UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

EL PROBLEMA DE MUESTREO DE ESPECIES DESDE UNA PERSPECTIVA BAYESIANA NO PARAMÉTRICA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

DIEGO IMANOL VALENZUELA FRANCO

DIRECTOR DE TESIS:
DR. RAMSÉS HUMBERTO MENA CHÁVEZ
2010

Hoja de Datos del Jurado

1. Datos del alumno
Valenzuela
Franco
Diego Imanol
51 71 68 78
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
406023526

2. Datos del tutor
Dr.
Ramsés Humberto
Mena
Chávez

3. Datos del sinodal 1
Dr.
Eduardo Arturo
Gutiérrez
Peña

4. Datos del sinodal 2
Dra.
Silvia
Ruiz-Velasco
Acosta

5. Datos del sinodal 3
Dr.
Carlos
Díaz
Ávalos

6. Datos del sinodal 4
M. en C.
Jésica
Hernández
Rojano

7. Datos del trabajo escrito.
El problema de muestro de especies desde una perspectiva bayesiana no paramétrica
81 p.
2010

Agradecimientos

A toda mi familia, por todo su apoyo, amor y cariño. A mis padres Diana y Eduardo, por todo el sacrificio y esfuerzo que han hecho por sus hijos. A mis abuelos, Tere y Carlos, que siempre han estado ahí. A mis hermanos Fermín y Ana, estoy orgulloso de ustedes.

A mis tías, Fabiola y Roxana, con su ayuda invaluable pude construir este camino. También a Carlos, Alicia, Libia, Olmo, Humberto y Hortencia.

A mis tías Verónica y Leticia, que sé siempre se acuerdan de mi.

A todos mis amigos, que incondicionalmente han estado a mi lado. A mi hermano Diego, a Victor, Sandra, Jorge, Yasmín, Maluk, Majo, Lidia, Anaid, Papu, Larissa, Ana Laura, Rosa, Fernanda; y a quien pueda estar dejando fuera, saben que no es por falta de aprecio.

A todos mis maestros que han sido parte de mi formación académica y personal. En especial a la Dra. Begoña Fernández y a la Dra. Ruth Fuentes.

Agradecimiento especial al Dr. Ramsés Mena por ser mi asesor en este trabajo, por su guía y consejos dentro y fuera de la tesis.

Por sus correcciones y comentarios al Dr. Eduardo Guitierrez, a la Dra. Silvia Ruiz-Velasco, al Dr. Carlos Díaz y a la Maestra Jésica Hernández; quienes aceptaron ser mis sinodales

Y gracias especiales a la facultad y a la UNAM. Es un orgullo enorme ser universitario.

¡GOYA! ¡GOYA!
¡CACHUN, CACHUN, RA, RA!
¡CACHUN, CACHUN, RA, RA!
¡GOYA!

¡¡UNIVERSIDAD!!

Índice general

Introducción

El muestreo y clasificación de especies ha sido un problema que ha interesado por mucho tiempo a la biología y ecología, pues es natural considerar la importancia que representa poder distinguir el número total de especies de cierta población biológica. En este caso la técnica de muestreo juega un papel importante, sobre todo si sólo tenemos acceso a una muestra de la población. Ahora pensemos en una segunda muestra hipotética, en este caso interesará predecir el número de nuevas especies que puedan aparecer, además de las ya observadas.

De forma resumida, lo anterior explica lo que es el problema de muestreo de especies (PME), tema sobre el cual se basará el presente trabajo. Existen diversos estudios que han revisado al PME desde diferentes puntos de vista. Se escogió el enfoque de la estadística bayesiana no paramétrica, área de estudio relativamente joven, debido principalmente a que ha mostrado buenos resultados en su aplicación con datos, sobre todo en labores predictivas.

Dentro de la tesis se explica con detalle el problema de muestreo de especies, en el capítulo uno se revisan las diversas aplicaciones que se le han dado al problema dentro y fuera del campo biológico, además de dar explicación sobre el problema que representa el manejo de poblaciones finitas e infinitas para muestreo e inferencia estadística.

Posteriormente, en el capítulo dos, se repasan estudios previos con enfoques varios que van desde la estadística clásica, pasando por el uso de procesos estocásticos hasta la inferencia bayesiana. Una parte muy importante se centra en explicar el sentido en el cual es utilizada la estadística bayesiana

no paramétrica para atacar el problema, así como el sustento teórico de las herramientas matemáticas utilizadas.

Una vez expuesta la teoría previa, en el capítulo tres, se procederá a revisar los enfoques más recientes que existen sobre el tema: se hablará sobre el proceso Dirichlet y el proceso Poisson-Dirichlet con dos parámetros, su utilidad y los estimadores bayesianos no paramétricos, dentro del contexto del PME, que se han logrado obtener a la fecha. Finalmente, y dado que el problema resurgió con un nuevo interés sobre el campo de la genética, se aplicarán en el capítulo cuatro dichos estimadores con datos reales obtenidos de muestras de marcadores de secuencia expresada [1] de librerías de ADN, para probar su funcionamiento y eficacia.

Se trata de un trabajo que pretende explicar y dar a conocer un problema probablemente poco difundido a nivel licenciatura, así como la investigación que se ha realizado sobre éste, ubicando sus alcances presentes y las posibilidades a futuro.

[1] También conocidos como *expressed sequence tags*, EST's, por su nombre en inglés

Capítulo 1

Preliminares

Por naturaleza, el hombre tiende a clasificar los elementos a su alrededor, probablemente sin un fin en particular más allá de mantener un 'orden' en su entorno, para luego estudiarlo y explicarlo. Los problemas surgen cuando el entorno crece y las cantidades a clasificar se escapan a las capacidades del observador o 'clasificador'. Métodos existen, y sobran, que facilitan llevar a buen término esta tarea; todo depende del contexto en que se trabaje y de la eficacia del método.

1.1. El problema de muestreo de especies (PME)

Sin más preámbulo, expliquemos el problema de muestreo de especies sobre el cual se trabajará:

Primero consideremos una población cualquiera cuyos elementos es posible clasificar en diversas clases disjuntas. Entendemos a una población, estadísticamente, como un conjunto de elementos de referencia sobre el que se realizan observaciones de estudio.

En principio nos es desconocida la cantidad de clases existentes. El poder llegar a estimar el número de clases en la población de interés, se convierte en un problema principal a resolver cuando no se tiene la información de toda la población.

Para esta tarea sólo tenemos disponible una muestra de la población, además de que se nos pueden presentar varias complicaciones, entre ellas el tamaño de la población total, el cual puede ser demasiado grande y, dependiendo del problema, teóricamente podríamos trabajar con poblaciones finitas o infinitas.

Dada la situación, se pueden escoger utilizar diferentes técnicas de muestreo. Por ejemplo, para el caso finito se puede aplicar un muestreo multinomial, que resulta en una técnica con reemplazo; o un muestreo sin reemplazo, como el muestreo hipergeométrico o Bernoulli. Si la población es infinita, se puede usar un muestreo multinomial o Bernoulli.

1.1.1. Estimando el número de clases

Supongamos que se obtiene una muestra de n elementos de una población particionada en K clases, donde K es un número desconocido. Se asume que las clases se podrán identificar fácilmente, por lo que cada elemento podrá ser emparentado con elementos de la misma clase.

Introduciendo formalmente la notación, la muestra obtenida, después de haber sido clasificada, puede representarse en el siguiente vector aleatorio de frecuencias: $\mathbf{n} := [n_1, \ldots, n_K]'$, donde n_i es el número de elementos de la i-ésima clase, $i = 1, \ldots, K$. Es importante notar que la i-ésima clase pertenece a la muestra si y sólo si $n_i > 0$, y no es posible conocer de la muestra cuales $n_i's$ son cero. Esto indica que $\mathbf{n}$ no es observable.

Ahora, si consideramos el siguiente vector aleatorio $\mathbf{k} := [k_1, \ldots, k_n]'$, donde k_j es el número de clases obtenidas j veces en la muestra; es decir, $k_j = \#\{n_i : n_i = j\}$, $j = 1, \ldots, n$; así obtendremos un vector observable. El problema es estimar K solamente a partir de $\mathbf{k}$.

Siguiendo con la notación anterior y considerando a c como el número total de clases en la muestra, tenemos que:

1. $k = \sum_{j=1}^{n} k_j$
2. $n = \sum_{i=1}^{K} n_i \Rightarrow n_i = \sum_{j=1}^{n} jk_j$

Es común considerar como hipótesis que todas las clases son del mismo tamaño, situación que se conoce también como tratamiento homogéneo. Además hay que notar que K es un número desconocido, dado desde el inicio y representa **todas** las clasificaciones de la población, que en otro contexto podremos llamar especies.

Existen varias teorías y técnicas de muestreo, dependiendo del contexto del problema, sin que se pueda afirmar que exista una forma de muestreo óptima. La elección final dependerá del problema o hipótesis planteadas, de los supuestos propuestos, la experiencia adquirida y la facilidad de uso de la herramienta.

La problemática del muestreo se centra en parte a la elección de la técnica adecuada, además de determinar un tamaño adecuado de muestra. En el caso de la estimación de clases, diferentes tipos de muestreo afectarán directamente los estimadores que aproximen el número de clases en una población.

Se le hace referencia como 'El problema de muestreo de especies' dadas las aplicaciones que se le han dado en el campo biológico y ecológico para la estimación del número de especies. Se trata de un problema con amplia historia y que recientemente ha vuelto a generar interés por su importancia en el campo genético.

1.1.2. Labor predictiva

Como segunda parte, consideremos una extensión al problema de estimar el número de clases. Suponiendo que se ha obtenido un estimador confiable del número de clases después de un primer muestreo, o partiendo del hecho de

que el problema permite conocer esta información, surgen los cuestionamientos: ¿existirán aún clases desconocidas?, si es así, ¿cuál es la probabilidad de encontrar nuevas clases y el valor esperado del número de nuevas clases al realizar un segundo muestreo?, y finalmente ¿cuál es la probabilidad de obtener una nueva clase en un tiempo determinado?.

En sí estos cuestionamientos pueden ser un tanto simples y vagos en primera instancia; para lograr mayor claridad habrá que sintetizar un poco las ideas:

Consideremos una muestra de tamaño n, digamos $X_1, X_2, \ldots, X_n$; suponemos que podemos exhibir $K_n \in 1, \ldots, n$ especies distintas. Como cada especie o clase distinta puede aparecer en más de una ocasión, cada una será exhibida con frecuencia aleatoria $(N_1, \ldots, N_{K_n})$; la frecuencia es aleatoria dado que se obtienen de una muestra aleatoria, donde tenemos que $\sum_{i=1}^{K_n} N_i = n$ c.s.

Ahora, dada la primer muestra, interesa estimar el número de nuevas especies que se podrán observar si se realiza un nuevo muestreo. Esto es, dado el muestreo de tamaño n y un nuevo muestreo de tamaño m, obtendremos $K_m^{(n)} := K_m - K_n$ especies distintas a las observadas en el primer muestreo.

Esto es lo que consideraremos finalmente como el problema de descubrimiento de nuevas especies. Y lo que interesará conocer es:

$$\mathbb{P}(K_m^{(n)} = k | K_n = j)$$

de donde podemos obtener

$$\mathbb{E}(K_m^{(n)} | K_n = j),$$

la probabilidad de descubrir k nuevas clases y el valor esperado del número de nuevas clases, respectivamente. A esto se le relaciona el problema de determinar el tamaño óptimo de la muestra, tanto para la muestra principal como para la adicional, sobre todo trabajando en un marco de muestreo de especies.

Ésta es la problemática principal sobre la cual tratará este trabajo. Al respecto existe una amplia documentación sobre los diferentes acercamientos

teóricos y metodológicos que se han llegado a implementar, de los cuales se comentará más adelante. En lo particular, este trabajo se centrará en el enfoque dado desde el punto de vista de la estadística bayesiana no paramétrica, teoría que ha probado tener buenos resultados en su aplicación a problemas reales y se basa en supuestos relativamente flexibles.

1.1.3. Dónde se aplica

El problema de muestreo de especies tiene diversas aplicaciones al mundo real; como ya fue mencionado, tiene importantes referencias sobre estudios ecológicos y biológicos. Sin embargo también existe ingerencia en otros campos, tales como educación, numismática, sistemas, bases de datos (resaltando el caso de datos duplicados), estudio de algoritmos, lingüística, entre otros.

¿En general qué es lo que interesa? En la mayoría de los casos a los investigadores, independientemente del área de estudio en que trabajen, les interesa conocer el número de 'especies perdidas' o especies desconocidas dado un muestro básico de la población de interés. Ésto con la intención de obtener una predicción acertada que indique si existe la necesidad de realizar un nuevo muestreo, lo cual adquiere importancia dado el costo que supone realizar muestras dependiendo del campo en que se trabaje.

En el caso de la lingüística es posible estimar el número total de palabras que un autor conoce pero no empleó en sus obras, y por lo tanto desconocemos. Tenemos un ejemplo llamativo mencionado por Efron y Thisted (1976) [7] y Efron (2003) [6], en el cual los autores se fijan en la obra escrita de William Shakepeare. Siendo un exponente emblemático de la lengua inglesa, resulta interesante preguntarse por el número de palabras (su vocabulario) que el autor sabía pero no utilizó. Dentro del problema de muestreo de especies se cuestiona por el número de especies no observadas, en este caso las especies se intercambian por las palabras.

De estudios previos, se sabe que dentro del *canon* de Shakespeare (la lista de obras que se ha demostrado fueron escritas por el autor) escribió un gran total de 884,647 palabras, las cuales podemos clasificar tal como se muestra en el cuadro 1.1, donde x es el número de veces en que apareció una

palabra distinta, es decir, tenemos que 14,376 palabras distintas aparecieron exactamente una vez, 4,343 aparecieron exactamente 2 veces, etc.

Así se obtiene que el autor de obras como Hamlet, utilizó un total de 31,534 palabras distintas. Utilizando un modelo bayesiano empírico, Efron encontró que la probabilidad, bajo su modelo, de encontrar una nueva palabra no existente dentro del *canon* de Shakespeare, en el supuesto de encontrar una nueva obra, resultó ser ≈ 0.016.

Cuadro 1.1: Frecuencia de las distintas palabras utilizadas por William Shakespeare

x	1	2	3	4	5	6	7	8	9	10	Total
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	7	12	9	8	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78
100+											846
											31534

Para la ecología, la aplicación del problema resulta tener un rango de implementación más amplio, puesto que la idea fundamental es encontrar especies desconocidas por el hombre en un cierto contexto. Ejemplos existen muchos, pues hipotéticamente podemos tomar cualquier región natural en la cual se sospeche de una alta densidad poblacional animal y/o vegetal. El PME puede ser aplicado con la intención de estimar el número de especies localizadas en esa región, además de poder predecir el número de especies no observadas; esto con el fin de ahorrar trabajos extensivos de observación, que en este caso puede consumir demasiados recursos.

Brevemente, como ejemplo práctico, podemos referirnos a la página electróni-

ca de la doctora Anne Chao (http://chao.stat.nthu.edu.tw/indexE.html), investigadora de la universidad nacional de Tsing Hua, en Taiwan. En ésta, al momento de escribir esta tesis, en el portal principal se explica un pequeño ejemplo: se tiene el dato que en Taiwan existen 458 especies de pájaros; en las cercanías del estuario de Ker-Yar se pudieron observar 155, pero la estimación regresa un total de 180 especies en el área, arrojando una diferencia de especies no observadas. Más allá de este ejemplo, en la página también se encuentra documentación interesante sobre el PME.

En el caso de la biología y de la microbiología, recientemente ha habido un renovado interés en el área de la genética. En un caso específico encontramos a los microarreglos, herramienta utilizada en la biología molecular que permite observar cambios en los niveles de expresión de los genes, esto es, qué tan activo resulta un gen en determinada célula.

Ligado a los microarreglos, tenemos otra aplicación en la genética. En este caso, encontramos lo que se denomina como marcador de secuencia expresada o *expresed sequence tags*, EST por su acrónimo en inglés, que en resumen son pequeñas sub-secuencias de una secuencia nucleotídica transcrita (codificante de una proteína o no), las cuales se obtienen de secuenciar librerías ADNc[1] consistente en millones de genes; siendo más específicos, se trata de clones consistentes en ADN que es complementario al ARNm[2]. Son utilizados para descubrimiento, predicción y mapeo de genes o determinación de secuencias[3], entre otros.

Dado que los EST's son pequeñas partes de información de ADN, y debido a costos, los investigadores se interesan en conocer estimadores para la probabilidad de descubrir un nuevo gen, con el fin de ayudar en la decisión de realizar, o no, secuencias adicionales. También, en casos donde la secuenciación ya no es viable, la estimación se utiliza para determinar el genoma.

[1] ADN complementario
[2] ARN mensajero
[3] como lo realizado para construir el mapa del gen humano

1.2. Población de una muestra

La palabra población se utiliza para denotar el agregado del cual la muestra a estudiar será obtenida. Así mismo, la población que será muestreada (*población muestra*) deberá coincidir con la población de la cual se desea obtener información (*población objetivo*). En ocasiones, por cuestiones de practicidad o conveniencia, la población muestra está más restringida que la población objetivo. Si es el caso, debe tomarse en cuenta que las conclusiones obtenidas serán sólo aplicables a la población muestra. El extender estas conclusiones a la población objetivo dependerá de otras fuentes de información.

El estudio, desde el punto de vista estadístico, de una población cualquiera podría resultar demasiado laborioso si no hiciéramos uso de la herramienta del muestreo, la cual permite obtener una muestra o subconjunto de casos o individuos de la población. Para que la muestra sea de utilidad, ésta tiene que permitir interpolar o inferir la mayoría de las propiedades de la población total. En otras palabras, la muestra tiene que ser representativa, de lo cual nunca estaremos totalmente seguros y dependerá directamente de la técnica de muestreo. Es importante verificar que la información a obtener es relevante para el estudio y que ninguna información está siendo omitida.

Resulta de gran importancia definir el tamaño de población sobre la que se esté trabajando, acotando: existen poblaciones finitas e infinitas. Definir este aspecto repercute directamente sobre el trabajo de investigación, puesto que cada una de ellas debe ser tratada de diferente manera aplicando distintos métodos y modelos estadísticos. En algunos casos de aplicación, el tema de los recursos influye junto con estos aspectos.

Al hablar de poblaciones finitas e infinitas, surge una duda básica: ¿cómo es posible pensar en la existencia de poblaciones infinitas? En el caso de la bioestadística, la duda es razonable, por ejemplo, pensando en el estudio de poblaciones de especies conocidas cuyo número es delimitado. Sin embargo existen casos en que el tamaño de población es inimaginable, muy grande, o por diversas cuestiones no contable; es cuando se hace notar la necesidad de suponer una población infinita.

1.2.1. Población finita

Se describe como aquella población cuyo número de elementos se mantiene finito durante un tiempo particular. Su número es imaginable y manejable, en otras palabras: contable.

Dentro de los estudios más comunes aplicados sobre este tipo de poblaciones se encuentran las encuestas y los cuestionarios, que más bien hacen referencia a estudios muestrales sobre la población, y los censos, que implican un estudio global de la misma. Una forma sencilla de estudio de estas poblaciones se da cuando se hace uso de muestras con reemplazo, lo cual no siempre sucede.

El estudio de poblaciones finitas generalmente (y es más cierto con poblaciones biológicas) se restringe a un tiempo en particular, se describen procesos para una determinada ventana de tiempo, lo cual quita cierta generalidad a su interpretación.

1.2.2. Población infinita

Como ya se mencionó, el estudio de una población infinita surge cuando se tienen poblaciones cuyo número de elementos es inimaginable y poco manejable; simplificando, se trata de aquella población cuya cantidad de elementos es no contable. Ejemplos los encontramos en economía, biología e ingeniería, entre otros.

El estudio para estas poblaciones se simplifica cuando en el muestreo las observaciones son estocásticamente independientes y siguen una misma distribución teórica.

En el caso infinito, la interpretación de su estudio tiene un carácter más amplio y general, al no estar ligado directamente a un sector o elemento en particular o a un tiempo específico.

El campo genético es un buen ejemplo de manejo de poblaciones infini-

tas; si bien el número de genes puede representar un número finito, en la aplicación considerar una población infinita puede facilitar ciertos cálculos y razonamientos.

Capítulo 2

Enfoques para modelar el PME

Existen diferentes modelos con los cuales es posible realizar un acercamiento al problema de muestreo de especies. La diversidad de los modelos parte de dos supuestos principales: tamaño de la población (finita o infinita) y los métodos de muestreo. Por otra parte, los modelos existentes se extienden a través de variadas ramas de la estadística: estadística clásica o frecuentista, el enfoque bayesiano, vía procesos de Poisson (solución más apegada a herramientas probabilísticas) y solución vía distribuciones aleatorias, este último relacionado con un enfoque bayesiano no paramétrico.

2.1. Enfoque frecuentista

La documentación de este apartado se basa ampliamente en el artículo elaborado por Bunge y Fitzpatrick (1993) [2], en el cual se describen y resumen los métodos con enfoque frecuentista más estudiados.

Para cada escenario se describen distintos modelos y estimadores para el valor de K (el número de clases en que se divide la población), señalando sus respectivas fortalezas y debilidades. A continuación se enlistarán dichos escenarios, de los cuales parten los modelos.

Población finita, muestreo hipergeométrico

Este modelo fue propuesto por Goodman (1979) [14]. Supongamos que la población es finita con tamaño conocido T. Sea N_i la variable que denotará el número de unidades en la clase i-ésima, con $i = 1, \ldots, K$, tenemos entonces que $\sum_{i=1}^{K} N_i = T$, y consideraremos $M = \text{máx}_{1 \leq i \leq K}\{N_i\}$. Si obtenemos una muestra de n elementos de forma aleatoria sin reemplazo, entonces $\mathbf{n}$ tendrá una distribución hipergeométrica múltiple, con respectiva función de masa

$$p_n(\mathbf{n}) = \binom{T}{n}^{-1} \times \prod_{i=1}^{K} \binom{N_i}{n_i}.$$

Recordemos que $\mathbf{n}$ es el vector aleatorio de la forma $[n_1, \ldots, n_K]'$, donde n_i es el número de elementos de la i-ésima clase, bajo la muestra de tamaño n $(T \geq n)$.

Para este modelo, si $n \geq M$, entonces existe un único estimador insesgado para K; en caso contrario no es posible conseguir tal estimador. El estimador (cuando $n \geq M$) es

$$\hat{K}_{GOODMAN1} = k + \sum_{j=1}^{n} (-1)^{j+1} \frac{(T-n+j-1)!(n-j)!}{(T-n-1)!n!} k_j,$$

donde k_j, j-ésimo elemento del vector observable $\mathbf{k}$:$[k_1, \ldots, k_n]'$, es el número de clases obtenidas j veces en la muestra y $k = \sum_{j=1}^{n} k_j$.

Aunque el estimador, que al ser único entonces también es un estimador insesgado uniformemente de mínima varianza (UMVUE), en algunos casos su varianza es demasiado grande, de tal forma que se llega a desechar su utilidad.

Tomando como base este estimador, Shlosser utilizó una aproximación asintótica en donde $T, n \to \infty$ de tal forma que $\frac{n}{T} \to q \in (0, 1)$. De esta forma obtuvo lo siguiente:

$$\hat{K}_{SHLOSSER} = k + k_1 \left(\sum_{i=1}^{n} iq(1-q)^{i-1} k_i \right)^{-1} \sum_{j=1}^{n} (1-q)^j k_j,$$

de donde $\hat{K}_{SHLOSSER} \geq k$.

En este caso no se calculó el sesgo o varianza del estimador, sin embargo en simulaciones con porcentajes de muestreo cercanos al 10 % funcionó razonablemente bien. Hay que agregar que un estimador insesgado, obtenido a partir de aproximaciones asintóticas, aparentemente mejora al UMVUE ya mencionado, aunque, esto no se comprobó formalmente y las simulaciones de Shlosser no fueron extensivas.

Población finita, muestreo Bernoulli

Supóngase ahora que los T elementos pertenecientes a la población se introducen a la muestra de forma independiente, cada uno con probabilidad p. De esta forma, el tamaño total de la muestra se puede tomar como una variable aleatoria binomial (T, p), donde la i-ésima clase contribuirá independientemente con un número aleatorio de elementos a partir de una variable aleatoria binomial (N_i, p). Esto suponiendo que las clases o especies son independientes.

En este caso la función de masa la podemos ver como

$$p_n(\mathbf{n}) = p^n(1-p)^{T-n} \prod_{i=1}^{K} \binom{N_i}{n_i},$$

donde $\sum_{i=1}^{K} n_i = n$. Suponiendo que se conoce el valor de p, Goodman aportó un estimador insesgado diferente para K:

$$\hat{K}_{GOODMAN2} = k + \sum_{j=1}^{n} (-1)^{j+1} \left(\frac{1-p}{p}\right)^j k_j.$$

La utilidad de este estimador resulta no ser muy buena, presentándose propiedades indeseables similares a $\hat{K}_{GOODMAN1}$.

Otra propuesta, de mayor utilidad, fue dada por Esty, que consideró el muestreo Bernoulli en el contexto de un modelo de 'superpoblación', donde

$N_1, ..., N_K$ son variables aleatorias independientes idénticamente distribuidas (v.a.i.i.d.) binomiales negativas con parámetros (θ_1, θ_2) y función de masa

$$p_{N_i}(N_i) = \Gamma(\theta_1 + N_i) \times \theta_2^{\theta_1}(1-\theta_2)^{N_i}/(\Gamma(\theta_1 N_i!).$$

Entonces $n_1, \ldots, n_k$ son variables aleatorias binomiales negativas $(\theta_1, \theta_2/(\theta_2 + p - \theta_2 p))$ i.i.d. Esto es un caso especial de una distribución invariante de la abundancia [1]. Bajo estas bases Esty produjo el siguiente estimador,

$$\hat{K}_{BN} = \frac{n}{\hat{\mu}},$$

para valores conocidos de θ_1, donde p y θ_2 están siendo implícitamente estimados, y donde $\hat{\mu}$ es la solución a la ecuación $n/k = \mu/(1-(1+\mu/\theta_1)^{-\theta_1})$. Sin embargo, a partir de simulaciones, se concluyó que $\hat{K}_{BN}$ no es muy recomendable.

Población infinita, muestreo multinomial

En este caso, supongamos que obtenemos una muestra aleatoria de n elementos a partir de una población infinita particionada en K clases bajo el siguiente vector de proporciones $\pi = [\pi_1, \ldots, \pi_K]'$, con $\sum_{i=1}^{K} \pi_i = 1$. En este caso, $\mathbf{n}$ tiene una distribución multinomial K-dimensional, con función de masa

$$p_n(\mathbf{n}) = \binom{N_i}{n_i, \ldots, n_K} \prod_{i=1}^{K} \pi_i^{n_i}, \sum_{i=1}^{K} n_i = n.$$

Población Infinita, muestra multinomial, clases del mismo tamaño

Al dividir el número de clases, podemos asumir que éstas son del mismo tamaño, o bien, cada clase tiene una distribución distinta. Como ya se mencionó anteriormente, es común asumir la hipótesis que acepta clases de igual tamaño, que aunque no se acerca a modelos reales, es bastante manejable.

[1] *Invariant abundance distribution*

Sea $H_{=} : \pi_1 = \cdots = \pi_K = K^{-1}$, que representa la hipótesis ya mencionada. Tenemos que problemas clásicos como «el coleccionista de cupones» son ejemplos de muestreo multinomial bajo $H_{=}$. Bunge y Fitzpatrick mencionan que es difícil dar un recuento de los modelos que estiman K de forma satisfactoria; en este caso particular sólo serán mencionados de forma resumida los resultados más importantes:

Estimador de Máxima Verosimilitud

En este caso se tiene que el estimador de máxima verosimilitud, para K, denotado como $\hat{K}_{EMV=}$, se puede obtener a partir de la solución de K^* dada la ecuación

$$k = K^*(1 - e^{-n/K^*}).$$

UMVUE

Si se conoce que $n \geq K$, enconces el estimador se denota como

$$\hat{K}_{UMVUE=} = \frac{S_{k,n+1}}{S_{k,n}},$$

donde $S_{i,j}$ es el número de Stirling de segunda especie [2]. Se ha mostrado que asintóticamente $\hat{K}_{EMV=} \approx \hat{K}_{UMVUE=}$.

Cobertura

La *cobertura*, U, de una muestra es la suma aleatoria de las π_i's correspondientes a las clases observadas. La cobertura se toma como la proporción de la población representada en la muestra aleatoria, lo cual nos sirve como indicador para determinar si se obtuvieron, en la muestra, todas (o casi todas) las clases de la muestra. Entonces

$$U = \sum_{i=i}^{K} \pi_i \, \mathbb{I}_{n_i > 0}.$$

[2]Se definen como la cantidad de maneras que existen de hacer una partición de un conjunto de n elementos en k subconjuntos.

Bajo $H_=$, $U = k/K$; si se obtiene un estimador $\hat{U}$, el estimador que aproxime a K será dado por $k/\hat{U}$. Good y Toulmin (1956) [13], propusieron el siguiente estimador $\hat{U_{GOOD}} = (1 - k/n)$. Entonces, bajo $H_=$, se obtiene:

$$\hat{K}_{COV=} = \frac{k}{\hat{U_{GOOD}}}.$$

Población infinita, muestra multinomial, modelos paramétricos

En las aplicaciones es poco realista pensar en clases del mismo tamaño; la dificultad se muestra cuando la variación entre tamaños es grande. Para trabajar este problema se han propuesto un par de modelos paramétricos.

El primero asume que las π_i's tienen una estructura funcional, dependiente de un número reducido de parámetros; es decir, $\pi_i = f(i; \theta, K)$, $i = 1, \ldots, K$, donde θ es un vector parámetro y f es decreciente en los valores de i. En este caso no existe un estimador concreto o que sea fácil de calcular.

En el segundo modelo paramétrico se puede aproximar el histograma de las π_i's por medio de una función de densidad que depende de algún parámetro θ. Sichel desarrolló un modelo de este estilo, utilizando la distribución gaussiana inversa generalizada [3]. Se encontró que la expresión obtenida dependía de 3 parámetros, uno de los cuales podía fijarse en casos de aplicación, obteniéndose la función de densidad de una gaussiana inversa como sigue

$$\psi(\pi; \theta_1, \theta_2) = \theta_1 \sqrt{\theta_2} exp\{\theta_1 - \frac{\pi}{\theta_2} - \frac{\theta_1^2 \theta_2}{(4\pi)}\}/(2\sqrt{Pi}\pi^{3/2}),$$

donde $Pi \approx 3.1416$.

Si la distribución de las π_i's está aproximadamente dada por $\psi(\pi; \theta_1, \theta_2)$, entonces $K \approx (\mathbb{E}_\psi(\pi))^{-1} = 2(\theta_1\theta_2)^{-1}$. Dado que la probabilidad de que una clase arbitraria puede aparecer j veces en una muestra aleatoria de tamaño n

[3] También conocida como la distribución Wald, es una familia de distribuciones continuas con soporte en $(0, \infty)$ consistente en 2 parámetros:
$f(x) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}} \mathbb{I}_{(0,\infty)}(x)$

está aproximadamente dada por una distribución Poisson ψ-mezclada, Sichel encontró los estimadores $(\hat{\theta_1}, \hat{\theta_2})$, y así obtuvo el estimador

$$\hat{K}_{SICHEL} = \frac{2}{\hat{\theta_1}\hat{\theta_2}}.$$

Hay que marcar que, dadas las pruebas realizadas sobre el estimador, el tamaño de muestra a utilizar debería ser mayor a 1500 elementos. De las dos opciones paramétricas, es esta última la que resulta mejor en aplicaciones.

Población infinita, muestra multinomial, modelos no paramétricos

Los modelos no paramétricos se centran en estimar K sin ningún supuesto acerca de π. Dado ésto, no se ha encontrado un estimador insesgado. Sin embargo Anne Chao (1984) [3], usando estimadores de los momentos de la forma $\mathbb{E}(k_i)$, obtuvo una cota inferior para estimadores no paramétricos de K,

$$\hat{K}_{CHAO1} = k + \frac{k_1^2}{2k_2}.$$

Posteriormente, Chao y Lee (1992) [4] usaron la idea de la cobertura para encontrar el estimador

$$\hat{K}_{CHAO2} = \frac{k}{\hat{u}_{GOOD}} + \frac{n(1-\hat{u}_{GOOD})}{\hat{u}_{GOOD}}\hat{\gamma}^2,$$

donde $\hat{\gamma}$ es el estimador del coeficiente de variación γ de las $\pi_i's$. Entonces el estimador anterior es una variación de $\hat{K}_{COV=}$ más una corrección sesgada que depende de $\hat{\gamma}$.

Población infinita, muestreo Poisson

Ahora suponga que el número de representantes de la i-ésima clase en la muestra son variables aleatorias independientes Poisson con media λ_i, $i =$

$1, \ldots, K$. En este caso la muestra total es una variable aleatoria Poisson con media $\lambda = \sum_{j=1}^{K} \lambda_j$, con función de masa

$$p_n(\mathbf{n}) = \exp -\lambda \prod_{i=1}^{K} \frac{\lambda_i^{n_i}}{n_i!}.$$

En este caso, para poder encontrar un estimador, habrá que asumir que las λ_i's son en sí una muestra aleatoria de alguna distribución F; entonces $\mathbb{E}(k) = C(1 - p_0(F))$, donde $p_0(F)$ es la probabilidad de que una variable aleatoria Poisson F-mezclada sea igual a cero. Así, dado un estimador $\hat{p_0(F)}$, el estimador de K es

$$\hat{K}_{POISSON} = \frac{k}{1 - \hat{p_0(F)}}.$$

Bajo un acercamiento empírico bayesiano no paramétrico, Efron y Thisted (1976) [7] presentaron ideas para aproximar un estimador de K. Sea $Kp_0(F)$ el valor esperado del número de clases no observadas; se propusieron formas para obtener un valor K^* y una distribución F^* tal que minimicen $Kp_0(F)$. Esto resulta en un una cota inferior estimada para K:

$$\hat{K}_{ET} = k_{adj} + K^* p_0(F^*),$$

donde k_{adj} es una versión ajustada de k. Se amplia la idea de este modelo más adelante.

Población infinita, muestreo Bernoulli múltiple

Teniendo una población infinita (bajo el mismo supuesto de estar particionada en K clases), supóngase que es observada en n ocasiones (o bien, por n observadores), en cada ocasión cada una de las clases es observada (o no). Dado ésto, la muestra puede ser representada por una matriz $[x_{ij}]$ de dimensión $K \times n$, donde la entrada $x_{ij} = 1$ si la i-ésima clase es observada por j-ésima ocasión, $i = 1, \ldots, K$ y $j = 1, \ldots, n$. Sólo filas con al menos un 1 son observadas, de hecho, k_l = número de filas con exactamente l 1's.

Sea un modelo en el que se consideran a las entradas x_{ij} todas independientes, con $\mathbb{P}(x_{ij} = 1) \equiv \pi_i$ y $j = 1, \ldots, n$. Entonces la contribución de cada clase a la muestra, $n_i = \sum_{j=1}^{n} x_{ij}$, es una variable aleatoria binomial con parámetros (n, π_i) y función de masa

$$p_n(\mathbf{n}) = \prod_{i=1}^{K} \binom{n}{n_i} \pi_i^{n_i} (1 - \pi_i)^{n - n_i},$$

donde las $\pi_i's$ las tomamos como una muestra aleatoria de alguna distribución F. Con estos supuestos se obtiene el estimador *jackknife* [4] $\hat{K}_{BOc}$, por ejemplo para $c = 1$

$$\hat{K}_{BO1} = k + \binom{n-1}{n} k_1.$$

2.2. Enfoque bayesiano

2.2.1. Modelo de Hill

Existen propuestas bayesianas desde el punto de vista paramétrico. Primero consideramos el modelo propuesto por Bruce M. Hill de la universidad del Michigan y que es resumido por Bunge y Fitzpatrick (1993) [2].

Tómese la siguiente función de distribución inicial:

$$(K, N_1, \ldots, N_K) : \prod (K, N_1, \ldots, N_K) =$$

$$\prod_{N_1, \ldots, N_K | K, T} (N_1, \ldots, N_K | K, T) \prod_{K,T} (K, T) = \binom{T-1}{K-1}^{-1} \prod_{K,T} (K, T),$$

donde $\prod_{K,T}(K, T)$ es una distribución arbitraria en $\mathbb{T} \times \mathbb{T}$. Esta expresión inicial o *a priori* no da mucha información sobre si cada partición dividida en C clases es igualmente probable.

[4] La técnica de *jackknife*, básicamente hace referencia a una técnica de remuestreo.

Basándose en una población finita a partir de un muestreo multinomial y el modelo anterior, hay que fijar T y tomar a

$$\prod_{K,T}(K,T) = \prod_{K}(K)$$

como una distribución binomial negativa truncada, tal que

$$\prod_{K}(K) \propto \binom{K+\theta_1-1}{K}\theta_2^K,$$

con $K = 1,\ldots,T$; $\theta_1 \in (0,\infty), \theta_2 \in (0,1]$. Con estos supuestos, es posible obtener una función *a posteriori* para el caso de poblaciones infinitas; esto mientras $T \to \infty$.

Se puede extender este modelo, tomando como distribución inicial

$$\prod_{K,\pi}(K,\pi) = D(\pi; K, \theta_3),$$

donde $D(\pi; K, \theta_3)$ es la densidad simétrica Dirichlet K-dimensional, con parámetros $\theta_3 \in (0,\infty)$ y $\prod_K(K)$ es una función *a priori* en $\mathbb{T}$. De lo anterior se deriva la siguiente función a posteriori $p_{K|\mathbf{k}}(K|\mathbf{k}) = p_{K|k}(K|k)$. En particular, si $\prod_K(K)$ es la distribución binomial negativa truncada, entonces

$$p_{K|\mathbf{k}}(K|\mathbf{k}) = p_{K|k}(K|k) \propto \theta_2^C \binom{K+\theta_1-1}{K}\binom{K}{k} / \binom{\theta_3 K+n-1}{n}.$$

Se consideró a la moda de $p_{K|k}(K|k)$ como el estimador de K.

2.2.2. El modelo de Efron y Thisted

Es importante mencionar el trabajo de Efron y Thisted (1976) [7] dada su relevancia histórica. El artículo *Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?* es uno de los primeros estudios

en tratar el problema de muestreo de las especies, por medio de un ejemplo práctico, a la vez de resultar interesante.

En el apartado 1.1.3 de esta tesis *Dónde se aplica*, se habló un poco sobre el problema, que básicamente se puede resumir de la siguiente forma: de las palabras que Shakepeare conocía como escritor, ¿cuántas palabras utilizó en sus obras? y ¿cuántas más pudo haber utilizado de haber escrito una nueva obra? Este también corresponde un primer acercamiento para resolver el problema de estimar el descubrimiento de una nueva especie dentro del PME.

De las obras de Shakespeare, se conoce que utilizó un total de 884,647 palabras; definamos a n_x como el número de diferentes tipos de palabras que aparecen exactamente x veces (vease la tabla 1.1), con $(x = 1, 2, \ldots)$. Entonces tendremos un total de $\sum_{x=1}^{\infty} n_x = 31,534$ diferentes tipos de palabras dentro del trabajo conocido de Shakespeare (o *canon*).

Ahora supongamos que se descubre una cantidad considerable de trabajos desconocidos del autor; poniéndolo en números supóngase un total de $884,647 * t$ palabras, donde t se entiende como una unidad de tiempo. Entonces ¿cuántos tipos diferentes de palabras Shakespeare conocía además de las 31,534? En este caso las especies no descubiertas son las palabras que Shakespeare sabía pero no utilizó. La primer cuestión es fácil de determinar revisando sus escritos, la segunda constituye el problema a resolver.

Utilizando un acercamiento bayesiano a partir de las ideas de Fisher (1943) [11], que provee una idea paramétrica, y el modelo no paramétrico de Good y Toulmin (1956) [13], se da un acercamiento a la solución del problema. A continuación se resumen las ideas y herramientas utilizadas por Efron y Thisted.

Modelo Básico

Partiendo de un modelo básico, supongamos que existen K especies y que después de realizar una captura (en otras palabras un muestreo), de tamaño n en una unidad de tiempo, entonces habremos capturado n_i individuos de la especie i, con $i = 1, \ldots, K$. Como en modelos anteriores, sólo habremos observado aquellos valores n_i mayores que cero. En este caso, asumimos que los individuos fueron capturados de acuerdo a una distribución Poisson con

media λ_i, es decir, se propone que el uso de procesos Poisson para la descripción de la captura. Los individuos aparecen en la muestra de acuerdo a un proceso Poisson no homogeneo.

Para el modelo habría que suponer el periodo de muestreo o captura desde el tiempo -1 hasta el tiempo 0. La idea es extrapolar las cuentas de $[-1, 0]$ a un tiempo t en el futuro. Sea $n_i(t)$ el número de veces que la especie i aparece en el periodo completo $[-1, t]$. El asumir un proceso Poisson implica lo siguiente:

(i) que $n_i(t)$ tiene una distribución Poisson con media λ_i,

(ii) dado $n_i(t)$, n es condicionalmente binomial con parámetros $(n_i(t), \frac{1}{(t+1)})$.

Sea $G(\lambda)$ la función de distribución empírica de los números $\lambda_1, \ldots, \lambda_K$. Además, sea k_j, $j = 1, \ldots, n$, el número de especies observadas exactamente j veces en $[-1, 0]$, donde

$$\eta_j = \mathbb{E}(k_j) = K \int_0^\infty (e^{-\lambda}\lambda^j/j!) dG(\lambda), \tag{2.1}$$

y sea $\Delta(t)$ el número esperado de especies observadas en $(0, t]$ pero no en [-1,0]:

$$\Delta(t) = K \int_0^\infty e^{-\lambda}(1 - e^{-\lambda t}) dG(\lambda). \tag{2.2}$$

Lo que se desea estimar es $\Delta(t) = K$, el número esperado de especies a obtener en las próximas t unidades de tiempo. No sobra remarcar que esta última meta es intrínseca a la descripción que se dio sobre el problema de muestreo de las especies.

Si se substituye la expansión

$$1 - e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \cdots$$

en (2.2) y lo comparamos con (2.1), obtenemos que

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \dots . \tag{2.3}$$

Si se asume convergencia en el lado derecho de la expresión anterior, es posible obtener el siguiente estimador insesgado de $\Delta(t)$

$$\hat{\Delta}(t) = k_1 t - k_2 t^2 + k_3 t^3 - \dots .$$

En el caso práctico de Shakespare, si $t = 1$ el estimador es $\hat{\Delta}(1) = 11430$.

Sin embargo la expresión (2.3) no es funcional para valores de t mayores a uno. El crecimiento geométrico de t^x produce oscilaciones considerables al aumentar el número de términos.

Modelo Binomial Negativo

De Fisher se toma el siguiente modelo. Al modelo básico se le agregan las siguientes suposiciones:

1. La función de distribución $G(\lambda)$ se aproxima por una distribución Gamma con función de densidad:

$$g_{\alpha\beta}(\lambda) = \{\beta^\alpha \Gamma(\alpha)\}^{-1} \lambda^{\alpha-1} e^{-\lambda/\beta}. \tag{2.4}$$

2. Los parámetros $\lambda_1, \dots, \lambda_K$ son independientes e idénticamente distribuidos con densidad $g_{\alpha\beta}(\lambda)$

Entonces de (2.1) se obtiene que

$$\eta_j = \eta_1 \frac{\Gamma(j+\alpha)}{j!\Gamma(1+\alpha)} \gamma^{j-1}, \tag{2.5}$$

para $\gamma^j = \beta/(1+\beta)$. Esta expresión es proporcional a la distribución binomial negativa con parámetros α y γ, escrita de tal forma para el problema de las especies, en cuyo caso $j = 0$ no se considera.

Es entonces que podemos escribir a la expresión (2.2) de la forma

$$\Delta(t) = \eta_1 \frac{\int_0^\infty e^{-\lambda}(1 - e^{-\lambda t})dG(\lambda)}{\int_0^\infty \lambda e^{-\lambda} dG(\lambda)}. \tag{2.6}$$

Si en la expresión anterior se sustituye $dG(\lambda)$ por la densidad definida en (2.4), obtendremos $\Delta_{\alpha\gamma}(t) = -\eta_1\{(1+\gamma t)^{-\alpha} - 1\}/(\gamma\alpha)$, a menos que $\alpha = 0$, en cuyo caso se tiene que $\Delta_{0\gamma}(t) = (\eta_1/\gamma)log(1+\gamma t)$.

Si $\alpha > 0$, $\Delta_{\alpha\gamma}(t)$ se acerca a su valor límite η_1/α mientras t se aproxima a infinito. El caso $\alpha \leq 0$ se tiene que $\Delta_{\alpha\gamma}(t)$ crece sin alguna cota mientras t se incrementa. Las oscilaciones infinitas de $g_{\alpha\beta}(\lambda)$ cerca de $\lambda = 0$ producen un número no acotado de nuevas especies mientras se examinan periodos cada vez más largos.

La idea de Fisher, de aplicar una distribución binomial negativa, da un mejor acercamiento al estimador $\Delta(t)$, sobre todo por el hecho de que $G(t)$ no está indeterminada.

Nos quedamos con estos modelos, remarcando que Efron y Thisted se extienden aplicando la transformación de Euler para mejorar a $\Delta(t)$ y evitar las oscilaciones que se obtienen en algunos puntos.

2.3. Solución vía procesos de Poisson

Dando pie ahora al problema de descubrir una nueva clase o una nueva especie, Mao (2004) [23] estudia un estimador basado en la *cobertura de muestreo*, de lo cual ya se tuvo mención previa en la sección 2.1 y es adjudicado al trabajo de Good y Toulmin.

Considérese nuestra ya conocida población compuesta por K_n clases disjuntas. Se obtuvo previamente una primer muestra o muestra básica de tamaño $n = s$, la cual es ampliada a una muestra de tamaño $m = (1+t)s$ con $t \geq 0$. La muestra básica corresponde al caso particular en que $t = 0$. Sea π_i

la proporción que ocupa la i-ésima clase en el muestreo, donde $\sum_{i=1}^{K_n} \pi_i = 1$. Finalmente, considérese $N_i(t)$ como el número de individuos en la i-ésima clase dentro de la muestra ampliada. Condicionando a la muestra de tamaño m, la probabilidad de descubrir una nueva clase es

$$U(t) = \sum_{i=1}^{K_n} \pi_i \mathbb{I}_{N_i(t)=0}.$$

Nótese que $1 - U(t)$ representa el total de proporciones de las clases ya identificadas en la muestra ampliada; en este caso a esto le llamamos *cobertura*. El problema que se presenta ahora es el de predecir o estimar $U(t)$ a partir de ideas de procesos de Poisson.

Si tomamos $N_i = N_i(0)$, y k_l es el número de clases que tienen l individuos en la muestra básica, con $l = 1, \ldots, n$, donde

$$k_l = \sum_{i=1}^{K_n} \mathbb{I}_{\{N_i(t)=l\}},$$

este caso se acerca a la idea de cobertura explicada anteriormente.

Las cantidades k_0 y K_n serán tratadas como cantidades desconocidas. Estaremos entonces interesados en predecir la probabilidad condicional $U(t)$ basada en la observación de k_x para $x \geq 1$. Un estimador propuesto por Good y Toulmin es el siguiente:

$$\hat{U}(t) = s^{-1} \sum_{x=1}^{\infty} (-t)^{x-1} x k_x,$$

que Mao (2004) [23] muestra puede ser visto como un estimador empírico bayesiano no paramétrico del valor esperado de la probabilidad de descubrir una nueva clase, en un modelo multinomial.

Mao (2004), y Mao y Lindsay (2002) [24], dieron un acercamiento sobre el problema de cobertura por medio de modelos de Poisson. En forma resumida, el número de individuos de cada clase se toma como un proceso Poisson con intensidad dada según la clase específica. En otras palabras, cada $X_i(t)$ se

asume como un proceso Poisson homogéneo con intensidad Λ_i y esperanza $(1+t)\Lambda_i$. La muestra básica y la ampliada consisten de individuos identificados en los intervalos de tiempo $(0,1]$ y $(0,1+t]$ respectivamente. Como resultado de la reinterpretación del problema tenemos que:

$$\pi_i = \Lambda_i / \sum_{j=1}^{C} \Lambda_j,$$

con $i = 1, 2, ...K_n$. Mientras que la representación de $U(t)$ queda como:

$$U(t) = \frac{\sum_{i=1}^{C} \Lambda_i \mathbb{I}_{X(i)=0}}{\sum_{i=1}^{K_n} \Lambda_i}.$$

Se dice que Λ_i surge como una variable aleatoria de una distribución F. El análisis del problema de descubrir una nueva clase es no paramétrico, en el sentido de que la distribución F puede ser cualquier distribución discreta en el intervalo $(0, \infty)$ con un número finito de puntos de soporte.

Mao (2004) menciona que la nueva estructura de $U(t)$ permite que el estudio del problema sea un tanto más sencillo. A partir de la introducción de modelos Poisson, se proponen nuevos estimadores de $U(t)$, basados en: inferencia sobre modelos multinomiales, modelos bayesianos empíricos y distribuciones empíricas, aproximación de momentos, aproximación de verosimilitud. Cada nuevo estimador tiene su complicación y su explicación es extensa, por lo que no se describirán a continuación. Éstos han sido descritos por Mao (2004) [23].

2.4. Solución vía distribuciones aleatorias

La solución al problema de muestreo de las especies vía distribuciones aleatorias se liga a la teoría de estadística bayesiana no paramétrica. Tal como se describe en Lijoi, Mena y Prünster (2007) [19]. Este apartado discute otro enfoque para encontrar una solución específica del PME sobre el descubrimiento de nuevas especies, descrito en el artículo mencionado. A diferencia de los modelos anteriores, éste trata la idea desde un punto de vista predictivo.

No sólo se trata de encontrar un estimador, el modelo involucra evaluar la probabilidad de descubrir un número desconocido de nuevas especies en una nueva muestra de una población, condicional al número de especies registradas en una muestra inicial o básica.

Consideremos una población de individuos que pueden ser agrupados en $K = N$ diferentes clases o especies. Retomando la idea de *cobertura de muestreo*, denotaremos a las proporciones desconocidas de individuos de la i-ésima especie como π_i. Ahora supongamos que se obtiene una muestra de tamaño n, y el número de especies identificadas determinado por $j = 1, \ldots, N$. Mientras que N_i representa el número de individuos de la población contenidos en la i-ésima clase. Como ya se mencionó, el interés elemental se centra en realizar una inferencia sobre el número de especies no observadas o no descubiertas, es decir

$$1 - U(n) = \sum_{i:n_i=0} \pi_i,$$

que resulta ser la proporción de especies o clases no observadas y donde $U(n)$ es la *cobertura de muestreo*.

El acercamiento a través de distribuciones aleatorias comienza al hacer a las probabilidades π_i elementos aleatorios. Enseguida suponemos que las observaciones realizadas a la población, y que denotaremos como $\{X_n\}_{n\geq 1}$, son independientes e idénticamente distribuidas dada una medida aleatoria $\tilde{P} = \sum \pi_i \delta_{X_i}$. La distribución correspondiente a $\tilde{P}$ representa a una distribución inicial utilizada en inferencia bayesiana no paramétrica. En aplicaciones, sobre todo en ciencias genómicas, si el número de clases es considerablemente grande, resulta apropiado suponer el tamaño de N infinito.

En este caso es necesaria una distribución inicial no paramétrica. Un ejemplo bien conocido de distribución inicial es el llamado proceso de Dirichlet.

2.4.1. Distribuciones iniciales tipo Gibbs

La siguiente construcción es un tratamiento no paramétrico que considera a una clase de distribuciones iniciales o *a priori* que inducen una partición

aleatoria, para observaciones tipo Gibbs. La conexión entre modelos de particiones aleatorias y estadística bayesiana no paramétrica se revisará en el siguiente capítulo.

Sea $\{X_n\}_{n\geq 1}$ una sucesión de observaciones intercambiables, cada una tomando valores en algún espacio $\mathbb{X}$. Suponemos que existe una medida aleatoria, $\tilde{P}$, cuya distribución se puede entender como una distribución inicial no paramétrica tal que

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n | \tilde{P}) = \prod_{i=1}^{n} \tilde{P}(A_i),$$

para cualquier $n \geq 1$ y cualquier subconjunto $A_1, \ldots, A_n$ de $\mathbb{X}$. Asumimos que $\tilde{P}$ es discreta con probabilidad uno y $\mathbb{E}[\tilde{P}(\cdot)] = P_0(\cdot)$, donde P_0 es no atómico, es decir $P_0(x) = 0$ para todo $x \in \mathbb{X}$. El número de distintas observaciones, representada por K_n, es un entero menor o igual a n, que identifica las K_n distintas especies registradas.

Cuando se observan $K_n = k$ diferentes especies se etiquetan como $X_1^*, \ldots, X_k^*$ y N_j representa el número de individuos que pertenecen a la j-ésima especie. Las distribuciones iniciales que serán consideradas inducen una distribución conjunta de K_n y del vector $(N_1, \ldots, N_{K_n})$ de la forma

$$\mathbb{P}[K_n = k \cap N_j = n_j, j = 1, \ldots, k] = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \qquad (2.7)$$

para alguna $\sigma \in (0, 1)$, y para algún conjunto de pesos no negativos

$$V_{n,k} : n \geq 1, 1 \leq k \leq n,$$

que satisfacen la ecuación

$$V_{n,k} = (n - k\sigma)V_{n+1,k} + V_{n+1,k+1}.$$

Esta distribución es invariante bajo permutaciones de $(n_1, \ldots, n_k)$. La probabilidad anterior genera particiones aleatorias intercambiables identificadas como tipo Gibbs, además de permitir obtener la siguiente distribución

predictiva para las observaciones de la muestra:

$$\mathbb{P}(X_{n+1} \in A | X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=i}^{k} (n_j - \sigma) \delta_{X_j^*}(A), \quad (2.8)$$

dado que $(X_1, \ldots, X_n)$ es una muestra de tamaño n con $K_n = k$ observaciones diferentes $(X_1^*, \ldots, X_k^*)$ y con respectivas frecuencias $n_1, \ldots, n_k$.

Es importante resaltar que el esquema de muestreo es tal que la probabilidad de muestrear una nueva especie depende solamente de n y de k, dado que una nueva especie va a ser observada. Además, cada partición tipo Gibbs determina de forma única una distribución inicial discreta no paramétrica. Algunas distribuciones iniciales que caen en esta clase y vale la pena mencionar son el proceso de Dirichlet, el proceso Poisson-Dirichlet y el proceso gaussiano inverso normalizado.

2.4.2. Estimando la probabilidad de descubrir una nueva especie

Consideremos una población compuesta idealmente por un número infinito de especies y una muestra aleatoria $X_1, \ldots, X_n$ de tamaño n, que se tomará como la muestra básica, donde tendremos que la muestra contiene $j \leq n$ valores distintos $X_1^*, \ldots, X_j^*$ que identificaran a cada especie distinta. La distribución del número de especies K_n presentes en la muestra, bajo la idea de que las X_i's son generadas por una distribución inicial tipo Gibbs, es

$$\mathbb{P}(K_n = k) = \frac{V_{n,k}}{\sigma} \mathcal{C}(n, k; \sigma)$$

donde $\mathcal{C}(n, k; \sigma)$ es un coeficiente factorial generalizado[5]. Ésta se puede interpretar como la distribución inicial del número de especies en la muestra a ser observada.

Como siguiente paso, se obtendrá una muestra adicional de m individuos, así obtendremos la «muestra ampliada» de tamaño $n + m$. Una vez que se

[5]Más adelante se dará una explicación sobre el coeficiente factorial generalizado

conoce el número de especies observada en la muestra básica y la frecuencia con la que cada especie ha sido observada, interesará conocer:

1. La probabilidad de observar nuevas especies en la muestra ampliada $X_{n+1}, \ldots, X_{n+m}$.

2. La probabilidad de observar una nueva especie en la extracción número $n+m+1$, sin haber observado los elementos anteriores de la muestra $X_{n+1}, \ldots, X_{n+m}$.

Denotamos como $X_j^{(1,n)} = (X_1, ..., X_n)$, una muestra básica de tamaño n que contiene j distintas especies, con $j = 1, \ldots, n$. Similarmente tendremos que $X^{(2,n)} = (X_{n+1}, ..., X_{n+m})$ es la segunda muestra de tamaño m, la cual se considera como no observada. Por último, tomamos la diferencia $K_m^{(n)} = K_m - K_n$ como el número de nuevas especies en $X^{(2,n)}$, entonces denotamos por $X_k^{(2,n)}$ a la muestra de tamaño m con $K_m^{(n)} = k$.

Por lo tanto, podemos reinterpretar el punto 1 como encontrar la probabilidad $\mathbb{P}(K_m^{(n)} = k | X_j^{(1,n)})$ que se puede reescribir como $\mathbb{P}(K_m^{(n)} = k | K_n = j)$, para cualquier $k = 0, 1, \ldots, m$ y para cualquier $j = 1, \ldots, n$, lo cual se puede entender como la distribución posterior del número de especies a ser observadas en una muestra de tamaño m. Lo anterior arroja la siguiente proposición:

Proposición 2.1 *Sea $\{X_n\}_{n\geq 1}$ una sucesión de observaciones intercambiables bajo una distribución inicial tipo Gibbs. Entonces, para toda $k \in 0, 1, \ldots, m$,*

$$\mathbb{P}(K_m^{(n)} = k | K_n = j) = \frac{V_{n+m,j+k}}{V_{n,j}} \frac{1}{\sigma^k} \mathcal{C}(m, k; \sigma, -n + j\sigma) \; j \in 1, \ldots, n. \quad (2.9)$$

De la función anterior se puede inferir que K_n resulta «suficiente» para predecir el número de distintas nuevas observaciones. Aquí $\mathcal{C}(m, k; \sigma, -n + j\sigma)$ es el coeficiente factorial generalizado no centralizado, del cual se puede encontrar una explicación en el apéndice de esta tesis.

Pasando ahora al punto 2, lo que se desea obtener es un estimador bayesiano para la probabilidad de descubrir una nueva especie en la extracción número $(n+m+1)$, dada la muestra básica $X_j^{(1,n)}$. Si suponemos que ya han sido observadas las dos muestras obtenidas, la probabilidad de descubrimiento estará dada por $\mathbb{P}(K_1^{n+m}|X_j^{(1,n)}, X_k^{(2,n)})$. Pero recordemos que el problema implica el no haber observado los elementos de la segunda muestra $X^{(2,n)}$, así que, tomando el hecho de que el número de distintas especies K^n es suficiente, la probabilidad de descubrimiento se puede reescribir como $\mathbb{P}(K_1^{n+m}|K_n = j, K_m^{(n)})$. Sin embargo, el haber omitido la observación de $X_k^{(2,n)}$ obliga a obtener un estimador para

$$D_m^{(n:j)} := \mathbb{P}(K_1^{n+m}|K_n = j, K_m^{(n)}).$$

Decimos que esta expresión representa una «probabilidad aleatoria» la cual obtiene su aleatoriedad de $K_m^{(n)}$.

El estimador representa una versión bayesiana no paramétrica del estimador de Good y Toulmin, es decir, se trata de un estimador para

$$U(n+m) = \sum_{i\geq 1} p_i \, \mathbb{I}_0(N_{i,n+m}),$$

la cobertura de muestreo para una muestra consistente en $n+m$ observaciones. Así tenemos una nueva proposición:

Proposición 2.2 *Sea una sucesión de variables aleatorias intercambiables bajo una distribución inicial tipo Gibbs. Entonces el estimador bayesiano, bajo una función de pérdida ajustada, de la probabilidad de observar una nueva especie en la extracción número* $(n+m+1)$*, condicional a la muestra básica* $X_j^{(1,n)}$ *con* j *distintas especies, está dado por*

$$\hat{D}_m^{(n,j)} = \sum_{k=0}^{m} \frac{V_{n+m+1,j+k+1}}{V_{n,j}} \frac{1}{\sigma^k} \mathcal{C}(m,k;\sigma,-n+j\sigma). \tag{2.10}$$

Los estimadores, obtenidos vía distribuciones aleatorias, tienen la ventaja de que las expresiones (2.1) y (2.2) pueden ser calculadas de forma exacta con pocos esfuerzos computacionales, obteniendo una expresión cerrada de los pesos $V_{n,k}$'s. Distribuciones iniciales tipo Gibbs que permiten tener casos favorables para los pesos $V_{n,k}$'s son los procesos de Dirichlet y Poisson-Dirichlet.

Capítulo 3

Enfoque bayesiano no paramétrico

A partir de este capítulo trabajaremos con variables aleatorias $X_1, X_2, \ldots$ en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, con Ω el espacio muestral, $\mathcal{F}$ una σ-álgebra y $\mathbb{P}$ una medida de probabilidad.

La inferencia bayesiana no paramétrica es un área relativamente joven que ha adquirido importancia en los últimos años. Esto se debe, en gran parte, a la flexibilidad que muestra en el modelado estadístico, comparado con la alternativa paramética; además del surgimiento de técnicas de simulación eficientes que le dan a los modelos no paramétricos fortaleza en problemas aplicados. Por mencionar algunas de las técnicas tenemos la simulación Monte Carlo vía Cadenas de Markov (MCMC, por sus siglas en inglés), que ha tenido un desarrollo importante en los últimos años.

Este capítulo se centra en mostrar la aportación de la estadística bayesiana no paramétrica al PME, vía los procesos Dirichlet y Poisson-Dirichlet, el primero considerado parteaguas en esta área de la estadística.

3.1. Teorema de Bruno de Finetti

A continuación se da un resumen al teorema de Bruno de Finetti, el cual es ampliamente utilizado en la teoría bayesiana y cuyo nombre lo obtiene de Bruno de Finetti, probabilista italiano. Una revisión más detallada sobre la teoría que envuelve las siguientes ideas pude encontrarse por ejemplo en Aldous (1985) [1] y en Fristedt et al. (1996) [12].

3.1.1. Concepto de intercambiabilidad

Una sucesión finita de variables aleatorias $(X_1, \ldots, X_n)$ se denomina *intercambiable* si

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}),$$

i.e., sus distribuciones finito-dimensionales son iguales en distribución, para toda permutación π del conjunto $\{1, \ldots, n\}$.

Por otro lado, si tenemos una sucesión infinita $(X_1, X_2, \ldots)$, se le denomina intercambiable si

$$(X_1, X_2, \ldots) \stackrel{\mathcal{D}}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots);$$

esto, para toda permutación π de $\{1, \ldots, n\}$, $n \geq 1$; en otras palabras, toda permutación para la cual $\#\{i : \pi(i) \neq i\} < \infty$.

Un ejemplo básico de intercambiabilidad puede observarse en ejercicios de muestreo. Supóngase el siguiente experimento: se tiene una urna que contiene n bolas marcadas como $(x_1, \ldots, x_n)$. Los resultados obtenidos de $(X_1, X_2, \ldots)$, una sucesión infinita de extracciones con reemplazo, representa una sucesión infinita intercambiable; mientras que los resultados provenientes de $(X_1, \ldots, X_n)$, sucesión consistente en n extracciones sin reemplazo, forman una sucesión finita intercambiable (o n-intercambiable).

Generalizando, en el primer caso, $\{X_i\}_{i \geq 1}$ representa una sucesión de variables aleatorias independientes idénticamente distribuidas (i.i.d) uniformes en $x_1, \ldots, x_n$; de donde vemos que cualquier sucesión i.i.d. es intercambiable.

En el segundo caso podemos escribir

$$(X_1, \ldots, X_n) = (x_{\pi^*(1)}, \ldots, x_{\pi^*(n)}),$$

donde π^* denota la permutación aleatoria uniforme en $\{1, \ldots, n\}$, es decir, $\mathbb{P}(\pi^* = \pi) = \frac{1}{n!}$ para cada π. De forma más general, sean $(Y_1, \ldots, Y_n)$ variables aleatorias arbitrarias y tomemos π^* independiente de $\{Y_i\}_{i\geq 1}$. Entonces

$$(X_1, \ldots, X_n) = (Y_{\pi^*(1)}, \ldots, Y_{\pi^*(n)})$$

define una sucesión n-intercambiable. Esto no es posible trasladar al caso infinito, dado que no podemos obtener una permutación uniforme de un conjunto infinito contable.

Finalmente, hay que resaltar el siguiente resultado: si $(X_1, X_n, \ldots)$ es una sucesión de variables aleatorias i.i.d., entonces éstas son intercambiables. El recíproco no es necesariamente cierto.

3.1.2. Mezclas de sucesiones i.i.d.

El teorema de Bruno de Finetti descrito con palabras básicamente se leería así: «una sucesión infinita intercambiable es una mezcla de variables aleatorias i.i.d.», lo cual representa el camino inverso a la idea de intercambiabilidad. Este es un primer acercamiento al teorema, pero habría que formalizar esta idea bajo una estructura matemática, sobre todo cuando la idea de 'mezcla de variables aleatorias i.i.d.' no es del todo clara y la obtención de dichas mezclas es clave para el teorema.

Sea una sucesión infinita de variables aleatorias $X^{(\infty)} = \{X_n\}_{n\geq 1}$, definida en algún espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, donde cada X_i toma valores en un espacio polaco[1] medible $\mathbb{X}$ con su respectiva σ-álgebra de Borel $\mathcal{X}$. Consideremos a $\mathcal{P}_{\mathbb{X}}$ como el espacio de todas las medidas de probabilidad en $(\mathbb{X}, \mathcal{X})$ y supongamos a P una distribución de probabilidad.

Entonces podemos describir a la sucesión $\{X_n\}_{n\geq 1}$ de la siguiente forma:

[1]Espacio topológico metrizable que es completo y separable; es decir, un espacio homeomórfico a un espacio métrico completo.

(i) Tomemos P de forma aleatoria de $\mathcal{P}_{\mathbb{X}}$;

(ii) entonces $\{X_n\}_{n\geq 1}$ será una serie i.i.d. con distribución Q.

Visto de una forma más general, considérese a $\mathbb{X}$ nuestro espacio polaco y a $\mathcal{P}_{\mathbb{X}}$; sea Q una distribución sobre $\mathcal{P}_{\mathbb{X}}$, entonces podemos reinterpretar (i) como

(i') Escogemos a P aleatoriamente de la distribución de probabilidad Q.

Lo anterior representa también una idea bayesiana, en la que se define una sucesión $\{X_i\}_{i\geq 1}$ como i.i.d. con respecto a P con una *distribución inicial* o *a priori* Q. Formalizando esta idea, podemos escribir

$$\mathbb{P}(X^{(\infty)} \in A) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i) Q(dP); \tag{3.1}$$

con $A = A_1 \times \cdots \times A_n \times \mathbb{X}^{\infty}$ y $A_i \in \mathcal{X}$ para cualquier $i = 1, \ldots, n$, y $\mathbb{X}^{\infty} = \mathbb{X} \times \mathbb{X} \times \cdots$. Esta expresión se trata de la distribución de una sucesión como una mezcla de variables aleatorias i.i.d. donde $X^{(\infty)} := \{X_i\}_{i=1}^{\infty}$, X_i con valores en $\mathbb{X}$.

En realidad, la idea anterior se trata de un caso especial de la siguiente idea más general que vale la pena mencionar: dada una familia $\{\mu_\gamma : \gamma \in \Gamma\}$ de distribuciones en un espacio S, llamaremos a una distribución ν una mezcla de las μ_γ si

$$\nu(\cdot) = \int_{\Gamma} \mu_\gamma(\cdot) \Theta(d\gamma),$$

para alguna distribución Θ en Γ.

3.1.3. Medidas aleatorias de probabilidad

Para definir cabalmente nuestra serie de variables aleatorias $\{X_n\}_{n\geq 1}$ como una mezcla i.i.d., necesitamos conocer lo que son las medidas aleatorias.

Definición Una medida de probabilidad aleatoria es simplemente una variable aleatoria $\mathcal{P}_{\mathbb{X}}$-valuada.

De forma más precisa, consideremos un campo P en $\mathcal{P}_{\mathbb{X}}$; el campo natural sería el generado por el mapeo

$$\theta \to \theta(A),$$

medible en $A \subset \mathbb{R}$. En otras palabras, una medida aleatoria es una distribución de probabilidad definida en el espacio de funciones de distribución.

De forma equivalente podemos decir que una medida aleatoria es una función $P(\omega, A)$, $\omega \in \Omega$, $A \subset \mathbb{X}$ tal que

(i) $P(\omega, \cdot)$ es una medida de probabilidad $\forall\ \omega \in \Omega$.

(ii) $P(\cdot, A)$ es una variable aleatoria $\forall\ A \subset \mathbb{R}$.

Considérese P_1 y P_2, con $P_1(\cdot, A) = P_2(\cdot, A)$ casi seguramente (c.s.) para cada $A \subset \mathbb{X}$, i.e. son iguales casi seguramente como variables aleatorias en $\mathcal{P}_{\mathbb{X}}$.

Antes de enunciar el Teorema de de Finetti, habría que mencionar algunas proposiciones y condiciones que se deben cumplir $\{X_n\}_{n\geq 1}$.

Definición Sea P una medida aleatoria y sea $X^\infty = \{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias. Decimos que X^∞ es una mezcla de variables aleatorias i.i.d.'s *moduladas* por P si

$$\mathbb{P}(X_i \in A_i, 1 \leq i \leq n | P) = \prod_{i\geq 1} P(\omega, A_i);$$

para todo $A_1, \ldots, A_n$ y $n \geq 1$. Ésto básicamente dice que la distribución de X es de la forma (3.1), donde Q es la distribución de P.

Lo anterior se puede reinterpretar dadas las condiciones siguientes:

Lema 3.1 *Sea $\mathcal{G} = \sigma(P)$, una σ-álgebra. Entonces X^∞ es una mezcla de variables aleatorias i.i.d.'s moduladas por P si y sólo si*

(i) Las variables aleatorias $\{X_n\}_{n\geq 1}$ son condicionalmente independientes dada $\mathcal{G}$, esto es $\mathbb{P}(X_i \in A_i, 1 \leq i \leq n|\mathcal{G}) = \prod_{i\geq 1} \mathbb{P}(X_i \in A_i|\mathcal{G})$.

(ii) La distribución condicional de X_i dada $\mathcal{G}$ es P; es decir, $\mathbb{P}(X_i \in A_i) = P(\omega, A_i)$.

Proposición 3.2 *Sea X^∞ sea una sucesión de variables aleatorias infinita intercambiable tomando valores en el espacio $(\mathbb{X}, \mathcal{X})$. Entonces*

$$P = \lim_{n\to\infty} P_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \text{ c.s.}$$

La distribución de P la conoceremos como la medida de de Finetti de X, donde P es la medida empírica límite de X.

En este caso, notemos que P_n converge débilmente a P c.s., es decir, $\mathbb{P}[P_n \to P] = 1$, con P la medida empírica de X.

Las condiciones anteriores aclaran un poco el panorama sobre la idea final que será el teorema que se enuncia a continuación.

3.1.4. Teorema de representación

Teorema 3.3 ***Teorema de representación 1.***
Sea $\mathcal{P}_\mathbb{X}$ el espacio de todas las medidas de probabilidad en el espacio $\mathbb{X}$. La distribución de una sucesión infinita de variables aleatorias intercambiables X^∞, está determinada de forma única por su medida de de Finetti dado lo siguiente:

(i) La sucesión X^∞ es condicionalmente i.i.d. dada la distribución de la medida empírica límite P, y

(ii) la distribución condicional de cada término de la sucesión es la distribución de P.

La familia de las medidas de de Finetti consisten en todas las medidas de probabilidad en $\mathcal{P}_{\mathbb{X}}$.

El teorema de Bruno de Finetti, otro nombre para el teorema de representación, tiene otra reinterpretación que permite entenderlo desde este otro enfoque:

Teorema 3.4 ***Teorema de representación 2.***
La sucesión infinita de variables aleatorias $X^{(\infty)}$ es intercambiable si y sólo si existe una medida de probabilidad Q sobre el espacio $\mathcal{P}_{\mathbb{X}}$ de todas las medidas de probabilidad en $\mathbb{X}$ tal que , para cualquier $n \geq 1$ y $A = A_1 \times \ldots \times A_n \times \mathbb{X}^{\infty}$, entonces

$$\mathbb{P}[X^{(\infty)} \in A] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i) Q(dP),$$

donde $A_i \in \mathcal{X}$ para todo $i \in \mathbb{N}$.

Entonces, si $X^{(\infty)}$ es intercambiable, Q es única y $P_n \overset{c.s}{\to} P$. En otras palabras, dada $P \sim Q$ las variables aleatorias X_i son i.i.d. y cada X_i pueden pensarse como «réplicas o repeticiones del mismo fenómeno».

Ésto nos remite directamente a la expresión (3.1) que ya nos había dado una idea clara del teorema. Es importante ver que la probabilidad Q es la *medida de de Finetti* de la sucesión de variables aleatorias $X^{(\infty)} = \{X_n\}_{n \geq 1}$, y como ya se mencionó, puede interpretarse como la distribución inicial en inferencia bayesiana. Cuando el soporte de Q es infinito dimensional entonces nos encontramos en un problema inferencial no paramétrico.

Existen varias familias de distribuciones iniciales Q, algunas de ellas bastante conocidas en aplicaciones de la estadística bayesiana no paramétrica. En específico, existen clases que generalizan el proceso Dirichlet y Poisson-Dirichlet de los cuales se hablará más adelante.

Una forma de reinterpretar el supuesto de intercambiabilidad sobre una sucesión de variables aleatorias $\{X_n\}_{n\geq 1}$ en términos de independencia condicional, se puede ver por medio del siguiente modelo:

$$X_i|\tilde{P} \overset{i.i.d.}{\sim} \tilde{P}$$

$$\tilde{P} \sim Q \tag{3.2}$$

En donde $\tilde{P}^n = \prod_{i=1}^{n} \tilde{P}$ representa la distribución condicional de $\{X_n\}_{n\geq 1}$ dado $\tilde{P}$. Entonces $\tilde{P}$ es alguna medida de probabilidad aleatoria en $(\Omega, \mathcal{F}, \mathbb{P})$, tomando valores en $\mathcal{P}_{\mathbb{X}}$.

Existen varios métodos para construir las medidas de probabilidad Q, entre los que nombramos: transformaciones de procesos estocásticos conocidos, representación directa infinito dimensional; un método importante y en el que detallaremos más adelante, es vía distribuciones predictivas obtenidas a partir de particiones intercambiables.

3.2. Proceso Poisson-Dirichlet

En el teorema de Bruno de Finetti, se sientan bases para poder entender el concepto de medidas aleatorias, que representan el preámbulo para adentrarnos en dos procesos importantes que se utilizan para la solución del problema de muestreo de especies (PME), el proceso Dirichlet y el Poisson-Dirichlet. A continuación se da un introducción a los procesos, revisando ideas presentadas por Ishwaran y James (2001) [16]; Navarrete, Quintana y Müller (2008) [22], y, Lijoi y Prünster (2009) [21].

El hecho de que gran mayoría de la literatura que refiere al proceso Poisson-Dirichlet haya aparecido fuera del ámbito estadístico, resultó en que ciertas propiedades que el proceso poseía y que lo hacían potencialmente útil como una distribución inicial dentro de la estadística bayesiana no paramétrica, fueran desapercibidas.

Junto con el proceso Dirichlet, al proceso Poisson-Dirichlet se le reconoce como un *modelo de muestreo de especies* (MMS) (*species sampling models* o SSMs).

Distribución Dirichlet

La distribución Dirichlet se trata de la versión multivariada de la distribución Beta. Consideremos la distribución de orden $K \geq 2$, sean $X = (X_1, \ldots, X_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$, con $\alpha_1, \ldots, \alpha_K > 0$, entonces su función de densidad se ve como

$$f(x_1, \ldots, x_{K-1}; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \text{ para toda } x_1, \ldots, x_{K-1} > 0,$$

que satisface $x_1 + \ldots + x_{K-1} \leq 1$, donde x_K es una abreviación de $1 - x_1 - \ldots - x_{K-1}$. La densidad es cero fuera del símplex [2] $(K-1)$-dimensional.

Aquí, $B(\alpha)$ es la función beta multinomial, que puede ser expresada en términos de la función gamma:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}, \quad \alpha = (\alpha_1, \ldots, \alpha_K).$$

3.2.1. El proceso Dirichlet

La existencia de un proceso con las distribuciones finito dimensionales del tipo Dirichlet fue establecida por Ferguson (1973) [10]. Se le conoce simplemente como proceso Dirichlet, y su importancia radica en su uso dentro de la estadística bayesiana no paramétrica que posteriormente llevaría a la obtención del proceso Poisson-Dirichlet. A continuación se da una pequeña descripción de la construcción del proceso.

[2] Un símplex o n-símplex es el análogo en n dimensiones de un triángulo. Más exactamente, un símplex es la envoltura convexa de un conjunto de $(n+1)$ puntos independientes afines en un espacio euclídeo de dimensión n o mayor.

Sea un espacio medible $(\mathbb{X}, \mathcal{X})$, y sea α una medida finita no negativa en $\mathcal{X}$. Entonces un proceso estocástico P indexado por elementos C de $\mathcal{X}$, se dice que es un **proceso Dirichlet** en $(\mathbb{X}, \mathcal{X})$ con parámetro α si para toda partición $(C_1, \ldots, C_k)$ de $\mathbb{X}$, el vector aleatorio $(P(C_1), \ldots, P(C_k))$ tiene una distribución Dirichlet con parámetros $(\alpha(C_1), \ldots, \alpha(C_k))$. P se puede considerar una medida de probabilidad aleatoria en el espacio medible.

Una propiedad atractiva, con la que nos podemos encontrar, es que si P es un proceso Dirichlet con parámetro α en $(\mathbb{X}, \mathcal{X})$, y si $X_1, \ldots, X_n$ es una muestra obtenida a partir de P, entonces la distribución posterior de P dada la muestra es también un proceso Dirichlet con parámetro $\alpha + \sum_{i=1}^{n} \delta_{X_i}$, donde δ_x es la medida de Dirac sobre x.

3.2.2. Distribuciones iniciales *stick-breaking*

Existen diferentes maneras de poder entender al proceso Poisson-Dirichlet, como por ejemplo a partir de distribuciones iniciales del tipo Gibbs, que se mencionaron en el apartado 2.4.1. Por otro lado, el proceso también se puede explicar por medio del concepto de distribuciones iniciales conocidas como *stick-breaking*. Se tratan de formas o clases especiales de medidas aleatorias de probabilidad con la característica de ser ricas y flexibles, las cuales pueden ser construidas a partir de sucesiones de variables aleatorias independientes con distribución beta. Como ejemplos de estas medidas aleatorias está el proceso Dirichlet, su variante de dos parámetros, el proceso Poisson-Dirichlet de dos parámetros y el proceso Beta de dos parámetros.

Las distribuciones iniciales *stick-breaking* son medidas aleatorias casi seguramente discretas P que pueden ser representadas como

$$P(\cdot) = \sum_{k=1}^{\infty} p_k \delta_{Z_k}(\cdot),$$

donde $\delta_{Z_k}(\cdot)$ denota la medida discreta concentrada en Z_k, o medida de Dirac. Además p_k son variables aleatorias (conocidas también como pesos aleatorios) de tal forma que son independientes de Z_k y tales que $0 \leq p_k \leq 1$ y $\sum_{k=1}^{\infty} p_k = 1$ casi seguramente. Se asume que los elementos Z_k son variables aleatorias

con distribución H sobre un espacio polaco medible $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, donde también se asume que H es no atómica (i.e. $H(y) = 0$ para toda $y \in \mathcal{Y}$).

Lo que diferencia a las distribuciones iniciales *stick-breaking* de medidas aleatorias generales de la forma (3.2), es el método de construcción de los pesos aleatorios. Consideremos a P una medida aleatoria denotada como una función $P_N(\mathbf{a}, \mathbf{b})$ de la forma (1) y

(1) $p_1 = V_1$

(2) $p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1})V_k$, $k \geq 2$,

donde V_k son variables aleatorias independientes $Beta(a_k, b_k)$ para parámetros $a_k, b_k > 0$, y $\mathbf{a} = (a_1, a_2, \ldots)$, $\mathbf{b} = (b_1, b_2, \ldots)$. La construcción de los pesos definidos en la lista anterior puede pensarse como un proceso *stick-breaking* o del 'rompimiento de una vara'; informalmente ésto se puede explicar como el proceso en que a cada paso, aleatoria e independientemente, rompemos una vara de longitud total uno, la longitud que resulta del pedazo roto la agregamos al valor p_k en el rompimiento número k.

La ventaja de las distribuciones iniciales *stick-breaking* es la utilidad que ofrecen a la estadística bayesiana no paramétrica. La construcción *stick-breaking* del proceso Poisson-Dirichlet resulta ser la forma más sencilla e intuitiva para definirlos.

3.2.3. El proceso Poisson-Dirichlet vía distribuciones iniciales *stick-breaking*

El proceso Poisson-Dirichlet de dos parámetros, es una familia de medidas aleatorias que fue introducida por Pitman (1995) [25], que a su vez se basa en el proceso PD de un parámetro propuesto en el trabajo de Kingman (1978) [17]. Se trata en una clase de modelos bastante populares que ha encontrado ciertas aplicaciones en áreas como teoría de excursión, combinatoria, modelos bayesianos mezcla y genética de población, particularmente en fragmentaciones y coalescencia.

A continuación se expresa la construcción del proceso Poisson-Dirichlet por medio de la idea de distribuciones iniciales *stick-breaking*.

Considérese dos parámetros (σ, θ) tales que $\sigma \in (0,1)$ y $\theta > -\sigma$, sea $\{V_k\}_{k\geq 1}$ una sucesión de variables aleatorias independientes, con

$$V_k \sim Beta(\theta + k\sigma, 1-\sigma).$$

En este caso definimos los pesos *stick-breaking* como

(1) $\tilde{p}_1 = V_1$,

(2) $\tilde{p}_j = V_j \prod_{i=1}^{j-1}(1-V_i)$

Por otra parte, supóngase que $\{Y_n\}_{n\geq 1}$ es una sucesión de variables independientes idénticamente distribuidas, que además son independientes de los pesos $\tilde{p}_i$ y cuya distribución denotada por P_0 es no atómica. Definiendo a δ_a como la función de masa (o medida discreta) concentrada en el punto a, la medida aleatoria discreta definida como

$$\tilde{P}_{\sigma,\theta} = \sum_{j\geq 1} \tilde{p}_j \delta_{Y_j}$$

es un proceso Poisson-Dirichlet con parámetros (σ, θ). Para conveniencia en la escritura, en adelante se referirá al proceso como $PD(\sigma, \theta)$.

3.3. Funciones de probabilidad sobre particiones intercambiables

Una vez que se ha introducido el proceso Poisson-Dirichlet, estamos a medio camino de comprender cómo utilizar este proceso como herramienta para poder dar una solución al PME. Sin embargo el proceso por sí sólo no da la solución, tenemos que hacer uso de cierta teoría que nos llevan a funciones de probabilidad predictivas que tienen una fuerte relación con funciones de probabilidad sobre particiones intercambiables.

Como se hizo referencia implícita en la sección 2.4 del capítulo 2, la solución al PME vía distribuciones aleatorias involucra el uso de funciones predictivas y particiones intercambiables.

Lo siguiente es sólo una idea básica, que necesariamente tendrá que ser ampliada:

Tomemos los siguientes elementos aleatorios K_n y $(N_1, \ldots, N_{K_n})$, los cuales son definidos en la sección 2.4.1 como parte de la solución al PME vía distribuciones aleatorias. Cuando su función de distribución conjunta (2.7) es calculada en algún punto $(k, n_1, \ldots, n_k)$ y ésta es invariante con respecto a cualquier permutación de los valores $(n_1, \ldots, n_k)$ sobre los enteros, entonces dicha distribución es conocida como una función de probabilidad sobre una partición intercambiable.

En general, cuando se tiene una función de distribución conjunta evaluada en algún punto y sucede lo antes descrito, entonces podemos hablar de funciones de probabilidad sobre particiones intercambiables. Una función de este estilo identifica la «ley» intrínseca en una partición aleatoria intercambiable $\tilde{\Pi}$ en el conjunto de los enteros positivos.

A continuación se presenta una concepción más amplia, comenzando por explicar qué son las particiones intercambiables y cómo se construyen; terminando con las funciones de nuestro interés.

3.3.1. Particiones intercambiables

Para un entero positivo n, una partición de n es una colección no ordenada de enteros positivos con suma n. Existen dos formas comunes de determinar una partición de n:

(1) Por una sucesión decreciente de términos,

$$n_{(1)} \geq n_{(2)} \geq \cdots \geq n_{(k)}$$

(2) Por el número de términos de varios tamaños,

$$m_j = \#i : n_{(i)},\ j = 1, \ldots, n,$$

donde $\sum m_j = k$, y $\sum j m_j = n$. Una partición aleatoria de n es una variable aleatoria que denotaremos por Π_n, con valores en el conjunto de todas las particiones de n.

Sea $\{1, \ldots, n\}$, un conjunto finito. Una partición de $\{1, \ldots, n\}$ es una colección no ordenada de subconjuntos disjuntos no vacíos de N_n, que llamaremos $\{C_i\}_{i\geq 1}$, con $\cup_i C_i = \{1, \ldots, n\}$, donde los conjuntos C_i se conocerán como clases de la partición. Ahora, dada una partición $\{C_i\}_{i\geq 1}$ de $\{1, \ldots, n\}$, ésta la consideraremos como una posible realización de Π_n.

Decimos que Π_n es intercambiable si la distribución de Π_n es invariante bajo todas las permutaciones de N_n.

Un resultado interesante que relacionan las particiones aleatorias con la estadística bayesiana no paramétrica es el siguiente: las particiones aleatorias determinan distribuciones iniciales discretas no paramétricas. Para terminar de comprender esto hay que explicar qué son las funciones de probabilidad sobre particiones intercambiables.

3.3.2. Funciones de probabilidad sobre particiones intercambiables

La realización de medidas de distribución discretas, en general, llevan a analizar las estructuras de partición sobre las observaciones que generan.

Para hablar de este tipo de funciones, hay que tomar en cuenta un espacio con respecto a las particiones intercambiables descritas en el apartado anterior, considerando la partición aleatoria Π_n. Tomemos una sucesión consistente en n observaciones $X_1, \ldots, X_n$ que asumen intercambiabilidad y pueden ser representadas por el modelo 3.2, entonces tenemos que la medida $\tilde{P}$, proveniente de dicho modelo, implica que deben existir empates en los datos, i.e. $\mathbb{P}[X_i = X_j] > 0$ para $i \neq j$.

En otras palabras, nos encontramos en un subespacio $\mathcal{P}_{\mathbb{X}}^d \subset \mathcal{P}_{\mathbb{X}}$, compuesto sólo de medidas atómicas. Entonces Π_n es tal que i y j pertenecen al mismo conjunto en la partición si y sólo si $X_i = X_j$.

Sea cualquier número $k \in \{1, \ldots, n\}$ y supongamos una partición $\{C_1, \ldots, C_k\}$ es una partición de $\{1, \ldots, n\}$ en k conjuntos C_i. Siendo que una partición aleatoria Π_n se puede considerar una variable aleatoria, una forma de obtener su función de distribución consiste en tomar las frecuencias de cada conjunto de la partición.

Es decir, para $n_i = \#C_i$, entonces $(n_i, \ldots, n_k) \in \{(n_1, \ldots, n_k) : n_i \geq 0, \sum_{i=1}^{k} n_i = n\}$ y entonces

$$\mathbb{P}[\Pi_n = \{C_1, \ldots, C_n\}] = p(|C_1|, \ldots, |C_j|), \tag{3.3}$$

para alguna función simétrica p de composiciones $(n_1, \ldots, n_k)$ de $\{1, \ldots, n\}$. Esta función se conoce como función de probabilidad sobre particiones intercambiables.

Definición Sea $\{X_n\}_{n\geq 1}$ una sucesión intercambiable. Entonces,

$$\{\Pi_k^{(n)} : 1 \leq k \leq n, n \geq 1\},$$

donde $\Pi_k^{(n)}(n_1, \ldots, n_k)$, proviene de la función p definida en (3.3), la conocemos como la función de probabilidad sobre particiones intercambiables (FPPI).

Notemos que esta función se satisface la regla de adición

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \Pi_k^{(n+1)}(n_1, \ldots, n_{j+1}, \ldots, n_k).$$

Se puede decir que cualquier función simétrica no negativa que satisface la regla anterior, es la FPPI sobre alguna sucesión intercambiable.

Aplicando a un contexto biológico, en este caso al problema de muestreo de especies, sea la sucesión $(X_1, \ldots, X_n)$ una muestra de una población biológica. No nos interesa mucho la realización de las $X_i's$, más bien interesa la probabilidad de observar $K_n = k$ distintas especies que reduce la muestra a $(X_1^*, \ldots, X_n^*)$ valores únicos con frecuencias $(N_1, \ldots, N_{K_n})$, donde $\sum_{i=1}^{k} N_i = n$.

Además podemos reinterpretar

$$N_j := \sum_{m=1}^{n} \mathbb{I}(X_m = X_j^*),$$

el número de veces que el j-ésimo valor distinto X_j^* aparece entre la sucesión $X_1, \ldots, X_n$. Entonces N_j también se puede ver como el número de elementos en la j-ésima clase de Π_n.

Distribuciones predictivas

La existencia de una FPPI conlleva a un sistema de distribuciones predictivas inducidas por Q (función *a priori* definida por el modelo 3.2). Además suponga $\Pi_k^{(n)}$ como la FPPI asociada; si $(X_1, \ldots, X_n)$ contiene k distintas valores $(X_1^*, \ldots, X_n^*)$ y n_j de esos valores son iguales a X_j^* entonces se tiene que

$$\mathbb{P}[X_{n+1} = \text{nueva especie} | X_1, \ldots, X_n] = \frac{\Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}$$

y

$$\mathbb{P}[X_{n+1} = X_j^* | X_1, \ldots, X_n] = \frac{\Pi_k^{(n+1)}(n_1, \ldots, n_{j+1}, \ldots, n_k)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}.$$

Fórmula de muestreo de Ewens

Probablemente el ejemplo más conocido de función de probabilidad sobre particiones intercambiables es la fórmula de muestreo de Ewens, desarrollada

por Warren J. Ewens (1972) [8], y la cual se considera una piedra angular en la teoría de la genética de poblaciones. A continuación se resume la idea de la fórmula y sus alcances.

Ewens especifica que, bajo ciertas condiciones, si una muestra aleatoria de n gametos es tomada de una población y clasificada de acuerdo al gen bajo un locus[3] en particular, entonces, si definimos el evento de que existan a_1 alelos representados una vez en la muestra, y que existan a_2 alelos representados dos veces, y continuando con la idea, la probabilidad del evento es

$$\mathbb{P}(a_1, \ldots, a_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{\alpha_j}}{j^{\alpha_j} a_j!},$$

para algún número $\theta > 0$, cuando $a_1, \ldots, a_n$ es una sucesión de enteros no negativos tales que

$$a_1 + 2a_2 + 3a_3 + \cdots + na_n = n.$$

Cabe mencionar que la probabilidad conjunta anterior corresponde a la probabilidad conjunta que se obtiene en la construcción del proceso de Dirichlet.

Las condiciones a las que se refiere el enunciado inicial son las siguientes:

(1) El tamaño de la muestra n es «pequeño» en comparación con el tamaño de la población total,

(2) la población está en equilibrio estadístico bajo mutación y deriva genética, además de que la injerencia del locus para la clasificación es insignificante,

(3) cada alelo mutante es «nuevo».

Sabemos que es una función de probabilidad sobre particiones intercambiables, dado que se trata de una distribución definida en el conjunto de todas

[3]Localización particular de un gen o secuencia de ADN en un cromosoma

las particiones de los primeros n enteros, invariante ante cualquier permutación de las particiones. La fórmula también es conocida como la distribución de Ewens.

La fórmula de Ewens surge naturalmente del 'proceso del restaurante chino', el cual resulta ser un método para obtención de funciones de probabilidad sobre particiones intercambiables.

3.4. Solución al PME

Las ideas propuestas a continuación provienen en gran mayoría del artículo de Favaro, Lijoi, Mena y Prünster (2009) [9]. Retomemos el problema de muestreo de las especies (PME) descrito en la sección 1.1 en el capítulo 1 y el problema específico tratado en la sección 2.4.

Sea $\{X_n\}_{n\geq 1}$ una sucesión de observaciones intercambiables, que toma valores en un espacio $\mathbb{X}$. $K_n \in \{1, \ldots, n\}$ identifica las distintas especies registradas en esta muestra de observaciones y el vector $(N_1, \ldots, N_{K_n})$, donde cada entrada N_j representa el número de individuos en la n-ésima observación, es tal que$\sum_{i=1}^{K_n} N_i = n$.

Al ser la sucesión $\{X_n\}_{n\geq 1}$ intercambiable, por el teorema de Bruno de Finetti $\{X_n\}_{n\geq 1}$ puede ser caracterizada por un modelo jerárquico, tomando a las X_n como elementos de una muestra aleatoria de alguna distribución $\tilde{P}$ y una distribución inicial Q en $\tilde{P}$. Esto se puede ver como 3.2:

$$X_i|\tilde{P} \overset{i.i.d.}{\sim} \tilde{P}$$

$$\tilde{P} \sim Q$$

Ya hemos visto dos formas de construir distribuciones iniciales. Por un lado, en la sección 2.4.1 se hace referencia a las distribuciones iniciales de Gibbs, y por otro, por medio de distribuciones iniciales *stick-breaking*. La

construcción de Gibbs permitió obtener la expresión general (2.7), donde los pesos no negativos $V_{n,k} : n \geq 1, 1 \leq k \leq n$ son no definidos.

Bajo el modelo anterior y tomando a $\tilde{P}$ como un proceso Poisson-Dirichlet $PD(\sigma, \theta)$ como una distribución inicial del tipo Gibbs, los pesos toman el siguiente valor

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}}.$$

Entonces (2.7) se reescribe como

$$\Pi^{(n)}_{(n_1,\ldots,n_k)} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k}(1 - \sigma)_{n_j - 1}, \tag{3.4}$$

la cual resulta ser la FPPI que caracteriza la partición aleatoria inducida por el $PD(\alpha, \theta)$. Ésta también es la forma de ver a la función de distribución de un proceso $PD(\sigma, \theta)$, bajo los supuestos anteriores, conocida como la fórmula de muestreo de Pitman. En la expresión anterior se tiene que, para todo entero no negativo N, decimos que $(a)_N = \frac{\Gamma(a+N)}{\Gamma(a)} = a(a + 1) \cdots (a + N - 1)$ es el n-ésimo factorial ascendente de a, con $(a)_0 \equiv 1$. Para $k = 0$ se conviene que $\prod_{i=1}^{0}(\theta + i\sigma) = 1$.

Analizando la expresión se puede llegar a la conclusión de que se compone por el producto de dos factores, el primero dependiente sólo en (n, k) el número de la muestra y el número de distintas clases o especies, mientras el segundo factor depende en las frecuencias $(N_1, \ldots, N_n)$ vía el producto $\prod_{i=1}^{k}(1 - \sigma)_{n_j - 1}$.

Nótese que si $\alpha \to 0$ entonces la FPPI se reduce a

$$\Pi^{(n)}_{(n_1,\ldots,n_k)} = \frac{\theta^k}{(\theta)_n} \prod_{j=1}^{k}\Gamma(n_j),$$

que coincide con la FPPI de un proceso Dirichlet.

Por otro lado, en la sección 2.4.1 también se introdujo la distribución predictiva (2.8) inducida por particiones aleatorias del tipo Gibbs. Entonces, bajo la perspectiva del PME, sea $(X_1, \ldots, X_n)$ una muestra poblacional

consistente en $K_n = k$ distintas especies $X_1^*, \ldots, X_k^*$ y n_j de ellas iguales a X_j^*, y considerando el valor dado a los pesos $V_{n,k}$, entonces la distribución predictiva asociada al proceso $PD(\sigma, \theta)$ se ve como

$$\mathbb{P}(X_{n+1} \in A | X_1, \ldots, X_n) = \frac{\theta + k\sigma}{\theta + n} P_0(A) + \frac{1}{\theta + n} \sum_{j=i}^{k} (n_j - \sigma) \delta_{X_j^*}(A).$$

Para el proceso $PD(\sigma, \theta)$ la probabilidad de observar un nuevo valor depende del número de distintas observaciones, en contraste al proceso Dirichlet. Otra diferencia se representa por el comportamiento asintótico del número de clases o especies K_n generado por las primeras n observaciones; entonces, si $n \to \infty$, tenemos que $K_n \sim S_{\sigma,\theta} n^{\sigma}$, donde $S_{\sigma,\theta}$ es una variable aleatoria positiva con densidad en $\mathbb{R}^+$ dependiente de σ y θ.

En comparación, si consideramos que las observaciones se rigen bajo un proceso Dirichlet, y si $n \to \infty$ entonces $K_n \sim \theta \log(n)$ c.s.; es decir, el número de distintas observaciones se incrementa con una tasa logarítmica.

Es fácil ver que bajo un proceso PD, en comparación al proceso Dirichlet, el número de distintas observaciones se incrementa con una tasa mayor, n^{σ}.

Ahora, sin perder de vista al modelo (3.2) y siendo que $\tilde{P}$ es un proceso Poisson Dirichlet (σ, θ), la cobertura de muestreo estará dada por

$$U(n) := \hat{C}_1^{(n,j)} = 1 - \frac{\theta + j\sigma}{\theta + n}. \tag{3.5}$$

Y la distribución del número de observaciones distintas, dentro de un muestreo de tamaño n, es

$$\mathbb{P}(K_n = k) = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{\sigma^k (\theta + 1)_{n-1}} \mathcal{C}(n, k; \sigma).$$

Así, la distribución de $K_m^{(n)} := K_m - K_n$, el número de especies distintas que serán observadas en una muestra adicional de tamaño m, condicional

a una muestra básica de tamaño n que ya nos presentó $K_n = k$ especies distintas, será:

$$\begin{aligned}\mathbb{P}_m^{(n,j)}(k) :=& \\ &= \mathbb{P}(K_m^{(n)} = k | K_n = j) \\ &= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1}(\theta+i\sigma)}{\sigma^k} \mathcal{C}(m,k;\sigma,-n+j\sigma),\end{aligned} \tag{3.6}$$

para $k = 0, \ldots, m$, donde $\mathcal{C}(m,k;\sigma,-n+j\sigma)$ es el coeficiente factorial generalizado no centralizado. Esta expresión es la clave para evaluar estimadores bayesianos útiles para la inferencia dentro del PME, la cual proviene directamente de la proposición 2.1 que aparece en la sección 2.4.2, en donde se presenta una expresión general para la distribución de $K_m^{(n)}$. Existe una demostración para el caso general indicado en (2.1) que podemos encontrar en Lijoi et al. (2007) [19]; pero para el caso del proceso $PD(\sigma,\theta)$ existe la siguiente

Demostración La demostración se basa en la idea de la representación de la distribución posterior o *a posteriori* de $\tilde{P}_{\sigma,\theta}$, dada una muestra $X_1, \ldots, X_n$ con distribución $\tilde{P}_{\sigma,\theta}$. Si las observaciones X_i son i.i.d. condicionales dada $\tilde{P}_{\sigma,\theta}$ y además sabemos que la muestra $X_1, \ldots, X_n$ contiene $j \leq n$ valores distintos $X_1^*, \ldots, X_j^*$, entonces

$$\tilde{P}_{\sigma,\theta}|(X_1,\ldots,X_n) \stackrel{d}{=} \sum_{i=1}^{j} w_i \delta_{X_i^*} + w_{j+1}\tilde{P}_{\sigma,\theta+j\sigma}, \tag{3.7}$$

donde los valores $(w_1 \ldots, w_j)$ surgen de una distribución Dirichlet con j variables y parámetros $(n_1 - \sigma, \ldots, n_j - \sigma, \theta + j\sigma)$, $n_i = \#r : X_r = X_j^*$ es la frecuencia de X_i^* en la muestra y $w_{j+1} = 1 - \sum_{i=1}^{j} w_i$.

Para poder llegar a la expresión deseada, utilizaremos la representación *a posteriori* que se definió en (3.7), y las propiedades distribucionales de K_i, para cualquier i. Entonces, de (3.7) se puede notar que, dado

$$w \sim Beta(\theta + j\sigma, n - j\sigma),$$

una observación que denotaremos como X_{n+i}, con $i = 1, \ldots, m$, no coincide con ninguna de las especies distintas $K_n = j$ observadas en la muestra básica

con probabilidad w. Así

$$\mathbb{P}(K_m^{(n)} = k|K_n = j) = \frac{\Gamma(\theta+n)}{\Gamma(\theta+j\sigma)\Gamma(n-j\sigma)}$$

$$\otimes \int_0^1 \mathbb{P}(K_m^{(n)} = k|K_n = j, w) \times\ w^{\theta+j\sigma-1}(1-w)^{n-j\sigma-1}dw.$$

Para poder obtener $K_m^{(n)} = k$, al menos k elementos de los m en la muestra $X_{n+1}, \ldots, X_{n+m}$, deben coincidir con las k nuevas especies distintas no observadas en las $K_n = j$ especies distintas de la muestra básica. Entonces

$$\mathbb{P}(K_m^{(n)} = k|K_n = j, w) = \sum_{i=k}^{m} \binom{m}{i} w^i(1-w)^{m-i}\mathbb{P}(K_i = k), \tag{3.8}$$

donde K_i es el número de especies distintas dentro de i observaciones generadas por un proceso $PD(\sigma, \theta + j\sigma)$.

En este caso se deriva que

$$\mathbb{P}(K_i = k) = \prod_{i=1}^{k-1} \frac{(\theta + j\sigma + l\sigma)}{\sigma^k(\theta+j\sigma+1)_{i=1}}\, \mathcal{C}(i,k;\sigma)\ \ i = k, \ldots, m, \tag{3.9}$$

donde el coeficiente factorial generalizado es $\mathcal{C}(i,k;\sigma) = \frac{1}{k!}\sum_{r=0}^{k}(-1)^r\binom{k}{r}(-r\sigma)_i$.

Si sustituimos (3.9) en (3.8) obtenemos

$$\begin{aligned}\mathbb{P}_m^{(n,j)}(k) = & \\ &= \frac{(\frac{\theta}{\sigma}+j)_k}{(\theta+n)_m}\sum_{i=k}^{m}\binom{m}{i}\mathcal{C}(i,k;\sigma)(n-j\sigma)_i \\ &= \frac{(\frac{\theta}{\sigma}+j)_k}{(\theta+n)_m}\mathcal{C}(i,k;\sigma,-n+j\sigma),\end{aligned}$$

donde el coeficiente factorial generalizado no central se ve como:

$$\mathcal{C}(i,k;\sigma,-n+j\sigma) = \frac{1}{k!}\sum_{r=0}^{k}(-1)^r\binom{k}{r}(n-\sigma(r+j))_m$$

. □

Una característica importante de (3.6), es que podemos obtener el siguiente estimador

$$\hat{\mathbb{E}}_m^{(n,j)} := \mathbb{E}[K_m^{(n)}|K_n = j] = \sum_{k=0}^{m} k\mathbb{P}_m^{(n,j)}(k), \tag{3.10}$$

el cual resulta ser el número esperado de nuevas especies, importante dentro del PME.

Además es posible obtener la probabilidad de descubrimiento, es decir, la probabilidad que en la observación $(n+m+1)$ se obtenga una nueva especie o clase, sin haber tenido que observar los m registros intermedios (entre las primeras n y $n+m+1$ observaciones). Esta probabilidad proviene de evaluar el $PD(\sigma, \theta)$ en la expresión general de la Proposición 2.2.

$$\hat{D}_m^{(n,j)} = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{k=0}^{m} \frac{\prod_{i=j}^{j+k}(\theta + i\sigma)}{\sigma^k} \mathcal{C}(m, k; \sigma, -n + j\sigma) \tag{3.11}$$

Es entonces, que la cobertura de muestreo después de realizar $n+m$ observaciones puede escribirse como $\hat{C}_m^{(n,j)} = 1 - \hat{D}_m^{(n,j)}$. Las expresiones anteriores de $\hat{\mathbb{E}}_m^{(n,j)}$ y $\hat{D}_m^{(n,j)}$ tienen la ventaja de ser explícitas y pueden ser evaluadas de forma exacta. El problema surge cuando el tamaño de la muestra adicional es demasiado grande, lo que causa que el cálculo computacional de evaluar (3.6) y (3.11) resulte ser una carga pesada. Sin embargo, es posible obtener versiones simplificadas de las dos ecuaciones mencionadas.

Como primer paso hay que considerar a los números de Stirling no centrales de segundo tipo cuya expresión es

$$S_{(r,i;\gamma)} = \frac{1}{i!} \sum_{l=0}^{i} (-1)^{i-l} \binom{i}{l} (l + \gamma)^r$$

para $r = 0, 1, \ldots$; $i = 0, \ldots, r$; y donde $S_(r, i; \gamma) = 0$ para $i = r+1, r+2, \ldots$. A partir de esta expresión es que podemos ver los momentos para $K_m^{(n)}$, dado K_n y así obtener expresiones simplificadas a partir de (3.10) y (3.11).

Proposición 3.5 *Bajo el modelo del proceso Poisson-Dirichlet de dos parámetros* (σ, θ)*, se obtiene*

$$\mathbb{E}[(K_m^{(n)})^r | K_n = j] = \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(j + \frac{\theta}{\sigma}\right)_{\nu} S_{(r,\nu;j+\frac{\theta}{\sigma})} \frac{(\theta + n + \nu\sigma)_m}{(\theta + n)_m} \quad (3.12)$$

recordando que, para todo entero no negativo N*, decimos que* $(a)_N$ *es el* n*-ésimo factorial ascendente de* a*.*

Entonces, un estimador bayesiano no paramétrico de $K_m^{(n)}$ momento derivado de (3.12)

$$\mathbb{E}[K_m^{(n)} | K_n = j] = \left(j + \frac{\theta}{\sigma}\right) \left\{ \frac{(\theta + n + \sigma)_m}{(\theta + n)_m} - 1 \right\}. \quad (3.13)$$

Este estimador será el más utilizado, pues es una versión simplificada de la expresión que indica el número esperado de nuevas especies, cuyo cálculo resulta ser bastante accesible en casos aplicados, como se verá más adelante.

Igualmente, la probabilidad de descubrimiento es posible verla como

$$\hat{D}_m^{(n,j)} = \frac{\theta + j\sigma}{\theta + n} \frac{(\theta + n + \sigma)_m}{(\theta + n + 1)_m}. \quad (3.14)$$

Tomando en cuenta que una forma de entender a la cobertura de muestreo (que se presentó en esta sección como 3.5) es en función de $\hat{D}_m^{(n,j)}$, también ésta se simplifica.

Interpretación probabilista de 3.13

Nótese que el estimador del número esperado de nuevas especies, como se tiene expresado en (3.13), puede ser interpretado desde el punto de vista probabilístico como

$$\mathbb{E}[K_m^{(n)} | K_n = j] = \mathbb{P}(X_{n+1} = \text{nueva especie} | K_n = j) \mathbb{E}_{\sigma,\theta+n}[K_m],$$

donde $\mathbb{E}_{\sigma,\theta+n}[K_m]$ es la esperanza no condicional del número de distintas especies, considerando m observaciones, con respecto a la distribución de probabilidad de un proceso $PD(\sigma, \theta + n)$. La obtención de los momentos de la distribución no condicional, $\mathbb{E}[K_n^r]$, es posible a partir de (3.12) si se considera a $n = j = 0$.

Capítulo 4

Aplicación

En los capítulos previos se ha delineado el camino para modelar el PME. Es entonces que, dada una población y una muestra de cierto tamaño, de la cual hemos podido clasificar sus elementos en categorías ajenas, ahora conocemos estimadores que surgieron a partir del proceso $PD(\sigma, \theta)$ y que nos permiten obtener 1) la probabilidad de encontrar nuevos elementos de clasificación única en una nueva muestra, 2) el número esperado de nuevas especies y 3) la probabilidad de descubrimiento.

La aplicación de dichos estimadores con datos reales es necesaria para corroborar su utilidad y demostrar que las herramientas funcionan, tomando las reservas necesarias que todo trabajo estadístico demanda. En este apartado la teoría prevía obtiene justificación, pero sobre todo, se cumple uno de los objetivos principales de esta tesis.

Para dar una idea sobre el funcionamiento de estos estimadores, tomaremos datos reales obtenidos a partir de marcadores de secuencia expresada o EST que, como se resumió en el apartado 1.1.3, resultan ser una herramienta importante para la identificación y descubrimiento de genes en distintos organismos. Los EST's [1] representan porciones de genes expresados, que para fines estadísticos podremos tomar como nuestra muestra.

[1] Se utilizará el acrónimo en inglés para unificar con la literatura existente.

Entonces, dados los datos de un EST de una cierta librería de ADNc, los procedimientos de estimación predictiva descritos por el PME (y resumidos en los tres puntos del primer párrafo) pueden ser aplicados, dado que estos estimadores permitirán tomar decisiones tales como: si es factible seguir secuenciando la librería para la obtención nuevos genes. Asimismo, es posible establecer el grado de redundancia de una librería EST que determinará la eficiencia del proceso de secuenciación.

Dado el hecho de que los métodos de secuenciación del ADN son bastante caros en su costo de realización, tener herramientas que permitan determinar la eficiencia de futuras implementaciones resulta fundamental para investigaciones actuales.

Los datos e ideas principales para representar las aplicaciones fueron obtenidos a partir de Lijoi, Mena y Prünster (2007) [20]. Los supuestos y resultados numéricos que aparecerán a continuación se extrajeron de mencionado artículo. No se describe explicitamente el desarrollo de las fórmulas que llevan a esos números, siendo que un resumen es suficiente para el análisis sobre los estimadores obtenidos.

Así mismo, no se realizaron pruebas ni simulaciones propias considerando que la bibliografía existente cubre apropiadamente los fines de este apartado. Por otra parte, al final de este capítulo, se hace referencia a los programas y paquetes computacionales utilizados para realizar las simulaciones originales.

Entrando en materia y con la intención de unificar la teoría con la aplicación consideremos lo siguiente:

a) **Cobertura**: en secuencias de ADN la cobertura puede verse como la proporción de genes en una librería representada en una muestra inicial; de forma equivalente se puede tomar como la probabilidad de que una nueva lectura de la librería no produzca un nuevo gen. Permite un primer acercamiento al análisis de la redundancia de datos en la librería.

Recuperemos la expresión (3.5) del capítulo anterior, la idea de cobertura sobre la cual nos orientaremos. Para obtener un estimador de la cobertura hacemos referencia directa a la probabilidad de descubri-

miento representada en (3.11), por lo que 1 menos la probabilidad de descubrimiento será el estimador deseado.

b) **Número esperado de nuevos genes**: después de observar una muestra EST inicial de tamaño n generada de una librería de ADNc, y estimada su cobertura, es posible realizar la predicción sobre futuras lecturas o muestras. Entonces interesa conocer el número esperado de nuevos genes únicos en una nueva muestra EST de tamaño m. Ésto nos da una medida más aproximada sobre la redundancia de la librería.

Hacemos referencia directa a la expresión (3.10) y a su forma simplificada (3.13).

c) **Tasa de descubrimiento**: resulta importante conocer la tasa a la cual la probabilidad de descubrir un nuevo gen decae mientras se realizan mayor número de observaciones o lecturas de una librería. Es decir, interesa determinar la probabilidad de que la lectura $(n+m+1)$ origine un nuevo gen, dada la muestra EST inicial de tamaño n, sin importarnos el resultado de realizar la muestra intermedia de tamaño m.

De nuevo nos remitimos a la expresión (3.11) de la probabilidad de descubrimiento, considerando que la tasa se podrá determinar considerando diferentes tamaños de la muestra inicial n.

Nótese entonces que b) y c) permiten dar una idea sobre el valor a tomar de m, el tamaño de la muestra adicional.

En la sección 2.1 se describieron varios métodos que parten de un enfoque frecuentista y que han tratado el PME; entre ellos resaltan las ideas propuestas por Good y Toulmin, de donde se obtiene un estimador bastante popular en estos ámbitos, el cual es estudiado por Mao (2004) [23] y retomado en Lijoi et al. (2009) [21] desde el punto de vista bayesiano no paramétrico.

La recopilación de las ideas frecuentistas nos permite razonar sus debilidades frente a las incógnitas que se nos presentan en el PME, en particular en la aplicación sobre secuencias genéticas. Entre estas debilidades, tenemos que varios estimadores sólo funcionan bajo ciertas hipótesis del tamaño de

las muestras. Una falla importante es la dificultad de los modelos frecuentistas para incorporar genes únicos aún no observados, es decir, elementos de clasificación única aún no muestreados. Esto conlleva comportamientos erráticos de los estimadores.

Por otro lado, la idea en la inferencia bayesiana se centra en tomar la información conocida y obtener datos que permitan predicción. Como ya se mencionó, los estimadores bayesianos no paramétricos que utilizaremos para la aplicación serán los que se obtuvieron a partir del proceso Poisson-Dirichlet de dos parámetros.

Aplicando al contexto genético, se considera que el proceso $PD(\sigma, \theta)$ es la distribución inicial de las proporciones de los genes dentro de la librería. Entonces, esto implica que las etiquetas de los genes de las secuencias son *intercambiables*, y que el orden de aparición de la etiquetas no influye en el cálculo de las probabilidades.

En la teoría consideramos a las etiquetas de las secuencias como infinitas. Sin embargo, en la práctica tenemos, como cota superior para el número de genes únicos que serán observados, el tamaño total de la librería secuenciada que siempre será finita.

4.1. Datos y análisis

Los datos que se aplican consisten en muestras de EST's obtenidas de librerías de ADN de dos diferentes organismos: el *Mastigamoeba balamuthi* un protista amitocondrial, de la cual se consideran librerías normalizadas y no normalizadas; y librerías de la *Naegleria gruberi*, obtenidas de cultivos bajo condiciones aeróbicas y anaeróbicas.

Cada muestra EST consiste en n lecturas con $K_n = k$ genes únicos y sus correspondientes frecuencias $n_1, \ldots, n_k$, es decir, n_i es el número de etiquetas que identifican al i-ésimo gen dentro de la muestra inicial, donde claramente $k \in 1, \ldots, n$ y $\sum_{i=1}^{k} n_i = n$. Asimismo, las lecturas pueden ser clasificadas de

acuerdo a su etiqueta que representa el nivel de expresión, de tal forma que

$$k_l \equiv \sum_{i=1}^{K_n} \mathbb{I}_{(\{n_i=l\})}, \text{ para } l = 1, 2, \ldots, s,$$

donde s es el máximo nivel de expresión para genes únicos.

Como nota adicional, es posible aplicar el llamado protocolo de 'normalización' a los datos obtenidos a partir de EST's, lo cual tiene como objetivo hacer uniformes las frecuencias de genes dentro de la librería, y con ésto mejorar la tasa de descubrimiento. Sin embargo este tipo de procedimientos resulta tener también un costo alto en la práctica. Los datos de uno de los ejemplos que tomaremos como referencia ya han sido tratados con dicho protocolo.

Cuadro 4.1: **Muestras de EST's clasificadas según su nivel de expresión**

Librería	l	1	2	3	4	5	6	7	8	9	10	11
Naeglaria Aeróbica		346	57	19	12	9	5	4	2	4	5	4
Naeglaria Anaeróbica		491	72	30	9	13	5	3	1	2	0	1
Mastigamoeba No normalizada		378	33	21	9	6	1	3	1	1	1	0
Mastigamoeba Normalizada		200	21	14	4	3	3	1	0	1	0	0
Librería	l	12	13	14	15	16	17	18	27	55	k	n
Naeglaria Aeróbica		1	0	0	0	1	1	1	1	1	473	959
Naeglaria Anaeróbica		0	1	3	0	0	0	0	0	0	631	969
Mastigamoeba No normalizada		0	1	0	5	0	0	0	0	0	460	715
Mastigamoeba Normalizada		0	0	1	0	0	0	0	0	0	248	363

El cuadro 4.1 caracteriza los datos de las 4 muestras EST, que organiza a los genes en 20 clasificaciones según su nivel de expresión, del 1 al 18, 27 y 55.

En el caso de la *Mastigamoeba* normalizada, para el primer nivel de expresión tenemos $r_1 = 200$, lo que quiere decir que 200 genes (únicos) aparecen sólo una vez, lo cual se puede representar como $n_1 = n_2 = \ldots = n_{200} = 1$; siguiendo ese proceso, para el segundo nivel de expresión tenemos $r_2 = 21$, 21 genes (únicos) aparecen 2 veces, entonces $n_{201} = \ldots = n_{221} = 2$. También de esta muestra tenemos que el total de genes encontrados es $n = 363$ y el número de genes únicos es $k = 248$.

Ahora analicemos los resultados que surgen después de aplicar las herramientas bayesianas no paramétricas para obtener la cobertura, el número esperado de nuevos genes y la tasa de descubrimiento.

Cobertura

Los estimados para la cobertura correspondientes a las muestras de la *Mastigamoeba* no normalizada (con $n = 715$) y normalizada ($n = 363$), son 0.47 y 0.45 respectivamente. Podemos ver que no existe gran diferencia entre las dos muestras; sin embargo, la 'normalización' es lo que hace la diferencia se puede concluir que una muestra inicial de la mitad del tamaño produce prácticamente la misma cobertura. En el caso que de las muestras de la *Naegleria* aeróbica ($n = 959$) y anaeróbica ($n = 969$) se obtiene 0.64 y 0.49 de cobertura respectivamente. En este caso la cobertura del cultivo aeróbico es mejor, pero es posible que ésto indique mayor redundancia en la librería.

Número esperado de nuevos genes y la tasa de descubrimiento

Recordemos que el número esperado de nuevos genes en una muestra adicional de tamaño m nos puede indicar una medida general de redundancia, mientras que la tasa de descubrimiento predice la tendencia a la cual la probabilidad de descubrimiento decae mientras más lecturas de la muestra son consideradas.

En la figura 4.1 se da la comparación entre métodos frecuentistas y bayesianos. En este caso, para las 2 muestras de la *Naegleria gruberi* se obtiene

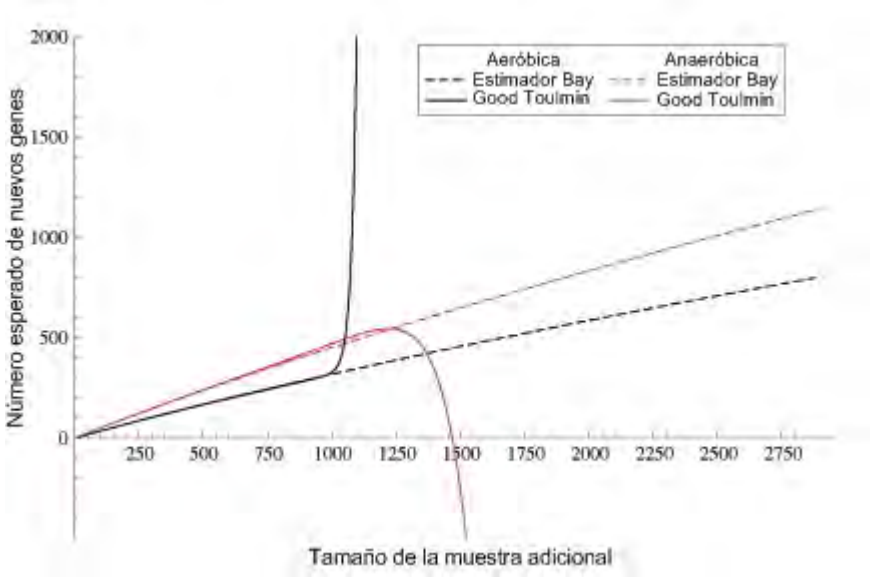

Figura 4.1: Número esperado de nuevos genes, comparación con estimador Good-Toulmin

el número esperado de nuevos genes, aplicando el estimador Good-Toulmin contra el estimador bayesiano no paramétrico basado en el proceso Poisson-Dirichlet. En la gráfica se registra el comportamiento de los estimadores para el número esperado de nuevos genes dado el número de la muestra adicional. De inmediato se puede observar que, si el tamaño de la muestra adicional m es más grande que el tamaño de la muestra inicial n, el estimador Good-Toulmin adquiere comportamientos erráticos. En general el estimador se comporta erróneamente para $n > 2m$.

En los cuadros 4.2 y 4.3 se muestran los principales resultados de las cuatro muestras; la primera se refiere a las muestras de la *Mastigamoeba* normalizada y no normalizada, mientras que la segunda agrupa los resultados de la *Naegleria* aeróbica y anaeróbica. Se toman varios casos en los que varía el tamaño de la muestra inicial.

En la primer columna aparece el porcentaje que se toma sobre el tamaño de la muestra inicial n, desde la mitad de la muestra hasta tres veces su tamaño original; la segunda despliega el tamaño de la muestra adicional

Cuadro 4.2: **Estimaciones para librería de la Mastigamoeba**

%n	m	Número esperado de nuevos genes	Probabilidad de descubrimiento
Mastigamoeba no normalizada			
50	358	180	0.481
100	715	346	0.452
150	1072	503	0.430
200	1430	654	0.412
250	1788	799	0.398
300	2145	939	0.386
Mastigamoeba normalizada			
50	182	94	0.493
100	363	180	0.456
150	544	260	0.428
200	726	336	0.406
250	908	408	0.389
300	1089	477	0.374

m que varía dependiendo del valor que indica la primer columna y de la librería; la tercera representa el número esperado de nuevos genes y la cuarta la probabilidad de descubrimiento de nuevos genes en la $n+m+1$ observación o lectura de la librería.

Resaltamos del cuadro 4.2 el hecho de que la muestra *Mastigamoeba* normalizada, en su probabilidad de descubrimiento o tasa de descubrimiento, se reduce de forma más precipitada que la no normalizada. Ésto se puede notar simplemente comparando el comportamiento del estimador, en la cuarta columna, dado el porcentaje de la muestra inicial; para el 50 % y 100 % la tasa de descubrimiento es mayor para la librería normalizada, a partir del 150 % de n la tasa es menor en comparación a la librería no normalizada. Sin embargo, tenemos que notar el tamaño de la muestra adicional dado el tamaño de la muestra inicial; para la muestra no normalizada el tamaño de m es más del doble que la normalizada para el 300 % de n, lo cual implica un cierto costo de aplicación.

La comparación de estas 2 librerías se centra en ayudar a decidir si se

Cuadro 4.3: **Estimaciones para librería de la Naegleria**

%n	m	Número esperado de nuevos genes	Probabilidad de descubrimiento
		Naegleria aeróbica	
50	480	162	0.318
100	959	307	0.290
150	1438	441	0.270
200	1918	566	0.254
250	2398	685	0.242
300	2877	798	0.231
		Naegleria anaeróbica	
50	484	231	0.450
100	969	440	0.412
150	1454	632	0.384
200	1938	812	0.362
250	2422	983	0.344
300	2907	1146	0.330

efectúa el protocolo de normalización, que ya se mencionó tiene costo propio. Entonces es notable que para esta decisión el tamaño de las muestras de las librerías en cuestión es una variable importante para tomar en cuenta.

Ahora, sobre las librerías de la *Naeglaria*, nos gustaría conocer cuál es mejor en el sentido de la obtención de nuevos genes. En primer lugar comparemos la segunda columna entre el caso aeróbico y anaeróbico. Vemos que el tamaño de la muestra adicional, para cada tamaño de muestra inicial, no varía demasiado. El número esperado de nuevos genes es mayor para la librería anaeróbica, mientras que la probabilidad de descubrimiento que se obtiene es sensiblemente mayor. Aunque en la librería aeróbica se encuentra la ventaja de tener menor decaimiento en la tasa de descubrimiento, este hecho se opaca al necesitar un tamaño de muestra adicional mayor para obtener un número esperado de genes 'aceptable'.

Como nota final, hay que resaltar el hecho de que, si bien los datos anteriores pueden resultar un tanto crudos, su análisis muestra que los estimadores obtenidos son útiles y confiables bajo situaciones reales complejas.

4.2. Parámetros del proceso $PD(\sigma, \theta)$

Es necesario aclarar un punto importante sobre la aplicación y cálculo de los estimadores. Se trata de la obtención del valor de los dos parámetros del proceso $PD(\sigma, \theta)$ sobre el cual se está modelando el problema. Es necesario encontrar σ y θ tales que maximicen la fórmula de muestreo de Pitman (3.4), bajo la muestra observada $(k, n_1, \ldots, n_k)$, es decir

$$(\sigma, \theta) = \arg \max_{(\sigma,\theta)} \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}.$$

Para las librerías *Naegleria gruberi*, los parámetros que maximizan resultan ser $(\hat{\sigma}, \hat{\theta}) = (0.67, 46.3)$ para el caso del cultivo aeróbico y $(\hat{\sigma}, \hat{\theta}) = (0.66, 155.5)$ para el caso anaeróbico. Por el otro lado, las librerías de la *Mastigamoeba balamuthi* se obtienen los parámetros $(\hat{\sigma}, \hat{\theta}) = (0.7, 57)$ y $(\hat{\sigma}, \hat{\theta}) = (0.77, 46)$, para los datos normalizados y no normalizados respectivamente.

La elección de los parámetros en realidad está determinada por los datos a muestrear. El valor de θ se relaciona al número de distintos genes observados en la muestra inicial n; entre más grande sea la relación k/n mayor será $\hat{\theta}$. De forma similar, σ se determina por la configuración de las frecuencias de aparición de genes únicos $n_1, \ldots, n_k$. Además, se puede probar que para un cierto valor de θ, el número esperado de valores de nuevos genes, es una función creciente de σ, es decir, mientras σ se incrementa se espera registrar un mayor número de nuevos genes en una muestra adicional m.

4.3. Algunos aspectos técnicos

Sin entrar en detalles, a continuación se describe el lenguaje y compiladores utilizados para la obtención de los datos previamente expuestos.

En particular consideremos a la expresión (3.6), clave en la inferencia del PME. Si n y m, resultan ser demasiado grandes, la expresión anterior resulta

dificil de calcular. Para esta labor se utilizaron compiladores del lenguaje C, librerías PARI y el sofware R instalados. PARI es una biblioteca que colecciona rutinas en C permitiendo cálculos rápidos, funciona en sistemas windows pero tiene mejor desempeño en sistemas basados en linux.

Se compila la función C/PARI *PDpkj.c* para generar una librería dinámica (i.e. $libPDpkj.so$). Esta función contiene el código en C que relaciona las librerías PARI. Incluímos la librería dinámica a una rutina en R, creando el archivo discovery.r. La librería resultante contiene en particular la función

$$Pkj(n, m, k, theta, sigma, j, precision),$$

la cual calcula (3.6). La notación de los argumentos proviene de la expresión original, el único desconocido es *precision*, el cual es un argumento de precisión utilizado por PARI.

Para poder calcular el valor esperado de (3.6) que podemos obtener de la expresión simplificada (3.13), es posible utilizar Maple o Mathematica definiendo la función
E:=(j,t,s,n,m)− >(k+t/s)(pochhammer(t+s+n,m)/pochhammer(t+n,m)-1)*,
donde t y s reemplazan a *theta* y *sigma*, respectivamente.

Conclusiones

El enfoque bayesesiano no paramétrico nos provee de un modelo probabilístico completo, que permite tomar información conocida y transformarla en información predictiva. Su implementación a la solución del problema de muestreo de especies ha quedado ampliamente justificado y ha demostrado no tener problemas notables ante cualquier cambio de las variables, hablando específicamente sobre el tamaño de las muestras.

Se ha logrado dar una idea concisa y completa sobre el modelo no paramétrico basado en el proceso Poisson-Dirichlet de dos parámetros, abarcando un contexto teórico amplio que describe las bases matemáticas necesarias para comprender la esencia de los estimadores obtenidos. Desde ideas probabilísticas básicas que envuelven la teoría de la estadística bayesiana, hasta la introducción de procesos que derivan funciones predictivas.

Por su parte, el contexto que se da sobre estudios previos, permite resaltar las ventajas del enfoque bayesiano sobre otros estimadores, evitando las debilidades de acercamientos frecuentistas y teniendo mejores resultados sobre aplicaciones comparado a los enfoques estocásticos. Otra fortaleza del modelo no paramétrico es que no utiliza ningún supuesto paramétrico para funcionar.

Uno de los puntos más importantes a resaltar, consiste en que la implementación del proceso Poisson-Dirichlet de dos parámetros origina estimadores completamente explícitos, en particular para el nuevo número de especies, la tasa de descubrimiento y la cobertura.

La aplicación de estos estimadores a cuatro librerías genéticas EST resultan en predicciones interesantes que, en nuestro análisis, son coherentes bajo secuencias o muestras adicionales, lo cual se buscaba y esperaba obtener dentro de los objetivos del trabajo.

Esta información sugiere que implementaciones futuras sobre otro tipo de librerías arrojarán resultados de gran valor para investigadores en estos campos, proveyendo guías para tomar decisiones tales como determinar si resulta viable volver a secuenciar una cierta librería, determinando además el tamaño óptimo de la nueva muestra EST; o también tomar la decisión de efectuar el llamado protocolo de normalización, lo cual, en términos vanos, representa ahorro de recursos.

Un tema complementario a los estimadores, del cual no se discute en esta tesis, es el uso de los intervalos HDP (highest posterior density intervals); que de forma simple se pueden explicar como la contraparte bayesiana de los intervalos de confianza en el caso frecuentista. Su importancia no se desestima, pues complementan los resultados dando cotas de acción y dando mayor rango de decisión en el análisis de los datos. Si se desea completar la teoría sobre su construcción, bajo el contexto de este trabajo, es posible investigarla en Favaro et. al. (2009) [9].

La tesis presenta versiones simplificadas de los estimadores deseados, que resultan ser bastante buenos en la aplicación. Sin embargo, ésto no implica que su estudio se haya detenido, pues toda herramienta matemática puede ser perfectible.

Apéndice A

Coeficiente factorial generalizado

Coeficiente factorial generalizado central

Para todo $n \geq 1$ y $k = 0, \ldots, n$, el coeficiente factorial generalizado $\mathcal{C}(n, k; \sigma)$ se define como el coeficiente factorial del k-ésimo orden de t en la expansión del factorial generalizado de n-ésimo orden de t con parámetro de escala denotado por σ, i.e.

$$(\sigma t)_n = \sum_{k=0}^{n} \mathcal{C}(n, k; \sigma)(t)_k.$$

De la determinación sobre la distribución del número de distintas especies K_n que aparecen en una muestra de tamaño n,

$$\mathcal{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n,$$

aclarando que $\mathcal{C}(0, 0; \sigma) = 1$ y $\mathcal{C}(n, 0; \sigma) = 0$.

Coeficiente factorial generalizado no central

Se define como el coeficiente factorial del k-ésimo orden de t en la expansión del factorial generalizado no central de n-ésimo orden de t, denotado como $\mathcal{C}(n,k;\sigma,\gamma)$, con parámetro de escala σ y parámetro no central γ, i.e.

$$(\sigma t-\gamma)_n=\sum_{k=0}^{n}\mathcal{C}(n,k;\sigma,\gamma)(t)_k.$$

Entonces

$$\mathcal{C}(n,k;\sigma,\gamma)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(-j\sigma-\gamma)_n.$$

Es posible obtener una expresión del Coeficiente factorial generalizado no central a partir de la expresión central

$$\mathcal{C}(n,k;\sigma,\gamma)=\sum_{s=k}^{n}\binom{n}{s}\mathcal{C}(s,k;\sigma)(-\gamma)_{n-s}.$$

Expresiones a partir de números de stirling

$$\lim_{\sigma\to 0}=\frac{\mathcal{C}(n,k;\sigma)}{\sigma^k}=|S_{n,k}|,$$

donde $|S_{n,k}|$ es el valor absoluto de los números de stirling de primera especie. Además

$$\lim_{\sigma\to 0}=\sum_{i=k}^{n}\binom{n}{i}|S_{i,k}|(-\gamma)_{n-i}.$$

Bibliografía

[1] Aldous, D. *Exchangeability and related topics*. École d'Été de Probabilités de Saint-Flour XIII - 1983. Lectures Notes in Mathematics 1117. Springer-Verlag, Heidelberg. Berlin. 1985. pp 1-198.

[2] Bunge, J., Fitzpatrick, M. Estimating the Number of Species: A Review. *Journal of the American Statistical Association* **88**. No 421. 1993. pp 364-37.

[3] Chao,A. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal od Statistics, Theory and Applications*. No 11. 1984. pp 265-270.

[4] Chao,A., Lee,S.M. Estimating the Number of Classes Via Sample Coverage. *Journal of the American Statistical Association*. No 87. 1992. pp 210-217.

[5] Chochran, W. G. Sampling techniques. Wiley. ed 3ra. New York. 1977. 426 p.

[6] Efron, B. Robbins. Empirical Bayes and Microarrays . *The Annals of Statistics* **31**. No 2. 2003. pp 366-378.

[7] Efron, B., Thisted, R. Estimating the number of unseen species: how many words did Shakespeare know?. *Biometrika* **63**. No 3. 1976. pp 435-47.

[8] Ewens, W. The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**. 1972. pp 87-112.

[9] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. Bayesian nonparametric inference for species variety with two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society Series B* **71**. 2009. pp 993-1008.

[10] Ferguson, Thomas. Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**. 1973. pp 209-230.

[11] Fisher, R.A., Corbet, A.S. Williams, C.B. The Relation Between the Number of Species and the Number of Individuals un a Random Sample of an Animal Population. *Journal of Animal Ecology*. No 12. 1943. pp 42-58.

[12] Fristedt, B., Gray, L. A modern approach to probability theory. Birkhäuser. Boston. 1996. 780 p

[13] Good, I.J., Toulmin, G.H. The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased. *Biometrika*. No 43. 1956. pp 45-63.

[14] Goodman, L.A. On the Estimation of the Number of Classes in a Population. *Annals of Mathematical Statistics* **20**. 1949. pp 572-579.

[15] Hansen, B., Pitman, J. Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters* **46**. No 3. 2000. pp 251-256.

[16] Ishwaran, H., James, L. F. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* **96**. No 453. 2001. pp 161-173.

[17] Kingman, J. F. C. Random Partitions in Population Genetics. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences* **361**. No 1704, 1978, pp 1-20.

[18] Kozak, M. Finite and Infinite Populations in Biological Statistics: Should We Distinguish Them?. *The Journal of American Science* **4**. No 1. 2008. pp 59-62.

[19] Lijoi, A., Mena, R. H., Prünster, I. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**. 2007. pp 769-786.

[20] Lijoi, A., Mena, R. H., Prünster, I. A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics* **11**. 2007. Artículo No 339.

[21] Lijoi, A., Prünster, I. Models beyond the dirichlet process. In Bayesian Nonparametrics (Hjort,N.L., Homes, C.C., Müller, P. and Walker, S.G., eds.). Cambridge University Press. forthcoming.

[22] Navarrete, C., Quintana, F.A., Müller, P. Some Issues on Nonparametric Bayesian Modeling Using Species Sampling Models. *Statistical Modelling International Journal* **8**. No 1. 2008. pp 3-21.

[23] Mao, C. X. Prediction of the conditional probability of discovering a new class. *Journal of the American Statistical Association* **99**. No 468. 2004. pp 1108-1118.

[24] Mao, C. X., Lindsay, B. G. A Poisson model for coverage problems with an application in genomic research. *Biometrika* **89**. No 3. 2002. pp 669-681.

[25] Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability Theory Related Fields* **102**. 1995. pp 145-158.