NÚMERO DE ESPECIES: TEORÍA UNIFICADA DE MUESTREO PARA POBLACIONES INFINITAS

Antonio Vargas Sabadías

Departamento de Economía y Empresa. Facultad de Ciencias Jurídico-Sociales. Universidad de Castilla-La Mancha (Campus de Toledo)

Palabras y frases claves: Procesos de punto, procesos de Poisson y de Polya. Muestreo secuencial. Números de ocupación. Superposición de procesos de punto.

RESUMEN

Consideramos una población en la que se ha establecido una partición que la divide en S clases. En numerosas ocasiones, interesa estimar, más que el tamaño de las clases, el propio número S de clases. Así, geólogos y biólogos pueden tener interés por averiguar el número de especies de una población de animales o plantas; a los lingüistas les puede interesar conocer el tamaño del vocabulario de un autor. El número de componentes conectadas de un grafo, el número de errores de un programa informático, el número de fenómenos astronómicos desconocidos,...son ejemplos de posibles aplicaciones. Se han publicado numerosos trabajos para estimar el número de clases de una población mediante diferentes procedimientos independientes, pero no se ha realizado aún un estudio unificado. Este problema, conocido como problema del número de especies, es ahora tratado como una superposición de S procesos homogéneos independientes de Poisson $P_1,...,P_S$, de razones $\lambda_1 ,..., \lambda_s$, modelo que se generaliza para valores aleatorios de los λ. Cuando la población es infinita, se desarrolla una teoría unificada de muestreo, que proporciona un estimador UMVUE de S, válido para un esquema de urnas secuencial (esquema de contagio) y para los tipos de muestreo aleatorio con y sin reemplazamiento, obteniéndose asimismo una estimación del error típico de muestreo.

1. INTRODUCCIÓN

Supongamos una población, entre cuyos elementos hay una partición formada por S clases, y nos interesa precisamente averiguar el propio número S de clases, que es desconocido, así como se desconoce a priori la identidad de cada clase.

Seleccionamos una muestra de tamaño T de la población, admitiendo que, una vez es seleccionado un elemento, la clase correspondiente puede ser identificada. El resultado del experimento estará re-

presentado por el vector aleatorio $\mathbf{N}=(N_1,N_2,...,N_D)$, donde N_j representa **"el número de elementos de la clase j-ésima que forman parte de la muestra"**, siendo D **"el número de clases o especies diferentes que hay en la muestra"**

La dificultad radica en que la clase j-ésima puede no aparecer en la muestra, en otras palabras: **N** no es observable, de modo que la clase j-ésima forma parte de la muestra si $N_j>0$.

En lugar de trabajar con **N**, se trabaja con otro vector $\mathbf{M}=(M_1, M_2,...,M_D)$, que sí es observable, y, donde M_r representa **"el número de especies que aparecen r veces en la muestra"**. Los M_r son los **"números de ocupación"** o **"frecuencias de frecuencias"**, como los llamó Good.

El problema fundamental es el de estimar S a partir de los M_r.

La diversidad D y el tamaño muestral T se pueden expresar en función de los M_r:

$$D = \sum_{r=1}^{T} M_r \quad y \quad T = \sum_{r=1}^{D} r\, M_r \tag{1.1}$$

Para precisar ideas, admitimos los siguientes axiomas:

Axioma I: La población consta de S especies (S es desconocido) con abundancias relativas

$$\vec{p}\,' = (p_1,p_2,...,p_S) \quad , \text{donde} \quad \sum_{j=1}^{S} p_j = 1,\ 0<p_j<1. \tag{1.2}$$

La abundancia relativa p_j es la probabilidad con que la clase j-ésima está en la población.

Axioma II: Se toma al azar una muestra de la población, en la que N_j representa **"el número de miembros de la j-ésima especie que son seleccionados en la muestra"**. Se supone que N_j sigue una distribución de Poisson con parámetro $\lambda_j=\kappa p_j$. Así

$$P(N_j = r/\lambda_j) = \frac{e^{-\lambda_j}\lambda_j^r}{r!} \tag{1.3}$$

La elección del axioma II supone admitir que

$$E(N_j) = \lambda_j = \kappa p_j \tag{1.4}$$

es decir, **"el número esperado de individuos de la especie j-ésima es proporcional a la abundancia relativa de esta especie en la población"**. La constante de proporcionalidad representa el número total de individuos capturados de entre todas las especies que han sido seleccionadas. Una vez admitido el axioma II, los distintos submodelos van a depender de la distribución de los λ_j .

Para representar el "**muestreo de las especies**" como una superposición de procesos de Poisson en el intervalo de tiempo (0,t], con t>0 (Bunge, 1993), sea N_k (t) "el número de individuos de la especie I_k que han sido seleccionadas en el período (0,t]". Entonces se verifica la siguiente proposición:

Proposición 1: La variable aleatoria N_k (t) define un proceso de Poisson homogéneo de media $\lambda_k t$.

Se puede extender el muestreo de Poisson haciendo que los λ_k sean, a su vez, variables aleatorias con una cierta distribución, como puede ser la distribución gamma. La elección de una distribución gamma para los λ_k nos va a conducir al modelo de Polya.

Axioma III: Los $\{\lambda_j\}_{j=1,2,...,S}$ son independientes y están idénticamente distribuidos con una distribución común gamma de parámetros (A,1/A), es decir:

$$f(\lambda) = \frac{A^A}{\Gamma(A)} \lambda^{A-1} e^{-\lambda A} \quad ,\lambda>0,\ 0<A<\infty,\ t>0. \tag{1.5}$$

De este modo, λ_j es tal que $E[\lambda_j]=1$ y $Var[\lambda_j]=1/A$.

La elección del axioma III nos lleva al siguiente resultado conocido:

Proposición 2: Si $N_j(t)$ es una variable que sigue una distribución de Poisson de media $\lambda_j t$, y suponemos que los λ_j siguen una distribución gamma $\Gamma(A,1/A)$, entonces la distribución compuesta es la binomial negativa de parámetros $BN\ [A,A/(t+A)]$, con

$$P_r = P\,[\,N_j(t) = r / \lambda\,] = \binom{A+r-1}{r} \left(\frac{t}{t+A}\right)^r \left(\frac{A}{t+A}\right)^A \tag{1.6}$$

Al ser N_k (t) el número de individuos de la clase I_k, N_k (t)+A representa el número de pruebas necesarias para obtener, por primera vez, A individuos de una clase distinta de I_k, lo que sucederá si, y sólo, si en la última prueba se obtiene un individuo que no pertenece a la clase I_k, y en las N_k (t)+A-1 pruebas anteriores habían aparecido N_k (t) individuos de la especie I_k.

Como se desconoce el número de especies que no aparecen en la muestra, debemos tomar la distribución truncada en cero, que es la distribución de $N_k(t)$ condicionada por $N_k(t)>0$, puesto que la especie I_k estará representada en la muestra si N_k (t)>0:

$$P_r^* = P\,[N_k(t) = r / \lambda,\ N_k(t)>0] = \binom{A+r-1}{r} \left(\frac{t}{t+A}\right)^r \left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{t+A}\right)^A} \tag{1.7}$$

Se verifican las siguientes propiedades para P^*_r:

A) $$\sum_{r=1}^{D} P_r^* = 1 \tag{1.8}$$

B) $$\sum_{r=1}^{D} P_r = (1-P_0)\sum_{r=1}^{D} P_r^* = 1-P_0 \tag{1.9}$$

Si no admitimos el axioma III, estamos ante un proceso puro de nacimiento, con tasa de crecimiento constante, mientras que admitir el axioma III nos lleva también a un proceso de nacimiento puro, pero con tasa de crecimiento variable (Feller, 1993).

La función generatriz de probabilidades de la distribución truncada es:

$$\Psi_{N_k^*(t)}(z) = \frac{1}{1-P_0}\left(\frac{p}{1-qz}\right)^A = \frac{1}{1-P_0}\Phi_{N_k(t)}(z)\ ,\ 0<p<1 \tag{1.10}$$

Los momentos respecto al origen de la distribución truncada correspondiente al proceso k-ésimo, se obtienen, por tanto, de multiplicar los momentos respecto al origen de la distribución no truncada

por $$\frac{1}{1-P_0} = \frac{1}{1-\left(\frac{A}{A+t}\right)^A} \tag{1.11}$$

Se obtiene así:

C) $$E\left[N_k^*(t)\right] = \frac{tA}{A}\frac{1}{1-P_0} = \frac{t}{1-P_0} \tag{1.12}$$

D) $$Var\left[N_k^*(t)\right] = \frac{t}{1-P_0}\left(1+t+\frac{t}{A}-\frac{t}{1-P_0}\right) \tag{1.13}$$

Una forma de estimar los parámetros de la población consiste en utilizar la media muestral como estimador de la media de la población. La media muestral es:

$$\bar{X} = \frac{1}{D}\sum_{k=1}^{D} kM_k = \frac{\hat{T}}{D} \tag{1.14}$$

que es la media de la distribución truncada; luego:

$$\frac{\hat{T}}{D} = \frac{t}{1-P_0} = \frac{St}{D} \tag{1.15}$$

El problema del **"número de especies"** no es sino un caso particular del clásico problema de **"esquema de urnas"** o de **"ocupación aleatoria de S celdas por n bolas"**. El esquema de urnas es un modo de trabajo conceptual con distribuciones estadísticas. Se considera una

población de n individuos (bolas en una urna), que son idénticas salvo en el color.

En una prueba simple, se selecciona una bola de la urna y se anota su color; la bola se devuelve entonces a la urna. Así se realizan más pruebas bajo condiciones idénticas a la primera. Si cada una de las bolas tiene la misma probabilidad de ser extraída en cada prueba, el experimento corresponde al muestreo aleatorio con reemplazamiento.

Si modificamos las reglas del esquema de urnas anterior, de forma que, cuando se selecciona una bola de un determinado color, se devuelven c+1 bolas del mismo tipo a la urna, tenemos el **"esquema de urnas de Polya"**, que, según los distintos valores de c, da lugar a los diferentes tipos de muestreo (Stuart, 1987):

a) *el esquema de contagio*: corresponde al caso c=1 y, cuando la población es infinita, está regido por la distribución binomial negativa;

b) el *muestreo aleatorio con reemplazamiento*: corresponde a c=0 y, cuando la población es infinita, está regido por la distribución de Poisson;

c) *el muestreo aleatorio sin reemplazamiento*: corresponde a c=-1, que, si la población es infinita, está regido por la binomial.

En vez de seleccionar una muestra de tamaño fijo, se puede alterar la regla de parada y elegir continuar con el muestreo hasta que se consiga obtener por primera vez el A-ésimo éxito. Este es el método de muestreo *secuencial:* se trata de contar el número de fracasos hasta que se obtiene el A-ésimo éxito.

El tamaño muestral T es variable, y nos interesa conocer la distribución del vector **M**.

El problema es equivalente a distribuir al azar una de las n bolas en una de las S celdas etiquetadas con los números 1,2,...,S, de modo que la probabilidad de que la bola caiga en la celda i-ésima es p_i.

2. DISTRIBUCIÓN DE LOS NÚMEROS DE OCUPACIÓN

Si definimos la función indicador $I_k(t)$, que toma el valor 1 si la celda k está ocupada, y cero si no lo está, se definen los **"números de ocupación"**, M_r, de la siguiente forma:

$$M_r = \sum_{k=1}^{S} I\,[N_k(t) = r] \qquad (2.1)$$

Los números de ocupación (M_r) son variables aleatorias definidas a partir de la función indicador, que tienen una distribución en el muestreo, que nos interesa conocer. Con este fin, vamos a estudiar su función generadora de probabilidades.

Al ser las M_r variables intercambiables, su función generadora de probabilidades será:

$$\phi_{M_r}(u) = E\left[u^{M_r \mid \sum_{k=1} M_r = D}\right] = E\left[u^{\sum_{k=1} I[N_k(t)=r]}\right] = \prod_{k=1}^{D}\left[(u-1)\frac{P_r}{1-P_0}+1\right] = \left[(u-1)\frac{P_r}{1-P_0}+1\right]^D \qquad (2.2)$$

que se trata de una distribución binomial de parámetros D y P_r*.

Como la función generadora de probabilidades determina, de manera única, la distribución, acabamos de demostrar la siguiente proposición:

Proposición 3: Los números de ocupación, M_r, son variables aleatorias independientes con una distribución binomial de parámetros $B(D,P_r/(1-P_0)$.

Como consecuencia de esta proposicón, se verifica el siguiente corolario:

Corolario 3.1: La distribución del vector aleatorio $M=(M_1,M_2,...,M_D)$ condicionado por $M_1+M_2+...+M_D=D$, es multinomial de parámetros:

$$M\ (D,P_1/(1-P_0),...,P_D\ /(1-P_0))=M(D,P^*_1,...,P^*_D)$$

Si estimamos P_r* mediante, $E[M_r]/D$, obtenemos una aproximación de la función generadora de probabilidades de M_r, que sigue una distribución binomial de parámetros $B(D,M_r/S)$.

Se tienen, por tanto, las siguientes propiedades, donde las esperanzas son condicionadas:

1. $$E[M_r] = \frac{DP_r}{1-P_0} = DP_r^* \qquad (2.3)$$

Esta propiedad también es cierta para r=0, es decir:

2. $$\hat{P}_0 = \frac{S-D}{S} \qquad (2.4)$$

Podemos, entonces, enunciar la siguiente proposición:

Proposición 4: $\frac{E[M_r]}{D}$ es un estimador insesgado de P_r*.

Designamos por $\hat{M}_r$ a $E[M_r]$, ya que se trata de una variable aleatoria.

3. En particular $$\hat{M}_1 = \frac{At}{A+t}(S-D) \qquad (2.5)$$

3. ESTIMADOR DE MÁXIMA VEROSIMILITUD

Proposición 5: $D/(1-\hat{P}_0)$ es un estimador insesgado de S, bajo la distribución truncada en cero.

En efecto: $$E(D)=S(1-P_0)=D. \tag{3.1}$$

Luego $\dfrac{D}{1-\hat{P}_0}$ es un estimador insesgado de S, cuya estimación depende de A.

Teniendo en cuenta que la distribución conjunta del vectror $(M_1,....,M_D)$ condicionada por $M_1+...+M_D=D$ es multinominal de parámetros $M(D,P_1/(1-P_0),...P_D/(1-P_0))$, la función de verosimilitud es:

$$L = P\left(M_1,M_2,...,M_D \mid \sum_{k=1}^{D} M_k = D \right) = \prod_{r=1}^{D} \frac{P_r^{M_r}}{1-P_0} \tag{3.2}$$

Tomando logaritmos en los dos miembros de (3.2) se obtiene:

$$\ln L = \sum_{r=1}^{D} M_r \ln P_r - D\ln(1-P_0) \tag{3.3}$$

Si estimamos t por T/S en las expresiones de P_r y P_0, y derivamos con respecto a S, queda:

$$\frac{\partial \ln L}{\partial S} = \sum_{r=1}^{D} M_r \frac{P_r'}{P_r} + D\frac{P_0'}{1-P_0} \tag{3.4}$$

donde hemos llamado $P_r' = \dfrac{d}{dS}P_r$ y $P_0' = \dfrac{d}{dS}P_0$.

Desarrollando (3.4), se obtiene:

$$\frac{\partial \ln L}{\partial S} = \frac{AqP_0}{S^2}\left(\frac{D}{1-P_0} - S \right) \tag{3.5}$$

Igualando a cero la derivada, resulta finalmente:

$$\hat{S} = \frac{D}{1-P_0}. \tag{3.6}$$

Tenemos, por tanto, la siguiente proposición:

Proposición 6: $D/(1-P_0)$ es un estimador de máxima verosimilitud de S.

La derivada del logaritmo de la función de verosimilitud para la distribución truncada viene dada por la expresión (3.5), donde el primer factor es independiente de las observaciones, lo que nos permite afirmar que la cota de Cramer-Rao es accesible (Kendall, 1987).

Se trata, por tanto, de un estimador uniformemente de mínima varianza (UMVUE). Como es insesgado, la varianza coincide con el inverso del factor que multiplica a [D/(1-P_0)-S] en (3.5), obteniéndose:

$$\hat{Var}(\hat{S}) = \frac{\hat{S}^2}{AqP_0} \tag{3.7}$$

Hemos demostrado la siguiente proposición:

Proposición 7: Una estimación del error típico de muestreo del estimador de máxima verosimilitud de S es

$$E\hat{T}M(\hat{S}) = \frac{\hat{S}}{\sqrt{\hat{M}_1}} \tag{3.8}$$

En efecto, tomando $\hat{M}_1$ como estimador de AqP_0 en (3.7) y extrayendo la raíz cuadrada, resulta (3.8).

4. ESTIMADORES DE S EN FUNCIÓN DEL PARÁMETRO

Antes de buscar estimadores de los parámetros que intervienen en esta distribución, nos interesa establecer algunas relaciones entre ellos. Así son inmediatas las siguientes:

1. $$t = \frac{A\hat{M}_1}{A(S-D)-\hat{M}_1} \tag{4.1}$$

2. $$S-D = \frac{\hat{M}_1}{A} + \frac{\hat{M}_1}{t} \tag{4.2}$$

3. $$\frac{\hat{M}_1}{\hat{T}} = pP_0 \tag{4.3}$$

Como consecuencia de la relación (4.3) anterior, podemos enunciar la siguiente proposición:

Proposición 8: En el muestreo por esquema de contagio, el estimador de Good-Turing (Good, 1956), $\frac{\hat{M}_1}{\hat{T}}$,es menor que P_0.

En efecto: Como 0<p<1, $p P_0 = \hat{M}_1/\hat{T}$ implica que $\hat{M}_1/\hat{T}$ es menor que P_0.

4. $$P_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}} \tag{4.4}$$

Para demostrar esta propiedad, basta con tomar logaritmos en ambos miembros de la relación (4.3) y tener en cuenta que $P_0=p^A$.

5. $$P_0 = \frac{\hat{M}_1}{\hat{T}} + \frac{\hat{M}_1}{SA} \tag{4.5}$$

En efecto: $\hat{M}_1 = \frac{DP_1}{1-P_0} = SP_1 = \frac{SAt}{A+t}P_0 \Rightarrow P_0 = \frac{[A+t]\hat{M}_1}{SAt} = \frac{A\hat{M}_1}{SAt} + \frac{\hat{M}_1}{SA} = \frac{\hat{M}_1}{\hat{T}} + \frac{\hat{M}_1}{SA}$

Si tenemos en cuenta que $S=D/(1-P_0)$ y las relaciones (4.4) y (4.5), se obtienen las expresiones (4.6) y (4.7):

I. $$\hat{S} = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}} \tag{4.6}$$

II. $$\hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\frac{1}{A} \tag{4.7}$$

El problema de estimar S pasa por estimar antes A.

5. COEFICIENTE DE VARIACIÓN DE PEARSON DE LAS p_k

Las abundancias relativas son, en este modelo, variables aleatorias, cuya distribución conviene determinar, puesto que del grado de heterogeneidad de su distribución va a depender el estimador de S, en concreto de su coeficiente de variación.

Proposición 9: La variable aleatoria $\sum_{j\neq i}^{S} \lambda_j$ sigue una distribución gamma $\Gamma\left((S-1)A, \frac{1}{A}\right)$

En efecto, se trata de la suma de S-1 variables aleatorias independientes, todas ellas con distribución gamma $\Gamma(A,1/A)$ por tanto también sigue una distribución gamma de parámetros:

((S-1)A),1/A).

Proposición 10: La variable aleatoria $p_i = \dfrac{\lambda_i}{\lambda_i + \sum_{j \neq i}^{S} \lambda_j} = \dfrac{\lambda_i}{\sum_{j=1}^{S} \lambda_j}$ sigue una distribución beta de parámetros B[A,(S-1)A)].

Las proposiciones 9 y 10 se deducen inmediatamente de las propiedades de las funciones gamma y beta (Rao, 1965).

Como las p_i siguen todas la misma distribución beta de parámetros B[A,(S-1)A], la varianza de p_i es:

$$\sigma^2_{p_i} = \frac{S-1}{S^2[SA+1]} \tag{5.1}$$

y, al ser $E[p_i]=1/S$, y el cuadrado del coeficiente de variación de Pearson de las p_i es:

$$\gamma^2 = \frac{S-1}{SA+1} \tag{5.2}$$

Si S es suficientemente grande y A finito, se verifica la siguiente proposición:

Proposición 11: Cuando S es suficientemente grande, el cuadrado del coeficiente de variación de Pearson de las p_k es aproximadamente igual al inverso del parámetro de la distribución:

$$\gamma^2 \approx \frac{1}{A} \tag{5.3}$$

En efecto: $\gamma^2 = \dfrac{S-1}{SA+1} = \dfrac{S}{SA+1} - \dfrac{1}{SA+1} \approx \dfrac{1}{A}$

puesto que, cuando S es suficientemente grande,

$$\frac{1}{SA+1} \to 0 \; y \; \frac{S}{SA+1} \to \frac{1}{A}$$

Podemos observar cómo la heterogeneidad de la distribución de las p_k es inversamente proporcional al parámetro. Cuanto mayor es A, más homogénea es la distribución, dándose el mayor grado de homogeneidad en el caso en que A tiende a infinito.

La función generatriz de momentos nos permite obtener la siguiente relación:

$$E\left[\sum_{k=2}^{D} k(k-1)M_k\right] = \frac{St^2(A+1)}{A} \tag{5.4}$$

Si designamos por $\hat{R}_2$ al primer miembro de (5.4), se obtiene la siguiente relación entre A, t y S:

$$St^2(A+1) = \hat{R}_2 A \tag{5.5}$$

En función del coeficiente de variación de Pearson de las p_k, si tenemos en cuenta que $\frac{A}{A+1} = \frac{1}{\gamma^2+1}$ resulta:

$$P_0 = \left(\frac{\hat{M}_1}{T}\right)^{\frac{1}{\gamma^2+1}} \tag{5.6}$$

Esta última relación muestra cómo la probabilidad de especies desconocidas depende del coeficiente de variación de Pearson de las p_k, de tal forma que, cuando el coeficiente de variación sea nulo, estaremos bajo la hipótesis de homogeneidad, que supone la equiprobabilidad de las p_k; en cambio, si la distribución es heterogénea, la hipótesis de equiprobabilidad no es admisible.

Estos resultados nos permiten expresar S en función del coeficiente de variación de Pearson de las p_k. Así, a partir de (4.6) y (4.7), se obtienen las expresiones:

$$S = \frac{D}{1-\left(\frac{\hat{M}_1}{T}\right)^{\frac{1}{\gamma^2+1}}} \tag{5.7}$$

$$\hat{S} = \frac{TD}{T-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\gamma^2 \tag{5.8}$$

Este último es el estimador de Chao (Chao, 1992). Se trata, por tanto, ahora de estimar el coeficiente de variación de Pearson de las p_k. La dificultad está en que el propio coeficiente de variación de Pearson, cuyo cuadrado es el inverso del coeficiente A, depende a su vez de S.

6. ESTIMADOR DEL COEFICIENTE DE VARIACIÓN (1/A)

Para conseguir un estimador de 1/A (cuadrado del coeficiente de variación de Pearson de las p_k), que sea independiente de S, vamos a utilizar el cuadrado del coeficiente de variación de Pearson muestral:

$$\hat{\gamma}^2 = \frac{D(\hat{R}_2+\hat{T})-\hat{T}^2}{\hat{T}^2} \tag{6.1}$$

cuyo error típico de muestreo es suficientemente pequeño ($Var(\gamma) \leq \hat{\gamma}^2/D$) (Stuart, 1987).

Como estimador de $\hat{\gamma}^2+1$, resulta:

$$\hat{\gamma}^2+1=\frac{D(\hat{R}_2+\hat{T})}{\hat{T}^2} \tag{6.2}$$

Substituyendo la estimación de 1/A en las expresiones (4.6) y (4.7), se obtienen los estimadores de S:

$$\hat{S}_1=\frac{\hat{T}D}{\hat{T}-\hat{M}_1}+\frac{\hat{M}_1}{\hat{T}-\hat{M}_1}\frac{D(\hat{R}_2+\hat{T})-\hat{T}^2}{\hat{T}} \tag{6.3}$$

y

$$\hat{S}_2=\frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)\frac{\hat{T}^2}{D(\hat{R}_2+\hat{T})}} \tag{6.4}$$

Una estimación del error típico de muestreo de estos estimadores viene dada por (3.8).

7. MUESTREO SIN REEMPLAZAMIENTO

Si el número medio $N_k(t)$ de individuos de la clase I_k que hay en el intervalo (0,1] es V, entones, la variable aleatoria $J_k(t)$, que proporciona el número de estos individuos que hay en el subintervalo (0,q] sigue una distribución binomial de parámetros B(A,q). $J_k(t)$ puede definirse como el proceso de punto que proporciona el número de individuos de la especie I_k que hay en el subintervalo (0,q].

Hemos supuesto que los sucesos que tienen lugar en los sucesivos puntos de tiempo son independientes, por lo que los $J_k(t)$ son también independientes entre sí. Cada uno de los $J_k(t)$ tiene una distribución binomial de parámetros B(A,q), que corresponde a un proceso puro de muerte con tasa de muerte lineal.

Como se desconoce el número de especies que no forman parte de la muestra, tomamos también la distribución truncada en cero. Tenemos así el proceso de Bernoulli $J_k^*(t)$, que verifica:

$$E[J_k^*(t)]=\frac{At}{A+t}\frac{1}{1-P_0}=\frac{AqS}{D}\quad\text{, de donde}$$

$$E\left[\sum_{k=1}^{D}J_k^*(t)\right]=AqS \tag{7.1}$$

Podemos enunciar la siguiente proposición:

Proposición 12: $J_k^*(t)$ es un proceso de Bernoulli, cuya distribución es binomial de parámetros B[A,q/(1-P_0)], que proporciona la probabilidad del número medio de individuos pertenecientes a la clase I_k que hay en el intervalo (0,q].

Resumiendo: el proceso de Polya, $N_k(t)$ proporciona el número de individuos de la clase I_k que hay en el intervalo (0,t] cuando se obtiene por primera vez A individuos pertenecientes a otra clase, habiéndose obtenido antes $N_k(t)+A-1$ individuos de la clase I_k. Nos permite estudiar el problema del número de especies cuando el modelo de muestreo corresponde al esquema de urnas de contagio (c=1).

El proceso de Bernoulli, $J_k(t)$, en cambio, proporciona el número de individuos de la especie I_k que hay en el subintervalo (0,q], y va a permitirnos estudiar el modelo de las especies cuando el muestreo es sin reemplazamiento (c=-1). La distribución que regula este esquema de muestreo es la binomial de parámetros B(A,q), y corresponde a un proceso de muerte con tasa de muerte lineal (Feller, 1993).

Podemos distinguir este modelo si observamos que, en él, el estimador de Good-Turing, $\hat{M}_1/\hat{T}$, es mayor que P_0, ya que

$$P_0 = \frac{\hat{M}_1}{T}q \tag{7.2}$$

Los estimadores no homogéneos de P_0 son ahora:

$$P_0 = \frac{\hat{M}_1}{\hat{T}} - \frac{\hat{M}_1}{SA} \quad \text{y} \quad P_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A-1}} \tag{7.3}$$

Estimando el coeficiente de variación, 1/A, mediante la expresión (6.1), se obtienen ahora para S:

$$\hat{S}_3 = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} - \frac{\hat{M}_1}{\hat{T}-\hat{M}_1}\,\frac{D(\hat{R}_2+\hat{T})-\hat{T}^2}{\hat{T}} \tag{7.4}$$

y

$$\hat{S}_4 = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{\hat{T}^2}{2\hat{T}^2-D(\hat{R}_2+\hat{T})}}} \tag{7.5}$$

8. MUESTREO CON REEMPLAZAMIENTO

En el muestreo con reemplazamiento, el parámetro A tiende a infinito, con lo que el estimador de P_0 coincide con el estimador de Darroch (basado en el de Good-Touring) (Darroch, 1980):

$$S_5 = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} \tag{8.1}$$

9. ANÁLISIS DE RESULTADOS

El análisis de estos modelos nos permite dar un criterio unificado al

problema de las especies cuando la población es infinita. Estas conclusiones confirman resultados ya conocidos (Jhonson, 1976); cada una de las tres distribuciones (binomial negativa, binomial y de Poisson) pueden ser consideradas como desarrollo de:

$$[(1+w)-w]^{-B}$$

siendo, en el caso de la binomial negativa, B>0 y w>0; para la binomial, -1<w<0 y B<0; la distribución de Poisson corresponde a un caso de límite intermedio, donde w tiende a cero y B tiende a infinito, con $Bw=\lambda$.

Por ello, según el modelo de que se trate, tendremos:

1) Si el muestreo es secuencial:

$$B>0 \text{ y } 0<w<1$$

Si hacemos A=B, Q=1+w y P=w, el término (r+1)-ésimo del desarrollo del binomio $(1+w-w)^{-B}$, que, en este caso, es $(Q-P)^{-A}$, resulta

$$\binom{A+r-1}{A-1}\left(\frac{P}{Q}\right)^{r}\left(1-\frac{P}{Q}\right)^{A}$$

que es la función de cuantía de la binomial negativa, de modo que, si hacemos ahora q=P/Q, resulta la binomial negativa en la expresión que hemos venido utilizando:

$$\binom{A+r-1}{A-1} q^{r} p^{A}$$

Estimando 1/A mediante (6.1), se obtienen, como estimadores de S, las expresiones (6.3) y (6.4).

2) Si el muestreo es completamente aleatorio sin reemplazamiento:

$$B<0 \text{ y } -1<w<0$$

Si hacemos A=-B>0, -w=q, con lo que p=1+w.

Entonces el término (r+1)-ésimo del desarrollo de $(1+w-w)^{-B}=(q+p)^{A}$ es

$$\binom{A}{r} q^{r} p^{A-r}$$

que es la función de cuantía de una distribución binomial de parámetros B(A,q).

Si estimamos 1/A a partir de (6.1), resulta:

$$\frac{1}{\hat{A}} = \frac{D(\hat{M}_2+\hat{T})}{\hat{T}^2}$$

y podemos utilizar, como estimadores de S, las expresiones de (7.4) y (7.5).

3) Si el muestreo es completamente aleatorio con reemplazamiento, A tiende a infinito, y, como estimador de S, resulta: (8.1). Este modelo de mestreo corresponde al de Maxwell-Boltzman.

Los distintos valores del parámetro A nos permiten analizar también algunos otros casos particulares:

a) A=1, que corresponde a la distribución de Bose-Einstein, en cuyo caso

$$P_0 = \left(\frac{\hat{M}_1}{\hat{T}} \right)^{\frac{1}{2}} \tag{9.1}$$

Se trata de una situación particular de muestreo secuencial, con un nivel de heterogeneidad del 100%.

b) A=2, que corresponde al modelo de Esty (Esty, 1986a), en cuyo caso

$$P_0 = \left(\frac{\hat{M}_1}{\hat{T}} \right)^{\frac{2}{3}} \tag{9.2}$$

También corresponde a una situación no homogénea con un nivel de heterogeneidad del 50%.

c) Si A=0, se estaría en una situación con grado máximo de heterogeneidad. La distribución adecuada es la de Ewens, y, como estimador de S, se puede utilizar:

$$S_6 = \frac{\hat{T}^2 \hat{M}_1}{(\hat{T}-\hat{M}_1)^2} \ln \frac{\hat{T}}{\hat{M}_1} \tag{9.3}$$

10. EJEMPLO

Vamos a aplicar estos resultados al ejemplo que plantea Fisher (Fisher, 1943), en que se pretende averiguar el número de especies de mariposas en Malaya a partir de los siguientes datos:

Nº de especies	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	240
Nº de ocupación	118	74	44	24	29	22	20	19	20	15	12	14	6	12	6	9	9	6	10	10	11	5	3	3

Tenemos: M_1=118, M_2=74, D=501, T=3306 y R_2=35350. De aquí se obtiene, como estimación de A:

A=1'2954

Vemos, en primer lugar, que no es admisible la hipótesis de homogeneidad, ya que el cuadrado del coeficiente de variación de Pearson

de las abundancias relativas se puede estimar por:

$$\hat{\gamma}^2 = \frac{1}{\hat{A}} = 0'7719 \Rightarrow \hat{\gamma} = 0'8786$$

El coeficiente de variación de Pearson, en porcentaje, es del 87'86%, lo que indica la heterogeneidad de la distribución.

Como estimación de S, obtenemos:

$$\hat{S} = 614$$

Si utilizamos la expresión (3.8),el error típico del estimador en el muestreo es aproximadamente igual a:

$$\hat{\sigma}_{\hat{S}} = \frac{614}{\sqrt{118}} = 56'52$$

REFERENCIAS BIBLIOGRÁFICAS

BUNGE, J. y FITZPATRICK, M.(1993). "Estimating the Number of Species: A Review", Journal of the American Statistical Association, Vol. 88, Nº421.

CHAO, A.(1981). "On Estimating the Probability of Discovering a New Species", The Annals of Statistics, 9, 1339-1342.

————(1984) "Nonparametric Estimation of the Number of Classes in a Population". Scandinavian Journal of Statistics. Theory and Applications, 11, 265-270.

————(1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability". Biometrica, 43, 783791.

CHAO, A. and LEE, S.M.(1992). "Estimating the Number of Classes via Sample Coverage", Journal of the American Statistical Association, Vol. 87, Nº 417.

CHAO, M.T. (1992) "From Animal Trapping to Type Token". Statistica Sinica, 2,189-201.

COX, D.R. and ISHAM, V.(1992). "Point Processes", Chapman & Hall, Ipswich.

DARROCH, J.N.(1958) "The Multiple-Recapture Census. Stimations of a Close Populations". Biometrica, 45, 343-359.

DARROCH, J.N., y Ratcliff,D. (1980). "A Note on Capture-Recapture Estimation", Biometrics, 36, 149-153.

ENGEN, S. (1977) "Comments on two different approaches to the analysis of species frequenc data". Biometrics, 33,205-213.

ESTY, W.W.(1982). "Confidence Intervals for the Coverage of Low Coverage Samples", The Annals of Statistics, 10, 190-196.

————(1983). "A Normal Limit Law for a Nomparametric Coverage Estimator". Mathematical Scientist, 10, 41-50.

————(1986a). "The Size of a Coinage". Numismatic Chronicle, 146, 185-215.

————(1986b). "The Efficiency of Good's Nomparametric Coverage Estimator". The Annals of Statistics, 14, 1257-1260.

FELLER, W.(1993). "Introducción a la Teoría de Probabilidades y sus Aplicaciones", I y II", Ed. Limusa, México.

FISHER, R.A., Corbet,A.S. and Williams,C.B. (1943). "The Relation between the Number of Species and the Number of Individuals in a Random Sample from an Animal Population". Journal of Animal Ecology, 12, 42-58.

GOOD, I.J.(1950). "Probabylity and the Weigghing of Evidence. London. Charles Griffin.

————(1953). "On the Population Frequencies of Species and the Stimation of Population Parameters", Biometrika, 40, 237-264. GOOD, I.J. and TOULMIN, G.H.(1956). "The Number of New Species and the Increase in Population Coverage. When a5-63.

HOLTS, L. (1981). "Some Assintotic Results for Incomplete Multinomial or Poisson Samples", Scandinavian Journal of Statistics, 8, 243-246.

————(1986). "On Birthday, Collectors', Occupancy and Others Classical Urn Problems", International Statistical Review, 54, 15-27.

JOHNSON, N.L. and KOTZ, S.(1977) "Urn Models and their Application, John Wiley, Nueva York.

LO, S.(1992) "From Species Problem to a General Coverage Problem Via a New Interpretation", The Annals of Statistics, 20 1094-1109.

ORD, J.K. and Whitmore,G.A.(1986). "The Poisson Inverse-Gaussian Distribution as a Model for Species Abundance". Communications in Statisics, Part A. Theory and Methods, 15, 853-871.

PARZEN, E. (1972). "Procesos Estocásticos", Ed. Paraninfo, Madrid.

PORT, S.C. (1993). "Theoretical Probability for Applications". John Wiley and Sons. New York.

RAO, J.N.K. and Wu,C.F.J. (1988). "Resampling Inference with Complex Survey Data". Journal of the American Statistical Assotiation. 83, 231-241.

RAO, C.R. (1965). "Linear Statistical Inference and its Applications", Ed. John Wiley, Nueva York.

ROHATGI, V.K. (1976) "An Introduction to Probability Theory and Mathematical Statistics", Ed. John Wiley, Nueva York.

————(1984). "Statistical Inference", Ed. John Wiley, Nueva York.

STUART, A & ORD J. K. (1987). "Kndall's Advanced Theory of Statistics"- Vol. 2. 17.17. Charles Griffin & Co Ltd. London. (4ª Ed.).

ZELTERMAN, D. (1981). "Robust Estimation in Truncated Discrete Distributions with Applications to Capture-Recapture Experiments". Journal of Statistical Planning and Inference, 18, 225-237.