee (1990) y Prieto (1998).

El objetivo principal de este artículo es comprobar la eficacia de los estimadores presentados por estos autores, así como comparalos con los de Darroch y Ratclif (1980), Esty (1985) y Chao y Shen-Ming Lee (1992) para el caso equiprobable. Un pequeño esquema que aclara la situación actual de como estimar K en una población infinita, según el enfoque de partida es el siguiente:

Esquema de estimación de K.



2./ Determinación del número de clusters en una población vía el cubrimiento muestral.

Considérese una población de elementos (personas, animales o cosas) que está subdividida en un número infinito y desconocido, K, de clusters. Se extrae una muestra aleatoria de tamaño n. Se denota por N_1, $N_2,\ldots,N_i,\ldots$, variables aleatorias que indican el número de clusters que son observados exactamente i veces, siendo n_1, $n_2,\ldots,n_i,\ldots$, sus valores observados, respectivamente. El problema consiste justamente en estimar K.

Un concepto relacionado con el problema es el cubrimiento muestral. Se define como la suma de las probabilidades de los clusters observados. Si la probabilidad de observar el cluster j es $p_j\geq 0$, y X_j es el número de observaciones del cluster j en la muestra, se tiene que el cubrimiento muestral viene dado por la expresión:

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

donde

$$I_j=\begin{cases}1 \text{ si } X_j\geq 1,\\ 0 \text{ si } X_j=0.\end{cases}$$

Considerando que las probabilidades de observación de cada cluster son iguales, es decir, $p_j=\frac{1}{K}$, $\forall$ j=1,...,K. Entonces C se puede expresar como:

$$C=\sum_{j=1}^{K}(1/K)I_j=\frac{D}{K},$$

donde D es el número de clusters observados en la muestra.

Despejando K y calculando un estimador para C, $\hat{C}$, se llega a un estimador para el número de clusters que subdividen a la población:

$$\hat{K}=\frac{D}{\hat{C}}. \qquad (1)$$

Ver Darroch y Ratcliff (1980).

Hay que subrayar, que estimar el cubrimiento bajo la hipótesis de equiprobabilidad, está relacionado con el problema clásico de ocupación. El problema consiste en distribuir aleatoriamente un número n de bolas en K celdas, y calcular la distribución de probabilidad del número de celdas vacías. Si cada celda representa un cluster, dicha probabilidad es una distribución exacta para el número de clusters no representados en la muestra, cuando todos los clusters son igualmente probables de representación en la población. Téngase en cuenta que la probabilidad que un cluster esté representado en la muestra es exactamente igual que la probabilidad original de ocupación por una bola en una celda, es decir, 1/K. Ver Feller (1972).

Por consiguiente, cabe ahora preguntarse por un estimador para C. Fue Good (1953) quien desarrollo el siguiente estimador:

$$\hat{C}=1-(n_1/n).$$

Un estimador para la varianza de $\hat{C}$ y la construcción de intervalos de confianza fue dado por Esty (1983).

Supongase ahora que los tamaños de los clusters son independientes e idénticamente distribuidos según una distribución binomial negativa con parámetros N (N>0) y r (0<r<1), tal que

$$p(\text{el tamaño del cluster } j=y)=\binom{y-1}{N-1} r^N(1-r)^{y-N}, \quad \begin{matrix} y=N,N+1,\ldots \\ j=1,\ldots,K.\end{matrix}$$

Esty (1985) obtuvo el siguiente estimador:

$$\hat{K}_{Esty}=\frac{D}{\hat{C}}+\frac{n(1-\hat{C})}{\hat{C}}\,\frac{1}{N}. \qquad (2)$$

Esty sugirió el valor de N=2 por su amplio uso en el campo de la numismática, estimando el parámetro de la distribución a partir de la información muestral.

Chao y Shen-Ming Lee (1992) asumen una muestra aleatoria de tamaño n, extraida de una población constituida por K clusters. Consideran la media de las probabilidades, $\bar{p}= \sum_i p_i/K$, y el coeficiente de variación de las probabilidades,

$$\gamma^2=[\sum_i (p_i-\bar{p})^2/K]^{1/2}/\bar{p}.$$

Nótese que si las probabilidades son iguales, γ^2 es igual a cero. El estimador obtenido fue:

$$\hat{K}_{C-S}= \frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}} \hat{\gamma}^2, \qquad (3)$$

siendo $\hat{\gamma}^2$ un estimador para γ^2 que viene expresado por:

$$\hat{\gamma}^2= \frac{\hat{K} \sum_{i=1}^{n} i(i-1)n_i}{n(n-1)} -1.$$

Obsérvese que tanto (2) como (3) contienen cuál es la magnitud del sesgo provisto por el estimador (1). Lo que sucede es que la magnitud del sesgo en (3) (segundo sumando) depende sólamente de unos parámetros muy determinados calculados a partir de la muestra: $\hat{\gamma}^2$ y $\hat{C}$; de manera que el sesgo decrece cuando el cubrimiento muestral crezca, o por el contrario, el sesgo crece cuando el coeficiente de variación sea grande. En cambio, en (2), el sesgo (segundo sumando) depende no sólamente del tamaño de la muestra y del cubrimiento muestral sino también de un parámetro de una determinada distribución de probabilidad que se ajusta a un valor, en este caso muy en concreto, por su buen comportamiento en la estimación de K en el campo de la numismática. Este valor dependerá de la población estudiada, siendo un punto abierto de investigación. Willson, Folks y Young (1984) estiman N mediante un procedimiento multipasos basado en el conocimiento de muestras de tamaño 50 y 100.

3./ Estimación del número de clusters, K, en una población heterogénea.

Supongase una muestra de tamaño n que es extraida con remplazamiento. Se denota por $p_j \geq 0$ la probabilidad de que cualquier observación pertenezca al cluster j-ésimo, con j=1,...,K, $\sum_{j=1}^{K} p_j=1$. Sea $\hat{K}_1$ el número de clusters observados en la muestra. Entonces Emigh (1983) prueba que la probabilidad de observar justamente $\hat{K}_1=c$ (con c<K) clusters es:

$$\text{Prob}(\hat{K}_1=c)= \sum_{m=1}^{c} \binom{K-m}{c-m} (-1)^{c-m} \sum_{j_1 \neq j_2 \neq \ldots \neq j_m} \left(\sum_{r=1}^{m} p_{j_r} \right)^n. \qquad (4)$$

La distribución de $\hat{K}_1$ depende de las probabilidades desconocidas de ocurrencia de los clusters p_j, j=1,...,K. El objetivo consiste en estimar dichas probabilidades para que sean utilizadas posteriormente en la maximización de la expresión (4) para un estimador para K. El algoritmo utilizado es el de la figura 1. Un estimador alternativo a $\hat{K}_E$ (estimador de Emigh) es el que se propone en Chao y Shen-Ming Lee (1990) , obtenido a partir de la desigualdad de Schwartz. Se prueba numéricamente que el error que comete es menor que el dado por $\hat{K}_E$, requeriendo además menos cálculos para su obtención. Definiendo la variable aleatoria indicadora

$$Z_j^{(i,n)}= \begin{cases} 1 \text{ si el cluster j es observado i veces en la muestra de tamaño n,} \\ 0 \text{ en otro caso,} \end{cases}$$

$$E(N_i)= \sum_{j=1}^{K} \binom{n}{i} p_j^i (1-p_j)^{n-i}. \qquad (5)$$

Recordando la desigualdad de Schwartz,

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

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

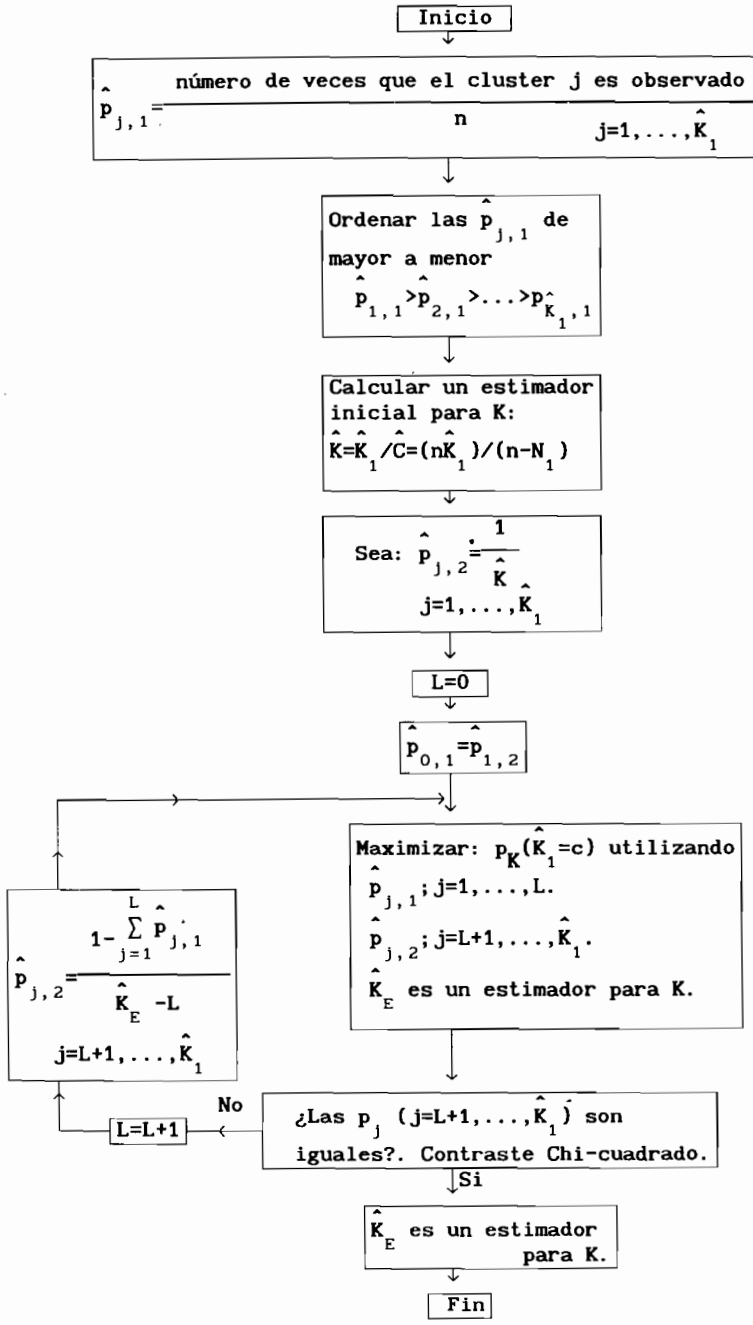
Es decir,

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

que es equivalente a:

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

Figura 1.



Utilizando (5) esta última desigualdad es equivalente a

$$E(N_0)\geq\frac{(E(N_1))^2(n-1)}{2nE(N_2)}. \qquad (6)$$

Por consiguiente, si $K=\hat{K}_1+N_0$ y se toma esperanzas,

$$E(K)=E(\hat{K}_1)+E(N_0).$$

Teniendo en cuenta la expresión (6)

$$\underline{K}_2=E(K)\geq E(\hat{K}_1)+\frac{n-1}{2n}\frac{E(N_1)^2}{E(N_2)}.$$

Reemplazando las esperanzas por los valores observados, y suponiendo que el valor observado de N_2 es mayor que 0, un estimador para K es:

$$\hat{\underline{K}}_2=\hat{K}_1+\frac{n-1}{2n}\frac{n_1^2}{n_2}. \qquad (7)$$

Se prueba también en Chao y Shen-Ming Lee (1990) que $(\hat{\underline{K}}_2-\underline{K}_2)$ converge a una distribución normal con media cero y varianza $\sigma^2(\hat{\underline{K}})$, siendo un estimador:

$$\hat{\sigma}^2(\hat{\underline{K}}_2)=\hat{K}_1+\left(\frac{n-1}{n}\right)^2\left\{\frac{n_1^3}{n_2^2}+\frac{n_1^4}{4n_2^3}+\frac{n}{n-1}\frac{2n_1^2}{n_2}\right\}-$$

$$-\left(\frac{n-1}{n}\right)^2 n_1^2\left(\frac{n-2}{n-1}-\frac{3n_1n_3}{2n_2^2}\right)^2\frac{1}{(\sum_{j=1}^{K}X_j)}, \qquad (8)$$

donde $\hat{k}_1$, n_1, n_2 y n_3 son los valores observados de $E(\hat{K}_1)$, $E(N_1)$, $E(N_2)$ y $E(N_3)$ respectivamente; y X_j el número de veces que se observa el cluster j en la muestra.

Obsérvese que (7) está formado por dos sumandos: una primera parte que trata de un estimador natural de K, $\hat{K}_1$, sesgado; y una segunda parte que trata de corregir el sesgo producido por $\hat{K}_1$. Justamente en Prieto (1998), el objetivo principal es corregir y ajustar $\hat{K}_1$ por su sesgo estimado aplicando la técnica bootstrap. El estimador viene dado por la fórmula:

$$\hat{K}_{Prieto}=\hat{K}_P=\hat{K}_1+\sum_{j=1}^{\hat{K}_1}(1-(n_j/t))^{nt}. \qquad (9)$$

donde n_j es el número de muestras (de las t de tamaño n extraidas con reemplazamiento de la población) donde el cluster j es observado. Allí se

calcula el estimador del sesgo, siendo:

$$S_e=\sum_{j=1}^{\hat{K}_1}(1-(n_j/t))^{nt}.$$

La esperanza de $\hat{K}_P$ es:

$$E(\hat{K}_P)=K-\sum_{j=1}^{K}\left\{(1-p_j)^{nt}-\sum_{j=1}^{K}\binom{t}{r}(1-(r/t))^{nt}p_j^r(1-p_j)^{t-r}\right\},$$

y su varianza es calculada utilizando el jackknife agrupado. Ver Prieto (1998).

A continuación se presentan los resultados numéricos obtenidos al estimar K mediante los métodos y procedimientos presentado en los apartados 2 y 3.

4./ Resultados numéricos.

Para comprobar como tabajan todos los estimadores presentados y obtener comparaciones y conclusiones, se han evaluado mediante métodos computacionales por simulación. El número de clusters ha sido fijado en 300 y 400. Las probabilidades de los distintos clusters pertenecen al intervalo [0,002; 0,01]. Se han considerado 7 casos posibles. En el primer caso se ha asumido las probabilidades iguales. En el segundo, los primeros 100 clusters tienen probabilidades 0,04 de ser observados y los 100 siguientes 0,006. Los siguientes casos van considerando poblaciones más heterogéneas. Ver tabla 1a) y tabla 1b). Se ha evaluado el error cuadrático medio producido en los 7 casos por los estimadores mediante la fórmula:

$$ECM=E((\mathbf{Est}-K)^2)=\frac{1}{50}\sum_{j=1}^{50}(\mathbf{Est}-K)^2;$$

así como el sesgo producido por dichos estimadores:

$$Sesgo(\mathbf{Est})=E(\mathbf{Est}-K)=\frac{1}{50}\sum_{j=1}^{50}(\mathbf{Est}-K).$$

Cada caso se ha simulado 50 veces, donde los datos reflejados en los gráficos son los promedios de los resultados. Nótese que **Est** debe ser sustituido en estas fórmulas justamente por los estimadores presentados.
Los resultados que se visualizan en los gráficos 1 y 2 corresponden a los estimadores equiprobables. Nótese que la evaluación de $\hat{K}_E$ requiere fijar el parámetro n de la distribución binomial negativa. Este ha sido fijado en 2. Se ha simulado muestras de tamaño 50 y 100. Los gráficos 1 y 2 están referidos a muestras de 100 elementos. De los resultados se obtiene:

- Obsérvese que $\hat{K}$ es cada vez más sesgado a medida que la población es más heterogénea. Obsérvese su excelente valor en el caso de una población homogénea.
- En poblaciones homogéneas tanto $\hat{K}_E$ como $\hat{K}_{C-S}$ trabajan muy bien.
- Si el tamaño muestral crece todos los estimadores son más eficientes. De ahí que los gráficos presentados estén referidos siempre a los tamaños muestrales más grandes simulados.
- A medida que el grado de la heterogeneidad es mayor, los estimadores equiprobables se alejan del verdadero valor de K.
- Los tres estimadores equiprobables bajoestiman en todos los casos el valor de K.
- Los estimadores de $\hat{K}_E$ (con $n=2$) y $\hat{K}_{C-S}$ son muy parejos, siendo éste último menos sesgado en poblaciones más heterogéneas.

Considérese ahora los estimadores para el caso no equiprobable. El algoritmo de la figura 1 para $\hat{K}_E$ se ha realizado simulando 5 ó 10 muestras de tamaño 50 y 100 elementos. En los gráfico 3 y 4 son los resultados de 10 muestras de tamaño 100, que es justamente donde $\hat{K}$ tiene menos sesgo. La evaluación de K mediante $\hat{K}$ consiste en simular una muestra aleatoria tamaño 50 y 100, y anotar los valores observados de las variables $\hat{K}_1$, N_1 y N_2 para utilizar la fórmula (11). Los gráficos 3 y 4 reflejan la solución para 100 elementos que es justamente donde existe más precisión por el estimador $\hat{K}_{-2}$. Por último la evaluación de $\hat{K}_P$, al igual que se ha hecho para $\hat{K}_E$, pasa por la simulación de muestras aleatorias (5 y 10), de tamaños 50 y 100. Los gráficos 3 y 4 reflejan el resultado para 10 muestras de tamaño 100, que es justamente donde también $\hat{K}_P$ tiene menos sesgo. Los resultados obtenidos se visualiza en los gráficos 3 y 4, indicando:

- Para una población con probabilidades de observación iguales, la estimación de K mediante $\hat{K}_{-2}$, K_E y $\hat{K}_P$ es muy buena.
- En poblaciones heterogéneas es preferible utilizar $\hat{K}_P$ que $\hat{K}_E$ o $\hat{K}_{-2}$, donde éste último siempre infra-estima K.
- El error de cada estimador aumenta a medida que la población se asume más heterogénea, aunque para $\hat{K}_P$ lo hace más débilmente. Siempre es preferible utilizar $\hat{K}_P$.
- El error estándar del estimador $\hat{K}_{-2}$ crece cuando el grado de heterogeneidad de la población también crece. Es el segundo mejor estimador en cuanto al sesgo producido.

Es lógico que si el error de estimación es grande, el error cuadrático medio sea también muy grande. Justamente esto sucede en las gráficas de las estimaciones en el caso equiprobable a medida que la población es más heterogénea. Ver figuras 5 y 6. Por el contrario, el error cuadrático medio es mucho menor para el caso no equiprobable. Nótese que el ECM perteneciente el de $\hat{K}_P$, pertenece a la gráfica donde las ordenadas toman menor valor. Ver figura 7. Observando la gráfica de los ECM para el caso K=400, ésto vuelve a suceder para $\hat{K}_P$. Ver figura 8. Al tratarse de distintos estimadores para un determinado valor (K), justamente el sesgo de los estimadores es el que nos da una valoración de cuanto fiable es cada uno. Que la población sea homogénea o heterogénea es una característica importante desde el punto de vista del método de la estimación. Se ha calculado un coeficiente, que nos mide cuánto de disperso están las probabilidades de la media de éstas. Viene expresado por la siguiente fórmula:

$$CV=\left(\frac{\sqrt{\sum_i (p_i-\bar{p})}}{\sqrt{n}}\right)/\bar{p}.$$

Es un valor comprendido entre 0 y 1. Fue utilizado en el apartado 2 para la obtención del estimador $\hat{K}_{C-S}$. Hay que notar que si CV es igual a cero, entonces los clusters son igualmente probables. A medida que CV se distancia más de 0, aproximándose a 1, la población es más heterogénea. Los gráficos presentados reflejan en el eje de abscisas los coeficientes de variación de las probabilidades de los clusters en los diferentes casos (7 casos). En el eje de ordenadas se ha escrito el sesgo producido por los estimadores. Aquí, hay que advertir que se han dividido las gráficas dependiendo: si son de tipo equiprobable o son de tipo no equiprobable, si K=300 o K=400. Los coeficientes de variación de los 7 casos vienen en la tabla 2.

Tabla 1 a). K=300

Casos	p_j
1	$\hat{p}_j=0,003$ j=1-300
2	$\hat{p}_j=0,00406$ j=1-150 $p_j=0,0026$ j=151-300
3	$p_j=0,004$ j=1-100 $p_j=0,003$ j=101-200 $p_j=0,003$ j=201-300
4	$p_j=0,004$ j=1-75 $\hat{p}_j=0,0026$ j=76-150 $p_j=0,0053$ j=151-225 $p_j=0,0013$ j=225-300
5	$p_j=0,0025$ j=1-60 $\hat{p}_j=0,00416$ j=61-120 $p_j=0,003$ j=121-180 $p_j=0,005$ j=180-240 $p_j=0,0016$ j=241-300
6	$p_j=0,0031$ j=1-38 $p_j=0,0039$ j=39-75 $p_j=0,0042$ j=76-113 $p_j=0,0027$ j=114-151 $p_j=0,005$ j=152-189 $p_j=0,0025$ j=190-227 $p_j=0,001$ j=228-265 $p_j=0,004362$ j=265-300
7	$p_j=0,0035$ j=1-23 $p_j=0,004$ j=24-47 $p_j=0,01$ j=48-70 $p_j=0,005$ j=70-93 $p_j=0,006$ j=93-116 $p_j=0,0065$ j=116-139 $p_j=0,007$ j=140-173 $p_j=0,0003$ j=174-196 $p_j=0,00003$ j=196-219 $p_j=0,000004$ j=220-243 $p_j=0,0019$ j=244-2266 $p_j=0,0008$ j=267-289 $p_j=0,01449$ j=290-300

Tabla 1 b). K=400

Casos	p_j
1	p_j=0,0025 j=1-400
2	p_j=0,003 j=1-200 p_j=0,002 j=200-400
3	p_j=0,00149 j=1-134 p_j=0,00265 j=135-266 p_j=0,00335 j=267-400
4	p_j=0,003 j=1-100 p_j=0,002 j=101-200 p_j=0,004 j=201-300 p_j=0,001 j=301-400

Casos	p_j
5	p_j=0,0025 j=1-80 p_j=0,003125 j=81-160 p_j=0,001875 j=161-240 p_j=0,00375 j=241-320 p_j=0,01125 j=321-400
6	p_j=0,006 j=1-50 p_j=0,002 j=51-100 p_j=0,004 j=101-150 p_j=0,0002 j=151-200 p_j=0,0007 j=201-250 p_j=0,0003 j=251-300 p_j=0,0036 j=301-350 p_j=0,0032 j=351-400

Casos	p_j
7	p_j=0,0023 j=1-30 p_j=0,0001 j=31-60 p_j=0,0025 j=61-90 p_j=0,003 j=90-120 p_j=0,0035 j=120-150 p_j=0,004 j=151-180 p_j=0,0045 j=181-210 p_j=0,005 j=211-240 p_j=0,00006 j=241-270 p_j=0,000043 j=271-300 p_j=0,002 j=301-330 p_j=0,00316 j=331-360 p_j=0,002372 j=360-400

Tabla 2.

K=300	1	2	3	4	5	6	7
CV	0	0,2748	0,3497	0,4583	0,5194	0,5941	0,7348
K=400	1	2	3	4	5	6	7
CV	0	0,2021	0,2964	0,3617	0,4482	0,5273	0,6971

Figura 1

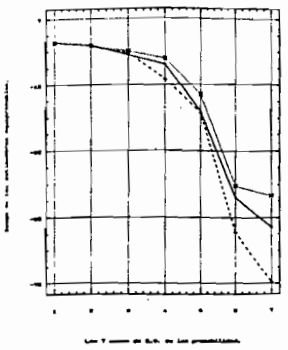


Figura 2

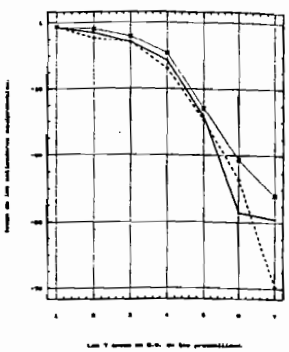


Figura 3

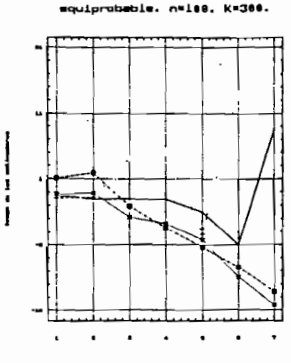


Figura 4

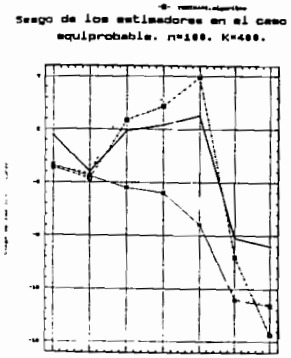


Figura 5

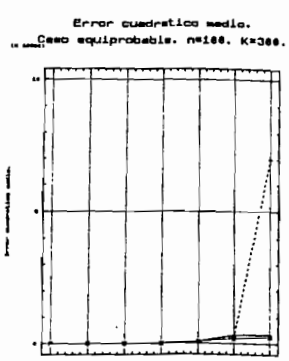


Figura 6

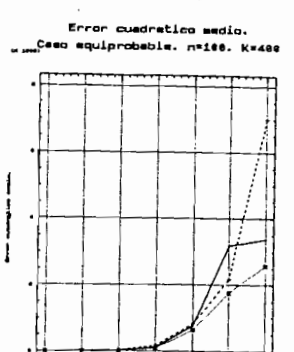


Figura 7

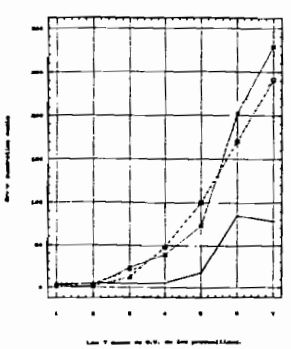
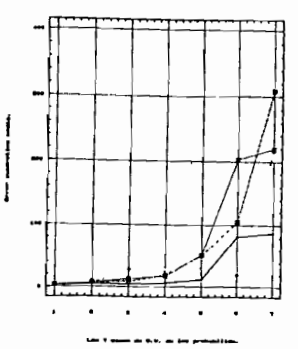


Figura 8



Bibliografía:

Betro, B y Zielinski, R. (1987). "A Monte Carlo study of a Bayesian Decision rule concerning the number of different values of a discrete random variable". Communications in Statistics, Part B- simulation and Computation, 16, 925-938.

Bickel, P.J. y Yahav, J.A. (1985). "On estimating the number of unseen species: How many executions were there?. "Technical Report 43, University of California, Berkeley, Dept. of Statistic.

Carothers, A.D. (1973). "Capture-recapture methods applied to a population with known parameters". Journal of Animal Ecology, 42, 125-146.

Cohen, A. y Sackrowitz, H.B. (1990). "Admisibility of estimators of the probability of unobserved outcomes." Annals of the Institute of Statistical Mathematics, 42, 623-636.

Chao, A. (1981). "On estimating the probability of discovering a new species". The Annals of Statistics, 9, 1339-1342.

Chao, A y Shen-Ming Lee. (1990). "Estimating the number of unseen species with frequency counts", Chinese Journal of Mathematics, 18, 335-351.

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

Clayton, M.K y Frees, E.W. (1987). "Nonparametric estimation of the probability of discovering a new species". Journal of the American Statistical Association, 82, 305-311.

Crowcroft, P. and Jeffers, J.N.R. (1961). "Variability in the behavior of wild house nice toward live traps". Proceeding of the Zoological Society, 137, 573-582.

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

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

Eberhardt, L.L. (1969). Population estimates from recapture frequencies". Journal of Wildlife Management, 33, 28-39.

Efron, B. y Thisted, R. (1976). "Estimating the number of unseen species: How many words did Shakespeare Know?. Biometrics 63, 435-447.

Emigh, T.H. (1983). "On the number of observed classes from a multinomial distribution". Biometrics, 39, 485-491.

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

Esty, W.W. (1982). "Confidence intervals for the coverage of low coverage samples". The Annals of Statistics, 10, 190-196.

Esty, W.W. (1983). "A normal limit law for a nonparametric estimator of the coverage of a random sample". The Annals of Statistics, 11, 905-912.

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.

Feller, W. (1972). "An introduction to Probability theory and its applications" 4rd ed. Wiley: New York.

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 an animal population". Journal of Animal Ecology, 12, 42-58.

Good, I.J. (1953). "On the population frequencies of species and the estimation of population parameters". Biometrika, 40, 237-264.

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

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

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

Lo, S. (1992). "From species problem to a general coverage problem via a new interpretation". The Annals of Statistics, 20, 1094-1109.

Marchand, J.P. y 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.

Prieto, J. J. (1998). "¿Cuántos clusters hay en una población?. Qüestiió, 22, 1, 69-82.

Tanaka, R. (1956). "On differential response to life traps of marked and unmarked populations". Animal of Zoology, Japan, 29, 44-51.

Tanaka, M.T. (1985). "New regression formula to estimate the whole population for recapture addicted small mammals". Research Population Ecology, 9, 83-94.

Tanton, M.T. (1965). "Problems of line-trapping and population estimation of the wood mouse". Journal of Animal Ecology, 34, 1-22.

Wilbur, M.M., and Landwehr, J.M. (1974). "The estimation of population size with equal and unequal risks of capture". Ecology, 55, 1339-1348.

Willson, L.J., Folks, J.L. y Young, J.H. (1984). "Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k". Biometrics, 40, 109-117.

Yatracos, Y.G. (1991). "On the species and related problems". Statistics & Probability Letters, 12, 209-212.

Young, H., Ness, J., and Emlen, J.T. (1952). "Heterogeneity of trap response in a population of house mice". Journal of Wildlife Management, 16, 169-180.