UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE MATEMÁTICA APLICADA
(BIOMATEMÁTICA)

PROBLEMA DE LAS ESPECIES: TRATAMIENTO UNIFICADO DESDE LA SUPERPOSICIÓN DE PROCESOS DE PUNTO.

Antonio Vargas Sabadías

Tesis Doctoral. Madrid,1997

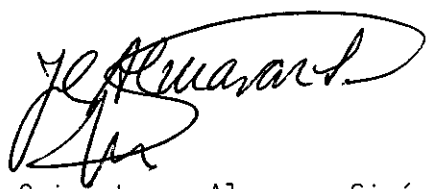
UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE MATEMÁTICA APLICADA
(BIOMATEMÁTICA)

PROBLEMA DE LAS ESPECIES: TRATAMIENTO UNIFICADO DESDE LA SUPERPOSICIÓN DE PROCESOS DE PUNTO.

Memoria presentada en la Universidad Complutense de Madrid para optar al Grado de Doctor por el Licenciado Antonio Vargas Sabadías.

La memoria titulada "*Problema de las especies: Tratamiento unificado desde la superposición de procesos de punto*", de la que es autor Antonio Vargas Sabadías, ha sido realizada en el Departamento de Matemática Aplicada (Biomatemática) de la Facultad de Ciencias Biológicas de la Universidad Complutense de Madrid, bajo mi dirección. A mi juicio, su contenido es científicamente valorable para constituir una tesis doctoral.

Madrid, 28 de Abril de 1997



Fdo.: Juan Crisóstomo Almaraz Simón, Doctor en Matemáticas y Profesor del Departamento de Matemática Aplicada (Biomatemática) de la Facultad de Ciencias Biológicas de la Universidad Complutense de Madrid.

ÍNDICE

INTRODUCCIÓN

Con frecuencia nos encontramos con una población sobre la que se ha establecido una partición que la divide en S clases, y el interés no es el de conocer el tamaño de las clases que forman la partición, sino averiguar el propio número S de clases de que consta dicha población.

Este tipo de problema se le plantea al biólogo y al geólogo cuando necesitan conocer el número de especies diferentes que hay en una población de animales o de plantas. Un problema concreto surge en Genética si se pretende averiguar el número de alelos neutros en un "locus", en una población, sabiendo el número de alelos neutros observados en una muestra. En numismática, se plantea el problema de conocer el número de troqueles que se utilizaron cuando se acuñaron determinadas monedas. Si esta información se complementa con el conocimiento del número de monedas emitidas por cada troquel, se pueden sacar conclusiones sobre la economía del país emisor en aquella época.

El número de errores en un programa informático, el número de ficheros repetidos en un archivo, el número de fenómenos astronómicos en observación son otros de los numerosos ejemplos en que es de aplicación el problema que nos ocupa.

Se trata de uno de los problemas clásicos de la Estadística, que no ha sido aún resuelto satisfactoriamente y que más dificultades ha encontrado, porque, cualquiera que sea el número de clases que se han observado, siempre cabe la posibilidad de un gran número de clases pequeñas que no han sido observadas.

Aunque han sido numerosos los procedimientos ensayados, lo han sido de modo independiente, por lo que se echa de menos un tratamiento unificado.

Los numerosos campos de aplicabilidad, la aparición, en los últimos años de nuevos enfoques (Lo(1983 y 1992), Chao(1990, 1991 y 1992), Bunge(1993) y nuevas técnicas,... nos indujeron a llevar a cabo un análisis del problema.

RESUMEN DE OBJETIVOS ALCANZADOS

RESUMEN DE OBJETIVOS ALCANZADOS

Para estimar el número de especies en una población, en la que hay establecida una partición, cuyas S clases son las distintas especies que la componen, se han diseñado numerosos modelos, aunque es poca la información comparativa disponible.

Se ha dedicado más atención a aquella situación en que las clases son todas del mismo tamaño, lo que supone admitir la hipótesis de equiprobabilidad u homogeneidad, es decir, todas las clases gozan de la misma probabilidad de ser elegidas.

Los estimadores discutidos hasta ahora, especialmente en una situación no homogénea, presentan inconvenientes que proporcionan dudas razonables acerca de las conclusiones experimentales.

El problema fundamental es que se desconoce el número de especies que no forman parte de la muestra, y, por consiguiente, la probabilidad de esas "*especies desconocidas*", y cabe la posibilidad de la existencia de especies raras.

El estimador más fiable, en el caso de una situación no homogénea, es el de Chao-Lee (1992).

Con el fin de conseguir una unificación en la teoría de muestreo, siguiendo las directrices marcadas por Bunge[1] y Fitzpatrick (1993), hemos sumergido el muestreo de las especies en una superposición de procesos de punto.

De este modo, cuando se recorre el intervalo (0,t], se recubre el muestreo de Poisson, mientras que, si se condiciona la superposición al número de sucesos que tienen lugar en el intervalo (0,t], se recubre el muestreo multinomial.

[1] Bunge, J. y Fitzpatrick, M. "Estimating the Number of Species: A Review". Journal of American Statistical Association. Marzo de 1993.

Asimismo, si se considera el número de especies que hay en un subintervalo (0,p] condicionado a que hayan tenido lugar A sucesos, se recubre el muestreo hipergeométrico.

Nuestro objetivo es, por una parte, conseguir hacer un tratamiento unificado del problema, y, por otra parte, explicar la relación existente entre los diversos estimadores, como puede ser la del estimador de Poisson y aquellos otros estimadores basados en el recubrimiento muestral como el de Darroch.

Suponemos una población infinita en la que hay establecida una partición definida por un número desconocido S de especies.

Suponemos además que las proporciones de las especies (también llamadas "*abundancias relativas*") están representadas por $p_1, p_2, \ldots, p_S$.

En un cierto instante, se toma una muestra aleatoria de la población, de modo que el tamaño muestral T quede dividido en D clases formadas por las especies representadas en la muestra.

Consideremos la partición de la población formada por la sucesión $M_1, M_2, \ldots, M_S$, donde M_k representa el número de especies que aparecen k veces en la muestra.

Designamos por D al número de especies diferentes que se dan en la muestra, al que llamaremos "*diversidad*", y por T al tamaño muestral, de forma que se cumplen las condiciones:

$$D = \sum_{j=1}^{T} M_j \qquad (0.1)$$

$$T = \sum_{j=1}^{D} jM_j \qquad (0.2)$$

Se define el recubrimiento muestral, C, como la suma de las probabilidades de las clases observadas, según veremos más adelante. Si designamos por C_0 la suma de las probabilidades de las especies desconocidas, es $C=1-C_0$.

Una primera estimación de C_0 es la de Good-Turing[2]:

$$\hat{C}_0 = \frac{\hat{M}_1}{T} \qquad (0.3)$$

La superposición de procesos de punto nos ha permitido diseñar dos grandes modelos: uno no paramétrico y otro paramétrico, que introducimos mediante un sistema de axiomas.

I. Modelo no paramétrico

Cuando observamos la superposición de procesos de Poisson homogéneos a lo largo del intervalo de tiempo (0,t], tenemos el modelo de muestreo de Poisson, mientras que, si condicionamos la superposición de procesos al tamaño muestral, se obtiene el modelo multinomial, siendo, en este caso, el tamaño muestral una constante y la diversidad una variable aleatoria, mientras que, en el modelo de Poisson, la diversidad es constante y el tamaño muestral variable.

A) Modelo de Poisson:

En el modelo de Poisson, la variable aleatoria $N_k(t)$, que representa el número de especies de la clase I_k que hay en la muestra en el instante t, se distribuye según una distribución de Poisson de media $\lambda_k t$. Estimando la media de la población a partir de la media muestral y apoyándonos en las propiedades de los primeros números de ocupación, se obtiene el estimador de Poisson para la probabilidad de especies desconocidas:

$$\hat{C}_0 = e^{-\frac{2\hat{M}_2}{\hat{M}_1}} \qquad (0.4)$$

de donde se obtiene el estimador de Poisson para S, versión de

[2] I.J.Good. "The Populations Frequencies of Species and the Estimations of Populations Parameters". Journal of the American Statistical Association. Diciembre de 1953. Volumen 40, Apartados 3 y 4.

Zelterman:

$$\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}} \qquad (0.5)$$

Desarrollando en serie de Taylor e^{-t} en un entorno de t=0, se obtiene un nuevo estimador de Poisson que depende del recubrimiento muestral:

$$\hat{S}_2 = \frac{D}{\frac{\hat{C}_1}{\hat{C}_0}\left(1-\frac{\hat{C}_1}{2\hat{C}_0}\right)} = \frac{D}{\frac{2\hat{M}_2}{\hat{M}_1}\left(1-\frac{\hat{M}_2}{\hat{M}_1}\right)} \qquad (0.6)$$

B) Modelo multinomial:

Tras comprobar que la condición necesaria y suficiente para que se cumpla la hipótesis de equiprobabilidad es que sea nulo el coeficiente de variación de Pearson de las p_k, surgen de modo natural, bajo la hipótesis de homogeneidad, los estimadores de Darroch (S_4):

$$\hat{S}_4 = \frac{D}{\hat{C}} \qquad (0.7)$$

y de máxima verosimilitud(S_3), que es la raíz de la ecuación:

$$\hat{S}_3 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}\cdot\frac{I}{S}}} \qquad (0.8)$$

Mediante un desarrollo en serie de Taylor de $e^{-p_k T}$ en un entorno de p_k=1/S, obtenemos el estimador de Chao-Lee para situaciones no homogéneas, que depende del recubrimiento muestral y del coeficiente de variación de Pearson de las p_k:

$$\hat{S}_5 = \frac{D}{\hat{C}} + \frac{\hat{M}_1}{\hat{C}}\hat{\gamma}^2 \qquad (0.9)$$

El estimador de Chao se basa en los métodos del recubrimiento muestral. Aplicando a este estimador la técnica de Lo, consistente en simular una prueba adicional, cuya probabilidad se deduce de los resultados de las pruebas anteriores, logramos reducir el sesgo del estimador de la probabilidad de especies desconocidas. Así tenemos como estimador de menor sesgo de P_0:

$$\hat{C}_0 = \frac{\hat{M}_1}{T} - \frac{2\hat{M}_2}{T(T+1)} \qquad (0.10)$$

que da lugar al estimador homogéneo:

$$\hat{S}_8 = \frac{\hat{D}T(T+1)}{(T+1)(T-\hat{M}_1)+2\hat{M}_2} \qquad (0.11)$$

y al estimador para situación no homogénea:

$$\hat{S}_{11} = \frac{\hat{D}T(T+1)}{(T+1)(T-\hat{M}_1)+2\hat{M}_2} + \frac{T(T+1)\hat{M}_1}{(T+1)(T-\hat{M}_1)+2\hat{M}_2}\hat{\gamma}^2 \qquad (0.12)$$

Haciendo el desarrollo en serie de Taylor tomando como variable el tiempo en un entorno de T/S, obtenemos un nuevo estimador para la situación no homogénea, que también depende del coeficiente de variación de Pearson de las abundancias relativas, pero que no se basa en el recubrimiento muestral, sino en la proporción de especies que hay en la muestra, D/T.

Procediendo como Chao, obtenemos un primer estimador para el caso homogéneo, que nos permite realizar una estimación previa de S, que utilizamos para estimar el coeficiente de variación de Pearson de las p_k.

Para el caso homogéneo, resulta:

$$\hat{S}_6 = \frac{\frac{T}{2}}{1-\frac{\hat{D}}{T}} \qquad (0.13)$$

Para situaciones no homogéneas:

$$\hat{S}_7 = \frac{T^3 \hat{R}_2}{4\,(T-\hat{D})^2\,(T-1)} \qquad (0.14)$$

donde $\hat{R}_2 = \frac{1}{T(T-1)} \sum_{k=2} k\,(k-1)\,E(M_k)$.

Aplicando la técnica de Lo a estos estimadores, se obtienen:

$$\hat{S}_{10} = \frac{T^2\,(T+1)}{2\,[\,(T+1-\hat{D}) - \hat{M}_1\,)\,]} \qquad (0.15)$$

y

$$\hat{S}_{11} = \frac{T^4\,(T+1)\,\hat{R}_2}{4\,[\,(T+1-\hat{D}) - \hat{M}_1]^2\,(T-1)} \qquad (0.16)$$

Si atendemos al comportamiento asintótico del proceso combinado, que sigue una distribución de Poisson de media M_1, conseguimos un intervalo de confianza para S en una situación homogénea del 95'55 %:

$$\frac{TD}{T-\hat{M}_1+2\sqrt{\hat{M}_1}} \leq S \leq \frac{TD}{T-\hat{M}_1-2\sqrt{\hat{M}_1}} \qquad (0.17)$$

Por otra parte, obtenemos la distribución aproximada del vector asintótico $M(M_1,\ldots,M_D)$, que es multinomial de parámetros $M_\alpha(T, M_1/S, \ldots, M_T/S)$, y que nos permite estimar la varianza de los diversos estimadores, puesto que, al ser todos ellos funciones diferenciables respecto de los M_k, se puede determinar su matriz de covarianzas.

II. Modelo paramétrico

Se puede llegar a este modelo a través de la distribución simétrica de Dirichlet, tal como lo hacen Keener[3], Rothman y Starr. Nuestro modelo parte de la distribución de Poisson compuesta con la distribución gamma, tal como procedió Esty (1985).

Este modelo nos ha permitido hacer un tratamiento unificado del problema de las especies. El modelo diseñado está regulado por la distribución binomial negativa de parámetros BN(A,p), que corresponde a un "esquema de contagio" (proceso de Polya para c=1):

$$P[N_j(t)=r/\lambda, N_k(t)>0] = \binom{A+r-1}{r}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{t+A}\right)^A} \quad (0.18)$$

Un mismo parámetro, A, va a determinar los distintos procesos de muestreo, dando lugar a los diferentes estimadores de S. Las abundancias relativas son ahora variables aleatorias, con distribución beta ß(A,(S-1)A), de donde se duduce que el cuadrado del coeficiente de variación de Pearson de las p_k es aproximadamente igual a 1/A.

El estimador de la probabilidad de especies desconocidas depende, por tanto, del coeficiente de variación de Pearson, obteniéndose:

$$\hat{P}_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}} = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{1}{\hat{\gamma}^2+1}} \quad (0.19)$$

1) Si el muestreo corresponde al esquema de contagio, se obtiene el estimador de Esty:

$$\hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\frac{1}{A} \quad (0.20)$$

[3] R.Keener, E.Rothman y N.Starr. "Distributions on Partitions". The Annals of Statistics. Volumen 15, nº 4, 1466-1481. 1987.

2) En el muestreo sin reemplazamiento, tenemos:

$$\hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} - \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\frac{1}{A} \qquad (0.21)$$

3) Cuando A tiende a infinito, corresponde al muestreo "con reemplazamiento", y resulta el estimador de Darroch:

$$\hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} \qquad (0.22)$$

A diferencia de Esty, que fija el valor del parámetro A previamente, asignándole el valor A=2, porque así conviene en numismática, pretendemos estimar A, teniendo en cuenta el comportamiento del proceso combinado en la superposición de procesos de punto.

Utilizando el razonamiento de Feller, hemos hallado la esperanza matemática del tiempo de espera hasta la primera renovación. Comparándola con el resultado obtenido en el análisis del proceso combinado, hemos obtenido un estimador de A:

$$\hat{A} = \frac{2\hat{M}_2}{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)-2\hat{M}_2} \qquad (0.23)$$

que proporciona una estimación del cuadrado del coeficiente de variación de Pearson de las p_k, independiente de la distribución de A:

$$\hat{\gamma}^2 = \frac{1}{\hat{A}} = \frac{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)-2\hat{M}_2}{2\hat{M}_2} \qquad (0.24)$$

Así se resuelve el problema de Chao-Lee, y no hay necesidad de realizar una estimación previa de S bajo la hipótesis de equiprobabilidad.

Substituyendo estas estimaciones, se consiguen los estimadores de S, para los diferentes tipos de muestreo.

A) Muestreo secuencial y completamente aleatorio sin reemplazamiento: Los estimadores de mayor interés son

$$\hat{S}_{13} = \frac{D}{1-\left(\frac{\hat{M}_1}{T}\right)^{\frac{2\hat{M}_2}{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)}}} \qquad (0.25)$$

$$\hat{S}_{12} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1^2(\ln\hat{T}-\ln\hat{M}_1)}{(\hat{T}-\hat{M}_1)\,2\hat{M}_2} \qquad (0.26)$$

que producen la misma estimación que el de Poisson-Zelterman en el modelo secuencial:

$$\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}} \qquad (0.27)$$

Aunque la expresión de los estimadores (0.25) y (0.26) es la misma para el muestreo secuencial y el muestreo completamente aleatorio sin reemplazamiento, el resultado no es el mismo, puesto que el segundo miembro de (0.26) toma un valor positivo en el muestreo secuencial, y es negativo en el muestreo completamente aleatorio sin reemplazamiento. La razón se fundamenta en el hecho de ser la estimación de A positiva en el muestreo secuencial, y, negativa cuando se trata del muestreo completamente aleatorio sin reemplazamiento.

También es diferente el resultado del estimador (0.25), puesto que el exponente en la binomial sería:

$$-A/(1-A).$$

B) Muestreo completamente aleatorio con reemplazamiento: Los estimadores de mayor interés son:

$$\hat{S}_4 = \frac{D}{1-\frac{\hat{M}_1}{T}} \qquad (0.29)$$

y el estimador de Poisson, que coincidirá con él:

$$\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}} \qquad (0.30)$$

Varianza de los estimadores:

Se demuestra que $D/(1-P_0)$ es un estimador de S de máxima verosimilitud.

Como estos estimadores son funciones diferenciables de los M_r, se pueden estimar las varianzas, siendo:

$$\sigma_S = \sum_{k=1} \sum_{h=1} \frac{\partial S}{\partial M_k} \frac{\partial S}{\partial M_h} \sigma_{kh} \qquad (0.31)$$

donde

$$\sigma_{kh} = \begin{cases} \hat{M}_k \left(1 - \frac{\hat{M}_k}{\hat{D}}\right) , & si\ k=h \\ -\frac{\hat{M}_k \hat{M}_h}{\hat{D}} , & si\ k \neq h \end{cases} \qquad (0.32)$$

En forma matricial, se puede expresar del siguiente modo:

$$\sigma_S^2 = L \Sigma L' \qquad (0.33)$$

siendo

$$L = \left(\frac{\partial S}{\partial M_1} , \frac{\partial S}{\partial M_2} , \ldots , \frac{\partial S}{\partial M_h} \right) \qquad (0.34)$$

L' es la traspuesta de L y $\Sigma = (\sigma_{kh})$ es la matriz de covarianzas de los M_k.

Para calcular estos parámetros, hay que considerar a D y a T como variables aleatorias dependientes de los M_k.

El criterio para decidir por uno u otro tipo de distribución está basado en la relación entre los estimadores de la probabilidad de especies desconocidas de Good-Turing y de Poisson-Zelterman:

$$\frac{\hat{M}_1}{T} \quad y \quad e^{-\frac{2\hat{M}_2}{\hat{M}_1}}$$

Estos resultados se pueden contrastar con el criterio de Katz, expuesto por Johnson y Kotz, y que se basa en el estimador:

$$\hat{K} = \frac{\hat{m}_2 - \hat{X}}{\hat{X}}$$

donde s^2 es la cuasivarianza, y $\hat{X}$ la media muestral.

En nuestro caso, este estimador es:

$$\hat{K} = \frac{D(\hat{T}+\hat{R}_2)}{\hat{T}(D-1)} - \frac{\hat{T}}{D}$$

de forma que la esperanza de este valor es aproximadamente I-1, y su varianza es aproximadamente 2/D, siendo I el índice de dispersión. Luego:

Si K>0, la distribución correspondiente es la binomial negativa;
Si K<0, se trata de la binomial;
Si K=0, la distribución que corresponde es la de Poisson.

Otras situaciones particulares son:

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

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

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, en cuyo caso

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

También corresponde a una situación no homogénea con un nivel de heterogeneidad del 67%.

c) A=0, que corresponde al modelo de Ewens, en cuyo caso la distribución adecuada es la distribución log-normal.

Se han llevado a cabo estudios del problema de las especies utilizando la distribución inversa-gauusiana como distribución de las abundancias relativas, por lo que añadimos un capítulo en que se resume el estado actual de la cuestión, que presenta una nueva forma de analizar el problema que nos ocupa.

PROBLEMA DE LA ABUNDANCIA DE ESPECIES

I. PROBLEMA DE LA ABUNDANCIA DE ESPECIES

1.1. Planteamiento

Supongamos una población, entre cuyos elementos hay una partición formada por S clases, y nos interesa precisamente averiguar el propio número S de clases.

La población puede ser finita o infinita. Si la población es finita, el muestreo puede realizarse con reemplazamiento (*muestreo multinomial*) o sin reemplazamiento (*muestreo hipergeométrico o de Bernoulli*). Si la población es infinita, el muestreo puede ser multinomial o de Bernoulli, o como resultado de la contribución aleatoria de Poisson de cada clase.

Dado un modelo de muestreo, se puede tratar de encontrar S mediante una formulación paramétrica o no paramétrica, y, en cada uno de los casos, se le puede dar un tratamiento bayesiano o no bayesiano.

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*".

Se trata de un problema que se resiste a una solución estadística, ya que cabe la posibilidad de que haya muchas especies "*raras*", que no son observadas.

Hay, para algunos modelos, estimadores apreciables, aunque aún está por hacerse un estudio global y comparativo de los diversos modelos.

1.2. Modelos teóricos

Se puede establecer una primera clasificación de los modelos teóricos de muestreo atendiendo al tamaño de la población.

Cuando la población es finita, tenemos el muestreo hipergeométrico y el muestreo de Bernoulli. Si la población es infinita, se distingue el muestreo multinomial, el muestreo de Poisson y el de Bernoulli..

En el modelo de muestreo multinomial, el vector aleatorio $\vec{N} = (N_1, N_2, \ldots, N_S)$ tiene una distribución multinomial con función masa de probabilidad dada por:

$$P(N_1 = n_1, \ldots, N_S = n_S) = \binom{T}{n_1, \ldots, n_S} \prod_{k=1}^{S} \pi_k^{n_k} \qquad (1.1)$$

siendo N_k el número de individuos de la clase k-ésima y $\sum_{k=1}^{S} n_k = T$.

En el modelo multinomial, se ha dedicado gran atención a aquella situación en que las clases son todas del mismo tamaño, situación que se conoce también como tratamiento homogéneo, que parte de la hipótesis de homogeneidad, es decir:

$$H_0 \equiv p_1 = p_2 = \ldots = p_S = \frac{1}{S}$$

Son numerosos los autores que han aplicado a este modelo las técnicas del recubrimiento muestral. Good, en 1953, a sugerencias de Turing, fue el primero que propuso, como estimador del recubrimiento muestral, $1-M_1/T$. Robbins, Darroch y Ratcliff, Lo y Chao, entre otros muchos, han desarrollado procedimientos basados en el recubrimiento muestral.

El modelo de Poisson fue propuesto por Fisher, Corbet y Willians en 1943, y parte de la hipótesis de que el número de representantes de la clase k-ésima de la muestra es una variable aleatoria de Poisson de media λ_k, siendo estas variables independientes. Entonces el tamaño muestral T es una variable aleatoria de media $\lambda t = \sum_{k=1}^{S} \lambda_k t$, con:

$$P[T=n] = e^{-\lambda t} \prod_{k=1}^{S} \frac{(\lambda_k t)^n}{n_k!} \qquad (1.2)$$

Cuando, además se supone que las λ_k son, a su vez, variables aleatorias, se obtiene el modelo paramétrico, que puede dar lugar a varios submodelos, según sea la distribución de las λ_k.

1.3. Tratamiento paramétrico

Son varios los modelos paramétricos que se han planteado. Uno de ellos es el que propone Keener[4], basado en la distribución simétrica de Dirichlet.

Keener parte de una población infinita en la que hay establecida una partición definida por un número desconocido S de especies.

Supone además que esta población evoluciona con el tiempo hasta que las proporciones de especies (también llamadas "*abundancias relativas*" y representadas por $p_1, p_2, \ldots, p_S$) alcanzan su equilibrio, que viene determinado por una distribución simétrica de las p_j, como es la distribución de Dirichlet D(A,...,A).

[4] Keener,R., Rothman,E. y Starr,N. "Distribution on Partitions". The Annals of Statistics, Vol. 15, Nª 4. Pág. 1466-1481. 1987.

En un cierto instante, antes de alcanzar el equilibrio, se toma una muestra aleatoria de la población, de modo que el tamaño muestral T quede dividido en D clases formadas por las especies representadas en la muestra.

De este modo, la distribución de esta partición es la distribución inducida por la mixtura de la distribución multinomial con la de Dirichlet, D(A,...,A).

En Genética, por ejemplo, si se consideran como especies los tipos de alelos en un "*locus*" genético, el modelo adecuado consiste en estimar los parámetros de una familia de distribuciones definidas sobre particiones.

Se considera la partición de la población formada por la sucesión $M_1, M_2, \ldots, M_S$, donde M_k representa el número de especies que aparecen k veces en la muestra.

Designando por D (*diversidad*) al número de especies diferentes que se dan en la muestra, y por T al tamaño muestral, se verifican las condiciones:

$$D = \sum_{j=1}^{T} M_j \tag{1.3}$$

$$T = \sum_{j=1}^{D} jM_j \tag{1.4}$$

Si se designa por N_i el número de especies que han sido capturadas de la clase I_i, admitiendo la distribución mixta de la multinomial con la de Dirichlet de parámetro A, se obtiene la distribución:

$$P[N_k = n_k, k=1,2,..] = \binom{S}{n_1 n_2 \ldots n_S} \prod_{k=1}^{S} p_k^{n_k} = \prod_{k=1}^{\infty} \frac{\binom{n_k+A+1}{n_k}}{\binom{T+SA+1}{T}} \tag{1.5}$$

de la que se deduce la ecuación para particiones, que, si se denota por G(k) el número de sucesos que tienen lugar k veces (número de especies que aparecen k veces en la muestra), es

$$P[G(k)=M_k,\ k=1,2,\ldots]=\binom{S}{D}\frac{D!}{\prod_{k=1}^{\infty}M_k!}\frac{\prod_{k=1}^{\infty}\binom{k+A+1}{k}^{M_k}}{\binom{T+SA+1}{T}} \qquad (1.6)$$

Hay, en esta distribución, dos parámetros desconocidos: A y S, que deben ser: A>-1 y S entero positivo. El parámetro A se puede interpretar como una medida del grado de variabilidad: cuanto más pequeño sea A, mayor será la heterogeneidad de las p_k.

Esta distribución puede presentar tres situaciones límites según los posibles valores de A. Si $A\to\infty$, la distribución (1.5) converge débilmente hacia la distribución de Maxwell-Boltzman, cuyo estudio realizaron Lewontin y Prout en 1953, y que hemos planteado desde otra perspectiva.

En esta situación, la distribución de los M_k quedaría en la forma:

$$P(M_k=m_k,\ k=1,2,\ldots)=\frac{S!\,T!}{(S-D)!\,S^T}\frac{1}{\prod_{k=1}^{S}M_k!\,(k!)^{M_k}} \qquad (1.7)$$

con las condiciones:

$$D=\sum_{k=1}^{T}M_k \quad \text{y} \quad T=\sum_{k=1}^{D}kM_k$$

Si A=1, se tiene la distribución de Bose-Einstein:

$$P(N_k=n_k,\ k=1,2,\ldots,S)=\frac{1}{\binom{S+T+1}{T}} \qquad (1.8)$$

donde N_k es el número de individuos de la especie k observados.

La tercera situación que es aquella en que $A \to 0$ y $S \to \infty$ de tal forma que $AS \to w$, cuya distribución converge débilmente a la distribución de Ewens:

$$P(G(k)=M_k, k=1,2,\ldots,S) = \frac{w^S}{\binom{T+w+1}{T}\prod_{k=1}^{S} M_k!^{M_k}} \qquad (1.9)$$

como puede verse en el artículo antes citado de Keener.

El tratamiento genérico se hace a partir de la distribución:

$$P(N_k=n_k, k=1,2,\ldots,S) = \frac{\prod_{k=1}^{S}\binom{A+k+1}{n_k}}{\binom{SA+T+1}{T}} \qquad (1.10)$$

que nos da la probabilidad conjunta de que el número de especies de la clase k que son observados sea n_k, k=1,2,...,S

A este mismo resultado se llega partiendo del modelo paramétrico que tiene como base la distribución binomial negativa, en el que nos vamos a apoyar.

Como los distintos modelos no son sino aproximaciones unos de los otros, vamos a tratar de conseguir tratarlos mediante una teoría de muestreo unificada, trabajo aún sin desarrollar.

En realidad, son dos los modelos estocásticos fundamentales diseñados:

1) el modelo multinomial no paramétrico debido a Good-Toulmin;

2) el modelo paramétrico de Fisher, Corbet y Willians.

1.4. Tratamiento unificado

La idea fundamental de este trabajo es la de hacer un estudio desde la construcción de una superposición de procesos independientes homogéneos de Poisson, con el fin de conseguir una representación de los diversos modelos, que permita realizar un estudio unificado, y, de este modo poder comparar los distintos resultados.

Comenzaremos por establecer una axiomática que nos sitúe claramente en cada uno de los modelos.

Supongamos que seleccionamos una muestra de tamaño T de una población que está formada por S clases, que forman una partición de dicha población, donde S es desconocido, así como se desconoce a priori la identidad de cada clase.

Admitimos, sin embargo, que, una vez es seleccionada una clase, ésta puede ser identificada.

El resultado del experimento puede ser representado, en teoría, por el vector aleatorio $\vec{N} = (N_1, N_2, \ldots, N_S)$, de modo que N_j representa "el número de elementos de la clase j-ésima que forman parte de la muestra".

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

En lugar de trabajar con $\vec{N}$, se trabaja con un vector como $\vec{M} = (M_1, M_2, \ldots, M_S)$, 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 las llamó Good.

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

Denotamos por D "*el número de clases o especies diferentes que hay en la muestra*", y lo llamamos "*diversidad*", según hemos señalado antes.

Con el fin de introducir los modelos multinomial y de Poisson de forma que sea posible unificar el estudio de los muestreos binomial, hipergeométrico, multinomial, de Poisson y de Bernoulli como aproximaciones unos de otros, vamos a sumergir el muestreo en una superposición de procesos de Poisson homogéneos, lo que equivale a admitir los siguientes axiomas:

Axioma I: La población consta de S especies (S es desconocido) con abundancias relativas $\vec{p}' = (p_1, p_2, \ldots, p_S)$, donde

$$\boxed{\sum_{j=1}^{S} p_j = 1, \quad 0<p_j<1} \tag{1.11}$$

Axioma II: Se selecciona 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

$$\boxed{P(N_j = r/\lambda_j) = P(r/\lambda_j) = \frac{e^{-\lambda_j}\lambda_j^r}{r!}} \tag{1.12}$$

La elección del axioma II supone admitir que

$$\boxed{E(N_j) = \lambda_j = \kappa p_j} \tag{1.13}$$

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, podemos definir el modelo como la superposición P de S procesos homogéneos independientes. Sean $Q_1, Q_2, \ldots, Q_S$ una sucesión de procesos de Poisson de intensidad 1, a cuyos tiempos entre llegadas vamos a designar por $Z_1, Z_2, \ldots$ Los Z_k son variables aleatorias independientes con distribución exponencial de media 1. Definamos a partir del proceso P una sucesión de procesos marcados. Para ello, a cada uno de los sucesos del proceso P le asociamos una marca $I_1, I_2, \ldots$ de modo que las marcas son independientes unas de otras, teniendo como distribución:

$$P[I_k = k] = p_k,\ k = 1, 2, \ldots, S \tag{1.14}$$

siendo las marcas también independientes de los tiempos entre llegadas.

De esta manera, tenemos los sucesos marcados I=k, que definen procesos de Poisson P_k de intensidad p_k, siendo los sucesos $P_1, P_2, \ldots, P_S$ independientes.

Definamos la variable aleatoria

$$H_{k,n} = \sum_{j=1}^{n} X_{k,j} \tag{1.15}$$

donde $X_{k,j}$ es el tiempo que transcurre desde que tiene lugar el suceso $I_{k,j}$ hasta que se verifica el $I_{k,j+1}$, siendo $X_{k,1}$ el tiempo que transcurre desde el origen de tiempos, 0, hasta que aparece la primera especie de la clase I_k, es decir, $H_{k,n}$ es la suma parcial n-ésima de las $N_{k,j}$, n=1,2,...

Las variables aleatorias $X_{k,j}$ son, por tanto, los tiempos entre llegadas correspondientes al proceso P_k, son independientes y tienen una distribución exponencial de media $1/\lambda_k$.

Sea t>0, y definamos la variable aleatoria $N_k(t)$ como "el número de elementos $H_{k,n}$ que hay en el intervalo (0,t] (lo que equivale al número de individuos de la especie I_k que han sido seleccionadas en el período (0,t]), podemos establecer la siguiente proposición:

Proposición 1.1: La variable aleatoria $N_k(t)$ tiene una distribución de Poisson de parámetro $\lambda_k t$.

Demostración:

$X_{k,1}, X_{k,2}, \ldots, X_{k,j}, \ldots$ son los tiempos entre llegadas correspondientes al proceso P_k, y

$$H_{k,n} = \sum_{j=1}^{n} X_{k,j} \; ; \; k = 1, 2, \ldots$$

los tiempos de espera correspondientes

a) Comprobemos, en primer lugar, que la proposición es cierta para $N_k(t)=0$. En efecto:

$$P[N_k(t)=0] = P[H_{k,0} \leq t] - P[H_{k,1} \leq t] = 1 - \lambda_k \int_t^{\infty} e^{-\lambda_k x} dx = 1 - e^{-\lambda_k t}$$

lo que demuestra la proposición para $N_k(t)=0$.

b) Supongamos ahora que n es un número entero positivo. Como los $X_{k,j}$ son no negativos, la sucesión de sumas parciales $H_{k,n}$ es una sucesión monótona no decreciente y se verifica que

$$P[N_k(t)=n] = P[H_{k,n} \leq t, H_{k,n+1} > t]$$

Como

$$P[H_{k,n} \leq t] = P[H_{k,n} \leq t, H_{k,n+1} > t] + P[H_{k,n+1} \leq t]$$

se deduce que

$$P[N_k(t)=n] = P[H_{k,n} \leq t] - P[H_{k,n+1} \leq t]$$

Ahora bien, al ser $H_{k,n}$ la suma de n variables aleatorias que siguen una distribución exponencial de media $1/\lambda_k$, los $H_{k,n}$ son independientes con una distribución gamma de parámetros $\Gamma(n, 1/\lambda_k)$, luego:

$$P[N_k(t)=n] = \int_0^t \frac{1}{\Gamma(n)} \lambda_k^n x^{n-1} e^{-\lambda_k x} dx - \int_0^t \frac{1}{\Gamma(n+1)} \lambda_k^{n+1} x^n e^{-\lambda_k x} dx =$$

$$= \frac{t^n \lambda_k^n e^{-\lambda_k t}}{n!} = \frac{(\lambda_k t)^n e^{-\lambda_k t}}{n!},$$

con lo cual queda demostrada la proposición.

Se puede extender el muestreo de Poisson haciendo que los λ_k sean, a su vez, variables aleatorias con una cierta distribución F, como puede ser la distribución gamma o la inversa gaussiana.

La elección de una distribución gamma para los λ_k nos lleva al modelo de Polya, que analizaremos con detalle, así como la elección de la inversa gaussiana conduce al modelo de Sichel. El axioma III nos lleva al modelo de Polya generalizado.

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

$$f(\lambda) = \frac{A^A}{\Gamma(A)} \lambda^{A-1} e^{-\lambda A}, \quad \lambda>0,\ 0<A<\infty,\ t>0. \tag{1.16}$$

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

Observación: El hecho de que los λ_j sean independientes no implica que lo sean los p_j, puesto que

$$p_j = \frac{\lambda_j}{\sum_{k=1}^{s} \lambda_k} \tag{1.17}$$

En efecto, como $\sum_{k=1}^{s} p_k = 1$ y $\lambda_k = \kappa p_k$, se verifica $\kappa = \sum_{k=1}^{s} \lambda_k$, de donde se deduce (1.17).

La elección del axioma III nos lleva al siguiente resultado:

Proposición 1.2: Si $N_j(t)$ es una variable que sigue una distribución de Poisson de parámetro $\lambda_j t$, y suponemos que los λ_j siguen una distribución gamma $\Gamma(A, 1/A)$, la distribución compuesta es la binomial negativa de parámetros $BN[A, A/(t+A)]$, con

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

En efecto, la distribución compuesta viene dada por:

$$P(N_j(t) = r/\lambda) = \frac{e^{-\lambda t}(\lambda t)^r}{r!} f(\lambda) =$$

$$= \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^r}{r!} \frac{A^A \lambda^{A-1} e^{-\lambda A}}{\Gamma(A)} d\lambda =$$

$$= \frac{A^A t^r}{r!\,\Gamma(A)} \int_0^\infty \lambda^{A+r-1} e^{-\lambda(A+t)} d\lambda = \frac{A^A t^r}{r!\,\Gamma(A)} \frac{1}{(A+t)^{A+r-1}} \int_0^\infty e^{-x} x^{A+r-1} \frac{dx}{A+t} =$$

$$= \frac{A^A t^r}{r!\,\Gamma(A)} \frac{1}{(A+t)^r} \int_0^\infty e^{-x} x^{A+r-1} dx = \frac{A^A \Gamma(A+r)}{r!\, t^A \Gamma(A)} \left(\frac{t}{A+t}\right)^{A+r} =$$

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

Se obtiene la distribución *binomial negativa*. Si denotamos por $N_k(t)$ el número de individuos de la clase I_k, $N_k(t)+A$ representa el número de individuos de la población necesarios para obtener A individuos de otras clases distintas de I_k, lo que

sucederá si, y sólo, si en la última prueba se obtiene un individuo que pertenece a una especie desconocida, y en las $y=N_k(t)+A-1$ pruebas anteriores habían aparecido $N_k(t)=r$ individuos de la especie I_k.

$P_{kr}(t)$ se interpreta como la probabilidad de que el A-ésimo éxito tenga lugar en el ensayo A+r. Utilizando el lenguaje propio del problema del número de especies, $P_{k,r}(t)$ es la probabilidad de que, en el instante t, se obtenga por primera vez A individuos de especies distintas de I_k, habiéndose obtenido r 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 en la muestra si $N_k(t)>0$:

$$P[N_j(t)=r/\lambda, N_k(t)>0] = \binom{A+r-1}{r}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{t+A}\right)^A} \quad (1.18)$$

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.

Si el axioma III se substituye por el axioma IV, obtenemos el modelo de Sichel o de la inversa gaussiana, cuya distribución truncada en el origen fue analizada por Sichel(1986) y Ord(1986).

Axioma IV: Los $\{\lambda_j\}_{j=1,2,\ldots,S}$ son independientes y están idénticamente distribuidos teniendo como distribución común la inversa Gaussiana de parámetros (μ,γ).

La distribución IG fue introducida por Schrödinger en 1915 para estudiar el primer tiempo de paso en el movimiento Browniano.

Su distribución truncada en el origen viene dada por:

$$P(R=r) = P[N_k(t)=r] = \frac{P_r}{1-P_0} = \frac{P_r}{\left[1-e^{\frac{\gamma}{\mu}-\frac{\gamma}{\mu_*}}\right]}, \quad r=1,2,\ldots \quad (1.19)$$

El modelo paramétrico nos va a permitir dar un tratamiento unificado de los distintos submodelos, que surgen aquí como casos particulares.

Para llevar a cabo un tratamiento unificado, necesitamos analizar los diferentes modelos desde distintos puntos de vista.

Así, vamos a ver:

A) Introducción de los diferentes tipos de muestreo desde el modelo teórico del esquema de urnas.
B) Generalización atendiendo al esquema de urnas de Polya.
C) Estudio de la distribución muestral.
D) Análisis del problema a partir de las ecuaciones diferenciales.

1.5. Esquema de urnas y modelo de muestreo

Un modo de trabajo conceptual con distribuciones estadísticas es el *esquema de urnas:* Se considera una población de N individuos (bolas en una urna), que son idénticas salvo en el color.

Suponemos que b=Np de las bolas son de un mismo color, por ejemplo, blanco, y que el resto a=Nq son azules, siendo p+q=1.

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 simple (o muestreo completamente aleatorio) con reemplazamiento.

Si se realiza una prueba simple durante un gran número de veces, podemos esperar que la proporción de veces que es seleccionada una bola blanca se aproxime a Np/N=p; es decir, las frecuencias relativas del número de bolas blancas seleccionadas serán f_0=q, f_1=p.

Ahora consideremos pares de pruebas; como las condiciones son idénticas en cada prueba, esperamos que dos bolas blancas tengan lugar con una frecuencia relativa de $(Np)^2/N^2=p^2=f_2$, $f_0=q^2p$ y $f_1=2pq$, ya que podemos seleccionar blanca bien en la primera , bien en la segunda extracción.

Generalizando, si llevamos a cabo n pruebas, las frecuencias relativas de j blancas y n-j azules serán:

$$f_j=\binom{n}{j}p^jq^{n-j}\ ,\ j=0,1,2,\ldots,n. \qquad (1.20)$$

El término p^jq^{n-j} da la frecuencia relativa de una sucesión específica de j blancas y n-j azules, siendo $\binom{n}{j}$ el número de sucesiones.

1.5.1. Aproximación de la binomial por la distribución de Poisson

Cuando la proporción p de éxitos en la población es muy pequeña, si el tamaño muestral es bastante grande como para que sea np apreciable cuando p es muy pequeño, la distribución binomial tiende a la de Poisson. Concretando, se tiene la siguiente proposición:

Proposición 1.3: Cuando $n\to\infty$, $p\to 0$, $np\to\lambda$, se verifica:

$$\binom{n}{r}p^rq^{n-r}\sim e^{-\lambda}\frac{\lambda^r}{r!}\ ,\ r=0,1,2,\ldots;\ \lambda>0.$$

En efecto:

$$\binom{n}{r}p^rq^{n-r}=\frac{n!}{(n-r)!\,r!}\,\frac{\lambda^r}{n^r}\left(1-\frac{\lambda}{n}\right)\sim\frac{\sqrt{(2\pi)\,e^{-n}n^{n+1/2}}}{\sqrt{(2\pi)\,(n-r)^{n-r+1/2}e^{-n+r}n^r}}$$

$$\sim \frac{1}{\left(1-\frac{r}{n}\right)^n e^r} \frac{\lambda^r}{r!} e^{-\lambda} \sim \frac{\lambda^r}{r!} e^{-\lambda}.$$

Hemos obtenido la distribución de Poisson, que viene dada por:

$$P_r = e^{-\lambda} \frac{\lambda^r}{r!} \ , \ r=0,1,2,\ldots; \lambda>0. \qquad (1.21)$$

Por tanto, "*cuando la población es infinita, el muestreo con reemplazamiento viene descrito por la distribución de Poisson*".

1.5.2. Esquema de urnas de Polya

Vamos a modificar 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. En tal caso, las pruebas sucesivas pueden dejar de ser independientes, y la frecuencia relativa de j éxitos en n pruebas es:

$$f_j = \binom{n}{j} \frac{Np\,(Np+jc-c)\,Nq\,(Nq+c)\ldots[Nq+(n-j-1)\,c]}{N\,(N+c)\ldots(N+nc-c)} \qquad (1.22)$$

Cuando c=0, (1.22) se reduce a (1.20), y tenemos el muestreo con reemplazamiento, cuya distribución, cuando N tiende a infinito, es la de Poisson, siendo la binomial cuando la población es finita.

1.5.2.1. Aproximación de la distribución hipergeométrica por la binomial

Cuando c=-1, no se devuelve ninguna bola, y tenemos el muestreo sin reemplazamiento. La distribución de frecuencias queda en la forma:

$$f_j = \frac{1}{N^{[n]}} (Np)^{[j]} (Nq)^{[n-j]} , \qquad (1.23)$$

donde $N^{[n]} = N(N-1)\ldots(N-n+1)$,

y j es tal que $máx(0, n-Nq) \leq j \leq mín(n, Np)$.

Como $N^{[n]} = \frac{N!}{(N-n)!}$, (1.23) se puede poner como:

$$f_j = \frac{\binom{Np}{j}\binom{Nq}{n-j}}{\binom{N}{n}} \qquad (1.24)$$

Esta es la distribución *hipergeométrica*. Cuando N tiende a infinito, (1.23) tiende a la binomial[5].

Luego, "*cuando la población es infinita, la distribución binomial describe el muestreo sin reemplazamiento*".

[5] Vijay K.Rohatgi. "Statistical Inference". Cap.6, apdo. 6.4. John Wiley & Sons. New York, 1984.

1.5.2.2. Aproximación de la hipergeométrica negativa por la binomial negativa

Cuando c=1 en el esquema de urnas, hay un elemento de "contagio", puesto que si tiene lugar un color particular, es más fácil que tengan lugar bolas del mismo color.

De (1.22) se deduce:

$$f_j = \frac{\binom{N}{j}(Np+j-1)^{[j]}(Nq+n-j-1)^{[n-j]}}{N+n-j)^{[n]}} = \frac{\binom{Np+j-1}{j}\binom{Nq+n-j-1}{n-j}}{\binom{N+n-j}{n}} \qquad (1.25)$$

donde j=0,1,...n. Se trata de la distribución *hipergeométrica negativa*.

Cuando N tiende a infinito, la distribución hipergeométrica negativa tiende a la binomial negativa.

Luego, "*cuando la población es infinita, el esquema de contagio es descrito por la distribución binomial negativa*".

1.5.3. Modelo de muestreo secuencial

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 el A-ésimo éxito. Este método de muestreo se denomina *secuencial* y tiene gran interés en medicina y en las pruebas de control de calidad en la industria: se trata de contar el número de fracasos hasta que se obtiene el A-ésimo éxito, (cada sucesión de pruebas debe terminar con el A-ésimo éxito).

Así se comporta el proceso de Polya, al que nos conducen los axiomas I, II y III. Corresponde a un esquema de contagio, cuya distribución es la binomial negativa de parámetros BN(A,p). La frecuencia de r fracasos es:

$$f_r = \binom{A+r-1}{r} p^A q^r, \quad r=0,1,2,... \qquad (1.26)$$

En el esquema de urnas de Polya, si se añaden c+1 bolas, según los distintos valores de c, se tienen los siguientes modelos de muestreo:

A) Población finita:

a) Cuando la población es finita y el tiempo discreto, el muestreo con reemplazamiento está regido por la distribución binomial de parámetros (T,p).
b) Si la población es finita, el muestreo sin reemplazamiento está regido por la distribución hipergeométrica:

$$P_r = \frac{\binom{Mp}{r}\binom{Nq}{n-r}}{\binom{N}{r}} \quad , \; r=0,1,2\ldots$$

c) Si la población es finita, el esquema de contagio está regido por la distribución hipergeométrica negativa.

B) Población infinita:

a) *esquema de contagio*: corresponde al caso c=1 y, cuando la población es infinita, está regido por la distribución binomial negativa;
b) *muestreo aleatorio simple con reemplazamiento*: corresponde a c=0 y, cuando la población es infinita, está regido por la distribución de Poisson;
c) *muestreo aleatorio simple sin reemplazamiento*: corresponde a c=-1, que, si la población es infinita, está regido por la binomial.

Los dos últimos surgen, como caso particular, de la binomial negativa.

Los modelos de muestreo finitos: Bernoulli, hipergeométrico e hipergeométrico multivariado pueden ser aproximados, según acabamos de ver, por los modelos de muestreo sin reemplazamiento (binomial), con reemplazamiento (Poisson) y esquema de contagio (binomial negativa).

1.6. Distribución muestral según el modelo

Consideremos una población formada por m objetos distintos a los que llamamos "*elementos*". Vamos a suponer que cada elemento puede ser identificado por uno de los números 1,2,...,m.

Supondremos también que cada elemento de la población puede presentar características adicionales, conocidas como "*atributos*", de modo que cada elemento de la población puede tomar un único valor de los diferentes atributos.

Si, en total, hay r atributos, entonces hay asociado a cada elemento x un vector r-dimensional, $\mathbf{h}(x)=(h_1(x),\ldots,h_r(x))$, donde $h_i(x)$ es el valor del i-ésimo atributo asociado al elemento x.

Una muestra de tamaño n es una selección de n elementos de la población. Los elementos que forman parte de la muestra no tienen por qué ser distintos. La muestra es con reemplazamiento si los elementos seleccionados son reemplazados cada vez. La muestra es sin reemplazamiento si los elementos seleccionados no son reemplazados.

Seleccionar una muestra de tamaño n con reemplazamiento es equivalente a seleccionar n muestras de tamaño 1, cada una procedente de la misma distribución de elementos.

1.6.1. Muestreo completamente aleatorio

Nos interesa conocer la distribución del vector $\mathbf{M}=(M_1,\ldots,M_n)$, que es uno de los problemas que nos interesa resolver.

Una muestra de tamaño n sin reemplazamiento puede obtenerse de dos formas distintas: 1) seleccionando los n elementos secuencialmente; o bien eligiéndolos todos a la vez.

Representaremos la muestra por la n-upla $(x_1,\ldots,x_n)$.

Cada elemento de la población puede tener diversas características llamadas atributos, de forma que, a cada muestra $(x_1,\ldots,x_n)$ le corresponde la n-upla formada por los vectores de atributos $(h(x_1),\ldots,h(x_n))$.

Se denomina *esquema de muestreo* a la regla mediante la cual se seleccionan los elementos de la población que van a formar parte de la muestra.

El esquema de muestreo induce sobre el conjunto

$$g^{(n)}=\{(x_1,\ldots,x_n) \mid x_i entero, 1\leq x_i\leq m\}$$

de todas las n-uplas formadas por elementos del conjunto $\{1,2,\ldots,m\}$, una distribución de probabilidad.

Cuando el muestreo es sin reemplazamiento, este conjunto se restringe al subconjunto $g_{(n)}$ de todas las n-uplas que son diferentes.

Hay m^n n-uplas con reemplazamiento y $(m)_n=m(m-1)\ldots(m-(n-1))$ sin reemplazamiento para una muestra de tamaño n.

Podemos, de esta forma, diseñar un modelo de muestreo por medio del vector aleatorio $\mathbf{X}=(X_1,\ldots,X_n)$, que toma valores en $g^{(n)}$, ó en $g_{(n)}$ y cuya distribución es la que corresponda al esquema de muestreo.

Definición 1.1: Una muestra de tamaño n *es una muestra completamente aleatoria* si el vector **X** se distribuye uniformemente sobre su conjunto de posibles valores, es decir, sobre el conjunto $g^{(n)}$, cuando el muestreo es con reemplazamiento, y sobre $g_{(n)}$ cuando el muestreo es sin reemplazamiento.

Según esta definición, en el muestreo completamente aleatorio con reemplazamiento, las variables aleatorias $X_1,\ldots,X_n$ son independientes y X_i tiene una distribución uniforme sobre el conjunto $\{1,2,\ldots,m\}$.

En el muestreo completamente aleatorio sin reemplazamiento, las variables aleatorias $X_1,\ldots,X_n$ están también distribuidas uniformemente sobre el conjunto $\{1,2,\ldots,m\}$, pero las variables $X_1,\ldots,X_n$ no son independientes, aunque sí son intercambiables, ya que las variables aleatorias $X_1,\ldots,X_n$ tienen como densidad conjunta la siguiente:

$$f(x_1,\ldots,x_n) = (m)_n^{-1} 1_{g(n)}(x_1,\ldots,x_n) \qquad (1.27)$$

La diferencia fundamental entre estos modelos de muestreo es que, en el muestreo con reemplazamiento, las variables aleatorias $X_1,X_2,\ldots,X_n$ son independientes.

En el muestreo completamente aleatorio sin reemplazamiento, la densidad marginal de $X_{r+1},\ldots,X_n$ dados $X_1=x_1,\ldots,X_n=x_n$, es:

$$f(x_{r+1},\ldots,x_n \mid x_1,\ldots,x_n) = (m-r)_{n-r} 1_{g(n)}(x_1,\ldots,x_n) \qquad (1.28)$$

El muestreo completamente aleatorio, tanto con reemplazamiento como sin reemplazamiento se da cuando los individuos de la población pueden estar clasificados en una de las D clases conocidas (I_1, $I_2,\ldots,I_D$) o en ninguna de éstas, sino en otra desconocida, a la que denotamos por I_0 y suponemos que las proporciones correspondientes en la población son $p_1,p_2,\ldots,p_D$ y P_0, entonces, las frecuencias relativas para una muestra formada por n individuos, en el muestreo con reemplazamiento, viene dada por términos de la forma:

$$f(x_0,x_1,\ldots,x_D) = \frac{n!}{x_0!x_1!\ldots x_D!} p_0^{x_0} p_1^{x_1}\ldots p_D^{x_D} \qquad (1.29)$$

que corresponden a los términos del desarrollo de

$$(P_0+p_1+\ldots+p_D)^n \qquad (1.30)$$

y cuya función característica viene dada por

$$\left(P_0 + p_1 e^{it_1} + \ldots + p_D e^{it_D}\right)^n, \qquad (1.31)$$

expresión que nos permite obtener los distintos momentos.

La distribución (1.29) es la distribución multinomial de parámetros $M_\alpha(n, p_1, \ldots, p_0)$, que corresponde al modelo de muestreo con reemplazamiento.

Cuando el tamaño muestral es grande, los muestreos con y sin reemplazamiento son esencialmente equivalentes[6], tendiendo la distribución hipergeométrica múltiple a la multinomial.

1.6.2. Muestreo secuencial

Se da el tipo de muestreo secuencial cuando suponemos que también en la población hay una partición formada por D+1 clases, pero admitimos que el muestreo continúa hasta que se obtienen A individuos de la clase desconocida I_0, en cuyo caso, las frecuencias relativas[7] son

$$f(X_1 = x_1, \ldots, X_D = x_D) = \frac{(x+A-1)!}{(A-1)!\,x_1!\ldots x_D!} p_1^{x_1} \ldots p_D^{x_D} P_0^A \qquad (1.32)$$

donde $x = \sum x_i$, siendo cada $x_i \geq 0$ y $P_0 = 1 - p_1 - p_2 - \ldots - p_D = 1 - p$.

La función característica viene dada por

$$P_0^A \left(1 - e^{-it_1} + \ldots + p_D e^{-it_D}\right)^{-A} \qquad (1.33)$$

Se trata de la distribución multinomial negativa de parámetros

[6] Vijay K. Rohatgi. "Statistical Inference". Cap.6. Apdo.6.7. Jhon Wiley & Sons. New York 1984.

[7] Alan Stuart & J.Keith Ord. "Kendall's Advanced Theory of Statistics" Vol.I, Cap.5. Charles Griffin & Company. London 1987.

$$MN(A, p_1, \ldots, p_D) \ , \ p_1 + \ldots + p_D = 1 - P_0 . \qquad (1.34)$$

que corresponde al modelo de muestreo secuencial.

Entonces podemos enunciar la siguiente proposición:

Proposición 1.4: En el muestreo secuencial se verifica:
1) el vector aleatorio **N** se distribuye según una multinomial negativa $MN(A, p_1, \ldots, p_D)$.

1.6.3. Definición de los números de ocupación y del recubrimiento muestral

Sea M_r el número de clases que tienen exactamente r elementos en la muestra. Se define entonces:

Definición 1.2:

$$M_r = \sum_{k=1}^{S} I[N_k(t) = r] \ , \ r = 1, \ldots, T \qquad (1.35)$$

donde I(X) es la función indicador y T el tamaño muestral. Los $M_r, r=1,2,\ldots$ son los *números de ocupación* (número de especies que tienen r representantes en la muestra).

Se define el *recubrimiento muestral C* como la suma de las probabilidades de las clases observadas. Si utilizamos la función indicador, es:

Definición 1.3:

$$C = \sum_{k=1}^{S} p_k I[N_k(t) > 0] \qquad (1.36)$$

Tanto C como los M_r son variables aleatorias que varían con la muestra.

Se define también C_0 (*suma de las probabilidades de las especies desconocidas*) como:

Definición 1.4:

$$C_0 = \sum_{k=1}^{S} p_k I[N_k(t) = 0] \tag{1.37}$$

Evidentemente se verifica: C_0=1-C.

Se define también la suma C_r de las *probabilidades de las especies que han aparecido r veces en la muestra* como:

Definición 1.5:

$$C_r = \sum_{k=1}^{S} p_k I[N_k(t) = r] \tag{1.38}$$

Entonces el recubrimiento muestral se puede expresar por:

$$C = \sum_{r=1} C_r = 1 - C_0 \tag{1.39}$$

Si el muestreo es sin reemplazamiento, la distribución de M es la hipergeométrica múltiple de parámetros $HM(n, M_1, \ldots, M_r)$, teniendo M_r la distribución hipergeométrica con parámetros $H(n, M_r, S)$.

Tenemos que estimar las M_r y las C_r, que no son parámetros, sino variables aleatorias. Veamos entonces qué se entiende por estimador insesgado de una variable aleatoria:

Definición 1.6: Un estimador W para estimar una variable aleatoria V se dice que es insesgado si, y sólo si E(estimador)=E(variable aleatoria).

Si W es un estimador de la variable aleatoria W, lo denotaremos por $\hat{w}$.

1.7. Introducción de los modelos a partir de sus ecuaciones diferenciales

Sea $N_k(t)$ el número de nacimientos (individuos) que aparecen en el intervalo (0,t]), y sea T_r el tiempo de espera hasta que tiene lugar el r-ésimo nacimiento en un proceso puro de nacimiento con intensidades $a_0, a_1, \ldots$
Entonces

$$P_r(t) = P[N_k(t) = r]$$

es la solución del sistema de ecuaciones diferenciales

$$\begin{aligned} P'_0(t) &= -a_0 P_0(t) \\ P'_r(t) &= -a_r P_r(t) + a_{r-1} P_{r-1}(t), \quad r \geq 1 \end{aligned} \tag{1.40}$$

con las condiciones iniciales

$$P_0(0) = 1, \quad P_r(0) = 0, \quad r \geq 1$$

y verificando $\sum \frac{1}{a_n} = +\infty$, con el fin de que $\sum P_r(t) = 1$.

1.7.1. Muestreo con reemplazamiento

Cuando $a_r = \lambda_r > 0$, siendo r un número natural, los procesos de nacimiento son procesos de Poisson con intensidad p_r. Entonces:

$N_k(t)$ sigue una distribución de Poisson de parámetros λ_r, y $H_{k,r}$ sigue una gamma de parámetros $(r, t/\lambda_k)$, es decir:

$$P[N_k(t) = r] = \frac{\lambda_k^r e^{-\lambda_k}}{r!}, \quad r \in \mathbb{N} \tag{1.41}$$

es la solución del sistema de ecuaciones diferenciales básicas, que corresponde a un proceso de nacimiento puro con tasa de nacimiento constante.

$$\begin{aligned} P'_0(t) &= -\lambda_r P_r(t) \\ P'_r(t) &= -\lambda_r P_r(t) + \lambda_r P_{r-1}(t), \quad r \geq 1 \end{aligned} \tag{1.42}$$

La función de densidad de los tiempos de espera es:

$$f_{H_{k,r}} = \frac{r!}{(r-1)!\,(T-r)!}\left(\frac{x}{t}\right)^{r-1}\left(1-\frac{x}{t}\right)^{T-r}\frac{1}{t}, \quad 0<x<t, \tag{1.43}$$

Luego, si llamamos $N_k^*(t) = N_k(t) \mid N_k(t) \geq 1$, $N_k^*(t)$ sigue la distribución de Poisson truncada en cero, de parámetro λ_k, y el tiempo de espera $H_{(k)}$ es el estadístico de orden k-ésimo de una muestra de tamaño T, procedentes de la distribución gamma anterior.

1.7.2. Muestreo sin reemplazamiento

Para un número natural A y $a_r = (A-r)\lambda_r$, $r=0,1,\ldots,A$, el proceso de nacimiento es equivalente a un proceso puro de muerte, es decir, un proceso de muerte puro con tasa de muerte lineal, y el esquema de urnas correspondiente es el muestreo aleatorio simple sin reemplazamiento.

Las ecuaciones del proceso, al que vamos a denotar por $J_k(t)$, son:

$$\begin{aligned} P'_0(t) &= -Ap_kP_0(t) + q_kP_1(t) \\ P'_r(t) &= -[rq_k(t) + (A-r)\,q_k]\,P_r(t) + (r+1)\,q_kP_{r+1}(t) + (A-r+1)\,p_kP_{r-1}(t), \quad 0 \leq r \leq A \end{aligned}$$

La solución[8] límite de estas ecuaciones es

$$P[J_k(t) = r] = \binom{n-A}{r} q^r (1-q)^{n-A-r}, \quad 0 \leq r \leq A-1 \tag{1.44}$$

[8] Feller, W. "Introducción a la Teoría de las Probabilidades y sus Aplicaciones", volumen I, capítulo XVII, apartado 7, e). LIMUSA. Méjico 1993.

donde q=t/(A+t).

Luego $J_k(t)$ es binomial de parámetros B(B,q), siendo B=n-A. Esta distribución se puede obtener también a partir de B procesos independientes de Poisson, cada uno de ellos con intensidad 1, en cuyo caso T_t será el número de procesos con al menos un suceso en el intervalo (0,t].

Una variable aleatoria con esta distribución también tendrá lugar si se toman B puntos al azar del intervalo (0,1] y se cuenta el número r de puntos que hay en el subintervalo (0,q] , o, lo que es equivalente, $J_k(t)$ representa el número de individuos de la especie I_k que hay en el subintervalo (0,p].

Como $J_k(t)$ debe ser mayor que cero, utilizaremos la distribución truncada en cero, por lo que la distribución que rige el modelo es:

$$P[J_k(t)=r]=\binom{B}{r}\frac{p^r(1-p)^{n-A-r}}{1-q^A}\ ,\quad 0<r\leq A \qquad (1.46)$$

Si A representa el tamaño inicial de la población, como los individuos mueren, el tamaño se reduce en este modelo.

1.7.3. Muestreo secuencial (Proceso de Polya)

El proceso de Polya se obtiene como paso al límite del esquema de urnas de Polya, y se trata de un proceso de nacimiento, no estacionario.

Las ecuaciones diferenciales que dan lugar al proceso son:

$$P'_r(t)=-\frac{A+r}{t}\lambda P_r(t)+\frac{A+r-1}{t}\lambda P_{r-1}(t)\ ,\ r\geq 1$$
$$P'_0(t)=0$$

con las condiciones iniciales

$$P_i(0)=1\ ,\ P_r(0)=0\ ,\ r\neq A$$

La solución a este sistema fue dada, en primer lugar, por Yule en la forma:

$$P_y(t) = \binom{y-1}{A-i} e^{i\lambda t}(1-e^{-\lambda t})^{y-r},\ y=A,A+1,\ldots \qquad (1.47)$$

Se trata por tanto, de una distribución binomial negativa de parámetros:

$$BN(A-i;i;e^{-\lambda t}) \qquad (1.48)$$

Esta distribución se generaliza dando lugar a la forma:

$$P[N_k(t)=r] = \binom{A+r-1}{A-1}\left(\frac{t}{A+t}\right)^r\left(\frac{A}{A+t}\right)^A,\ r=1,2,\ldots \qquad (1.49)$$

en que la vamos a utilizar, y que corresponde a un proceso de nacimiento puro, con tasa de nacimiento lineal.

1.8. Esquema de trabajo

Hemos diseñado dos modelos de muestreo: uno no paramétrico y otro paramétrico, del que se deduce el no paramétrico como caso particular. Por motivos metodológicos, vamos a desarrollar, en primer lugar, el modelo no paramétrico, aunque más adelante veremos cómo surge del paramétrico. La distribución binomial negativa de parámetros (A,p) regula el modelo paramétrico, siendo A el parámetro que mide el grado de heterogeneidad de la distribución.

En los apartados 1.5 y 1.6, vimos cómo los modelos de muestreo hipergeométrico, de Bernoulli e hipergeométrico multivariado pueden ser aproximados por las distribuciones binomial, de Poisson y binomial negativa.

Por ello, nos vamos a limitar a estudiar el problema de las especies en poblaciones con un número infinito de elementos. Seguiremos el siguiente orden:

1. **Modelo no paramétrico**:

A) Muestreo de Poisson.
B) Muestreo multinomial.

2. **Modelo paramétrico**:

A) Muestreo secuenccial, que está descrito por la distribución binomial negativa; se obtiene componiendo la distribución de Poisson con la gamma.
   a) Esquema de contagio (distribución binomial negativa)
   b) Muestreo con reemplazamiento, regido por la distribución de Poisson. Muestreo de Maxwell-Boltzman.
   c) Muestreo sin reemplazamiento, que regula la distribución binomial.
   d) Muestreo de Ewens (A=0).
   e) Muestreo de Bose-Einstein (A=1).

3. Nuevo modelo paramétrico: La distribución inversa gaussiana.

1.9. Resumen del capítulo

Se plantea el problema de las especies, analizando los principales modelos diseñados. Observamos la no existencia de un desarrollo unificado, y, siguiendo las directrices marcadas por J.Bunge y M.Fitzpatrick, tratamos de representar el problema de las especies desde la superposición de procesos de punto. Con esta finalidad, se definen los conceptos fundamentales y se establece una axiomática que nos sitúa con precisión en uno u otro de los dos modelos que vamos a diseñar:

a) un modelo no paramétrico, que está regulado por la distribución de Poisson;

b) un modelo paramétrico, regulado por la distribución binomial negativa.

El modelo no paramétrico surgirá más adelante del paramétrico como caso particular, cuando el valor del parámetro tiende a infinito.

Partiendo del concepto de esquema de urnas y del esquema de urnas de Polya, se obtienen las distribuciones que regulan los modelos, comprobando cómo el muestreo hipergeométrico, de Bernoulli e hipergeométrico multivariado se pueden aproximar por las distribuciones binomial, de Poisson y binomial negativa, respectivamente.

Se analizan los modelos también desde sus distribuciones muestrales, comprobando cómo, en el muestreo completamente aleatorio con reemplazamiento, la distribución muestral es la multinomial. Asimismo, cuando la muestra es suficientemente grande, el muestreo sin reemplazamiento tiende también a la distribución multinomial.

En el modelo de muestreo secuencial la distribución muestral es la multinomial negativa.

También son analizados los tres modelos de muestreo a partir de sus ecuaciones diferenciales, concluyendo que el modelo secuencial corresponde a un proceso puro de nacimiento con tasa de crecimiento lineal, el muestreo completamente aleatorio a un proceso de Poisson, y el muestreo sin reemplazamiento es un proceso puro de muerte.

De acuerdo con estos resultados, se ha planteado el esquema que se debe de seguir en los capítulos posteriores.

El muestreo secuencial tiene como distribución la multinomial negativa, el muestreo completamente aleatorio con reemplazamiento, la de Poisson, y el muestreo completamente aleatorio sin reemplazamiento, la binomial.

MODELO NO PARAMÉTRICO

II. MODELO NO PARAMÉTRICO

2.1. Modelos de muestreo multinomial y de Poisson

Hemos definido el modelo genérico como la superposición de S procesos homogéneos independientes $P_1, P_2, \ldots, P_S$, en el sentido de que cada clase I_i contribuye proporcionando elementos al muestreo de acuerdo con el proceso P_i.

El proceso de muestreo consiste en seleccionar W_D especies, D de las cuales son diferentes (D es la diversidad y W_D va a coincidir con el tamaño muestral), de modo que la especie I_1 aparece n_1 veces, la especie I_2 aparece n_2 veces,..., la especie I_D aparece n_D veces.

La admisión de los axiomas I, II y III nos permite, según vimos en el capítulo primero, sunergir el muestreo en una superposición de procesos de punto, en este caso, de procesos de Poisson, de forma que se introducen paralelamente los modelos de Poisson y multinomial, que van a estar regulados, respectivamente, por las distribuciones

$$P_{k,r} = \frac{(\lambda_k t)^r e^{-\lambda_k t}}{r!(1-e^{-\lambda_k t})} \quad \text{y} \quad P_{k,r} = \frac{(p_k T)^r e^{-p_k T}}{r!(1-e^{-p_k T})}$$

según veremos, donde T es el tamaño muestral.

2.2. Modelo de Poisson

El modelo de Poisson fue propouesto por Fisher, Coorbet y Willians en 1943 en un trabajo sobre las especies de *Lepidoptera*. En este modelo, el tamaño muestral es una variable aleatoria con una distribución de Poisson de media $\lambda t = \sum_{j=1}^{S} \lambda_j t$, de modo que el número de individuos de cada clase o especie va a seguir una distribución de Poisson de media $\lambda_j t$, $j=1,2,\ldots,S$.

En el primer capítulo, definimos la variable aleatoria $N_k(t)$ como "el número de individuos de la especie I_k que han sido seleccionados en el período (0,t]", lo que nos permitió establecer la proposición 1.1:

> **Proposición 1.1**: La variable aleatoria $N_k(t)$ tiene una distribución de Poisson de media $\lambda_k t$.

2.2.1. Propiedad aditiva de la distribución de Poisson

Las siguientes proposiciones son el fundamento del modelo de Poisson:

Proposición 2.1: Si $N_1(t),\ldots N_S(t)$ son S variables aleatorias independientes, todas con distribución de Poisson de parámetro $\lambda_k t$, k=1,2,...,S; la variable aleatoria $T=N_1(t)+\ldots+N_S(t)$ sigue también una distribución de Poisson de parámetro λt, en donde $\lambda_1+\ldots+\lambda_S=\lambda$.

Se demuestra esta proposición teniendo en cuenta que la suma de S variables aleatorias independientes de Poisson es una variable de Poisson, cuyo parámetro es la suma de los parámetros de cada una de las variables.

También es cierta la recíproca:

Proposición 2.2: Si el número de especies, T, que hay en la muestra se distribuye según una distribución de Poisson de media λt, el número, $N_k(t)$, de individuos de la clase I_k que hay en el intervalo (0,t] sigue una distribución de Poisson de media $\lambda_k t$, siendo las variables $N_k(t)$ independientes y $\lambda = \sum_{k=1}^{S} \lambda_k$.

Demostración: Supongamos que el número de especies T que hay en la muestra es una variable aleatoria y no un número fijo, como sucede cuando observamos los procesos hasta que ha transcurrido un período de tiempo t fijo.

Por el teorema de la probabilidad total, sabemos que el vector aleatorio $(N_1(t),\ldots,N_S(t))$ tiene como distribución la multinomial dada por

$$P(N_1(t)=x_1,\ldots,N_S(t)=x_S) = \frac{(x_1+\ldots+x_S)!}{x_1!\ldots x_S!}\pi_1^{x_1}\ldots\pi_S^{x_S}P(T=x_1+\ldots+x_S) \quad (2.1)$$

con $\pi_k = \dfrac{\lambda_k}{\sum_{j=1}^{S}\lambda_j}$, ya que es evidente que:

$$P(N_1(t)=x_1,\ldots,N_S(t)=x_S \mid T=n) = 0$$

a menos que $n=x_1+\ldots+x_S$.

Consideremos a T como una variable aleatoria que tiene una distribución de Poisson de media λt, es decir:

$$P(T=n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}1_{[0,..1,...]}(n), \quad t>0$$

Como $\lambda = \lambda_1+\ldots+\lambda_S$, se verifica que:

$$e^{-\lambda t} = e^{-\lambda_1 t-\ldots-\lambda_S t} = e^{-\lambda_1 t}\ldots e^{-\lambda_S t}$$

Entonces la expresión (2.1) queda en la forma:

$$P(N_1(t)=x_1,\ldots,N_S(t)=x_S) =$$

$$= \frac{(x_1+\ldots+x_S)!}{x_1!\ldots x_S!}\left(\frac{\lambda_1}{\lambda}\right)^{x_1}\ldots\left(\frac{\lambda_S}{\lambda}\right)^{x_S} e^{-\lambda_1 t}\ldots e^{-\lambda_S t}\frac{(\lambda t)^{x_1+\ldots+x_S}}{(x_1+\ldots+x_S)!} =$$

$$= \prod_{i=1}^{S} \frac{(\lambda_i t)^{x_i}}{x_i!} e^{-\lambda_i t}$$

Se trata de la distribución conjunta de S variables aleatorias independientes, con distribución de Poisson de parámetros $\lambda_1 t,\ldots,\lambda_S t$, respectivamente. Luego $N_1(t)....,N_S(t)$ son independientes, y $N_k(t)$ tiene distribución de Poisson de media $\lambda_k t$.

Este resultado nos permite afirmar que:

"Cuando observamos la superposición de S procesos durante el intervalo (0,t], recubrimos el muestreo de Poisson con parámetros $\lambda_1 t,\ldots,\lambda_S t$, "

La distribución que rige este modelo, al ser desconocido el número de especies que no forman parte de la muestra, es la distribución de Poisson truncada en cero, a la que designamos por $N_k^*(t)$ (es $N_k(t)$ condicionada por $N_k(t)>0$). Su función masa de probabilidad es:

$$P^*_{k,r} = P(N^*_k(t)=r) = \frac{(\lambda_k t)^r e^{-\lambda_k t}}{r!\,(1-e^{-\lambda_k t})}\ ,\ r=1,2,\ldots \qquad (2.2)$$

2.2.2. Muestreo con reemplazamiento

El modelo formado por la superposición de S procesos de Poisson independientes, corresponde a un esquema de urnas con un muestreo aleatorio con reemplazamiento.

En efecto:

$$P_{k,r} = P[N_k(t) = n_k] = \frac{(\lambda_k t)^{n_k} e^{-\lambda_k t}}{n_k!}$$

es la solución del sistema de ecuaciones diferenciales:

$$P_{k,0}(t) = -\lambda_k P_{k,0}(t)$$
$$P_{k,r}(t) = -\lambda_k P_{k,r}(t) + \lambda_k P_{k,r-1}(t), \quad r \geq 0$$

con las condiciones iniciales

$$P_{k,0}(0) = 1 \quad , P_{k,r}(0) = 0 \quad , r \geq 1.$$

$\{N_k(t), t \geq 0\}$ es un proceso de Poisson de media λ_k, cuyos tiempos entre llegadas, X_{ki}, van a ser independientes y con una misma distribución exponencial de media $1/\lambda_k$. Los tiempos de espera, $H_{kn}=X_{k1}+X_{k2}+\ldots+X_{kn}$, serán también independientes con una distribución gamma de parámetros $(n, 1/\lambda_k)$.

El proceso de Poisson satisface las ecuaciones prospectivas:

$$P'_{i,r}(t) = -\lambda_k P_{i,r}(t) + \lambda_k P_{i,r-1}(t)$$

y las retrospectivas:

$$P'_{i,r}(t) = -\lambda_k P_{i,r}(t) + \lambda_k P_{i+1,r}(t)$$

2.2.3. Distribución de los tiempos entre llegadas

Los tiempos entre llegadas, $X_{k,1}$, $X_{k,2}$,... correspondientes a cada proceso $\{N_k(t), t\geq 0\}$ son independientes, teniendo todos la misma distribución: exponencial de media $1/\lambda_k$, cuya función de densidad es, por tanto:

$$f_{X_{k,i}}(x) = \lambda_k e^{-\lambda_k x} \qquad (2.3)$$

siendo $E(X_{k,i}) = \frac{1}{\lambda_k}$ y $Var(X_{k,i}) = \frac{1}{\lambda_k^2}$. .

Esto significa que, comenzando por un origen de tiempo arbitrario, los subsiguientes puntos tienen lugar en los instantes $X_{k1}, X_{k2}, \ldots$, de modo que las variables aleatorias X_{ki} son independientes y tienen todas la misma distribución.

X_{k1} es el tiempo que transcurre, en el proceso P_k, desde el origen de tiempos, 0, hasta que tiene lugar el primer suceso, y, para j>1, X_{kj} es el tiempo que transcurre, en el proceso P_k, desde que tiene lugar el suceso (j-1)-ésimo hasta el j-ésimo.

2.2.4. Distribución de los tiempos de espera

Los tiempos de espera, $H_{k,n}$, representan el tiempo transcurrido hasta que han tenido lugar n sucesos, de modo que

$$H_{k,n} = X_{k1} + X_{k,2} + \ldots + X_{k,n} \ , \ n \geq 1 .$$

Al ser sumas de n variables aleatorias independientes con distribución exponencial de media $1/\lambda_k$, los tiempos de espera son también variables aleatorias independientes con distribución gamma de parámetros $(n, 1/\lambda_k)$.

Luego la función de densidad de los tiempos de espera viene dada por:

$$f_{H_{k,n}}(t) = \frac{\lambda_k t^{n-1}}{(n-1)!} e^{-\lambda_k t}$$

siendo

$$E[H_{k,n_k}] = \frac{n_k}{\lambda_k} \quad , \quad Var[H_{k,n_k}] = \frac{n_k}{\lambda_k^2} .$$

Los tiempos entre llegadas se expresan en función de los tiempos de espera por las relaciones:

$$X_{k,1} = H_{k,1} \, , X_{k,2} = H_{k,2} - H_{k,1} , \ldots , X_{k,n} = H_{k,n} - H_{k,n-1}$$

La relación[1] entre los procesos y los tiempos de espera viene dada por:

$$\boxed{N_k(t) \leq n \Leftrightarrow H_{k,n+1} > t} \qquad (2.4)$$

de donde se deduce que

$$\boxed{N_k(t) = n \Leftrightarrow H_{k,n} \leq t \;\; y \;\; H_{k,n+1} > t} \qquad (2.5)$$

lo que nos dice que "en el intervalo de tiempo (0,t] tienen lugar exactamente n sucesos si, y sólo si el tiempo de espera hasta que ha tenido lugar el suceso n-ésimo es menor o igual que t, y el tiempo de espera hasta el suceso (n+1)-ésimo es mayor que t".

[1] E.Parzen. "Procesos Estocásticos". Cap.4, apdo.4.5. Paraninfo. Madrid 1972.

De las últimas relaciones se deducen:

$$P[N_k(t) \leq n] = P[H_{k,n+1} > t] \, , \; n=0,1,2,\ldots$$

$$P[N_k(t) = n] = P[H_{k,n} \leq t] - P[H_{k,n+1} \leq t] \, , \; n=1,2,\ldots$$

$$P[N_k(t) = 0] = 1 - P[H_{k,1} \leq t]$$

Estas relaciones se pueden exponer por medio de las funciones de distribución:

$$F_{N_k(t)} = 1 - F_{H_{k,n+1}}(t) \, , \; n=0,1,2,\ldots$$

$$p_{N_k(t)} = F_{H_{k,n}}(t) - F_{H_{k,n+1}}(t) \, , \; n=1,2,\ldots$$

$$p_{N_k(0)} = 1 - F_{H_{k,1}}(t)$$

2.2.5. Propiedades de la distribución

Según hemos apuntado antes, al ser desconoccidas las especies que no aparecen en la muestra, utilizamos la distribución truncada en cero:

$$P\left[N_k^*(t) = r\right] = \frac{(\lambda_k t)^r e^{-\lambda_k t}}{r! \, (1 - e^{-\lambda_k t})} \qquad (2.6)$$

donde $N_k^*(t) = N_k(t) \mid N_k(t) > 0$. Se cumplen, por tanto, las siguientes propiedades:

1. $$\sum_{r=1}^{S} P_{k,r} = 1 - P_0 \qquad (2.7)$$

2. $$\sum_{r=0}^{S} P_{k,r} = 1 \qquad (2.8)$$

3. $$e^{-\lambda_k t} = P_0 \qquad (2.9)$$

4. $$E[N_k^*(t)] = \frac{\lambda_k t}{1 - e^{-\lambda_k t}} \qquad (2.10)$$

5. $$E[N_k^*(t)]^2 = \frac{\lambda_k t}{1 - e^{-\lambda_k t}} + \frac{(\lambda t_k)^2}{1 - e^{-\lambda_k t}} \qquad (2.11)$$

6. $$\mu_2 = Var[N_k^*(t)] = \frac{\lambda_k t}{1 - e^{-\lambda_k t}} \left(1 + \lambda_k t - \frac{\lambda_k t}{1 - e^{-\lambda_k t}} \right)$$

7. Índice de dispersión:

$$\frac{\mu_2}{\alpha_1} = 1 + \lambda_k t - \frac{\lambda_k t}{1 - e^{-\lambda_k t}} \qquad (2.12)$$

8. $$E\left[N_k^*(t)\right]^3 = \frac{(\lambda_k t)^3}{1 - e^{-\lambda_k t}} \left(1 - \frac{3}{1 - e^{-\lambda_k t}} + \frac{2}{1 - e^{-3\lambda_k t}} \right) \qquad (2.13)$$

2.3. Modelo de Poisson

Un proceso estocástico puede ser observado bien en un período de tiempo fijo, bien hasta que han tenido lugar un número fijo de sucesos. En el primer caso, es aleatorio el número de sucesos que tienen lugar durante el período fijo de tiempo, es decir, es aleatorio el tamaño muestral T, mientras que, en el último caso, es aleatoria la amplitud del intervalo de tiempo.

Para situarnos en un proceso de Poisson, supongamos que es observado durante un intervalo de tiempo fijo (0,L].

Supongamos que se han observado T sucesos durante este período de tiempo en los instantes $0<t_1<t_2<..<t_T<L$.

Los tiempos entre llegadas $T_r=t_r-t_{r-1}$, (r=1,2,...,T) son independientes y tienen todos la misma distribución exponencial.

La probabilidad de que aparezca una especie desconocida es:

$$P_0 = e^{-\lambda_k L} \tag{2.14}$$

Tomando logaritmos en ambos miembros y sumando desde K=1 hasta S, resulta:

$$-\lambda_k L = \ln P_0 \Rightarrow -\lambda L = S \ln P_0 \Rightarrow P_0 = e^{\frac{-\lambda L}{S}} \tag{2.15}$$

La distribución del proceso puede ponerse entonces en la forma:

$$P_r = \left(\frac{\lambda L}{S}\right)^r \frac{e^{\frac{-\lambda L}{S}}}{r!\left(1-e^{\frac{-\lambda L}{S}}\right)} \tag{2.16}$$

Los números de ocupación en tal modelo, se pueden expresar por:

$$E[M_r] = S\left(\frac{\lambda L}{S}\right)^r \frac{e^{\frac{-\lambda L}{S}}}{r!\left(1-e^{\frac{-\lambda L}{S}}\right)} \tag{2.17}$$

Particularizando para los dos primeros números de ocupación, dividiendo miembro a miembro e igualando en las expresiones anteriores, se obtiene:

$$\frac{\lambda L}{S} = \frac{2\hat{M}_2}{\hat{M}_1} \tag{2.18}$$

de donde resulta el estimador de Poisson (versión de Zelterman):

$$\boxed{\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}}} \tag{2.19}$$

2.3.1. Nuevo estimador de Poisson

Si desarrollamos en serie e^{-t} en un entorno de t=0, y particularizamos para $t=\lambda L/S$, obtenemos también un estimador de P_0:

$$e^{\frac{-\lambda L}{S}} = 1 - \frac{\lambda L}{S} + \frac{(\lambda L)^2}{2S^2} + \frac{(\lambda L)^3 e^{-\theta t}}{6S^3} \quad , \; 0<\theta<1 \tag{2.20}$$

donde el resto es menor que $\frac{1}{6}\left(\frac{\lambda L}{S}\right)^3$

Teniendo en cuenta (2.19) y la relación con los dos primeros números de ocupación, se obtiene el estimador de Poisson en la nueva expresión:

$$\hat{S}_2 = \frac{D}{\frac{2\hat{M}_2}{\hat{M}_1}\left(1 - \frac{\hat{M}_2}{\hat{M}_1}\right)} \qquad (2.21)$$

Este estimador puede también expresarse en función de los recubrimientos muestrales, quedando en la forma:

$$\hat{S}_2 = \frac{D}{\frac{\hat{C}_1}{\hat{C}_0}\left(1 - \frac{\hat{C}_1}{2\hat{C}_0}\right)} \qquad (2.22)$$

Se plantea también, con frecuencia, el interrogante de si un conjunto de sucesos son realmente de Poisson.

Para determinar si un proceso es de Poisson, se puede realizar un contraste de bondad de ajuste. También se puede utilizar un procedimiento cuyo fundamento es la relación entre la realización de sucesos de Poisson y la distribución uniforme.

Sean $0<t_1<t_2<\ldots<t_{nk}<L$ las épocas en que tuvieron lugar los sucesos de Poisson durante (0,L], y consideremos la expresión

$$T_r = \sum_{j=1}^{r} t_j \qquad (2.23)$$

donde T_r es la suma de r variables aleatorias distribuidas uniformemente en el intervalo (0,L]. Si observamos que la media y la varianza de una variable aleatoria uniforme en (0,L] vienen dadas por L/2 y $L^2/12$, respectivamente, tenemos:

$$E[T_r] = \frac{rL}{2} \quad y \quad V[T_r] = \frac{rL^2}{12} \qquad (2.24)$$

Por tanto, para n grande, utilizando el teorema central del límite, llegamos a la conclusión de que

$$Z = \frac{T_r - rL/2}{\sqrt{rL^2/12}} \tag{2.25}$$

sigue una distribución normal N(0,1).

2.4. Muestreo multinomial

Cuando un proceso de Poisson es observado hasta que han tenido lugar un número n fijo de sucesos, la variable aleatoria T_k(tiempo de espera hasta que aparecen k individuos de la especie I_k) es la suma de k variables aleatorias exponenciales, luego su distribución es gamma:

$$f_{T_k}(x) = \frac{e^{-\lambda_k x} x^{n_k-1}}{(n_k-1)\,!\,(1-e^{-\lambda_k})^{n_k}} \quad , 0<x<\infty \tag{2.26}$$

y, por consiguiente, $1/T_k$ es una variable aleatoria, cuya esperanza matemática es:

$$E\left(\frac{1}{T_k}\right) = \int_0^\infty \frac{1}{t} \frac{e^{-\frac{\lambda_k}{1-e^{-\lambda_k}}} p_k^{n_k} t^{n-1}}{(n-1)\,!}\, dt = \tag{2.27}$$

$$= \frac{\lambda_k}{(n-1)\,!\,(1-e^{-\lambda_k})} \Gamma(n-1) = \frac{\lambda_k}{(n_k-1)\,(1-e^{-\lambda_k})} \tag{2.28}$$

siendo la varianza:

$$V\left(\frac{1}{T_k}\right) = E\left(\frac{1}{T_k^2}\right) - \left[E\left(\frac{1}{T_k}\right)\right]^2 = \frac{t^2}{(n_k-1)\,(n_k-2)} - \frac{t^2}{(n_k-1)^2} \tag{2.29}$$

La proposición 2.3, ya conocida, que condiciona la superposión de procesos al tamaño muestral, nos conduce al modelo multinomial:

Proposición 2.3: Si las variables aleatorias $N_1, N_2, \ldots, N_S$ son independientes y tales que cada N_i se distribuye según una distribución de Poisson de parámetro λ_i, entonces la distribución de $(N_1, N_2, \ldots, N_S)$ condicionada por $T = \sum_{i=1}^{S} N_i$, es multinomial de parámetros

$$T, \pi_1 = \frac{\lambda_i}{\sum_{j=1}^{S} \lambda_j}, \ldots, \pi_S = \frac{\lambda_S}{\sum_{j=1}^{S} \lambda_j}. \qquad (2.30)$$

Luego:

"Cuando se condiciona la distribución al número de sucesos T que tienen lugar en el intervalo de tiempo (0,t], se recubre el muestreo multinomial".

2.4.1. Aproximación de la distribución de Poisson a la multinomial

Para determinar la distribución que va a regular el modelo, nos servimos de la siguiente proposición enunciada por Sidney C. Port[2]:

Proposición 2.4: Sea el vector aleatorio $(N_1, N_2, \ldots, N_S)$ que sigue una distribución multinomial $M_2(n, p_1(n), \ldots, p_S(n))$, siendo $p(n) = p_1(n) + \ldots + p_S(n)$, y sean N_i^*, $i = 1, 2, \ldots, S$ variables aleatorias con distribución de Poisson de parámetros $np_i(n)$, siendo los N_i^* independientes. Entonces, si $np_i(n)$ tiende a λ_i, se verifica que:

$$\lim_{n \to \infty} P[N_1 = x_1, \ldots, N_S = x_S] = \prod_{k=1}^{S} \frac{\lambda_k^{x_k} e^{-\lambda_k}}{x_k!} \qquad (2.31)$$

donde T=n es ahora el tamaño muestral.

[2] S.C. Port. "Theoretical Probability for Applications". Cap. 29. 29.7. John Wiley & Sons. Nueva York-1993.

La demostración requiere los cuatro siguientes pasos:

I. $$\left(1-\frac{x}{n}\right)\prod_{i=1}^{s}\frac{[np_i(n)]^{x_i}}{x_i!}(1-p(n))^n \le P[N_1=x_1,\ldots,N_s=x_s] \le$$

$$\le [1-p(n)]^{n-x}\prod_{i=1}^{s}\frac{[np_i(n)]^{x_i}}{x_i!}$$

donde

$$x=x_1+\ldots+x_s,\ x_i\ge 0,\ entero.$$

Esto es así por ser

$$P[N_1=x_1,\ldots,N_s=x_s]=\frac{n!}{(n-x)!}\prod_{k=1}^{s}\frac{[np_k(n)]^{x_k}}{x_k!}[1-p(n)]^{n-x}$$

donde $np_k(n)=n\dfrac{\overline{\lambda_k}}{\sum_{j=1}^{s}\lambda_j}=p_kT.$

II. Si 0<t<1, entonces $e^{-\frac{t}{(1-t)}}\le 1-t\le e^{-t}$

Utilizando los dos resultados anteriores, obtenemos:

$$\left(1-\frac{x}{n}\right)e^{-np(n)^2[1-p(n)]^{-1}}\le\frac{P(N_1=x_1,\ldots,N_s=x_s)}{P(\overline{N_1^*=x_1,\ldots,N_s^*=x_s})}\le(1-p(n))^{-x}$$

III. Si $np_i(n)^2$ tiende a cero, para cada i, se verifica que

$$\lim_{n\to\infty}\left(1-\frac{x}{n}\right)e^{-np(n)^2[1-p(n)]^{-1}}=1$$

y

$$\lim_{T\to\infty}(1-p(n))^{-x}=1.$$

Por lo tanto:

$$\lim_{T\to\infty} \frac{P[N_1 = x_1, \ldots, N_S = x_S]}{P[N_1^* = x_1, \ldots, N_S^* = x_S]} = 1$$

IV. Entonces, como además sucede que $np_i(n) = n\frac{\lambda_i}{\sum_{k=1}^{S}\lambda_i} \to \lambda_i$, se verifica:

$$\lim_{T\to\infty} P[N_1 = x_1, \ldots, N_S = x_S] = \prod_{k=1}^{S} \frac{\lambda_k^{x_k} e^{-\lambda_k}}{x_k!}$$

Como corolario de esta proposición resulta la siguiente proposición fundamental en el desarrollo del modelo de muestreo multinomial que planteamos, puesto que nos dice la distribución que va a regularlo.

Proposición 2.5: La distribución que regula el modelo multinomial es la distribución de Poisson de parámetro p_kT, siendo ahora T el tamaño muestral.

En efecto:

Ya vimos que $(N_1(t), N_2(t), \ldots, N_S(t))$ sigue una distribución multinomial de parámetros $(T, p_1(T), \ldots, p_S(T))$, siendo

$$p_i(T) = \frac{\lambda_i}{\sum_{j=1}^{D}\lambda_j} = \frac{\lambda_i}{T}$$

de tal modo que $\lambda_i = T\,p_i(T) = p_i T$.

Se obtiene así la distribución:

$$\lim_{n\to\infty} P[N_1(T)=x_1,\ldots,N_D(T)=x_D]=\prod_{k=1}^{D}\frac{(p_kT)^{x_k}e^{-p_kT}}{x_k!} \qquad (2.32)$$

En este modelo, el tamaño muestral T es constante, ya que hemos condicionado el proceso a la suma de los $N_k(t)$, que es T.

2.4.2. Probabilidades de especies desconocidas en el muestreo multinomial

La distribución de $N_k^*(t)$ condicionada por $\sum_{k=1}^{D} N_k^*(t)=T$, que corresponde al muestreo con reemplazamiento, es:

$$P_{k,r}^*=P[N_k(T)=r/N_k(T)>0]=\frac{(p_kT)^r e^{-p_kT}}{r!(1-e^{-p_kT})} \qquad (2.33)$$

Como $P_0=e^{-p_kT}$, resulta:

$$SP_0=\sum_{k=1}^{S}e^{-p_kT}\Rightarrow P_0=\frac{1}{S}\sum_{k=1}^{S}e^{-p_kT}$$

Luego:

$$P_0=\frac{1}{S}\sum_{k=1}^{S}e^{-p_kT} \qquad (2.34)$$

Para estimar P_0, buscamos un estimador de C_0, o del recubrimiento muestral $1-C_0$. Goods[1] fue el primero en proponer, a sugerencias de Touring, el siguiente estimador del recubrimiento muestral:

$$\hat{C} = 1 - \frac{\hat{M}_1}{T}$$

$\hat{C}_0 = 1-\hat{C} = \hat{M}_1/T$ es un estimador de las probabilidades de las especies desconocidas (que no aparecen en la muestra), según vamos a ver. Luego $\hat{C}_0$ permite estimar la probabilidad del número de especies que no aparecen en la muestra a partir del número de especies que aparecen una sola vez en la muestra y del tamaño de ésta. $\hat{C}_0 = \hat{M}_1/T$ es estimador de C_0, pero debemos tener en cuenta que C_0 no es un parámetro en el sentido habitual, sino en el sentido en que lo define Lo[2].

De acuerdo con esta definición, Robbins, realizando una prueba adicional en el mismo experimento, encuentra un nuevo estimador insesgado de la probabilidad de especies desconocidas, propiedad que no cumple el estimador de Good, y demuestra la siguiente proposición:

Proposición 2.7: $\hat{C}_0 = \hat{M}_1/T+1$ es un estimador insesgado de C_0.

Robbins afirma que "W es un *buen estimador* de V porque, siendo

U=V-W, es E[U]=0 y $E[U^2]<1/(T+1)$.

[1] Good, I.J. "The number of new species and the increase in Population Coverage, when a Sample is Increased" Biometrica, Vol. 40, 1953 y Good, I.J y Toulmin". Vol. 46, 1956, antes citadas y Toulmin GH. The Biometrica, 43, pàg. 45-63

[2] Robbins, H. (1968) "Estimating the Total Probability of the Unobserved Outcomes of an Experiment". The Annals of Mathematical Statistics, Vol. 39, Pág. 256-257. Lo, S. (1992) "From the Species Problem to a General Coverage Problem Via a New Interpretation". The Annals of Statistics, 20, 1094-1109.

2.4.3. Propiedades de los números de ocupación

1. $$\boxed{E(M_r) = \sum_{k=1}^{S} (p_{k,r}T)^r \frac{e^{-p_k T}}{r!}} \qquad (2.35)$$

En efecto, teniendo en cuenta la definición de M_r y las propiedades de la esperanza matemática, se obtiene:

$$E[M_r] = \sum_{k=1}^{S} E[N_k(T) = r] = \sum_{k=1}^{S} \frac{(p_k T)^r}{r!} e^{-p_k T}$$

Luego tomamos a M_r como estimador de $E[M_r]$, lo que expresamos mediante $\hat{M}_r$.

2. $$\boxed{\hat{C}_r = \frac{r+1}{T} \hat{M}_{r+1}} \qquad (2.36)$$

En efecto:

$$E(C_r) = \sum_{k=1}^{S} p_k E[I[N_k(T) = r] = \frac{r+1}{T} \sum_{k=1}^{S} \frac{(p_k T)^{r+1}}{(r+1)!} e^{-p_k t} = \frac{r+1}{T} \hat{M}_{r+1}$$

3. En particular:

$$E(C_0) = \frac{\hat{M}_1}{T} \quad , \quad E(C_1) = \frac{2\hat{M}_2}{T} \quad , \quad E(C_2) = \frac{3\hat{M}_3}{T}$$

4. $$\boxed{\hat{M}_r = \frac{T}{r} \hat{C}_{r-1}} \qquad (2.37)$$

Se deduce inmediatamente de (2.36).

5 $$\boxed{E[C] = 1 - \frac{\hat{M}_1}{T}} \qquad (2.38)$$

Basta con tener en cuenta que $C=1-C_0$.

6. $$\boxed{\sum_{k=1}^{S} p_k^r = \frac{1}{T^{(r)}} \sum_{k=r}^{S} k^{(r)} E(M_r)} \qquad (2.39)$$

donde

$T^{(r)} = T(T-1)(T-2)\ldots(T-(r-1))$ y $k^{(r)} = k(k-1)(k-2)\ldots(k-(r-1))$, $r=1,2,3,..$

Esta expresión fue ya encontrada por Good[3].

Luego, como estimador de $\sum_{i=1}^{S} p_i^k$ tenemos:

7. $$\boxed{\hat{H}_k = \frac{1}{T^{(k)}} \sum_{j=k} j^{(k)} E(M_j)} \qquad (2.40)$$

y, en particular:

$$\hat{H}_2 = \frac{1}{T(T-1)} \sum_{k=2} k(k-1) E(M_k) \qquad (2.41)$$

Si designamos por $\hat{R}_k = \sum_{j=k+2} j^{(k)} E(M_k)$, podemos expresar

$$\boxed{\hat{H}_2 = \frac{\hat{R}_2}{T(T-1)}} \qquad (2.42)$$

[3] Good, I.J. "The Population Frequenties of Species and the Estimation of Population Parameters". Biometrica, Vol. 40, Pág. 237-264. 1953, y Good, I.J. y Toulmin, G.H. "The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased". Biometrica, Vol. 43, Pág. 45-63, 1956.

9. $$\boxed{E[D] = S - \sum_{k=1}^{S} e^{-p_k T} = S\left[1 - \frac{1}{S}\sum_{k=1}^{S} e^{-p_k T}\right]} \qquad (2.43)$$

En efecto:

$$E[D] = E\left[\sum_{k=1}^{T} M_r\right] = \sum_{r=1}^{T}\sum_{k=1}^{S} \frac{(p_k T)^r}{r!} e^{-p_k T} = \sum_{k=1}^{S} e^{-p_k T}\left[\sum_{r=1}^{T} \frac{(p_k T)^r}{r!}\right] =$$

$$= \sum_{k=1}^{S} e^{-p_k T}\left[\sum_{r=0}^{T} \frac{(p_k T)^r}{r!} - 1\right] = \sum_{k=1}^{S} e^{-p_k T} e^{p_k T} - \sum_{k=1}^{S} e^{-p_k T} = S - \sum_{k=1}^{S} e^{-p_k T} \ , \ c.q.d.$$

11. De (2.43) se deduce inmediatamente que $D/(1-P_0)$ es un estimador insesgado de S:

$$\boxed{E\left[\frac{D}{1-P_0}\right] = S} \qquad (2.44)$$

2.4.4. Función generadora de momentos de M_r

La función generadora de momentos de los M_r viene dada por:

$$\phi_{M_r}(u) = E\left[e^{uM_r}\right] = E\left[e^{u\sum_{k=1}^{S} I[N_k(t)=r]}\right] = \prod_{k=1}^{S}\left[e^{u} P_{k,r} + 1 - P_{k,r}\right] =$$

$$\phi_{M_r}(u) = E\left[e^{uM_r}\right] = E\left[e^{u\sum_{k=1}^{S} I[N_k(t)=r]}\right] = \prod_{k=1}^{S}\left[e^{u} P_{k,r} + 1 - P_{k,r}\right] =$$

$$= \prod_{k=1}^{S} \left[1+(e^{u}-1)P_{k,r}\right] = \left[1+(e^{u}-1)\,P_{k,r}\right]^{S}$$

Luego la función generadora de momentos de M_r es

$$\phi_{M_r}(u) = \left[1+(e^{u}-1)\,P_{k,r}\right]^{S} \qquad (2.45)$$

Se trata de una distribución binomial de parámetros $B(S,P_r)$. Tomando esperanzas, resulta:

$$E[\hat{M}_r] = SP_r$$

de donde podemos estimar P_r mediante:

$$\hat{P}_r = \frac{\hat{M}_r}{S}$$

Podemos enunciar, por tanto, la siguiente proposición:

Proposición 2.8: Los números de ocupación, M_r, en este modelo, tienen una distribución binomial de parámetros: (S,P_r).

Se verifican las siguientes propiedades:

I. $$E[M_r] = P_r S = \hat{M}_r \qquad (2.46)$$

II. $$E\left[M_r(M_r-1)\right] = P_r^2 S(S-1) \qquad (2.47)$$

III. $$Var\,[M_r] = \hat{M}_r\left(1-\frac{\hat{M}_r}{S}\right) \qquad (2.48)$$

Por ser la distribución de los M_r binomial, se verifica la siguiente proposición:

Proposición 2.9: La distribución del vector de posición $(M_1,\ldots,M_T)$ tiene una distribución multinomial $M_1(T,P_1,\ldots,P_T)$. Estimando los P_r mediante M_r/S, la distribución de $(M_1,\ldots,M_T)$ es aproximadamente multinomial de parámetros $M\left(T,\frac{\hat{M}_1}{S},\ldots,\frac{\hat{M}_T}{S}\right)$.

2.4.5. Covarianzas

Veamos cuál es la función generadora de momentos de la variable aleatoria bidimensional (M_h,M_k).

$$\phi_{M_h+M_k}(u,v) = \left[e^{uM_h+vM_k}\right] = E\left[e^{uI[N_j(T)=h]+vI[N_j(T)=k]}\right] =$$

al ser las variables intercambiables

$$= \prod_{h=1}^{S}\prod_{k=1}^{S}\left[(e^u-1)P_h^*(1-P_k^*)+(e^v-1)P_k^*(1-P_k^*)+(e^u+e^v-1)P_h^*P_k^*-(e^{u+v}-1)P_h^*P_k^*+1\right]$$

Tomando logaritmos en los dos miembros, resulta:

$$\ln\phi_{M_h+M_k}(u,v) = (e^u-1)\hat{M}_h+(e^v-1)\hat{M}_k-(e^{u+v}-1)\frac{\hat{M}_h\hat{M}_k}{S}$$

Luego la función generatriz de momentos es:

$$\boxed{\phi_{M_h+M_k}(u,v) = e^{\left\{(e^u-1)\hat{M}_h+(e^v-1)\hat{M}_k-(e^u+e^v-1)\frac{\hat{M}_h\hat{M}_k}{S}\right\}}} \qquad (2.49)$$

que es la función generadora de momentos de una distribución bivariante de Poisson, que verifica, por tanto:

I. $$Cov(M_h,M_k) = -\frac{\hat{M}_h\hat{M}_k}{S} \qquad (2.50)$$

II. $$Var[M_h] = \hat{M}_h\left(1-\frac{\hat{M}_h}{\hat{S}}\right) \qquad (2.51)$$

2.5. Coeficente de Variación de las p_k

En las aplicaciones prácticas rara vez se da la hipótesis de homogeneidad, que consiste en admitir que "*todas las clases tienen el mismo tamaño y todos los individuos de la población la misma probabilidad de ser seleccionados*", lo que representamos por

$$H_= : p_1 = p_2 = \ldots p_S = \frac{1}{S}$$

Para encontrar un estimador para una situación no homogénea, es necesario tener en cuenta la variación entre las probabilidades de las clases. El coeficiente de variación de Pearson es una buena medida de la heterogeneidad de la distribución de las p_k.

Se define el coeficiente de variación de Pearson de las p_k como:

$$\hat{\gamma} = S\sqrt{\frac{1}{S}\sum_{k=1}^{S}\left(p_k-\frac{1}{S}\right)^2} \qquad (2.52)$$

El cuadrado del coeficiente de variación de Pearson es, por tanto, igual a S por la varianza de las p_k, y, por tanto, igual al producto de S por la suma de los cuadrados de las p_k menos 1:

$$\hat{\gamma}^2+1 = S\sum_{i=1}^{S} p_i^2 \qquad (2.53)$$

Como $H_2=R_2/(T(T-1))$ es un estimador insesgado de $\sum_{i=1}^{S} p_i^2$,

teniendo en cuenta la relación anterior, se obtiene:

$$\hat{\gamma}^2+1 = \frac{S\hat{R}_2}{T(T-1)} \qquad (2.54)$$

de donde resulta, como estimador[4] del cuadrado del coeficiente de variación de Pearson:

$$\hat{\gamma}^2 = máx\left\{\frac{\hat{S}}{T(T-1)}\sum_{k=1}^{S} k(k-1)\,E(M_k) -1, 0\right\} \qquad (2.55)$$

2.6. Estimadores de S

Vamos a considerar dos situaciones, según se admita o no la hipótesis de homogeneidad:

$$H_=: p_1=p_2=\ldots p_S=1/S$$

2.6.1. Situación homogénea. Muestreo multinomial

Cuando las abundancias relativas son iguales, lo que expresaremos diciendo que se cumple la hipótesis $H_=$ de "homogeneidad", es decir, cuando se verifican las igualdades $p_1=p_2=\ldots=p_S=1/S$, el problema ha sido ampliamente tratado y resuelto satisfactoriamente en algunos casos.

[4] Chao, A. y Lee,S. "Estmating the Number of Classes via Sample Coverage". Journal of the Americal Statistical Association. Vol. 87, Nº 417, Pág. 210-217. Marzo de 1992.

Podemos citar, entre otros muchos, a Lewontin y Prout(1956), Darroch(1958), Harris(1968) y Holst(1981),...

Dos estimadores son los que destacan por su importancia: el estimador de Darroch y el estimador de máxima verosimilitud.

Proposición 2.9: "Bajo la hipótesis de homogeneidad, los momentos centrales de las p_k son todos nulos.

En efecto:

$$m_k = \sum_{k=1}^{S} \left(p_k - \frac{1}{S}\right)^k = 0, \quad si \; p_1 = p_2 = \ldots = p_S = \frac{1}{S}$$

Proposición 2.10: Bajo la hipótesis de homogeneidad, se verifican las siguientes relaciones:

$$A) \; \frac{1}{S}\sum_{k=1}^{S} e^{-p_k T} = e^{-\frac{T}{S}} \; ; \quad B) \; \frac{\hat{M}_1}{T} = e^{-\frac{T}{S}}$$

Demostración:

La relación A) es inmediata. En cuanto a B), basta con tener en cuenta que

$$E[M_1] = \sum_{k=1}^{S} \frac{T}{S} e^{-\frac{T}{S}} = T e^{-\frac{T}{S}} \Rightarrow \frac{\hat{M}_1}{T} = e^{-\frac{T}{S}}$$

Proposición 2.11: La condición necesaria y suficiente para que se verifique la hipótesis de homogeneidad es que sea nulo el coeficiente de variación de Pearson de las p_k.

Demostración: Basta con tener en cuenta la proposición 2.5 y la definición del coeficiente de variación.

2.6.1.1. Estimador de máxima verosimilitud

El estimador conocido como de máxima verosimilitud es la solución $\hat{S}_1$ de la ecuación:

$$\boxed{D = S\left[1-e^{-\frac{T}{S}}\right]} \qquad (2.56)$$

Se trata del estimador de máxima verosimilitud para la distribución de particiones de Maxwell-Boltzman, que es la distribución para particiones en el modelo multinomial cuando se admite la hipótesis de homogeneidad, cuyo estudio realizamos a continuación.

2.6.1.2. Distribución de Maxwell-Boltzman

La distribución de $\{N_k(t),\ k=1,2,\ldots,D\}$ se obtiene a partir de la distribución de los $N_k(t)$ condicionada por $\sum_{k=1}^{S} N_k(t) = T$, luego

$$P[N_k(t)=n_k] = P[N_k(t)=n_k, k=1,2,\ldots,D \,/\, \sum_{k=1}^{S} N_k(t) = T] =$$

$$= \frac{\prod_{k=1}^{S} (p_k t)^{n_k} \frac{e^{-p_k t}}{n_k!} (p_k t)^{T-n_k} \frac{e^{-p_k t}}{(T-n_k)!}}{(p_k t)^{T} \frac{e^{-p_k t}}{T!}} =$$

$$= \frac{D!}{\prod_{k=1}^{S} n_k!} \prod_{k=1}^{S} p_k^{n_k}$$

Si se admite la hipótesis de homogeneidad: $p_1=p_2=\ldots=p_S=1/S$, se obtiene:

$$= \frac{D!}{\prod_{k=1}^{D} n_k!} \left(\frac{1}{S}\right)^{T}, \; n_k \neq 0$$

Esta distribución es la probabilidad de obtener una muestra de una población en la que hay una partición formada por S clases de modo que en dicha muestra hay n_k objetos de la población. Ahora bien, sólo D, de entre las S clases; estarán representadas en la muestra, luego habrá otras S-D clases no representadas. No todas las muestras se distinguen unas de otras.

No se sabe qué conjunto de D clases de las S que componen la población se encuentran en la muestra, siendo el número de muestras con D clases diferentes:

$$\frac{S!}{D!\,(S-D)!}$$

Además sucede que no se sabe cuál de entre las D clases seleccionadas está representada n_k veces. El número de formas en que D clases pueden constituir una partición de modo que M_k clases aparezcan k veces es:

$$\frac{D!}{\prod M_k!}$$

Según estos razonamientos, la probabilidad de obtener una muestra distinguible es:

$$P = \left[\frac{T!}{\prod_{k=1}^{D} n_k! \prod_{k=1}^{D} M_k!}\right]\left[\frac{S!}{(S-D)!\, S^T}\right] \qquad (2.57)$$

Hemos obtenido así la función de verosimilitud. Se trata de la distribución correspondiente al muestreo multinomial bajo la hipótesis de homogeneidad (distribución de Maxwell-Boltzman), cuyo estudio realizaron Lewontin[5] y Prout en 1956.

Veamos que la solución S_1 de la ecuación

$$D = S\left[1 - e^{-\frac{T}{S}}\right]$$

es un estimador suficente y de máxima verosimilitud para S.

Que es suficiente resulta evidente si observamos la expresión (2.57) de la función de verosimilitud, que aparece descompuesta en producto de dos factores: el factor de la derecha depende de S y de D, y el factor de la izquierda, que depende solamente de las observaciones.

Veamos que se trata de un estimador de máxima verosimilitud, siguiendo un razonamiento diferente al de Lewntin y Prout.

Tomando logaritmos, resulta la función:

$$\ln P = \ln\Gamma(S+1) + \ln\Gamma(T+1) - \ln\Gamma(S-D+1) -$$

$$-T\ln S - \sum_{k=1}^{D} \ln\Gamma(M_k+1) - \sum_{k=1}^{D} \ln\Gamma(n_k+1) \qquad (2.58)$$

Los últimos sumandos no dependen de S, por lo que derivando con respecto a S, se obtiene:

$$\partial\frac{\ln(P)}{\partial S} = \frac{\Gamma'(S+1)}{\Gamma(S+1)} - \frac{\Gamma'(S-D+1)}{\Gamma(S-D+1)} - \frac{T}{S}$$

[5] Lewontin, R.C. and Prout, T. "Estimation of the Number of Diferent Classes in a Population". Biometrics, 12, pág. 211-223. Junio de 1956.

Igualando a cero, y, teniendo en cuenta que

$$\frac{\Gamma'(x+1)}{\Gamma(x+1)} = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{x}-c$$

siendo c la constante de Euler-Mascheroni, resulta:

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{S}-c-1-\frac{1}{2}-\cdots-\frac{1}{S-D}+c-\frac{T}{S}=0$$

de donde se deduce:

$$\frac{1}{S-D+1}+\frac{1}{S-D+2}+\cdots+\frac{1}{S}=\sum_{j=1}^{D}\frac{1}{S-D+j}=-\frac{T}{S}$$

Teniendo en cuenta la relación[6]:

$$\sum_{j=1}^{D}\frac{1}{S-D+j} \approx \int_{S-D+\frac{1}{2}}^{S+\frac{1}{2}} x^{-1}dx = \ln\frac{S+\frac{1}{2}}{S-D+\frac{1}{2}}$$

resulta:

$$\ln\frac{\hat{S}+\frac{1}{2}}{\hat{S}-\hat{D}+\frac{1}{2}}=\frac{T}{\hat{S}}$$

Como

$$\frac{\hat{S}+\frac{1}{2}}{\hat{S}-\hat{D}+\frac{1}{2}} \simeq \frac{\hat{S}}{\hat{S}-\hat{D}}$$

podemos poner

[6] Feller,W. "Introducción a la Teoría de Probabilidades y sus Aplicaciones", Volumen I, Capítulo IX, apartado 3. Méjico, 1993.

$$\frac{\hat{S}}{\hat{S}-\hat{D}} = e^{\frac{T}{\hat{S}}} \Rightarrow \frac{\hat{S}-\hat{D}}{\hat{S}} = e^{-\frac{T}{\hat{S}}} \Rightarrow 1-\frac{\hat{D}}{\hat{S}} = e^{-\frac{T}{\hat{S}}}$$

de donde finalmente:

$$\boxed{\hat{D} = \hat{S}_3\left[1-e^{-\frac{T}{\hat{S}_3}}\right]} \qquad (2.59)$$

cuya solución es el estimador de máxima verosimilitud para, al que vamos a dentar por S_3.

2.6.1.3. Varianza del estimador de máxima verosimilitud

La derivada primera del logaritmo neperiano de la función de verosimilitud queda en la forma:

$$\frac{\partial \ln P}{\partial \hat{S}} = \ln\hat{S} - \ln(\hat{S}-\hat{D}) - \frac{T}{\hat{S}}$$

puesto que satisface las condiciones de Cramer-Rao.

La varianza del estimador viene dada, en general, por

$$Var(\hat{S}) = \frac{-1}{E\left(\frac{\partial^2 \ln P}{\partial \hat{S}^2}\right)}$$

Ahora bien: $$\frac{\partial^2 \ln P}{\partial \hat{S}^2} = \frac{1}{\hat{S}} - \frac{1}{\hat{S}-\hat{D}} + \frac{T}{\hat{S}^2}$$

Entonces

$$E\left[\frac{\partial^2 \ln P}{\partial \hat{S}^2}\right] = \frac{1}{\hat{S}} + \frac{T}{\hat{S}^2} - E\left[\frac{1}{\hat{S}-\hat{D}}\right] = \frac{1}{\hat{S}} + \frac{T}{\hat{S}^2} - E\left[\hat{S}e^{-\frac{T}{\hat{S}}}\right] = \frac{1}{\hat{S}} + \frac{T}{\hat{S}^2} - \frac{1}{\hat{S}e^{-\frac{T}{\hat{S}}}}$$

Luego

$$Var(\hat{S}_3) = \frac{-1}{\frac{1}{\hat{S}} + \frac{T}{\hat{S}^2} + \frac{1}{\hat{S}} e^{-\frac{T}{\hat{S}}}} = \frac{\hat{S}}{e^{\frac{T}{\hat{S}}} - \frac{T}{\hat{S}} - 1}$$

resultando la expresión que dieron Lewontin y Prout:

$$\boxed{var(\hat{S}_3) = \frac{\hat{S}_3}{e^{\frac{T}{\hat{S}_3}} - \frac{T}{\hat{S}_3} - 1}} \qquad (2.60)$$

2.6.1.4. Estimador de Darroch

Si utlizamos conjuntamente los resultados A) y B) de la proposición 2.8, obtenemos: .

$$E(D) = \hat{S} - \sum_{k=1} e^{-p_k T} = \hat{S} - \frac{\hat{S}\hat{M}_1}{T} = \hat{S}\left(1 - \frac{\hat{M}_1}{T}\right) = \hat{S}\hat{C}$$

de donde, si denotamos a este nuevo estimador por S_4, es:

$$\boxed{\hat{S}_4 = \frac{\hat{D}}{\hat{C}}}$$

que puede ponerse en la forma

$$\boxed{\hat{S}_4 = \frac{T\hat{D}}{T - \hat{M}_1}} \qquad (2.61)$$

Se trata del estimador de Darroch, que está basado en el estimador del recubrimiento muestral de Good-Touring.

Más adelante veremos una estimación de la varianza, que se obtiene fácilmente al estar formado el estimador por combinaciones lineales de los números de ocupación.

W.W. Esty[7] estudió la eficiencia de los estimadores de Darroch y de máxima verosimilitud, analizando los estimadores de los respectivos recubrimientos muestrales:

$$\bar{C}=1-e^{-\frac{T}{S}} \quad y \quad \ddot{C}=1-\frac{\hat{M}_1}{T}$$

concluyendo que el estimador de Good-Touring es "*bastante eficiente cuando se le compara con el de máxima verosimilitud*".

2.7. Tiempo de espera hasta la primera renovación

Aún cuando la superposición de procesos no fuera un proceso de renovación, se puede obtener la distribución del tiempo de espera hasta la primera etapa de renovacón después de la etapa cero siguiendo el razonamiento que hace Feller[8]:

En efecto, la función de valor medio de cada proceso es:

$$m_k(t) = E[N_k(t)] = p_k T \sim \frac{t}{\mu_X}$$

luego

$$\frac{1}{\mu_X} = p_k \frac{T}{t} \Rightarrow \frac{1}{\mu_1} + \ldots + \frac{1}{\mu_D} = \frac{T}{t} \sum_{k=1}^{S} p_k = \frac{T}{t} = \frac{1}{\alpha} \qquad (2.62)$$

de donde el tiempo de espera W en el proceso cumulativo es el más pequeño de los tiempos de espera W_k, por lo que

[7] Esty, W. W. "The Annals of Statistics", Vol. 14, N° 3, 1257-1260; 1966.

[8] Feller, W. "Introducción a la Teoría de Probabilidades y sus aplicaciones". Volumen II. LIMUSA, México 1989.

$$P[W>t] = 1-P[W\leq t] = 1-(1-e^{-T}) = e^{-T} \qquad (2.63)$$

Entonces

$$P[W\leq t] = 1-e^{-T} \qquad (2.64)$$

Además, en el proceso cumulativo, se verifica:

$$E[W] = \frac{1}{\alpha} = \frac{T}{t} \Rightarrow t = T\alpha \qquad (2.65)$$

Como $P[W>t]=(P_0)^S$, se verifica que

$$\boxed{\hat{P}_0 = e^{-\frac{T}{S}}} \qquad (2.66)$$

2.8. Proceso combinado

Teniendo en cuenta la caracterización de los procesos de Poisson con relación a la superposición de procesos que hace Samuels[10], a saber: "*La superposición de dos procesos de renovación independientes es un proceso de renovación si, y sólo si los procesos componentes son procesos de Poisson*", el proceso combinado es un proceso de renovación.

Además, el proceso combinado, N(t), *puede aproximarse*[11] *por la suma de un gran número de variables independientes indicador, de modo que la distribución de N(t) tiende a una distribución de Poisson, cuya función generatriz de momentos viene dada por:*

[10] Samuels,S.M. "A Characterization of the Poisson Processes". J. Appl. Prob. 11, 72-85. 1974.

[11] Cox,D.R. & Isham,V. "Point Processes", capítulo 4, apartado 4.5. Chapman & Hall, Londres 1992.

$$\boxed{\psi_{N(t)}(z) = E(z^{N(t)}) = \prod_{k=1}^{S} \left\{1 - \frac{W}{S}(1-z)\right\}} \qquad (2.67)$$

siendo

$$\frac{W}{S} = P\left[N_k^*(t) = 1\right], \ \forall k. \qquad (2.68)$$

de modo que, cuando S tiende a infinito, la expresión anterior tiende a la función generadora de probabilidades de una variable de Poisson de parámetro w.

Por lo tanto, en nuestro caso será:

$$\psi_{N(t)}(z) = E(z^{N(t)}) = \prod_{k=1}^{S} \left\{1 - p_k T e^{-p_k T}(1-z)\right\} \qquad (2.69)$$

Tomando logaritmos:

$$\ln[\phi_{N(t)}(z)] = \sum_{k=1}^{S} \ln\left\{1 + p_k T e^{-p_k T}(z-1)\right\} =$$

$$= \sum_{k=1}^{S} \left\{p_k T e^{-p_k T}\right\}(z-1) = T\sum_{k=1}^{S} p_k e^{-p_k T}(z-1) =$$

$$= T\sum_{k=1}^{S} p_k e^{-p_k T}(z-1) = T\hat{C}_0(z-1) = \hat{M}_1(z-1)$$

Luego la función generadora de probabilidades del proceso combinado es:

$$\boxed{\phi_{N(T)}(z) = e^{-\hat{M}_1(1-z)} = e^{-\hat{C}_0 T(1-z)}} \qquad (2.70)$$

que es la función generatriz de probabilidades de un proceso de Poisson de razón $M_1 = C_0 T$.

Entonces, como

$$\boxed{m(t) = \hat{C}_0 T = \frac{t}{\mu_X}} \qquad (2.71)$$

los tiempos entre llegadas, X, correspondientes al proceso combinado siguen una distribución exponencial de media $1/C_0=T/M_1$, y los tiempos de espera hasta que tienen lugar r sucesos siguen una distribución gamma de parámetros $(r,T/M_1)$, de modo que $E[W_1]=T/M_1$, donde W_1 es el tiempo trascurrido hasta obtener 1 especie diferente.

Como $N(t)=M_1$, los M_1-1 instantes $T_{(1)},\ldots,T_{(M1)}$-1 del intervalo (0,t] en los que ocurren los sucesos son variables aleatorias que tienen la misma distribución que si fueran los parámetros correspondientes a M_1 variables independientes $U_1,\ldots U_{M1}$, distribuídas uniformemente en el intervalo (0,t].

Se dice que $T_{(1)},\ldots,T_{(M1)}$ son los parámetros correspondientes a $U_1,\ldots U_{M1}$ si $T_{(1)}$ es el menor de los valores entre $U_1,\ldots U_{M1}$. $T_{(2)}$ es el segundo valor más pequeño, y así sucesivamente, siendo $T_{(M1)}$ el valor mayor entre $U_1,\ldots U_{M1}$. Los M_1 instantes $T_{(1)},\ldots,T_{(M1)}$ del intervalo $(0,M_1]$ en los que tienen lugar los sucesos, considerados como variables aleatorias no ordenadas, están distribuidos uniformemente y son independientes en el intervalo $(0,M_1]$.

La variable aleatoria $T_{(r)}$ y la variable aleatoria W_r, que representa el tiempo de espera hasta que tiene lugar el r-ésimo suceso, son dos notaciones diferentes para expresar el mismo concepto.

Entonces el estadístiico de orden h-ésimo de $T_{(1)},\ldots,T_{(M1)}$, y, por tanto, de $W_1,\ldots,W_{M1}$ es el estadístico de orden h-ésimo de $T_{(1)},\ldots,T_{(M1)}$, que son variables aleatorias distribuidas uniformemente en $(0,M_1]$. Luego el estadístico de orden $T_{(h)}$ sigue una distribución beta.

2.8.1. Comportamiento asintótico. Intervalo de confianza para S

Acabamos de ver que el proceso combinado N(t) se comporta como un proceso de Poisson homogéneo de media M_1, lo que nos dice que en el período (0,t] se producen M_1 renovaciones, es decir, aparecen M_1 especies diferentes.

Ahora bien, los procesos de punto renovados verifican las siguientes propiedades asintóticas[12] para valores grandes de t:

Proposición 2.12: Sea $N_k(t)$ un proceso de punto renovado, con tiempos entre llegadas X_i, independientes y todos con la misma distribución que una variable aleatoria T.

Una expresión asintótica para la media m(t)=E[N(t)], si $\mu = E[T] < \infty$, es:

$$\lim_{t\to\infty} \frac{m(t)}{t} = \frac{1}{\mu}. \qquad (2.72)$$

y una expresión asintótica para la varianza Var[N(t)],si son finitas $\mu = E[T]$ y $\sigma^2 = Var[T]$, es

$$\lim_{t\to\infty} \frac{Var[N(t)]}{t} = \frac{\sigma^2}{\mu^3}. \qquad (2.73)$$

Entonces la normalidad asintótica del proceso renovado, si $E[T^2] < \infty$, viene dada por

[12] Parzen,E. "Procesos Estocásticos". Capít. 5. apart. 5.3. Ed.Paraninfo. Madrid 1972.

$$\lim_{t\to\infty} P\left[\frac{N(t)-\frac{t}{\mu}}{\sqrt{t\frac{\sigma^2}{\mu^3}}} \le x\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}y^2}dy \qquad (2.74)$$

Luego N(t) está distribuido asintóticamente con media y varianzas asintóticas dadas por (2.72) y (2.73)[13].

Aplicando este resultado a nuestro proceso, como

$$\frac{t}{\mu} = \hat{M}_1 \quad y \quad \frac{t\sigma^2}{\mu^3} = \hat{M}_1$$

resulta:

$$P\left[\left|\frac{N(t)-\hat{M}_1}{\sqrt{\hat{M}_1}}\right| \le z_{\alpha/2}\right] = 1-\alpha \qquad (2.75)$$

De aquí, si tomamos $\alpha = 0'0445$, se obtiene el intervalo de confianza del 95'55% para N(t):

$$\left[\hat{M}_1 - 2\sqrt{\hat{M}_1} \le N(t) \le \hat{M}_1 + 2\sqrt{\hat{M}_1}\right] \qquad (2.76)$$

Teniendo en cuenta que $E[N(t)] = \hat{M}_1$, dividiendo por el tamaño muestral T, es:

$$\frac{\hat{M}_1 - 2\sqrt{\hat{M}_1}}{T} \le \hat{C}_0 \le \frac{\hat{M}_1 + 2\sqrt{\hat{M}_1}}{T}$$

Restando de 1, resulta:

[13] Parzen,E. "Procesos Estocásticos". Capítulo 4, 4A. Ed. Paraninfo. Madrid 1972.

$$\frac{T-\hat{M}_1-2\sqrt{\hat{M}_1}}{T} \leq \hat{C} \leq \frac{T-\hat{M}_1+2\sqrt{\hat{M}_1}}{T}$$

Invirtiendo y multiplicando por D, se obtiene finalmente:

$$\boxed{\frac{\hat{D}}{\frac{T-\hat{M}_1+2\sqrt{\hat{M}_1}}{T}} \leq \frac{\hat{D}}{\hat{C}} \leq \frac{\hat{D}}{\frac{T-\hat{M}_1-2\sqrt{\hat{M}_1}}{T}}} \qquad (2.77)$$

que es un intervalo de confianza del 95'55% para S, bajo la hipótesis de equiprobabilidad.

2.9. Muestreo multinomial. Situación no homogénea

La situación homogénea ha merecido más atención. Sin embargo, en la práctica, lo más probable es que nos encontremos con situaciones no homogéneas, siendo las clases no equiprobables.

Para resolver el problema en una situación no homogénea, es conveniente tener en cuenta la variación entre las probabilidades de las diferentes clases, cuyo estudio nos lo van a proporcionar los momentos centrales de las abundancias relativas. Necesitamos, por tanto, conocer la distribución de las p_k.

En principio, en este modelo, las abundancias relativas son constantes.

Se puede generalizar el modelo, haciendo que las abundancias relativas sean variables aleatorias, lo que haremos en el proximo capítulo, donde supondremos que las λ_k siguen una distribución gamma.

2.9.1. Estimador de Chao para situación no homogénea

Se trata de obtener el estimador de S en función del coeficiente de variación de Pearson de las p_k.

Vamos a considerar el estimador del recubrimiento muestral $\hat{C} = \frac{\hat{M}_1}{T}$. Partiendo ahora de la expresión

$$\boxed{\frac{\hat{D}}{\hat{C}} = S + \frac{S\left[1 - \frac{1}{S}\sum_{k=1}^{S} e^{-p_k T}\right] - S\hat{C}}{\hat{C}}} \qquad (2.78)$$

que se transforma en:

$$\frac{\hat{D}}{\hat{C}} = S + \frac{S(1-\hat{C}) - \sum_{k=1}^{S} e^{-p_k T}}{\hat{C}}$$

al ser $1-\hat{C} = \frac{\hat{M}_1}{T}$, resulta:

$$\boxed{S = \frac{\hat{D}}{\hat{C}} - \frac{S\frac{\hat{M}_1}{T} - \sum_{k=1}^{S} e^{-p_k T}}{\hat{C}}} \qquad (2.79)$$

Teniendo en cuenta que

$$E\left[\frac{M_1}{T}\right] = \sum_{k=1}^{S} p_k e^{-p_k T}$$

y substituyendo este valor en (2.79), resulta:

$$\boxed{S = \frac{\hat{D}}{\hat{C}} - \frac{S\sum_{k=1}^{S} p_k e^{-p_k T} - \sum_{k=1}^{S} e^{-p_k T}}{\hat{C}} = \frac{\hat{D}}{\hat{C}} - \frac{S\sum_{k=1}^{S} e^{-p_k T}\left(p_k - \frac{1}{S}\right)}{\hat{C}}} \quad (2.80)$$

Desarrollando $e^{-p_k T}$ en serie de Taylor en un entorno simétrico de centro $p_k=1/S$, se obtiene:

$$e^{-p_k T} = e^{-\frac{T}{S}}\left[1 - T\left(p_k - \frac{1}{S}\right)\right] + \frac{T^2 e^{-\theta T}}{2}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)^2, \; 0<\theta<1$$

y, por tanto:

$$S\sum_{k=1}^{S} e^{-p_k T}\left(p_k - \frac{1}{S}\right) =$$

$$= Se^{-\frac{T}{S}}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right) - STe^{-\frac{T}{S}}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)^2 + \frac{ST^2 e^{-\theta T}}{2}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)\left(P_k - \frac{1}{S}\right)^2 =$$

$$= -Te^{-\frac{T}{S}} S\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)^2 + \frac{ST^2 e^{-\theta T}}{2}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)^3 =$$

$$= -Te^{-\frac{T}{S}}\gamma^2 + \frac{T^2 e^{-\theta T}}{2S^2}\sum_{k=1}^{S}(1 - Sp_k)^3 = -\hat{M}_1\gamma^2 - \frac{T^2 e^{-\theta T}}{2S^2}\sum_{k=1}^{S}(1 - Sp_k)^3 =$$

$$= -\hat{M}_1\gamma^2 - \frac{T^2 e^{-\theta T}}{2S^2}\sum_{k=1}^{S}\left(p_k - \frac{1}{S}\right)^3, \; 0<\theta<1.$$

donde el resto

$$K = \frac{T^2 e^{-\theta T}}{2S^2} \sum_{k=1}^{S} (1-Sp_k)^3, \quad 0<\theta<1 \tag{2.81}$$

es función del momento de tercer orden de las p_k.

Acotemos el resto:

$$|K| = \frac{T^2 e^{-\theta T}}{2S^2} \sum_{k=1}^{S} |1-Sp_k|^3 \leq \frac{T^2}{2S^2} \sum_{k=1}^{S} |1-Sp|^3 \leq \frac{T^2}{2S^2} \sum_{k=1}^{S} e^{-3Sp_i}$$

La última desigualdad es cierta por verificarse la relación

$$1-x \leq e^{-x}, \quad \forall x.$$

Entonces:

$$|K| \leq \frac{T^2}{2S^2} Se^{-3} \leq \frac{T^2}{2Se^3} \approx \frac{T\ln(T/\hat{M}_1)}{2e^3}$$

y este último término nos proporciona una cota del error, al estimar S, suficientemente pequeña.

Como nuevo estimador de S, al que denotaremos por S_5, ahora para una situación no homogénea, hemos obtenido:

$$\hat{S}_5 = \frac{\hat{D}}{\hat{C}} + \frac{T(1-\hat{C})}{\hat{C}} \hat{\gamma}^2 \tag{2.82}$$

Se trata del estimador de Chao, que depende del coeficiente de variación de Pearson. Es necesario, para evaluarlo, tal como hace Chao, realizar una estimación previa de S utilizando un estimador homogéneo como el de Darroch, con lo cual obtenemos:

$$\hat{S}_5 = \frac{T\hat{D}}{T-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\left(\frac{\hat{D}\hat{R}_2}{(T-\hat{M}_1)(T-1)} - 1\right) \tag{2.83}$$

2.9.2. Estimador basado en la proporción de especies

Veamos un nuevo estimador para la situación no homogénea, que depende también del coeficiente de variación de Pearson, pero que, en lugar de estar basado en el recubrimiento muestral, se define a partir de la proporción de especies que aparecen repetidas en la muestra:(T-D)/D.

Partimos de la relación

$$1 - P_0 = 1 - \frac{1}{S}\sum_{k=1}^{S} e^{-p_k T} \qquad (2.84)$$

Como tratamos con procesos estocásticos, cuyo parámetro t es continuo, podemos desarrollar la suma en serie de Taylor en un entorno de t=0, y particularizando en t=T (condicionando al tamaño muestral T), puesto que se trata de una exponencial, que converge en toda la recta real, se obtiene:

$$1-\frac{1}{S}\sum_{k=1}^{S} e^{-p_k T} = 1-\frac{1}{S}\sum_{k=1}^{S}\left[1-p_k T+\frac{T^2}{2!}p_k^2-\frac{T^3}{3!}p_k^3 e^{-p_k T\theta}\right],\ 0<1<\theta \qquad (2.85)$$

por tanto

$$\frac{\hat{D}}{S} = \frac{T}{S} - \frac{T^2}{2S}\sum_{k=1}^{S} p_k^2 + \frac{T^3}{6S}\sum_{k=1}^{S} p_k^3 e^{-\theta p_k T}$$

donde

$$K = \frac{T^3}{6S}\sum_{k=1}^{S} p_k^3 e^{-\theta p_k T}\ ,\ 0<\theta<1.$$

Teniendo en cuenta que

$$\sum_{k=1}^{S} p_k^2 = \frac{\gamma^2+1}{S}$$

se obtiene:

$$\frac{\hat{D}}{S} = \frac{T}{S} - \frac{T^2}{2}\,\frac{\gamma^2+1}{S^2} \qquad (2.86)$$

de donde, si denotamos a este estimador por S_7, queda:

$$\hat{S}_7 = \frac{T^2}{2\,(T-\hat{D})}\,(\hat{\gamma}^2+1) \qquad (2.87)$$

que también puede expresarse en la forma:

$$\hat{S}_7 = \frac{\frac{T}{2}}{1-\frac{\hat{D}}{T}} + \frac{\frac{T}{2}}{1-\frac{\hat{D}}{T}}\,\hat{\gamma}^2 \qquad (2.88)$$

Si suponemos que se verifica la hipótesis de homogeneidad, es decir, que todas las p_k son iguales, será nulo el coeficiente de variación, y, por tanto, se obtiene el estimador homogéneo basado en la proporción de especies, D/T, que hay en la muestra:

$$\hat{S}_6 = \frac{\frac{T}{2}}{1-\frac{\hat{D}}{T}} \qquad (2.89)$$

Utilizando (2.88) para hacer una estimación previa de S, que nos permita estimar el coeficiente de variación de las p_k, resulta:

$$\hat{S}_7 = \frac{T^3 \hat{R}_2}{4\,(T-\hat{D})^2\,(T-1)} \qquad (2.90)$$

2.9.2.1. Acotación del resto

Partimos de

$$K = \frac{T^3}{6}\sum_{k=1}^{S} p_k^3 e^{-p_k T\theta} \quad , 0<\theta<1.$$

Entonces

$$K \le \frac{T^3}{6S}\sum_{k=1}^{S} p_k^3 \le \frac{T^3}{6S^3} \le \frac{T}{6S}\left(\frac{T}{S}\right)^2 \approx \frac{T}{6S}\left(\ln\frac{T}{\hat{M}_1}\right)^2$$

de donde, una cota superior del resto, al estimar S, es:

$$H = SK \le \frac{T}{6}\left(\ln\frac{T}{\hat{M}_1}\right)^2$$

2.10. Estimadores de menor sesgo para el recubrimiento muestral

Vamos a utilizar el método de eliminación de información de Lo para reducir el sesgo del estimador del recubrimiento muestral.

2.10.1. Nuevo estimador utilizando el método de Lo

El método desarrollado por Shaw Haw Lo[14] consigue reducir el sesgo del error en la estimación del estimador de Good-Toulmin, lo que nos ha permitido reducir el sesgo de los estimadores de Darroch y de Chao.

[14] Shaw-Hwa Lo. "The Annals of Statistics". Vol. 20, N° 2, 1094-1109; 1992.

La idea de Lo es crear información acerca de una cantidad aleatoria desconocida, eliminando una observación de la muestra de dimensión T en un cierto instante dado, y comparar la observación eliminada con el resto de las T-1 informaciones restantes.

Fundamentalmente lo consigue realizando una nueva experiencia apoyado en la equivalencia del "problema de estimar la probabilidad total de especies desconocidas" y el de "estimar la probabilidad de que, en una prueba posterior, se obtenga una especie desconocida condicionada por los resultados anteriores".

Por tanto, una vez se han obtenido los T resultados de una muestra de tamaño T, se detiene el proceso. Entonces se simula una nueva prueba, la (T+1)-ésima, y se observa el resultado en función de los anteriores. Desconocemos lo que sucede en esta nueva prueba, por lo que debemos estimar el valor esperado a partir de los datos de que disponemos en las T pruebas realizadas.

2.10.2. Estimadores de la probabilidad de especies desconocidas

Comenzamos por exponer el proceso que sigue Lo para estimar la probabilidad total de especies desconocidas.

Suponemos que hay M_1 especies que han aparecido una sola vez y M_2 que han aparecido 2 veces.

La probabilidad de que aparezca una especie desconocida se estima por $C_0=M_1/T$; la probabilidad de que aparezca una especie que ha aparecido una sola vez se estima por $C_1=2M_2/T$; y la probabilidad de que aparezca una especie que ya ha aparecido más de una vez, se estima por $1-C_0-C_1=1-M_1/T-2M_2/T$.

Entonces, si designamos por M'_1 el número de especies que aparecen una sola vez en la muestra de tamaño T+1, podemos estimar la esperanza de M'_1 en función de que la especie obtenida haya aparecido una sola vez, más de una vez o ninguna.

Tendremos:

$M'_1 = M_1+1$, con probabilidad $\frac{M_1}{T}$, si no había aparecido:

$M'_1 = M_1$, con probabilidad $1-\frac{M_1}{T}-\frac{2M_2}{T}$, si apareció más de una vez:

$M'_1 = M_1-1$, con probabilidad $\frac{2M_2}{T}$, si apareció una sola vez.

El valor esperado de M'_1 condicionado por los valores $M_1, M_2, \ldots$ obtenidos es:

$$E(\hat{M}'_1/(M_1,M_2,..)) = (M_1+1)\frac{M_1}{T}+M_1\left(1-\frac{M_1}{T}-\frac{2M_2}{T}\right)+(M_1-1)\frac{2M_2}{T} =$$

$$M_1+\frac{M_1}{T}-\frac{2M_2}{T}$$

Si utilizamos $M_1+\frac{M_1}{T}-\frac{2M_2}{T}$ como estimador de M'_1, se obtiene el estimador de C_0:

$$\hat{G}_0 = \frac{1}{T+1}E[M_1/M_1,M_2,..] = \frac{1}{T+1}\left(\hat{M}_1+\frac{\hat{M}_1}{T}-\frac{2\hat{M}_2}{T}\right) = \frac{\hat{M}_1}{T+1}+\frac{\hat{M}_1-2\hat{M}_2}{T(T+1)} =$$

$$= \frac{\hat{M}_1}{T}-\frac{2\hat{M}_2}{T(T+1)}$$

El estimador mejorado para C_0 es, por tanto:

$$\boxed{\hat{G}_0 = \frac{\hat{M}_1}{T}-\frac{2\hat{M}_2}{T(T+1)}} \qquad (2.91)$$

Este estimador no es insesgado para estimar C_0. Sin embargo reduce el sesgo de $\hat{G}_0$, al estimar C_0, en el orden de $o(T^{-2})$.

En efecto, el sesgo de la (T-1)-estimación de orden para estimar la probabilidad de que el resultado de la siguiente prueba pertenezca a la muestra de tamaño T, viene dado por:

$$E\left[\frac{M_1}{T}-\frac{M_1+\delta}{T+1}\right]$$

donde

- δ vale 1, con probabilidad $\frac{\hat{M}_1}{T}$, si la especie obtenida no había salido antes,
- δ vale 0, con probabilidad $1-\frac{\hat{M}_1}{T}-\frac{2M_2}{T}$ si la especie obtenida había salido al menos dos veces,
- δ vale -1, con probabilidad $\frac{2M_2}{T}$, si la especie obtenida había salido una sola vez.

El sesgo del error es, por tanto:

$$E\left[\frac{M_1}{T}-\frac{M_1+\delta}{T+1}\right]=\frac{E(M_1)}{T}-\frac{E(M_1)}{T+1}-\frac{E(\delta)}{T+1}=$$

$$=\frac{E[M_1]\,(T+1)-E[M_1]\,T}{T(T+1)}-\frac{1}{T+1}\left[\frac{E[M_1]}{T}+(-1)\,\frac{2E[M_2]}{T}\right]=$$

$$\frac{E[M_1]}{T(T+1)}-\frac{E[M_1]}{T(T+1)}+\frac{2E[M_2]}{T(T+1)}\le\frac{2}{T+1}=O(T^{-1}) \qquad (2.92)$$

La desigualdad tiene lugar por ser $E[M_2]=\frac{T}{2}E[C_1]$, con $\hat{C}_1<1$.

En cambio, el sesgo de la (T+1)-estimación es del orden de T^{-2}. En efecto, el sesgo de la nueva estimación será ahora:

$$E\left[\frac{M_1}{T}-\frac{2M_2}{T}-\delta\right]=E\left[\frac{M_1}{T}-\frac{2M_2}{T}-\frac{M_1'}{T+1}+\frac{2M_2'}{T+1}\right] \qquad (2.93)$$

donde M_2' representa el número de especies que aparecen dos veces entre todas las pruebas.

Si tenemos en cuenta que:

$$|M_1-M_1'|\leq 1 \quad y \quad |M_2-2M_2'|\leq 2$$

con probabilidad 1, la espresión (2.93) está acotada por 3/(T(T+1)), y, por tanto, el error es de orden (T^{-2}).

Mediante el mismo razonamiento, se llega a un estimador para C_r con menor sesgo, siendo éste:

$$\boxed{\hat{G}_r=\frac{(r+1)\,\hat{M}_{r+1}}{T+1}+\frac{(r+1)\,[(r+1)\,\hat{M}_{r+1}-(r+2)\,\hat{M}_{r+2}]}{T(T+1)}} \qquad (2.94)$$

2.10.2.1. Estimadores de Darroch-Lo y Chao-Lo

Utilizando (2.91), para esimar P_0, los estimadores de Darroch, $\hat{S}_4=\frac{\hat{D}}{\hat{C}}$, y Chao, $\hat{S}_5=\frac{\hat{D}}{\hat{C}}+\frac{\hat{M}_1}{\hat{C}}\hat{\gamma}^2$, mejorados con la técnica de Lo, que se obtienen, son:

$$\hat{S}_8=\frac{\hat{D}}{1-\frac{\hat{M}_1}{T}+\frac{2\hat{M}_2}{T(T+1)}}=\frac{\hat{D}T(T+1)}{(T+1)\,(T-\hat{M}_1)+2\hat{M}_2}$$

Obtenemos, de este modo, los estimadores que vamos a denominar de Darroch-Lo y Chao-Lo, respectivamente, y que son más precisos que los de Darroch y Chao en cuanto al sesgo:

$$\hat{S}_8 = \frac{\hat{D}T(T+1)}{(T+1)(T-\hat{M}_1)+2\hat{M}_2} \qquad (2.95)$$

y

$$\hat{S}_9 = \frac{DT(T+1)}{(T+1)(T-\hat{M}_1)+2\hat{M}_2} + \frac{T(T+1)\hat{M}_1}{(T+1)(T-\hat{M}_1)+2\hat{M}_2}\left(\frac{\hat{D}\hat{R}_2}{(T+1)(T-\hat{M}_1)}-1\right) \qquad (2.96)$$

2.10.3. Estimador de la proporción de especies

Vamos a utilizar el criterio de Lo para mejorar la estimación de especies que son extraídas al menos una vez en una muestra de tamaño T+1. Si designamos por D' el número de especies diferentes que aparecen en la muestra de tamaño T+1, la esperanza de D' se puede estimar en función de si la especie obtenida había salido ya o no había sido obtenida antes.

La diversidad, D, tiene como estimación D', siendo:

D'=D+1, con probabilidad M_1/T, si la especie obtenida no había aparecido antes,

D'=D, con probabilidad $1-M_1/T$, si la especie obtenida había salido antes.

El valor esperado de D', condicionado por los resultados de la muestra de tamaño T es, por lo tanto:

$$E(D/(M_1,M_2,..)) = \frac{1}{T+1}\left((\hat{D}+1)\left(\frac{\hat{M}_1}{T}\right)+\hat{D}\left(1-\frac{\hat{M}_1}{T}\right)\right) = \frac{\hat{D}}{T+1}+\frac{\hat{M}_1}{T(T+1)}$$

es decir:

$$F_0 = \frac{\hat{D}}{T+1}+\frac{\hat{M}_1}{T(T+1)} \qquad (2.97)$$

El sesgo del error, como sucede en la estimación de C_0, se reduce también, siendo del orden de (T^{-2}).

Entonces, la proporción de especies que aparecen al menos una vez en la muestra de tamaño T+1, reduciendo el sesgo del error, es:

$$1-F_0 = 1-\frac{\hat{D}}{T+1}-\frac{\hat{M}_1}{T(T+1)} \qquad (2.98)$$

2.10.3.1. Estimador de S basada en F_0

Se trata ahora de utilizar F_0 para estimar S en las expresiones (2.88) y (2.89). El estimador que se obtiene para el número de especies tomando F_0, es:

$$\hat{S}_{11} = \frac{T^2(T+1)}{2\,[T(T+1-\hat{D})-\hat{M}_1]}\,(\hat{\gamma}^2+1) \qquad (2.99)$$

Bajo la hipótesis de homogeneidad, resulta el estimador:

$$\hat{S}_{10} = \frac{T^2(T+1)}{2\,[T(T+1-\hat{D})-\hat{M}_1]} \qquad (2.100)$$

Tomando $SR_2/(T(T-1))$ como estimador de γ^2+1 , se obtiene:

$$\hat{S}_{11} = \frac{T^4(T+1)\,\hat{R}_2}{4\,[T(T+1-\hat{D})-\hat{M}_1]^2\,(T-1)} \qquad (2.101)$$

2.11. Cálculo de la varianza de los estimadores de S

Los estimadors obtenidos son todos funciones diferenciables de los M_r, lo que nos permite estimar la varianza y, por consiguiente, el error típico de los mismos. Para ello, necesitamos hallar la matriz de covarianzas σ_{xy}, y la varianza del estimador de S vendrá dada por:

$$\hat{\sigma}_{\hat{S}} = \sum_{k=1} \sum_{h=1} \frac{\partial S}{\partial M_k} \frac{\partial S}{\partial M_h} \sigma_{kh} \qquad (2.102)$$

donde

$$\sigma_{kh} = \begin{cases} \hat{M}_k \left(1 - \frac{\hat{M}_k}{\hat{S}}\right) & , \; si \; k=h \\ -\frac{\hat{M}_k \hat{M}_h}{\hat{S}} & , \; si \; k \neq h \end{cases} \qquad (2.103)$$

En forma matricial, se puede expresar del siguiente modo:

$$\sigma_S^2 = L \Sigma L' \qquad (2.104)$$

siendo

$$L = \left(\frac{\partial S}{\partial M_1} , \frac{\partial S}{\partial M_2} , \ldots , \frac{\partial S}{\partial M_h} \right) \qquad (2.105)$$

L' es la traspuesta de L y $\Sigma = (\sigma_{kh})$ es la matriz de covarianzas de los M_k.

2.11.1. Expresión de los estimadores para el cálculo

Para calcular estos parámetros, considerando D y T como variables aleatorias dependientes de los M_k, los distintos estimadores deben expresarse en función de los M_r. El estimador de Darroch-Lo queda en la forma:

$$S_4 = \frac{(\sum kM_k)(\sum M_k)(\sum k(k-1)M_k)}{(\sum (k-1)M_k)(\sum kM_k - 1)} \qquad (2.106)$$

2.12. Ejemplo

Para hacer un estudio comparativo, nos hemos servido del propio ejemplo que utiliza Chao, que lo toma, a su vez, de Holst.

Se trata de un problema de numismática: Ha sido descubierto un tesoro formado por 204 monedas, que han sido clasificadas según los diferentes tipos de troqueles utilizados en el proceso de emisión, y se pretende averiguar el número de troqueles utilizados en el propio período de emisión de las monedas.

Las frecuencias obtenidas para el reverso han sido: 156 aparecen una sola vez, 19 aparecen dos veces, 2 aparecen tres veces y 1 aparece cuatro veces. En el anverso, las frecuencias vienen dadas por: 141 una vez, 26 dos veces, 8 tres veces, 2 cuatro veces, 1 cinco veces, 1 seis veces y 1 siete veces. El siguiente cuadro resumen recoge los valores obtenidos por los estimadores de máxima verosimilitud, Darroch, Chao y estimadores basados en las proporciones.

	Anverso			Reverso		
Estimador	S	CV^2	Error	S	CV^2	Error
S_3	731	0'084	130'6	256	0'335	24'8
S_4	756	0'132	145'1	282	0'465	64'9
S_5	844	"	186'1	378	"	65'2
$*S_6$	800	0'197	141'8	330	0'721	30'7
$*S_7$	958	"	158'7	568	"	70'9
S_8	753	0'127	142'8	281	0'465	32'7
S_9	838	"	183'1	374	"	64'3
$*S_{10}$	797	0'193	139'6	329	0'700	30'5
$*S_{11}$	946	"	156'7	563	"	70'5

Los intervalos de confianza del 44'5% son:

Para el anverso: (497, 1537)

para el reverso, (235,352)

Todos los estimadores homogéneos evaluados caen dentro de los intervalos. Así:

731, 756, 800, 754, 797 están comprendidos entre 497 y 1537.

256, 282, 330, 281, 329 están comprendidos entre 235 y 352.

Observamos cómo, en el anverso (donde el coeficiente de variación de Pearson es pequeño), en los estimadores basados en la proporción (marcados con *), es menor el error típico que en los de Darroch y Chao.

2.13. Resumen del capítulo

Se introduce el modelo no paramétrico como la superposición de S procesos de punto, que son S procesos de Poisson homogéneos, que dan lugar a los dos grandes esquemas de muestreo: esquema de Poisson cuando se observan los procesos durante un intervalo fijo de tiempo (0,t], y el modelo multinomial, que se obtiene cuando se continúa el proceso hasta que tiene lugar un número fijo de sucesos.

En el primer caso, se obtiene la distribución de Poisson de media $\lambda_k t$, y, en el segundo, se demuestra que la distribución tiende a la de Poisson de parámetro $p_k T$.

Estimando la media poblacional a partir de la media de la muestra, y utilizando las propiedades de los dos primeros números de ocupación, se obtiene el estimador de la probabilidad desconocida:

$$\hat{P}_0 = e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

que da lugar al estimador de Poisson (verión de Zelterman):

$$\hat{S} = \frac{D}{1-e^{\frac{-2\hat{M}_2}{\hat{M}_1}}}$$

Desarrollando e^{-t} en serie de Taylor en un entorno de t=0, y particularizando para $t=\lambda/S$, resulta un nuevo estimador de Poisson, que depende de C_0 y C_1:

$$\hat{S} = \frac{D}{\frac{\hat{C}_1}{\hat{C}_0}\left(1-\frac{\hat{C}_1}{2\hat{C}_0}\right)}$$

y que permite obtener fácilmente una expresión de la varianza, ya que C_0, C_1 y D son funciones diferenciables de los M_j.

Se estudian las propiedades fundamentales del modelo multinomial, concluyendo:

1. El vector $M=(M_1,\ldots,M_D)$ tiene una distribución multinomial de parámetros $M(T,M_1/T,\ldots,M_T/T)$.

2. Estimando los P_r por M_1/S, la distribución del vectror M es multinomial $M(T,\hat{M}_1,\ldots,\hat{M}_T)$.

3. Las p_k son constantes, siendo condición necesaria y suficiente para que se verifique la hipótesis de equiprobabilidad que sea nulo el coeficiene de variación de las p_k.

Surgen, de modo natural, los estimadores homogéneos de Darroch y de máxima verosimilitud. Este último es el que coresponde a la distribución de Maxwell-Boltzman, que se deduce de modo diferente al que emplearon Lewontin y Prout, dando también la expresión de la varianza.

Analizando el proceso combinado, se llega a construir un intervalo de confianza para S, bajo la hipótesis de homogeneidad:

$$\frac{TD}{T-\hat{M}_1+2\sqrt{\hat{M}_1}} \leq \hat{S} \leq \frac{TD}{T-\hat{M}_1-2\sqrt{\hat{M}_1}}$$

Desarrollando en serie de Taylor $e^{-p_k T}$ en un entorno de $p_k=1/S$, se obtiene, de modo sencillo, el estimador de Chao, que depende del coeficiente de variación de Pearson de las p_k, el cual, a su vez depende de S. Por ello, es necesario hacer previamente una estimación de S a partir de un estimador homogéneo como el de Darroch, que se elige por ser, según demostró Esty "*bastante eficiente cuando se le compara con el de máxima verosimilitud*".

El estimador de Chao se basa en el recubrimiento muestral C, siendo:

$$\hat{S} = \frac{\hat{D}}{\hat{C}} + \frac{T(1-\hat{C})}{\hat{C}}\hat{\gamma}^2$$

Desarrollando también $e^{-p_k T}$ en serie de Taylor, pero ahora, utilizando $t=p_k T$ como variable, se obtiene un nuevo estimador para una situación no homogénea, que no depende del recubrimiento muestral, sino de la proporción de especies en la muestra:

$$\hat{S} = \frac{\frac{T}{2}}{1-\frac{D}{T}}(\hat{\gamma}^2+1)$$

Finalmente, hemos aplicado los métodos de Lo para reducir el sesgo en los estimadores de P_0 y de F_0, obteniendo como estimadores de S de menor sesgo para el caso no homgéneo:

$$\hat{S}_9 = \frac{DT(T+1)}{(T+1)(T-\hat{M}_1)+2\hat{M}_2} + \frac{T(T+1)\hat{M}_1}{(T+1)(T-\hat{M}_1)+2\hat{M}_2}\left(\frac{\hat{D}\hat{R}_2}{(T+1)(T-\hat{M}_1)}-1\right)$$

y

$$\hat{S}_{11} = \frac{T^4(T+1)\hat{R}_2}{4[T(T+1-\hat{D})-\hat{M}_1]^2(T-1)}$$

MODELO PARAMÉTRICO

III MODELO PARAMÉTRICO

3.1. Introducción

Hemos representado el "muestreo de las especies" como una superposición de procesos de Poisson en el intervalo de tiempo (0,t], con t>0, que, al ser desconocidas las especies que no forman parte de la muestra, debemos truncarla en el origen. Representamos por $N_k^*(t)$ la distribución truncada en el origen y por $N_k(t)$ la distribución ordinaria. Sus expresiones son:

$$P[N_k(t)=r] = \frac{(\lambda_k t)^r \, e^{-\lambda_k t}}{r!} \qquad (3.1)$$

que proporciona la probabilidad de obtener r individuos de la especie I_k en el intervalo (0,t], y

$$P[N_k^*(t)=r] = P[N_k(t)=r \mid N_k(t)\geq 1] = \frac{(\lambda_k t)^r \, e^{-\lambda_k t}}{r!\,(1-e^{-\lambda_k t})} \qquad (3.2)$$

que es la probabilidad de obtener r individuos de la especie I_k en el intervalo (0,t], cuando ya se ha obtenido al menos un individuo de la clase I_k.

Se puede generalizar el modelo haciendo que las λ_k sean, a su vez, variables aleatorias con una distribución común como puede ser la gamma de parámetros (A,1/A) dando lugar al modelo de Polya.

3.2. Proceso de Polya

La elección de una distribución gamma de parámetros (A,1/A) para los λ_j nos lleva al proceso de Polya, que pasamos a analizar con detalle.

El proceso de Polya se obtiene fundamentalmente de tomar, como distribución de los λ_j, una distribución gamma, de tal modo que la distribución compuesta es la binomial negativa, que viene dada por:

$$P(N_k(t)=r/\lambda) = \frac{\Gamma(A+r)}{r!\Gamma(A)}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \quad , \quad r=1,2,\ldots \qquad (3.3)$$

Vamos a suponer que las diversas especies constituyen una partición de la población, formada por D+1 clases: $I_1, I_2, \ldots, I_D$, cuyas especies son conocidas, y una clase I_0, compuesta por las especies desconocidas (que no forman parte de la muestra).

Si realizamos un experimento consistente en la selección aleatoria de individuos de la población, y realizamos, no un número fijo de pruebas, sino que continuamos el proceso hasta que se obtienen por primera vez A especies de la clase I_0, habiéndose obtenido antes r-1 individuos de la clase I_k, estamos ante el tipo de muestreo secuencial, cuya distribución conjunta corresponde a la distribución multinomial negativa.

La distribución (3.3) puede, entonces, ser entendida como la probabilidad del número de fracasos que preceden a A éxitos, siendo $N_k(t)$+A el número total de pruebas necesarias para conseguir A éxitos, lo que sucede si en la última prueba resulta éxito, y, en las A+$N_k(t)$-1 previas ha habido exactamente $N_k(t)$ fracasos en un intervalo de tiempo (0,t].

Se trata de un modelo de muestreo secuencial, en el cual, en lugar de seleccionar una muestra de tamaño fijo T, alterando la regla de parada, se continúa el muestreo hasta que se consigue el A-ésimo éxito.

Vamos a considerar *"éxito" obtener un individuo perteneciente a la clase* I_0. De esta forma, la probabilidad de que $N_k(t)$ sea igual a r es la probabilidad de obtener r individuos de la clase I_k en A+r pruebas.

Se trata del proceso de Polya, modelo al que conduce el esquema de urnas de Polya. Es un "proceso de renovación ordinario", aunque también puede ser considerado como "un proceso puro de nacimiento, no estacionario[1]"

Este proceso se puede analizar mediante el esquema de urnas de Polya, que consiste en una urna compuesta por b bolas blancas y a azules. Se extrae una bola al azar, que resulta ser de un determinado color, y se reemplaza añadiendo a la urna c bolas más del mismo color. Se extrae una nueva bola de la urna y se repite el mismo procedimiento.

Después de cada extracción el número de bolas del color extraído se incrementa en c bolas más, mientras que el número de bolas del otro color no cambia.

La probabilidad de que, en $n=n_1+n_2$ extracciones, resulten n_1 bolas blancas y n_2 bolas azules viene dada por:

$$P_{n_1,n} = \frac{\binom{-b/c}{n_1}\binom{-a/c}{n_2}}{\binom{-(b+a)/c}{n}} \qquad (3.4)$$

Si hacemos

$$\frac{b}{b+a}=p\,,\quad \frac{r}{b+a}=q\quad y\quad \frac{c}{b+a}=\gamma \qquad (3.5)$$

queda:

[1] Feller,W., Introducción a la Teoría de las Probabilidades y sus Aplicaciones". Volumen I, Capítulo XVII. Limusa, México, 1993.

$$P_{n_1,n} = \frac{\binom{-p/\gamma}{n_1}\binom{-q/\gamma}{n_2}}{\binom{-1/\gamma}{n}} \qquad (3.6)$$

expresión que es válida para valores arbitrarios p>0, q>0 y γ>0, no necesariamente racionales, pero verificando p+q=1.

Si $n\to\infty, p\to 0$ y $\gamma\to 0$, de modo que $np\to\lambda, n\gamma\to\rho^{-1}$, entonces, para n fijo, resulta:

$$\boxed{P_{n_1,n} = \binom{A+n_1}{n_1}\left(\frac{\rho}{1+\rho}\right)^A\left(\frac{1}{1+\rho}\right)^{n_1}} \qquad (3.7)$$

El proceso de Polya se obtiene, por tanto, como paso al límite del esquema de urnas de Polya, y se trata de un proceso puro de nacimiento, no estacionario. Las ecuaciones diferenciales que dan lugar al proceso son:

$$\begin{aligned} P'_r(t) &= -\frac{A+r}{t}\lambda P_r(t) + \frac{A+r-1}{t}\lambda P_{r-1}(t), \quad r\geq 1 \\ P'_0(t) &= 0 \end{aligned} \qquad (3.8)$$

con las condiciones iniciales

$$P_i(0) = 1 \ , \ P_r(0) = 0 \ , \ r\neq A \qquad (3.9)$$

La solución a este sistema fue dada, en primer lugar, por Yule en la forma:

$$\boxed{P_r(t) = \binom{A+r-1}{r} e^{(-\lambda t)^A}(1-e^{-\lambda t})^r, \ r>A} \qquad (3.10)$$

Se trata, por tanto, de una distribución binomial negativa de parámetros:

$$BN(r;A;e^{-\lambda t})$$

o, lo que es equivalente: BN(A,p), a la que llegamos a partir de la distribución de Poisson compuesta con la distribución gamma de las λ_k.

Cuando la función de densidad es, como en nuestro caso, una gamma de parámetros A y 1/A, se obtiene la forma límite[1] de la distribución de Polya para el valor de c=1, según se vio en el apartado 1.5:

$$P_r = \binom{A+r-1}{A-1}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A, \quad r=1,2,\ldots \tag{3.11}$$

Como es desconocido el número de especies que no aparecen en la muestra, vamos a tomar la distribución truncada en cero:

$$P_r^* = \binom{A+r-1}{A-1}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{A+t}\right)^A}, \quad r=1,2,\ldots \tag{3.12}$$

donde llamamos

$$P_r^* = \frac{P_r}{1-P_0} = P[U_k(t)=r] = P[N_k(t)=r \mid N_k(t)>0] \tag{3.13}$$

siendo P_0 la suma de las probabilidades de las especies desconocidas.

El caso particular en que c=1 corresponde al esquema de contagio.

Si c=0, tenemos el muestreo con reemplazamiento (regido por la distribución de Poisson).

Si c=-1, se trata de muestreo sin reemplazamiento, cuya distribución es la binomial, puesto que hemos supuesto que la población es infinita.

3.2.1. Función generatriz de probabilidades

Se puede interpretar $P[N_k(t)]$ como la distribución de probabilidad de obtener $N_k(t)=r$ individuos de la especie I_k, habiéndose realizado $A+N_k(t)$ pruebas.

Proposición 3.1: La función generatriz de probabilidades de la distribución binomial negativa de parámetros A y p viene dada por:

$$\boxed{\phi_{N_k(t)}(z) = p^A(1-qz)^{-A} = \ , \ 0<p<1.} \qquad (3.14)$$

siendo: $p = \frac{A}{A+t}$, $q=1-p=\frac{t}{A+t}$

A partir de la función generatriz de probabilidades se deducen los primeros momentos de esta distribución:

1. $\alpha_1 = E[N_k(t)] = \frac{At}{A} = t$ (3.15)

2. $\alpha_2 = E\{[N_k(t)]^2\} = \frac{t[t+A(t+1)]}{A}$ (3.16)

3. $\alpha_3 = E\{[N_k(t)]^3\} = t+3t^2+t^3+\frac{2t^3}{A^2}+\frac{3t^2}{A}+\frac{3t^3}{A}$ (3.17)

La varianza viene dada por:

4. $\mu_2 = \sigma^2 = Var[N_k(t)] = \frac{(A+t)\,t}{A}$ (3.18)

y el índice de dispersión es:

5. $I_k = \frac{A+t}{A} = \frac{1}{p} > 1$ (3.19)

Observamos una característica de la binomial negativa, que le diferencia de las distribuciones binomial y de Poisson: *"el índice de dispersión es mayor que 1"*.

La admisión del axioma III nos ha permitido construir este modelo paramétrico, cuya distribución, al ser desconocido el número de individuos que no forman parte de la muestra, la utilizaremos truncada en cero, denotándola por $N_k^*(t)$.

P_r^* denota la probabilidad del número de representantes de la clase I_r en un proceso en el que, de A+r individuos seleccionados, ya ha aparecido uno de la especie I_k en el intervalo [0,t].

Se verifican las siguientes propiedades para P_r^*:

I. $$\boxed{\sum_{r=1}^{D} P_r^* = 1} \qquad (3.20)$$

II. $$\boxed{\sum_{r=1}^{D} P_r = (1-P_0)\sum_{r=1}^{D} P_r^* = 1-P_0} \qquad (3.21)$$

3.2.2. Momentos de la distribución truncada ($N_k^*(t)$)

La función generatriz de probabilidades de la distribución truncada es:

$$\boxed{\Psi_{N_k^*(t)}(z) = \frac{1}{1-P_0}\left(\frac{p}{1-qz}\right)^A = \frac{1}{1-P_0}\Phi_{N_k(t)}(z) \quad , \quad 0<p<1} \qquad (3.22)$$

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

$$\frac{1}{1-P_0} = \frac{1}{1-\left(\frac{A}{A+t}\right)^A} \tag{3.23}$$

Se obtiene así:

1. $$E\left[N_k^*(t)\right] = \frac{tA}{A}\,\frac{1}{1-P_0} = \frac{t}{1-P_0} \tag{3.24}$$

2. $$E\left[\left(N_k^*(t)\right)^2\right] = \frac{t}{1-P_0}\left(1+t+\frac{t}{A}\right) \tag{3.25}$$

3. $$E\left[N_k^*(t)^3\right] = \frac{t}{1-P_0}\left(t+t^2+t^3+\frac{2t^3}{A^2}+\frac{2t^2}{A}+\frac{3t^3}{A}\right) \tag{3.26}$$

4. $$Var\left[N_k^*(t)\right] = \frac{t}{1-P_0}\left(1+t+\frac{t}{A}-\frac{t}{1-P_0}\right) \tag{3.27}$$

5. $$I(t) = 1+t+\frac{t}{A}-\frac{t}{1-P_0} \tag{3.28}$$

También se puede calcular aproximadamente:

6. $$E\left[\frac{1}{N_k^*(t)}\right] \approx \left[t-\frac{A+t}{A}\right]^{-1} \tag{3.29}$$

7. $$Var\left[\frac{1}{N_k^*(t)}\right] \approx \frac{A+t}{A}\left[t-1-\frac{t}{A}\right]^{-2}\left[t-2-\frac{2t}{A}\right]^{-1} \tag{3.30}$$

Si estimamos la media de la población mediante la media muestral, los primeros momentos respecto al origen y la varianza de la distribución truncada, quedan en la forma:

$$I'.\quad E\left[U_k^*(t)\right] = \frac{t}{1-P_0} = \frac{tS}{D} = \frac{\hat{T}}{D} \tag{3.31}$$

$$II'.\quad E\left[\left(N_k^*(t)\right)^2\right] = \left(t+t^2+\frac{t^2}{A}\right)\frac{1}{1-P_0} = \frac{T}{D}+\frac{T^2}{DS}+\frac{T^2}{DSA} = \frac{T}{D}\left(1+\frac{T}{S}+\frac{T}{SA}\right) \tag{3.32}$$

$$III'.\quad Var\left[N_k^*(t)\right] = \frac{T}{D}\left(1-\frac{T}{D}+\frac{T}{S}+\frac{T}{SA}\right) \tag{3.33}$$

3.2.3. Distribución del tamaño muestral (T)

En el modelo de muestreo secuencial, el tamaño muestral, T, es una variable aleatoria. Veamos su distribución:

T es suma de D variables aleatorias independientes, todas ellas con la misma distribución binomial negativa BN(A,p). Entonces, si tenemos en cuenta las propiedades fundamentales de la binomial negativa: *"La suma de variables aleatorias independientes con distribución binomial negativa B(A,p) es una variable aleatoria también con distribución binomial negativa B(DA,p)"*, según se demuestra en el apéndice teórico.

3.3. Propiedades de los números de ocupación y del recubrimiento para la distribución truncada

Teniendo en cuenta la propiedad:

$$\boxed{E[M_r] = E_\lambda\left[E(M_r|\lambda)\right]} \tag{3.34}$$

fijada la diversidad D, resulta:

$$E[M_r] = DP[N_k^*(t) = r] = SP[N_k(t) = r] \tag{3.35}$$

Se obtienen así los siguientes resultados:

1. $$E(M_r) = \frac{DP_r}{1-P_0} = DP_r^* \quad r=0,1,2,\ldots \tag{3.36}$$

En efecto:

$$E(M_r) = E\left(\sum_{k=1}^{S} I[N_k(t)=r]\right) = \sum_{k=1}^{D} P[N_k^*(t)=r] = \frac{DP_r}{1-P_0} = SP_r = DP_r^*$$

Como consecuencia de esta relación, se pueden estimar P_r y P_r^* por medio de:

2. $$\hat{P}_r = \frac{E(M_r)\ (1-P_0)}{D}, \quad r=0,1,2,\ldots \tag{3.37}$$

3. $$\hat{P}_r^* = \frac{\hat{P}_r}{1-P_0} = \frac{E[M_r]}{D}, \quad r=1,2,\ldots \tag{3.38}$$

Podemos, entonces, enunciar la siguiente proposición:

Proposición 3.2: $\frac{E[M_r]}{D}$ es un estimador insesgado de P_r^*.

Veamos que la propiedad 1 también es cierta para r=0, es decir:

4. $$\hat{P}_0 = \frac{S-D}{S} \tag{3.39}$$

En efecto:

$$P_0 = 1 - \sum_{r=1}^{D} P_r = 1 - \frac{1}{S}\sum_{r=1}^{D} M_r = 1 - \frac{D}{S} = \frac{S-D}{S}$$

Si designamos por M_0=S-D, el número de especies desconocidas, tenemos (3.39).

De (3.39) se deduce inmediatamente:

5. $$\boxed{1 - \hat{P}_0 = \frac{D}{S}} \qquad (3.40)$$

Estas propiedades nos proporcionan una familia de estimadores de S, ya que las esperanzas de los números de ocupación son proporcionales a los términos del desarrollo de la distribución binomial negativa, siendo S la constante de proporcionalidad:

$$\hat{S} = \frac{E[M_r]}{P_r} = \frac{DP_r^*}{P_r} = \frac{D}{1-P_0}, \quad r=0,1,2,\ldots$$

3.3.1. Esperanza de D

Proposición 3.3: $\frac{D}{1-\hat{P}_0}$ es un estimador insesgado de S, bajo la distribución truncada en cero:

$$\boxed{E(D) = S(1-P_0) = D} \qquad (3.41)$$

Lo pone de manifiesto la igualdad (3.40).

$D/(1-\hat{P}_0)$ es un estimador insesgado de S, que va a depender de t y de A; este último parámetro determina el coeficiente de variación de Pearson de las p_k, y, por consiguiente, el grado de heterogeneidad de su distribución, según confirmaremos más adelante.

3.4. Vector de ocupación y números de ocupación

Veamos cuál es la distribución del vector $\mathbf{N}=(N_1(t),\ldots,N_D(t))$. Para ello, vamos a remodelar el esquema clásico de urnas, dándole un nuevo enfoque. El esquema de urnas consiste en distribuir una bola al azar en una de las S celdas etiquetadas con los números $1,2,\ldots,S$, de modo que la probabilidad de que la bola caiga en la celda i-ésima es p_i.

El nuevo enfoque consiste en observar el valor de un vector aleatorio $\mathbf{X}$, que puede estar formado por uno de los vectores unitarios $\mathbf{e}_1,\ldots,\mathbf{e}_S$, de modo que $P(\mathbf{X}=\mathbf{e}_i)=p_i$.

Identificarèmos el suceso $\mathbf{X}_i=\mathbf{e}_i$ con el resultado de que "la bola caiga en la celda i-ésima".

Las repeticiones independientes de este experimento vienen descritas por vectores aleatorios independientes y con la misma distribución $\mathbf{X}_1,\mathbf{X}_2,\ldots$, donde $\mathbf{X}_i$ es el resultado de la i-ésima prueba.

Definición 3.1: El vector aleatorio $\mathbf{N}_n(t)$ en el instante t, se define como la suma

$$\mathbf{N}_n(t)=\mathbf{X}_1(t)+\ldots+\mathbf{X}_n(t),$$

donde la componente j-ésima de $\mathbf{N}_n(t)$ es $N_{nj}(t)=N_{1j}(t)+\ldots+N_{nj}(t)$, y representa el número (de entre las n primeras bolas) que hay en la celda j en el instante t. $\mathbf{N}_n(t)$ se conoce como "*vector de ocupación*".

La distribución del vector de ocupación se concentra en un conjunto finito S_n, que se define como:

$$S_n=\left\{x\in\mathbb{R}^S \mid x_j=1,2,\ldots,n\ ,\ \sum_{j=1}^{S} x_j=n\right\}$$

En el modelo de muestreo secuencial, el vector aleatorio $\mathbf{N}_n(t)$ sigue la distribución multinomial negativa.

Nos interesa considerar el número N_k de bolas que que hay en la celda k en el instante L+H. Sea, por tanto, L+H el tiempo transcurrido (ó número de pruebas necesarias) para tener, por primera vez, H especies de la clase desconocida I_0.

Nos interesa estudiar la distribución del vector aleatorio $N_n=N_1+\ldots+N_D$, donde N_k representa el número de individuos de la especie k que hay en el instante L+H.

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:

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

Sea L+H el tiempo, o lo que es equivalente, el número de pruebas requeridas para conseguir, por primera vez, H individuos de la clase I_0, y sea $N_k(t)$ el número de individuos de la clase I_k en el instante L+H.

3.5. Distribución de los números de ocupación

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.

Para conseguirlo, es conveniente tener en cuenta las consideraciones de Engen[2], que afirma que la media y la varianza de los M_r vienen condicionadas por las S variables $(\lambda_1,\ldots,\lambda_S)$, de modo que la función de densidad $f(\lambda)$ ha de ser tomada como una aproximación, por lo que:

$$E[M_r \mid \lambda] \doteq SP_r = DP_r^*$$

[2] Engen, S. "Comments on two different approachs to the Analysis of Species Frequenty Distribution. Biometrica, 33, 205-213.

Esta aproximación está en concordancia con las consideraciones de Engen y es la que proporciona la menor varianza.

3.5.1. Función generadora de probabilidades de M_r

Al ser las M_r variables intercambiables, se verifica que:

$$\phi_{M_r}(u) = E\left[u^{M_r \mid \sum_{k=1} M_r = D}\right] = E\left[u^{\sum_{k=1} I[N_k(t)=r]}\right] =$$

$$= \prod_{r=1}^{D} \left[u\frac{P_r}{1-P_0} + 1 - \frac{P_r}{1-P_0}\right] = \tag{3.42}$$

$$= \prod_{k=1}^{D} \left[(u-1)\frac{P_r}{1-P_0} + 1\right] = \left[(u-1)\frac{P_r}{1-P_0} + 1\right]^D = \tag{3.43}$$

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:

Propsición 3.5: 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 $M=(M_1,M_2,\ldots,M_D)$ condicionado por $M_1+M_2+\ldots+M_D=D$, es multinomial $M_\alpha(D,P_1/(1-P_0)),\ldots,P_D/(1-P_0))$.

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. $$E[M_r] = \frac{DP_r}{(1-P_0)} = SP_r \qquad (3.44)$$

2. $$E[M_r(M_r-1)] = \frac{D(D-1)P_r^2}{(1-P_0)^2} \qquad (3.45)$$

3. $$E[M_r^2] = \frac{D(D-1)P_r^2}{(1-P_0)^2} + \frac{DP_r}{1-P_0} \qquad (3.46)$$

4. $$Var(M_r) = \frac{DP_r}{1-P_0} - \frac{DP_r^2}{(1-P_0)^2} = \frac{DP_r}{1-P_0}\left(1-\frac{P_r}{1-P_0}\right) = \hat{M}_r\left(1-\frac{\hat{M}_r}{D}\right) \qquad (3.47)$$

5. $$E[M_rM_k] = \frac{D(D-1)P_rP_k}{(1-P_0)^2}, \quad si\ r\neq k \qquad (3.48)$$

6. $$Cov[M_rM_k] = -\frac{\hat{M}_r\hat{M}_k}{D} \qquad (3.49)$$

Veamos ahora cuál es la función de verosimilitud. Para ello, partimos del conocimiento de la distribución de los M_r.

3.6. Función de verosimilitud

Teniendo en cuenta que los M_r son independientes y se distribuyen según una distribución binomial de parámetros $B(D,P_r/(1-P_0))$, la distribución conjunta del vector $(M_1,\ldots,M_D)$ condicionada por $M_1+\ldots+M_D=D$ es multinomial de parámetros $M(D,P_1/(1-P_0),\ldots,P_D/(1-P_0))$, luego la función de verosimilitud es:

$$L \equiv P\left(G(1)=M_1,\ldots,G(D)=M_D \mid \sum_{k=1}^{D} M_k = D\right) = \prod_{r=1}^{D} \frac{P_r^{M_r}}{1-P_0} \qquad (3.50)$$

3.6.1. Estimador de máxima verosimilitud

Tomando logaritmos en los dos miembros de (3.50) se obtiene:

$$\ln L = \sum_{r=1}^{D} M_r \ln P_r - D\ln(1-P_0) \qquad (3.51)$$

Derivando con respecto a S, queda:

$$\frac{\partial \ln L}{\partial S} = \sum_{r=1}^{D} M_r \frac{P_r'}{P_r} + D\frac{P_0'}{1-P_0} \qquad (3.52)$$

donde hemos llamado $P_r' = \frac{d}{dS} P_r$

Si estimamos t por T/S en las expresiones de P_r y P_0, resulta la relación:

$$\frac{P_r'}{P_r} = \frac{A(Tt-rS)}{S(SA+T)} \qquad (3.53)$$

de donde se deduce:

$$\sum_{r=1}^{D} M_r \frac{P_r'}{P_r} = \frac{TA(D-S)}{S(SA+T)} \qquad (3.54)$$

Por otra parte tenemos:

$$\frac{DP_0'}{1-P_0} = \frac{DTAP_0}{S[SA+T](1-P_0)} \qquad (3.55)$$

Igualando a cero la derivada, resulta:

$$-\frac{TA(D-S)}{S[SA+T]} = \frac{DTAP_0}{S[SA+T](1-P_0)} \qquad (3.56)$$

de donde, teniendo en cuenta que $P_0 = \frac{S-D}{S}$ y $\hat{M}_1 = \frac{AT(S-D)}{SA+T}$, se obtiene:

$$\frac{TAP_0}{SA+T}\left(\frac{D}{S(1-P_0)} - 1\right) = \frac{AqP_0}{S}\left(\frac{D}{1-P_0} - S\right) \qquad (3.57)$$

Hemos obtenido la siguiente relación:

$$\frac{\partial \ln L}{\partial S} = \frac{AqP_0}{S}\left(\frac{D}{1-P_0} - S\right) \qquad (3.58)$$

Igualando a cero la derivada, resulta finalmente:

$$\hat{S} = \frac{D}{1-P_0}, \ c.q.d. \qquad (3.59)$$

Hemos demostrado la siguiente proposición:

Proposición 3.6: $D/(1-P_0)$ es un estimador de máxima verosimilitud de S.

3.6.2. Varianza del estimador

La derivada del logaritmo de la función de verosimilitud para la distribución truncada viene dada por la expresión (3.58), donde el primer factor es independiente de las observaciones, lo que nos permite afirmar que la cota de Cramer-Rao es accesible. Se trata, por tanto, de un estimador uniformemente de mínima varianza[3] (U.M.V.), siendo la varianza el inverso del factor que multiplica a [D/(1-P_0)-S], es decir:

$$\hat{Var}(\hat{S}) = \frac{\hat{S}}{AqP_0}$$

donde $\hat{S} = \frac{D}{1-P_0}$ y $\hat{M}_1 = AqP_0$.

Hemos demostrado la siguiente proposición:

Proposición 3.7: Una estimación de la varianza del estimador de máxima verosimilitud de S es

$$\boxed{\hat{Var}(\hat{S}) = \frac{\hat{S}^2}{\hat{M}_1}} \qquad (3.60)$$

3.7. Distribución del número de pruebas

Proposición 3.8: Sea A un número entero, y designemos como "éxito" el suceso "obtener un individuo de la clase I_0". Sea $V_{k,A}(t)=N_k(t)+A$ "*el número Y de pruebas necesarias para conseguir por primera vez el A-ésimo éxito en una sucesión de pruebas de Bernoulli, todas con probabilidad p de éxito*".

[3] M. Kendall y A. Stuart. "The Advanced Theory of Statistics", Vol. 2. 17.17. Charles Griffin & Co Ltd. London 1977 (4ª Ed.).

Entonces, la variable $V_k(t)=N_{k,A}(t)+A$ tiene como distribución la binomial negativa de parámetros (A,p) en la forma:

$$P[V_k(t)=y]=\binom{y-1}{A-1}\left(\frac{t}{t+A}\right)^{y-A}\left(\frac{A}{t+A}\right)^{A},\quad y=A,A+1,\ldots$$

Es otra expresión de la binomial negativa de parámetros (A,p). Cualquiera de las dos acepciones nos llevará al mismo estimador.

Demostración: Consideremos el suceso $\{V_{k,A}=A+r\}$. Este puede tener lugar si, y sólo si la prueba A+r es un éxito, habiendo resultado exactamente r fracasos en las primeras A+r-1 pruebas. Entonces se verifica:

$$P[V_{k,A}(t)=A+r]=P(V_{k,A+r-1}(t)=A-1,\,V_{k,A+r}(t)=1)=$$

$$=P(V_{k,A+r-1}=A-1)\,P(V_{k,A+r}=1)=\binom{A+r-1}{A-1}p^{A-1}(1-p)^{r}=\binom{Y-1}{A-1}(1-p)^{Y-A}p^{A}$$

Denotaremos a la distribución del número de pruebas truncada en cero por $V_k^*(t)$, con lo que se obtiene:

$$P[V_k^*(t)=y]=\binom{y-1}{A-1}\left(\frac{t}{t+A}\right)^{y-A}\left(\frac{A}{t+A}\right)^{A}\frac{1}{1-\left(\frac{A}{t+A}\right)^{A}},\quad y=A,A+1,\ldots$$

Proposición 3.9: La función generatriz de probabilidades de la distribución del número de pruebas, $V_k^*(t)$, es la binomial negativa truncada en cero, que viene dada por:

$$M_{V_k^*(t)}(z)=(pz)^{A}(1-qz)^{-A}\frac{1}{1-P_0},\quad 0<p<1 \qquad (3.61)$$

siendo también: $p=\frac{A}{A+t}$, $q=1-p=\frac{t}{A+t}$

Los primeros momentos de la distribución de los tiempos de espera son ahora:

$$E[V_k^*(t)] = E[N_k^*(t)] + \frac{A}{1-P_0} = \frac{t}{1-P_0} + \frac{A}{1-P_0} = \frac{t+A}{1-P_0} \quad (3.62)$$

$(t+A)/(1-P_0)=S(t+A)/D$ es el número medio de pruebas esperado para conseguir St/D individuos de la clase I_k y SA/D individuos de la clase I_0.

$$E[V_k^*(t)^2] = \frac{t}{1-P_0}\left(1+t+\frac{t}{A}\right) \quad (3.63)$$

y la varianza viene dada por la expresión:

$$\boxed{Var\left[V_k^*(t)\right] = \frac{t}{1-P_0}\left(1+t+\frac{t}{A}-\frac{t}{1-P_0}\right)} \quad (3.64)$$

3.7.1. Tiempos de espera y tiempos entre llegadas

$N_k^*(t)=V_k^+-A$ es el número de fracasos que preceden al A-ésimo éxito en una sucesión de $V_k^*(t)$ pruebas de Bernoulli con probabilidad de éxito p, que se puede expresar como

$$V_A^*(t) = X_1+X_2+\ldots+X_A \quad (3.65)$$

donde $X_1, X_2, \ldots, X_A$ son variables aleatorias independientes,

todas ellas con una misma distribución geométrica de parámetro $p=A/(A+t)$, por lo que será $P(X_{k,j}=r)=pq^{A-1}$ la probabilidad, truncada en cero, del tiempo de espera necesario para obtener un individuo de la clase I_0.

Así, $X_{k,1}$ es el tiempo de espera necesario para obtener el primer éxito (primer individuo perteneciente a la clase I_0 después de la etapa cero en el proceso $N_k^*(t)$), $X_{k,2}$ es el tiempo de espera para un nuevo individuo de la clase I_0, y así sucesivamente.

$V_k^*(t)$ es el número de pruebas necesarias hasta que se obtiene el A-ésimo individuo perteneciente a la clase I_0, con la condición de que ha aparecido al menos un individuo de la clase I_k.

Los $X_{k,h}$ se pueden poner:

$$X_{k,1} = V_{k,1}, X_{k,2} = V_{k,2} - V_{k,1}, \ldots, X_{k,A} = V_{k,A} - V_{k,A-1} \qquad (3.66)$$

Proposición 3.10: Los $X_{k,h}$ son los tiempos entre llegadas desde que aparece el (h-1)-ésimo individuo de la clsae I_0 hasta que aparece el h-ésimo, y siguen una distribución geométrica de razón p=A/(A+t), cuya función generatriz de probabilidades es:

$$M(z) = p(1-qz)^{-1} \ , \ p = \frac{A}{A+t} \quad y \quad q = 1-p.$$

En efecto: Sean $x_1, x_2, \ldots, x_D$ números enteros positivos. El suceso

$[X_{k^*,1}^A = x_1, X_{k,2}^* = x_2, \ldots, X_{k,D}^* = x_D] = [V_{k,1}^* = x_1, V_{k,2}^* = x_1 + x_2, \ldots, V_{k,D} = x_1 + x_2 + \ldots + x_D]$

tiene lugar si, y sólo si las pruebas $x_1, x_1+x_2, \ldots, x_1+x_2+\ldots+x_D$ son éxitos y el resto de las primeras pruebas $x_1, \ldots x_D$ son fracasos.

Así:

$$P(X_1 = x_1, \ldots, X_D = x_D) = p^D (1-p)^{x_1+\ldots+x_D-D} = \prod_{k=1}^{D} p(1-p)^{x_i-1}$$

Es evidente que $P(X_1 = x_1, \ldots, X_D = x_D) = 0$, si $x_1, \ldots x_D$ no son todos enteros positivos. De este modo:

$$P(X_1=x_1,\ldots,X_D=x_D)=\prod_{k=1}^{D} p(1-p)^{x_i-1}1_{(1,2,..)}(x_k)$$

Como se trata de un producto de funciones de los x_k solamente, $X_1,\ldots,X_D$ son independientes.

Al ser $P(X_1=x_1)=P(X_D=x_D)=p(1-p)^{x-1}1_{(1,2,..)}(x)$, $X_{k,1},\ldots,X_{k,D}$ tienen la misma distribución que $V_{k,1}$. Por tanto:

$$E[X_{K,h}]=\frac{A+t}{A} \quad y \quad Var(X_{k,h})=\frac{1-p}{p^2}=\frac{t(A+t)}{A^2}$$

3.8. Relación de la distribución beta con la binomial negativa y la binomial.

Si se necesitan al menos n pruebas para obtener el A-ésimo éxito, entonces el número de éxitos en n pruebas es A como máximo.

Si $W_k(t)$ sigue una distribución binomial negativa con parámetros BN(A,p), y $C_k(t)$ es una distribución binomial con parámetros B(n, p), se dan las siguientes relaciones entre ellas[4]:

$$\boxed{P[V_k(t)\leq n]=P[C_k(t)\geq A]}$$

y

$$\boxed{P[V_k(t)>n]=P[C_k(t)<A]}$$

donde n=A+r.

[4] Rohatgi, Vijay K. "Statistical Inference". Capitulo 6, apartado 6.5. John Wiley & Sons. New York 1984.

Además, la binomial negativa está relacionada con la distribución beta por medio de la siguiente expresión[5]:

$$\boxed{P[\beta_n(t) \leq p] = P[C_n(t) \geq A]}$$

donde $\beta_n(t)$ es una beta de parámetros $\beta(A,n-A+1)$, y $C_n(t)$ es binomial de parámetros: $B(n,p)$.

Entonces, los tiempos de espera verifican:

$$P[V_k(t) \leq n] = 1-P[V_k(t) > n] = 1-P[C_k(t) < A] =$$

$$= P[C_k(t) \geq A] = P[\beta_k(t) \leq p]$$

Como $\beta_k(t)$ sigue una distribución beta de parámetros $\beta(A,A+k)$ en este caso, se tiene:

$$P[V_k(t) \leq A+1] = 1-P[\beta_1(t) < p] = \frac{1}{B(A,1)}\int_0^p x^{A-1}(1-x)^0 dx = p^A = P_0.$$

Llegamos de nuevo a la expresión de P_0 en función de p.

$$\boxed{P_0 = p^A} \qquad (3.67)$$

3.9. Conexión con el muestreo sin reemplazamiento

Si consideramos el número medio de individuos de la clase I_k que hay en el intervalo (0,1], 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).

[5] Rohatgi, Vijay K. "Statistical Inference" John Wiley & Sons. Capítulo 6, apartado 7.5. New York 1984.

$J_k(t)$ 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].

Como hemos supuesto que los sucesos que tienen lugar en los sucesivos puntos de tiempo son independientes, los $J_k(t)$ son también independientes entre sí.

La distribución de $J_k(t)$ es binomial de parámetros B(A,q), y 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:

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

Entonces

$$\boxed{E\left[\sum_{k=1}^{D} J_k^*(t)\right] = SAq} \qquad (3.68)$$

Podemos enunciar la siguiente proposición:

Proposición 3.11: $J_k^*(t)$ 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 que hay en el intervalo (0,t] cuando se obtiene por primera vez A individuos pertenecientes a la clase I_0, habiéndose obtenido antes $N_k(t)$. 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), que analizaremos en el capítulo 4.

3.10. Relaciones entre los diversos parámetros

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

1. $$\boxed{t = \frac{A\hat{M}_1}{A(S-D) - \hat{M}_1}} \qquad (3.69)$$

En efecto:

$$\frac{\hat{M}_1}{D} = P_1^* = \frac{At}{A+t}\frac{S-D}{D} \rightarrow \frac{t}{A+t} = \frac{\hat{M}_1}{A(S-D)} \rightarrow \frac{t}{A} = \frac{\hat{M}_1}{A(S-D)-\hat{M}_1} \rightarrow t = \frac{A\hat{M}_1}{A(S-D)-\hat{M}_1}$$

2. $$\boxed{S-D = \frac{\hat{M}_1}{A} + \frac{\hat{M}_1}{t}} \qquad (3.70)$$

De (3.69) se deduce:

$$(S-D)A - \hat{M}_1 = \frac{A\hat{M}_1}{t} \rightarrow (S-D)A = \hat{M}_1 + \frac{A\hat{M}_1}{t} \rightarrow S-D = \frac{\hat{M}_1}{A} + \frac{\hat{M}_1}{t}$$

De la expresión de los P_r resulta:

3. $$\boxed{\frac{\hat{M}_{r+1}}{\hat{M}_r} = \frac{A+r}{r+1}\frac{t}{A+t}} \qquad (3.71)$$

En particular, para r=1, se obtiene:

$$\frac{\hat{M}_2}{\hat{M}_1} = \frac{(A+1)}{2}\,\frac{t}{A+t} \Rightarrow \frac{t}{A+t} = \frac{2\hat{M}_2}{\hat{M}_1\,(A+1)}$$

Podemos, de este modo, expresar p y q en función únicamente del parámetro A:

4. $$\boxed{q = \frac{t}{A+t} = \frac{2\hat{M}_2}{\hat{M}_1\,(A+1)}} \qquad (3.72)$$

5. $$\boxed{p = \frac{A}{A+t} = \frac{\hat{M}_1\,(A+1) - 2\hat{M}_2}{\hat{M}_1\,(A+1)}} \qquad (3.73)$$

6. Invirtiendo la igualdad anterior, se obtiene:

$$\boxed{\frac{1}{t} + \frac{1}{A} = \frac{\hat{M}_1}{2\hat{M}_2}\,\frac{A+1}{A}} \qquad (3.74)$$

7. Despejando 1/t en la relación anterior, queda:

$$\boxed{\frac{1}{t} = \frac{\hat{M}_1}{2\hat{M}_2} + \frac{\hat{M}_1 - 2\hat{M}_2}{2\hat{M}_2}\,\frac{1}{A}} \qquad (3.75)$$

que relaciona loa dos parámetros, A y t.

8. Teniendo en cuenta (3.70) y (3.74), resulta:

$$\boxed{S-D = \frac{\hat{M}_1}{t} + \frac{\hat{M}_1}{A} = \frac{\hat{M}_1^2}{2\hat{M}_2}\,\frac{A+1}{A}} \qquad (3.76)$$

que nos va a proporcionar uno de los estimadores de Chao (el estimador basado en los números de ocupación).

9. Para r=0, de (3.71) resulta:

$$\boxed{q = \frac{t}{A+t} = \frac{\hat{M}_1}{(S-D)\,A}} \qquad (3.77)$$

3.11. Estimación de la media de la población a partir de la media de la muestra

La distribución que regula el proceso es la binomial negativa de parámetros BN(A,p), con p=A/(A+t). Al ser desconocido el número de especies que no aparecen en la muestra, la hemos truncado en cero:

$$P_r = \binom{A+r-1}{A-1}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{A+t}\right)^A}\,,\quad r=1,2,\ldots$$

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:

$$\bar{X} = \frac{1}{D}\sum_{k=1}^{D} kM_k = \frac{\hat{f}}{D}$$

que es la media de la distribución truncada; luego:

$$\frac{\hat{f}}{D} = \frac{t}{1-P_0} = \frac{St}{D} \qquad (3.78)$$

lo que nos dice que T/S es un estimador de S, donde t es el tiempo de espera hasta obtener SA+T individuos en el instante t

3.12. Relaciones cuando se estima t por T/S

Si t=TS se tiene:

1. $$\boxed{q = \frac{t}{A+t} = \frac{\hat{T}}{SA+\hat{T}}} \qquad (3.79)$$

2. $$\boxed{p = \frac{A}{A+t} = \frac{SA}{SA+\hat{T}}} \qquad (3.80)$$

3. $$\boxed{\frac{\hat{M}_1}{\hat{T}} = pP_0} \qquad (3.81)$$

Esta relación resulta inmediata, ya que

$$P_0 = \frac{\hat{M}_1}{\hat{T}}\left(\frac{SA+\hat{T}}{SA}\right) = \frac{\hat{M}_1}{\hat{T}}\,\frac{1}{p}$$

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

Proposición 3.12: En el muestreo secuencial, el estimador de Good-Turing, M_1/T, es menor que P_0.

En efecto: Como $0<p<1$, $pP_0=M_1/T$ implica que M_1/T es menor que P_0.

4. $$\boxed{P_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}} \qquad (3.82)$$

Para demostrar esta propiedad, basta con tomar logaritmos en ambos miembros de la relación anterior:

$$\ln P_0 = \ln\left(\frac{\hat{M}_1}{\hat{T}}\right) - \ln p \;\Rightarrow\; A\ln p + \ln p = \ln\left(\frac{\hat{M}_1}{\hat{T}}\right) \;\Rightarrow$$

$$\Rightarrow\; (A+1)\ln p = \ln\left(\frac{\hat{M}_1}{\hat{T}}\right) \;\Rightarrow\; p = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{1}{A+1}}$$

Ahora bien, si tenemos en cuenta que $P_0 = p^A$, resulta:

$$P_0 = p^A = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}$$

Veamos un estimador del número, M_0, de especies desconocidas:

5. $$\boxed{\hat{M}_0 = S - D = \frac{\hat{M}_1\,(SA+\hat{T})}{\hat{T}A}} \qquad (3.83)$$

$$\ln P_0 = A\ln\frac{SA}{SA+\hat{T}}$$

6. Si t=T/S

$$\boxed{\frac{\hat{T}}{SA+\hat{T}} = \frac{2\hat{M}_2}{\hat{M}_1\,(A+1)}} \qquad (3.84)$$

7. $$\boxed{P_0 = \frac{\hat{M}_1}{\hat{T}} + \frac{\hat{M}_1}{SA}} \qquad (3.85)$$

En efecto:

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

3.13. Estimadores de S en función del parámetro A

Si tenemos en cuenta que $S=D/(1-P_0)$ y la relaciones (3.83) y (3.84), se obtienen dos expresiones, (3.87) y (3.88), para estimar S, que van a proporcionar una de las soluciones más interesantes de nuestro problema:

I. $$\boxed{\hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\frac{1}{A}} \qquad (3.86)$$

En efecto:

$$\frac{S-D}{S} = \frac{\hat{M}_1}{\hat{T}} + \frac{\hat{M}_1}{\hat{T}-\hat{M}_1}\frac{1}{A} \Rightarrow 1-\frac{D}{S} = \frac{\hat{M}_1}{\hat{T}} + \frac{\hat{M}_1}{SA} \Rightarrow 1-\frac{D}{S} = \frac{S\hat{M}_1A+\hat{T}\hat{M}_1}{\hat{T}SA} \cdot$$

$$\Rightarrow \hat{T}SA-\hat{T}DA = S\hat{M}_1A+\hat{T}\hat{M}_1 \Rightarrow SA(\hat{T}-\hat{M}_1) = \hat{T}DA+\hat{T}\hat{M}_1 \Rightarrow$$

$$\Rightarrow S = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\,\frac{1}{A}$$

II.
$$\boxed{S = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}}} \qquad (3.87)$$

Multiplicando por A en los dos miembros de (3.86), obtenemos:

III.
$$\boxed{SA = \frac{\hat{T}DA}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}} \qquad (3.88)$$

De (3.75) se obtiene, para S, el estimador:

IV.
$$\boxed{S = \frac{\hat{T}\hat{M}_1}{2\hat{M}_2} + \frac{\hat{T}(\hat{M}_1-2\hat{M}_2)}{2\hat{M}_2}\,\frac{1}{A}} \qquad (3.89)$$

3.14. Distribución de las abundancias relativas (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 3.13: 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 3.14: La variable aleatoria $\dfrac{\lambda_i}{\lambda_i + \sum_{j \neq i}^{S} \lambda_j} = \dfrac{\lambda_j}{\sum_{j=1}^{S} \lambda_j}$

sigue una distribución beta de parámetros: B[A,(S-1)A)].

Corolario 3.1: Según nuestro modelo, es $p_i = \dfrac{\lambda_i}{\sum_{j=1}^{S} \lambda_j}$, luego cada una de las p_i se distribuye según una beta de parámetros B[A,(S-1)A].

La proposición 3.14 y el corolario 3.1 se deducen inmediatamente de las propiedades de las funciones gamma y beta[7].

Al seguir las p_i una distribución beta de parámetros B[A,(S-1)A], se verifican las siguientes propiedades:

I. $$E[p_i] = \frac{A}{A+(S-1)A} = \frac{A}{SA} = \frac{1}{S} \qquad (3.96)$$

II. $$VAR(p_i) = \frac{A(S-1)A}{[SA]^2[SA+1]} = \frac{(S-1)A^2}{S^2A^2[SA+1]} = \frac{S-1}{S^2[SA+1]} \qquad (3.97)$$

III. $$E[p_i^2] = \frac{S-1}{S^2[SA+1]} + \frac{1}{S^2} = \frac{1}{S^2}\left[\frac{S-1}{SA} + \frac{SA+1}{SA+1}\right] = \frac{A+1}{S[SA+1]} \qquad (3.98)$$

IV. $$E\left[\sum_{k=1}^{S} p_i^2\right] = \frac{A+1}{SA+1} \qquad (3.99)$$

[7] Rao.C.R. (1985). "Linear Statistical Inference and its Applications". Cáp. 3, apdo. 3a. New Delhi.

3.15. Coeficiente de variación de Pearson de las p_i

El cuadrado del coeficiente de variación de Pearson viene definido por

$$\boxed{\gamma^2 = \frac{\sigma^2_{p_i}}{\left(\frac{1}{S}\right)^2}} \qquad (3.100)$$

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

$$\boxed{\sigma^2_{p_i} = \frac{S-1}{S^2\,[SA+1]}} \qquad (3.101)$$

luego:

$$\boxed{\gamma^2 = \frac{S-1}{SA+1}} \qquad (3.102)$$

Por otra parte

$$\gamma^2+1 = \frac{\cancel{S-1}}{SA+1} + 1 = \frac{\cancel{S+SA}}{SA+1} = \frac{S\,(A+1)}{SA+1}$$

Ahora bien, como

$$\gamma^2+1 = \frac{S(A+1)}{SA+1} = S\sum_{i=1}^{S} p_i^2$$

tenemos también la relación:

$$\boxed{\gamma^2+1 = S\sum_{i=1}^{S} p_i^2} \qquad (3.103)$$

de donde resulta:

$$\gamma^2 = \frac{1}{A} - \frac{\sum_{i=1}^{S} p_i^2}{A} \qquad (3.104)$$

En efecto:

$$\gamma^2 = \frac{S-1}{SA+1} \quad y \quad \sum_{i=1}^{S} p_i^2 = \frac{A+1}{SA+1}$$

Despejando SA+1 e igualando:

$$\frac{S-1}{\gamma^2} = \frac{A+1}{\sum_{i=1}^{S} p_i^2} \rightarrow \gamma^2 (A+1) = S\sum_{k=1}^{S} p_k^2 - \sum_{k=1}^{S} p_k^2 = \gamma^2 + 1 - \sum_{k=1}^{S} p_k^2$$

que da lugar a (3.104).

Si S es suficientemente grande y A finito, el último sumando de la expresión anterior tiende a cero. En efecto:

$$\gamma^2 = \frac{S-1}{SA+1} = \frac{S}{SA+1} - \frac{1}{SA+1} \approx \frac{1}{A}$$

puesto que, cuando S tiende a infinito

$$\frac{1}{SA+1} \rightarrow 0 \quad y \quad \frac{S}{SA+1} \rightarrow \frac{1}{A}$$

lo que nos permite enunciar la siguiente proposición:

Proposición 3.15: Cuando la población es suficientemente grande, el cuadrado del coeficiente de variación de Pearson es igual al inverso del parámetro de la distribución:

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

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.

Las conclusiones serán válidas para el modelo de Bernoulli.

3.16. Relaciones basadas en el coeficiente de variación de las p_k

La función generatriz de momentos nos permite obtener las siguientes relaciones:

1. $$E\left[\sum_{k=2}^{D} k(k-1)\,M_k\right] = \frac{St^2(A+1)}{A} \qquad (3.105)$$

Demostración:

$$E\left[\sum_{k=2}^{D} k(k-1)\,M_k\right] = \sum_{k=2}^{D} k(k-1)\,E(M_k) = \sum_{k=1}^{D} k^2 E(M_k) - \sum_{k=1}^{D} kE(M_k) =$$

$$= DE[U_k(t)^2] - DE[U_k(t)] = D\left(t+t^2+\frac{t^2}{A}\right)\frac{1}{1-P_0} - D\frac{t}{1-P_0} = St+St^2+\frac{St^2}{A} - St =$$

$$= St^2\left(\frac{A+1}{A}\right)$$

Si designamos por R_2 al primer miembro de (3.105), a partir de la segunda propiedad se obtiene la siguiente relación entre A, t y S:

$$S t^2 (A+1) = \hat{R}_2 A \qquad (3.106)$$

2. Si hacemos T=St, tenemos:

$$S = \frac{T^2}{\hat{R}_2} \frac{(A+1)}{A} \qquad (3.107)$$

que puede expresarse también en la forma:

$$S = \frac{\hat{T}^2}{\hat{R}_2} + \frac{\hat{T}^2}{\hat{R}_2} \frac{1}{A} \qquad (3.108)$$

En efecto:

$$S t^2 = \frac{T^2}{S} \Rightarrow \frac{T^2 (A+1)}{S} = \hat{R}_2 A$$

de donde se deduce inmediatamente (3.108).

Despejando SA en (3.108), se deduce:

3. $$SA = \frac{\hat{T}^2 (A+1)}{\hat{R}_2} \qquad (3.109)$$

Esto permite expresar $M_1/(SA)$ en función del parámetro A solamente:

4. $$\frac{\hat{M}_1}{SA} = \frac{\hat{M}_1 \hat{R}_2}{\hat{T}^2 (A+1)} \qquad (3.110)$$

Llevando este valor a (3.85), se obtienen el estimador insesgado de P_0 en las formas:

5. $$P_0 = \frac{\hat{M}_1}{T} + \frac{\hat{M}_1 \hat{R}_2}{\hat{T}^2 (A+1)} \qquad (3.111)$$

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

6. $$P_0 = \left(\frac{\hat{M}_1}{T}\right)^{\frac{1}{\gamma^2+1}} \qquad (3.112)$$

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, pero, si la distribución es muy heterogénea, la hipótesis de equiprobabilidad no es admisible.

Estos resultados nos permiten expresar S en función del coeficiente de variación de Pearson en el muestreo multinomial. Así obtenemos las expresiones:

7. $$S = \frac{D}{1-\left(\frac{\hat{M}_1}{T}\right)^{\frac{1}{\gamma^2+1}}} \qquad (3.113)$$

8. $$\hat{S} = \frac{TD}{T-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\gamma^2 \qquad (3.114)$$

Se trata, por tanto, ahora de estimar el coeficiente de variación de Pearson.

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.

9. $$\sum_{k=3}^{S} k(k-1)(k-2)E[M_k] = \frac{St^2A[A^2t+4(4A+5t)+8A+6t]}{A^3(1-P_0)} \quad (3.115)$$

10. Si t=T/S, resulta:

$$\sum_{k=3}^{S} k(k-1)(k-2)E[M_k] = \frac{T^2[A^2T+2A(12S-T)-16S-27T]}{A^2DS} \quad (3.116)$$

3.17. Superposición de procesos (Proceso combinado)

Tenemos una población formada por S procesos independientes de punto, que cumplen las condiciones de regularidad dadas por Cox[8] e Isham, que, en resumen son:

1) Los puntos de cada uno de los procesos componentes están adecuadamente espaciados, de modo que en cada subintervalo en que dividamos el intervalo (0,t] , hay al menos un punto de cada proceso con una probabilidad alta; esto significa que las épocas de renovación de cada proceso individual deben de ser extraordinariamente raras, de modo que el efecto cumulativo se deba a muchas causas pequeñas; por ello, para cada k, debe ser $P(X_k=r)$ pequeño y μ_X grande.

2) Ninguno de los procesos domina al resto.

[8] Cox, D.R. e Isham, V. "Point Processes". Chapman & Hall. Londres-1992.

La superposición de procesos de Polya truncados en el origen satisface estas condiciones. Entonces, el proceso combinado, N(t), puede aproximarse[9] por la suma de un gran número de variables independientes indicador, de modo que la distribución de N(t) tiende a una distribución de Poisson, cuya función generadora de probabilidades tiene por expresión:

$$\psi_{N(t)}(z) = E(z^{N(t)}) = \prod_{k=1}^{D} \left\{ 1 - \frac{\hat{M}_1}{D}(1-z) \right\} \qquad (3.117)$$

siendo

$$\frac{\hat{M}_1}{D} = P[U_k(t) = 1], \ \forall k.$$

de modo que, cuando D tiende a infinito, la expresión anterior tiende a la función generadora de probabilidades de una variable de Poisson de parámetro $\hat{M}_1 = \frac{A\,t\,(S-D)}{A+t}$

La familia de funciones generadoras de probabilidades del proceso combinado N(t) es:

$$\phi_{N(T)}(z) = e^{-\frac{At(S-D)}{A+t}(1-z)} \qquad (3.118)$$

Se trata de la función generatriz de probabilidades de un proceso de Poisson cuya razón es M_1=A(S-D)/(A+t). Corresponde a un proceso de Poisson no homogéneo.

[9] "Point Processes", capítulo 4, apartado 4.5. "D.R. Cox y V. Isham", Chapman & Hall, Londres 1992.

En efecto: Tenemos D procesos, $(U_k(t),\ (k=1,2,\ldots,D))$, independientes procedentes de una población en la que hay establecida una partición formada por S procesos. Los procesos $U_k(t)$ cumplen:

$$P[U_k(t)=1]=\frac{\hat{M}_1}{D},\ k=1,2,\ldots,D$$

Entonces, la función generatriz de probabilidades de la superposición es el producto

$$\Phi_{N(T)}(z)=\prod_{k=1}^{D}\left\{1-\frac{\hat{M}_1}{D}(1-z)\right\}=\left\{1-\frac{\hat{M}_1}{D}(1-z)\right\}^{D}=$$

$$=\left\{1+\frac{\hat{M}_1}{D}(z-1)\right\}^{D} \qquad (3.119)$$

que asintóticamente converge a la distribución de Poisson de razón M_1.

En efecto, tomando logaritmos, resulta:

$$\ln\{\Phi_{N(t)}(z)\}=D\ln\left\{1+\frac{\hat{M}_1}{D}(z-1)\right\}=\hat{M}_1(z-1)=\frac{At(S-D)}{A+t}(z-1) \qquad (3.120)$$

obteniéndose así la función generatriz de probabilidades de un proceso de Poisson de media $M_1=At(S-D)/(A+t)$.

$$\boxed{\Phi_{N(t)}(z)=e^{-\frac{At(S-D)}{A+t}(z-1)}} \qquad (3.121)$$

Acabamos de demostrar la siguiente proposición:

Proposición 3.16: Si N(t) es el proceso combinado, resultado de la superposición de D procesos de Polya, que cumplen las condiciones de regularidad, N(t) es un proceso de Poisson no homogéneo, que tiene como media:

$$M_1=At(S-D)/(A+t)$$

3.18. Tiempo de espera hasta la primera renovación

Siguiendo, como en el capítulo anterior, el razonamiento de Feller[10], aún cuando la superposición de procesos de Polya no sea un proceso de renovación, se puede obtener la distribución del tiempo de espera hasta la primera etapa de renovacón después de la etapa cero.

En efecto, la función de valor medio de cada proceso es:

$$m_k(t) = E[N_k(t)] = t = \frac{t}{\mu_X}$$

luego

$$\mu = \frac{t}{t} = 1 \Rightarrow \frac{1}{\mu_1} + \ldots + \frac{1}{\mu_D} = S = \frac{1}{\alpha} \qquad (3.122)$$

Supongamos un estado estable, es decir, los sucesos han estado ocurriendo durante mucho tiempo. Para el tiempo de espera W_k en la época de renovación más próxima en el k-ésimo proceso se tiene entonces aproximadamente

$$P[W_k(t) \leq t] = \frac{t}{\mu_k}$$

[10] "Introducción a la Teoría de Probabilidades y sus aplicaciones", volumen II. W. Feller, LIMUSA, México 1989.

El tiempo de espera para la primera renovación en el proceso cumulativo, W, es el más pequeño de los tiempos de espera W_k, por lo que

$$\left(1-\frac{t}{\mu_x}\right)\cdots\left(1-\frac{t}{\mu_s}\right)e^{-t\alpha}=e^{-\frac{t}{S}}$$

Tenemos, por tanto, que:

$$P[W>t]=1-P[W\leq t]\approx 1-(1-e^{-\frac{t}{S}})=e^{-\frac{t}{S}} \qquad (3.123)$$

Entonces

$$P[W\leq t]\approx 1-e^{-\frac{t}{S}} \qquad (3.124)$$

Ahora bien, si es N(t) el proceso cumulativo, se verifica:

$$P_0=\{P[N(t)=0]\}^S=e^{-\frac{tS}{S}}=e^{-t} \qquad (3.125)$$

luego, si nos fijamos en el comportamiento del proceso compuesto, N(t), obtenemos el estimador de P_0:

$$\boxed{\hat{P}_0=e^{-t}} \qquad (3.126)$$

donde t es el tiempo de espera necesario para conseguir, por primera vez, A individuos de la clase I_0.

La función de densidad del tiempo de espera hasta la primera renovación es:

$$\boxed{f_W(t)=e^{-t}} \qquad (3.127)$$

cuya esperanza matemática es:

$$\boxed{E[W] = 1} \qquad (3.128)$$

3.19. Estimador de Poisson

Si tenemos en cuenta que

$$\hat{M}_1 = \frac{A\,t\,(S-D)}{A+t}$$

podemos considerar el proceso combinado como proceso de Poisson no homgéneo, y aplicar la siguiente proposicón:

Proposición[11] **3.17:** Sea T_k el tiempo de espera hasta que tiene lugar el suceso k-ésimo en un proceso de Poisson no homogéneo, N(t), con función de valor medio continua, m(t). Con la condición de que N(t)=k, los k instantes $t_1,t_2,\ldots,t_k$ del intervalo (0,t] en que ocurren los sucesos, son variables aleatorias que tienen la misma distribución que si fueran los parámetros ordenados correspondientes a k variables aleatorias independientes $U_1,U_2,\ldots,U_k$, con función de distribución:

$$\boxed{F_{U_j}(u) = \frac{m(u)}{m(t)}\ ,\ 0 \le u \le t.} \qquad (3.129)$$

En nuestro caso, al ser m(t)=M_1, los M_1 instantes $t_1,t_2,\ldots,t_{M1}$ del intervalo (0,t] en que tienen lugar los sucesos son variables aleatorias con la misma distribución que los parámetros ordenados correspondientes a M_1 variables aleatorias $U_1,U_2,\ldots,U_{M1}$, con función de distribución:

[11] E.Parzen. "Procesos Estocásticos". Cap.4, apdo. 4.4. Paraninfo. Madrid 1972.

$$\boxed{F_{U_j}(u) = \frac{u\,(A+t)}{t\,(A+u)}\ ,\ 0 \le u \le t.} \qquad (3.130)$$

La función de densidad de U_j es, por consiguiente:

$$\boxed{f_{U_j}(u) = \frac{A\,(A+t)}{t\,(A+u)^2}\ ,\ 0 \le u \le t.} \qquad (3.131)$$

y la esperanza matemática:

$$\boxed{E[U_j] = \frac{A\,(A+t)}{t}\ln\left(\frac{A+t}{A}\right) - A} \qquad (3.132)$$

Teniendo en cuenta que

$$\frac{A\,(A+t)}{t}\ln\left(\frac{A+t}{A}\right) - A = -\ln(P_0)\ \frac{A+t}{t} - A \qquad (3.133)$$

y, dado que $E[U_j]=E[W]=1$, por (3.128), se verifica:

$$-\ln(P_0)\ \frac{A+t}{t} - A = 1 \qquad (3.134)$$

de donde se deduce:

$$-\ln(P_0) = \frac{t\,(A+1)}{A+t} \qquad (3.135)$$

y, por tanto:

$$P_0 = e^{-\frac{t\,(A+1)}{A+t}} \qquad (3.136)$$

Como, por otra parte, es

$$\frac{t\,(A+1)}{A+t} = \frac{2\,\hat{M}_2}{\hat{M}_1} \tag{3.137}$$

se obtiene finalmente, como estimador de P_0:

$$\boxed{P_0 = e^{-\frac{2\hat{M}_2}{\hat{M}_1}}} \tag{3.138}$$

Se trata de un estimador de la probabilidad de especies desconocidas, que nos lleva al estimador de Poisson (versión de Zeltermann):

$$\boxed{\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}}} \tag{3.139}$$

Este estimador de S no depende del parámetro de la distribución, por lo que podemos servirnos de él para estimar A.

3.20. Estimador dependiente del recubrimiento muestral

En el apartado 3.5, encontramos el estimador, para el modelo multinomial, de P_0:

$$\boxed{\hat{P}_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}} \tag{3.140}$$

que podemos igualar al de Poisson para obtener así la relación entre el estimador de P_0 del modelo de Poisson y la de aquellos otros modelos que se basan en el recubrimiento muestral:

$$e^{-\frac{2\hat{M}_2}{\hat{M}_1}} = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}} \qquad (3.141)$$

En concreto, (3.141) nos muestra la relación entre los estimadores para P_0 de Poisson y de Darroch. Podemos observar, en primer lugar, que, si A tiende a infinito, los estimadores de Poisson y de Darroch coinciden.

De (3.141) se deduce también:

$$-\frac{2\hat{M}_2}{\hat{M}_1} = \frac{A}{A+1}\ln\frac{\hat{M}_1}{\hat{T}} \qquad (3.142)$$

y de aquí resulta:

$$\frac{A+1}{A} = \frac{-\ln\frac{\hat{M}_1}{\hat{T}}}{\frac{2\hat{M}_2}{\hat{M}_1}} \qquad (3.143)$$

o lo que es equivalente:

$$\frac{A+1}{A} = \frac{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)}{2\hat{M}_2} \qquad (3.144)$$

De esta última relación se obtiene, para A:

$$\hat{A} = \frac{2\hat{M}_2}{\hat{M}_1 (\ln\hat{T} - \ln\hat{M}_1) - 2\hat{M}_2} \qquad (3.145)$$

y, para estimar 1/A, su inversa:

$$\hat{\gamma}^2 = \frac{1}{\hat{A}} = \frac{\hat{M}_1 (\ln\hat{T} - \ln\hat{M}_1) - 2\hat{M}_2}{2\hat{M}_2} \qquad (3.146)$$

Si estimamos 1/A mediante (3.146) y llevamos esta estimación a (3.86) y (3.87), se obtienen los estimadores más interesantes de S, para el modelo de muestreo por "esquema de contagio", junto al estimador de Poisson:

$$\hat{S}_{12} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1} \frac{\hat{M}_1 (\ln\hat{T} - \ln\hat{M}_1) - 2\hat{M}_2}{2\hat{M}_2} \qquad (3.147)$$

y

$$\hat{S}_{13} = \frac{D}{1 - \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{2\hat{M}_2}{\hat{M}_1 (\ln\hat{T} - \ln\hat{M}_1)}}} \qquad (3.148)$$

3.21. Otros estimadores dependientes del parámetro

Como estimador del cuadrado del coeficiente de variación de Pearson, 1/A, hemos obtenido:

$$\frac{1}{\hat{A}} = \frac{\hat{M}_1 (\ln T - \ln\hat{M}_1) - 2\hat{M}_2}{2\hat{M}_2}$$

Además de los estimadores seleccionados en el apartado 3.16, cuyas propiedades les dan preferncia sobre el resto, hemos encontrado otros tres estimadores de S, que dependen también del parámetro A, en concreto de 1/A (cuadrado del coeficiente de variación de Pearson de las p_k), válidos, por lo tanto, para situaciones no homogéneas.

Hemos comprobado la idea expuesta por Chao y Lee[12]: "*Si estimamos el número de clases sin estimar la variación entre las probabilidades de las clases, lo que obtendremos es una cota inferior que nos lleva a la situación de equiprobabilidad*".

En función de, estos tres estimadores son:

$$\hat{S} = \frac{T\hat{M}_1}{2\hat{M}_2} + \frac{T(\hat{M}_1 - 2\hat{M}_2)}{2\hat{M}_2}\,\frac{1}{A}\,, \quad \hat{S} = D + \frac{\hat{M}_1^2}{2\hat{M}_2}\,\frac{A+1}{A} \text{ y } \hat{S} = \frac{\hat{T}^2}{\hat{R}_2}\,\frac{A+1}{A}$$

y, en función de la estimación de A, quedan en la forma:

1. $$\hat{S}_{14} = \frac{T\hat{M}_1}{2\hat{M}_2} + \frac{T(\hat{M}_1 - 2\hat{M}_2)\,[\hat{M}_1(\ln T - \ln \hat{M}_1) - 2\hat{M}_2]}{4\hat{M}_2^2} \qquad (3.149)$$

2. $$\hat{S}_{15} = D + \frac{\hat{M}_1^3(\ln T - \ln \hat{M}_1)]}{4\hat{M}_2^2} \qquad (3.150)$$

3. $$\hat{S}_{16} = \frac{T^2\hat{M}_1(\ln T - \ln \hat{M}_1)}{2\hat{R}_2\hat{M}_2} \qquad (3.151)$$

[12] Anne Chao y Shen-Ming Lee. "Estimating the Number of Classes via Sample Coverage". Vol.87, nº 417. ASAJA. Marzo de 1992.

3.21.1. Error típico de los estimadores

Para obtener una expresión aproximada del error típico de estos estimadores, ponemos, tanto D como T, en función de los números de ocupación. S es así una función de $(M_1, M_2, \ldots, M_h)$, que es diferenciable.

Necesitamos la matriz de covarianzas de los M_k, que se obtiene a partir de la función generadora de probabilidades.

3.21.2. Cálculo de la varianza de los estimadores de S

Los números de ocupación (M_r) son variables aleatorias definidas a partir de la función indicador, cuya función generadora de momentos analizamos en el apartado 3.5, viendo sus propiedades fundamentales, que nos permiten obtener la matriz de covarianzas y, por tanto, una estimación de la varianza de los diversos estimadores de S, cuando son funciones diferenciables de los M_k. Las varianzas vienen dadas por

$$\sigma_S = \sum_{k=1} \sum_{h=1} \frac{\partial S}{\partial M_k} \frac{\partial S}{\partial M_h} \sigma_{kh} \qquad (3.152)$$

donde

$$\sigma_{kh} = \begin{cases} \hat{M}_k \left(1 - \frac{\hat{M}_k}{\hat{D}}\right) , & si\ k=h \\ -\frac{\hat{M}_k \hat{M}_h}{\hat{D}} , & si\ k \neq h \end{cases} \qquad (3.153)$$

En forma matricial, se puede expresar del siguiente modo:

$$\boxed{\sigma_S^2 = L\Sigma L'} \qquad (3.154)$$

siendo

$$\boxed{L = \left(\frac{\partial S}{\partial M_1}, \frac{\partial S}{\partial M_2}, \ldots, \frac{\partial S}{\partial M_h} \right)} \qquad (3.155)$$

L' es la traspuesta de L y $\Sigma = (\sigma_{kh})$ es la matriz de covarianzas de los M_k.

Para calcular estos parámetros, hay que considerar a D y a T como variables aleatorias dependientes de los M_k.

3.22. Resumen del capítulo

Se llega al proceso generalizado de Polya a través de la distribución mixta de Poisson, haciendo que las $\tilde{\lambda}_k$ se distribuyan según una gamma de parámetros (A,1/A). El proceso de Polya es el límite al que tiende el esquema de urnas de Polya. Se trata de un proceso puro de nacimiento, con tasa de nacimiento lineal, que corresponde al caso c=1 en el esquema de urnas de Polya, es decir, a un esquema de muestreo secuencial.

La distribución que regula el proceso es la binomial negativa de parámetros BN(A,q), con p=t/(A+t), truncada en cero:

$$P_r^* = \binom{A+r-1}{A-1} q^r p^A \frac{1}{1-p^A}, \quad r = 0,1,\ldots$$

A partir de la función generatriz de probabilidades se obtienen la media, los primeros momentos y las distribuciones de los números de ocupación. De las propiedades de estos últimos se obtiene una familia de estimadores de S:

$$\hat{S} = \frac{E[M_r]}{P_r} = \frac{D}{1-P_0}$$

demostrando que $D/(1-P_0)$ es un estimador insesgado de S.

El vector aleatorio, cuyas componentes son los números de ocupación, $M=(M_1,\ldots,M_D)$ se distribuye según una multinomial negativa, y, cuando se le condiciona por $M_1+\ldots+M_D=D$, sigue una multinomial de parámetros $M_\alpha(D,M_1/D,\ldots,M_D/D)$.

Resultados importantes son:

1. $\hat{P}_0 = \left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}$ es un estimador insesgado de P_0.

2. Dos estimadores insesgados de S son:

$$\hat{S} = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{A}{A+1}}} \quad y \quad \hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\frac{1}{A}$$

3. Las abundancias relativas son variables aleatorias con una distribución beta de parámetros $\beta[A,(S-1)A]$.

4. El cuadrado del coeficiente de variación de Pearson de las P_k es aproximadamente igual a 1/A.

5. La función de máxima verosimilitud de S es:

$$L(S) = \prod_{k=1}^{D} \frac{P_r^{M_r}}{1-P_0}$$

6. $D/(1-P_0)$ es un estimador insesgado de S.

El problema fundamental es entonces el de estimar el parámetro A. Para conseguir una estimación independiente de S, hemos

considerado el proceso combinado, que, por una parte, en virtud de la proposición de Cox[13] e Isham, tiende a un proceso de Poisson de media M_1. Así, hemos conseguido obtener la distribución del tiempo de espera hasta la primera renovación.

A esta misma distribución hemos llegado, utilizando el razonamiento que hace Feller[14].

Relacionando ambos resultados, se obtiene para P_0, como estimador independiente de A, el estimador de Zelterman:

$$\hat{P}_0 = e^{-\frac{2\hat{M}_2}{\hat{M}_1}}$$

que, además de proporcionarnos el estimador de Poisson para S:

$$\hat{S}_P = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}}$$

cuya varianza es:

$$Var(\hat{S}_P) = \frac{\hat{S}^2}{\sqrt{\hat{M}_1}}$$

nos permite obtener un estimador de A independiente de S:

$$\hat{A} = \frac{2\hat{M}_2}{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)-2\hat{M}_2}$$

[13] D.R.Cox & V.Isham. "Point Processes". Cap. 4, apdo. 4.5. Chapman & Hall.Ipswich 1980.

[14] W.Feller. "Introducción a la teoría de probabilidades y sus aplicaciones". Vol.II, Cap.11, apdo.5. LIMUSA. Méjico 1993.

Esta estimación de A nos proporciona los estimadores de S:

$$\hat{S} = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{2\hat{M}_2}{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)}}} \quad \text{y} \quad \hat{S} = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} + \frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\,\frac{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)-2\hat{M}_2}{2\hat{M}_2}$$

En cuanto a la relación con los otros modelos de muestreo, cuando A tiende a infinito, la binomial negativa tiende a la distribución de Poisson, que regula el muestreo completamente aleatorio con reemplazamiento.

El muestreo sin reemplazamiento es equivalente a tomar A puntos al azar del intervalo (0,1] y contar el número de puntos que hay en el subintervalo (0,q].

MUESTREO SIN REEMPLAZAMIENTO

IV MUESTREO SIN REEMPLAZAMIENTO

4.1. Proceso de Bernoulli

En el capítulo 3, definíamos el proceso $J_k(t)$, que sigue una distribución binomial de parámetros A y q, donde A es el opuesto del parámetro A que se obtiene en la binomial negativa. Denotamos por $J_k^*(t)$ a la distribución truncada en cero, y tenemos un proceso de Bernoulli, cuya función masa de probabilidad viene dada por:

$$P_r^* = \binom{A}{r} \frac{q^r (1-q)^{A-r}}{1-p^A} \quad , r=1,2,\ldots \tag{4.1}$$

que proporciona la probabilidad del número de individuos de la clase I_k en el intervalo (0,q]. $J_k(t)$ corresponde a un proceso puro de muerte, según veremos en el siguiente apartado, que nos va a permitir estudiar el problema del número de especies cuando el esquema de muestreo es sin reemplazamiento.

Este planteamiento es equivalente a tomar A puntos al azar del intervalo (0,1] y contar el número de puntos en el intervalo (0,q].

4.2. Proceso puro de muerte

Dada una población inicial de tamaño A>0, los individuos mueren con una cierta razón de proporcionalidad, pudiendo su número, en ocasiones, llegar a ser cero. Si el tamaño de la población es n, sea μ_n la proporción de muerte, que definimos del siguiente modo:

La probabilidad de que haya una muerte en el intervalo $(t, t+\Delta t)$ es $\mu_n \Delta t + o(\Delta t)$, y la probabilidad de que no haya ninguna muerte es $1-\mu_n \Delta t + o(\Delta t)$, y todas las restantes posibilidades tienen una probabilidad del orden de $o(\Delta t)$.

También vamos a suponer que la ocurrencia de una muerte en el intervalo $(t, t+\Delta t)$ es independiente del tiempo desde que tuvo lugar la última muerte.

Para tal proceso, vamos a deducir[1] las ecuaciones diferenciales como sigue. Admitiendo que las transiciones tienen lugar en los intervalos (0,t] y $(t, t+\Delta t)$ tenemos:

$$P_k(t+\Delta t) = [1-\mu_k \Delta t + o(\Delta t)]$$

$$P_n(t+\Delta t) = P_n(t)\,[1-\mu_k \Delta t + o(\Delta t)] + \mu_{n+1} \Delta t P_{n+1}(t) + o(\Delta t)\,, \quad n<A$$

dando

$$P'_A(t) = -\mu_A P_A(t)$$

$$P'_A(t) = \mu_n P_n(t) + \mu_{n+1} P_{n+1}(t)$$

En el caso particular en que $\mu_n = n$, la solución a estas ecuaciones toma la forma:

$$P_n(t) = \binom{A}{n} e^{-nt} (1-e^{-t})^{A-n}\,, \quad n \leq A$$

que es una binomial con parámetro B(A,q), siendo $q = e^{-t}$.

[1] Narayan Bhat.U.(1987). "Elements of Applied Stochastic Processes". Cap.7, apdo. 7. John Willy. New York.

La distribución del tamaño de la población en el instante t, puede obtenerse considerando a los miembros originarios de la población como iniciadores de una sucesión de Bernoulli.

Supongamos que cada miembro originario tiene una vida distribuida exponencialmente con media $1/\mu$. Entonces cada individuo de la población original tiene una duración de vida que está distribuida exponencialmente con media $1/\mu$, y, para cada individuo vivo en el instante t, la probabilidad de que muera en el intervalo de tiempo $(t, t+\Delta t)$ es $\mu\Delta t+o(\Delta t)$.

Suponiendo que la muerte de un individuo de la población es un suceso independiente, para n individuos vivos en el instante t, la distribución de probabilidad del número de muertes durante $(t, t+\Delta t)$ puede ponerse como sigue:

$$P\{\text{ninguno muera durante } (t, t+\Delta t)\} = [1-\mu\Delta t+o(\Delta t)]^n =$$

$$= 1-n\mu\Delta t+o(\Delta t)$$

$$P\{\text{alguien muera durante } (t, t+\Delta t)\} =$$

$$= n[\mu\Delta t+o(\Delta t)][1-\mu\Delta t+o(\Delta t)]^{n-1} =$$

$$= n\mu\Delta t+o(\Delta t)$$

y $$P\{\text{mueran k (k>1) durante } (t, t+\Delta t)\} =$$

$$\binom{n}{k} n[\mu\Delta t+o(\Delta t)]^k[1-\mu\Delta t+o(\Delta t)]^{n-k} = o(\Delta t).$$

Estas probabilidades conducen a los postulados de un proceso de muerte, en el que la probabilidad de una muerte antes del instante t es la probabilidad de que la vida de dicho individuo no pase de t y viene dada por $1-e^{-t}$, y la probabilidad de que un individuo no muera en el instante t se puede expresar como e^{-t}.

Entonces la distribución del proceso viene dada por:

$$P[J_k^*(t)=r]=\binom{A}{r}\frac{(e^{-t})^r(1-e^{-t})^{A-r}}{1-(1-e^{-t})^A} \quad , \; r\leq A \tag{4.2}$$

siendo $P_0=[1-e^{-t}]^A$. (4.3)

Si llamamos q a e^{-t}, resulta:

$$\boxed{P[J_k^*(t)=r]=\binom{A}{r}\frac{q^r p^{A-r}}{1-p^A} \quad , \; r=0,1,2,\ldots} \tag{4.4}$$

4.3. Modelo de muestreo

Para un número natural A y $a_r=(A-r)p_r$, r=0,1,2,...,A, el proceso (4.4) es equivalente a un proceso de muerte con tasa de crecimiento lineal, y el esquema de urnas correspondiente es el muestreo completamente aleatorio sin reemplazamiento.

Esta distribución se puede obtener también a partir de A+r procesos independientes de Poisson, cada uno de ellos con intensidad 1, en cuyo caso $T_{(r)}$ será el instante en que tiene lugar el r-ésimo suceso, o bien, $T_{(r)}$ será el número de procesos con al menos un elemento en el intervalo (0,t].

4.4. Función generatriz de probabilidades

La función generatriz de probabilidades de la distribución binomial que regula el proceso $\{J_k^*(t)\}$ es:

$$\boxed{\Psi_{J_k^*(t)}(z)=\frac{1}{1-P_0}(qz+1-q)^A \; , \; 0<q<1} \tag{4.5}$$

Los primeros momentos para la distribución sin truncar son:

1. $$E[J_k(t)] = Aq \tag{4.6}$$

2. $$Var[J_k(t)] = Apq \tag{4.7}$$

3. $$E[J_k(t)]^2 = Apq + A^2q^2 \tag{4.8}$$

4. $$I[J_k(t)] = p < 1 \tag{4.9}$$

Observamos cómo el índice de dispersión, para esta distribución, es menor que 1.

Los primeros 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

$$\frac{1}{1-P_0} = \frac{1}{1-p^A} = \frac{S}{D} \tag{4.10}$$

Se obtiene así:

1'. $$E[J_k^*(t)] = \frac{SAq}{D} \tag{4.11}$$

2'. $$Var[J_k^*(t)] = \frac{SAq}{D}\left(1+q(A-1)-\frac{SAq}{D}\right) \tag{4.12}$$

3'. $$E[J_k^*(t)]^2 = \frac{SAq}{D}(1+q(A-1)) \tag{4.13}$$

4'. $$I[J_k^*(t)] = 1+q(A-1)-\frac{SAq}{D} \tag{4.14}$$

4.5. Muestreo hipergeométrico y de Bernoulli

En el capítulo primero, vimos que, si la población tiende a infinito, la distribución hipergeométrica tiende a la binomial. Luego, cada proceso sigue una binomial, y se verifica, por tanto, la siguiente proposición:

Proposición 4.2: Si $J_1(t),\ldots,J_D(t)$ son independientes y siguen una distribución binomial B(A,q), la distribución del vector $(J_1(t),\ldots,J_D(t))$ condicionada por $\sum_{k=1}^{D} J_k(t) = T$ viene dada por:

$$P\left[J_1(t)=n_1, J_2(t)=n_2, \ldots \mid \sum_{k=1}^{D} J_k(t)=T\right] = \frac{\prod_{k=1}^{D}\binom{A}{n_k}}{\binom{DA}{T}} \qquad (4.15)$$

Se trata de la distribución hipergeométrica multivariada. La demostración de esta proposición es inmediata, ya que:

$$P\left[J_1(t)=n_1, J_2(t)=n_2, \ldots \mid \sum_{k=1}^{D} J_k(t)=T\right] = \frac{\prod_{k=1}^{D}\binom{A}{n_k} q^{n_k}(1-q)^{A-n_k}}{\binom{DA}{T} q^{T}(1-q)^{DA-T}} =$$

$$= \frac{\prod_{k=1}^{D}\binom{A}{n_k}}{\binom{DA}{T}}$$

4.6. Distribución en el muestreo

En el primer capítulo, vimos que, cuando el tamaño muestral es grande, la distribución muestral del muestreo completamente aleatorio sin reemplazamiento tiende a la multinomial de parámetros:

$$M(D, P_1, \ldots, P_D) \tag{4.16}$$

Se verifican, por tanto, las siguientes propiedaes:

1. $$E[M_r] = \hat{M}_r = SP_r = DP_r^* \tag{4.17}$$

2. $$Var[M_r] = M_r\left(1 - \frac{M_r}{D}\right) \tag{4.18}$$

3. $$E[M_r(M_r-1)] = \frac{D(D-1)P_r^2}{(1-P_0)^2} \tag{4.19}$$

Teniendo en cuenta la primera propiedad, resulta:

4. $$\boxed{\frac{\hat{M}_{r+1}}{\hat{M}_r} = \frac{(A-r)q}{(r+1)p}} \tag{4.20}$$

En particular, para r=1, se tiene:

5. $$\frac{t}{A} = \frac{2\hat{M}_2}{\hat{M}_1(A-1)} \tag{4.21}$$

y, para r=0, resulta:

6. $$t = \frac{\hat{M}_1}{S-D} \tag{4.22}$$

7. $$\boxed{\hat{R}_2 = \frac{T^2(A-1)}{SA}} \tag{4.23}$$

Demostración:

$$\hat{R}_2 = E\left[\sum_{k=2}^{S} k(k-1)M_k\right] = E\left[(J_{kA}(t))^2\right] - E\left[J_{kA}(t)\right] = D\left[\frac{T^2}{SD} - \frac{T^2}{DSA}\right] = \frac{T^2}{S} - \frac{T^2}{SA}$$

$$\Rightarrow \hat{R}_2 = \frac{T^2}{S}\,\frac{A-1}{A}$$

Si tenemos en cuenta que $\hat{M}_r = SP_r = DP_r^*$, podemos enunciar la siguiente proposición:

Proposición 4.3: Si G(k) es el número de especies que figuran k veces en la muestra, la distribución de (G(1),...,G(D)) es aproximadamente multinomial $M_\alpha(D;M_1/D,\ldots,M_D/D)$.

4.7. Estimadores de P_0

I. Estimador de P_0:

$$\boxed{P_0 = \frac{\hat{M}_1}{T} - \frac{\hat{M}_1}{SA}} \tag{4.24}$$

Esta relación es también inmediata:

$$\frac{p}{q} = \frac{SA-T}{T} \rightarrow M_0 = \frac{\hat{M}_1}{A}\frac{SA-T}{T} \rightarrow P_0 = \frac{\hat{M}_1}{T} - \frac{\hat{M}_1}{SA}$$

Despejando S en la expresión anterior, se obtiene:

II. $$\boxed{P_0 = \frac{\hat{M}_1}{T}p} \qquad (4.25)$$

Veámoslo:

$$P_0 = \frac{\hat{M}_1}{T}\left(1 - \frac{T}{SA}\right) = \frac{\hat{M}_1}{T}p$$

La relación anterior nos permite enunciar la siguiente proposición:

Proposición 4.4: En el muestreo completamente aleatorio sin reemplazamiento, el estimador de Good-Turing, M_1/T, es mayor que P_0.
Es suficiente tener en cuenta que $0<p<1$ en la relación (4.25).

III. $$\boxed{P_0 = \left(\frac{\hat{M}_1}{T}\right)^{\frac{A}{A-1}}} \qquad (4.26)$$

En efecto:

$$P_0 = p^A = \frac{\hat{M}_1}{T}p \rightarrow (A-1)\ln p = \ln\left(\frac{\hat{M}_1}{T}\right) \rightarrow p = \left(\frac{\hat{M}_1}{T}\right)^{\frac{1}{A-1}} \rightarrow P_0 = \left(\frac{\hat{M}_1}{T}\right)^{\frac{A}{A-1}}$$

4.8. Estimadores de S en función del parámetro

De la expresión anterior se deduce el estimador de S:

A) $$S = \frac{D}{1-\left(\frac{\hat{M}_1}{T}\right)^{\frac{A}{A-1}}} \qquad (4.27)$$

Despejando S en (4.24), se obtiene:

B) $$S = \frac{TD}{T-\hat{M}_1} - \frac{T\hat{M}_1}{T-\hat{M}_1}\,\frac{1}{A} \qquad (4.28)$$

C) $$S = \frac{T\hat{M}_1}{2\hat{M}_2} - \frac{T(\hat{M}_1-2\hat{M}_2)}{2\hat{M}_2}\,\frac{1}{A} \qquad (4.29)$$

Como q/(1-q)=t/(SA+T), resulta:

$$\frac{2\hat{M}_2}{\hat{M}_1} = \frac{(A-1)T}{SA+T}$$

de donde:

$$S = \frac{(A-1)T\hat{M}_1}{2\hat{M}_2 A} - \frac{T}{A} = \frac{T\hat{M}_1}{2\hat{M}_2} - \left(\frac{\hat{M}_1}{2\hat{M}_2}-1\right)\frac{T}{A} = \frac{T\hat{M}_1}{2\hat{M}_2} - \frac{T(\hat{M}_1-2\hat{M}_2)}{2\hat{M}_2}\,\frac{1}{A}$$

Despejando S en la expresión (4.23), resulta:

D)
$$\hat{S} = \frac{T^2}{\hat{R}_2}\,\frac{A-1}{A} \qquad (4.30)$$

E)
$$S = D + \frac{\hat{M}_1^2}{2\hat{M}_2}\,\frac{A-1}{A} \qquad (4.31)$$

En efecto, en las relaciones anteriores (4.21) y (4.22), despejando q/p e igualando, se obtiene:

$$\frac{q}{p} = \frac{\hat{M}_1}{(S-D)A} = \frac{2\hat{M}_2}{\hat{M}_1(A-1)} \rightarrow S-D = \frac{\hat{M}_1^2}{2\hat{M}_2}\,\frac{A-1}{A}$$

de donde se deduce inmediatamente 4.31.

4.9. Estimadores de S

Como , en este modelo, E[M_1] tiene la misma expresión que en el modelo secuencial, podemos estimar A por medio de (4.39), que proporciona, para el muestreo sin reemplazamiento, las mismas expresiones, para los estimadores de S, que en el muestreo secuencial:

I.
$$\hat{S}_{17} = \frac{D}{1-\left(\frac{\hat{M}_1}{T}\right)^{\frac{2\hat{M}_2}{\hat{M}_1(\ln T-\ln\hat{M}_1)}}} \qquad (4.36)$$

II.
$$\hat{S}_{18} = \frac{TD}{T-\hat{M}_1} + \frac{T\hat{M}_1}{T-\hat{M}_1}\,\frac{\hat{M}_1(\ln T-\ln\hat{M}_1)}{2\hat{M}_2} \qquad (4.37)$$

III.
$$\hat{S}_{19} = \frac{T\hat{M}_1}{2\hat{M}_2} + \frac{T(\hat{M}_1-2\hat{M}_2)}{2\hat{M}_2}\,\frac{\hat{M}_1(\ln T-\ln\hat{M}_1)}{2\hat{M}_2} \qquad (4.34)$$

IV. $$\hat{S}_{20} = D + \frac{\hat{M}_1^2}{2\hat{M}_2} + \frac{\hat{M}_1^2}{2\hat{M}_2}\,\frac{\hat{M}_1\,(\ln T - \ln \hat{M}_1)}{2\hat{M}_2} \qquad (4.38)$$

V. $$\hat{S}_{21} = \frac{T^2}{\hat{R}_2} + \frac{T^2}{\hat{R}_2}\,\frac{\hat{M}_1\,(\ln T - \ln \hat{M}_1)}{2\hat{M}_2} \qquad (4.39)$$

Aunque hemos obtenido las mismas expresiones, el resultado es diferente. En efecto, al estimar el parámetro, se obtiene para A un valor negativo, por lo que deberíamos haber tomado B=-A como parámetro de la binomial. Por este motivo, si mantenemos el mismo nombre para el parámetro, resulta la misma expresión, pero los segundos términos de los cuatro últimos estimadores serán negativos. El mismo efecto se obtiene en el estimador (4.36).

4.10. Resumen del capítulo

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.

Cuando el tamaño muestral es suficientemente grande, los dos tipos de muestreo completamente aleatorio (con reemplazamiento y sin reemplazamiento), tienen la misma distribución muestral.

Podemos distinguir este modelo si observamos que, en él, el estimador de Good-Turing, M_1/T, es mayor que P_0, ya que

$$P_0 = \frac{\hat{M}_1}{T}\,p$$

En la práctica, el criterio que adoptamos es el de comparar los estimadores de las probabilidades desconocidas de Good-Turing y de Poisson-Zelterman:

Si $\frac{\hat{M}_1}{T} > e^{-\frac{2\hat{M}_2}{\hat{M}_1}}$ tenemos el muestreo completamente aleatrio sin reemplazamiento.

Se puede contrastar este criterio con el de Katz, que consiste en estimar el índice de dispersión de la población a partir del índice de dispersión de la muestra.

Se obtienen, para estimar S, las mismas expresiones del muestreo secuencial, aunque el resultado es diferente. En este modelo, el valor de S será inferior al valor que se obtiene bajo la hipótesis de homogeneidad.

MUESTREO CON REEMPLAZAMIENTO

V MUESTREO CON REEMPLAZAMIENTO

5.1. Relación entre la binomial negativa y la distribución de Poisson

Cuando A tiende a infinito, es decir cuando la distribución de las p_k es homogénea, la binomial negativa tiende a la distribución de Poisson, según vemos en la siguiente proposición:

Proposición[1] **5.1:** Sea BN(A,r,p) la distribución binomial negativa. Cuando $A \to \infty$ y $Aq \to \lambda \Rightarrow BN(k,A,p) \to e^{-\lambda}\frac{\lambda^k}{k!}$

En efecto:

La función generadora de probabilidades de la binomial negativa es:

$$\phi(z) = \left(\frac{p}{1-qz}\right)^A \qquad (5.1)$$

Sea λ fijo. Cuando $p \to 1$, $q \to 0$ y $A \to \infty$ de modo que $q \sim \lambda/A$ resulta

[1] W. Feller "Introducción a la Teoría de Probabilidades y sus Aplicaciones". Vol. I, Cap. XI, 6. LIMUSA. México, 1993.

$$\left(\frac{p}{1-qz}\right)^A = \left(\frac{1-\frac{\lambda}{A}}{1-\frac{\lambda z}{A}}\right)^A \qquad (5.2)$$

Tomando logaritmos, el segundo miembro tiende a $e^{-\lambda(1-z)}$, que es la función generadora de probabilidades de la distribución de Poisson, cuya función masa de probabilidades truncada en cero es:

$$\boxed{P_k^* = \frac{e^{-\lambda}\lambda^k}{k!\,(1-e^{-\lambda})}} \qquad (5.3)$$

5.2. Función generatriz de probabilidades

La función generatriz de probabilidades de la distribución de Poisson truncada que regula el proceso $\{N_k^*(t)\}$ es:

$$\boxed{\Psi_{N_k^*(t)}(z) = \frac{1}{1-P_0} e^{-\lambda(1-s)} \;,\; \lambda > 0 \;,\; s \in \mathbb{R}.} \qquad (5.4)$$

Los primeros momentos para la distribución sin truncar son:

1. $$E[N_k(t)] = \lambda \qquad (5.5)$$

2. $$Var[H_k(t)] = \lambda \qquad (5.6)$$

3. $$E[N_k(t)]^2 = \lambda - \lambda^2 \qquad (5.7)$$

4. $$I[N_k(t)] = 1 \tag{5.8}$$

Observamos cómo el índice de dispersión, para esta distribución, es menor a 1.

Los primeros 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

$$\frac{1}{1-P_0} = \frac{1}{1-p^A} = \frac{S}{D} \tag{5.9}$$

Se obtiene así:

1'. $$E[N_k^*(t)] = \frac{\lambda S}{D} \tag{5.10}$$

2'. $$Var[N_k^*(t)] = \frac{1}{1-P_0}\left(1+\lambda-\frac{\lambda}{1-P_0}\right) \tag{5.11}$$

3'. $$E[N_k^*(t)]^2 = \frac{\lambda+\lambda^2}{1-P_0} \tag{5.12}$$

4'. $$I[N_k^*(t)] = 1+\lambda-\frac{\lambda}{1-P_0} \tag{5.13}$$

5.3. Propiedades de los números de ocupación

La distribución que regula el muestreo con reemplazamiento es la distribución de Poisson de parámetro λ y se cumplen las siguientes propiedades:

1. $$E(M_r) = \sum_{k=1}^{S} E\{I[N_k(T) = r]\} = \frac{D\lambda^r e^{-\lambda}}{r!\,(1-e^{-\lambda})} \tag{5.14}$$

En particular:

2. $$\boxed{E[M_1] = \frac{D\lambda e^{-\lambda}}{1-e^{-\lambda}}} \qquad (5.15)$$

3. $$\boxed{E[M_2] = \frac{D\lambda^2 e^{-\lambda}}{2\,(1-e^{-\lambda})}} \qquad (5.16)$$

De (5.15) se deduce:

4. $$\boxed{1-e^{-\lambda} = \frac{D\lambda}{D\lambda+\hat{M}_1}} \qquad (5.17)$$

En efecto:

$$\hat{M}_1 - \hat{M}_1 e^{-\lambda} = D\lambda e^{-\lambda} \Rightarrow e^{-\lambda} = \frac{\hat{M}_1}{D\lambda+\hat{M}_1} \Rightarrow 1-e^{-\lambda} = \frac{D\lambda}{D\lambda+\hat{M}_1}$$

De la propiedad anterior resulta:

5. $$\boxed{S = D + \frac{\hat{M}_1}{\lambda}} \qquad (5.18)$$

Es inmediato, ya que

$$S = \frac{D}{1-e^{-\lambda}} = \frac{D\lambda+\hat{M}_1}{\lambda} = D + \frac{\hat{M}_1}{\lambda}$$

De (5.16) se deduce:

6. $$\boxed{1-e^{-\lambda}=\frac{D\lambda^2}{D\lambda^2+2\hat{M}_2}} \qquad (5.19)$$

Veámoslo:

$$2\hat{M}_2-2\hat{M}_2 e^{-\lambda}=D\lambda^2 e^{-\lambda} \Rightarrow e^{-\lambda}=\frac{2\hat{M}_2}{D\lambda^2+2\hat{M}_2} \Rightarrow 1-e^{-\lambda}=\frac{D\lambda^2}{D\lambda^2+2\hat{M}_2}$$

Por otra parte:

7. $$\boxed{\lambda=\frac{2\hat{M}_2}{\hat{M}_1}} \qquad (5.20)$$

En efecto, igualando (5.17) y (5.19), se obtiene (5.20).

8. $$\boxed{e^{-\lambda}=e^{\frac{-2\hat{M}_2}{\hat{M}_1}}} \qquad (5.21)$$

5.4. Estimadores de S

Este modelo se da cuando A tiende a infinito, por lo que el coeficiente de variación de Pearson de las p_k se anula. Luego, como estimadores de S, tendremos los estimadores homogéneos.

Así, de las relaciones (5.8), (5.20) y (5.21) se obtienen los siguientes estimadores para S:

I. $$\boxed{\hat{S}_{22}=\frac{T\hat{M}_1}{2\hat{M}_2}} \qquad (5.22)$$

puesto que:

$$E[N_k(t)]=\frac{\lambda S}{D}=\frac{T}{D} \Rightarrow \lambda S=T \Rightarrow \lambda=\frac{T}{S}$$

Entonces:

$$\frac{2\hat{M}_2}{\hat{M}_1} = \frac{T}{S} \Rightarrow S = \frac{T\hat{M}_1}{2\hat{M}_2}$$

II. $$\boxed{\hat{S}_{23} = D + \frac{\hat{M}_1^2}{2\hat{M}_2}} \qquad (5.23)$$

En efecto:

$$\hat{S} = D + \frac{\hat{M}_1}{\hat{\lambda}} = D + \frac{\hat{M}_1^2}{2\hat{M}_2}$$

III. $$\boxed{\hat{S}_{24} = \frac{T^2}{\hat{R}_2}} \qquad (5.24)$$

Veámoslo:

$$E[R_2] = E[U_k(t)^2] - E[U_k(t)] = D(\lambda+\lambda^2)\frac{S}{D} - \lambda\frac{S}{D} = D\lambda^2\frac{S}{D} = \frac{T^2}{S} \Rightarrow S = \frac{T^2}{\hat{R}_2}$$

Si utlizamos la esperanza matemática de M_1 en el estimador de máxima verosimilitud, obtenemos:

$$\boxed{\hat{P}_0 = e^{-\frac{T}{S}} = \frac{\hat{M}_1}{T}}$$

de donde:

IV. $$\boxed{\hat{S}_{25} = \frac{\hat{D}}{\hat{C}} = \frac{T\hat{D}}{T-\hat{M}_1}} \qquad (5.25)$$

Se trata una vez más del estimador de Darroch, que está basado en el estimador del recubrimiento muestral de Good-Touring.

5.4.1. Estimador de Poisson

Se conoce como estimador de Poisson el estimador de la forma

$$\hat{S} = \frac{D}{1-\hat{p}_0(F)} \tag{5.26}$$

donde $\hat{p}_0(F)$ es un estimador de "la probabilidad de que una variable aleatoria con la distribución mixta de Poisson es igual a cero"

Para esta distribución, el estimador de "la probabilidad de que el número de especies sea cero" es

$$\hat{P}_0 = e^{-\lambda} \tag{5.27}$$

Si utilizamos (5.20) para estimar t, obtenemos el estimador de P_0:

$$\hat{P}_0 = e^{-\frac{2\hat{M}_2}{\hat{M}_1}} \tag{5.28}$$

Llevando este valor al estimador de Poisson, se obtiene:

$$\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}} \tag{5.29}$$

Se trata del estimador de Poisson-Zelterman, cuya varianza puede obtenerse utilizando la expresión (3.60).

5.5. Distribución de Maxwell-Boltzman

La binomial negativa, en las condiciones indicadas, cuando A tiende a infinito, converge a la distribución de Poisson de parámetro λ.

Si condicionamos la distribución de $(N_1(t),\ldots,N_D(t))$ por $\sum_{k=1}^{D} N_k(t) = T$, se obtiene la distribución multinomial:

$$P = P[N_k(t) = n_k, k=1,2,\ldots,D \,/\, \sum_{k=1}^{S} N_k(t) = T] =$$

$$= \frac{T!}{\prod_{k=1}^{D} n_k!} \prod_{k=1}^{D} p_k^{n_k},\ n_k \neq 0 \ y\ \lambda = p_k t$$

Si se admite la hipótesis de homogeneidad: $p_1=p_2=\ldots=p_S=1/S$, siendo el número de muestras con D clases diferentes, en el capítulo segundo vimos que la distribución correspondiente al muestreo multinomial es la de Maxwell-Boltzman, que proporciona, como estimación de S, la solución S_3 de la ecuación

$$D = \hat{S}\left[1-e^{-\frac{T}{\hat{S}}}\right] \qquad (5.31)$$

cuya varianza, según vimos es:

$$var(\hat{S}_3) = \frac{\hat{S}_3}{e^{\frac{T}{\hat{S}_3}} - \frac{T}{\hat{S}_3} - 1} \qquad (5.32)$$

(5.32)W.W. Esty[2] estudió la eficiencia de los estimadores de Darroch y de máxima verosimilitud, analizando los estimadores de los respectivos recubrimientos muestrales:

$$\overline{C}=1-e^{-\frac{T}{S}} \quad y \quad \ddot{C}=1-\frac{M_1}{T}$$

concluyendo que el estimador de Good-Touring es "*bastante eficiente cuando se le compara con el de máxima verosimilitud*".

6.6. Resumen del capítulo

La distribución que regula el muestreo completamente aleatorio con reemplazamiento, cuando la población es infinita, es la distribución de Poisson de parámetro λ, siendo $q \sim \lambda/A$, cuando p tiende a 1 y A tiende a infinito.

La distribución para particiones correspondiente es la distribución de Maxwell-Boltzman, que da lugar al estimador de máxima verosimilitud, S_1, solución de la ecuación:

$$\hat{S}_3 = \frac{\hat{D}}{1-e^{-\frac{T}{\hat{S}_3}}}$$

Como A tiende a infinito, se verifica la hipótesis de equiprobabilidad, y los estimadores de S son todos homogéneos. Los estimadores de menor varianza son el de Darroch y el de Poisson, que, en este modelo, deben dar estimaciones muy próximas.

[2] Esty, W. W. "The Annals of Statistics", Vol. 14, N° 3, 1257-1260; 1966.

MUESTREO DE BOSE-EINSTEIN Y DE EWENS

VI MUESTREO DE BOSE-EINSTEIN Y DE EWENS

6.1. Convergencia de la binomial negativa a la serie logarítmica (Modelo de Ewens)

El modelo paramétrico presenta tres casos extremos, uno de los cuales es aquel en que A tiende a infinito, que da origen a la distribución de Maxwell-Boltzman, que estudiamos con detalle en el segundo y quinto capítulo.

Los otros casos extremos son aquellos en que A tiende a cero ó A=1, y que van a dar lugar a las distribuciones de Ewens y Bose-Einstein.

Proposición[1] **6.1:** Sea BN(r;A,p) la distribución binomial negativa truncada en cero.

Cuando $A \to 0 \Rightarrow BN(r;A,p) \to -\alpha \frac{q^r}{r}$, $r=1,2,3,\ldots$

donde $\alpha = \frac{1}{\log(1-q)}$

En efecto:

$$BN(r;A,p) = \frac{A(A+r-1)^{[r-1]}}{r!} p^A q^r \frac{1}{1-p^A}$$

[1] A. Stuart & J. Keith Ord. "Kendall's Advanced Theory of Statistics". Vol. 1, Cap. 5, Apdo. 5.28. Charles Griffin & Co. Ltd. Belfast 1983.

Teniendo en cuenta que

$$\lim_{A\to 0} \frac{A}{1-p^A} = \frac{-1}{\log(p)} \qquad (6.1)$$

para lo cual basta con aplicar la regla de l'Hôpital, resulta:

$$\lim_{A\to 0} BN(r;A,p) = -\alpha \frac{q^r}{r} \quad , \quad r=1,2,\ldots \qquad (6.2)$$

Se trata de la serie logarítmica, cuya función masa de probabilidad es:

$$\boxed{P_r = -\alpha \frac{q^r}{r} \quad , \quad r=1,2,\ldots} \qquad (6.3)$$

y cuya función generatriz de probabilidades viene dada por:

$$\boxed{P(z) = \frac{1}{\log(1-q)} \log(1-qz)} \qquad (6.4)$$

El modelo de Ewens se caracteriza por tener como distribución a la "serie logarítmica".

Los primeros momentos respecto al origen y la varianza son:

$$a_1 = \frac{-\alpha q}{p} \; ; \; a_2 = \frac{-\alpha q}{p^2} \; ; \; a_3 = \frac{-\alpha q(1+q)}{p^3} \; ; \; a_4 = \frac{-\alpha q(1+4q+q^2)}{p^4}$$

$$\mu_2 = Var(X) = -(1+\alpha q)\frac{\alpha q}{p^2} \; ; \; \mu_3 = -\alpha q(1+q+3\alpha q+2\alpha^2) \; ;$$

$$\mu_4 = \frac{-\alpha q(1+4q+q^2)+4\alpha q(1-q)+6\alpha^2 q^2}{p^4}$$

A) Por una parte, tenemos que la esperanza de M_r es:

$$\boxed{\hat{M}_r = -\frac{D\alpha}{r} q^r} \qquad (6.5)$$

En particular:

$$E(M_1) = -\alpha D q$$

$$E(M_2) = -\alpha D \frac{q^2}{2}$$

de donde, dividiendo miembro a miembro, resulta:

$$\boxed{\frac{2\hat{M}_2}{\hat{M}_1} = q} \qquad (6.6)$$

Como $-\alpha = \frac{\hat{M}_1}{Dq} = \frac{\hat{M}_1^2}{2\hat{M}_2 D}$

se deduce:

$$\boxed{P_r = \frac{\hat{M}_1}{Dr}\left(\frac{2\hat{M}_2}{\hat{M}_1}\right)^{r-1}} \qquad (6.7)$$

B) Si utilizamos la media muestral para estimar la media de la población:

$$-\alpha \frac{q}{p} = \frac{T}{D} \quad \Rightarrow \quad -\alpha = \frac{Tp}{Dq} = \frac{T(\hat{M}_1 - 2\hat{M}_2)}{D\hat{M}_1} \qquad (6.8)$$

Entonces:

$$\hat{M}_1 = D\frac{Tp}{Dq}q = Tp \quad \Rightarrow \quad p = \frac{\hat{M}_1}{T} \quad y \quad q = \frac{T-\hat{M}_1}{T}$$

con lo cual:

$$\boxed{P_r = \frac{\hat{M}_1}{Dr}\left(\frac{T-\hat{M}_1}{T}\right)^{r-1}} \qquad (6.9)$$

Luego se obtiene así el estimador de S:

$$\boxed{S_{26} = \frac{T^2\hat{M}_1}{(T-\hat{M}_1)^2}\ln\frac{T}{\hat{M}_1}} \qquad (6.10)$$

6.2. Distribución de Bose-Einstein

La distribución de Bose-Einstein se obtiene para el valor de A=1, en cuyo caso se verifica:

Proposición 6.2: Sea BN(r;A,p) la distribución binomial negativa truncada en cero. Cuando $A \to 1 \Rightarrow BN(r;A,p) \to pq^r$ donde $q = \frac{t}{A+t}$

En efecto: Basta con hacer A=1 y se obtiene:

$$\boxed{P_r = p(1-p)^{r-1}} \qquad (6.11)$$

Se trata de una distribución geométrica, que caracteriza al modelo de Bose-Einstein. Es, por tanto, la distribución de los

tiempos entre llegadas correspondientes a un proceso de Bernoulli cuya distribución es:

$$P_r = \binom{A}{r} \frac{p^r (1-p)^{A-r}}{1-q^A} \quad , r=1,2,\ldots \tag{6.12}$$

que, al ser desconocidas las especies que no figuran en la muestra, la truncamos en cero.

Esta distribución nos proporciona la probabilidad del número de individuos de la clase I_0 que quedan en el instante t=T.

Si hallamos la esperanza matemática de M_1 y M_2 resulta:

$$E(M_1) = Dp \quad \rightarrow \quad p = \frac{\hat{M}_1}{D} \quad y \quad q = \frac{D-\hat{M}_1}{D}$$

$$E(M_2) = p(1-p)$$

Dividiendo miembro a miembro las dos igualdades anteriores e igualando, se obtiene:

$$\frac{\hat{M}_2}{\hat{M}_1} = 1-p$$

de donde:

$$P_r = \frac{\hat{M}_1 - \hat{M}_2}{\hat{M}_1} \left(\frac{\hat{M}_2}{\hat{M}_1} \right)^{r-1} \tag{6.13}$$

y

$$P_r = \frac{\hat{M}_1}{D} \left(\frac{D-\hat{M}_1}{D} \right)^{r-1} \tag{6.14}$$

De (6.13) y (6.14) se obtienen los estimadores de S:

$$\boxed{S_{27} = \frac{D^2}{D-\hat{M}_1}} \qquad (6.15)$$

y

$$\boxed{S_{28} = \frac{D\hat{M}_1}{\hat{M}_2}} \qquad (6.16)$$

6.3. Resumen del capítulo

Las distribuciones de Bose-Einstein y de Ewens son, en realidad dos casos particulares(extremos) del modelo paramétrico.

Cuando A=0, se obtiene la serie logarítmica, cuya distribución caracteriza al modelo de Ewens, y viene dada por:

$$P_r = -\alpha \frac{q^r}{r} \quad , \quad r=1,2,\ldots$$

para la que hemos obtenido el estimador de S:

$$S_{11} = \frac{T^2\hat{M}_1}{(T-\hat{M}_1)^2} \ln \frac{T}{\hat{M}_1}$$

La distribución de Bose-Einstein se obtiene para otro caso límite del parámetro, A=1, y se caracteriza por tender a la distribución geométrica:

$$P_r = p\,(1-p)^{r-1}$$

MODO DE ACTUAR EN EL MODELO PARAMÉTRICO

VII MODO DE ACTUACIÓN EN EL MODELO PARAMÉTRICO

7.1. Análisis de resultados

La relación que existe entre los estimadores de Poisson y de Darroch nos permite dar un criterio unificado al problema de las especies cuando la población es infinita.

En efecto, el análisis de los tres modelos estudiados y el de los casos extremos nos ha permitido llegar a las siguientes conclusiones:

Proposición 7.1: El estimador de las probabilidades desconocidas de Good-Turing es:

A) Menor que el de Zelterman, en el muestreo secuencial.
B) Mayor que el de Zelterman, si el muestreo es completamente aleatorio sin reemplazamiento.
C) Igual que el de Zelterman, si el muestreo es completamente aleatorio con reemplazamiento.

En efecto: Hemos comprobado en el capítulo 4 que:

$$e^{\frac{\hat{M}_1}{T}} < e^{\frac{-2\hat{M}_2}{\hat{M}_1}}, \text{ si el muestreo es secuencial,}$$

$$\frac{\hat{M}_1}{T} > e^{\frac{-2\hat{M}_2}{\hat{M}_1}}, \text{ si el muestreo es c.a. sin reemplazamiento, y}$$

$$\frac{\hat{M}_1}{T} = e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$, si el muestreo es c.a. con reemplazamiento.

Como consecuencia de esta proposición, se tiene la siguiente:

Proposición 7.2: El valor del parámetro A es:

A) A>0, en el muestreo secuencial,
B) A<0, en el muestreo c.a. sin reemplazamiento, y
C) A tiende a infinito, en el muestreo c.a. con reemplazamiento.

Demostración:

En efecto: Si el muestreo es secuencial, sabemos que:

$$\frac{\hat{M}_1}{T} < e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

pero esta relación es equivalente a afirmar que $\hat{A} > 0$, puesto que:

$$\frac{\hat{M}_1}{T} < e^{\frac{-2\hat{M}_2}{\hat{M}_1}} \Leftrightarrow \ln\hat{M}_1 - \ln T < \frac{-2\hat{M}_2}{\hat{M}_1} \Leftrightarrow \hat{M}_1(\ln T - \ln\hat{M}_1) > 2\hat{M}_2 \Leftrightarrow$$

$$\Leftrightarrow \frac{\hat{M}_1(\ln T - \ln\hat{M}_1)}{2\hat{M}_2} > 1 \Leftrightarrow \frac{A+1}{A} > 0 \Leftrightarrow 1+\frac{1}{A} > 0 \Leftrightarrow \frac{1}{A} > 0 \Leftrightarrow A > 0$$

2) Si el muestreo es aleatorio simple con reemplazamiento, por la proposición 7.1, sabemos que

$$\frac{\hat{M}_1}{T} = e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

Entonces:

$$\frac{\hat{M}_1}{T} = e^{\frac{-2\hat{M}_2}{\hat{M}_1}} \Leftrightarrow \ln\hat{M}_1 - \ln T = \frac{-2\hat{M}_2}{\hat{M}_1} \Leftrightarrow \hat{M}_1(\ln T - \ln\hat{M}_1) = 2\hat{M}_2 \Leftrightarrow$$

$$\Leftrightarrow \frac{\hat{M}_1(\ln T - \ln\hat{M}_1)}{2\hat{M}_2} = 1 \Leftrightarrow \frac{1}{A} = 0 \Leftrightarrow A \to \infty$$

3) Si el muestreo es completamente aleatorio sin reemplazamiento, por la proposición 7.1, sabemos que

$$\frac{\hat{M}_1}{T} > e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

luego

$$\frac{\hat{M}_1}{T} > e^{\frac{-2\hat{M}_2}{\hat{M}_1}} \Leftrightarrow \ln\hat{M}_1 - \ln T > \frac{-2\hat{M}_2}{\hat{M}_1} \Leftrightarrow \hat{M}_1(\ln T - \ln\hat{M}_1) < 2\hat{M}_2 \Leftrightarrow$$

$$\frac{\hat{M}_1(\ln T - \ln\hat{M}_1)}{2\hat{M}_2} < 1 \Leftrightarrow \frac{A+1}{A} < 1 \Leftrightarrow \frac{1}{A} < 0 \Leftrightarrow A < 0$$

Estas conclusiones confirman los resultados obtenidos ya por N.L.Jhonson[1] y S.Kotz, que comprobaron cómo cada una de las tres

[1] Johnson,N.L & Kotz,S.(1976). "Discrete Distributions". John Wiley & Sons. New York.

distribuciones (binomial negativa, binomial y de Poisson) pueden ser consideradas como desarrollo de:

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

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

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

A) Si el muestreo es secuencial:

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

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

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

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

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

Tomando como estimador de A:

$$A=\frac{2\hat{M}_2}{\hat{M}_1(\ln T-\ln\hat{M}_1)-2\hat{M}_2}$$

pdemos tomar, como estimador de S, uno cuaquiera de los dos sigientes:

$$\hat{S}_{12}=\frac{\hat{T}D}{\hat{T}-\hat{M}_1}+\frac{\hat{T}\hat{M}_1}{\hat{T}-\hat{M}_1}\,\frac{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)-2\hat{M}_2}{2\hat{M}_2}$$

o bien:

$$\hat{S}_{13} = \frac{D}{1-\left(\frac{\hat{M}_1}{\hat{T}}\right)^{\frac{2\hat{M}_2}{\hat{M}_1(\ln\hat{T}-\ln\hat{M}_1)}}}$$

También se puede tomar el estimador de Poisson:

$$\hat{S}_1 = \frac{D}{1-e^{-\frac{2\hat{M}_2}{\hat{M}_1}}}$$

Las estimaciones que proporcionan estos tres estimadores son muy próximas.

B) Si el muestreo es completamente aleatorio sin reemplazamiento:

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

Si hacemos B=-A>0, -w=p, con lo que q=1-w.
Entonces el término (r+1)-ésimo del desarrollo de $(1+w-w)^{-A}=(q+p)^B$ es

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

que es la función masa de probabilidad de una distribución binomial de parámetros B(B,q).

Si estimamos B por

$$\hat{B} = -A = -\frac{2\hat{M}_2}{\hat{M}_1(\ln T-\ln\hat{M}_1)-2\hat{M}_2}$$

al ser A<0 en este modelo, es B>0, y podemos utilizar, como estimadores de S, las mismas expresiones de S_{12} y S_{13}, si bien el segundo término de S_{12} será ahora negativo. Al no haber cambiado

de nombre el parámetro, A, su estimación será negativa. También S_{13} dará una estimación inferior a la del modelo secuencial.

C) Si el muestreo es completamente aleatorio con reemplazamiento, A tiende a infinito, y, como estimadores de S, se toman:

$$\hat{S}_4 = \frac{\hat{T}D}{\hat{T}-\hat{M}_1} \quad \text{ó} \quad \hat{S}_{22} = \frac{T\hat{M}_1}{2\hat{M}_2}$$

7.2. Criterio de Katz

Para seleccionar el modelo, los resultados anteriores obtenidos de la comparación de los estimadores de Good-Turing y de Zelterman para la probabilidad de especies desconocidas, se pueden contrastar por el criterio de Katz, también expuesto por Johnson y Kotz, y que se basa en la estimación muestral del estimador:

$$\hat{K} = \frac{\hat{m}_2 - \hat{X}}{\hat{X}}$$

donde s^2 es la cuasivarianza, y $\hat{X}$ la media muestral.

En nuestro caso, este estimador es:

$$\hat{K} = \frac{D(\hat{T}+\hat{R}_2)}{\hat{T}(D-1)} - \frac{\hat{T}}{D}$$

de forma que:

Si K>0, la distribución correspondiente es la binomial negativa;

Si K<0, se trata de la binomial;

Si K=0, la distribución que corresponde es la de Poisson.

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

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

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

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, en cuyo caso

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

También corresponde a una situación no homogénea con un nivel de heterogeneidad del 70%.

c) A=0, que corresponde al modelo de Ewens, en cuyo caso la distribución adecuada es la distribución log-normal.

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 podría utilizar:

$$S_{11} = \frac{T^2 \hat{M}_1}{(T-\hat{M}_1)^2} \ln \frac{T}{\hat{M}_1}$$

7.3. Ejemplo

Vamos a aplicar estos resultados al ejemplo de Holst sobre numismática que proponemos en el capítulo II.

A) Anverso de las monedas:

$$\frac{\hat{M}_1}{T} = 0'7647 < 0'7838 = e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

luego corresponde a un esquema de muestreo secuencial, en que:

$$A = 9'87 \Rightarrow \hat{\gamma}^2 = \frac{1}{A} = 0'101 \Rightarrow \hat{\gamma} = 0'318$$

El criterio de Katz confirma este resultado, ya que:

$$\frac{D(\hat{T}+\hat{R}_2)}{\hat{T}(D-1)} - \frac{\hat{T}}{D} = 0'165 > 0$$

El coeficiente de variación de Pearson es 0'318x100=31'8%

Cualquiera de los estimadores que hemos propuesto proporciona para S un valor aproximado al de Zelterman. Tomemos S_{12}:

$$\hat{S}_{12} = 823$$

Si utilizamos la expresión de la varianza del apartado 3.6.2, el error típico del estimador en el muestreo es aproximadamente igual a:

$$s_{\hat{S}_{12}} = \sqrt{\frac{823^2}{156}} = 65'89$$

B) Reverso de las monedas:

$$\frac{\hat{M}_1}{T} = 0'5 < 0'6 = e^{\frac{-2\hat{M}_2}{\hat{M}_1}}$$

También corresponde a un esquema de muestreo secuencial, en que:

$$A = 2'78 \Rightarrow \hat{\gamma}^2 = \frac{1}{A} = 0'359 \Rightarrow \hat{\gamma} = 0'59$$

El criterio de Katz confirma el resultado:

$$\frac{D(\hat{T}+\hat{R}_2)}{\hat{T}(D-1)} = \frac{\hat{T}}{D} = 0,626 > 0$$

El coeficiente de variación de Pearson es 0'59x100=59%

Cualquiera de los dos estimadores que hemos propuesto proporciona para S:

$$\hat{\sigma}_{S_{12}} = 355$$

Como estimador del error típico, resulta:

$$\hat{\sigma}_{S_{12}} = \sqrt{\frac{355^2}{102}} = 35'15$$

MODELO DE SICHEL: INVERSA GAUSSIANA

VIII. MODELO DE SICHEL (INVERSA GAUSSIANA)

8.1. Introducción

Sichel[1] y más tarde Ord[2] y Whitmore, en lugar de la distribución gamma, utilizaron la distribución inversa de Gauss o distribución *"inversa gaussiana (IG)"* en el problema del número de especies, truncándola también en el origen.

El problema fundamental es el de diseñar un modelo adecuado que nos permita estimar S a partir de los M_r, admitiendo los axiomas I y II, pero, en lugar del III, se admite ahora el axioma IV:

Axioma IV: Los $\{\lambda_j\}_{j=1,2,\ldots,s}$ son independientes y están idénticamente distribuidos siendo la inversa Gaussiana de parámetros (μ,γ) la distribución común.

La distribución IG fue introducida por Schrödinger en 1915 para estudiar el primer tiempo de paso en el movimiento Browniano.

[1] H.S. Sichel. "Assintotiscs Efficiencies of three mrthods of estimation for the inverse Gaussian-Poisson distribution". Biometrika, 69, 467-472. 1986.

[2] J.K. Ord & G.A. Whitmore. "The Poisson-Inverse Gaussian Distribution as a Model for Species Abundance". Cmmunications in Statisrics. Part A.-Theory and Methods, 15, 853-871. 1986.

El problema que se planteó Schrödinger consistía en el movimiento de una partícula, restringido a una línea recta, comenzando en x=0 en el instante t=0.

Su desplazamiento, después de un cierto tiempo t, es la suma total de los desplazamientos debidos a la influencia del campo, que es vt y el desplazamiento γ debido a la influencia del movimiento browniano. Este último desplazamiento se distribuye de acuerdo con la ley de Gauss de media 0 y varianza t/α.

Así un desplazamiento de γ a $\gamma+d\gamma$ tiene de probabilidad:

$$\sqrt{\frac{\alpha}{\pi t}}\, e^{-\frac{\alpha\lambda^2}{t}}\, d\gamma$$

siendo α la mitad de la varianza $\frac{1}{2}\sigma^2$

Ahora el desplazamiento total x es:

$$x = vt+\gamma$$

de modo que la probabilidad de que el desplazamiento quede entre x y x+dx es

$$\sqrt{\frac{\alpha}{\pi t}}\, e^{-\frac{\alpha\,(x-vt)^2}{t}}\, dx$$

Si tomamos la variable aleatoria λ , la función de densidad de la IG es

$$f(\lambda) = f(\lambda/\mu,\gamma) = \begin{cases} \left(\frac{\gamma}{2\pi\lambda^3}\right)^{1/2} \frac{e^{-\gamma(\lambda-\mu)^2}}{2\mu^2\lambda}, & si\ \lambda>0 \\ 0, & si\ \lambda\le 0 \end{cases} \qquad (8.1)$$

donde $\gamma>0$ y $\mu>0$.

La media de esta variable es μ y la varianza $\frac{\mu^3}{\gamma}$.

El modelo de Sichel consiste en tomar la inversa gaussiana (IG) como distribución de los λ_j , de modo que la distribución compuesta, inversa gaussiana de Poisson (IGP), va a ser la representación de $P(r)=P(Y_j=r)$.

La relación entre los λ_j y los p_j viene dada por:

$$p_j = \frac{\lambda_j}{\sum_{k=1}^{s} \lambda_k} \tag{8.2}$$

8.2. Distribución inversa gaussiana

La elección del axioma III nos lleva al siguiente resultado:

Proposición 8.1: Si $N_k(t)$ es una variable que sigue una distribución de Poisson de parámetro λ , es decir:

$$P[N_k(t)=r] = \int_0^\infty \frac{(\lambda t)^r}{r!} e^{-\lambda t} f(\lambda \mid \mu, \gamma)\, d\lambda$$

y λ sigue una distribución IG, la distribución de Poisson que utiliza la inversa gaussiana como distribución mixta tiene como distribución de probabilidades:

$$P[R=r] = P[N_k(t)=r] = \int_0^\infty \frac{(\lambda t)^r}{r!} e^{-\lambda t} f(\lambda \mid \mu, \gamma)\, d\lambda$$

que resulta:

$$P[R=r] = \frac{m_r^*}{r!} e^{\frac{\gamma}{\mu} - \frac{\gamma}{m^*}}, \quad r=0,1,2,\ldots \tag{8.3}$$

donde m_r^* es el momento de orden r con respecto al origen de la IG de parámetros (μ, γ) y $\mu_* = \left(\frac{1}{\mu^2} + \frac{2}{\gamma}\right)^{-\frac{1}{2}}$.

De la expresión (8.3) se deduce que, para hallar las probabilidades de la IGP, es necesario hallar los momentos de orden r respecto al origen. Vamos a llamar:

$$m_0^* = 1$$

$$m_1^* = \mu_* = \frac{1}{\sqrt{\frac{1}{\mu^2} + \frac{2}{\gamma}}}$$

$$m_r^* = \mu_*^r \left(\frac{2\theta}{\pi}\right)^{1/2} e^{\theta} K_{r-1/2}(\theta) = \mu_*^r \sum_{i=0}^{r-1} \frac{(r-1+j)!}{j!\,(r-1-j)} (2\theta)^{-j}, \quad r=1,2,\ldots$$

donde $\theta = \frac{\lambda t}{\mu_*}$ y $K_\nu(z)$ designa la función de Bessel modificada de segunda especie de orden ν y con argumento z.

Además estos momentos verifican las relaciones recurrentes, que van a facilitar el cálculo:

$$m_{r+1}^* = \mu_*^2 \left(m_{r-1}^* + \frac{(2r-1)\, m_r^*}{\gamma}\right), \quad r=1,2,\ldots$$

de donde se obtienen las relaciones de recurencia entre las probabilidades de la IGP:

$$\boxed{P_{r+1} = \mu_\nu^2 \left[\frac{P_{r-1}}{r(r-1)} + \frac{2r-1}{\gamma(r+1)} P_r\right], \quad r=1,2,\ldots} \qquad (8.4)$$

Los momentos respecto al origen se obtienen de las relaciones de recurrencia de las probabilidades anteriores.

Cuando se aplica para hallar el número de especies, debe de utilizarse, del mismo modo que en los anteriores modelos, la *distribución truncada en el origen*,

$$P(R=r) = P[N_k(t)=r] = \frac{P_r}{1-P_0} = \frac{P_r}{\left[1-e^{\frac{r}{\mu}-\frac{\gamma}{\mu_*}}\right]}, \quad r=1,2,\ldots \tag{8.5}$$

En particular, cuando μ tiende a infinito, la función de densidad de la IGP tiende a la función:

$$f(\lambda) = \left(\frac{\gamma}{2\pi\lambda^3}\right)^{\frac{1}{2}} - e^{-\frac{\gamma}{2\lambda}} \tag{8.6}$$

que es tal que $1/\lambda$es el cuadrado de una variable normal de media cero. En este caso, se verifican las siguientes propiedades:

$$\begin{aligned} \mu_* &= \left(\frac{\gamma}{2}\right)^{\frac{1}{2}} \\ P_r &= m_r^* \frac{e^{-(2\gamma)^{1/2}}}{r!} \\ m_{r+1}^* &= \frac{\gamma}{2} m_{r-1}^* + \frac{2r-1}{2} m_r^* \\ m_0^* &= 1 \\ m_1^* &= \mu_* \end{aligned} \tag{8.7}$$

En este último caso, no existe la esperanza matemática de la variable. Veamos, entonces, cómo se deducen los estimadores de μ y τ en la inversa gaussiana truncada en cero.

A partir de la fórmula de recurrencia (8.4) se obtiene:

$$\frac{\partial P_r}{\partial \mu} = \frac{\gamma (r+1) P_{(r+1)}}{\mu^3} - \frac{\gamma P_r}{\mu^2} \tag{8.8}$$

y

$$\frac{\partial P_r}{\partial \gamma} = \frac{(\gamma + r\mu) P_r}{\gamma \mu} - \frac{\gamma + \mu^2 (r+1)}{\gamma \mu^2} P_{r+1} \quad \forall\, r \geq 0. \tag{8.9}$$

Tomando logaritmos, resulta la función:

$$L = \sum_{r=1}^{\infty} \hat{M}_r \ln P_r - D \ln (1 - P_0) \tag{8.10}$$

Derivando esta expresión con respecto a μ y γ, se obtiene:

$$\frac{\partial L}{\partial \gamma} = \frac{\gamma}{\mu^3} \frac{S_1 - D\mu + D(\hat{M}_1 - \mu \hat{M})}{1 - P_0} \tag{8.11}$$

y

$$\frac{\partial L}{\partial \gamma} = \frac{1}{\mu^2 \gamma} \left[D\gamma\mu + D\bar{r}\mu^2 - (\mu^2 - \gamma)\, S_1 + \frac{D}{1 - P_0} (\gamma \mu M_0 - \hat{M}_1 \gamma - \hat{M}_1 \mu^2) \right] \tag{8.12}$$

donde

$$S_1 = \sum_{r=1}^{\infty} (r+1) \hat{M}_r \frac{P_{r+1}}{P_r} \tag{8.13}$$

Igualando a cero (8.11) y (8.12), se pueden obtener por métodos numéricos de aproximación, las estimaciones de los parámetros. Hay que tener cuidado, ya que, en algúncaso raro, podría obtenerse algún valor fuera del espacio paramétrico.

Se obtiene la misma expresión que obteníamos en el modelo paramétrico para la función de máxima verosimilitud de S:

$$L = \frac{D!}{\prod M_k!} \frac{\prod_{k=1}^{D} P_k^{M_k}}{(1-P_0)^D} \qquad (8.14)$$

de donde se obtiene el estimador insesgado de máxima verosimilitud de S:

$$\hat{S} = \frac{D}{1-\hat{P}_0} \qquad (8.15)$$

La varianza del estimador también se estima a partir de la matriz de covarianzas de los M_r.

El proceso para hallar la estimación de S consiste en hallar las soluciones de (8.11) y (8.12) por métodos iterativos partiendo de los valores iniciales de (8.7).

APÉNDICE TEÓRICO

APÉNDICE TEÓRICO

A.1. Descripción de un proceso estocástico

Se denomina *proceso estocástico* al experimento aleatorio que se desarrolla en el tiempo de una manera controlada por medio de las leyes probabilísticas. En otras palabras, un proceso estocástico se puede describir mejor como una familia $\{N(t), t \in T\}$ de variables aleatorias indexadas, cuyo índice t recorre un conjunto T.

El conjunto de índices T se denomina "*espacio paramétrico*". Cuando T toma valores enteros, se dice que tenemos un proceso *de parámetro discreto*, mientras que si T es continuo, se considera un proceso *de parámetro continuo*.

Los valores N(t) que toma el proceso se denominan "*estados*" y el conjunto E de los posibles estados "*espacio de estados*".

Cuando el proceso es discreto, denotamos el parámetro por n, representando el proceso por $\{X_n, n=0,1,2,...\}$. El parámetro natural en la mayoría de los problemas que se presentan suele ser el tiempo, por lo que, en general, cuando nos refiramos a éste, lo llamaremos "*parámetro de tiempo*". Sin embargo, hay ocasiones en que el parámetro es el área de una superficie o el volumen de una zona del espacio.

Atendiendo a la naturaleza del espacio de estados, los procesos estocás-ticos se clasifican en "*cadenas de Markov*" si el espacio de estados es discreto, y "*procesos de Markov*" cuando es continuo.

Según sea el parámetro discreto o continuo, la cadena o el proceso de Markov se dice "*de parámetro discreto*" o "*de parámetro continuo*".

Dado un valor del parámetro t, el proceso estocástico $\{N(t)\}$ es una variable aleatoria simple, cuya distribución de probabilidad se puede obtener del mismo modo que la de cualquier otra variable aleatoria. Ahora bien, como t varía en un espacio paramétrico T, la simple distribución para un determinado t dado no proporciona información suficiente acerca del proceso.

Para tener información completa del proceso, necesitamos conocer la distribución conjunta de las variables aleatorias de la familia $\{N(t), t\in T\}$.

Cuando t es continuo no es posible obtener tal distribución conjunta puesto que el número de miembros de la familia es infinito.

Parece adecuado, ante estas circunstancias, suponer que el conportamiento de estos procesos puede conocerse estudiándolo en un conjunto discreto de puntos.

Sea $(t_1, t_2, \ldots t_n)$, con $t_1<t_2<\ldots<t_n$ pertenecientes a T. Entonces, la distribución conjunta del proceso N(t) en estos puntos puede definirse como:

$$P[N(t_1)\leq x_1, \ldots, N(t_n)\leq x_n] \qquad (A.1)$$

La forma más simple para esta distribución se obtiene cuando las variables aleatorias son independientes, en cuyo caso viene dada por el producto de las distribuciones individuales, reduciéndose el estudio del proceso al estudio de una variable aleatoria simple.

A veces no se tiene una descripción completa del proceso al no ser posible el conocimiento de la distribución conjunta. En tales casos, sin embargo, es posible conseguir la información necesaria a partir de las "*funciones de distribución de transición*".

Se trata de funciones de distribución condicionales que se basan en cierta información disponible para un valor específico del parámetro de tiempo del proceso.

Sean t_0 y t_1 dos puntos de T tales que $t_0 \leq t_1$. Entonces se define la fun-ción de distribución de transición condicional como

$$F(x_0, x_1; t_0, t_1) = P[N(t_1) \leq x_1 \mid N(t_0) = x_0] \qquad (A.2)$$

Cuando el proceso estocástico tiene los espacios paramétricos y de estado discretos, se definen las probabilidades de transición como

$$P_{ij}^{(m,n)} = P(X_n = j \mid X_m = i) \ , n \geq m. \qquad (A.3)$$

Para definir el proceso de Poisson, necesitamos introducir unos conceptos previos:

Definición A.1: El proceso $\{N(t), t \in T\}$ de parámetro continuo *tiene incrementos independientes* si N(0)=0, y, para cualquier colección de índices $t_1, t_2, \ldots, t_n$, las variables aleatorias

$$N(t_1)-N(t_0), \ldots, N(t_n)-N(t_{n-1})$$

son independientes.

Definición A.2: Se dice que el proceso $\{N(t), t \in T\}$ *es homogéneo en el tiempo* (o de parámetro homogéneo) si la función de distribución de transición dada por (A.2) depende sólo de la diferencia t_1-t_0 en lugar de depender de t_0 y de t_1. Entonces se tiene:

$$F(x_0, x; t_0, t_0+t) = F(x_0, x; 0, t) \qquad (A.4)$$

para $t_0 \in T$.

Representaremos la expresión (A.4) por $F(x_0, x; t)$. La expresión correspondiente para el proceso discreto $\{X_n, n=0,1,2\ldots\}$ vendría dada por $P_{ij}^{(n)}$.

Sea F(x,t) la distribución incondicional del proceso N(t) definida como:

$$F(x,t) = P[N(t) \leq x] \tag{A.5}$$

Sea también $f(x_0)$ la densidad de probabilidad de N(0). Ahora puede determinarse F(x,t) a partir de la densidad de probabilidad $F(x_0,x;t)$ a través de la sencilla relación

$$F(x,t) = \int_{x_0 \in E} F(x_0,x;t)\, f(x_0)\, dx_0 \tag{A.6}$$

donde E representa el espacio de estados.

La relación correspondiente, en el caso discreto, tiene la forma:

$$\rho_j^{(n)} = \sum_{i \in S} p_i^{(0)}\, P_{ij}^{(n)} \tag{A.7}$$

donde

$$\rho_j^{(n)} = P[X_n = j] \tag{A.8}$$

A.2. Procesos de Markov

Los procesos estocásticos que tienen lugar en la mayor parte de las situaciones que se dan en la vida real son tales que, para un conjunto de parámetros $t_1,t_2,\ldots,t_n$ de T, las variables aleatorias $X(t_1),X(t_2),\ldots,X(t_n)$ muestran cierto grado de dependencia.

El tipo más simple de dependencia es la "*dependencia de primer orden*, que se conoce como "*dependencia de Markov*", que puede definirse como sigue:

Consideremos un conjunto de puntos finito (o infinito contable) $(t_1,t_2,\ldots,t_n)$, $t_0<t_1<t_2<\ldots<t_n<t$ y t,t_r pertenecientes a T, siendo T el espacio paramétrico del proceso {N(t)}.

Definición A.3: La dependencia que nos muestra el proceso $\{N(t), t \in T\}$ es "*dependencia de Markov*" si la distribución condicional de N(t) para ciertos valores dados de $N(t_1),\ldots,N(t_n)$ depende sólo de $N(t_n)$, que es el último valor conocido del proceso, es decir:

$$P[X(t) \leq x \mid X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0] = \quad (A.9)$$

$$= P[X(t) \leq x \mid X(t_n) = x_n] = F(x_n, x; t_n, t) \quad (A.10)$$

Definición A.4: Se denominan "*procesos de Markov*" aquellos procesos que poseen la dependencia de Markov. Por lo tanto, si, en un proceso de Markov, se conoce el estado para un valor específico del parámetro t, esta información es suficiente para predecir el comportamiento del proceso fuera de este punto.

Como consecuencia de (A.10), se cumple también la siguiente relación:

$$F(x_0, x; t_0, t) = \int_{y \in E} F(y, x; \tau, t)\, dF(x_0, y; t_0, \tau) \quad (A.11)$$

donde $t_0 < \tau < t$ y E es el espacio de estados del proceso N(t).

Cuando el proceso estocástico tiene un espacio de estados y un espacio paramétrico discretos, resulta:

para $n > n_1 > n_2 > \ldots > n_k$, siendo n y $n_1, n_2, \ldots, n_k$ pertenecientes al espacio paramétrico,

$$P[X_n = j \mid X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_k) = P(X_n = j \mid X_{n_1} = i_1) = P_{ij}^{(n_1, n)}$$

Utilizando esta propiedad, para $m < r < n$ resulta:

$$P_{i,j}^{(n_1, n)} = \sum_{k \in E} P(X_n = j \mid X_m = i) = P(X_n = j \mid X_r = k)\, P(X_r = k \mid X_m = i) = \sum_{k \in E} P_{ik}^{(m, x)}\, P_{kj}^{(r, n)}$$

donde E es el espacio de estados del proceso.

Las ecuaciones anteiores se denominan "*ecuaciones de Chapman-Kolmogorov*" para el proceso, que son fundamentales para el enfoque que demos al analizar los procesos de Markov.

A.3. Procesos de punto

Definición A.5: Un proceso estocástico, cuyas variables aleatorias sólo pueden tomar los valores enteros 0,1,2,... se denomina *proceso de valores enteros o de punto*.

La idea principal de los procesos de punto consiste en el estudio de colecciones aleatorias de ocurrencias puntuales. Vamos a suponer que los puntos tienen lugar a lo largo del eje de tiempo, en principio, aunque también se puede considerar que los puntos tienen lugar en una cierta región del espacio.

El proceso de punto más simple es aquel en que los puntos tienen lugar de manera totalmente aleatoria.

Se entiende por proceso de punto un proceso de valores enteros $\{N(t), t\in T\}$ que cuenta el número de puntos en un intervalo, estando distribuidos los puntos por medio de un mecanismo estocástico determinado.

Los puntos van a representar los instantes $\tau_1, \tau_2, ..$ en los que han tenido lugar los sucesos de un determinado carácter específico, siendo $0<\tau_1<\tau_2<...$

Se denominan tiempos sucesivos entre llegadas a las variables aleatorias

$$T_1=\tau_1, T_2=\tau_2-\tau_1, \ldots, T_n=\tau_n-\tau_{n-1}, \ldots \qquad (A.12)$$

Si representamos por N(t) el número de puntos en el intervalo (0,t], entonces $\{N(t), t\in T\}$ es el proceso de punto de la serie.

Se puede definir un proceso de punto de varias formas, siendo las más usuales las que lo definen:

a) directamente, como un proceso con incrementos independientes;
b) deduciendo sus propiedades a partir de las hipótesis realizadas sobre los tiempos entre llegadas.

A.4. Procesos de Poisson

El "*proceso de Poisson*" es un proceso de punto cuyo espacio de estados es discreto y cuyo espacio paramétrico es continuo. Introduimos el proceso directamente.

Definición A.6: Se dice que un proceso de valores enteros $\{N(t), t\in T\}$ es un *proceso de Poisson* con número de ocurrencias (ó intensidad) ν si se cumplen las siguientes hipótesis:

I. $\{N(t), t\in T\}$ tiene incrementos estacionarios independientes.

II. Para instantes cualesquiera s y t, el número N(t)-N(s) de veces que ha tenido lugar el suceso en el intervalo de s a t tiene una distribución de Poisson de media $\nu(t-s)$.

Luego, para k=0,1,2,... es

$$P[N(t)-N(s)=k] = e^{-\nu(t-s)}\frac{[\nu(t-s)]^k}{k!}, \qquad (A.13)$$

$$E[N(t)-N(s)] = \nu(t-s) \ , \ Var[N(t)-N(s)] = \nu(t-s). \qquad (A.14)$$

ν representa el número medio de ocurrencias por unidad de tiempo de los sucesos que se cuentan.

Aquellos sucesos, cuya función de conteo N(.) sea un proceso que tiene lugar de acuerdo con un proceso de Poisson con un número medio de ocurrencias ν, se dirá que son procesos del tipo de Poisson con intensidad ν.

A.4.1. Procesos de Poisson en el tiempo

Sea el proceso N(t), que representa el número de veces que un suceso tiene lugar en el intervalo de tiempo (0,t]. Definamos, para s<t:

$$P_{i,j}(s,t) = P\{N(t) = j \mid P[N(s) = i]\} \qquad (A.15)$$

Supongamos que tiene lugar a lo largo del intervalo de tiempo (0,t], y sea, para t>0, N(t)="número de sucesos que ocurren en el intervalo (0,t]" de modo que, para h>0, arbitrariamente pequeño, N(t+h)-N(t) sólo toma valores no negativos.

Entonces decimos que el proceso de punto {N(t), t>0} es un proceso de Poisson si satisface los siguientes axiomas:

Axioma I: N(0)=0, ya que se empieza a contar en el instante t=0.

Axioma II: El proceso N(t) tiene incrementos independientes.

Axioma III: Para todo t>0, 0<P[N(t)>0]<1.

Axioma IV: Para todo t>0, se verifica

$$\lim_{h \to 0} \frac{P[N(t+h) - N(t) \geq 2]}{P[N(t+h) - N(t) = 1]} = 0 \qquad (A.16)$$

Axioma V: El proceso de punto N(t) tiene incrementos estacionarios.

El axioma III significa que, en un intervalo cualquiera, tan pequeño como se desee, hay una probabilidad positiva de que ocurra un suceso, aunque no hay certeza.

El axioma IV dice que, en intervalos infinitamente pequeños, no pueden tener lugar simultáneamente dos o más sucesos.

De un proceso de Poisson nos interesa, en principio, el número total N(t) de ocurrencias en un intervalo de tiempo de amplitud t.

Todas las ocurrencias son de la misma clase, situándose cada ocurrencia en un punto del eje de tiempo.

Se supone que las fuerzas e influencias que rigen el fenómeno se conservan constantes, de modo que la probabilidad de cualquier suceso particular es la misma en todo intervalo de tiempo de duración (o amplitud) t e independientes del desarrollo pasado del proceso. Esto significa que se trata de *un proceso de Markov homogéneo en el tiempo*. Vamos ahora a deducir las probabilidades básicas:

$$P_n(t)=P[N(t)=n] \qquad (A.17)$$

Supongamos que se trata de un proceso temporal. Elegimos, entonces, un origen de medición del tiempo, decimos que "*en la época t>0 el sistema está en el estado E_n si han tenido lugar exactamente n saltos entre 0 y t*".

Entonces, $P_n(t)$ es igual a la probabilidad del estado E_n en la época t, pero $P_n(t)$ puede ser descrito también como la probabilidad de transición desde un estado arbitrario E_j en una época arbitraria s al estado E_{j+n} en la época s+t. Siguiendo a Feller, vamos a traducir esta descripción informal del proceso a propiedades de las probabilidades $P_n(t)$.

Para ello, dividamos un intervalo de tiempo de longitud igual a la unidad en N subintervalos de longitud h=1/N. La probabilidad de algún salto dentro de estos subintervalos es igual a $1-P_0(h)$, y, de esta manera, el número esperado de subintervalos que dan cabida a algún salto es igual a $h^{-1}[1-P_0(h)]$.

Cuando h tiende a cero, este número converge al número esperado de saltos dentro de cualquier intervalo de longitud unitaria. Luego se puede suponer que:

$$\exists \lambda > 0 \mid h^{-1}[1-P_0(h)] \rightarrow \lambda \qquad (A.18)$$

Un salto siempre debe conducir desde un estado E_j al estado vecino E_{j+1}, luego el número esparado de subintervalos (de longitud h) que dan cabida a más de un salto debe tender a cero.

Luego, cuando $h \to 0$, $h^{-1}[1-P_0(h)-P_1(h)] \to 0$

$$P_0(h) = 1-\lambda h+o(h) \qquad (A.19)$$

(A.19) es equivalente a $P_1(h) = \lambda h+o(h)$.

A.4.2. Deducción de la ley que regula el proceso

1. El proceso comienza en la época cero desde el estado E_0.

2. Las transiciones directas desde un estado E_j sólo son posibles a E_{j+1}.

3. Cualquiera que sea el estado E_j en la época t, la probabilidad de un salto dentro de un intervalo corto de tiempo entre t y t+h es igual a

$$\lambda h+O(h) \qquad (A.20)$$

mientras que la probabilidad de más de un salto es o(h).

A partir de estos postulados se demuestra formalmente que

$$P_n(t) = (\lambda t)^n \frac{e^{-\lambda t}}{n!} \qquad (A.21)$$

Con el fin de demostrarlo, supongamos, en primer lugar, que $n \geq 1$, y consideremos el suceso de que en la época t+h, el sistema está en el estado E_n.
La probabilidad de este suceso es igual a $P_n(t+h)$ y puede tener lugar de tres maneras distintas:

1) En la época t, el sistema puede estar en el estado E_n y no hay ningún salto. La probabilidad de esta situación es:

$$P_n(t)\,P_0(h) = P_n(t)\,[1-\lambda h] + o(h) \qquad (A.22)$$

2) La segunda posibilidad es que, en la época t, el sistema se encuentre en E_{n-1} y ocurra exactamente un salto entre t y t+h. La probabilidad de esta situación es:

$$P_{n-1}(t)\,\lambda h + o(h) \qquad (A.23)$$

3) En la época t, cualquier otro estado requiere más de un salto entre t y t+h, y la probabilidad de ese suceso es o(h).

Como consecuencia se debe cumplir que

$$P_n(t+h) = P_n(t)\,(1-\lambda h) + P_{n-1}(t)\,\lambda h + o(h) \qquad (A.24)$$

lo cual puede expresarse en la forma

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h} \qquad (A.25)$$

Cuando $h \to 0 \Rightarrow o(h) \to 0$, y el límite es:

$$P'_n(t) = -\lambda P_n(t) + \lambda P_{n-1}(t) \quad , n \geq 1 \qquad (A.26)$$

Para n=0, no se presentan la segunda ni la tercera contingencias que se mencionaron antes y, por lo tanto, la última fórmula se reemplaza por

$$P_0(t+h) = P_0(t)\,(1-\lambda h) + o(h) \qquad (A.27)$$

que tiende a

$$\boxed{P'_0(t) = \lambda P_0(t)} \qquad (A.28)$$

A partir de este resultado y de que $P_0(t)=1$, obtenemos:

$$P_0(t) = e^{-\lambda t} \qquad (A.29)$$

Si substituimos este resultado en (A.26), obtenemos una ecuación diferencial ordinaria en $P_1(t)$:

$$P'_1(t) = -\lambda P_1(t) + \lambda e^{-\lambda t} \qquad \text{(A.30)}$$

que, con $P_1(0)=0$, nos da finalmente:

$$P_1(t) = \lambda t e^{-\lambda t} \qquad \text{(A.31)}$$

Procediendo de igual modo, obtenemos todos los términos $P_n(t)$; así:

$$\boxed{P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}} \qquad \text{(A.32)}$$

A.5. Tiempos entre llegadas y tiempo de espera

Cuando se observan sucesos que tienen lugar en el tiempo, en vez de contar el número de veces que tiene lugar un determinado suceso, se acostumbra a medir el período de tiempo transcurrido hasta que tiene lugar un número fijo de sucesos. La variable aleatoria W_r, que representa el tiempo transcurrido hasta que se registran r sucesos, es llamada *tiempo de espera* hasta el suceso r-ésimo. Entonces se definen también los tiempos entre llegadas:

Dados unos sucesos que tienen lugar en el intervalo (0,t), se definen los *tiempos entre llegadas sucesivos*: $T_1, T_2, \ldots$ del modo siguiente:

T_1 es el tiempo que transcurre desde el instante 0 hasta que tiene lugar el primer suceso, y, para k>1, T_k es el tiempo que transcurre desde que tiene lugar el (k-1)-ésimo hasta el k-ésimo suceso.

Los tiempos entre llegadas están relacionados con los tiempos de espera por:

$$T_1=W_1,\ T_2=W_2-W_1,\ \ldots,T_r=W_r-W_{r-1},\ \ldots \qquad \text{(A.33)}$$

y

$$W_r=T_1+T_2+\ldots+T_r,\ r>0. \qquad (A.34)$$

La relación entre los procesos y los tiempos de espera viene dada por:

$$\boxed{N(t)\leq n \Leftrightarrow W_{n+1}>t} \qquad (A.35)$$

A.5.1. Distribución de los tiempos entre llegadas

La siguiente proposición muestra la distribución de los tiempos entre llegadas:

Proposición A.3: Los tiempos entre llegadas, T_1, T_2,... correspondientes al proceso de Poisson $\{N(t), t\geq 0\}$ son variables aleatorias independientes y tienen la misma distribución exponencial de media $1/\lambda$.

La función de densidad es, por tanto:

$$f_{T_n}(x)=\lambda e^{-\lambda x},\ x\geq 0;\ n=1,2,\ldots \qquad (A.36)$$

siendo $E(T_k)=\frac{1}{\lambda}$ y $Var(T_k)=\frac{1}{\lambda^2}$. (A.37)

Esto significa que, comenzando por un origen de tiempo arbitrario, los subsiguientes puntos tienen lugar en los instantes $T_1, T_2,\ldots$ de modo que las variables aleatorias T_i son independientes y tienen todas la misma distribución.

Demostración: Teniendo en cuenta la equivalencia de los conjuntos

$$N(t)=0 \ y \ T_1>t$$

resulta:

$$e^{-\lambda t} = P[N(t)=0] = P[T_1 > t] \qquad (A.38)$$

Luego T_1 tiene una distribución exponencial de media $1/\lambda$

Si designamos por $f_1(x)$ la densidad de probabilidad de T_1, tenemos:

$$P[T_2 > t] = \int_u P[T_2 > t \mid T_1 = u]\, f_1(u)\, du = \int_u P[T(t+u) - T(u) = 0]\, f_1(u)\, du =$$

$$= e^{-\lambda t} \int_u f_1(u) = e^{-\lambda t}$$

lo que demuestra que T_2 es también exponencial de media $1/\lambda$ y que es inde-pendiente de T_1. Para demostrarlo, hemos supuesto la homogeneidad del proceso de Poisson. Si ahora llamamos $f_2(x)$ a la función de densidad de $T_3=T_1+T_2$, por un razonamiento similar legamos a que T_3 es exponencial también con la misma media, y, de este modo se demostraría la proposición por inducción.

Supongamos ahora que t es un punto en el que tiene lugar un suceso. Sea R(t) el tiempo transcurrido hasta que tiene lugar un nuevo suceso, y sea S(t) el tiempo transcurrido desde que tuvo lugar el último suceso. Debemos determinar la distribución de R(t) y de S(t) para tener un conocimiento completo del proceso.

Proposición A.4: La distribución de R(t) es independiente de t y viene dada por

$$P[R(t) \leq x] = 1 - e^{-\lambda x} \qquad (A.39)$$

Demostración: Sea τ, $(0 \leq \tau < t)$el instante en que tuvo lugar el último suceso. El suceso {R(t)>x} implica la equivalencia de los dos sucesos:

$$\{R(t) > x\} \;\; y \;\; \{T_n > t - \tau + x \mid T_n > t - \tau\} \qquad (A.40)$$

dando

$$P[R(t) > x] = P(T_n > t-\tau+x \mid T_n > t-\tau) =$$

$$= \frac{P(T_n > t-\tau+x)}{P(T_n > t-\tau)} = \frac{e^{-\lambda(t-\tau+x)}}{e^{-\lambda(t-\tau)}} = e^{-\lambda x}$$

Proposición A.5: La distribución S(t) tiene una concentración de probabilidad en t y viene dada por

$$P[S(t) = t] = e^{-\lambda t} \qquad (A.41)$$

$$P[S(t) \leq t] = 1 \text{ en otro caso } e^{-\lambda x}, 0 \leq t \qquad (A.42)$$

Supongamos que hay al menos un suceso que tiene lugar en el intervalo (0,t]. Entonces x es tal que al menos hay un suceso que tiene lugar en el intervalo (t-x,t]. Por lo tanto, podemos escribir:

$$P[S(t) \leq x] = \text{P[Al menos hay un suceso en (t-x,t)]}=$$

$$=1-\text{P[Ningún suceso tiene lugar en (t-x,t)]}=$$

$$1-e^{-\lambda x} \qquad (A.43)$$

Las variables R(t) y S(t) son conocidas como "*ecuaciones prospectivas y retrospectivas*"

Como corolario de la proposición A.2, se obtiene la siguiente propiedad de la función exponencial:

Corolario A.1: Sea Z una variable aleatoria con distribución exponencial y sean $s, t \geq 0$. Entonces:

$$P[T > t+s \mid T > s] = \frac{P(T > t+s)}{P(T > s)} = \frac{e^{-(t+s)}}{e^{-s}} = e^{-t} = P[T > t] \qquad (A.44)$$

Esta propiedad, que se conoce como "*falta de memoria de la función exponencial*" o "*propiedad de Markov*", en el caso contiuo, caracteriza a la distribución exponencial, ya que es ésta la única distribución continua con esta propiedad.
En virtud de esta propiedad, cuando trabajamos con procesos de Poisson, resulta intranscendente si ha tenido o no lugar el último suceso.

A.5.2. Distribución de los tiempos de espera

Los tiempos de espera, W_r, representan el tiempo transcurrido hasta que han tenido lugar r sucesos, de modo que

$$W_r = T_1 + T_2 + \ldots + T_r \ , \ r \geq 1 . \qquad (A.45)$$

Al ser sumas de r variables aleatorias independientes con distribución exponencial de media $1/\lambda$, los tiempos de espera son también variables aleatorias independientes con distribución gamma de parámetros $\Gamma(r, 1/\lambda)$. Luego la función de densidad de los tiempos de espera viene dada por:

$$f_{W_r}(t) = \frac{t^{r-1} e^{-\lambda t}}{(r-1)!} \qquad (A.46)$$

La función de distribución de los tiempos de espera W_r es:

$$F_{W_r}(t) = 1 - e^{-\lambda t}\left(1 + \lambda t + \ldots + \frac{(\lambda t)^{r-1}}{(r-1)!}\right), t>0; \qquad (A.47)$$

La demostración resulta inmediata teniendo en cuenta que:

$$1 - F_{W_r}(t) = P[W_r > t] = P[N(t) < r] = \sum_{k=1}^{r-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!}; \qquad (A.48)$$

La función característica es, por tanto:

$$\Phi_{W_r}(u) = \left(1 - i\frac{u}{\lambda}\right)^{-r}; \qquad (A.49)$$

de donde resultan inmediatamente:

$$E(W_r) = \frac{r}{\lambda}; \qquad (A.50)$$

$$Var(W_r) = \frac{r}{\lambda^2}; \qquad (A.51)$$

A.6. Distribución binomial negativa. Proceso de Polya

El proceso de Polya se obtiene como paso al límite del esquema de urnas de Polya, y se trata de un proceso puro de nacimiento, no estacionario. Las ecuaciones diferenciales que dan lugar al proceso son:

$$\begin{aligned} P'_r(t) &= -\frac{A+r}{t}\lambda P_r(t) + \frac{A+r-1}{t}\lambda P_{r-1}(t), \quad r \geq 1 \\ P'_0(t) &= 0 \end{aligned} \qquad (A.52)$$

con las condiciones iniciales

$$P_i(0) = 1, \quad P_r(0) = 0, \quad r \neq A \qquad (A.53)$$

Llegamos a esta distribución a partir de la distribución de Poisson compuesta con la distribución gamma de las λ_k.

Cuando la función de densidad es, como en nuestro caso, una gamma de parámetros A y t/A, se obtiene la forma límite[1] de la distribución de Polya para el valor de c=1:

$$P_r = \binom{A+r-1}{A-1}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A, \quad r=1,2,\ldots \qquad (A.54)$$

que vamos a tomar truncada en cero, ya que desconocemos las especies que no forman parte de la muestra:

$$P_r^* = \binom{A+r-1}{A-1}\left(\frac{t}{t+A}\right)^r\left(\frac{A}{t+A}\right)^A \frac{1}{1-\left(\frac{A}{A+t}\right)^A}\ ,\ r=1,2,\ldots \qquad (A.55)$$

donde llamamos

$$P_r^* = \frac{P_r}{1-P_0} = P[U_k(t)=r] = P[N_k(t)=r \mid N_k(t)>0] \qquad (A.56)$$

siendo P_0 la suma de las probabilidades de las especies desconocidas.

La suma de variables aleatorias independientes con distribución binomial negativa es una variable aleatoria también con distribución binomial negativa según vamos a ver.

Proposición A.6: Si $N_i(t)$ y $N_j(t)$ son independientes y siguen ambas una distribución binomial negativa de parámetros BN(A,p), la suma $N_i(t)+N_j(t)$ se distribuye también según una distribución binomial negativa BN(2A,p). Generalizando ese resultado a D procesos, se obtiene:

Proposición A.7: Si $N_1(t)$, $N_2(t)$,...,$N_D(t)$ son D variables aleatorias independientes, todas con la misma distribución binomial negativa BN(A,p), la suma $N_1(t)+N_2(t)+\ldots+N_D(t)$ tiene una distribución binomial negativa de parámetros BN(DA,p).

Demostración: La función generadora de probabilidades de $N_k(t)$ es:

$$\left(\frac{p}{1-(1-p)u}\right)^A \qquad (A.57)$$

La función generadora de probabilidades de la suma de las D variables independientes $N_k(t)$ es el producto de las funciones generadoras, luego:

$$\Phi_{S_T}(u) = \prod_{k=1}^{D} \left(\frac{p}{1-(1-p)\,u} \right)^{A} = \left(\frac{p}{1-(1-p)\,u} \right)^{DA} \qquad (A.58)$$

que es la función generadora de probabilidades de una distribución binomial negativa con parámetros (DA,p).

Proposición A.8: Si $N_1(t),\ldots,N_D(t)$ son independientes y siguen una distribución binomial negativa BN(A,p), la distribución de $(N_1(t),\ldots,N_D(t))$ condicionada por $\sum_{k=1}^{D} N_k(t) = T$ es

$$P\left[N_1(t) = n_1, N_2(t) = n_2, \ldots \mid \sum_{k=1}^{D} N_k(t) = T \right] = \frac{\prod_{k=1}^{D} \binom{A+n_k-1}{A-1}}{\binom{DA+T-1}{T}} \qquad (A.59)$$

Demostración:

$$P\left[N_1(t) = n_1, N_2(t) = n_2, \ldots \mid \sum_{k=1}^{D} N_k(t) = T \right] =$$

$$= \frac{P\left[N_1(t) = n_1, \ldots, N_D(t) = T - \sum_{k=1}^{D-1} N_k(t) \right]}{P\left[\sum_{k=1}^{D} N_k(t) = T \right]} =$$

$$= \frac{\prod_{k=1}^{D-1} \left[\binom{A+x_k-1}{A-1} p^A (1-p)^{x_k} \frac{1}{1-p^A} \right] \left[\binom{A+T-\sum N_k(t)-1}{A-1} \frac{p^A (1-p)^{T-\sum_{k=1}^{D-1} N_k(t)}}{1-p^A} \right]}{\binom{DA+T-1}{T} p^{DA} (1-p)^T \frac{1}{(1-p^A)^D}} =$$

$$= \frac{\prod_{k=1}^{D} \binom{A+x_k-1}{A-1}}{\binom{DA+T-1}{T}} \qquad (A.60)$$

Proposición A.9 Si $\vec{H}=(H_1,\ldots,H_D)$ sigue una distribución multinomial negativa $MN(A;p_1,p_2,\ldots,p_D)$, se verifica:

1) $L=\sum_{k=1}^{D} H_k$ sigue una distribución binomial negativa $BN(A;1-p)$,

donde $1-p=P_0$.

2) La distribución del vector $\vec{H}=(H_1,\ldots,H_D)$ condicionada por $L=n$, es multinomial $M_\alpha(n,h_1,\ldots,h_0)$, con parámetros $h_i=\frac{p_i}{p}$.

Demostración: Sea $x\geq 0$ un número entero. Para hallar la probabilidad $P(L=x)$, necesitamos sumar la función de densidad conjunta, que es multinomial, sobre todos los números enteros no negativos $x_1,\ldots,x_0$, cuya suma es x.

En virtud del desarrollo en serie de la multinomial, resulta:

$$\sum_{x_1}\cdots\sum_{x_D} \binom{A+x-1}{x} P_0^{H} \frac{x!}{x_1!\ldots x_D!} p^{x_1}\ldots p_D^{x_D} =$$

$$= \binom{A+x-1}{x} P_0^{A}(1-P_0)^{x}\ ,\ x=x_1+\ldots+x_D.$$

lo que demuestra 1).

Veamos la demostración de la segunda parte:

Si $x_1+\ldots+x_D=x$, será:

$$P[H_1=x_1,\ldots,H_D=x_D\ ,\ H=x] = P[H_1=x_1,\ldots,H_D=x_D]$$

y

$$P[H_1=x_1,\ldots,H_D=x_D,T=x] = 0,\ si\ x_1+\ldots+x_D\neq x.$$

Entonces, para $x_1+\ldots+x_D=x$, se verifica:

$$P[H_1=x_1,\ldots,H_D=x_D\ ,\ L=x] = \frac{x!}{x_1!\ldots x_D!}\prod_{i=1}^{D}\left(\frac{\rho_i}{\rho}\right)^{x_i} =$$

$$= \frac{x!}{x_1!\ldots x_D!}\prod_{i=1}^{D}\left(\frac{\hat{M}_i}{D}\right)^{x_i}$$

A.6.1. Inmersión del modelo en el muestreo multinomial

Queremos detener el proceso en el instante L+A, es decir, cuando aparezcan por primera vez A individuos de la clase I_0. Sea H_i el número de individuos de la clase I_i que hay en el instante L+A.

Nos interesa conocer las distribuciones que afectan a H_k y a L en el instante de parada. Para ello, hacemos uso de las siguientes proposiciones que establece Sidney C. Port[1]:

Sea, entonces, L+A el tiempo necesario (ó, lo que es igual) el número de pruebas necesarias par tener A individuos de la especie I_0, siendo H_i el número de individuos de la clase I_i en el instante L+A.

[1] Sidney C. Port. "Theoretical Probability for Applications". Capítulo 30. John Wiley & Sons, INC. New York-1994.

Proposición A.10: L sigue una distribución binomial negativa $BN(A,P_0)$, con A>0. Sea $\vec{H}=(H_1,\ldots,H_D)$ un vector aleatorio sobre R^D tal que la distribución de $\vec{H}=(H_1,\ldots,H_D)$ condicionada por L=n es multinomial $M_1(n,h_1+..+h_0)$, para todo $n\geq 0$.

Entonces, para $1\leq k\leq D$, $(H_1,\ldots,H_K)$ sigue una distribución multinomial negativa $MN(A,q_1+\ldots+q_k)$, donde

$$q_j=\frac{(1-P_0)\,h_j}{1-(1-P_0)\,(1-h)}\ ,\ y\ h=h_1+\ldots+h_k.$$

Demostración: Como la distribución de $\vec{H}=(H_1,\ldots,H_D)$ condicionada por L=n, es multinomial $M_\alpha(n,p_1,\ldots,p_D)$, la distribución de $(H_1,\ldots,H_D)$ condicionada por L=n, es multinomial $M_\beta(n,h_1,\ldots,h_k)$. Sean $x_1,\ldots,x_k$ números enteros no negativos y sea $x=x_1+\ldots+x_k$ su suma. Sea $p=p_1+\ldots+p_k$. Entonces, para $x_1,\ldots,x_k$ fijos, se verifica:

$$P(H_1=x_1,\ldots,H_k=x_k\mid L=n)=$$

$$=\frac{n!}{x_1!,\ldots,x_n!\,(n-x)!}\,p_1^{x_1}\ldots p_k^{x_k}(1-h)^{n-x}1_{[n\geq x]}$$

Por el teorema de la probabilidad total:

$$P(X_1=x_1,\ldots,X_k=x_k)=\sum_{k=1}^{\infty}P(X_1=x_1,\ldots,X_k=x_k\mid L=n)\,P(L=n)=$$

Para evaluar la última suma, consideremos el desarrollo

$$(1-t)^{-A}=\sum_{n=x}^{\infty}\binom{A+n-1}{n}t^n$$

válido para $0 \le t < 1$. Diferenciando x veces en ambos miembros, se obtiene:

$$(A+x-x)_x (1-t) \sum_{n=x}^{\infty} \binom{A+n-1}{n} (n)_x t^{n-x}$$

Como $n!/(n-x)! = (n)_x$, si hacemos $t = P_0(1-p)$ en la relación anterior, resulta que la suma es

$$(A+x-1)_x (1-\rho(1-p))^{-A-x}$$

De este modo

$$P(H_1 = x_1, \ldots, H_k = x_k) = P_0 [A(1-P_0))^{-A-x} p_1^{x_1} \ldots p_k^{x_k} \frac{x!}{x_1! \ldots x_k!} \frac{(A+x-1)_x}{x!} =$$

$$= \binom{A+x-1}{x} \frac{x!}{x_1! \ldots x_k!} \prod_{i=1}^{k} \left[\frac{(1-P_0) h_i}{1-(1-P_0)(1-h)} \right]^{x_i} (1-P_0)^A P_0$$

lo que prueba la proposición.

Proposición A.11: Sea $\vec{H} = (H_1, \ldots, H_D)$ un vector con distribución multinomial negativa $MN(A; p_1, \ldots, p_D)$.

Entonces, para $1 \le k \le r$, $\vec{H} = (H_1, \ldots, H_k)$ sigue una distribución multinomial negativa $MN(A, \tau_1, \ldots, \tau_k)$, donde $\tau_j = \dfrac{p_j}{1 - \sum_{i=k+1}^{D} p_i}$.

Demostración: Por la proposición A.10, la distribución de

$(H_1,\ldots,H_D)$, condicionada por L=n, donde $L=H_1+\ldots+H_r$, es multinomial $M_\alpha(n,p_1,\ldots,p_D)$ con $p_i=\frac{p_i}{1-P_0}$, teniendo L una distribución binomial negativa $BN(A,P_0)$.

El vector $(H_1,\ldots,H_k)$ tiene una distribución multinomial negativa $MN(A,q_1,\ldots,q_k)$ con $q_j=\frac{(1-P_0)h_j}{1-(1-P_0)(1-h)}$ y $h=h_1+\ldots+h_k$.

Como $(1-P_0)h_i=p_i$, se sigue que $q_j=\frac{p_j}{1-\sum_{i=k+1}^{D}p_i}$, c.q.d.

Proposición A.12: Sea L+A el número de pruebas necesarias para tener por primera vez exactamente A individuos de la especie I_0, y sea H_i el número de individuos de la clase i después de la prueba L+A. Entonces $\vec{H}=(H_1,..,H_D)$ se distribuye según una multinomial negativa $MN(A,h_1,\ldots,h_D)$.

Demostración: La sucesión de pruebas tales que hay un éxito si aparece un individuo de la especie I_{D+1} y un fracaso si no es así, constituye una sucesión de pruebas de Bernoulli con probabilidad de éxito P_0, y L+A es el instante del H-ésimo éxito para estas pruebas de Bernoulli. Según vimos antes, L sigue una distribución binomial negativa $BN(A,P_0)$.

Consideremos el suceso $[H_1=x_1,\ldots,H_D=x_D, L=n]$. Es evidente que se trata de un suceso imposible, a no ser que $x_1,\ldots,x_D$ sean números enteros no negativos cuya suma sea $x=x_1+\ldots+x_D=L$.
En tal caso, la prueba L+A debe dar como resultado un individuo de la especie I_0, mientras que en las L+A-1 pruebas anteriores, exactamente x individuos deben pertenecer a la especie

I_i , $1 \le i \le r$, y exactamente A-1 individuos a la especie I_0. Así:

$$P[H_1=x_1,\ldots,X_r=x_r \mid L=n] = p_{r+1}\frac{(L+A-1)!}{x_1!\ldots x_r!\,(A-1)!}p_1^{x_1},\ldots,p_r^{x_r}P_0^{A-1}$$

Como

$$P[L=n] = \binom{A+n-1}{n}P_0^A(1-P_0)^n$$

vemos que la distribución de $\vec{H}=(H_1,\ldots,H_D)$ condicionada por T=n,

es multinomial $M_1\left(n,\frac{h_1}{1-P_0},\ldots,\frac{h_r}{1-P_0}\right)$

Entonces, por la proposición A.10, $\vec{H}=(H_1,\ldots,H_D)$ sigue una distribución multinomial negativa $MN(A,h_1,\ldots,h_D)$.

A.7. Superposición de procesos de renovación

Una sucesión de variables aleatorias S_n constituye un *proceso de renovaión* si es de la forma

$$S_n = T_1+T_2+\ldots+T_n$$

siendo las T_i variables aleatorias mutuamente independientes con una distribución común F tal que F(0)=0.

Además de las T_k, se puede definir una variable no negativa S_0 con una distribución propia F_0.

Entonces resulta:

$$S_n = S_0+T_1+T_2+\ldots+T_n$$

El proceso de renovación se llama "*puro*" si S_0=0 y 0 cuenta como la ocurrencia número cero. En otro caso, se llama "*diferido*".

A la esperanza $\mu=E(T_k)$ se le llama "*tiempo medio de recurrencia*".

En la mayor parte de las aplicaciones, las T_k se suelen interpretar como "*tiempos de espera*", en cuyo caso a las S_n se les llama "*épocas de renovación*".

Dados n procesos de renovación, se puede formar un nuevo proceso combinando todas sus épocas de renovación en una sucesión. En general, el nuevo proceso no es de renovación, pero se puede calcular el tiempo de espera W hasta la primera renovación fácilmente.

Se demuestra que, bajo condiciones bastante generales como las que estableció Grigelionis en 1963, la distribución de W es aproximadamente exponencial, de modo que el proceso combinado está muy próximo a un proceso de Poisson.

1. W. Feller. "Introducción a la Teoría de la Probabilidades y sus Aplicaciones". Vol. II, Cap. 2, apdo. 6. LIMUSA. México, 3ª reimpresión 1993.

BIBLIOGRAFÍA

ARNOLD,B.C. y Beaver,R.J. (1988). "Estimation of the Number of Classes in a Population". Biometrical Journal, 30, 413-424.

ATKINSON,A.C., y YEH,I. (1982). "Inference for Sichel's Compound Poisson Distribution". Journal of the American Statistical Association", 77, 153-158.

ATKINSON,A.C., y YEH,I. (1988). "Estimation of the Number of Classes in a Population", Biometrical Journal, 30, 413-424.

BARNDORFF-NIELSEN,O.E. and COX,D.R. (1989). "Asymptotic Techniques for Use in Statistics". Chapman and Hall. London.

BERRY,D.A. and LINDGREN,B. (1990). "Statistics. Theory and Methods". Brooks/Cole Publishing Company. Pacific Grove.

BETRO,B. y ZIELINSKI,R. (1987). "A Monte Carlo of a Bayessian Decission Rule Concerning the Number of Different Values of a Discrete Random Variable". Communications in Statistics, Part B-Simulatuon and Computation, 16. 925-938.

BICKET,P.J. Y YAHAV,J.A. (1985) "On Estimating the Number of Unseen Species: How Many Executions Were There?". Technical Repport, nº 43, Universidad de California. Berkeley. Dept. of Statistics.

----------(1985). "On estimating the total Probability of the Unnonservet Outcomes of an Experiment". (Discurso pronunciado en un Simpposio en honor de H. Robbins. Vol. 8, ed. J.Van Rizin).

BICKET,P.J. Y YAHAV,J.A. (1988). "On Estimating the Number of Unseen Species and System Relility. De Statistical Decisions. Theory and Related Topics IV (Vol. 2). Ed. S.S.

BOENDER,C.G.E. and RINNOOY KAN, A.H.G. (1987). "A Multinomial Bayessian Approach to the Estimation of Population and Vocabulary Size". Biometrika, 74, 849-856.
California.

BRAINERd,R. (1972). "On the Relation Between Types and Tockens in Literay Text". Journal of Applied Probability. 9, 507-518.

BROWN,I.D. (1955). "Some Notes on the Coinage of Elisabeth I with Special Reference to Her Hammered Silver". British Journal of Numismatics, 28, 568-603.

BUNGE,J. y FITZPATRICK,M.(1993). "Estimating the Number of Species: A Review", Journal of the American Statistical Association, Vol. 88, N°421.

BURREL,Q. (1989). "On the Growth of Bibliographies with Time: Un Exercise in Bibliometrics". Journal of Documentation, 45, 302-317.

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". Biometrics, 43, 783-791.

CHAO,A. y LEE,S.M.(1992). "Estimating the Number of Classes via Sample Coverage", Journal of the American Statistical Association, Vol. 87, N° 417.

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

Chapman,D.G. (1951) "Some Properties of the Hipergeometric Distribution with Applications to Zoological Sample Censuses". University of California Pubications in Statistics; vol.1, pag. 131-160.

CLAYTON,M.K. y FREES,E.W.(1987). "Nomparametric Stimation of the Probability of Discovering a New Species", Journal of the American Statistical Association, 82, 305-311.

COHEN,A. y SACKOWITZ,H.B.(1990). "Admissibility of Stimators of the Probability of Unobserved Outcomes", Annals of the Institute of Statistical Mathematics, 42, 623-636.

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

CRAIG,C.C. (1958). "Use of Market Specimens in Estimating Populations". Biometrika, 40, 783-791.

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

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

EFROM,R. and THISTED,R. (1975). "Estimating the Number of Unseen Species.(How many words did Shakespeare Know?). Technical Report N° 9, Stanford University. Division of Biostatistics.

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.

---------(1985). Estimation of the Number of Classes in a Population an the Coverage of a Sample", 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". J.Animal Ecology, 12, 42-58.

GIBBONS,J.D. y CHAKRABORTI,S.(1992). Nonparametric Statistical Inference, Marcel Dekker,Inc., Nueva York.

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. TOULMIN,G.H.(1956). "The Number of New Species and the Increase in Population Coverage. When a Sample is Increased", Biometrika, 43, 45-63.

GOODMAN,L.A.(1949). "On the Stimation of the Number of Classes in a Population", Annals of Mathematical Statistics, 20, 572-579.

GUPTA S. S. y BERGER, J.O. (1977). New York: Springer-Verlag, pág. 265-271.

HALL,P. (1991). "Bahadur Representation for Uniform Resampling and Importance Resampling with Applications to Assyntotic Relative Efficiency". THe Annals of Statistics, 19, 1062-1072.

HARRIS,B. (1959). "Determining Bounds on Integral with Applications to Cataloging Problems". Annals of Mathematical Statistics, 30, 521-548.

----------(1968). "Statistical Inference in the Clasical Occupancy Problem Umbiased Estimation of the Number of Classes". Journal of the American Statistical Association, 63, 837-847.

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.

HOU,W and Ozsoyoglu,G. (1991). "Statistical Estimators for Aggregate Relational Algebra Queries". ACM Transactions on Database Systems, 16, 600-654.

HOU,W, Ozsoyoglu,G. and Taneja, B.K. (1988). "Statistical Estimators for Relational Algebra Expressions", in Proceding of the ACM Symposium on Principles of Database Systems", pág. 276-287.

IVCHENKO,G.I. and Timonina,B.K. (1983) "Estimating the Size of a Finite Population". Theory of Probability and its Applications, 27, 403-406.

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

KALANTAR,A.H. (1987). "Using the Hiperbolic Model to Estimate Species Richness". Mathematical Geology, 19, 151-154.

KALININ,V.M. (1987). Functional Related to the Poisson Distributions and the Statistical Structura of a Test". Proceeding of the SteKlov Institute of Mathematics, 79, 6-19.

KEENER,R., Roothman,E. and Starr,N. (1987). "Distributions on Partitions". The Annals of Statistics. 15,1466-1481.

KNOT,M. (1967). "Models for Cataloguing Problems". The Annals of Mathematical Statistics, 38, 1255-1260.

KORWAR,R.M. (1988). "On the Observed Number od Classes from Multivariate Power Series and Hipergeometric Distributions". Sankhya the Indian Journal of Statistics, 50, 39-59.

KOVER,J.G., RAO,J.N.K. and WU,C.F.J. (1988). "Bootstrap and Other Metods for Measure Errors in Survey Estimates". Canadian Journal of Statistics, 16, 25-45.

LAST,G. and BRANDT,A. (1995). "Market Point Processes on the Real Line". Springer. New York.

LE,J. (1989). On Assyntotics for the NPMLE of the Probability of Discovering a New Species and an Adaptive Stopping Rule in Two-Stage Searches". Dissertation in the University of Winsconsin-Madison, Dep. of Statistics.

LEWINS,W.A. and Joanes, D.N. (1984). "Bayesian Estimations of the Number of Species". Biometrika, 40, 323-328.

LEWONTIN,R.C. and PROUT,T. (1956). "Estimation of the Number of Different Classes in a Population". Biometrika, 12, 211-223.

LLOYD,C.J. and YIP,P. (1991). "A Unification of Inference From Capture-Recapture Studies Through Maringale in Stimating Funtions". Ed. V.P. Godambe, Oxford,U.K.:Clarendon Press, pág. 65-88.

LO,H. y Wani,J.K.(1983). "Maximun Likelihood Stimation of the Parameters of the Invariant Abundance Distribution", Biometrics, 39, 977-986.

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

MACDONALD,J.D. (1987). "Condominium". Philadelphia: J.B.Lippincot.

MANN,Ch.C. (1991) "Exstinction: Are Ecologists Crying Wolf?". Science, 253, 709-824.

MARCHAND,J.P. and SCHROECK,F.E. (1982). "On the Estimation of the Number of Equally Likely Classes in a Population". Communications in Statistics. Part A-Theory and Methods, 11, 1139-1146.

MACNEIL,D. (1973). "Estimating an Author's Vocabulary". Journal of the American Statistical Association, 68, 92-96.

NAJOCK,D. (1973). Boostrap Experiments for the Evaluation of Espected Values and Variances of Vocabulary Sizes", in Methodes Quantitatives et Informatiques dans L'etude des Textes. Geneva:Slatkine-Champion, pág. 658-670.

Nayack,T.K. (1989). "A Note on Esyimating the Number of Errors in a System by Recapture Samplig". Statistics and Probability Letters, 7, 191-194.

NEE,S, HARVEY,P.H. and MAY,R.M. (1991). "Lifting the Veil on Abundance Patterns". Proceeding of the Royal Society of London, Ser.B, 243, 161-163.

NARAYAN,U.B. (1984). "Elements of Applied Stochastic Processes". John Wiley and Sons. New York.

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.

PALMER,M.W. (1990). "The Estimation of Species Richness by Extrapolation". Ecology, 71, 1195-1198.

PAPOULIS,A. (1984). "Probability, Random Variables, and Stochastics Processes". McGRAW-HILL. New York.

PATIL,G.P. and TAILLIE,C. (1982). Diversity as a Concept and its Measurement". Journal of the American Association, 77, 548-567.

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

POLLOT,K.H. (1991). "Moddelling Capture, Recapture and Renoval Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present and Future". Journal of the American Statistical Association, 86, 225-238.

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

PRESTON,F.W. (1948). The Commonness and Rarity of Species". Ecology, 29, 254-283.

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.

RIOS,S. (1967). "Métodos Estadísticos. Ediciones del Castillo. Madrid.

ROBBINS,H.E.(1968). "Estimating the Total Probability of the Unobserved Outcomes of an Speriment". Annals of Mathematical Statistics, 39, 256-257.

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.

SANKARAN,M. (1968). "Mixtures by the Inverse Gaussian Distribution". Sankhya B, 30, 455-458.

SCHROECK,F.E. (1981). "Tabulated Results of the Estimation of the Number of Dies of a Coin and the Analysis of a Hoard of Copper Falus of Taimur Shah". Numismatic Circular, 89, 37-38.

SHLOSER,A. (1981). On Estimation of the Size of Dictionary of a Long Text on the Basis of a Sample". Engineering Cybernetics, 19, 97-102.

SESHADRI,V.(1993). "The Inverse Gaussian Distribution", Clarendon Press, Oxford.

SHABAN,S.A. (1981). "Computation on the Poisson-Inverse Gaussian Distribution". Commun. Statistics, A10, 1389-1399.

SICHEL,H.S.(1975). "On a Distribution Law for Word Frequencies", Journal of the American Statistical Association, 72, 542-547.

----------(1982). "Assyntotics Efficiencies of Three Methods of Estimation for the Inverse Gaussian-Poisson Distributions". Biometrika, 49, 467-472.

----------(1986a). The GIPG Distribution Model with Applications to Physics Literature". Czechoslovak Journal of Physics, Ser. B 36, 133-137.

----------(1986b). "Parameter Estimation for a Word Frequency Distribution Based on Occupancy Theorhy". Communications in Statistics, Part A-Theory and Methods, 15, 935-949.

----------(1986c). "Word Frequency Distributions and Type-Token Characteristics". Mathematical Scientist, 11, 45-72.

----------(1992a). Anatomy of the Generalized Inverse Gaussian-Poisson Distribution with Special Applications to Bibliometrics Studies". Information Processing and Management, 28, 5-17.

----------(1992b). "Note on a Strongly Unimodal Bibliometric Size Frequency Distribution". Journal of the American Society for Information Science, 43, 299-303.

SLOCOMB,J., STAUFFER,B and DICKSON,K.L. (1977). "On Fitting the Trincated Lognormal Distribution to Species-Abundance Data Ussing Maximum Likelihood Esimation". Ecology, 58, 693-696.

STAM,A.J. (1987). "Statistical Problem in Ancien Numismatics". Statistica Neerlandica, 41, 151-173.

STUARD,A. and Ord,J.K. (1987). "Kendall's Advanced Theory of Statistics". Vol. 1 y 2. Charles Griffin. Londres.

SUBRAHMANIAM,K. (1990). "A Primer in Probability". Marcel Dekker. New York.

THISTED,R. and Effrom,B. (1987). "Did Shakespeare Write a Newly-Discovery Poem?". Biometrika, 74, 445-455.

WANI,J.K. y LO,H.P.(1983), A Characterization of Invariant Power-Series Abundance Distributions", Canadian Journal of Statistics, 11, 317-323, 1983.

WHITMORE,G.A. (1983). "A Regression Method for Censored Inversa-Gaussian Data". Canad. J. Statist.., 11, 305-315.

WILLIANS,C.B. (1964). "Patterns in the Balance of Nature". London and New York. Academic Press.

YIP,P. (1989). "An Inference Procedure for a Capture and Recapture Experiment with Time-Dependent Capture Probabilities". Biometrics, 45, 471-479.

--------(1991a). "Estimating Population Size From a Capture-Recapture Experiment with Known Renovals". Theoretical population Biology, 40, 1-13.

--------(1991b). "A Martingale Estimating Ecuation for a Capture-Recapture Experiment in Discrete Time". Biometrics, 47, 1081-1088.

--------(1991c). "A Method of Inference for a Capture-Recapture in Discrete Time with Variable Capture probabilities". Communications in Statistics-Stochastics Models, 7, 343-362.

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

ZIELINSKI,R. (1981). "A Statistical Estimate of the Structure of Multi-Extremal problems". Mathematical Prggraming, 21, 348-356.