UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

PROCESO DE POISSON-DIRICHLET: ENTROPÍA Y TIEMPOS DE APARICIÓN DE NUEVAS ESPECIES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

JAVIER IGNACIO SANTIBÁÑEZ MOLINA

PROFESOR GUÍA:
SERVET MARTÍNEZ AGUILERA

MIEMBROS DE LA COMISIÓN:
ALEJANDRO MAASS SEPÚLVEDA
JAIME SAN MARTÍN ARISTEGUI
PABLO MARQUET ITURRIAGA
SEBASTIÁN DONOSO FUENTES
THIERRY HUILLET

Este trabajo ha sido parcialmente financiado por CMM ANID PIA AFB170001, CMM ANID BASAL ACE210010 y CMM ANID BASAL FB210005

SANTIAGO DE CHILE
2022

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS Y MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: JAVIER IGNACIO SANTIBÁÑEZ MOLINA
FECHA: 2022
PROF. GUÍA: SERVET MARTÍNEZ AGUILERA

PROCESO DE POISSON-DIRICHLET: ENTROPÍA Y TIEMPOS DE APARICIÓN DE NUEVAS ESPECIES

En este trabajo se estudia el muestreo de especies y el proceso de aparición de nuevas especies, donde se tiene una secuencia de elementos que se clasifican de acuerdo a la especie a la cual pertenece. La cantidad de especies real es desconocida y se puede considerar infinita. Para esto, se estudia el Proceso de Poisson-Dirichlet de dos parámetros, que es un modelo Bayesiano no paramétrico que permite generar particiones aleatorias de infinitas componentes. Este proceso tiene dos propiedades muy importantes: es parcialmente conjugado consigo mismo y es capaz de modelar distribuciones con ley potencia. Bajo este contexto, se utiliza la entropía como medida de diversidad en la muestra. Por un lado, el resultado principal de esta tesis es una relación y unas desigualdades para la diferencia ponderada de la entropía esperada posterior entre dos pasos sucesivos en el caso del Proceso de Poisson-Dirichlet. Por otro lado, se estudia la distribución de los tiempos entre apariciones de nuevas especies, donde se obtiene un resultado para el tiempo esperado entre apariciones.

In this work we study the sampling of species and the process of appearance of new species, where we have a sequence of elements that are classified according to the species to which they belong. The actual number of species is unknown and can be considered infinite. For this, we study the two-parameter Poisson-Dirichlet Process, which is a Bayesian nonparametric model that allows to generate random partitions of infinite components. This process has two very important properties: it is partially conjugate with itself and it is able to model power-law distributions. In this context, entropy is used as a measure of diversity in the sample. On the one hand, the main result of this thesis is a relation and inequalities for the weighted difference of the posterior expected entropy between two successive steps in the case of the Poisson-Dirichlet Process. On the other hand, we study the distribution of times between appearances of new species, where we obtain a result for the expected time between appearances.

"*Do you know who you really are?*
Are you sure it's really you?"
Queens of the Stone Age (2013).

Agradecimientos

En primer lugar me gustaría agradecer a mi familia, que siempre me ha entregado apoyo y cariño. En particular, a mi mamá y a mi papá, gracias por todo lo que me han entregado a lo largo de todos estos años. También a mi hermana Paula, a la Juany y a mis perros, por ser parte de mi vida. A mis amigos más íntimos[1], los Chill & Chabe: Cucho, Li, Nico y Seba, el mejor grupo de amigos para crecer juntos y enfrentar la vida.

A mis amigos del DIM, los Chacras: Azócar, Barri, Benja, Bruno, Choco, Magali, Majo, Nico, Pelela, Plablo, Vicho y Vivi, gracias por hacer el paso por la carrera un agrado, lleno de matraca, comida, risas y juegos, con especial cariño a los tres chiflados, al igual que a la gente de la 437. También a los amigos que hice los primeros años de universidad, que a pesar de estar en diferentes carreras y las juntas no son tan seguidas como antes, el cariño siempre sigue ahí; los cabros del cucajón, el casi-podcast, el programa de radio, la película y las bandas que hicimos: Cáscaras sin Papa, Sailors of the Dead y Cáscara.

Muchas gracias a Servet Martínez, por ser un gran profesor, que me motivó a seguir el área de probabilidades, también por haberme dado la oportunidad de trabajar junto a él en un cuerpo docente por varios semestres, y por haber sido mi profesor guía en este arduo proceso. Además, agradezco a los miembros de la comisión por haber aceptado ser parte de esta y por sus comentarios y sugerencias, también a todos los académicos que fueron parte de mi proceso educativo.

Finalmente, agradecer a todos los músicos y artistas que he escuchado a lo largo de los años, su música ha sido esencial para mi vida.

[1] hermanos, fratelli, beratnas

Tabla de Contenido

Introducción

En este trabajo se intenta entender en mayor detalle un problema que surge desde el ámbito de la ecología, que es el muestreo de especies y el proceso de aparición de nuevas especies, donde uno de los desafíos es estimar de alguna manera la probabilidad de observar un elemento de una especie que no se ha visto antes.

Bajo este contexto, una de las primeras soluciones de este problema es desde un enfoque frecuentista, que fue expuesta por Good en 1953 [1], con el trabajo en conjunto de Turing. Cabe destacar, que este modelo también es utilizado en el procesamiento de lenguaje natural, por ejemplo para estudiar la frecuencia de palabras en un documento. De hecho, el modelo fue desarrollado por Turing durante la Segunda Guerra Mundial, para descifrar mensajes de guerra. Se considera una muestra aleatoria de tamaño N de una población de animales de distintas especies, donde la especie j-ésima fue observada n_j veces con $j = 1, \ldots, k$, donde k es el número de especies distintas observadas. Luego, el estimador n_j/N para la frecuencia poblacional, o la probabilidad de aparición de la especie j-ésima, no es un buen estimador, ya que no considera el caso en que la siguiente muestra sea de una especie totalmente nueva, es decir, para la especie $k + 1$ se asignaría una probabilidad 0 de ocurrir.

Para evitar esto, el estimador de Good-Turing reduce la probabilidad total de los elementos observados, con el fin de asignar un estimador con probabilidad no nula a los elementos de especies que aún no se han observado. Esto mediante el valor m_l, que es el número de especies con frecuencia l en la muestra, o también conocido como la frecuencia de la frecuencia l. Se sabe que el estimador de Good-Turing no se comporta bien cuando l es grande. Esto se puede mejorar con ciertas reglas de suavizado, pero pese a esto, el estimador no logra comportarse bien en todos los casos (esto se explica con más detalle en la Sección 1.6).

Una solución más reciente a este problema es bajo el enfoque Bayesiano, es decir, donde los parámetros se modelan como variables aleatorias. En este contexto, se tiene inicialmente el Proceso de Dirichlet de un parámetro definido por Ferguson en 1973 [2]. El proceso en el cual nos enfocaremos en esta tesis es la generalización de dos parámetros, conocido como el Proceso de Poisson-Dirichlet planteado por Pitman y Yor en 1997 [3]. Este proceso es también un modelo no paramétrico, por lo que es capaz de modelar particiones aleatorias, con un número infinito de componentes del intervalo $[0, 1]$. Como en la población que se está muestreando, el número total de especies es desconocido, bajo el contexto del Proceso de Poisson-Dirichlet, podemos suponer que en la población hay infinitas numerables especies, es decir, a medida que se observan datos siempre pueden ir apareciendo nuevas especies de manera indefinida. La probabilidad de aparición de una nueva especie es siempre positiva. Así, este proceso permite trabajar la muestra aunque se desconozca el verdadero número de especies.

Suponiendo que la muestra observada proviene de un Proceso de Poisson-Dirichlet, una de las propiedades más importantes de este modelo es que la distribución posterior es una mezcla de una distribución finita Dirichlet y un Proceso de Poisson-Dirichlet. Esto permite estudiar la distribución de las especies de la partición, a medida que se observan nuevos elementos simplemente actualizando la distribución posterior. Además, otra propiedad interesante de este proceso, es que permite modelar distribuciones del tipo ley potencia (*power-law*), las cuales son muy comunes de observar en datos biológicos y ecológicos.

Gracias a estas características se decide trabajar en el contexto del Proceso de Poisson-Dirichlet. En el último tiempo este proceso ha tenido aplicaciones en muchos otros ámbitos aparte de la ecología, debido en parte al gran tamaño de datos que se están recolectando y la capacidad de cómputo que se tiene. Por ejemplo, se pueden encontrar aplicaciones en genética [4], procesamiento de lenguaje natural [5] e incluso en finanzas [6].

En este trabajo se desea estudiar la entropía del Proceso de Poisson-Dirichlet. Considerando que se tiene una muestra de una población, que se puede suponer infinita de animales de distintas (también infinitas) especies, la entropía corresponde a una forma de medir la información o la diversidad de la comunidad que se tiene dada la muestra. En [7] se tiene un notable resultado para estimar la entropía usando la esperanza de la distribución posterior en el caso Bayesiano, en particular para el Proceso de Poisson-Dirichlet.

Otro concepto que se desea estudiar en el contexto de este proceso, es la distribución de los tiempos entre apariciones de nuevas especies. En [8] se tienen los principales resultados sobre esta distribución para el caso de dos parámetros, mientras que en [9] se analiza el tiempo esperado entre apariciones para el caso de un parámetro en el Proceso de Dirichlet.

Dado lo anterior, el objetivo principal de esta tesis es obtener una relación para la variación de la entropía esperada posterior, entre dos pasos sucesivos para el Proceso de Poisson-Dirichlet, basados en el resultado expuesto en [7] e inspirados en lo hecho en el caso frecuentista en [10]. Esto para entender la cantidad de información o diversidad que aporta la aparición de una nueva especie en la muestra.

El propósito de este trabajo es también entender la secuencia de los tiempos entre apariciones de nuevas especies, y encontrar el tiempo esperado entre apariciones en el caso de dos parámetros, mediante los resultados obtenidos en [8] y [9]. Además de esto, se desea comprender la relación del Proceso de Poisson-Dirichlet con otros modelos, en particular con el modelo de Good-Turing estudiando lo expuesto en [11].

La estructura de esta tesis es de la siguiente manera: primero en el Capítulo 1 se expone el Proceso de Poisson-Dirichlet (PDP) y sus propiedades fundamentales, además de una sección con la relación de este proceso y el modelo de Good-Turing. Segundo, en el Capítulo 2 se estudia la entropía, y en especial la entropía Bayesiana para el caso del PDP. Posteriormente, el Capítulo 3 expone la definición y propiedades de los tiempos entre apariciones, se demuestran algunos resultados de la distribución de Waring y se presenta un resultado de esta tesis que es el tiempo esperado entre apariciones. Finalmente, en el Capítulo 4 se revisa la variación de la entropía en el caso frecuentista y se presenta el resultado principal de esta tesis, que es una relación y unas desigualdades para la variación de la entropía para el PDP.

Capítulo 1

Proceso de Poisson-Dirichlet

Este capítulo está dedicado a las definiciones y propiedades principales relacionadas al Proceso de Poisson-Dirichlet, de acuerdo al interés de este trabajo de tesis. Para esto, hemos tomado como inspiración lo expuesto en [12], [13], [14], [15], [5] y [7]. Muchos de los resultados serán enunciados sin demostración, ya que estas se pueden encontrar en la mayoría de las referencias recién mencionadas. Además, cuando sea pertinente se dará la referencia exacta para encontrar los resultados o en caso de que el lector desee profundizar en el tema.

La estructura de este capítulo es la siguiente: primero en la Sección 1.1 se definen los modelos Bayesianos no paramétricos, en particular los modelos de mezcla. Segundo, en la Sección 1.2 se introduce el Proceso de Stick-breaking. Después en la Sección 1.3 se define el Proceso de Poisson-Dirichlet para luego en la Sección 1.4 mostrar las propiedades principales de este proceso además de definir el Proceso del Restaurante Chino. En la Sección 1.5 se muestran simulaciones de los procesos anteriores. Finalmente, en la Sección 1.6 se expone el modelo de Good-Turing y su relación con el Proceso de Poisson-Dirichlet.

1.1. Modelos Bayesianos no paramétricos

Los modelos probabilísticos son usados para modelar la distribución de datos observados. Los modelos clásicos *paramétricos* usan un número fijo y finito de parámetros. Si el espacio de parámetros tiene dimensión infinita entonces se le llama modelo *no paramétrico*. Por ejemplo, una regresión lineal en $\mathbb{R}^2$ corresponde a un modelo paramétrico, esto pues se tiene que el espacio de parámetros es de dimensión 2, ya que las funciones lineales son representadas solo por la pendiente y el intercepto. En el caso que los datos observados sean no lineales se puede considerar el conjunto de todas las funciones (continuas y dos veces diferenciable, si se desea) en $\mathbb{R}^2$, lo que hace que el espacio de parámetros sea de dimensión infinita, es decir, se tendría un modelo no paramétrico.

Los modelos *Bayesianos* son aquellos donde los parámetros se modelan como variables aleatorias. En otras palabras, el enfoque Bayesiano toma la incertidumbre del valor del parámetro y la expresa como aleatoriedad. En estos modelos se asume una cierta distribución sobre los parámetros denominada distribución *prior*, la cual se escoge para intentar capturar la información de los datos observados o por conocimientos previos que se tengan. A la distribución después de haber visto los datos, suponiendo que vienen de la distribución prior, se le llama distribución *posterior*. Luego, los modelos Bayesianos no paramétricos son modelos

donde los parámetros son variables aleatorias y cuyo espacio de parámetros tiene dimensión infinita. Que un modelo sea Bayesiano permite evitar tener que escoger un solo valor para cada parámetro y que sea no paramétrico permite al modelo aumentar la dimensionalidad de los parámetros al observar nuevos datos.

Para definir una distribución en un espacio de dimensión infinita, se necesita la noción de un *modelo de mezcla*. En este tipo de modelos se necesita una medida base y este entrega una medida discreta. Se recuerda que la notación $X \sim G$ representa que la variable aleatoria X distribuye según G, donde G denota una medida de probabilidad (o más adelante denota alguna distribución conocida).

Definición 1.1 (Modelo de mezcla) *Dada G una medida de probabilidad sobre un espacio medible $\mathcal{X}$. Sea $\phi = (\phi_k)_{k\geq 1}$ tal que $\phi_k \sim G$ e independientes, además sea $\pi = (\pi_k)_{k\geq 1}$ a valores reales no negativos independiente de ϕ tal que $\sum_{k\in\mathbb{N}} \pi_k = 1$. Luego*

$$\Theta(\cdot) = \sum_{k\in\mathbb{N}} \pi_k \delta_{\phi_k}(\cdot) \tag{1.1}$$

es una medida discreta aleatoria (sobre $\mathcal{X}$) que corresponde a un modelo de mezcla. Donde δ es la medida de Dirac, es decir, $\delta_{\phi_k}(A) = \begin{cases} 1 & \phi_k \in A \\ 0 & \phi_k \notin A \end{cases}$ para todo conjunto medible A de $\mathcal{X}$.

Este modelo recibe el nombre de mezcla por ser la suma ponderada de distintas medidas. Los valores de ϕ se llaman los *átomos* del modelo y como son variables aleatorias este modelo también corresponde a un modelo Bayesiano. Los valores de π son los *pesos* del modelo los cuales también son variables aleatorias, si se tiene que hay un número finito de valores π_k distintos de cero sería un modelo de mezcla finito, el caso de interés es cuando se tiene un número infinito, que corresponde a un modelo no paramétrico.

Cuando la medida base G es continua, por ejemplo una Normal$(0, 1)$, se tiene que todos los átomos son distintos casi seguramente. La propiedad general se llama *no atómica*, es decir, $G(\{A\}) = 0$ para todo $A \in \mathcal{X}$. La propiedad contraria ocurre cuando la distribución es discreta, entonces $G(\{A\}) > 0$ para todo $A \in \mathcal{X}$ y puede ocurrir que $\phi_k = \phi_l$ para $k \neq l$. En lo que sigue de esta tesis se considera que siempre se trabaja con una medida no atómica.

Para construir este tipo de modelo, se escoge la medida G y se toman muestras independientes idénticamente distribuidas $\phi_1, \phi_2, \cdots \sim_{iid} G$. Para los pesos es más complejo, se podrían tomar muestras de una distribución en $[0, 1]$ pero con esto los pesos no suman necesariamente 1. En el caso finito esto se puede solucionar normalizando las variables, para M componentes se toman variables $\beta_1, \ldots, \beta_M$ en $[0, \infty)$ y se define

$$\pi_k := \frac{\beta_k}{T} \quad \text{donde} \quad T := \beta_1 + \cdots + \beta_M.$$

Con esto se tiene claramente que los pesos cumplen que suman 1. Para el caso infinito, se necesita más cuidado y para esto se usa el proceso de Stick-breaking que se introduce en la siguiente sección.

1.2. Proceso de Stick-breaking

La idea del proceso de Stick-breaking consiste en considerar el intervalo $[0,1]$ como un palo que se va rompiendo, de manera aleatoria, para obtener los pesos. El primer peso π_1 se obtiene de alguna distribución en $[0,1]$, para el peso siguiente la idea es que tome valores en el segmento restante del intervalo (palo), o sea que π_2 tome valores en $[0, 1-\pi_1]$. Se sigue así y π_k, dado que ya se obtuvieron los primeros $k-1$ valores, debe ser una distribución en $[0, 1-(\pi_1+\cdots+\pi_{k-1})]$. Con esto, se puede definir formalmente el proceso de Stick-breaking, donde las distribuciones son en $[0,1]$ y se escalan para que sean del segmento del intervalo correspondiente.

Definición 1.2 (Proceso de Stick-breaking) *Dadas distribuciones $(H_k)_{k\geq 1}$ en $[0,1]$ se genera la secuencia $\pi = (\pi_k)_{k\geq 1}$ a partir de $\beta_1 \sim H_1, \beta_2 \sim H_2, \ldots,$ de forma independiente, como*

$$\pi_1 := \beta_1, \quad \pi_k := \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \quad k \geq 2. \tag{1.2}$$

La secuencia π se llama proceso de Stick-breaking generado por $(H_k)_{k\geq 1}$ y cumple que $\sum_{k\geq 1}\pi_k = 1$ casi seguramente.

Para ver cuando la suma de los pesos es 1 casi seguramente (c.s.), primero notamos que las sumas parciales cumplen

$$\sum_{k=1}^{n}\pi_k = 1 - \prod_{j=1}^{n}(1-\beta_j),$$

lo cual no es difícil de probar. Por inducción el caso base es $\pi_1 = \beta_1 = 1-(1-\beta_1)$. Luego, $\sum_{k=1}^{n}\pi_k = \sum_{k=1}^{n-1}\pi_k + \pi_n = 1 - \prod_{j=1}^{n-1}(1-\beta_j) + \beta_n\prod_{j=1}^{n-1}(1-\beta_j) = 1-\prod_{j=1}^{n}(1-\beta_j)$. De esta manera, se tiene que $\sum_{k=1}^{\infty}\pi_k = 1$ c.s. si y solo si $\prod_{j=1}^{\infty}(1-\beta_j) = 0$ c.s. Para esto notamos que $\prod_{j=1}^{n}(1-\beta_j) \geq 0$ y es un producto decreciente, entonces $\prod_{j=1}^{\infty}(1-\beta_j) = 0$ c.s. si y solo si $\mathbb{E}(\prod_{j=1}^{\infty}(1-\beta_j)) = 0$. Luego, por independencia $\mathbb{E}(\prod_{j=1}^{\infty}(1-\beta_j)) = \prod_{j=1}^{\infty}(1-\mathbb{E}(\beta_j)) = 0$, como $0 < \mathbb{E}(\beta_j) < 1$, esto último se cumple si y solo si $\sum_{j=1}^{\infty}\mathbb{E}(\beta_j) = \infty$. Es decir, se tiene que

$$\sum_{k=1}^{\infty}\pi_k = 1 \text{ c.s.} \quad \text{si y solo si} \quad \sum_{j=1}^{\infty}\mathbb{E}(\beta_j) = \infty. \tag{1.3}$$

Con esto, ya se pueden generar las dos secuencias infinitas ϕ y π para construir la medida Θ dada por (1.1). Un caso particular de este proceso es la distribución Griffiths-Engen-McCloskey (GEM) que será de gran utilidad para definir el Proceso de Poisson-Dirichlet.

Definición 1.3 (Distribución GEM) *Para $0 \leq d < 1$ y $\alpha > -d$ a la secuencia π generada por (1.2) para las distribuciones $H_k = Beta(1-d, \alpha+kd)$ define la llamada distribución GEM de parámetros d y α, abreviado GEM(d,α).*

Para $\pi \sim \text{GEM}(d,\alpha)$ verifiquemos que se cumple (1.3). En efecto, usando la esperanza de una variable beta, que en este caso es $\mathbb{E}(\beta_j) = \frac{1-d}{1-d+\alpha+jd}$, y la relación $(j-1)(1+\alpha)+(j-1)d \geq$

$1 + \alpha + (j-1)d$, se tiene que

$$\frac{1-d}{1+\alpha+d}\sum_{j=1}^{\infty}\frac{1}{j-1} \leq \sum_{j=1}^{\infty}\frac{1-d}{1+\alpha+(j-1)d} = \sum_{j=1}^{\infty}\mathbb{E}(\beta_j).$$

Como en el lado izquierdo aparece la serie armónica, que se sabe diverge, entonces $\sum_{j=1}^{\infty}\mathbb{E}(\beta_j)$ también diverge y así se cumple $\sum_{k=1}^{\infty}\pi_k = 1$ c.s. para la distribución GEM.

Cabe destacar que en la literatura se puede encontrar esta distribución, y posteriormente también el Proceso de Poisson-Dirichlet, con el parámetro θ (nuestro α), y el parámetro α (nuestro d).

1.3. Proceso de Poisson-Dirichlet

Con todo lo anterior ya podemos definir el Proceso de Poisson-Dirichlet.

Definición 1.4 (Proceso de Poisson-Dirichlet) *Dada G una medida de probabilidad sobre un espacio medible $\mathcal{X}$. Para $0 \leq d < 1$ y $\alpha > -d$ sea $\pi \sim GEM(d, \alpha)$. Además sea $\phi = (\phi_k)_k$ para $k = 1, 2 \ldots$ una secuencia independiente que distribuyen según G e independiente de π. Luego, a la medida discreta aleatoria*

$$\Theta(\cdot) = \sum_{k\geq 1}\pi_k\delta_{\phi_k}(\cdot) \tag{1.4}$$

se le llama Proceso de Poisson-Dirichlet con medida base G y parámetros d y α, abreviado $PDP(d, \alpha, G)$.

Hay veces donde no es necesario explicitar la medida base y se escribe simplemente $PDP(d, \alpha)$, esto ocurre generalmente cuando se desea tomar una muestra del proceso y, de las clases observadas, solo interesa el momento cuando aparecen y su frecuencia. El parámetro d se denomina el parámetro de *descuento* y α el parámetro de *concentración*. Este último controla cuanta probabilidad se concentra en las primeras muestras, para α pequeño los primeros pesos son los que tienen la mayor cantidad de masa de probabilidad, mientras que para α grande la probabilidad está más esparcida y así los pesos π_k son más uniformes. El parámetro de descuento d controla la cola de la distribución de los pesos, para $d = 0$ se obtienen colas exponenciales: $\pi_k \propto e^{-k/\alpha}$ y para d más cerca de 1 se obtienen colas con ley potencia (*power-law*): $\pi_k \propto k^{-1/d}$. La importancia de las distribuciones con ley potencia es que son más comunes y acorde a la realidad, por ejemplo se pueden encontrar en la frecuencia de palabras en el idioma inglés, el tamaño de ciudades, el tamaño de cuerpos de agua (charcos, lagos, océanos) en la naturaleza, etc (ver [12], [16]).

El caso especial cuando $d = 0$ lleva el nombre de Proceso de Dirichlet y aparte de tener un comportamiento distinto para la cola de la distribución de los pesos, cumple varias propiedades interesantes y similares al Proceso de Poisson-Dirichlet. En general se define de otra manera: como la extensión infinita de una distribución de Dirichlet. La forma de construcción a partir del proceso de stick-breaking es igualmente válida y equivalente a la siguiente definición.

Definición 1.5 (Proceso de Dirichlet) *Sean G una medida de probabilidad sobre un espacio medible $\mathcal{X}$ y $\alpha > 0$. Se dice que una medida de probabilidad Θ es un proceso de Dirichlet con medida base G y parámetro de concentración α, denotado como $\Theta \sim DP(\alpha, G)$ si para toda partición finita medible $(A_1, \ldots, A_r)$ de $\mathcal{X}$, se tiene que el vector $(\Theta(A_1), \ldots, \Theta(A_r))$ sigue una distribución de Dirichlet de parámetro $(\alpha G(A_1), \ldots, \alpha G(A_r))$.*

1.4. Propiedades principales y Proceso del Restaurante Chino

Para entender de mejor manera el comportamiento de los parámetros involucrados en el Proceso de Poisson-Dirichlet, se estudia la esperanza y varianza.

Proposición 1.1 (ver [17]) *Sea $\Theta \sim PDP(d, \alpha, G)$, para todo conjunto medible $A \subseteq \mathcal{X}$ se tiene que*

$$\mathbb{E}(\Theta(A)) = G(A) \quad y \quad \mathrm{Var}(\Theta(A)) = G(A)(1 - G(A))\frac{1-d}{1+\alpha}. \tag{1.5}$$

Demostración. Para la demostración se necesita una propiedad de los pesos $\pi \sim GEM(d, \alpha)$, que permite trabajar la esperanza sobre π en una esperanza del primer y segundo peso (de mayor tamaño), la cual se puede encontrar en [7], donde f y g son funciones cualquiera

$$\mathbb{E}\left(\sum_{k=1}^{\infty} f(\pi_k)\right) = \mathbb{E}\left(\frac{f(\pi_1)}{\pi_1}\right), \tag{1.6}$$

$$\mathbb{E}\left(\sum_{i,j\neq i} g(\pi_i, \pi_j)\right) = \mathbb{E}\left(\frac{g(\pi_1, \pi_2)}{\pi_1 \pi_2}(1 - \pi_1)\right). \tag{1.7}$$

Luego, sea $A \subseteq \mathcal{X}$, la esperanza es

$$\begin{aligned}\mathbb{E}(\Theta(A)) &= \mathbb{E}\left(\sum_{k\geq 1} \pi_k \delta_{\phi_k}(A)\right) = \sum_{k\geq 1} \mathbb{E}(\pi_k)\mathbb{E}(\delta_{\phi_k}(A)) = \sum_{k\geq 1} \mathbb{E}(\pi_k)\mathbb{P}(\phi_k \in A) \\ &= \sum_{k\geq 1} \mathbb{E}(\pi_k) G(A) = G(A)\mathbb{E}\left(\sum_{k\geq 1} \pi_k\right) = G(A).\end{aligned}$$

Donde se usa (1.6). Para la varianza, veamos el momento de segundo orden

$$\begin{aligned}\mathbb{E}(\Theta(A)^2) &= \mathbb{E}\left(\left(\sum_{k\geq 1} \pi_k \delta_{\phi_k}(A)\right)^2\right) = \mathbb{E}\left(\sum_{k\geq 1} \pi_k^2 \delta_{\phi_k}(A)\right) + \mathbb{E}\left(\sum_{i,j\neq i} \pi_i \pi_j \delta_{\phi_i}(A)\delta_{\phi_j}(A)\right) \\ &= G(A)\mathbb{E}(\pi_1) + G(A)^2\mathbb{E}(1 - \pi_1),\end{aligned}$$

donde se usan las ecuaciones (1.6) y (1.7). Como $\pi_1 \sim \mathrm{Beta}(1-d, \alpha+d)$, entonces $\mathbb{E}(\pi_1) = \frac{1-d}{1+\alpha}$. Así,

$$\mathbb{E}(\Theta(A)^2) = G(A)\mathbb{E}(\pi_1) + G(A)^2\mathbb{E}(1 - \pi_1) = \frac{1-d}{1+\alpha}(G(A) - G(A)^2) + G(A)^2.$$

Por lo tanto,

$$\mathrm{Var}(\Theta(A)) = \mathbb{E}(\Theta(A)^2) - \mathbb{E}(\Theta(A))^2 = G(A)(1-G(A))\frac{1-d}{1+\alpha}.$$

□

De la ecuación (1.5) tenemos que la media del proceso es la medida base y que, para d fijo, el parámetro α se puede entender como el inverso de la varianza. Por un lado, para α tendiendo a 0 las muestras se concentran todas en un solo valor, por el otro lado, para α grande el proceso se concentra más alrededor de su media, es decir, las muestras que se tomen serán más similares a la distribución de la medida base. Esto último también ocurre cuando d se acerca a 1, como se enuncia a continuación.

Proposición 1.2 (ver [17]) *Sea $\Theta \sim PDP(d, \alpha, G)$, para todo conjunto medible $A \subseteq \mathcal{X}$ se tiene que,*

$$\Theta(A) \xrightarrow{p} G(A)$$

cuando $\alpha \to \infty$ (para d fijo) o cuando $d \to 1$ (para α fijo), donde $\xrightarrow{p}$ denota la convergencia en probabilidad.

Tomar muestras de (1.4) no es una tarea fácil, la primera muestra X_1 se obtiene de forma aleatoria dependiendo de los pesos π, supongamos el asociado a π_{k_1} el cual a su vez tiene asociado un átomo ϕ_{k_1}, que sabemos distribuye según G, que será la muestra X_1 observada y corresponde a la clase o especie representante de X_1 que denotaremos X_1^*. Como Θ es una medida discreta, entonces siempre es posible sacar nuevamente el átomo ϕ_{k_1} del proceso, si la segunda muestra X_2 es este átomo entonces se agrupan en la misma clase de la primera especie observada X_1^*, pero si es un nuevo átomo ϕ_{k_2} entonces esta muestra pertenece a una nueva clase que se denota X_2^*.

Esto continua hasta tener una muestra de tamaño N y donde cada átomo o especie tiene un número n_j de veces que fue observado, con $j = 1, \ldots, k$ donde k es el número total de especies distintas observadas. En este proceso de tomar muestras, los índices obtenidos $k_1, k_2, \ldots$ se deducen de los datos observados, por lo que pueden ser ignorados y enumerarlos a medida que se van observando las especies correspondientes como $X_1^*, \ldots, X_k^*$.

Entonces, consideremos en lo que sigue $\Theta \sim \mathrm{PDP}(d, \alpha, G)$ y una secuencia de muestras independientes $X_1, \ldots, X_N$ de Θ. Para los k valores distintos que aparecen denotemos sus clases de equivalencia (o especies) como $X_1^*, \ldots, X_k^*$, como fue explicado anteriormente, y sus respectivas frecuencias como $n_1, \ldots, n_k$, es decir, $n_j = \sum_{i=1}^{N} \mathbb{1}_{\{X_i \in X_j^*\}}$ y además $N = \sum_{j=1}^{k} n_j$. Así, se tiene la probabilidad condicional para una nueva observación dada la muestra de tamaño N ya observada y los parámetros del PDP.

Proposición 1.3 (ver [13]) *Considerando $\mathbf{X}_N = (X_1, \ldots, X_N)$, la probabilidad condicional de una nueva observación corresponde a*

$$\mathbb{P}(X_{N+1} | \mathbf{X}_N, d, \alpha, G) = \frac{\alpha + kd}{\alpha + N}\delta_{G(\cdot)} + \sum_{j=1}^{k} \frac{n_j - d}{\alpha + N}\delta_{X_j^*}(\cdot). \tag{1.8}$$

Para esta probabilidad tenemos que la nueva observación forma parte de una especie (clase) ya observada $X_{N+1} \in X_j^*$ con probabilidad $\frac{n_j - d}{\alpha + N}$, o aparece una nueva especie tomando una muestra según ley G con probabilidad $\frac{\alpha + kd}{\alpha + N}$. Notamos que la frecuencia juega un rol fundamental pues, para d cercano a cero, la probabilidad de estar en una especie con hartos elementos es bastante alta, es decir, se cumple el fenómeno de que los ricos se vuelven más ricos. Al aumentar el valor de d se puede suavizar este fenómeno, aumentando la probabilidad de descubrir nuevas especies pero de menor tamaño cada una.

El Proceso de Poisson-Dirichlet, viendo la secuencia generada por la ecuación (1.8), tiene una representación llamada el Proceso del Restaurante Chino. En esta analogía se considera un restaurante con un número ilimitado de mesas y de asientos para cada mesa.

- Un *cliente* llega al restaurante y ve k *mesas* ocupadas donde n_j personas se sientan en la mesa j disfrutando de la *comida* X_j^*.
- Este cliente puede empezar su propia mesa con probabilidad $\frac{\alpha + kd}{\alpha + N}$ y recibir una nueva comida X_{k+1}^* del menú G.
- De lo contrario, va a una de las k mesas existentes con probabilidad $\frac{n_j - d}{\alpha + N}$ y disfruta de la comida X_j^*.

Luego, el Proceso del Restaurante Chino (CRP) se define sobre las particiones que se generan.

Definición 1.6 (Proceso del Restaurante Chino) *Para $0 \leq d < 1$ y $\alpha > -d$, se genera una secuencia de enteros $m_1, m_2, \ldots$ a partir de*

$$\mathbb{P}(m_{N+1} | \mathbf{m}_N, d, \alpha) = \frac{\alpha + kd}{\alpha + N} \delta_{\{m_{N+1} = k+1\}} + \sum_{j=1}^{k} \frac{n_j - d}{\alpha + N} \delta_{\{m_{N+1} = j\}}.$$

La secuencia de enteros generada corresponde al Proceso del Restaurante Chino, denotado $CRP(d, \alpha)$.

Este proceso es de gran utilidad como alternativa para tomar muestras de un Proceso de Poisson-Dirichlet donde el vector π es desconocido. La analogía de los clientes y las mesas se puede entender también como los animales observados y las especies mencionado previamente.

En lo anterior, se puede considerar la variable aleatoria K como el números de especies (o mesas) distintas en la muestra, donde la realización $K = k$ se asume generalmente sin ser mencionada explícitamente. Luego, esta variable tiene una distribución y un valor esperado que es interesante de mencionar. Para esto, se necesita la notación del *símbolo de Pochhammer* $(x)_N$ que corresponde a $(x)_N := x(x+1)\ldots(x+N-1) = \Gamma(x+N)/\Gamma(x)$ y $(x)_0 = 1$, para x un número real positivo y N entero no negativo, además del *símbolo de Pochhammer con incremento* $(x|y)_N := x(x+y)\ldots(x+(N-1)y)$ y $(x|0)_N = x^N$.

Proposición 1.4 (ver [13]) *Para una muestra de un PDP de parámetros (d, α), donde solo*

se conoce que es de tamaño $N \geq 1$, se tiene que la distribución condicional de K es

$$\mathbb{P}(K = k|N, d, \alpha) = \frac{(\alpha|d)_k}{(\alpha)_N} S^N_{k,d}, \quad k = 1, \ldots, N. \tag{1.9}$$

La ecuación (1.9) se obtiene al integrar sobre todas las posibles particiones de tamaño k para una muestra de tamaño N. El valor $S^N_{k,d}$ corresponde a una generalización de los números de Stirling cuya fórmula se puede encontrar en [13] junto a diferentes maneras de computarlo.

Proposición 1.5 (ver [13]) *Considerando una muestra de un PDP de parámetros (d, α) donde solo se conoce que es de tamaño $N \geq 1$, el valor esperado de la cantidad de especies para $d > 0$ viene dado por*

$$\begin{aligned}\mathbb{E}(K|N, d, \alpha) &= \frac{\alpha}{d}\frac{(\alpha + d)_N}{(\alpha)_N} - \frac{\alpha}{d} \\ &\approx \frac{\alpha}{d}\left(1 + \frac{N}{\alpha}\right)^d \exp\left(\frac{dN}{2\alpha(\alpha + N)}\right) - \frac{\alpha}{d}, \quad \textit{para } N, \alpha \gg d.\end{aligned}$$

Donde $\approx$ significa aproximadamente igual y $\gg$ mucho mayor. Para $d = 0$

$$\begin{aligned}\mathbb{E}(K|N, d, \alpha) &= \alpha(\psi(\alpha + N) - \psi(\alpha)) \\ &\approx \alpha \log\left(1 + \frac{N}{\alpha}\right), \quad \textit{para } N, \alpha \gg 0.\end{aligned}$$

Donde $\psi(\cdot)$ corresponde a la función digamma $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$ la cual será explicada con más detalle más adelante. Con esto, tenemos que la cantidad de especies esperada para el Proceso de Dirichlet ($d = 0$) es de orden $\mathcal{O}(\log N)$ y para el Proceso de Poisson-Dirichlet es $\mathcal{O}(N^d)$, lo que nos confirma la intuición sobre la rapidez con la que van apareciendo nuevas especies a medida que se observan nuevos datos.

1.5. Simulaciones

En esta sección se muestran simulaciones realizadas para entender mejor el comportamiento del Proceso de Poisson-Dirichlet, además del proceso de Stick-breaking y del Restaurante Chino. Como los pesos $(\pi_k)_{k\geq 1}$ son un vector infinito para poder generarlos hay que tener algunas consideraciones.

En primer lugar, para el Proceso de Stick-breaking se considera el caso particular de la distribución GEM. Para d y α, definidos previamente, se toman muestras $\beta_k \sim \text{Beta}(1 - d, \alpha + kd)$ y se construyen los pesos mediante la ecuación (1.2). Esto continua hasta que el peso restante del intervalo es insignificante, del orden 10^{-6}, o hasta que ya se hayan generado 1000 valores de π_k, esto último pues para ciertos valores la cantidad de muestras necesarias para que el peso restante sea pequeño es muy grande, por lo que el tiempo de ejecución también lo es.

El algoritmo explicado claramente coincide con la distribución GEM cuando no se con-

sideran las dos reglas de detención, pero en la práctica obviamente se necesita un vector finito, por lo que este algoritmo es una aproximación no exacta de $(\pi_k)_{k\geq 1}$. En [18] plantean algoritmos de simulación exacta para estos pesos mediante una descomposición del vector, en [17] también platean un algoritmo que es más eficiente con respecto a la aproximación de Stick-breaking. A pesar de esto, se decide implementar este método por su simpleza y rápida ejecución.

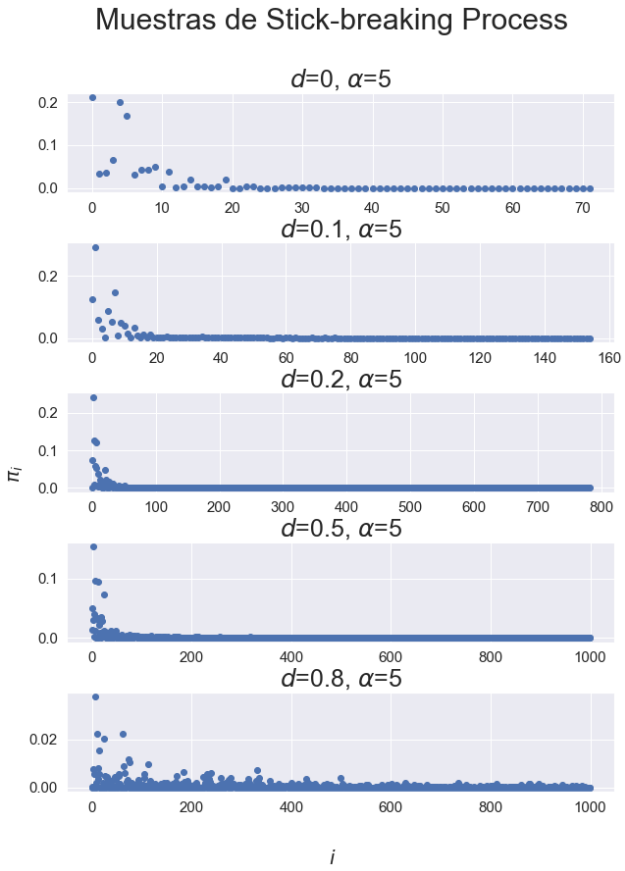


Figura 1.1: Muestras del Proceso de Stick-breaking considerando la distribución GEM, para $\alpha = 5$ y distintos d.

En la Figura 1.1 se tienen muestras tomadas del Proceso de Stick-breaking para la distribución GEM con $\alpha = 5$ y distintos valores de d. Notamos que en todos los casos la mayor parte del peso se concentra en los primero valores y mientras más aumenta d se generan más π_k hasta llegar al máximo definido de 1000. Además, se tiene que los primeros pesos van disminuyendo en valor a medida que aumenta d, de un peso de 0.2 para $d = 0$ hasta 0.02 para $d = 0.8$. Todo esto nos confirma el comportamiento del parámetro d, pensando en el Proceso de Poisson-Dirichlet, ya que mientras más se acerca a 1 aparecen más clases pero serán de menor tamaño pues el peso, o probabilidad de estar en una clase, es pequeño.

Con los pesos listos lo segundo que se hace es generar muestras de un Proceso de Poisson-Dirichlet, donde basta definir una medida base y tomar muestras de esta para obtener los átomos $(\phi_k)_{k\geq 1}$, y así poder tomar muestras del PDP. Esto se observa en la Figura 1.2 donde se tienen muestras del Proceso de Poisson-Dirichlet para una medida base Normal$(0, 1)$,

parámetro de descuento $d = 0.2$ y distintos valores de α para 1000 muestras cada uno. Se tiene que para los α pequeños las muestras se concentran prácticamente en un solo valor, además mientras más crece α empiezan a aparecer más clases y las muestras se aproximan a una distribución normal que en este caso corresponde a la distribución de la medida base, lo que valida lo expuesto en la Proposición 1.2.

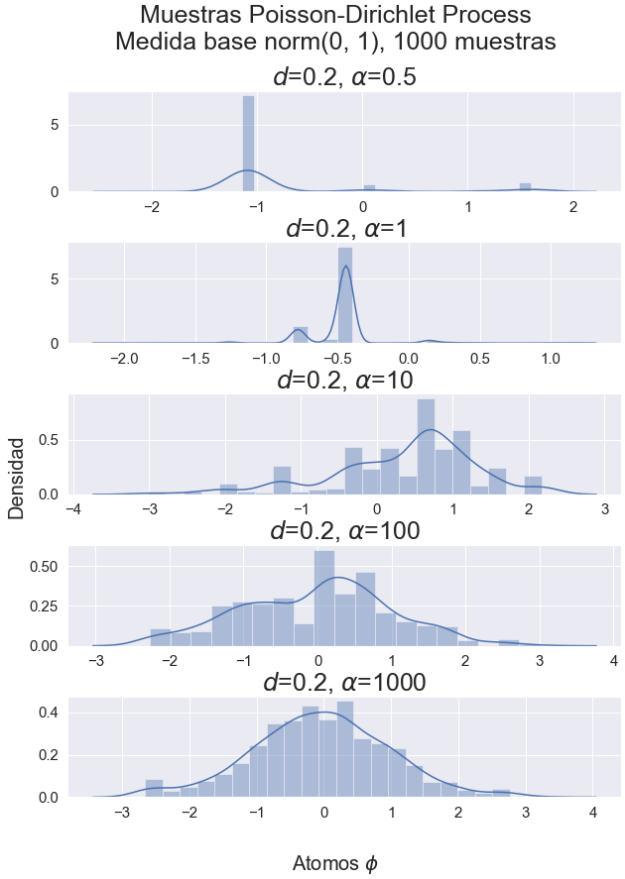
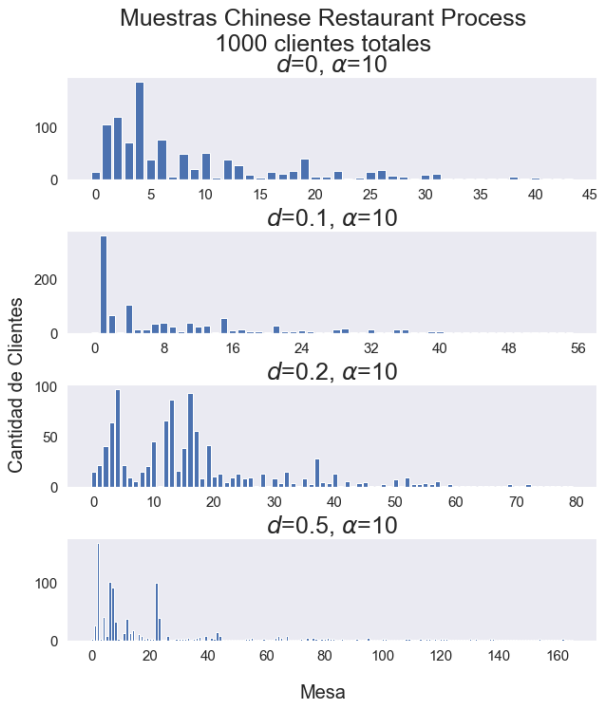


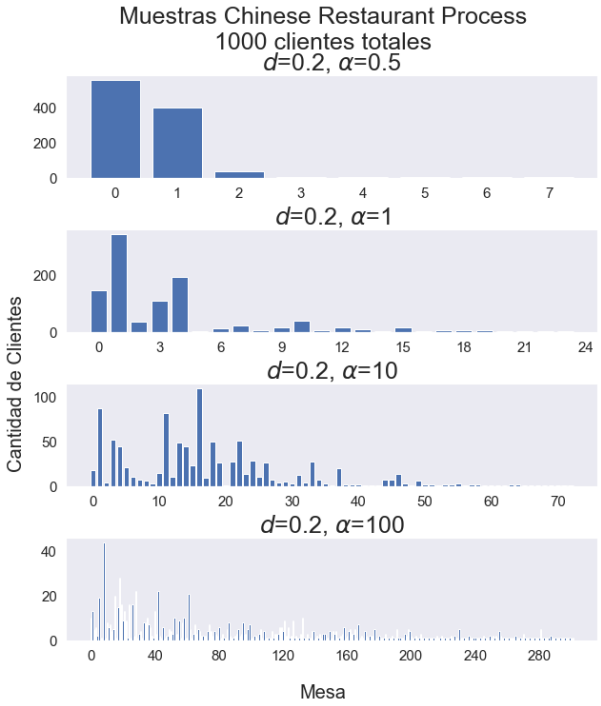
Figura 1.2: Muestras del Proceso de Poisson-Dirichlet, con medida base Normal$(0, 1)$, $d = 0.2$ y distintos valores de α, para 1000 muestras cada uno.

Por último, se deciden generar muestras del Proceso del Restaurante Chino. Para d y α previamente definidos se asume que el primer cliente se sienta en la primera mesa (mesa 0) y el resto de los clientes se generan de acuerdo a la probabilidad dada por la Definición 1.6 del CRP, es decir tienen probabilidad $\frac{n_j - d}{\alpha + N}$ de sentarse en la mesa ya existente j donde hay n_j clientes o probabilidad $\frac{\alpha + kd}{\alpha + N}$ de sentarse en una nueva mesa, hasta llegar a un número deseado de clientes. En la Figura 1.3 se tienen diferentes gráficos para muestras del Proceso del Restaurante Chino. Por un lado, para la Figura 1.3.a se tienen muestras para $\alpha = 10$ y distintos valores de d con 1000 clientes generados en cada uno, de esto podemos observar el comportamiento que tiene el parámetro de descuento: en general las primeras mesas son las que más tienen clientes pero aún así hay una gran cantidad de mesas, sobre todo al aumentar d, ya que para $d = 0$ hay cerca de 40 mesas y para $d = 0.5$ existen aproximadamente 160 mesas, pero con una cantidad muy pequeña de clientes para las mesas que están en la cola de la distribución. Por el otro lado, en la Figura 1.3.b se tienen muestras para $d = 0.2$ y distintos valores de α también para 1000 clientes cada uno. En este caso se tiene un

comportamiento más o menos similar de que las primeras mesas tienen más clientes y que al aumentar α también aumenta la cantidad de mesas, para $\alpha = 0.5$ hay 7 mesas mientras que para $\alpha = 100$ hay cerca de 280. La diferencia es que la cantidad de clientes para las primeras mesas disminuye drásticamente al aumentar α: de 400 a 40 aproximadamente, lo que se ve reflejado también en la cola de la distribución de las mesas con cantidades de clientes más o menos similares pero aún así pequeñas.



(a) Muestras del CRP con $\alpha = 10$ y distintos valores de d, para 1000 clientes cada uno.



(b) Muestras del CRP con $d = 0.2$ y distintos valores de α, para 1000 clientes cada uno.

Figura 1.3: Muestras del Proceso del Restaurante Chino.

1.6. Good-Turing y PDP

El problema de estimar la probabilidad de descubrimiento de nuevas especies tiene diferentes enfoques, donde el método frecuentista más conocido es el modelo de Good-Turing. En esta sección se define formalmente la probabilidad de descubrimiento, el modelo de Good-Turing, además de su estimador suavizado y la relación que tiene este modelo con el Proceso de Poisson-Dirichlet (ver [19], [11], [20], [21] y [4]).

Definición 1.7 *Sea $(X_i)_{i=1}^N$ una muestra de una población de infinitas especies $(X_i^*)_{i\geq 1}$ con proporciones desconocidas $(\pi_i)_{i\geq 1}$. Se tiene la probabilidad $D_{N,c}(l)$ de observar, en la $(N+c+1)$-ésima extracción, una especie con frecuencia $l \geq 0$ (o el número de veces que aparece es l) en la muestra ampliada de tamaño $N+c$, donde la muestra adicional no ha sido observada.*

$$D_{N,c}(l) = \sum_{i\geq 1} \pi_i \mathbb{1}_l(n_{i,N+c}), \tag{1.10}$$

donde $n_{i,N+c}$ es la frecuencia de la especie X_i^* en la muestra ampliada.

Notamos que $D_{N,c}(0)$ representa la proporción de especies aún no vistas, o en otras palabras la probabilidad de descubrir una nueva especie en la extracción $(N+c+1)$. A la probabilidad dada por la ecuación (1.10) se le llama probabilidad de $(c;l)$-descubrimiento, la cual típicamente se estudia para decidir el tamaño de la muestra adicional que se debe tomar. En el caso de este trabajo, se desea estudiar la probabilidad de $(0;l)$-descubrimiento para estimar la probabilidad de descubrir una nueva especie, por lo que usaremos la notación simplificada $D_N(l)$ y probabilidad de l-descubrimiento cuando $c=0$.

Por un lado tenemos el enfoque frecuentista de Good-Turing el cual, considerando una muestra $X_1,\ldots,X_N$ y una hipótesis $\mathcal{H}$ para los pesos π, toma en cuenta la variable aleatoria $M_{l,N}$ que corresponde al número de especies con frecuencia l en la muestra observada de tamaño N, también denominado la frecuencia de la frecuencia l. Luego el estimador de $D_N(l)$ es

$$\check{D}_N(l;\mathcal{H}) = (l+1)\frac{\mathbb{E}_{\mathcal{H}}(M_{l+1,N})}{N},$$

donde $\mathbb{E}_{\mathcal{H}}$ denota la esperanza con respecto a la distribución dada por $\mathcal{H}$. Para prescindir de esta hipótesis, se propone un estimador para N grande dado por

$$\check{D}_N(l;\mathcal{H}) \approx \check{D}_N(l) = (l+1)\frac{m_{l+1,N}}{N}, \tag{1.11}$$

donde $m_{l,N}$ corresponde al valor observado de $M_{l,N}$, es decir, el número de especies con frecuencia l en la muestra observada, el cual se denotará simplemente m_l cuando no haya confusión con el tamaño de la muestra, con esto se tiene que $N=\sum_{l\geq 1} l m_l$. El estimador $\check{D}_N(l)$ dado en la ecuación (1.11) es conocido como el *estimador de Good-Turing*. Notamos que este estimador tiene una característica bastante peculiar ya que depende de m_{l+1} y no de m_l como se hubiese intuido para estimar la probabilidad de l-descubrimiento. Además, para este estimador la probabilidad de observar una nueva especie en la siguiente observación es m_1/N, lo que nos dice que esta probabilidad depende únicamente de la cantidad de especies para las que se observó un solo individuo.

El estimador $\check{D}_N(l)$ se comporta bien solo para l suficientemente pequeño, ya que para l grande se tiene un comportamiento irregular para los valores m_l. Puede ocurrir que se observan $m_l>0$ y $m_{l+1}=0$, que entrega un estimador absurdo de $\check{D}_N(l)=0$. Para superar esto, se decide *suavizar* la secuencia de los m_l en una serie más regular. Luego si m_l' corresponde al valor de m_l bajo alguna regla de suavizado $\mathcal{S}$, entonces el estimador

$$\check{D}_N(l;\mathcal{S}) = (l+1)\frac{m_{l+1}'}{N}$$

es una mejor aproximación que $\check{D}_N(l)$. Para reglas de suavizado se puede considerar que m_l', como función de l, sea aproximadamente parabólica; o considerar el estimador de Good-Turing dado por (1.11) para l pequeños y para el resto una linea recta.

Por otro lado está el enfoque Bayesiano que busca la aleatorización de los pesos π, esto bajo el contexto ya planteado del Proceso de Poisson-Dirichlet. Sea una muestra $X_1,\ldots,X_N$ de un

$PDP(d, \alpha)$, donde se han observado $K = k$ especies distintas con frecuencia de frecuencias $m_1, \ldots, m_N$. Luego, el estimador Bayesiano no paramétrico de $D_{N,c}(0)$ para $c \geq 0$ es

$$\hat{D}_{N,c}(0) = \frac{\alpha + kd}{\alpha + N} \frac{(\alpha + N + d)_c}{(\alpha + N + 1)_c}. \tag{1.12}$$

Tenemos entonces que el estimador de la probabilidad de 0-descubrimiento es $(\alpha+kd)/(\alpha+N)$ que corresponde a lo mismo que se tiene en la ecuación (1.8) cuando aparece una nueva especie para la siguiente observación. Además, el estimador para la probabilidad de l-descubrimiento, para $l = 1, \ldots, N$, está dado por

$$\hat{D}_N(l) = (l - d) \frac{m_l}{\alpha + N}. \tag{1.13}$$

De acuerdo a los estimadores dados por (1.12) y (1.13) el enfoque Bayesiano no paramétrico bajo el contexto del Proceso de Poisson-Dirichlet tiene algunas ventajas notables con respecto al enfoque de Good-Turing: los estimadores son exactos, es decir, no es necesario aproximar para N grande; el estimador $\breve{D}_N(0)$ está en función de m_1 mientras que $\hat{D}_N(0)$ está en función de k; y el estimador $\hat{D}_N(l)$ depende de m_l, no de m_{l+1} como en Good-Turing, así se evita tener que usar reglas de suavizado para prevenir estimaciones absurdas dado el comportamiento irregular de los m_l.

La relación que existe entre los estimadores de Good-Turing y de Poisson-Dirichlet fue planteada por primera vez en [11], donde muestran que el estimador Bayesiano $\hat{D}_N(l)$ es asintóticamente equivalente, cuando el tamaño de la muestra N tiende a infinito, al estimador $\breve{D}_N(l; \mathcal{S})$ de Good-Turing con alguna regla de suavizado. La regla que se utiliza viene de la relación asintótica entre las variables $M_{l,N}$, el número de especies con frecuencia l, y K_N, el número de especies distintas, donde el subíndice se usa para explicitar la dependencia con el tamaño de la muestra N. Específicamente, se tiene que $A_N \overset{c.s.}{\approx} B_N$ cuando $N \to \infty$ significa que $\lim_{N\to\infty} A_N/B_N = 1$ casi seguramente, es decir A_N y B_N son asintóticamente equivalentes c.s. para N tendiendo a infinito. Luego, para $d > 0$ se tiene que

$$M_{l,N} \overset{c.s.}{\approx} \frac{d(1-d)_{l-1}}{l!} K_N,$$

o en otras palabras cuando N se acerca a infinito el número de especies con frecuencia l se convierte en una proporción $d(1-d)_{l-1}/l!$ del número total de especies. El siguiente teorema, incluyendo el caso $d = 0$ donde $M_{l,N}$ cumple una relación diferente, muestra la relación asintótica entre los estimadores $\hat{D}_N(l)$ y $\breve{D}_N(l; \mathcal{S})$.

Teorema 1.1 (ver [11]) *Dada una muestra $\mathbf{X}_N$ de tamaño N de un $PDP(d, \alpha)$ y $K_N = k_N$ especies con frecuencia de frecuencias $m_{1,N}, \ldots, m_{N,N}$. Entonces, para $N \to \infty$ y para $l = 0, \ldots, N$, se tiene que*

(i) para $d \in (0, 1)$ y $\alpha > -d$

$$\hat{D}_N(l) \approx (l + 1) \frac{m_{l+1,N}}{N} \approx (l + 1) \frac{\frac{d(1-d)_l}{(l+1)!} k_N}{N}, \tag{1.14}$$

(ii) para $d = 0$ *y* $\alpha > 0$

$$\hat{D}_N(l) \approx (l+1)\frac{m_{l+1,N}}{N} \approx (l+1)\frac{\frac{\alpha}{l+1}}{N}. \tag{1.15}$$

La primera aproximación en las ecuaciones (1.14) y (1.15) muestran que, para N suficientemente grande, $\hat{D}_N(l)$ es asintóticamente igual al estimador de Good-Turing $\check{D}_N(l)$; mientras que la segunda equivalencia muestran como suavizar los $m_{l,N}$ que aparecen en la primera aproximación.

Para el caso del Proceso de Dirichlet, tenemos que la regla de suavizado en (1.15) no depende del número total de especies k_N en la muestra. Esta característica claramente no es deseable ya que si se quiere inferir el número de especies en una muestra futura, este valor no dependería de la cantidad de especies en la muestra actual. En cambio en el caso del Proceso de Poisson-Dirichlet, la regla de suavizado en (1.14) sí depende de k_N, es decir, corresponde a un mejor estimador ya que utiliza la información que se tiene de la cantidad de especies en la muestra inicial.

Capítulo 2

Entropía Bayesiana

En este capitulo se explica el concepto de entropía, en particular la entropía Bayesiana, además de la función digamma junto a sus propiedades principales. Esto con el fin de exponer la entropía esperada en el caso del Proceso de Poisson-Dirichlet. Todo esto inspirados en [22], [23], [24] y la referencia principal [7].

Específicamente, en la Sección 2.1 se revisa el concepto de entropía junto a su estimador clásico. En la Sección 2.2 se estudia la función digamma y varias de sus propiedades. Después, en la Sección 2.3 se expone la entropía Bayesiana además de un ejemplo en el caso que se tienen dos especies y el estimador de Nemenman-Shafee-Bialek. Por último, en la Sección 2.4 se estudia la entropía Bayesiana en el contexto del Proceso de Poisson-Dirichlet, donde se obtiene el primer resultado de esta tesis, la Proposición 2.5, que corresponde a una cota superior de la entropía esperada posterior para el PDP.

2.1. Estimación de la entropía

La entropía es una forma de medir la "cantidad de información" de una variable aleatoria o de una distribución. Donde la información se obtiene al observar los valores que puede tomar la variable, y la cantidad hace referencia a la "sorpresa" que trae este valor cuando es observado. Para un valor que ocurre con muy baja probabilidad, este trae harta sorpresa al ser observado. En este trabajo, podemos considerar una comunidad biológica con un gran número de especies, que se puede considerar incluso infinito. Cuando se toma una muestra de esta comunidad y se clasifican de acuerdo a las especies que se observaron, puede que algunas especies más raras no hayan sido descubiertas, lo que no quiere decir que no existan. Entonces, se desea estudiar la entropía para entender la diversidad de la comunidad y de cómo varía esta al observar nuevos elementos, cuando se desconoce el número de especies y su abundancia. Para esto, se tiene que estimar de alguna manera la entropía usando los datos observados.

Supongamos que hay M especies en una comunidad, donde M puede ser finito cuando se conoce el número de especies o se puede considerar infinito cuando se desconoce, y sea $(\pi_i)_{i=1}^M$ una distribución discreta desconocida que corresponde a la proporción de cada especie o a la probabilidad de descubrimiento de cada especie, donde $\sum_{i=1}^M \pi_i = 1$. Luego, la entropía está

definida por

$$H(\pi) = -\sum_{i=1}^{M} \pi_i \log(\pi_i). \tag{2.1}$$

Aunque la entropía es conocida como una cantidad teórica, su estimación precisa a partir de los datos es una parte importante en muchas aplicaciones. Consideremos ahora una muestra $\mathbf{X}_N = (X_1, \ldots, X_N)$ tomada de la distribución anterior donde cada elemento es clasificado en alguna especie, con esta muestra podemos estimar el valor de la entropía. Cuando el número de las especies es conocido y finito, la técnica de estimación de la entropía más sencilla y directa de usar es considerar la distribución empírica de π, que se obtiene de las frecuencias n_i de cada especie en la muestra, es decir,

$$\hat{\pi}_i = \frac{n_i}{N}, \qquad i = 1, \ldots, M,$$

con lo que se obtiene el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) de la entropía

$$\hat{H}_{MLE} = -\sum_{i=1}^{M} \frac{n_i}{N} \log\left(\frac{n_i}{N}\right). \tag{2.2}$$

A pesar de ser un estimador simple y bastante intuitivo, este es un estimador sesgado, es decir, la esperanza de $\hat{H}_{MLE}$ no es igual al parámetro que se desea estimar H, sobre todo en el caso $N \ll M$ cuando varias especies aún no han sido observadas. Una propiedad importante que cumple este estimador es que es consistente, o sea mientras más grande es el tamaño de muestra N el estimador se aproxima a H.

En el contexto de la ecología, el verdadero número de especies es generalmente desconocido y algunas especies raras pueden no ser descubiertas al tomar una muestra. Hay diferentes enfoques que se pueden tomar para esto, por ejemplo Chao y Shen (ver [24]) proponen un estimador que utiliza el modelo de Good-Turing para tomar en cuenta la probabilidad de especies no observadas en la muestra.

En general se estima la entropía directamente en vez de la distribución completa, esto pues no siempre se tienen datos suficientes para estimar la distribución, o porque no se conoce realmente el número de especies. Por esto, el enfoque Bayesiano para estimar la entropía es una buena estrategia ya que permite asumir una distribución prior y luego hacer la estimación usando la distribución posterior.

2.2. Función digamma

Para poder continuar con la entropía Bayesiana, se necesita primero introducir la función digamma y sus principales propiedades, ya que será necesario en lo que sigue (ver [25], [26]). Esta función se define como la derivada logarítmica de la función gamma.

$$\psi(x) = \frac{d}{dx} \ln(\Gamma(x)) = \frac{\Gamma(x)'}{\Gamma(x)}, \tag{2.3}$$

donde $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ es la función gamma. La cual cumple la relación $\Gamma(x+1) = x\Gamma(x)$, con esto tenemos que la función digamma cumple la siguiente ecuación de recurrencia

$$\psi(x+1) = \psi(x) + \frac{1}{x}. \tag{2.4}$$

Otras propiedades interesantes de esta función es cuando se trabaja en los reales positivos, donde se tiene que la función digamma ψ es una función creciente. Además, tiene una representación integral explícita

$$\psi(x) = \int_0^\infty \left(\frac{e^{-t}}{t} - \frac{e^{-xt}}{1-e^{-t}} \right) dt. \tag{2.5}$$

La función digamma tiene una cota superior y una cota inferior para $x > 0$, con respecto a unas funciones del logaritmo. Estas cotas permiten obtener un intervalo donde se puede encontrar $\psi(x)$ cuando no se puede calcular directamente la función. Las cuales están dadas por la siguiente relación

$$\ln(x) - \frac{1}{x} \leq \psi(x) \leq \ln(x) - \frac{1}{2x}, \quad x > 0. \tag{2.6}$$

Por último, dada la cota superior se tiene que para x suficientemente grande se puede considerar la siguiente aproximación para la función digamma

$$\psi(x) \approx \ln(x) - \frac{1}{2x}. \tag{2.7}$$

2.3. Entropía Bayesiana

Consideremos nuevamente una muestra $\mathbf{X}_N = (X_1, \ldots, X_N)$ tomada de la distribución discreta desconocida π. Nos centramos en el caso que la cantidad de especies distintas observadas es pequeño en comparación con el número total desconocido (incluso infinito) de posibles especies. Mientras que la verdadera distribución en cualquier base de datos es indiscutiblemente finita, formular un modelo sobre un espacio infinito dimensional permite al enfoque Bayesiano ser flexibles acerca de la verdadera cardinalidad.

El enfoque Bayesiano para estimar la entropía requiere asumir una distribución prior para π y con eso inferir H usando la distribución posterior. Se tiene que el estimador de Bayes de mínimos cuadrados toma la forma de

$$\hat{H} = \mathbb{E}(H|\mathbf{X}_N) = \int H(\pi)p(H|\pi)p(\pi|\mathbf{X}_N)d\pi, \tag{2.8}$$

donde $p(\pi|\mathbf{X}_N)$ es la posterior sobre π bajo alguna prior $p(\pi)$ y verosimilitud discreta $p(\mathbf{X}_N|\pi)$. Si se asume una distribución prior sobre π de algún parámetro a, entonces el valor de interés es la entropía esperada posterior $\mathbb{E}(H|\mathbf{X}_N, a)$.

2.3.1. Entropía Bayesiana para dos especies

Para entender mejor cómo encontrar la entropía Bayesiana, veamos un ejemplo cuando se tienen solo 2 especies. Consideremos la distribución generada por $\pi \sim \text{Beta}(a, b)$ con $a, b > 0$, es decir con función de densidad

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad x \in [0,1],$$

donde $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ es la función beta, con lo cual $\pi \in [0,1]$. Luego el intervalo $[0,1]$ tiene la partición π y $(1-\pi)$ lo que corresponde a la distribución prior, para dos especies. De las propiedades de la distribución beta, sabemos que $(1-\pi) \sim \text{Beta}(b,a)$, para calcular la entropía primero debemos probar la siguiente propiedad.

Proposición 2.1 *Para $\pi \sim$ Beta(a,b) con $a, b > 0$, se tiene que*

$$\mathbb{E}(\pi \ln(\pi)) = \frac{a}{a+b}(\psi(a+1) - \psi(a+b+1)). \tag{2.9}$$

Demostración. La esperanza se toma con respecto a la función de densidad de π,

$$\mathbb{E}(\pi \ln(\pi)) = \int_0^1 x \ln(x) \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} dx = \frac{1}{B(a,b)} \int_0^1 \ln(x) x^a (1-x)^{b-1} dx.$$

Notamos que $\ln(x)x^a = \frac{\partial x^a}{\partial a}$, en efecto $x^a = e^{a\ln(x)}$ y luego se tiene que $\frac{\partial x^a}{\partial a} = e^{a\ln(x)}\ln(x) = x^a \ln(x)$. Así,

$$\begin{aligned}
\mathbb{E}(\pi \ln(\pi)) &= \frac{1}{B(a,b)} \int_0^1 \frac{\partial x^a}{\partial a}(1-x)^{b-1} dx \\
&= \frac{1}{B(a,b)} \frac{\partial}{\partial a} \left(\int_0^1 x^a (1-x)^{b-1} dx \right),
\end{aligned}$$

usando la representación integral de la función beta se tiene que

$$\mathbb{E}(\pi \ln(\pi)) = \frac{1}{B(a,b)} \frac{\partial}{\partial a}(B(a+1,b)).$$

Desarrollando la derivada

$$\begin{aligned}
\frac{\partial}{\partial a}(B(a+1,b)) &= \frac{\partial}{\partial a}\left(\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \right) \\
&= \frac{\Gamma'(a+1)\Gamma(b)\Gamma(a+1+b) - \Gamma(a+1)\Gamma(b)\Gamma'(a+1+b)}{\Gamma(a+1+b)^2} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \left(\frac{\Gamma'(a+1)}{\Gamma(a+1)} - \frac{\Gamma'(a+1+b)}{\Gamma(a+1+b)} \right) \\
&= B(a+1,b)(\psi(a+1) - \psi(a+1+b)).
\end{aligned}$$

Luego en la entropía basta expandir las funciones beta y utilizar la ecuación de recurrencia

de la función gamma

$$\begin{aligned}\mathbb{E}(\pi \ln(\pi)) &= \frac{1}{B(a,b)} B(a+1,b)(\psi(a+1) - \psi(a+1+b)) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} (\psi(a+1) - \psi(a+1+b)) \\ &= \frac{\Gamma(a+b)a\Gamma(a)}{\Gamma(a)(a+b)\Gamma(a+b)} (\psi(a+1) - \psi(a+1+b)) \\ &= \frac{a}{a+b} (\psi(a+1) - \psi(a+1+b)).\end{aligned}$$

Con lo que se obtiene lo deseado. □

La generalización de la distribución Beta es la llamada distribución de Dirichlet, que es de orden $k \geq 2$ y parámetros $a_1, \ldots, a_k \geq 0$ y se denota Dirichlet$(a_1, \ldots, a_k)$, con función de densidad

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\sum_{i=1}^k a_i)}{\prod_{i=1}^k \Gamma(a_i)} \prod_{i=1}^k x_i^{a_i - 1},$$

donde $(x_i)_{i=1}^k$ pertenece al $k-1$ símplex estándar, es decir, cumple que $\sum_{i=1}^k x_i = 1$ y $x_1, \ldots, x_k \geq 0$. Un caso particular de la distribución de Dirichlet es cuando todos los parámetros tienen el mismo valor, a esta distribución se le llama distribución de Dirichlet simétrica, se denota simplemente Dirichlet(a) y tiene densidad

$$f(x_1, \ldots, x_k) = \frac{\Gamma(ak)}{\Gamma(a)^k} \prod_{i=1}^k x_i^{a-1}.$$

Volviendo a la partición del intervalo $[0,1]$, si consideramos $\vec{\pi} = (\pi, 1-\pi)$ entonces se tiene que $\vec{\pi} \sim$ Dirichlet(a,b). Así la entropía en este caso es con respecto a esta distribución y está dada por

$$H = H(\vec{\pi}) = -\pi \ln(\pi) - (1-\pi)\ln(1-\pi),$$

luego la entropía esperada prior es

$$\begin{aligned}\mathbb{E}(H|a,b) &= -\mathbb{E}(\pi \ln(\pi)) - \mathbb{E}((1-\pi)\ln(1-\pi)) \\ &= -\frac{a}{a+b}(\psi(a+1) - \psi(a+b+1)) - \frac{b}{a+b}(\psi(b+1) - \psi(a+b+1)),\end{aligned}$$

donde se usó la ecuación (2.9) para π y para $(1-\pi)$. Por lo que la entropía esperada prior, considerando una distribución Dirichlet(a,b), corresponde a

$$\mathbb{E}(H|a,b) = \psi(a+b+1) - \frac{a}{a+b}\psi(a+1) - \frac{b}{a+b}\psi(b+1). \tag{2.10}$$

La pregunta ahora es qué pasa con la entropía al observar datos de esta distribución. La primera observación X_1 pertenece a la especie 1 con probabilidad π o a la especie 2 con probabilidad $1-\pi$, y esto igual para el resto de las N observaciones que denotamos $\mathbf{X}_N = (X_1, \ldots, X_N)$, es decir, siguen una distribución categórica.

La distribución categórica es una distribución discreta que describe los posibles resultados de una variable aleatoria que puede tomar una de $k > 0$ categorías posibles, con la probabilidad de cada categoría especificada por separado, es decir, los parámetros son las probabilidades $p_1, \ldots, p_k$ con $\sum_{i=1}^{k} p_i = 1$. La función de densidad (discreta) está dada por $p(x = i) = p_i$ y se denota Categórica$(p_1, \ldots, p_k)$. Luego en nuestro caso, cada observación X_j con $j = 1, \ldots, N$, puede pertenecer a la especie $i = 1$ con probabilidad $p_1 = \pi$, o a la especie $i = 2$ con probabilidad $p_2 = 1 - \pi$. Lo cual denotamos como

$$\mathbf{X}_N = (X_1, \ldots, X_N)|\vec{\pi} \sim \text{Categórica}(\pi, 1 - \pi).$$

La principal razón para usar la distribución de Dirichlet para este tipo de problemas, es que esta distribución es la *conjugada prior* de la distribución categórica. Que una distribución sea la conjugada prior, significa que la distribución posterior $p(\theta|x)$ es parte de la misma familia de distribuciones de probabilidad que la distribución prior $p(\theta)$, donde θ es algún parámetro y x es la muestra tomada a partir de la distribución prior.

En este caso, esto se traduce en que la distribución después de haber visto las observaciones $\mathbf{X}_N$ también es una distribución Dirichlet. Formalmente tenemos que

$$\begin{aligned} \vec{\pi} &\sim \text{Dirichlet}(a, b) \\ \mathbf{X}_N|\vec{\pi} &\sim \text{Categórica}(\pi, 1 - \pi), \end{aligned}$$

entonces la distribución posterior $\vec{\pi}_{post}$ cumple que

$$\vec{\pi}_{post} = \vec{\pi}|\mathbf{X}_N, a, b \sim \text{Dirichlet}(a + n_1, b + n_2), \tag{2.11}$$

donde n_j es la cantidad de observaciones que están en la especie $j = 1, 2$ y así $N = n_1 + n_2$.

Notamos de la ecuación (2.11) que las frecuencias tienen un rol importante para la distribución posterior, ya que mientras más grandes (o mientras más crece N) se tiene que los valores asumidos inicialmente, a y b, tienen menos impacto en las probabilidades posteriores de pertenecer a cada especie.

La entropía posterior en este caso corresponde a $H(\vec{\pi}_{post})$, la ecuación (2.10) es la esperanza de la entropía para una Dirichlet(a, b) y en este caso la queremos para una Dirichlet$(a+n_1, b+n_2)$ por lo que basta reemplazar con estos nuevos parámetros. Luego, la entropía esperada posterior en este caso está dada por

$$\mathbb{E}(H|\mathbf{X}_N, a, b) = \psi(a+b+N+1) - \frac{a+n_1}{a+b+N}\psi(a+n_1+1) - \frac{b+n_2}{a+b+N}\psi(b+n_2+1). \tag{2.12}$$

Tenemos de la ecuación (2.12) que en las fracciones que aparecen ocurre lo mismo que se mencionó antes, que los valores de n_1, n_2 y N tienen más peso, a medida que se observan más datos, que los parámetros iniciales a, b. Además, con este ejemplo notamos que la función digamma juega un rol importante en la entropía Bayesiana prior y posterior, incluso cuando se tienen solo dos especies.

2.3.2. Estimador de Nemenman-Shafee-Bialek

Para el caso de un número finito de dos o más especies, existe el estimador de Nemenman-Shafee-Bialek (NSB) [27]. Este estimador considera una distribución de Dirichlet simétrica, donde para las k especies se toma el mismo parámetro $a > 0$ (en vez de usar $a_1, \ldots, a_k$), como fue mencionado anteriormente. Este estimador utiliza un prior como mezcla de distribuciones Dirichlet simétricas

$$p(\pi) = \int p_{\mathrm{Dir}}(\pi|a)p(a)da,$$

donde $p_{Dir}(\pi|a)$ denota una Dirichlet simétrica de parámetro a que es la prior sobre π. Luego el estimador Bayesiano, retomando la ecuación (2.8), bajo el prior NSB corresponde a

$$\begin{aligned}\hat{H}_{NSB} = \mathbb{E}(H|\mathbf{X}_N) &= \int\int H(\pi)p(\pi|\mathbf{X}_N, a)p(a|\mathbf{X}_N)d\pi da \\ &= \int \mathbb{E}(H|\mathbf{X}_N, a)\frac{p(\mathbf{X}_N|a)p(a)}{p(\mathbf{X}_N)}da.\end{aligned}$$

El beneficio de este estimador es que es rápido de computar, sobre todo porque la integración es para un solo parámetro a. Además, la entropía esperada posterior $\mathbb{E}(H|\mathbf{X}_N, a)$ tiene una forma bastante esperable con respecto a lo que se obtuvo en el caso de dos especies con dos parámetros en (2.12),

$$\mathbb{E}(H|\mathbf{X}_N, a) = \psi(ka + N + 1) - \sum_{i=1}^{k} \frac{a + n_i}{ka + N}\psi(a + n_i + 1).$$

2.4. Entropía para el Proceso de Poisson-Dirichlet

En esta sección tomaremos el Proceso de Poisson-Dirichlet como prior pues permite trabajar sobre distribuciones discretas desconocidas que sean infinito numerable. Recordemos que este proceso tiene la forma $\sum_{i\geq 1} \pi_i \delta_{\phi_i}$, donde en adelante nos vamos a referir, abusando de la terminología, a los pesos π y al proceso con el mismo símbolo $PDP(d, \alpha)$. Además, como estamos trabajando con una medida base no atómica la podemos ignorar, lo que permite calcular la entropía del proceso como la entropía de los pesos, es decir $H(\pi)$ como en la definición inicial (2.1).

La principal ventaja de usar el Proceso de Poisson-Dirichlet es que la distribución posterior sobre la entropía tiene una forma explícita para los momentos. Además, sabemos que las muestras de un PDP tienen colas con ley potencia, que es una característica muy útil ya que es común observarla en datos biológicos y ecológicos. Usar este proceso no siempre trae beneficios, ya que al tener los parámetros fijos se impone una distribución prior estrecha sobre la entropía, es decir, la prior determina fuertemente la estimación de la entropía, dando lugar a un posible sesgo e intervalos de credibilidad posterior también estrechos cuando se tiene un conjunto de datos pequeño.

Para encontrar la esperanza del vector infinito π, se usa la propiedad del PDP de ser invariante bajo un muestreo de pesos ordenados, que permite convertir la esperanza sobre π (infinito) en una esperanza de una dimensión con respecto a la distribución del primer peso de mayor tamaño, dado por la ecuación (1.6). Gracias a esto se tiene la entropía esperada

prior en el caso del Proceso de Poisson-Dirichlet.

Proposición 2.2 (ver [7]) *Considerando una distribución prior $\pi \sim PDP(d, \alpha)$, entonces el valor esperado de $H(\pi)$ está dado por*

$$\mathbb{E}(H|d, \alpha) = \psi(\alpha + 1) - \psi(1 - d). \tag{2.13}$$

Notamos que la ecuación (2.13) es análoga a la ecuación (2.10) en el caso de dos especies para la distribución de Dirichlet. Al igual que en el caso de dos especies, nos interesa encontrar la entropía esperada después de haber visto una muestra de esta distribución. Para esto necesitamos entender el comportamiento de la distribución posterior.

Consideremos una muestra $\mathbf{X}_N = (X_1, \dots, X_N)$ de la distribución prior $\pi \sim PDP(d, \alpha)$, con k especies distintas y frecuencias $n_1, \dots, n_k$. El Proceso de Poisson-Dirichlet cumple una propiedad similar a la usada en el caso de dos especies, razón adicional para usar este proceso de prior, que la distribución de Dirichlet es la conjugada prior de la distribución categórica. Luego en este caso, la siguiente proposición nos dice que la distribución posterior toma la forma de una mezcla de una distribución Dirichlet, para los elementos observados, y un Proceso de Poisson-Dirichlet, para los elementos no observados.

Proposición 2.3 (ver [7]) *La distribución posterior es $\pi_{post} = \pi|\mathbf{X}_N, d, \alpha = (p_1, \dots, p_k, p_*\pi')$ donde*

$$\begin{aligned}(p_1, \dots, p_k, p_*) &\sim \text{Dirichlet}(n_1 - d, \dots, n_k - d, \alpha + kd)\\ \pi' = (\pi'_1, \pi'_2, \dots) &\sim PDP(d, \alpha + kd).\end{aligned}$$

De esta proposición se tiene que la probabilidad de pertenecer a una especie j ya observada es p_j con $j = 1, \dots, k$; y la de pertenecer a una nueva especie, o la probabilidad de que aparezca una nueva especie, es $p_* = 1 - \sum_{j=1}^{k} p_j$, donde estas probabilidades dependen de las frecuencias n_j y de k respectivamente. Además, dentro de la probabilidad de que aparezca una nueva especie, ser de alguna especie específica l (aún desconocida) está determinado por el peso π'_l. Que la distribución posterior sea una mezcla de una distribución Dirichlet y un PDP tiene relación con la mezcla de distribuciones que aparecen en la probabilidad condicional (1.8) cuando se tiene la medida base de forma explícita.

Luego, la entropía esperada posterior en el caso del Proceso de Poisson-Dirichlet, es el valor esperado de $H(\pi_{post})$, donde la distribución posterior π_{post} está dada por la Proposición 2.3. Este resultado, enunciado a continuación, corresponde a la propiedad principal del PDP que se utiliza para desarrollar lo que sigue de la tesis.

Proposición 2.4 (ver [7]) *La entropía esperada posterior corresponde a*

$$\mathbb{E}(H|\mathbf{X}_N, d, \alpha) = \psi(\alpha+N+1) - \frac{\alpha + kd}{\alpha + N}\psi(1-d) - \frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right). \tag{2.14}$$

De este valor de la entropía esperada posterior notamos que se tiene la siguiente cota superior, que es un primer resultado de este trabajo.

Proposición 2.5 *Para una muestra $\mathbf{X}_N$ de un PDP(d, α), con k especies distintas, se tiene*

$$\mathbb{E}(H|\mathbf{X}_N, d, \alpha) \leq \psi(\alpha + N + 1) - \frac{\alpha + k}{\alpha + N}\psi(1 - d) - \frac{1}{(\alpha + N)(1 - d)}. \tag{2.15}$$

Demostración. Como $n_i \geq 1$ para todo $i = 1, \ldots, k$, entonces $(1 - d)\psi(1 - d + 1) \leq (n_i - d)\psi(n_i - d + 1)$ y así

$$-\frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right) \leq -\frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(1 - d)\psi(1 - d + 1)\right).$$

Luego,

$$\begin{aligned}
\mathbb{E}(H|\mathbf{X}_N, d, \alpha) &\leq \psi(\alpha + N + 1) - \frac{\alpha + kd}{\alpha + N}\psi(1 - d) - \frac{1}{\alpha + N}(k - kd)\psi(1 - d + 1) \\
&= \psi(\alpha + N + 1) - \frac{\alpha + kd}{\alpha + N}\psi(1 - d) - \frac{1}{\alpha + N}(k - kd)\left(\psi(1 - d) + \frac{1}{1 - d}\right) \\
&= \psi(\alpha + N + 1) - \frac{\alpha + k}{\alpha + N}\psi(1 - d) - \frac{1}{(\alpha + N)(1 - d)}.
\end{aligned}$$

□

Más aún, la igualdad en (2.15) se alcanza si y solo si las N muestras son todas de distintas especies, es decir, $k = N$ y $n_i = 1$ para todo $i = 1, \ldots, k$. Esto corresponde a la entropía esperada posterior calculada en el lado derecho de la última ecuación en la demostración, donde se usa $n_i = 1$ para la comparación. Por lo que basta reemplazar k por N para obtener el valor en este caso.

A pesar de que (2.14) es un estimador Bayesiano de la entropía, es útil asociarle cierta propiedad frecuentista. Se sabe que el estimador MLE de la entropía (2.2) es consistente para cualquier distribución (finita o infinita numerable), así pues se tiene que $\mathbb{E}(H|\mathbf{X}_N, d, \alpha)$ es consistente ya que converge al estimador MLE, bajo cierta condición como muestra el teorema a continuación. La notación $A_N \xrightarrow{P} c$ significa que la sucesión de variables aleatorias A_N converge a alguna constante c en probabilidad.

Teorema 2.1 (ver [7]) *Para una muestra $\mathbf{X}_N$ tomada de una distribución discreta fija $\pi \sim PDP(d, \alpha)$, con K_N especies distintas tal que $\frac{K_N}{N} \xrightarrow{P} 0$, entonces*

$$|\mathbb{E}(H|\mathbf{X}_N, d, \alpha) - \mathbb{E}(H_{MLE}|\mathbf{X}_N)| \to 0.$$

Notamos que la hipótesis sobre la proporción de la cantidad de especies y el tamaño de la muestra parece ser completamente natural, ya que ir observando nuevas muestras es más rápido que el aumento en la cantidad de especies en la muestra, a pesar de que se tenga que K_N tiende a infinito casi seguramente. Por último, lo que dice este teorema es que la entropía esperada posterior en el caso del Proceso de Poisson-Dirichlet, para cada valor de d y α, tiene el mismo comportamiento que $\hat{H}_{MLE}$, en particular la consistencia.

Capítulo 3

Tiempos de Aparición

En este capítulo nos alejamos de la entropía para estudiar los tiempos de aparición de una muestra dada por el Proceso de Poisson-Dirichlet, en particular se estudian los tiempos entre apariciones. En la Sección 3.1 se definen las variables necesarias para el tiempo entre apariciones. La Sección 3.2 se podría considerar como un paréntesis en los tiempos entre apariciones para introducir la distribución de Waring y sus principales propiedades, que serán necesarias en lo que sigue, donde se muestra el desarrollo de las demostraciones pues estas no están en la bibliografía revisada para el Proceso de Poisson-Dirichlet de dos parámetros. En la Sección 3.3 se expone el tiempo esperado entre apariciones en la Proposición 3.5, que es un resultado de esta tesis, el cual corresponde a la generalización del caso de un solo parámetro. Todo esto tomando lo expuesto en [8], [28] y [9] para los tiempos entre apariciones. Para la distribución de Waring se considera también [29], [30], [31] y [32].

3.1. Tiempos de Aparición

Para esta sección consideremos nuevamente una distribución $\pi \sim PDP(d, \alpha)$ y una muestra $\mathbf{X}_N = (X_1, \ldots, X_N)$ de esta distribución. En los capítulos anteriores la muestra que se usaba no consideraba el orden en que se observan los elementos. En el contexto de los tiempos entre apariciones, que se pueden entender como el intervalo o tiempo transcurrido entre nuevas especies, se hace necesario usar explícitamente el orden en que aparecen estas observaciones.

Para poder entender los tiempos entre apariciones de especies necesitamos definir algunas variables aleatorias. Primero, consideremos la secuencia $(B_1, \ldots, B_N)$ dada por

$$B_1 = 1, \qquad B_j = \prod_{i=1}^{j-1} \mathbb{1}_{\{X_j \neq X_i\}}, \quad j = 2, \ldots, N.$$

Es decir, $B_j = 1$ cuando X_j es una nueva especie (distinta a todos los X_i anteriores), y 0 si no lo es. Con esto se tiene una secuencia binaria que permite destacar los instantes en los cuales aparecen nuevas especies en la muestra $\mathbf{X}_N$. Además, la probabilidad condicional de B_j dado valores anteriores de la secuencia viene dada, para $j = 1, \ldots, N$, por

$$\mathbb{P}(B_{j+1}=1 \mid B_1=b_1,\ldots,B_j=b_j)=\frac{\alpha+d\sum_{i=1}^{j}b_i}{\alpha+j}$$
$$\mathbb{P}(B_{j+1}=0 \mid B_1=b_1,\ldots,B_j=b_j)=\frac{j-d\sum_{i=1}^{j}b_i}{\alpha+j},$$

donde $b_i \in \{0,1\}$ para todo $i=1,\ldots,j$, que es análoga a la probabilidad condicional que se tiene para las muestras, solo que para la secuencia binaria. El número total de especies distintas en la muestra, K_N, también se puede definir mediante esta secuencia, contando cada vez que hay una nueva especie.

$$K_N=\sum_{j=1}^{N}B_j.$$

Si consideramos que hay $K_N=k$ especies en la muestra, la variable principal es el *tiempo de aparición*, que es el momento cuando aparece la j-ésima especie por primera vez

$$V_1=1, \qquad V_j=\min\left\{\ell:\sum_{i=1}^{\ell}B_i=j\right\}, \quad j=2,\ldots,k.$$

Con esta variable se tiene una secuencia más específica de los momentos en que aparece cada especie en la muestra. Además, X_{V_j} corresponde al primer elemento observado de la especie j, el cual se puede tomar como el representante de la especie.

Así, se puede definir el *tiempo entre apariciones* (o tiempo de espera) de una nueva especie como el número de observaciones entre especies, o el tiempo transcurrido después de observar una especie hasta observar una nueva, el cual está definido por

$$T_j=V_{j+1}-V_j, \quad 1\leq j<k \quad \text{y} \quad T_1=N \text{ si } k=1.$$

Si se observaron las especies $(X_i^*)_{i=1}^k$ entonces T_1 es el número de observaciones antes de X_2^* y cuando $k>1$, T_j es el número de observaciones entre X_j^* y X_{j+1}^*, incluido el primero. En el caso $k>1$ se tiene que $T_j=t_j$ para $j=1,\ldots,k-1$ es equivalente a $X_{t_1+\cdots+t_{j-1}+1}\in X_j^*$ para $j=2,\ldots,k$; es decir, los tiempos entre apariciones también permiten identificar la primera observación que pertenece a cada especie, además se cumple que $t_1+\cdots+t_{k-1}<N$. Se tiene por convención la definición del tiempo de próxima aparición (o tiempo de aún-espera) como

$$T_k=N-V_k+1 \text{ cuando } k>1, \text{ y } T_1=N \text{ si no.}$$

Con estas variables definidas se tiene un teorema importante demostrado en [8] que corresponde a la distribución conjunta del número total de especies en la muestra y los tiempos entre apariciones.

Teorema 3.1 (ver [8]) *Sean $k,t_1,\ldots,t_{k-1}$ valores en $\mathbb{N}$. Se define $\bar{t}_r=\sum_{j=1}^{r}t_j$ y $\bar{t}_0=0$, luego la distribución conjunta de K_N y $(T_j)_{j=1}^{k-1}$ es*

$$\mathbb{P}(K_N=k,T_1=t_1,\ldots,T_{k-1}=t_{k-1})=\frac{(\alpha|d)_k}{(\alpha)_N}\prod_{i=1}^{k}(1-id+\bar{t}_{i-1})_{t_i-1},$$

donde $t_k = N - \bar{t}_{k-1} > 0$.

Para poder estudiar la distribución conjunta solo de los tiempos entre apariciones y así la distribución condicional dado los tiempos anteriores, para luego entender el tiempo esperado, es necesario introducir la distribución de Waring.

3.2. Distribución de Waring

En esta sección se define la distribución de Waring que corresponde a la distribución de una variable aleatoria discreta con valores en $\{0, 1, \dots\}$, también se define la versión acotada de esta distribución. Junto a esto se enuncian las esperanzas de estas variables, además se desarrollan las demostraciones de estas esperanzas pues en la referencia principal [9] solo se encontraban enunciadas sin demostración, excepto por una propiedad como indicación para dirigir el cálculo.

En [29] se encuentra la esperanza de la distribución de Waring generalizada, que se obtiene mediante la función generadora de momentos, a pesar de esto se decide desarrollar la demostración usando la indicación. Para el caso de la distribución de Waring acotada, no se encuentra la demostración en la bibliografía revisada y de hecho nos parece que es más complicado que solo aplicar la indicación expuesta en [9], lo que es una razón adicional para hacer el desarrollo en este trabajo.

Partimos entonces con la definición de la distribución de Waring.

Definición 3.1 *Sean* $b > a > 0$ *valores reales, se dice que* W *distribuye según una distribución de Waring de parámetros* b *y* a*, denotado* $W \sim$ *Waring*(b, a) *si tiene la siguiente densidad discreta, para* $x \in \{0, 1, \dots\}$

$$\begin{aligned}\mathbb{P}(W = x) &= (b-a)\frac{(a)_x}{(b)_{x+1}} \\ &= (b-a)\frac{\Gamma(a+x)\Gamma(b)}{\Gamma(a)\Gamma(b+x+1)}.\end{aligned}$$

Una propiedad importante de esta distribución, es que se puede escribir como una distribución beta geométrica $W \sim$ Geométrica(Beta$(a, b-a)$). Donde $G \sim$ Geométrica(p) está dada por la densidad $\mathbb{P}(G = x) = p^x(1-p),\ x \geq 0$ (ver [9]), es decir, G distribuye como Geom$(1-p)-1$, con Geom(p) la variable aleatoria geométrica usual. En otras palabras, la distribución de Waring es el número de éxitos necesarios hasta obtener 1 fracaso en una secuencia de lanzamientos independientes Bernoulli con probabilidad de éxito p, donde p a su vez también es una variable aleatoria, según una distribución beta:

$$\begin{aligned}p &\sim \text{Beta}(a, b-a) \\ W|p &\sim \text{Geométrica}(p).\end{aligned}$$

Esta propiedad se puede entender como un caso particular de la distribución beta binomial negativa (BNB por sus siglas en inglés, ver [30]), donde $Z \sim \text{BNB}(r, \alpha, \beta)$ si satisface $Z|p \sim$ Binomial-Negativa(r, p) y $p \sim$ Beta(α, β); la cual se utiliza para conteo cuando hay colas

pesadas. Así, en este caso se tiene que la geométrica definida anteriormente corresponde a la binomial negativa para $r = 1$ fracaso. De este modo, el cálculo de la esperanza de la distribución de Waring es menos complicado. El valor esperado y la demostración se detallan en la siguiente proposición.

Proposición 3.1 (ver [9]) *Para $W \sim$ Waring(b, a) se tiene que*

$$\mathbb{E}(W) = \begin{cases} \frac{a}{b-a-1} & b-a>1 \\ \infty & \textit{si no.} \end{cases}$$

Demostración. Por lo mencionado anteriormente se tiene que $\mathbb{E}(W) = \mathbb{E}(\mathbb{E}(W|p))$, entonces consideremos $G \sim$ Geométrica(p) y calculemos $\mathbb{E}(G)$ para luego tomar la esperanza de esta y aleatorizar p con respecto a una distribución beta.

$$\begin{aligned} \mathbb{E}(G) &= \sum_{x\geq 0} xp^x(1-p) = p(1-p)\sum_{x\geq 0} xp^{x-1} = p(1-p)\sum_{x\geq 0}\frac{d}{dp}(p^x) \\ &= p(1-p)\frac{d}{dp}\left(\sum_{x\geq 0} p^x\right) = p(1-p)\frac{d}{dp}\left(\frac{1}{1-p}\right) = p(1-p)\frac{1}{(1-p)^2} \\ &= \frac{p}{1-p}. \end{aligned}$$

Con esto se tiene que $\mathbb{E}(W) = \mathbb{E}(p/(1-p))$ donde $p \sim \text{Beta}(a, b-a)$. Así,

$$\begin{aligned} \mathbb{E}(W) = \mathbb{E}\left(\frac{p}{1-p}\right) &= \int_0^1 \frac{x}{1-x}\frac{1}{B(a,b-a)}x^{a-1}(1-x)^{b-a-1}dx \\ &= \frac{1}{B(a,b-a)}\int_0^1 x^{a+1-1}(1-x)^{b-a-1-1}dx. \end{aligned}$$

Esta integral es finita solo cuando $a+1>0$ y $b-a-1>0$, como la primera desigualdad siempre se tiene, el cálculo se continua para $b-a>1$. Además, esta integral corresponde a una función beta, entonces

$$\begin{aligned} \mathbb{E}(W) &= \frac{1}{B(a,b-a)}B(a+1,b-a-1) \\ &= \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\frac{\Gamma(a+1)\Gamma(b-a-1)}{\Gamma(b)} \\ &= \frac{a\Gamma(a)\Gamma(b-a-1)}{\Gamma(a)(b-a-1)\Gamma(b-a-1)} \\ &= \frac{a}{b-a-1}. \end{aligned}$$

Por lo tanto se concluye el resultado deseado. □

Como se desea trabajar con los tiempos entre apariciones, se hace necesario modificar esta distribución, ya que los tiempos siempre estarán acotados por el tamaño de la muestra. Si se reagrupan los eventos $W \geq n$ para algún n, se obtiene la distribución de Waring acotada, la cual denotaremos $\widetilde{W}^n$ para explicitar el valor n hasta donde está acotada.

Definición 3.2 *Sean $b > a > 0$ valores reales y $n \in \mathbb{N}$, se dice que $\widetilde{W}^n$ distribuye según una distribución de Waring acotada de parámetros n, b y a, denotada $\widetilde{W}^n \sim A\text{-}Waring(n, b, a)$, si tiene la siguiente densidad discreta*

$$\mathbb{P}(\widetilde{W}^n = x) = \begin{cases} (b-a)\frac{(a)_x}{(b)_{x+1}} & si\ x \in \{0, 1, \ldots, n-1\} \\ \frac{(a)_x}{(b)_x} & si\ x = n. \end{cases}$$

Esta nueva variable aleatoria cumple una propiedad análoga a la anterior de ser una distribución beta geométrica, solo que para una distribución geométrica acotada, que denotaremos $\widetilde{G}^n$, con densidad $\mathbb{P}(\widetilde{G}^n = x) = p^x(1-p)$ para $x \in \{0, 1, \ldots, n-1\}$ y $\mathbb{P}(\widetilde{G}^n = x) = p^x$ si $x = n$ (ver [9]). Luego la esperanza se puede calcular de forma similar a la anterior para uno de los caso, y está dada por:

Proposición 3.2 *Para $\widetilde{W}^n \sim A\text{-}Waring(n, b, a)$ se tiene que, donde ψ es la función digamma (2.3),*

$$\mathbb{E}(\widetilde{W}^n) = \begin{cases} \frac{a}{b-a-1}\left(1 - \frac{(a+1)_n}{(b)_n}\right) & b - a \neq 1 \\ (b-1)[\psi(b+n) - \psi(b)] & b = a+1. \end{cases} \tag{3.1}$$

Demostración. Veamos primero la esperanza de $\widetilde{G}^n$, una geométrica acotada de parámetros n y p, cuyo cálculo es similar al caso no acotado.

$$\begin{aligned} \mathbb{E}(\widetilde{G}^n) &= \sum_{x=0}^{n-1} xp^x(1-p) + np^n = p(1-p)\frac{d}{dp}\left(\sum_{x=0}^{n-1} p^x\right) + np^n \\ &= p(1-p)\frac{d}{dp}\left(\frac{1-p^n}{1-p}\right) + np^n = p(1-p)\frac{-np^{n-1}(1-p) + (1-p^n)}{(1-p)^2} + np^n \\ &= p\frac{-np^{n-1} + np^n + 1 - p^n}{1-p} + \frac{np^n(1-p)}{1-p} = \frac{-np^n + np^{n+1} + p - p^{n+1} + np^n - np^{n+1}}{1-p} \\ &= \frac{p}{1-p} - \frac{p^{n+1}}{1-p}. \end{aligned}$$

Así, la esperanza de $\widetilde{W}^n$ viene dada por $\mathbb{E}(\widetilde{W}^n) = \mathbb{E}\left(\frac{p}{1-p} - \frac{p^{n+1}}{1-p}\right)$ donde $p \sim \text{Beta}(a, b-a)$. Veamos primero el caso $b - a > 1$, donde notamos que aparece la esperanza de una Waring W que ya conocemos $\mathbb{E}(p/(1-p)) = \mathbb{E}(W) = \frac{a}{b-a-1}$, el cálculo del otro término sigue de forma similar

$$\begin{aligned} \mathbb{E}\left(\frac{p^{n+1}}{1-p}\right) &= \frac{1}{B(a, b-a)}\int_0^1 \frac{x^{n+1}}{1-x}x^{a-1}(1-x)^{b-a-1}dx \\ &= \frac{1}{B(a, b-a)}\int_0^1 x^{a+n+1-1}(1-x)^{b-a-1-1}dx = \frac{1}{B(a, b-a)}B(a+n+1, b-a-1) \\ &= \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\frac{\Gamma(a+n+1)\Gamma(b-a-1)}{\Gamma(b+n)} \\ &= \frac{a\Gamma(b)}{\Gamma(a+1)(b-a-1)\Gamma(b-a-1)}\frac{\Gamma(a+n+1)\Gamma(b-a-1)}{\Gamma(b+n)} \\ &= \frac{a}{b-a-1}\frac{(a+1)_n}{(b)_n}. \end{aligned}$$

Luego la esperanza de $\widetilde{W}^n$ en el caso $b-a>1$ es

$$\mathbb{E}(\widetilde{W}^n)=\mathbb{E}\left(\frac{p}{1-p}\right)-\mathbb{E}\left(\frac{p^{n+1}}{1-p}\right)=\frac{a}{b-a-1}\left(1-\frac{(a+1)_n}{(b)_n}\right).$$

A diferencia de la distribución de Waring (no acotada) se tiene que la esperanza sí existe para el caso $b-a\leq 1$. Veamos el caso borde $b-a=1$, en efecto se tiene

$$\mathbb{E}(\widetilde{W}^n)=\mathbb{E}\left(\frac{p}{1-p}-\frac{p^{n+1}}{1-p}\right)=\frac{1}{B(a,b-a)}\int_0^1\left(\frac{x}{1-x}-\frac{x^{n+1}}{1-x}\right)x^{a-1}(1-x)^{b-a-1}dx,$$

reemplazando con $b=a+1$ se tiene que

$$\mathbb{E}(\widetilde{W}^n)=\frac{1}{B(b-1,1)}\int_0^1\left(\frac{x}{1-x}-\frac{x^{n+1}}{1-x}\right)x^{b-2}dx=(b-1)\left(\int_0^1\frac{x^{b-1}}{1-x}dx-\int_0^1\frac{x^{b+n-1}}{1-x}dx\right).$$

Tomemos la primera integral y consideremos el cambio $x=e^{-t}$, entonces

$$\int_0^1\frac{x^{b-1}}{1-x}dx=\int_\infty^0\frac{e^{-t(b-1)}}{1-e^{-t}}(-e^{-t})dt=\int_0^\infty\frac{e^{-bt}}{1-e^{-t}}dt.$$

De manera análoga para la otra integral se tiene que

$$\int_0^1\frac{x^{b+n-1}}{1-x}dx=\int_0^\infty\frac{e^{-(b+n)t}}{1-e^{-t}}dt.$$

Luego, restando estas integrales se tiene que

$$\begin{aligned}\int_0^\infty\frac{e^{-bt}}{1-e^{-t}}dt-\int_0^\infty\frac{e^{-(b+n)t}}{1-e^{-t}}dt&=\int_0^\infty\frac{e^{-bt}}{1-e^{-t}}dt-\int_0^\infty\frac{e^{-(b+n)t}}{1-e^{-t}}dt+\int_0^\infty\frac{e^{-t}}{t}dt-\int_0^\infty\frac{e^{-t}}{t}dt\\&=\int_0^\infty\left(\frac{e^{-t}}{t}-\frac{e^{-(b+n)t}}{1-e^{-t}}\right)dt-\int_0^\infty\left(\frac{e^{-t}}{t}-\frac{e^{-bt}}{1-e^{-t}}\right)dt\\&=\psi(b+n)-\psi(b).\end{aligned}$$

Donde se utiliza la representación integral de la función digamma (2.5). Por lo tanto en el caso $b=a+1$ se tiene

$$\mathbb{E}(\widetilde{W}^n)=(b-1)[\psi(b+n)-\psi(b)].$$

Por último, el caso $b-a<1$ se calcula de manera diferente. Como $\widetilde{W}^n$ es una variable aleatoria acotada esta debe tener esperanza finita, pero a diferencia del caso $b-a>1$ la representación integral de la función beta no existe para valores negativos por lo que hay que tomar otro camino. Para esto probemos primero la siguiente igualdad, por inducción.

$$\sum_{x=0}^{n-1}\frac{x\Gamma(a+x)}{\Gamma(b+x+1)}=\frac{b(b+1)\Gamma(a+1)\Gamma(b+n+1)-(b+n)\Gamma(b+2)\Gamma(a+n)(bn+a-an)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+1)}. \tag{3.2}$$

- El caso $n=1$. La suma del lado izquierdo es claramente 0, veamos que pasa con el numerador del lado derecho.

$$
\begin{aligned}
&b(b+1)\Gamma(a+1)\Gamma(b+1+1) - (b+1)\Gamma(b+2)\Gamma(a+1)(b+a-a)\\
&= b(b+1)\Gamma(a+1)\Gamma(b+2) - (b+1)\Gamma(b+2)\Gamma(a+1)b\\
&= 0.
\end{aligned}
$$

Por lo que se cumple el caso base.

- Antes de ver el paso inductivo, veamos que ocurre con $n = 2$. El lado izquierdo está conformado por el valor $x = 1$, que es simplemente $\frac{\Gamma(a+1)}{\Gamma(b+2)}$, por lo que el lado derecho también debe serlo. En efecto, usando la recursión de la función gamma, tenemos que $(b+2)\Gamma(b+2) = \Gamma(b+3)$ y $\Gamma(a+2) = (a+1)\Gamma(a+1)$ y entonces el lado derecho es

$$
\begin{aligned}
&\frac{b(b+1)\Gamma(a+1)\Gamma(b+3) - (b+2)\Gamma(b+2)\Gamma(a+2)(2b+a-2a)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+3)}\\
&= \frac{b(b+1)\Gamma(a+1)\Gamma(b+3) - \Gamma(b+3)(a+1)\Gamma(a+1)(2b-a)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+3)}\\
&= \frac{\Gamma(a+1)}{\Gamma(b+2)}\frac{b(b+1) - (a+1)(2b-a)}{(a-b)(a-b+1)},
\end{aligned}
$$

donde ya aparece la fracción que se desea, por lo que basta ver que la otra es 1, desarrollando y reordenando se tiene que

$$
\frac{b^2+b-2ab+a^2-2b+a}{a^2-2ab+a+b^2-b} = \frac{a^2+b^2-2ab+a-b}{a^2+b^2-2ab+a-b} = 1.
$$

Es decir, se cumple la igualdad (3.2) para $n = 2$.

- Supongamos que se tiene la igualdad (3.2) para $n-1$ y probemos que se cumple también para n.

$$
\sum_{x=0}^{n}\frac{x\Gamma(a+x)}{\Gamma(b+x+1)} = \sum_{x=0}^{n-1}\frac{x\Gamma(a+x)}{\Gamma(b+x+1)} + \frac{n\Gamma(a+n)}{\Gamma(b+n+1)},
$$

usando la hipótesis inductiva, y multiplicando por $(b+n+1)$ en el numerador y denominador para hacer aparecer $\Gamma(b+n+2)$, se obtiene

$$
\begin{aligned}
&\sum_{x=0}^{n}\frac{x\Gamma(a+x)}{\Gamma(b+x+1)}\\
&= \frac{b(b+1)\Gamma(a+1)\Gamma(b+n+1) - (b+n)\Gamma(b+2)\Gamma(a+n)(bn+a-an)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+1)} + \frac{n\Gamma(a+n)}{\Gamma(b+n+1)}\\
&= \frac{b(b+1)\Gamma(a+1)\Gamma(b+n+2) - (b+n+1)(b+n)\Gamma(b+2)\Gamma(a+n)(bn+a-an)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+2)}\\
&\quad + \frac{(b+n+1)n\Gamma(a+n)}{\Gamma(b+n+2)}\\
&= \frac{b(b+1)\Gamma(a+1)\Gamma(b+n+2) - (b+n+1)(b+n)\Gamma(b+2)\Gamma(a+n)(bn+a-an)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+2)}\\
&\quad + \frac{(a-b)(a-b+1)\Gamma(b+2)(b+n+1)n\Gamma(a+n)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+2)}.
\end{aligned}
$$

Donde notamos que ya tenemos listo el denominador y el primer término del numerador de la fracción, trabajemos el resto del numerador

$$\begin{aligned}
&-(b+n+1)\Gamma(b+2)\Gamma(a+n)[(b+n)(bn+a-an)-n(a-b)(a-b+1)] \\
&= -(b+n+1)\Gamma(b+2)\Gamma(a+n)[b^2n+ab-abn+bn^2+an-an^2 \\
&\qquad\qquad -a^2n+abn-an+abn-b^2n+bn] \\
&= -(b+n+1)\Gamma(b+2)\Gamma(a+n)[ab-a^2n+abn+bn^2-an^2+bn] \\
&= -(b+n+1)\Gamma(b+2)\Gamma(a+n)[a(b-an+bn)+n(bn-an+n)] \\
&= -(b+n+1)\Gamma(b+2)\Gamma(a+n)(a+n)(b-an+bn) \\
&= -(b+n+1)\Gamma(b+2)\Gamma(a+n+1)(b-an+bn),
\end{aligned}$$

donde se usa la recursión de la función gamma para hacer aparecer $\Gamma(a+n+1)$. Además, reordenando el último valor como $b-an+bn = b(n+1)+a-a(n+1)$, se obtiene el término faltante del numerador. Es decir, para n se tiene que

$$\begin{aligned}
&\sum_{x=0}^{n} \frac{x\Gamma(a+x)}{\Gamma(b+x+1)} \\
&= \frac{b(b+1)\Gamma(a+1)\Gamma(b+n+2)-(b+n+1)\Gamma(b+2)\Gamma(a+n+1)(b(n+1)+a-a(n+1))}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+2)}
\end{aligned}$$

Con lo cual queda demostrada por inducción la igualdad dada por (3.2).

Continuando con la demostración de la esperanza de $\widetilde{W}^n$ para el caso $b-a<1$, la esperanza discreta está dada por

$$\begin{aligned}
\mathbb{E}(\widetilde{W}^n) &= \sum_{x=0}^{n} x\mathbb{P}(\widetilde{W}^n = x) \\
&= \sum_{x=0}^{n-1} x(b-a)\frac{(a)_x}{(b)_{x+1}} + n\frac{(a)_n}{(b)_n} \\
&= \sum_{x=0}^{n-1} x(b-a)\frac{\Gamma(a+x)\Gamma(b)}{\Gamma(a)\Gamma(b+x+1)} + n\frac{(a)_n}{(b)_n} \\
&= (b-a)\frac{\Gamma(b)}{\Gamma(a)}\sum_{x=0}^{n-1}\frac{x\Gamma(a+x)}{\Gamma(b+x+1)} + n\frac{(a)_n}{(b)_n}.
\end{aligned}$$

Así, usando el resultado de la ecuación (3.2), se tiene que el primer término de la igualdad anterior se puede separar en dos. El primero de estos es

$$\begin{aligned}
(b-a)\frac{\Gamma(b)}{\Gamma(a)}\frac{b(b+1)\Gamma(a+1)\Gamma(b+n+1)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+1)} &= \frac{\Gamma(b+1)(b+1)a\Gamma(a)}{\Gamma(a)(b-a-1)\Gamma(b+2)} \\
&= \frac{\Gamma(b+2)a}{(b-a-1)\Gamma(b+2)} \\
&= \frac{a}{b-a-1}.
\end{aligned}$$

El otro corresponde a

$$\begin{aligned}
&-(b-a)\frac{\Gamma(b)}{\Gamma(a)}\frac{(b+n)\Gamma(b+2)\Gamma(a+n)(bn+a-an)}{(a-b)(a-b+1)\Gamma(b+2)\Gamma(b+n+1)}\\
&=-\frac{1}{b-a-1}\frac{\Gamma(b)\Gamma(a+n)(bn+a-an)}{\Gamma(a)\Gamma(b+n)}\\
&=-\frac{1}{b-a-1}\frac{\Gamma(a+n)}{(b)_n\Gamma(a)}(bn+a-an)\\
&=-\frac{1}{b-a-1}\frac{\Gamma(a+n+1)a}{(b)_n(a+n)\Gamma(a+1)}(bn+a-an)\\
&=-\frac{1}{b-a-1}\frac{(a+1)_n}{(b)_n}\frac{a}{a+n}(bn+a-an),
\end{aligned}$$

juntando esto con el término $n(a)_n/(b)_n$ que falta de la esperanza de $\widetilde{W}^n$, nos queda

$$\begin{aligned}
&-\frac{1}{b-a-1}\frac{(a+1)_n}{(b)_n}\frac{a}{a+n}(bn+a-an)+n\frac{\Gamma(a+n)}{\Gamma(a)(b)_n}\\
&=-\frac{1}{b-a-1}\frac{(a+1)_n}{(b)_n}\left[\frac{a}{a+n}(bn+a-an)-(b-a-1)\frac{\Gamma(a+1)}{\Gamma(a+1+n)}\frac{n\Gamma(a+n)}{\Gamma(a)}\right]\\
&=-\frac{1}{b-a-1}\frac{(a+1)_n}{(b)_n}\left[\frac{a}{a+n}(bn+a-an)-(b-a-1)\frac{an}{a+n}\right]\\
&=-\frac{a}{b-a-1}\frac{(a+1)_n}{(b)_n}\frac{1}{a+n}[bn+a-an-bn+an+n]\\
&=-\frac{a}{b-a-1}\frac{(a+1)_n}{(b)_n}\frac{1}{a+n}[a+n]\\
&=-\frac{a}{b-a-1}\frac{(a+1)_n}{(b)_n}.
\end{aligned}$$

Con esto, sumando los dos valores encontrados, se llega a que la esperanza de $\widetilde{W}^n$ en el caso $b-a<1$ es

$$\begin{aligned}
\mathbb{E}(\widetilde{W}^n)&=\frac{a}{b-a-1}-\frac{a}{b-a-1}\frac{(a+1)_n}{(b)_n}\\
&=\frac{a}{b-a-1}\left(1-\frac{(a+1)_n}{(b)_n}\right).
\end{aligned}$$

Con lo cual se concluye el resultado. □

Notamos que en la esperanza de la distribución de Waring acotada (3.1), el primer término coincide, en el caso $b-a>1$, con la esperanza de la distribución de Waring. Lo más interesante es que para el caso $b-a<1$ la fórmula es la misma al caso $b-a>1$, además que en el caso borde vuelve a aparecer la función digamma que hemos visto en otros valores esperados.

Cabe destacar también que el caso borde, $b-a=1$, en (3.1) corresponde a la derivada del caso general $b-a\neq 1$. En efecto, si consideramos $h=b-a-1$ entonces

$$\frac{a}{b-a-1}\left(1-\frac{(a+1)_n}{(b)_n}\right)=\frac{a}{h}\left(1-\frac{\Gamma(a+1+n)\Gamma(b)}{\Gamma(a+1)\Gamma(b+n)}\right)$$
$$=(b-1-h)\left(\frac{1}{h}-\frac{\Gamma(b+n-h)}{h\Gamma(b+n)}\frac{\Gamma(b)}{\Gamma(b-h)}\right).$$

Notando que la función digamma se puede escribir como

$$\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}=\frac{\lim_{h\to 0}\frac{\Gamma(x)-\Gamma(x-h)}{h}}{\Gamma(x)}=\lim_{h\to 0}\left(\frac{1}{h}-\frac{\Gamma(x-h)}{h\Gamma(x)}\right),$$

entonces

$$\begin{aligned}\frac{1}{h}-\frac{\Gamma(b+n-h)}{h\Gamma(b+n)}\frac{\Gamma(b)}{\Gamma(b-h)}&=\left(\frac{1}{h}-\frac{\Gamma(b+n-h)}{h\Gamma(b+n)}\right)\frac{\Gamma(b)}{\Gamma(b-h)}+\frac{1}{h}-\frac{\Gamma(b)}{h\Gamma(b-h)}\\&=\left(\frac{1}{h}-\frac{\Gamma(b+n-h)}{h\Gamma(b+n)}\right)\frac{\Gamma(b)}{\Gamma(b-h)}+\frac{1}{h}\left(\frac{\Gamma(b-h)-\Gamma(b)}{\Gamma(b-h)}\right)\\&=\left(\frac{1}{h}-\frac{\Gamma(b+n-h)}{h\Gamma(b+n)}\right)\frac{\Gamma(b)}{\Gamma(b-h)}-\frac{\frac{\Gamma(b)-\Gamma(b-h)}{h}}{\Gamma(b-h)}.\end{aligned}$$

Luego, tomando límite con h tendiendo a 0, que es realmente $b-a\to 1$, recuperamos la función digamma

$$\psi(b+n)\frac{\Gamma(b)}{\Gamma(b)}-\frac{\Gamma'(b)}{\Gamma(b)},$$

y juntando esto con el valor $\lim_{h\to 0}(b-1-h)=b-1$ que faltaba se obtiene

$$(b-1)(\psi(b+n)-\psi(b)),$$

que es justamente el caso borde $b-a=1$ de la esperanza de la distribución de Waring acotada (3.1).

3.3. Tiempo esperado entre apariciones

Continuando con los tiempos entre apariciones, para estudiar el tiempo esperado es necesario tener la distribución conjunta de los tiempos entre apariciones, donde se sigue considerando la muestra $\mathbf{X}_N$ de un $PDP(d,\alpha)$.

Proposición 3.3 (ver [8]) *Para r tal que $1\leq r<N$ y $t_1,\ldots,t_r\in\mathbb{N}$ la distribución conjunta de los tiempos entre apariciones $(T_j)_{j=1}^r$ está dada por*

$$\mathbb{P}(T_1=t_1,\ldots,T_r=t_r)=\frac{(\alpha|d)_{r+1}}{(\alpha)_{\bar{t}_r+1}}\prod_{i=1}^{r}(1-id+\bar{t}_{i-1})_{t_i-1},\tag{3.3}$$

dado que $\bar{t}_r=\sum_{j=1}^r t_j<N$ y se considera que $\bar{t}_0=0$.

En el caso particular $k = 1$, se tiene por convención que $T_1 = N$ y

$$\mathbb{P}(T_1 = N) = \frac{(1-d)_{N-1}}{(1+\alpha)_{N-1}}.$$

Cuando se tiene el caso usual de $k > 1$ o cuando $T_1 < N$ se obtiene de (3.3) que la distribución del primer tiempo entre apariciones está dada por

$$\mathbb{P}(T_1 = t_1) = (\alpha + d)\frac{(1-d)_{t_1-1}}{(1+\alpha)_{t_1}},$$

esto notando que $(\alpha)_{t_1+1} = \alpha(\alpha+1)_{t_1}$. Con lo cual se tiene que $(T_1 - 1) \sim$ A-Waring$(N - 1, 1 + \alpha, 1 - d)$. A continuación se tiene una proposición que extiende este resultado para el tiempo T_r condicionado a los tiempos anteriores.

Proposición 3.4 (ver [8]) *Para $r, t_1, \ldots, t_{r-1} \in \mathbb{N}$ tal que $1 < r \leq N$ y $\bar{t}_{r-1} < N$ se tiene que la distribución condicional del tiempo entre apariciones T_r dados los tiempos anteriores se define*

- *Para $r < k$, $1 \leq t_r < \bar{t}_{r-1}$*

$$\mathbb{P}(T_r = t_r | \{T_j = t_j\}_{j=1}^{r-1}) = (\alpha + rd)\frac{(\bar{t}_{r-1} + 1 - rd)_{t_r-1}}{(\bar{t}_{r-1} + 1 + \alpha)_{t_r}},$$

- *para $r = k$ o $t_r = N - \bar{t}_{r-1}$*

$$\mathbb{P}(T_r = t_r | \{T_j = t_j\}_{j=1}^{r-1}) = \frac{(\bar{t}_{r-1} + 1 - rd)_{t_r-1}}{(\bar{t}_{r-1} + 1 + \alpha)_{t_r-1}}.$$

Es decir, $T_r - 1$ condicionado a los tiempos anteriores $\{T_j = t_j\}_{j=1}^{r-1}$ tiene una distribución A-Waring$(N - \bar{t}_{r-1} - 1, \bar{t}_{r-1} + 1 + \alpha, \bar{t}_{r-1} + 1 - rd)$.

De esta proposición tenemos que para la distribución condicional no se necesita explícitamente el valor de cada tiempo entre apariciones, solo la suma de los anteriores. Así, se puede definir la variable

$$\overline{T}_r = \sum_{j=1}^{r} T_j$$

y se tiene que $(T_r - 1 \mid \overline{T}_{r-1} = \bar{t}_{r-1}) \sim$ A-Waring$(N - \bar{t}_{r-1} - 1, \bar{t}_{r-1} + 1 + \alpha, \bar{t}_{r-1} + 1 - rd)$.

Al tener esta distribución, podemos usar la ecuación (3.1) de la esperanza de la distribución Waring acotada de la sección anterior para obtener el tiempo esperado entre apariciones de la especie r condicionado al valor de $\overline{T}_{r-1} = \bar{t}_{r-1}$, que por comodidad se deja simplemente como $T_r \mid \bar{t}_{r-1}$.

Proposición 3.5 *Para $1 \leq r < k$ y $\bar{t}_{r-1}$ tal que $\bar{t}_0 = 0$, se tiene que*

$$\mathbb{E}(T_r \mid \bar{t}_{r-1}) = \begin{cases} 1 + \frac{\bar{t}_{r-1}+1-rd}{\alpha+rd-1}\left(1 - \frac{(\bar{t}_{r-1}+2-rd)_{N-\bar{t}_{r-1}-1}}{(\bar{t}_{r-1}+1+\alpha)_{N-\bar{t}_{r-1}-1}}\right) & \alpha + rd \neq 1 \\ 1 + (\bar{t}_{r-1}+\alpha)[\psi(\alpha+N) - \psi(\bar{t}_{r-1}+1+\alpha)] & \alpha + rd = 1. \end{cases} \tag{3.4}$$

Notamos de este resultado que el valor esperado está dado por el caso $\alpha + rd \neq 1$ prácticamente siempre, aunque puede ocurrir que se cumpla $\alpha + rd = 1$ para ciertos valores de α y d pero sería solo para a lo más un valor de r. Esto último tomando en cuenta el Proceso de Poisson-Dirichlet $d \neq 0$, pues si se toma $d = 0$ y $\alpha = 1$ se tendría siempre el segundo caso para la esperanza.

La ecuación (3.4) es un resultado de esta tesis, obtenido a partir de lo expuesto en [8] de la proposición anterior y lo demostrado en la Proposición 3.2, dada por [9]. Además, este resultado coincide con la esperanza condicional obtenida por Huillet [9] en el caso del Proceso de Dirichlet $d = 0$ (y $\alpha \neq 1$), la cual se obtiene de otra manera: a partir de la distribución de Dirichlet y tomando el límite de Kingman. Dicho de otra forma, el tiempo esperado entre apariciones en (3.4) corresponde a la generalización del caso de un solo parámetro.

Capítulo 4

Variación de la entropía y aparición de especies

Volviendo al estudio de la entropía, este capítulo está dedicado al resultado principal de esta tesis, que es una relación y unas desigualdades para la variación de la entropía entre dos pasos sucesivos para el Proceso de Poisson-Dirichlet.

Para esto, se estudia primero la variación de la entropía en el caso frecuentista y su relación con la aparición de nuevas especies en la Sección 4.1, para esto se sigue estrechamente las notas preliminares de Servet Martínez [10], dando más detalle en algunas partes de la demostración principal. En la Sección 4.2, inspirados por lo hecho en el caso frecuentista, se estudia la variación de la entropía Bayesiana en el caso del Proceso de Poisson-Dirichlet, obteniendo así el resultado principal en el Teorema 4.1, además de distintos corolarios a partir de esto.

4.1. Variación de la entropía frecuentista

Como fue mencionado, en esta sección se estudia la variación de la entropía en el caso frecuentista usando [10]. Para poder hacer esto, se explica la notación usada y se expone el resultado principal que es una desigualdad para una diferencia ponderada de la entropía entre dos pasos sucesivos, además de la demostración de esta variación.

El contexto es el mismo que se ha considerado en los capítulos anteriores, es decir, una secuencia de elementos $X_1, \ldots, X_n$ que se van clasificando en alguna clase (o especie) a medida que son observados. En el paso n se tiene que se han recolectado n elementos que denotaremos $I_n = \{X_1, \ldots, X_n\}$, lo que corresponde a agregar el elemento X_n a I_{n-1} (donde $I_0 = \emptyset$). Sea $\sim$ la relación de poner dos elementos juntos, es decir, que son parte de la misma clase. En cada paso n el conjunto I_n está dividido en una familia $\mathbb{J}(I_n)$ de las clases de equivalencia disjuntas: todos los elementos de la clase $J \in \mathbb{J}(I_n)$ están en relación $\sim$ entre ellos y ninguno de los elementos en $I_n \setminus J$ está en relación $\sim$ con los elementos de J. En este proceso siempre es posible que aparezcan nuevas clases.

Denotemos $\mathbb{J}_n = \mathbb{J}(I_n)$ a la familia de clases de equivalencia disjuntas. Las clases que no contienen a X_n se mantienen sin cambios cuando se pasa de $\mathbb{J}_{n-1}$ a $\mathbb{J}_n$, si X_n está en relación con algún elemento de alguna clase de equivalencia $J \in \mathbb{J}_{n-1}$ entonces se agrega para obtener la clase $J \cup \{X_n\} \in \mathbb{J}_n$, en cambio si X_n no está en relación $\sim$ con ninguno de los elementos

$X_1, \ldots, X_{n-1}$ entonces se crea una nueva clase en el paso n dada por el singleton $\{X_n\} \in \mathbb{J}_n$.

Sea $J_n(X_l)$ la clase de equivalencia de X_l en $\mathbb{J}_n$. Para $n \geq l$ se tiene claramente que $X_l \in J_n(X_l)$ y si $n < l$ se tiene $J_n(X_l) = \emptyset$ cuando ninguno de los elementos $X_1, \ldots, X_n$ está en relación $\sim$ con X_l. Entonces

$$J_n(X_n) = J_{n-1}(X_n) \cup \{X_n\} \text{ y } J_n(X_m) = J_{n-1}(X_m) \text{ si } X_m \not\sim X_n.$$

Para $J \in \mathbb{J}_n$ su peso $|J|_n$ es el número de pasos antes de o en n cuando un objeto en J es recolectado.

$$|J|_n = |\{l \in \{1, \ldots, n\} : X_l \in J\}.$$

Así,

$$|J_n(X_n)|_n = |J_{n-1}(X_n)|_n + 1 \text{ y } |J_n(X_l)|_n = |J_{n-1}(X_l)|_{n-1} \text{ para } X_l \not\sim X_n.$$

Además por definición se tiene

$$n = \sum_{J \in \mathbb{J}_n} |J|_n.$$

Con esto la entropía en el paso n es

$$H_n = H(I_n) = -\sum_{J \in \mathbb{J}_n} \frac{|J|_n}{n} \log\left(\frac{|J|_n}{n}\right), \tag{4.1}$$

donde se tiene por continuidad que $0 \log(0) = 0$, así se puede agregar la clase vacía si se desea. En (4.1) el valor $|J|_n/n$ es la frecuencia de J en el paso n y $-\log(|J|_n/n)$ es la información dada por esta clase, esta ecuación es análoga a (2.2) mostrada previamente. Además, se tiene la siguiente desigualdad

$$H_n \leq \log n,$$

y la igualdad se alcanza si y solo si cada par de elementos en I_n no están relacionados $\sim$, esto es $X_i \not\sim X_j$ para todo $1 \leq i < j \leq n$. De hecho, en este caso $|\mathbb{J}_n| = n$, donde cada clase $J \in \mathbb{J}_n$ es un singleton con $|J|_n = 1$ y así $H_n = -\sum_{i=1}^{n} 1/n \log(1/n) = \log n$. Se sabe que en todos los otros casos la desigualdad es estricta, donde notamos también que $H_1 = 0$.

Para estudiar la secuencia de entropía $(H_n : n \geq 1)$ es de utilidad denotar por $J_n^* = J_n(X_n) \in \mathbb{J}_n$ a la clase de $\mathbb{J}_n$ que contiene a X_n, y $\ell_n = |J_n^*|_n$ como el número de elementos que tiene. Con la notación ya definida, se tiene el siguiente resultado.

Proposición 4.1 *Para $n \geq 1$ sea*

$$\Delta_{n+1} = (n+1)(\log(n+1) - H_{n+1}) - n(\log(n) - H_n).$$

Entonces, las siguientes relaciones se cumplen

$$\forall n \geq 1, \quad \Delta_{n+1} = \ell_{n+1} \log(\ell_{n+1}) - (\ell_{n+1} - 1) \log(\ell_{n+1} - 1) \geq 0. \tag{4.2}$$

Más aún, Δ_{n+1} se hace cero cuando aparece una nueva clase en $n+1$,

$$\Delta_{n+1} = 0 \Leftrightarrow \ell_{n+1} = 1. \tag{4.3}$$

Demostración. La ecuación (4.3) se obtiene de (4.2) y el hecho de que $\ell \geq 1$ satisface la relación $\ell \log(\ell) - (\ell - 1)\log(\ell - 1) = 0$ si y solo si $\ell = 1$. Además como la función log es creciente se tiene que $\ell \log(\ell) - (\ell - 1)\log(\ell - 1)$ es siempre mayor o igual a 0 para $\ell \geq 1$, con lo cual se tiene la desigualdad en (4.2). Demostremos entonces la relación principal de (4.2), tenemos que la entropía en el paso $n + 1$ es

$$\begin{aligned} H_{n+1} &= -\sum_{J\in\mathbb{J}_{n+1}} \frac{|J|_{n+1}}{n+1} \log\left(\frac{|J|_{n+1}}{n+1}\right) \\ &= -\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n+1}}{n+1} \log\left(\frac{|J|_{n+1}}{n+1}\right) - \frac{|J^*_{n+1}|_{n+1}}{n+1} \log\left(\frac{|J^*_{n+1}|_{n+1}}{n+1}\right) \\ &= -\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n+1} \log\left(\frac{|J|_{n}}{n+1}\right) - \frac{\ell_{n+1}}{n+1} \log\left(\frac{\ell_{n+1}}{n+1}\right), \end{aligned}$$

donde en la última suma se usa $|J|_n$ en vez de $|J|_{n+1}$ porque en los J que se están sumando ya no se está considerando el elemento X_{n+1} pues se quita la clase de equivalencia que lo contiene J^*_{n+1}. Luego,

$$H_{n+1} = -\frac{n}{n+1} \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} \log\left(\frac{|J|_{n}}{n+1}\right) - \frac{\ell_{n+1}}{n+1} \log\left(\frac{\ell_{n+1}}{n+1}\right),$$

desarrollando esta sumatoria

$$\begin{aligned} &\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} \log\left(\frac{|J|_{n}}{n+1}\right) \\ &= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} \log(|J|_n) - \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} \log(n+1) \\ &= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} \log(|J|_n) - \log(n+1) \left(\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n}\right), \end{aligned}$$

notamos que

$$\begin{aligned} \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} &= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_{n}}{n} + \frac{\ell_{n+1}}{n} - \frac{\ell_{n+1}}{n} \\ &= \frac{1}{n} \sum_{J\in\mathbb{J}_{n+1}} |J|_{n+1} - \frac{\ell_{n+1}}{n} \\ &= \frac{1}{n}(n+1) - \frac{\ell_{n+1}}{n} \\ &= 1 - \frac{\ell_{n+1} - 1}{n} \end{aligned}$$

y así

$$\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n+1}\right)$$
$$= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log(|J|_n) - \log(n+1)\left(1-\frac{\ell_{n+1}-1}{n}\right).$$

Agregando el término faltante para tener $\log\left(\frac{|J|_n}{n}\right)$ en la suma y usando lo calculado anteriormente se tiene que

$$\begin{aligned}
&\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n+1}\right)\\
&= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log(|J|_n) - \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log(n)\\
&\quad + \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log(n) - \log(n+1)\left(1-\frac{\ell_{n+1}-1}{n}\right)\\
&= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n}\right) + (\log(n)-\log(n+1))\left(1-\frac{\ell_{n+1}-1}{n}\right).
\end{aligned}$$

Luego, reemplazando esto en la entropía H_{n+1} se obtiene

$$\begin{aligned}
H_{n+1} = &-\frac{n}{n+1}\left(\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n}\right)\right)\\
&- \frac{n}{n+1}(\log(n)-\log(n+1))\left(1-\frac{\ell_{n+1}-1}{n}\right) - \frac{\ell_{n+1}}{n+1}\log\left(\frac{\ell_{n+1}}{n+1}\right),
\end{aligned}$$

donde notamos que la suma que queda es muy similar a la entropía H_n excepto que falta una clase de equivalencia por sumar. De hecho si se suma el valor correspondiente a la clase faltante pero sin considerar el valor X_{n+1} o el paso $n+1$, es decir, la clase $J^*_{n+1}\setminus\{X_{n+1}\}$ que tiene peso $\ell_{n+1}-1$, entonces se vuelve a considerar la clase faltante en la suma y esto hasta el paso n,

$$\begin{aligned}
&\sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n}\right)\\
&= \sum_{J\in\mathbb{J}_{n+1}\setminus\{J^*_{n+1}\}} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n}\right) + \frac{\ell_{n+1}-1}{n}\log\left(\frac{\ell_{n+1}-1}{n}\right) - \frac{\ell_{n+1}-1}{n}\log\left(\frac{\ell_{n+1}-1}{n}\right)\\
&= \sum_{J\in\mathbb{J}_n} \frac{|J|_n}{n}\log\left(\frac{|J|_n}{n}\right) - \frac{\ell_{n+1}-1}{n}\log\left(\frac{\ell_{n+1}-1}{n}\right)\\
&= -H_n - \frac{\ell_{n+1}-1}{n}\log\left(\frac{\ell_{n+1}-1}{n}\right).
\end{aligned}$$

Así,

$$H_{n+1} = \frac{n}{n+1}\left(H_n + \frac{\ell_{n+1}-1}{n}\log\left(\frac{\ell_{n+1}-1}{n}\right)\right) - \frac{n}{n+1}(\log(n)-\log(n+1))\left(1-\frac{\ell_{n+1}-1}{n}\right) - \frac{\ell_{n+1}}{n+1}\log\left(\frac{\ell_{n+1}}{n+1}\right)$$

Ahora,

$$\begin{aligned}
&(n+1)H_{n+1} - nH_n \\
&= (\ell_{n+1}-1)\log(\ell_{n+1}-1) - (\ell_{n+1}-1)\log(n) \\
&\quad - n\log(n) + n\log(n+1) + \log(n)(\ell_{n+1}-1) - \log(n+1)(\ell_{n+1}-1) \\
&\quad - \ell_{n+1}\log(\ell_{n+1}) + \ell_{n+1}\log(n+1) \\
&= (\ell_{n+1}-1)\log(\ell_{n+1}-1) - n\log(n) + (n+1)\log(n+1) - \ell_{n+1}\log(\ell_{n+1}).
\end{aligned}$$

Entonces,

$$\begin{aligned}
(n+1)(\log(n+1)-H_{n+1}) - n(\log(n)-H_n) &= \ell_{n+1}\log(\ell_{n+1}) - (\ell_{n+1}-1)\log(\ell_{n+1}-1) \\
\Delta_{n+1} &= \ell_{n+1}\log(\ell_{n+1}) - (\ell_{n+1}-1)\log(\ell_{n+1}-1).
\end{aligned}$$

Por lo tanto, la fórmula (4.2) queda demostrada. □

Como $\log(n)-H_n \geq 0$ entonces el valor $\Delta_{n+1} = (n+1)(\log(n+1)-H_{n+1})-n(\log(n)-H_n)$ es una diferencia entre dos números no negativos, más aún, hemos visto que este valor es siempre no negativo y además que se hace cero solo cuando emerge una nueva clase, esto último para $n \geq 1$. En el caso $n = 0$, y usando que $H_1 = 0$, se tiene que $\Delta_1 = (n+1)(\log(n+1)-H_{n+1})|_{n=0} = 0$, lo cual es consistente con el hecho de que en el paso 1 siempre aparece una nueva clase.

Esta proposición muestra que uno puede recuperar el tamaño de la clase que contiene al objeto recolectado en el paso n. Notamos que el paso de H_n a H_{n+1} requiere un cambio global de las frecuencias de las clases y también un cambio global de la información asociada a estas clases.

4.2. Variación de la entropía Bayesiana

De acuerdo al análisis realizado en la sección anterior, si bien el número de clases en la muestra permanece finito, pueden aparecer nuevas clases a medida que se recolectan nuevos elementos, sin embargo no se tiene una probabilidad definida con la que aparecen estas nuevas clases. Dado esto y para considerar el caso en que hayan infinitas clases, se realiza un análisis de la variación de la entropía en el caso Bayesiano, en particular en el contexto del Proceso de Poisson-Dirichlet.

Dada una muestra $\mathbf{X}_N = (X_1, \ldots, X_N)$ de un Proceso de Poisson-Dirichlet de parámetros d, α, se tiene que la entropía esperada posterior está dada por la ecuación (2.14), que se repite a continuación,

$$\mathbb{E}(H|\mathbf{X}_N, d, \alpha) = \psi(\alpha + N + 1) - \frac{\alpha + kd}{\alpha + N}\psi(1 - d) - \frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right).$$

Como se desea estudiar la variación con respecto al paso siguiente, se define la entropía posterior al recolectar un nuevo elemento X_{N+1}, que puede estar en una especie X_j^* ya vista (con $j = 1, \ldots, k$) o no estar en ninguna de las especies de la muestra. En la muestra se asume que hay $K = k$ especies distintas, esta dependencia al valor k no se explicita pero siempre está considerada (igual para d, α). Así, la entropía queda definida por una nueva distribución posterior, a la cual se le calcula la esperanza, que denotamos de la siguiente manera

$$\begin{aligned}
\hat{H}_N &= \mathbb{E}(H|\mathbf{X}_N, d, \alpha) \\
\hat{H}_{N+1,\neq} &= \mathbb{E}(H|\mathbf{X}_N, X_{N+1} \notin X_j^* \ \forall j, d, \alpha) \\
\hat{H}_{N+1,j} &= \mathbb{E}(H|\mathbf{X}_N, X_{N+1} \in X_j^*, d, \alpha) \quad \text{para algún } j = 1, \ldots, k.
\end{aligned}$$

Con esto se puede enunciar el resultado principal de esta tesis, que corresponde a una relación para la diferencia ponderada de la entropía esperada posterior, entre dos pasos sucesivos, en el caso del Proceso de Poisson-Dirichlet, y un par de desigualdades con respecto a esta variación.

Teorema 4.1 *Para $N \geq 1$ sean*

$$\begin{aligned}
\Delta_{N+1,\neq} &= (\alpha + N + 1)\hat{H}_{N+1,\neq} - (\alpha + N)\hat{H}_N \\
\Delta_{N+1,j} &= (\alpha + N + 1)\hat{H}_{N+1,j} - (\alpha + N)\hat{H}_N \quad \textit{para algún } j = 1, \ldots, k.
\end{aligned}$$

Entonces, se cumplen las siguientes relaciones para todo $N \geq 1$

$$\Delta_{N+1,\neq} = \psi(\alpha + N + 1) - \psi(1 - d) > 0 \tag{4.4}$$
$$\Delta_{N+1,j} = \psi(\alpha + N + 1) - \psi(n_j + 1 - d) > 0 \quad \textit{para algún } j = 1, \ldots, k. \tag{4.5}$$

Más aún, se tienen las siguientes desigualdades para todo $j = 1, \ldots, k$

$$\Delta_{N+1,j} < \Delta_{N+1,\neq}, \tag{4.6}$$

y para $j^* = \arg\max_{j=1,\ldots,k} n_j$

$$\Delta_{N+1,j^*} \leq \Delta_{N+1,j}. \tag{4.7}$$

Demostración. Primero notamos que los valores en (4.4) y (4.5) sean mayores que 0 viene del hecho de que los parámetros del Proceso de Poisson-Dirichlet cumplen que $\alpha > -d$ y que la función digamma ψ es una función creciente en los reales positivos, además que las frecuencias cumplen $1 \leq n_j \leq N$ para todo $j = 1, \ldots, k$. Así,

$$\begin{aligned}
&\alpha + N + 1 > \alpha + 1 > 1 - d \\
&\psi(\alpha + N + 1) > \psi(1 - d)
\end{aligned}$$

y

$$\alpha + N + 1 \geq \alpha + n_j + 1 > n_j + 1 - d$$
$$\psi(\alpha + N + 1) > \psi(n_j + 1 - d).$$

Veamos entonces la demostración para la igualdad en la ecuación (4.4). En este caso la entropía esperada posterior $\hat{H}_{N+1,\neq}$ se calcula mediante (2.14) solo que considerando que la muestra ahora es de tamaño $N+1$, el número de especies aumenta a $k+1$ y las frecuencias n_j se mantienen para $j = 1, \ldots, k$, pero aparece una frecuencia para la nueva especie $n_{k+1} = 1$. Entonces,

$$\begin{aligned}
&\hat{H}_{N+1,\neq} \\
&= \psi(\alpha + N + 1 + 1) - \frac{\alpha + (k+1)d}{\alpha + N + 1}\psi(1-d) - \frac{1}{\alpha + N + 1}\left(\sum_{i=1}^{k+1}(n_i - d)\psi(n_i - d + 1)\right),
\end{aligned}$$

usando la relación de recurrencia (2.4) de la función digamma, $\psi(x+1) = \psi(x) + 1/x$, para el primer valor y sacando el término $k+1$ de la sumatoria se tiene que

$$\begin{aligned}
&\hat{H}_{N+1,\neq} \\
&= \psi(\alpha + N + 1) + \frac{1}{\alpha + N + 1} \\
&\quad - \frac{\alpha + kd}{\alpha + N + 1}\psi(1-d) - \frac{d}{\alpha + N + 1}\psi(1-d) \\
&\quad - \frac{1}{\alpha + N + 1}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right) - \frac{1}{\alpha + N + 1}(n_{k+1} - d)\psi(n_{k+1} - d + 1),
\end{aligned}$$

ahora la idea es juntar algunos términos para que aparezca la entropía en N, en efecto

$$\begin{aligned}
&\hat{H}_{N+1,\neq} \\
&= \frac{\alpha + N + 1}{\alpha + N + 1}\psi(\alpha + N + 1) + \frac{1}{\alpha + N + 1} \\
&\quad - \frac{\alpha + N}{\alpha + N + 1}\frac{\alpha + kd}{\alpha + N}\psi(1-d) - \frac{d}{\alpha + N + 1}\psi(1-d) \\
&\quad - \frac{\alpha + N}{\alpha + N + 1}\frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right) - \frac{1}{\alpha + N + 1}(1-d)\psi(1-d+1) \\
&= \frac{\alpha + N}{\alpha + N + 1}\left[\psi(\alpha + N + 1) - \frac{\alpha + kd}{\alpha + N}\psi(1-d) - \frac{1}{\alpha + N}\left(\sum_{i=1}^{k}(n_i - d)\psi(n_i - d + 1)\right)\right] \\
&\quad + \frac{1}{\alpha + N + 1}\psi(\alpha + N + 1) + \frac{1}{\alpha + N + 1} \\
&\quad - \frac{d}{\alpha + N + 1}\psi(1-d) - \frac{1}{\alpha + N + 1}(1-d)\psi(1-d+1),
\end{aligned}$$

entonces

$$\begin{aligned}
&\hat{H}_{N+1,\neq} \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) + \frac{1}{\alpha+N+1} \\
&\quad - \frac{d}{\alpha+N+1}\psi(1-d) - \frac{1}{\alpha+N+1}(1-d)\psi(1-d+1).
\end{aligned}$$

El último término se puede expandir usando la recurrencia de la función digamma para obtener

$$\begin{aligned}
\frac{1}{\alpha+N+1}(1-d)\psi(1-d+1) &= \frac{1}{\alpha+N+1}(1-d)\left(\psi(1-d) + \frac{1}{1-d}\right) \\
&= \frac{1}{\alpha+N+1}(1-d)\psi(1-d) + \frac{1}{\alpha+N+1}.
\end{aligned}$$

Luego reemplazando

$$\begin{aligned}
&\hat{H}_{N+1,\neq} \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) + \frac{1}{\alpha+N+1} \\
&\quad - \frac{d}{\alpha+N+1}\psi(1-d) - \frac{1}{\alpha+N+1}(1-d)\psi(1-d) - \frac{1}{\alpha+N+1} \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) - \frac{1}{\alpha+N+1}\psi(1-d)(d+(1-d)) \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) - \frac{1}{\alpha+N+1}\psi(1-d).
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
(\alpha+N+1)\hat{H}_{N+1,\neq} - (\alpha+N)\hat{H}_N &= \psi(\alpha+N+1) - \psi(1-d) \\
\Delta_{N+1,\neq} &= \psi(\alpha+N+1) - \psi(1-d),
\end{aligned}$$

con lo cual queda demostrada la ecuación (4.4).

Demostremos ahora la ecuación (4.5). La demostración sigue la misma idea que el caso anterior, es decir, calcular la entropía esperada posterior $\hat{H}_{N+1,j}$ mediante la ecuación (2.14). Sea $j \in \{1,\ldots,k\}$, en este caso el tamaño de muestra también es $N+1$ pero el número de especies se mantiene en k y lo que cambia es la frecuencia para la especie X_j^* que pasa de n_j a ser n_j+1, y se mantienen las mismas frecuencias para el resto de las especies. Entonces, teniendo en cuenta el comportamiento de la especie j-ésima en la sumatoria, se tiene que la

entropía esperada posterior es

$$\begin{aligned}
&\hat{H}_{N+1,j} \\
&= \psi(\alpha+N+1+1) - \frac{\alpha+kd}{\alpha+N+1}\psi(1-d) \\
&\quad - \frac{1}{\alpha+N+1}\left(\sum_{i=1}^{k}(n_i-d)\psi(n_i-d+1)\right. \\
&\qquad\qquad \left. - (n_j-d)\psi(n_j-d+1) + (n_j+1-d)\psi(n_j+1-d+1)\right).
\end{aligned}$$

Para los primeros términos se trabaja de la misma forma que antes,

$$\begin{aligned}
&\hat{H}_{N+1,j} \\
&= \frac{\alpha+N+1}{\alpha+N+1}\psi(\alpha+N+1) + \frac{1}{\alpha+N+1} - \frac{\alpha+N}{\alpha+N+1}\frac{\alpha+kd}{\alpha+N}\psi(1-d) \\
&\quad - \frac{\alpha+N}{\alpha+N+1}\frac{1}{\alpha+N}\left(\sum_{i=1}^{k}(n_i-d)\psi(n_i-d+1)\right) \\
&\quad - \frac{1}{\alpha+N+1}\left(-(n_j-d)\psi(n_j-d+1) + (n_j+1-d)\psi(n_j+1-d+1)\right),
\end{aligned}$$

entonces

$$\begin{aligned}
&\hat{H}_{N+1,j} \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) + \frac{1}{\alpha+N+1} \\
&\quad - \frac{1}{\alpha+N+1}\left(-(n_j-d)\psi(n_j-d+1) + (n_j+1-d)\psi(n_j+1-d+1)\right).
\end{aligned}$$

Desarrollando el último término, usando la recurrencia de la función digamma

$$\begin{aligned}
&- (n_j-d)\psi(n_j-d+1) + (n_j+1-d)\psi(n_j+1-d+1) \\
&\quad = -(n_j-d)\psi(n_j-d+1) + (n_j+1-d)\left(\psi(n_j+1-d) + \frac{1}{n_j+1-d}\right) \\
&\quad = -(n_j-d)\psi(n_j-d+1) + (n_j-d)\psi(n_j+1-d) + \psi(n_j+1-d) + 1 \\
&\quad = \psi(n_j+1-d) + 1.
\end{aligned}$$

Luego, reemplazando

$$\begin{aligned}
&\hat{H}_{N+1,j} \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) + \frac{1}{\alpha+N+1} \\
&\quad - \frac{1}{\alpha+N+1}(\psi(n_j+1-d)+1) \\
&= \frac{\alpha+N}{\alpha+N+1}\hat{H}_N + \frac{1}{\alpha+N+1}\psi(\alpha+N+1) - \frac{1}{\alpha+N+1}\psi(n_j+1-d).
\end{aligned}$$

Así,

$$(\alpha+N+1)\hat{H}_{N+1,j}-(\alpha+N)\hat{H}_N=\psi(\alpha+N+1)-\psi(n_j+1-d)$$
$$\Delta_{N+1,j}=\psi(\alpha+N+1)-\psi(n_j+1-d),$$

por lo tanto queda demostrada la ecuación (4.5).

Finalmente, demostremos las desigualdades (4.6) y (4.7). Sea $j\in\{1,\ldots,k\}$, la primera desigualdad viene del hecho de que $1-d<n_j+1-d$ y como la función digamma es creciente, entonces $\psi(1-d)<\psi(n_j+1-d)$ y así

$$\psi(\alpha+N+1)-\psi(n_j+1-d)<\psi(\alpha+N+1)-\psi(1-d)$$
$$\Delta_{N+1,j}<\Delta_{N+1,\neq}.$$

La otra desigualdad es porque se tiene $n_j\leq n_{j^*}$, y así de forma similar a la primera desigualdad

$$\psi(\alpha+N+1)-\psi(n_{j^*}+1-d)\leq\psi(\alpha+N+1)-\psi(n_j+1-d)$$
$$\Delta_{N+1,j^*}\leq\Delta_{N+1,j}.$$

□

De este teorema tenemos una forma más directa de calcular la variación de la entropía esperada posterior para dos pasos consecutivos, en vez de calcular $\hat{H}_{N+1,j}$ o $\hat{H}_{N+1,\neq}$ y $\hat{H}_N$ para obtener la variación. Estas dos ecuaciones (4.4) y (4.5) son claramente muy similares, de hecho si se considera que en la muestra inicial de tamaño N la frecuencia de la (posible) nueva especie es $n_{k+1}=0$ entonces se podría considerar simplemente el valor $\psi(\alpha+N+1)-\psi(n_j+1-d)$ para $j=1,\ldots,k,k+1$ para el cálculo de la variación de la entropía, es decir, este valor solo depende de la frecuencia actualizada de la especie del nuevo elemento observado (y obviamente depende también de $d,\alpha,\ k$ y el tamaño N, que se están considerando fijos).

Además, las ecuaciones (4.4) y (4.5) son las versiones Bayesianas, en particular del Proceso de Poisson-Dirichlet, a la variación frecuentista dada por la ecuación (4.2). En ambas aparece el número de elementos de la especie actualizada, pero en este caso aparece la función digamma en vez del logaritmo. A pesar de que en este caso no se tiene una caracterización como en (4.3) para la aparición de una nueva especie, igual se obtiene una forma explícita de calcular la variación cuando aparece una nueva especie.

Con respecto a las desigualdades, por un lado, la ecuación (4.6) nos dice que, para una muestra de tamaño N, al observar un nuevo elemento, la variación de la entropía cuando este elemento es de una nueva especie es mayor al aumento de entropía en el caso que este elemento sea de una especie ya observada. Este comportamiento verifica la intuición, pues al agregar un elemento de una especie totalmente nueva este entrega en media más información (con respecto a la entropía) de la muestra que haber agregado un elemento de una especie conocida, o en otras palabras, se tiene más diversidad en la muestra al ver una nueva especie.

Por el otro lado, la ecuación (4.7) nos muestra que si se agrega un elemento a la especie que tiene la mayor cantidad de elementos, la cantidad de información o diversidad que se

agrega es menor en media que haberlo agregado a cualquiera de las otras especies conocidas. Lo cual también es algo esperable, ya que este caso no entrega mucha información adicional sobre la muestra.

Un corolario que se puede obtener del Teorema 4.1 son cotas para la variación de la entropía.

Corolario 4.1 *Considerando que inicialmente $n_{k+1} = 0$, entonces*

$$\Delta_{N+1,j} = \psi(\alpha + N + 1) - \psi(n_j + 1 - d), \quad \textit{para algún } j = 1, \ldots, k, k+1.$$

Se tiene que

$$\Delta_{N+1,j} \geq \ln(\alpha + N + 1) - \frac{1}{\alpha + N + 1} - \ln(n_j + 1 - d) + \frac{1}{2(n_j + 1 - d)} \tag{4.8}$$

$$\Delta_{N+1,j} \leq \ln(\alpha + N + 1) - \frac{1}{2(\alpha + N + 1)} - \ln(n_j + 1 - d) + \frac{1}{n_j + 1 - d}. \tag{4.9}$$

Demostración. Estas cotas se obtiene de las cotas de la función digamma dadas por (2.6), es decir,

$$\ln(x) - \frac{1}{x} \leq \psi(x) \leq \ln(x) - \frac{1}{2x}.$$

Con esto se pueden acotar ambos términos de $\Delta_{N+1,j}$,

$$\ln(\alpha + N + 1) - \frac{1}{\alpha + N + 1} \leq \psi(\alpha + N + 1) \leq \ln(\alpha + N + 1) - \frac{1}{2(\alpha + N + 1)},$$

y

$$-\ln(n_j + 1 - d) + \frac{1}{2(n_j + 1 - d)} \leq -\psi(n_j + 1 - d) \leq -\ln(n_j + 1 - d) + \frac{1}{n_j + 1 - d}.$$

Con lo cual se concluyen las cotas en (4.8) y (4.9). □

Estas cotas permiten determinar un intervalo en donde se encuentra la variación de la entropía $\Delta_{N+1,j}$ cuando no se tiene una forma directa de calcular la función digamma. Además, se tiene que el largo del intervalo es $\frac{1}{2(\alpha + N + 1)} + \frac{1}{2(n_j + 1 - d)}$, es decir, a medida que aumenta el tamaño de la muestra N (o la frecuencia n_j para $j = 1, \ldots, k$) el intervalo se hace más pequeño, por lo que se tiene mayor precisión para el valor de la variación de la entropía.

Por último, cuando el tamaño de la muestra N es suficientemente grande, se hace un análisis del comportamiento de la aproximación de la variación de la entropía.

Corolario 4.2 *Se tienen las siguientes aproximaciones para N y n_j suficientemente grandes para $j = 1, \ldots, k$*

$$\Delta_{N+1,\neq} \approx \ln(\alpha + N + 1) - \frac{1}{2(\alpha + N + 1)} \tag{4.10}$$

$$\Delta_{N+1,j} \approx \ln(\alpha + N + 1) - \frac{1}{2(\alpha + N + 1)} - \ln(n_j + 1 - d) + \frac{1}{2(n_j + 1 - d)}. \tag{4.11}$$

Demostración. Estas aproximaciones se obtienen de manera directa usando la aproximación de la función digamma dada por (2.7). □

Notamos de la ecuación (4.10) que $\Delta_{N+1,\neq}$ tiende a infinito cuando N tiende a infinito, es decir, cuando la muestra es extremadamente grande se tiene que la variación de la entropía al ver una nueva especie es también extremadamente grande, esto significa que observar una nueva especie en esta etapa entrega mucha información de la diversidad de la muestra. En cambio cuando N y n_j tienden a infinito (con n_j cercano a N) se tiene de la ecuación (4.11) que $\Delta_{N+1,j}$ tiende a 0, o sea en esta etapa la información que aporta observar un nuevo elemento de una especie conocida es nula.

Conclusión

El estudio del muestreo de especies y el proceso de aparición de nuevas especies ha sido estudiado por varios años, desde distintos enfoques y obteniendo diversos resultados. En este trabajo nos enfocamos en el modelo Bayesiano con el Proceso de Poisson-Dirichlet, revisando la definición y construcción de este, además de sus propiedades fundamentales y su interesante relación con el modelo de Good-Turing. Se dio mayor énfasis a la entropía del proceso y los tiempos entre apariciones, para comprender el comportamiento de aparición de nuevas especies.

Por un lado, gracias al estudio de la entropía en el contexto del Proceso de Poisson-Dirichlet, utilizando los resultados expuestos en [7], se obtuvo una cota superior de la entropía esperada posterior en la Proposición 2.5. Asimismo, se logró el objetivo principal de esta tesis, que corresponde a la relación y desigualdades para la variación de la entropía esperada posterior, dadas por el Teorema 4.1. Esta relación nos permite obtener de forma más directa la variación de la entropía para dos pasos consecutivos.

Además, gracias a las desigualdades obtenidas, podemos interpretar de mejor manera los valores que se obtienen de esta variación, y entender la cantidad de información o diversidad que aporta observar un nuevo elemento en la muestra. En particular, se tiene que al agregar un nuevo elemento de una especie que no se ha visto antes, este entrega en media más información (o diversidad) a la muestra que haber visto un elemento de una especie ya conocida. También, si se observa un nuevo elemento de la especie con la mayor frecuencia, entonces este entrega en media la menor cantidad de información que haber observado un elemento de cualquier otra especie.

Por otro lado, con el análisis realizado para la distribución de los tiempos entre apariciones, usando los resultados de [8] y [9], y las demostraciones para la distribución de Waring desarrolladas en la Sección 3.2, se logró obtener, y entender de dónde proviene, la relación para el tiempo esperado entre apariciones para el caso del Proceso de Poisson-Dirichlet de dos parámetros. Este resultado, dado por la Proposición 3.5, corresponde a una generalización del caso de un solo parámetro.

Los resultados obtenidos en esta tesis son de gran interés para entender de mejor manera el proceso de aparición de nuevas especies en el Proceso de Poisson-Dirichlet, sin embargo los resultados obtenidos son para parámetros d y α fijados previamente. Además, en el caso de la entropía, imponer una distribución prior puede dar un sesgo en la entropía esperada posterior, sobre todo cuando se tienen pocos datos, lo que significaría también un sesgo para la variación de la entropía.

Como posibles extensiones de este trabajo, se puede buscar una relación entre la entropía esperada posterior y la distribución de los tiempos entre apariciones o los tiempos de aparición. Debido a que el resultado principal de la variación de la entropía es para dos pasos consecutivos, otra posible extensión sería analizar la variación para un tamaño mayor en la siguiente muestra. Esto puede traer más dificultades, ya que con el enfoque usado en este trabajo, se necesitaría asumir la especie a la cual corresponde cada una de las nuevas observaciones. Para esto, se podría usar la probabilidad de descubrimiento (1.10) para un cierto tamaño $c \geq 1$ de la próxima muestra a observar, y los estimadores estudiados en esa sección. Esto significaría tener que promediar en las observaciones entre $N+1$ y $N+c+1$.

Finalmente, como estos resultados están enfocados en el Proceso de Poisson-Dirichlet desde la teoría, es interesante pensar cómo serían aplicados en la práctica con datos reales de comunidades ecológicas. Por ejemplo, dada una muestra, se puede calcular la entropía en cada momento que se observa un elemento de alguna especie −el problema de decidir si este elemento pertenece a una nueva especie, puede ser un problema difícil en algunas áreas (como en genética), lo cual requiere una investigación adicional profunda−, y estudiar como se comporta la variación de la entropía para una próxima observación. Si la variación de la entropía todavía es significativa, se puede decidir tomar una nueva muestra para ampliar la muestra original, y así tener más información sobre la diversidad de la muestra y, por lo tanto, de la comunidad que se desea estudiar. Para poder hacer esto, es necesario hacer una implementación estadística compleja, para entender qué representa una nueva especie dados los parámetros del modelo.

Bibliografía

[1] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

[2] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[3] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, pp. 855–900, 1997.

[4] S. Favaro, A. Lijoi, and I. Prünster, "A new estimator of the discovery probability," *Biometrics*, vol. 68, no. 4, pp. 1188–1196, 2012.

[5] N. Sharif-Razavian and A. Zollmann, "An overview of nonparametric bayesian models and applications to natural language processing," *Science*, pp. 71–93, 2008.

[6] S. Sosnovskiy, "On financial applications of the two-parameter Poisson-Dirichlet distribution," *arXiv 1501.01954*, 2015.

[7] E. Archer, I. M. Park, and J. W. Pillow, "Bayesian entropy estimation for countable discrete distributions," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2833–2868, 2014.

[8] H. Yamato, M. Sibuya, and T. Nomachi, "Ordered sample from two-parameter GEM distribution," *Statistics & probability letters*, vol. 55, no. 1, pp. 19–27, 2001.

[9] T. Huillet, "Unordered and ordered sample from Dirichlet distribution," *Annals of the Institute of Statistical Mathematics*, vol. 57, no. 3, pp. 597–616, 2005.

[10] S. Martínez A., "An elementary entropy inequality for a sequence of objects," 2021. Notes CMM.

[11] S. Favaro, B. Nipoti, and Y. W. Teh, "Rediscovery of Good–Turing estimators via bayesian nonparametrics," *Biometrics*, vol. 72, no. 1, pp. 136–145, 2016.

[12] P. Orbanz, "Lecture notes on bayesian nonparametrics," *Journal of Mathematical Psychology*, vol. 56, pp. 1–12, 2014.

[13] W. Buntine and M. Hutter, "A bayesian view of the Poisson-Dirichlet process," *arXiv 1007.0296*, 2012.

[14] Y. W. Teh, "Dirichlet process," 2010.

[15] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," *Encyclopedia of machine learning*, vol. 1, 2010.

[16] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[17] L. Al Labadi and M. Zarepour, "On simulations form the two-parameter Poisson-Dirichlet process and the Normalized Inverse-Gaussian process," *arXiv 1209.5359*, 2012.

[18] A. Dassios and J. Zhang, "Exact simulation of two-parameter Poisson-Dirichlet random variables," *Electronic Journal of Probability*, vol. 26, pp. 1–20, 2021.

[19] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[20] J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh, "Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptiotics," *Statistica Sinica*, pp. 839–858, 2017.

[21] A. Lijoi, R. H. Mena, and I. Prünster, "Bayesian nonparametric estimation of the probability of discovering new species," *Biometrika*, vol. 94, no. 4, pp. 769–786, 2007.

[22] E. Archer, I. M. Park, and J. Pillow, "Bayesian estimation of discrete entropy with mixtures of stick-breaking priors," *Advances in Neural Information Processing Systems*, vol. 25, pp. 2015–2023, 2012.

[23] S. Vajapeyam, "Understanding Shannon's entropy metric for information," *arXiv 1405.2061*, 2014.

[24] A. Chao and T.-J. Shen, "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample," *Environmental and ecological statistics*, vol. 10, no. 4, pp. 429–443, 2003.

[25] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover, 1972.

[26] H. Alzer, "On some inequalities for the gamma and psi functions," *Mathematics of computation*, vol. 66, no. 217, pp. 373–389, 1997.

[27] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," *Advances in Neural Information Processing Systems 14*, pp. 471–478, 2002.

[28] H. Yamato, "On the Donnelly-Tavare-Griffiths formula associated with coalescent," *Ann. Statist*, vol. 1, pp. 353–355, 1997.

[29] J. O. Irwin, "The generalized Waring distribution. Part I," *Journal of the Royal Statistical Society: Series A (General)*, vol. 138, no. 1, pp. 18–31, 1975.

[30] Z. Wang, "One mixed negative binomial distribution with application," *Journal of Statistical Planning and Inference*, vol. 141, no. 3, pp. 1153–1160, 2011.

[31] M. Sibuya, "Generalized hypergeometric, digamma and trigamma distributions," *Annals of the Institute of Statistical Mathematics*, vol. 31, no. 3, pp. 373–390, 1979.

[32] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate discrete distributions*, vol. 444. John Wiley & Sons, 2005.