QÜESTIIÓ, vol. 22, 1, p. 69-82, 1998

¿CUÁNTOS CLUSTERS HAY EN UNA POBLACIÓN?

JUAN JOSÉ PRIETO MARTÍNEZ*

Sea una población cerrada formada por un número desconocido K y finito de clusters. El método bootstrap es utilizado para estimar el número de clusters que constituyen una población. Se propone un estimador para K, el cual es ajustado y corregido por su sesgo estimado mediante el método bootstrap de Efron (1979). La varianza del «estimador bootstrap» se calcula por el método jackknife agrupado. Mediante simulación, el estimador es comparado con el de Bickel y Yahav (1985).

How many clusters are there in a population?

Clasificación AMS: 162G05

Palabras clave: Número de clusters, Bootstrap, Jackknife agrupado.

* Juan José Prieto Martínez. Universidad Carlos III de Madrid. c/Madrid, 126. 28903 Getafe.

– Recibido en marzo de 1996.
– Aceptado en septiembre de 1997.

1. INTRODUCCIÓN

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

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 Ratclif (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. 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, constituidas 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). Fue 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 (1992) propone una técnica de estimación no paramétrica, pero utilizando también la idea del cubrimiento muestral. Bickel y Yahav (1985) propone un método no paramétrico para una población heterogénea sin utilizar el concepto de cubrimiento muestral.

La propuesta de este artículo es justamente plantear una técnica de estimación no paramétrica alternativa al estimador de Bickel (1985), sin necesidad de plantear un modelo de probabilidad ni de recurrir al concepto de cubrimiento muestral, demostrándose por métodos computacionales que el estimador propuesto es menos sesgado que el estimador de Bickel y Yahav (1985).

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 clusters existentes. Se propone inicialmente un estimador sesgado, el cual es corregido por su sesgo estimado mediante el método bootstrap de Efron (1979). En el apartado 2 se describe detalladamente dicho método, el cual es aplicado en el siguiente apartado para obtener el estimador ajustado. También se calcula su esperanza matemática y su varianza, esta última mediante el método jackknife agrupado. En el último apartado se refleja un estudio realizado por simulación para el estimador propuesto. Además de dar su valor, es comparado con el valor del estimador de Bickel y Yahav (1985), bajo distintas distribuciones de probabilidad. Se presentan seis casos posibles de una población que pasa de ser totalmente homogénea a ser heterogénea, comprobando la eficiencia del estimador propuesto frente a la del estimador de Bickel y Yahav. Finalmente se da el valor de su desviación típica aplicando el método jackknife generalizado.

2. MÉTODO BOOTSTRAP

Efron (1979) desarrolla el bootstrap como un método afín al jackknife, el cual requiere métodos de simulación para la estimación de un parámetro y de su varianza.

Considérese que $x_1, x_2, \ldots, x_n$ son observaciones independientes e idénticamente distribuidas de una función de distribución F desconocida. El procedimiento bootstrap sigue los siguientes pasos:

1. Construir la función de distribución de probabilidad empírica poniendo masa $1/n$ en cada una de las x_i, para $i = 1, \ldots, n$.
2. Extraer una submuestra de tamaño n con reemplazamiento de la muestra inicial, denominándose muestra bootstrap.
3. Calcular el estimador de θ basado en la muestra bootstrap.
4. Repetir los pasos N veces hasta conseguir N estimadores de θ, denotándolos por $\hat{\theta}_{(i)}$, $i = 1, \ldots, N$.
5. Calcular:
$$B_n(\theta) = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{(i)}$$
y
$$\hat{V}\left(B_n(\theta)\right) = \frac{1}{N-1}\sum_{i=1}^{N} \left(\hat{\theta}_{(i)} - B_n(\theta)\right)^2.$$

Efron sugiere que N sea un valor del intervalo $[50, 200]$ para generar estimadores adecuados de θ.

3. ESTIMACIÓN NO PARAMÉTRICA DEL NÚMERO DE CLUSTERS

3.1. El estimador

Sea una población constituida por K clusters. Se extraen t muestras de tamaño n cada una con reemplazamiento de dicha población. La probabilidad de observar el cluster j en cualquiera de las t muestras es p_j, con $j = 1,\ldots,K$. Se considerará como una sola muestra inicial la unión de las $(n \cdot t)$ observaciones procedentes de las t muestras. El único fin que tiene extraer t muestras es para saber con qué probabilidad estimada un cluster es observado. Nótese que:

$$\hat{p}_j = \frac{\text{número de muestras en que el cluster } j \text{ es observado}}{\text{número total de muestras}}$$

Sea K_1 el número de clusters diferentes observados en las t muestras. Naturalmente K_1 es un estimador sesgado. El objetivo principal es corregir y ajustar K_1 mediante su sesgo estimado. Para ello se extrae una submuestra de tamaño $(n \cdot t)$ con reemplazamiento de la muestra inicial (se denomina muestra bootstrap). Considerando la variable aleatoria indicatriz:

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

se define

$$K_2 = \sum_{j=1}^{K_1} I_j.$$

como el número de clusters observados en la muestra bootstrap.

Bajo el muestreo bootstrap, el valor esperado de K_2 viene dado por:

$$E_B(K_2) = E\left(\sum_{j=1}^{K_1} I_j\right) = \sum_{j=1}^{K_1} E_B(I_j).$$

Ahora, como las t muestras se realizan con reemplazamiento,

$$E_B(I_j) = \text{Prob}_B(I_j = 1) = 1 - \text{Prob}_B(I_j = 0) = 1 - (1 - (n_j/t))^{nt},$$

donde n_j es el número de muestras en que el cluster j es observado. Obsérvese que esta expresión es un valor muestral y no la esperanza matemática de la variable aleatoria indicatriz, la cual participa en el cálculo del sesgo producido por K_2. Por consiguiente, un estimador del sesgo producido viene dado por:

$$S_e(K_2) = \sum_{j=1}^{K_1} (1 - (n_j/t))^{nt},$$

y el estimador bootstrap para K es:

$$\boxed{\hat{K} = K_1 + \sum_{j=1}^{K_1} (1 - (n_j/t))^{nt}} \tag{1}$$

Nótese que el valor más pequeño de $\hat{K}$ es K_1 si y sólo si cada cluster de los K_1 es observado en las t muestras $(n_j = t)$. En cambio, si $n_j = 1$ (para $j = 1, \ldots, K_1$), entonces el valor máximo de $\hat{K}$ es $K_1 + K_1 (1 (1/t))^{nt} = K_1 \left(1 (t1/t)^{nt}\right)$.

3.2. La esperanza matemática de $\hat{K}$

Tomando esperanzas en (1),

$$E(\hat{K}) = E\left(K_1 + \sum_{j=1}^{K_1} (1 - (n_j/t))^{nt}\right).$$

Ahora,

$$E\left(\sum_{j=1}^{K_1} (1 - (n_j/t))^{nt}\right) = E\left(\sum_{j=1} I_j (1 - (n_j/t))^{nt}\right) =$$

$$= \sum_{j=1}^{K} \sum_{r=1}^{t} (1 - (r/t))^{nt} \binom{t}{r} p_j^r (1 - p_j)^{t-r}.$$

Por otro lado, si

$$\phi_j = \begin{cases} 1 & \text{si el cluster } j \text{ es observado en alguna de las } t \text{ muestras iniciales.} \\ 0 & \text{en otro caso.} \end{cases}$$

$$E(K_1) = E\left(\sum_{j=1}^{K} \phi_j\right) = \sum_{j=1}^{K} p(\phi_j = 1) = \sum_{j=1}^{K} (1 - p(\phi_j = 0)) =$$

$$= \sum_{j=1}^{K} \left(1 - (1 - p_j)^{nt}\right) = K - \sum_{j=1}^{K} (1 - p_j)^{nt}. \tag{2}$$

Así,

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

$$= K - \sum_{j=1}^{K} \left\{ (1 - p_j)^{nt} - \sum_{r=1}^{t} \binom{t}{r} (1 - (r/t))^{nt} p_j^r (1 - p_j)^{t-r} \right\}. \tag{3}$$

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

Cada muestra de tamaño n se divide en g grupos de tamaño h de manera que se van eliminando observaciones en bloque de cada muestra; es decir, primero se eliminan $x_1^{1)}, x_2^{1)}, \ldots, x_h^{1)}$ de la primera muestra, $x_1^{2)}, x_2^{2)}, \ldots, x_h^{2)}$ de la segunda muestra, etc; luego se eliminan $x_{h+1}^{1)}, \ldots, x_{2h}^{1)}$ de la primera muestra, $x_{h+1}^{2)}, \ldots, x_{2h}^{2)}$ de la segunda muestra, etc; en definitiva, se elimina de cada muestra uno cualquiera de los grupos y se recalcula el estimador propuesto denotado por $\hat{K}_{(i)}$. Definiendo

$$\hat{K}_{(\cdot)} = \frac{1}{g}\sum_i K_{(i)},$$

la varianza de $\hat{K}$ viene dada por

$$\mathrm{Var}(\hat{K}) = \frac{g-1}{g}\sum_i \left(K_{(i)} - K_{(\cdot)}\right)^2.$$

Este procedimiento de cálculo de la varianza se denomina método jackknife agrupado. Ver Efron (1982).

4. RESULTADOS NUMÉRICOS

Considérese una muestra aleatoria de tamaño n. Definiendo

$$I_j = \begin{cases} 1 & \text{si } X_j > 0 \\ 0 & \text{si } X_1 = 0 \end{cases},$$

donde X_j es el número de veces que el cluster j es observado, se tiene que:

$$D = \sum_{j=1}^{K} I_j$$

es el número de clusters observados; y

$$f_i = \sum_{j=1}^{K} I\left(X_j = i\right), \qquad i = 0, 1, \ldots, n,$$

es el número de clusters que son observados exactamente i veces en la muestra n.

Entonces, una cota inferior para K obtenida por Bickel y Yahav (1985) es

$$\hat{K}_{BY} = D + (f_1/n)\left(\left(\frac{\sum_{j=1}^{K} X_j}{f_1}\right)^{\frac{1}{n-1}} - 1\right)^{-1}.$$

Entonces, para comprobar la eficacia del estimador propuesto $\hat{K}$ con respecto al estimador $\hat{K}_{BY}$, se han evaluado mediante métodos computacionales por simulación. La evaluación de $\hat{K}$ ha sido llevada a cabo simulando t muestras aleatorias (en particular 1 y 5 muestras) de tamaño 50 y 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 han 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. Se han restringido los tamaños de la muestras a 50 y 100 por motivos que se comentan a continuación. Para saber si el método para el cálculo de la varianza del estimador propuesto es apropiado para conocer su error estándar, también se han realizado métodos computacionales tales que los grupos en que se ha dividido las muestras de tamaño 50 y 100 son de 5 y 10 elementos respectivamente.

Los resultados obtenidos indican que:

- Para poblaciones con probabilidades de observación iguales, la estima de $\hat{K}$ es preferible a la de $\hat{K}_{BY}$.
- En poblaciones heterogéneas es preferible utilizar $\hat{K}$ que $\hat{K}_{BY}$, donde éste siempre infraestima K, y en cambio $\hat{K}$, con $t = 1$, no siempre da un valor por defecto o por exceso. Sin embargo, cuando $\boldsymbol{t}$ y/o $\boldsymbol{n}$ crece el sesgo es por defecto. Su justificación viene a continuación.
- Para cualquier población, el sesgo estimado por $\hat{K}$ cuando se toma pocas muestras es siempre menor que cuando se toman muchas muestras. Además, fijado t, es preferible no considerar un tamaño muestral excesivamente grande. Esto es debido a que el segundo sumando de $\hat{K}$ está formado por sumas de valores comprendidos en el intervalo (0,1) con exponente $(n \times t)$. Ver (1).
- Para cualquier población, el sesgo cometido por $\hat{K}_{BY}$ cuando $n = 50$ es siempre mayor que cuando $n = 100$.
- El sesgo de cada estimador aumenta cuando la población es más heterogénea, aunque para $\hat{K}$ (con $t = 1$) lo hace más débilmente. Por otro lado, obsérvese que es preferible utilizar $\hat{K}$ (cuando t crece) que $\hat{K}_{BY}$.

– 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	$\hat{K}_{BY}$	N° de muestras	Estimador Bootstrap	$V(\hat{K})$
1	50	$p_j = 0.005$	130.15	1	211.05	29.36
		$j = 1 - 200$		5	152.60	
	100		156.69	1	219.98	34.36
				5	172.74	
2	50	$p_j = 0.004$	128.67	1	195.26	28.12
		$j = 1 - 100$		5	145.97	
		$p_j = 0.006$				
		$j = 101 - 200$				
	100		154.35	1	210.09	31.35
				5	142.35	
3	50	$p_j = 0.0035$	126.90	1	197.56	25.56
		$j = 1 - 90$		5	149.23	
		$p_j = 0.0045$				
		$j = 91 - 180$				
		$p_j = 0.014$				
		$j = 181 - 200$				
	100		143.04	1	203.81	30.54
				5	141.79	
4	50	$p_j = 0.01$	131.15	1	212.32	29.63
		$j = 1 - 10$		5	166.32	
		$p_j = 0.004$				
		$j = 11 - 100$				
	100	$p_j = 0.003$	149.33	1	224.36	33.71
		$j = 101 - 190$		5	155.97	
		$p_j = 0.023$				
		$j = 191 - 200$				

Tabla 1 (cont.)

Casos	n	p_j	$\hat{K}_{BY}$	N° de muestras	Estimador Bootstrap	$V(\hat{K})$
5	50	$p_j = 0.0035$	128.47	1	209.36	32.23
		$j = 1 - 50$		5	155.86	
		$p_j = 0.006$				
		$j = 51 - 100$				
		$p_j = 0.002$				
		$j = 101 - 125$				
	100	$p_j = 0.009$	144.73	1	189.49	39.36
		$j = 126 - 150$		5	148.21	
		$p_j = 0.005$				
		$j = 151 - 200$				
6	50	$p_j = 0.006$	149.52	1	216.52	36.23
		$j = 1 - 25$		5	174.25	
		$p_j = 0.0025$				
		$j = 26 - 50$				
		$p_j = 0.009$				
		$j = 51 - 75$				
	100	$p_j = 0.008$	162.19	1	222.89	42.97
		$j = 76 - 100$		5	172.12	
		$p_j = 0.001$				
		$j = 101 - 125$				
		$p_j = 0.002$				
		$j = 126 - 150$				
		$p_j = 0.005$				
		$j = 151 - 175$				
		$p_j = 0.004$				
		$j = 176 - 200$				

REFERENCIAS

[1] **Bickel, P.J.** y **Yahav, J.A.** (1985). «On estimating the number of unseen species: How many executions were there?»*Technical Report*, **43**, Department of Statistics, University of California, Berkeley.

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

[3] **Darroch, J.N.** (1958). «The multiple recapture census I: Estimation of a closed population». *BiometriKa*, **40**, 343-359.

[4] **Darroch, J.N.** y **Ratcliff, D.** (1980). «A note on capturerecapture estimation». *Biometrika*, **45**, 343-359.

[5] **Efron, B.** (1979). «Bootstrap methods: Another look at the jackknife». *The Annals of Statistics*, **7**, 1-26.

[6] **Efron, B.** (1982). *The jackknife, the bootstrap and the others resampling plans.* SIAM. Philadelphia.

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

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

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

[10] **Esty, W.W.** (1986). «The size of a coinage». *Numismatic Chronicle*, **146**, 185-215.

[11] **Good, I.J.** (1953). «On the population frequencies of species and the estimation of population parameters». *Biometrika*, **40**, 237-264.

[12] **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.

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

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

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

[16] **McNeil, D.** (1973). «Estimating an Author's vocabulary». *Journal of the American Statistical Association*, **68, 341**, 92-97.

[17] **Marchand, J.P.** y **Schroech, P.E.** (1982). «On the estimation of the number of equally likely classes in a population». *Communications in Statistics, Part ATheory and Methods*, **11**, 1139-1146.

ENGLISH SUMMARY

HOW MANY CLUSTERS ARE THERE IN A POPULATION?

JUAN JOSÉ PRIETO MARTÍNEZ*

Let a closed population with a unknow finite number K of clusters. The bootstrap method is used to estimate the number of cluster in a population. An estimate for K is proposed, adjusted and bias corrected by mean the bootstrap. Grouped jackknife is used to calculate its variance. The performance of the proposed estimator is investigated by means of Monte Carlo simulations and it is compared with Bickel and Yahav (1985).

AMS Classification: 162G05

Keywords: Number of clusters, Bootstrap, Grouped Jackknife.

* Juan José Prieto Martínez. Universidad Carlos III de Madrid. c/Madrid, 126. 28903 Getafe.

– Received March 1996.
– Accepted September 1997.

Assume that there is unknown number K of different clusters in a population. Suppose t samples of size n are taken with replacement. Denote p_j the probability that any observation belong to the jth cluster, $j =, \ldots, K$, $\sum_{j=1}^{K} p = 1$. If $(n \times t)$ observations is only one sample, then an estimator of p_j is:

$$\hat{p}_j = \frac{\text{number of samples where the cluster } j \text{ is observed}}{\text{number of samples}}$$

Let be K_1 the number of distinct clusters observed in the t samples. K_1 is a biased estimator. Now the goal is correct and adjust K_1 for its estimated bias.

Then a subsample of size $(n \times t)$ with replacement is taken of the t samples. This is name «the bootstrap sample». If

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

$K_2 = \sum_{j=1}^{K_1} I_j$ is the number of clusters observed in the bootstrap sample.

Under bootstrap sampling the expected value of K_2 is given by:

$$E_B(K_2) = \sum_{j=1}^{K_1} \left[1 - (1 - (n_j/t))^{nt}\right],$$

with n_j the number of samples where the cluster j is observed.

The estimator of the bias is then:

$$S_e(K_2) = \sum_{j=1}^{K_1} (1 - (n_j/t))^{nt}$$

and a (bootstrap) estimator of K is:

$$\hat{K} = K_1 + \sum_{j=1}^{K_1} (1 - (n_j/t))^{nt}.$$

The expectation of $\hat{K}$ is:

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

and its variance is calculate using the grouped jackknife.

The performance of the proposed estimator is investigated by means of Monte Carlo simulations and it is compared with Bickel and Yahav (1985) ($\hat{K}_{BY}$). The simulations results indicate that $\hat{K}$ is better than $\hat{K}_{BY}$.