Distribución de la riqueza de especies: una aproximación teórica y empírica mediante la distribución Beta.

Tamara Belén Rojas Leiva

14 de diciembre de 2022

Profesor Guía

Mauricio Tejo Arriagada Ph.D.

Instituto de Estadística, Universidad de Valparaíso

Proyecto de titulación para optar al:

grado académico de: *Magíster en Estadística*

Índice general

Agradecimientos **3**

Resumen **4**

Objetivos **5**

1. Contexto ecológico **6**
1.1. Dinámica de especies: conceptos básicos y contexto histórico del problema 6

2. Modelos estocásticos **10**
2.1. Cadenas de Markov 11
2.2. El proceso de Poisson 13
2.3. Procesos de nacimiento y muerte 15
2.3.1. Distribución de la riqueza de especies: el modelo de MacArthur. y Wilson . . 17
2.4. Procesos de difusión 18

3. El modelo propuesto **24**
3.1. El modelo 24
3.1.1. Conexión de (3.6) con el modelo de MacArthur y Wilson 26

4. Análisis estadístico de datos **27**
4.1. El modelo estadístico 27
4.2. Métodos de estimación de parámetros 28
4.2.1. Estimación de parámetros en la distribución beta 28
4.2.2. Estimación de parámetros en la distribución beta cero-inflada 30
4.2.3. Estimación de parámetros en la distribución beta-binomial 32
4.2.4. Prueba de bondad de ajuste de Kolmogórov-Smirnov 33
4.3. Análisis de datos 34
4.3.1. Análisis datos distribución beta y beta cero-inflada 34
4.3.2. Análisis datos distribución beta binomial 36
4.3.3. Conclusiones de los resultados obtenidos 38
4.4. Discusión y futuras consideraciones 38

Referencias **39**

Agradecimientos

Agradezco a profesor Mauricio Tejo quien aceptó trabajar conmigo y dedicó su tiempo en dirigir este estudio brindándome todo su apoyo.

A Rolando Rebolledo por su disponibilidad y colaboración durante todo este proceso.

A Pablo Marquet por facilitar las muestras con las que se desarrolló este estudio.

A Héctor Araya quien nos ayudó con la implementación de nuestra estimación del pool.

A mi familia por su confianza, por su apoyo incondicional y su constante motivación en los momentos que más lo necesité.

A todos aquellos que de una u otra manera influyeron en el desarrollo de este proceso entregándome su apoyo en cada momento.

Resumen

En este trabajo abordaremos una problemática de modelamiento probabilístico/estadístico en ecología. Se propondrá un modelo probabilístico para describir la dinámica de la riqueza de especies en metacomunidades, y luego desde allí se derivará un modelo estadístico para el ajuste de la distribución de la riqueza de especies en estado estacionario. Por tanto, el objetivo de este trabajo es proponer un modelo estadístico para el ajuste de la distribución de especies, desde argumentos teóricos probabilísticos, que a su vez se interpretan desde el punto de vista ecológico. Concretamente, se propondrá la distribución Beta como modelo para la distribución de la riqueza relativa de especies, la cual surge como la distribución estacionaria de un proceso de difusión en $[0, 1]$. Este proceso de difusión surge a su vez como una aproximación continua de un proceso de nacimiento y muerte que modela la proporción de especies en un espacio de estados discreto. La evaluación del ajuste del modelo se llevará a cabo haciendo uso de datos reales recientemente publicados, y en donde se analizarán distintas técnicas de estimación de parámetros y bondad de ajuste. Como resultado de la validación de este, se espera entregar información relevante desde el punto de vista ecológico, dada la vinculación de los parámetros de nuestro modelo con cantidades relevantes e interpretables desde aquel punto de vista.

Objetivos

Objetivo general

Proponer la distribución Beta como modelo para la distribución de la riqueza relativa de especies a partir de una aproximación por difusión de un proceso de nacimiento y muerte.

Objetivos específicos

1. Establecer las condiciones técnicas para aproximar el proceso de riqueza de especies modelado como un proceso de nacimiento y muerte hacia un proceso de difusión en $[0, 1]$.

2. Examinar las condiciones teóricas y ecológicas para que la distribución estacionaria del proceso de difusión resultante sea una Beta.

3. Estudiar distintos métodos de estimación de parámetros para la distribución Beta (estimación de momentos y estimación de máxima verosimilitud), y establecer métodos de estimación de parámetros para la distribución Beta-Binomial como modelo de riqueza absoluta.

4. Evaluar la bondad de ajuste de estos modelos en datos reales.

Capítulo 1

Contexto ecológico

En este capítulo se entregarán conceptos básicos y fundamentales en ecología de poblaciones, así como también se hará una revisión histórica acerca de algunos modelos en ecología matemática para la descripción de la dinámica de especies.

1.1. Dinámica de especies: conceptos básicos y contexto histórico del problema

La ecología matemática es la disciplina científica dedicada al estudio de sistemas ecológicos utilizando modelos matemáticos. Basados en suposiciones biológicamente razonables, el planteamiento de modelos matemáticos eficaces pueden contribuir al descubrimiento de conocimientos novedosos y no intuitivos sobre los procesos naturales, así como mejorar la comprensión del mundo natural al revelar las condiciones y los procesos biológicos fundamentales que inciden en dinámica de las poblaciones de especies.

Los resultados teóricos a menudo se pueden contrastar mediante estudios empíricos o de observación, logrando así evaluar la calidad del modelo teórico propuesto para describir, comprender e inferir el mundo ecológico, el cual es diverso, complejo y "ruidoso".

El objeto de estudio de la ecología matemática es principalmente la dinámica de poblaciones de especies, en sus distintos niveles de interacción y complejidad, aunque en muchas ocasiones esto va a acompañado de un fuerte ejercicio matemático/computacional o demostraciones matemáticas complejas que muestren o evidencien que los modelos matemáticos propuestos realmente describan lo que uno pretende que describan. En cuanto a las poblaciones de especies, podemos hacer distinción de tres niveles o escalas de enfoque anidadas, como se muestra en Figura 1.1.

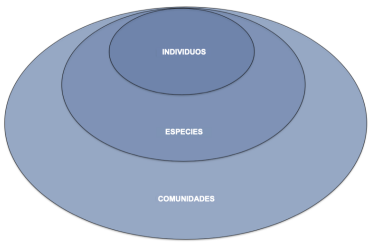


Figura 1.1: Relación anidada entre individuos, especies y comunidades. Figura obtenida desde (Marquet et al., 2020).

Dado un cierto conjunto de especies bajo estudio u observación sobre algún área geográfica delimitada (por ejemplo, el conjunto de animales vertebrados que viven en algún archipiélago determinado), se distingue a los individuos como las unidades representantes de cada especie, mientras que las comunidades representan subregiones del área geográfica delimitada que contienen (o que potencialmente pueden contener) a un conjunto de estas especies (del ejemplo anterior, cada isla del archipiélago representaría una comunidad). Al conjunto de todas estas comunidades que forman el área geográfica delimitada bajo estudio u observación se le conoce como metacomunidad (en el mismo ejemplo, esta correspondería al archipiélago).

Al número de individuos por especie se le conoce como abundancia de especies; mientras que al número de especies en la metacomunidad se le conoce como riqueza de especies. Sin embargo, en algunas ocasiones, y por motivo de la descripción matemática que se pueda realizar del fenómeno bajo estudio, se suelen identificar estos elementos de la metacomunidad utilizando graduaciones mas convenientes, como por ejemplo identificar la abundancia de especies con la biomasa de especies (cantidad total de materia viva presente en una comunidad), transformando la abundancia de una graduación discreta a una continua, como también considerar la riqueza relativa de especies en vez de la riqueza de especies; esto es, considerar la proporción de especies por comunidad con respecto al máximo de especies (asociadas a cierta taxa de interés) que pueden estar presentes en la metacomunidad, según sus condiciones ecológicas. A este número máximo se le conoce como el *pool* de especies.

Desde los trabajos pioneros de Humboldt, Darwin y Wallace, comprender los factores que impulsan los cambios en el número de especies que coexisten en las comunidades locales ha sido un objetivo importante de la ecología y la biogeografía. A pesar de esto, todavía no tenemos una teoría formal para describir los cambios en la riqueza de especies, ya que esta es una variable difícil de medir debido a los múltiples factores ecológicos que inciden en ella (Marquet et al., 2020).

Uno de los primeros intentos de derivar la probabilidad de observar s especies en una comunidad local fue el de Barton y David (1959) en el contexto de un modelo de urna simple. De hecho, se centraron en los cambios en el número de especies coexistentes como un proceso similar a la asignación de bolas de k colores diferentes entre N cajas idénticas. Aunque su principal objetivo era obtener una prueba para la asociación de especies, también derivaron la expresión para la ocurrencia esperada de s especies bajo el supuesto de independencia, donde la presencia de una especie es independiente de la presencia de otra especie.

En un contexto ecológico, diferentes colores corresponden a diferentes especies y las diferentes cajas a diferentes comunidades locales o muestras. Bajo el supuesto de que todas las especies se comportan independientemente unas de otras y tienen la misma abundancia (es decir, tienen la misma probabilidad de formar parte de una comunidad dada), $P(s)$ sigue una distribución binomial:

$$P(s) = \binom{K}{s} p^s (1-p)^{K-s} \quad ; s = 0, 1, ..., K \tag{1.1}$$

donde K corresponde al conjunto de especies disponibles para colonizar una comunidad local, p es la probabilidad de encontrar cualquier especie dada en una comunidad local y s es la riqueza de la comunidad local. Además, muestran que (1.1) es una buena aproximación siempre que el valor de p no varíe demasiado [ver Barton y David (1959), E. Pielou (1975)]. En caso contrario, como consecuencia de, por ejemplo, especies que tienen diferentes abundancias, entonces la distribución de la riqueza se puede aproximar mediante una distribución binomial con función generadora de probabilidad dada por:

$$G(z) = (Q + Pz)^\theta \quad \text{con} \quad P = \bar{p} + \frac{var(p)}{\bar{p}}, \quad Q = 1 - P, \quad \text{y} \quad \theta = \frac{K}{1 + \frac{var(p)}{\bar{p}^2}}, \tag{1.2}$$

donde $\bar{p} = (1/K)\sum_{i=1}^{K} p_i$ y $var(p) = (1/k)\sum_{i=1}^{K}(p_i - \bar{p})^2$.

Estas expresiones se han utilizado comúnmente para evaluar el número esperado de muestras o comunidades con diferentes números de especies para evaluar la asociación positiva o negativa entre especies en toda la comunidad [E. Pielou (1975), D. Pielou y Pielou (1967), Strong Jr (1982)]; es decir, un número inesperadamente grande de muestras o comunidades que contienen más o menos especies de las esperadas, respectivamente.

El segundo intento provino de MacArthur y Wilson (1963), (1967) bajo la "teoría de la biogeografía insular"(TBI), una de las primeras teorías neutrales en ecología comunitaria. Esta vez, sin embargo, el interés no fue proporcionar una descripción estadística de la expectativa de encontrar especies en una comunidad dada, sino comprender cómo varía la cantidad de especies según los procesos de colonización y extinción. Estos autores describieron el cambio en el número de especies que se encuentra en una sola isla y mostraron que esto se puede estudiar como un simple proceso de nacimiento/muerte. Para hacer esto, definieron $P_s(t)$ como la probabilidad que en el momento t una isla focal contenga s especies λ_s se define como la tasa de inmigración de nuevas especies a la isla (ya sea de la reserva o de otras islas) cuando s están presentes, y μ_s como la tasa de extinción de especies en la isla cuando hay s. La forma funcional de λ_s y μ_s corresponden a las curvas de intersección del modelo gráfico de TBI; es decir, λ_s es una función decreciente de s mientras que μ_s aumenta con s. Se supone que estas dos tasas son las mismas para cualquier especie. Por lo tanto, la probabilidad de observar s especies en una isla, o un conjunto de islas idénticas, sigue la ecuación maestra:

$$\frac{dP_s(t)}{dt} = P_{s-1}(t)\lambda_{s-1} + P_{s+1}(t)\mu_{s+1} - P_s(t)(\lambda_s + \mu_s), \tag{1.3}$$

donde s es el número de especies, λ_s es la tasa de colonización o especiación que aumenta de s a $s+1$ y μ_s es la tasa de extinción que disminuye de s a $s-1$. La ecuación (1.3) conduce a un estado estacionario o condición de equilibrio que satisface $P_s = P_0 \prod_{i=1}^{s} \frac{\lambda_{i-1}}{\mu_i}$. La distribución invariante del proceso, si existe, dependerá de la forma funcional de λ_s y μ_s (detalles de esto se verán más adelante).

Por otra parte, la teoría neutral de la ecología derivó la distribución de abundancia de especies utilizando el siguiente proceso de nacimiento/muerte (Volkov, Banavar, Hubbell, y Maritan, 2003) que expresa la probabilidad de que la k-ésima especie contenga n individuos:

$$\frac{dP_{n,k}(t)}{dt} = P_{n+1,k}(t)d_{n+1,k} + P_{n-1,k}(t)b_{n-1,k} - P_{n,k}(t)(d_{n,k} + b_{n,k}), \tag{1.4}$$

donde n es el número de individuos, b_n es la tasa de natalidad que aumenta de n a $n+1$ y d_n es la tasa de mortalidad que disminuye de n a $n-1$. Si asumimos que todas las especies tienen las mismas tasas de natalidad y muerte, y que estas son proporcionales a la densidad (es decir, $b_n = bn$ y $d_n = dn$), se obtiene que la distribución de equilibrio es la serie logarítmica de Fisher (Kendall, 1948). Observe que la ecuación (1.4) es idéntica en estructura a la ecuación (1.3). Sin embargo, ambos modelos no pueden ser "verdaderos" a la vez. En efecto, tenemos que la riqueza de especies, $S(\cdot)$, es un observable de la abundancia de especies $\{N_k(\cdot)\}_{k=1,...,K}$:

$$S(t) = \sum_{k=1}^{K} 1_{(0,\infty)} N_k(t),$$

y luego,

$$\{S(t) = s\} = \bigcup_{k_1,...,k_s} \{(N_{k_1}(t), ..., N_{k_s}(t)) \in (0,\infty)^s\}.$$

Por lo tanto, bajo (1.4), $S(\cdot)$ ya no es Markoviano y, por lo tanto, sus estados no pueden seguir la ecuación maestra (1.3) (las consideraciones técnicas se verán en el capítulo siguiente).

Curiosamente, MacArthur y Wilson (1963), (1967) no estaban interesados en obtener explícitamente una distribución estacionaria, que sería nuestra distribución de riqueza de especies, ni tampoco la teoría neutral de Hubbell, aunque era consciente del problema de conectar la abundancia con el número de especies en una comunidad determinada (Hubbell, 2011). Para resolver la ecuación (1.3) es necesario especificar una condición inicial P_0 y algunas condiciones de contorno, que en este caso se pueden deducir al notar que cuando no hay especies presentes en la isla, $\mu_0 = 0$ y $\lambda_0 \neq 0$, y cuando la isla está saturada de especies, es decir, el número de especies es igual al *pool* K, entonces $\mu_K \neq 0$ y $\lambda_K = 0$. Por lo tanto, el proceso estocástico asociado con el número de especies está confinado entre los dos estados reflectantes 0 y K. Una vez establecidas estas condiciones, el paso crucial es definir la forma funcional de las tasas de extinción y colonización.

Goel y Richter-Dyn (2016) consideraron dos escenarios para estas tasas: (1) $\lambda_s = \lambda(K - s)$ y $\mu_s = \mu s$ (2) $\lambda_s = \lambda(K - s)^2$ y $\mu_s = \mu s^2$. El primer escenario considera que debido a que el área y, por lo tanto, la cantidad de recursos presentes en la isla son fijos, y a medida que aumenta el número de especies en la isla el tamaño promedio de la población de cualquier especie dada disminuye proporcionalmente. De manera similar, es razonable suponer que la probabilidad que una nueva especie esté presente en la isla dependa de las especies ya presentes, porque cuantas más especies se establezcan en la isla menores serán las posibilidades de que un nuevo individuo inmigrante pertenezca a una nueva especie, por lo tanto, la tasa de inmigración debería disminuir monótonamente a medida que aumenta el número de especies establecidas en la isla, obteniendo así $\lambda_K = 0$. El escenario 2 considera las dependencias no lineales en ambos colonización y extinción apelando al hecho que cuando s es pequeño es probable que el número de especies que colonizan la isla sea mayor, ya que puede que las nuevas especies en dispersión lleguen muy rápido, lo que provocará una mayor disminución inicial en la tasa de inmigración, que más tarde se nivela a medida que aumenta s. De esta manera, la extinción puede no ser la misma para cada especie como se supuso en el escenario 1, pero probablemente aumentará a medida que aumente el número de especies debido a interacciones interespecíficas negativas [MacArthur y Wilson (1963), (1967)].

Como muestran Goel y Richter-Dyn (2016) cuando $t \to \infty$ en el escenario 1, $P_s(t)$ tiende a

$$P_s = \binom{K}{s} \left(\frac{\lambda}{\lambda+\mu}\right)^s \left(\frac{\mu}{\lambda+\mu}\right)^{K-s}; s = 0, 1, ..., K \tag{1.5}$$

La ecuación (1.5) es la misma que la ecuación (1.1) donde la probabilidad de éxito p corresponde a la probabilidad de que ocurra un evento de inmigración y $1 - p$ corresponde a la probabilidad de un evento de extinción. Para el caso del escenario 2, la distribución de equilibrio o de estado estacionario es:

$$P_s^* = \frac{\binom{K}{s}^2 \left(\frac{\lambda}{\mu}\right)^s}{\sum_{s=0}^{K} \binom{K}{s}^2 \left(\frac{\lambda}{\mu}\right)^s} \tag{1.6}$$

En este trabajo de tesis ahondaremos en este tipo de modelos, proponiendo una generalización de tales, lo cual implicará un uso de herramientas más avanzadas de procesos estocásticos. Es por ello que antes de comenzar con la descripción de nuestra propuesta, en el siguiente capítulo daremos una breve descripción de los principales elementos técnicos de procesos estocásticos a utilizar.

Capítulo 2

Modelos estocásticos

En este capítulo haremos una revisión de los modelos estocásticos Markovianos que consideraremos para construir nuestro modelo de dinámica de especies.

Muchos sistemas biológicos cambian de un estado a otro con el tiempo. Por ejemplo, los nervios cambian de inactivos a excitados y viceversa, las células cambian de sanas a enfermas, o una población de plantas reemplaza a otra. Si bien los cambios entre estados pueden ser inciertos, no obstante, se pueden asignar probabilidades de transición de un estado al siguiente. Si conocemos las probabilidades de transición entre estados, podemos evaluar los cambios en el sistema a lo largo del tiempo.

Para la formulación matemática de un modelo con resultados inciertos, primero se definen las cantidades matemáticas que entrarán en el modelo. Si Ω es un espacio muestral (es decir, la colección de todos los resultados posibles de un experimento), $\mathcal{F}$ es una sigma-álgebra (colección de subconjuntos medibles), P una medida de probabilidad, y X es una función de valor real definida sobre los elementos de Ω, entonces X es una variable aleatoria definida sobre el espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Por ejemplo, si X fuera la longitud a la horquilla de un pez capturado en centímetros, entonces Ω podríamos tomarlo como el conjunto $\mathbb{R}_+$ y $\mathcal{F}$ su conjunto de Borel.

Podemos seguir el cambio en una variable aleatoria a medida que aumenta un parámetro, como el tiempo. Una familia de variables aleatorias $\{X(t)\}$, indexadas por un parámetro t, se denomina proceso estocástico. Comenzaremos este capítulo con un ejemplo de un proceso estocástico "sin memoria" (proceso de Markov), llamado cadena de Markov. Luego, avanzaremos hacia el estudio de procesos Markovianos más elaborados que nos permiten abarcar el modelamiento de una mayor generalidad de fenómenos. Podemos clasificar estos en procesos o modelos a tiempo discreto o a tiempo continuo, como también será de relevancia clasificar el espacio de estados del proceso bajo estudio, el cual hará referencia a que si este tomará valores en conjuntos finitos, contables o no numerables.

La hipótesis de "no memoria" o de Markovianidad estará presente durante todo este capítulo, la cual nos ofrecerá un muy elaborado y estudiado cuerpo teórico para desarrollar modelos matemáticos estocásticos. Aunque esta hipótesis de "no memoria" o de Markovianidad podría ser cuestionable, la ecología matemática se ha basado en general en expresar estos tipos de modelos, resultando en muchos casos bastante razonables en la descripción de los fenómenos que se pretender representar.

2.1. Cadenas de Markov

Dentro de los procesos estocásticos más simples que se pueden concebir están aquellos que pueden caracterizarse completamente por su estado actual, y donde los estados pasados de una variable no afectan los resultados futuros. Un proceso estocástico $\{X(t)\}$ se denomina proceso de Markov si este es independiente de su pasado dado su estado actual. Si tenemos una secuencia de tiempos discretos $... < t_{i-1} < t_i < t_{i+1} < ...$, entonces un proceso es de Markov si este satisface:

$$P(X(t_{i+1}) \in A \mid X(t_i) = x_i, X(t_{i-1}) = x_{i-1}, ...) = P(X(t_{i+1}) \in A \mid X(t_i) = x_i),$$

donde A es un evento cualquiera. Es decir, la probabilidad de cualquier evento futuro es independiente de su "historia" dado el estado actual. Por ejemplo, si suponemos que una partícula va saltando de unidad en unidad por la línea recta de acuerdo al lanzamiento de una moneda, donde "cara" significa un desplazamiento de una unidad hacia la derecha y "sello" un desplazamiento de una unidad hacia la izquierda, entonces la probabilidad que la partícula en el lanzamiento "$n+1$-ésimo" se encuentre en cierta posición de la recta dado su historial de posiciones anteriores, solo dependerá de la posición actual en el n-ésimo instante.

Clásicamente una cadena de Markov es un modelo que describe la progresión de un proceso de Markov de un paso de tiempo a otro sobre un conjunto de estados finito. Para ilustrar este modelo, podemos considerar el siguiente ejemplo extraído desde De Vries, Hillen, Lewis, Müller, y Schönfisch (2006). Considere una población compuesta por roble rojo y nogal, en que en cualquier punto espacial el espacio muestral de posibles resultados es $\Omega = \{RO, HI\}$, donde RO representa roble rojo (abreviatura del nombre en inglés *red oak*) y HI representa nogal (abreviatura del nombre en inglés *hickory*). Suponemos que la vida útil de los dos árboles es similar. En cada generación, el roble rojo puede ser reemplazado por sí mismo o por un nogal, y el nogal puede ser reemplazado por sí mismo o un roble rojo. Este es un proceso de Markov, con el índice t indicando la generación. Si suponemos que cuando un roble rojo muere, es igualmente probable que sea reemplazado por un nogal o un roble rojo, y que cuando muere un nogal, tiene 0,74 de probabilidad de ser reemplazado por un roble rojo y 0,26 de ser reemplazado por un nogal. Estas transiciones se expresan mediante una matriz de transición $\boldsymbol{P} = (p_{ij})$, la cual es una matriz estocástica (es decir, que los elementos de cada columna suman 1, $\sum_i p_{ij} = 1$) cuyos componentes expresan estas cantidades:

$$\boldsymbol{P} = \begin{pmatrix} 0,5 & 0,74 \\ 0,5 & 0,26 \end{pmatrix}.$$

Así $p_{11} = 0,5$ y $p_{21} = 0,5$ son las probabilidades que un roble rojo sea reemplazado por un roble rojo y por un nogal, respectivamente, y $p_{12} = 0,74$ y $p_{22} = 0,26$ son las probabilidades que un nogal sea reemplazado por un roble rojo y por un nogal, respectivamente.

Para seguir los cambios en el sistema anterior a lo largo del tiempo, definimos un vector $\boldsymbol{u}_t = (o_t, h_t)^T$ que describe la probabilidad de la presencia de roble rojo y nogal en un lugar determinado del bosque después de t generaciones. En el caso de un bosque grande y homogéneo, el mismo modelo de transición se aplicaría en todos los puntos del espacio. Por tanto, o_t y h_t pueden interpretarse como las proporciones de roble rojo y nogal en un gran ecosistema forestal estadísticamente homogéneo. Si suponemos que el bosque es inicialmente 50 % de roble rojo y 50 % de nogal, entonces $\boldsymbol{u}_0 = (0,5, 0,5)^T$. Para encontrar $\boldsymbol{u}_1$, calculamos de la siguiente manera, utilizado probabilidad total:

o_1 = Proporción de roble rojo al tiempo 0 $\times$ probabilidad que un roble rojo sea reempalzado por un roble rojo

+ Proporción de nogal al tiempo 0 $\times$ probabilidad que un nogal sea reemplazado por un roble rojo.
$= 0,5 \times 0,5 + 0,5 \times 0,74 = 0,62$;

h_1 = Proporción de roble rojo al tiempo 0 $\times$ probabilidad que un roble rojo sea reemplazado por un nogal + Proporción de nogal al tiempo 0 $\times$ probabilidad que un nogal sea reemplazado por nogal.
$= 0,5 \times 0,5 + 0,5 \times 0,26 = 0,38$.

En términos matriciales, esto se puede escribir como:

$$\boldsymbol{u}_1 = \boldsymbol{P}\boldsymbol{u}_0.$$

De manera análoga, podemos observar que:

$$\boldsymbol{u}_2 = \boldsymbol{P}\boldsymbol{u}_1 = \boldsymbol{P}^2\boldsymbol{u}_0,$$

$$\boldsymbol{u}_3 = \boldsymbol{P}\boldsymbol{u}_2 = \boldsymbol{P}^3\boldsymbol{u}_0,$$

etc.

Decimos que el bosque ha alcanzado un equilibrio $\boldsymbol{u}^*$ si:

$$\boldsymbol{P}\boldsymbol{u}^* = \boldsymbol{u}^*.$$

Notar acá que $\boldsymbol{u}^*$ es un vector propio correspondiente al valor propio $\lambda = 1$. Luego, para calcular el vector propio debemos resolver:

$$(\boldsymbol{P} - I)\boldsymbol{u}^* = \boldsymbol{0} \Longleftrightarrow \begin{pmatrix} -0,5 & 0,74 \\ 0,5 & -0,74 \end{pmatrix} \begin{pmatrix} o^* \\ h^* \end{pmatrix} = \boldsymbol{0},$$

cuya solución es $o^* = 0,597$ y $h^* = 0,403$.
Estos resultados se pueden extender sin problemas para mayores dimensiones, considerando una matriz de transición $\boldsymbol{P} = (p_{ij})$, con $i, j = 1, ..., n$, y donde los p_{ij}'s representan la probabilidad de transición de un estado j a un estado i, con n estados posibles del sistema, y tales que $\sum_i p_{ij} = 1$, para todo $j = 1, ..., n$. Luego, si el vector n-dimensional $\boldsymbol{u}_t$ representa el estado del sistema al tiempo t (es decir, cada componente indica la probabilidad de la presencia o manifestación de cierta característica o cualidad inserto en algún ambiente al tiempo t, y son tales que la suma de las componentes es 1 para todo t), entonces el modelo general de cadena de Markov queda caracterizado por el sistema dinámico a tiempo discreto:

$$\boldsymbol{u}_{t+1} = \boldsymbol{P}\boldsymbol{u}_t,$$

dado un estado inicial $\boldsymbol{u}_0$. La garantía de la existencia de un equilibrio o estado estacionario como distribución límite queda establecido en el siguiente teorema.

Teorema 2.1.1 *Si alguna potencia de $\boldsymbol{P}$ tiene todas las entradas positivas, entonces, para cualquier vector de probabilidad $\boldsymbol{u}_0$, se tiene que, dado el modelo $\boldsymbol{u}_{t+1} = \boldsymbol{P}\boldsymbol{u}_t$, $\boldsymbol{u}_t \to \boldsymbol{u}^*$ cuando $t \to \infty$, donde $\boldsymbol{u}^*$ satisface $\boldsymbol{P}\boldsymbol{u}^* = \boldsymbol{u}^*$.*

El requisito de que alguna potencia de $\boldsymbol{P}$ tenga todas las entradas positivas ($\boldsymbol{P}$ es primitiva) asegura que, dados suficientes pasos de tiempo, uno pueda pasar de cualquier estado a cualquier otro estado y, por lo tanto, el resultado es independiente del estado original $\boldsymbol{u}_0$. Por otro lado, es sabido que no todas las distribuciones estacionarias son distribuciones límites. Por ejemplo, en un modelo con dos estados, si consideramos la matriz de transición

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

tenemos que el sistema tiene estado estacionario $\boldsymbol{u}^* = (0,5, 0,5)^T$. Sin embargo, $\boldsymbol{P}^t$ no tiene comportamiento límite, ya que esta va rotando de una potencia a otra entre P y la identidad I, lo cual ilustra un comportamiento periódico de la cadena de Markov.

Comentarios importantes:

- Tal como hemos definido el modelo de cadena de Markov, consideramos que la matriz de transición $\boldsymbol{P}$ es igual entre cualquier par de generaciones sucesivas. Sin embargo, existen modelos de cadena de Markov en la cual esta matriz de transición puede depender del instante en la cual se realiza la transición. Por ejemplo, para cualquier par de estados (i, j) y cualquier instante t, tenemos que $p_{ij} = P(X(t + 1) = i \mid X(t) = j)$. Si esta cantidad no depende del instante t (como en el caso estudiado), decimos que la cadena de Markov es (temporalmente) homogénea. En caso contrario, diremos que esta es una cadena de Markov no homogénea o inhomogénea, y los resultados del teorema anterior no aplican.

- Decimos que una cadena de Markov es irreducible si desde cualquier estado se puede acceder a otro. Por otro lado, podemos definir el periodo de un estado j de una cadena de Markov homogénea como el número mínimo de pasos que se requiere para acceder de j a j. Así, si este número de pasos es 1, y esto se satisface para todos los estado del sistema, decimos que la cadena de Markov es aperiódica. Luego, el teorema anterior se puede re-escribir como: Si una cadena de Markov homogénea con matriz de transición $\boldsymbol{P}$ es irreducible y aperiódica, entonces para cualquier vector de probabilidad $\boldsymbol{u}_0$, se tiene que, dado el modelo $\boldsymbol{u}_{t+1} = \boldsymbol{P}\boldsymbol{u}_t$, $\boldsymbol{u}_t \to \boldsymbol{u}^*$ cuando $t \to \infty$, donde $\boldsymbol{u}^*$ satisface $\boldsymbol{P}\boldsymbol{u}^* = \boldsymbol{u}^*$. La demostración de este resultado se deriva del teorema de Perron-Frobenius.

- Se definen también cadenas de Markov con espacio de estados contables, como también cadenas de Markov a tiempo continuo, siendo en estos últimos casos usualmente denominados como procesos Markovianos de saltos. A los procesos más generales que poseen espacios de estados continuos los consideraremos dentro de la generalidad de los procesos de Markov que veremos más adelante.

2.2. El proceso de Poisson

A continuación, introduciremos un primer proceso estocástico Markoviano a tiempo continuo, que es de amplio uso dentro del modelamiento estocástico: el proceso de Poisson.

Supongamos que estamos interesados en modelar los tiempos de llegadas de insectos polinizadores a una planta floreciente. Sean T_1, T_2, T_3,..., los tiempos de llegadas en orden secuencial. Luego, tenemos que los tiempos entre llegadas estan dados por T_1, $T_2 - T_1$, $T_3 - T_2$,.... Asumamos que tales tiempos entre llegadas son variables aleatorias independientes e idénticamente distribuidas, con

distribución exponencial de parámetro λ. Es posible mostrar que el tiempo de la n-ésima llegada tiene una distribución de probabilidad cuya densidad está dada por:

$$f_{n,\lambda}(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} \exp\{-\lambda x\},\ x > 0 \tag{2.1}$$

la cual corresponde a la función de densidad gamma de parámetros n y λ. En efecto, esta aparece del siguiente hecho: tenemos que el tiempo de la n-ésima se puede escribir como $T_n = T_1 + (T_2 - T_1) + (T_3 - T_2) + ... + (T_n - T_{n-1})$; es decir, mediante la suma de n variables aleatorias independientes e idénticamente distribuidas, con distribución exponencial de parámetro λ. Así, notar que en particular $f_{1,\lambda}(x)$ no es mas que función de densidad exponencial de parámetro λ.

Ahora, sea $\{N(t)\}$ el proceso que cuenta el número de insectos que ha llegado a la planta al tiempo t. Luego, se tiene que:

$$\begin{aligned}
&P(N(t) = 0) = P(T_1 > t) = \exp\{-\lambda t\}.\\
&P(N(t) < 2) = P(T_2 > t) = \int_t^\infty f_{2,\lambda}(x)dx = \lambda t \exp\{-\lambda t\} + \exp\{-\lambda t\}.\\
&P(N(t) < 3) = P(T_3 > t) = \int_t^\infty f_{3,\lambda}(x)dx = \frac{(\lambda t)^2}{2!}\exp\{-\lambda t\} + \lambda t \exp\{-\lambda t\} + \exp\{-\lambda t\}.\\
&\qquad\vdots\\
&P(N(t) < n) = P(T_n > t) = \sum_{k=0}^{n-1} \exp\{-\lambda t\}\frac{(\lambda t)^k}{k!} =: F_{N(t)}(n-1),
\end{aligned}$$

donde $F_{N(t)}(n-1) = P(N(t) \leq n-1)$ corresponde a la función de distribución acumulada de una variable con distribución de Poisson de parámetro λt.

Así, si los tiempos entre llegadas son independientes e idénticamente distribuidos, con distribución exponencial de parámetro λ, entonces el número de llegadas hasta el tiempo t sigue una distribución de Poisson de parámetro λt,

$$\mathbb{P}(N(t) = n) = \exp\{-\lambda t\}\frac{(\lambda t)^n}{n!},$$

$n = 0, 1, 2, ...$ Se tiene entonces que su esperanza, $E(N(t))$, y varianza, $V(N(t))$, están dadas por $E(N(t)) = \lambda t = V(N(t))$.

Algunas propiedades importantes de este proceso son:

- Para $s < t$, $N(s)$ y $N(t) - N(s)$ son independientes.
- Para un intervalo "pequeño" de tiempo Δ, la probabilidad que una llegada ocurra en $(t, t+\Delta]$ es aproximadamente proporcional al largo del intervalo (Δ),

$$\begin{aligned}
\lim_{\Delta\to 0} \frac{P(N((t,t+\Delta]) = 1)}{\Delta} &= \lim_{\Delta\to 0} \frac{1}{\Delta}\exp\{-\lambda\Delta\}\lambda\Delta = \lambda,\\
\lim_{\Delta\to 0} \frac{P(N((t,t+\Delta]) > 1)}{\Delta} &= 0,
\end{aligned}$$

  donde $N((t, t+\Delta]) = N(t+\Delta) - N(t) \sim Poisson(\lambda\Delta)$.

En la sección siguiente se introducen los procesos de nacimiento y muerte con los que generalizamos los procesos de saltos markovianos.

2.3. Procesos de nacimiento y muerte

Las poblaciones están sujetas a dos tipos principales de estocasticidad. La estocasticidad ambiental se refiere a la variación e incertidumbre en las condiciones ambientales en las que se encuentra una población. Estas condiciones incluyen efectos de temperatura, lluvia, competencia de otras especies, etc. La estocasticidad demográfica se refiere a la variación e incertidumbre que surgen del comportamiento impredecible de los individuos que componen una población. Esto último es relevante cuando el tamaño de la población es "pequeño". Aquí, las poblaciones con una tasa de crecimiento neto positivo aún pueden extinguirse debido a una "racha de mala suerte", en la que no se reproducen suficientes individuos antes de morir. En esta sección, consideramos cómo modelar la estocasticidad demográfica en tiempo continuo utilizando un modelo de nacimiento y muerte. Aquí se supone que los individuos actúan independientemente unos de otros, por lo que no hay términos de interacción no lineal en las ecuaciones.

Usualmente se considera partir el estudio de estos modelos con un modelo lineal de nacimiento puro, esto es, despreciando la muerte. La linealidad hace referencia a la linealidad de la tasa de ocurrencia (en este caso, la tasa de nacimiento) con respecto al tiempo. Sin embargo, este tipo de modelo, donde el primer nacimiento da origen al primer individuo de la población, corresponde al proceso de Poisson visto en la sección anterior, así que los modelos de nacimiento y muerte que veremos a continuación corresponden a extensiones naturales de este proceso.

Cuando ampliamos el análisis de la sección anterior a una población de individuos que nacen a una tasa b y mueren a una tasa d, las transiciones para n individuos en un intervalo de tiempo de duración Δ (pequeño) quedan descritas por:

$$\begin{aligned} P(\text{1 nacimiento en } [t, t+\Delta]) &= nb\Delta + o(\Delta), \\ P(\text{1 muerte en } [t, t+\Delta]) &= nd\Delta + o(\Delta), \\ P(\text{0 nacimiento o muerte en } [t, t+\Delta]) &= 1 - n(b+d)\Delta + o(\Delta), \end{aligned}$$

y la probabilidad de tener más de un nacimiento o una muerte en $[t, t+\Delta]$ es de $o(\Delta)$. Luego, según lo anterior, si definimos por $p_n(t+\Delta)$ como la probabilidad de contar con una población de n individuos al instante $t+\Delta$, tenemos que:

$$p_n(t+\Delta) = (n-1)b\Delta p_{n-1}(t) + (n+1)d\Delta p_{n+1}(t) + [1 - n\Delta(b+d)]p_n(t) + o(\Delta),$$

por lo que

$$\lim_{\Delta \to 0} \frac{p_n(t+\Delta) - p_n(t)}{\Delta} = \frac{dp_n(t)}{dt} = (n-1)bp_{n-1}(t) + (n+1)dp_{n+1}(t) - n(b+d)p_n(t). \tag{2.2}$$

A esta última ecuación se le conoce como ecuación maestra de un proceso de nacimiento y muerte lineal. Usualmente se le denomina ecuación maestra a ecuaciones de este tipo, que describen las probabilidades de transición de un proceso Markoviano de saltos. Para que queden estas determinadas, se debe contar con una condición inicial $p_n(0)$ dada. Es decir, si inicialmente el tamaño de la población es de $n_0 > 0$ individuos, entonces $p_n(0) = \delta_{n_0}$.

Este modelo puede ser generalizado considerando funciones de tasas de manera más general, pudiendo permitir, entre otras cosas, relaciones no lineales entre las tasas y el tamaño de la población. En efecto, pudimos haber derivado la ecuación maestra anterior considerando ahora que la relación entre las tasas de nacimiento y muerte y los tamaños de población se representan de manera más general como b_n y d_n, respectivamente. Es decir, en el caso lineal, la tasa de nacimiento en una población de n individuos a $n+1$ individuos es $b_n = nb$, y la tasa de muerte en una población de n

individuos a $n-1$ individuos es $d_n = nd$. Luego, podemos re-plantear la ecuación (2.2), ahora de manera más general como:

$$\frac{dp_n(t)}{dt} = b_{n-1}p_{n-1}(t) + d_{n+1}p_{n+1}(t) - (b_n + d_n)p_n(t). \tag{2.3}$$

Usando esta última expresión, derivaremos a continuación el promedio y la varianza de un proceso de nacimiento y muerte, como también su distribución o ley estacionaria.

Usando la misma notación que en el proceso de Poisson, ahora en este caso más general, definamos por $N(t)$ el número de individuos presentes en la comunidad focal al tiempo t. Luego, tenemos que el k-ésimo momento está dado por:

$$E(N(t)^k) = \sum_n n^k p_n(t).$$

Por un lado, tenemos que cuando $k = 1$, lo anterior satisface la ecuación diferencial:

$$\frac{dE(N(t))}{dt} = [\sum_n nb_{n-1}p_{n-1} - \sum_n nb_n p_n(t)] + [\sum_n nd_{n+1}p_{n+1} - \sum_n nd_n p_n(t)] = \sum_n b_n p_n(t) - \sum_n d_n p_n(t),$$

bajo la condición inicial $E(N(0)) = n_0$. Por otro lado, se puede mostrar que la ecuación del k-ésimo momento está dada por:

$$\frac{dE(N(t)^k)}{dt} = \sum_n [(n+1)^k - n^k]b_n p_n(t) - \sum_n [n^k - (n-1)^k]d_n p_n(t),$$

bajo la condición inicial $E(N(0)^k) = n_0^k$. Así, la varianza del proceso queda dada entonces por $V(N(t)) = E(N(t)^2) - E(N(t))^2$.

En Goel y Richter-Dyn (2016), se pueden hallar algunos resultados explícitos de medias y varianzas de algunos modelos de nacimiento y muerte particulares, los cuales surgen de acuerdo a cómo se escogen las funciones de tasas b_n y d_n. Podemos notar que en el caso lineal, la ecuación diferencial para la esperanza queda reducida a:

$$\frac{dE(N(t))}{dt} = (b-d)E(N(t)),$$

cuya solución entonces queda dada por $E(N(t)) = n_0 e^{(b-d)t}$. El cálculo de la varianza es más complicado, pero se puede obtener utilizando la función generadora de probabilidad ver [Sección 5.6.2 de De Vries et al. (2006) para su cálculo explícito]. Esta está dada por $V(N(t)) = \frac{n_0 e^{(b-d)t}[b/d+1][e^{(b-d)t}-1]}{b/d-1}$, si $b/d \neq 1$.

Ahora bien, la estabilidad de un proceso de nacimiento y muerte se estudia a través de la distribución límite de (2.3), la cual nos da origen a la distribución o ley estacionaria del proceso $\lim_{t\to\infty} p_n(t) = p_n^*$. Como esta ya no depende del tiempo, se puede hallar igualando (2.3) a cero, cuya solución resultante será aquella ley estacionaria. Para hallarla, primero debemos establecer algunas condiciones de borde adicionales con respecto a las tasas.

Cuando en una comunidad local, en un cierto instante, no contamos con individuos de nuestra población objetivo, podemos considerar aún $b_0 > 0$, cantidad que se interpreta como la tasa de

inmigración de un individuo, o tasa de colonización, si nuestra población objetivo es la riqueza de especies. Luego, la cantidad b_n puede interpretarse en general como la tasa de nacimiento/colonización en una población con n elementos. Así también, por otro lado, d_n puede interpretarse en general como la tasa de muerte o extinción de una especie si la población objetivo es la riqueza de especies, en una población con n elementos. Por tanto, sumaremos como condiciones de borde de la ecuación (2.3), $b_0 > 0$ y $d_0 = 0$.
Bajo estas condiciones, entonces procedemos a obtener la distribución estacionaria de un proceso de nacimiento y muerte, igualando (2.3) a cero:

$$0 = b_{n-1}p^*_{n-1} + d_{n+1}p^*_{n+1} - (b_n + d_n)p^*_n \iff d_{n+1}p^*_{n+1} - d_n p^*_n = b_n p^*_n - b_{n-1}p^*_{n-1}.$$

Tomando la suma en esta última ecuación, por todos los estados hasta un estado $i-1 \geq 0$, tenemos que:

$$d_i p^*_i = b_{i-1}p^*_{i-1} \iff p^*_i = \frac{b_{i-1}}{d_i}p^*_{i-1},$$

lo cual de manera recursiva obtenemos

$$p^*_i = p^*_0 \prod_{n=1}^{i} \frac{b_{n-1}}{d_n}, \tag{2.4}$$

donde p^*_0 se obtiene a partir de la condición de normalización $\sum_n p^*_n = 1$,

$$1 = \sum_n p^*_n = p^*_0 + p^*_0 \sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{b_{n-1}}{d_n} \iff p^*_0 = \left(1 + \sum_{i=1}^{\infty} \prod_{n=1}^{i} \frac{b_{n-1}}{d_n}\right)^{-1}.$$

2.3.1. Distribución de la riqueza de especies: el modelo de MacArthur. y Wilson

A continuación estudiaremos como un caso especial de un proceso de nacimiento y muerte, el modelo de riqueza de especies de MacArthur y Wilson (1963), (1967), el cual bajo su "teoría de biogeografía insular", presenta uno de los primeros modelos establecidos bajo la teoría neutral en comunidades ecológicas. La teoría neutral en comunidades ecológicas establece que los individuos de todas las especies de algún determinado nivel trófico, presentes en una cierta comunidad, presentan una equivalencia funcional; esto es, poseen equivalentes tasas de nacimiento/colonización y muerte/extinción o migración.

Tal como se vió en la Sección 1.1, el interés no fue proporcionar una descripción estadística de la expectativa de encontrar s especies en una comunidad dada, sino comprender cómo varía s dependiendo de los procesos de colonización y extinción.

Estableciendo así que la probabilidad de observar s especies en una isla, o un conjunto de islas idénticas, sigue la ecuación maestra:

$$\frac{dP_s(t)}{dt} = \lambda_{s-1}P_{s-1}(t) + \mu_{s+1}P_{s+1}(t) - (\lambda_s + \mu_s)P_s(t),$$

donde s es el número de especies, λ_s es la tasa de colonización o especiación que aumenta de s a $s+1$ especies y μ_s es la tasa de extinción que disminuye s a $s-1$. Acá se asume que existe un número máximo de especies posibles en la comunidad (es decir, un *pool*), la cual denotaremos por K. Luego, se tienen como condiciones de borde $\lambda_K = \mu_0 = 0$.

Por lo visto anteriormente, tenemos que la distribución estacionaria estará dada por $P_s^* = P_0^* \prod_{n=1}^{s} \frac{\lambda_{n-1}}{\mu_n}$. Una elección de funciones de tasa más sencilla, pero que a su vez da origen a un modelo que comúnmente se plantea en ecología para modelar la riqueza de especies (Marquet et al., 2020), es la relación lineal; esto es, tomando $\lambda_s = \lambda(K - s)$ y $\mu_s = \mu s$. Colocando estas funciones de tasa en la distribución estacionaria anterior, tenemos que:

$$P_s^* = P_0^* \prod_{n=1}^{s} \frac{\lambda_{n-1}}{\mu_n} = P_0^* \prod_{n=1}^{s} \frac{\lambda(K - [n-1])}{\mu n} = P_0^* \frac{\lambda^s}{\mu^s} \prod_{n=1}^{s} \frac{K - (n-1)}{n}$$

$$= P_0^* \frac{\lambda^s}{\mu^s} \frac{K!}{(K-s)!s!}.$$

Notar ahora que, para satisfacer la condición $\sum_{s=0}^{K} P_s^* = 1$, se tiene que $P_0^* = \left(\frac{\mu}{\mu+\lambda}\right)^K = \left(1 - \frac{\lambda}{\mu+\lambda}\right)^K$. En efecto,

$$P_s^* = \left(\frac{\mu}{\mu+\lambda}\right)^K \frac{\lambda^s}{\mu^s} \frac{K!}{(K-s)!s!}$$

$$= \frac{K!}{(K-s)!s!} \left(\frac{\lambda}{\mu+\lambda}\right)^s \left(1 - \frac{\lambda}{\mu+\lambda}\right)^{K-s}. \tag{2.5}$$

Esto es, la distribución estacionaria resultante de la riqueza es una distribución binomial con "probabilidad de éxito" igual a $p = \lambda/(\lambda + \mu)$.

2.4. Procesos de difusión

En esta sección estudiaremos algunos procesos Markovianos de espacio de estados continuos, a tiempo continuo, conocidos como procesos de difusión. Introduciremos estos procesos vía límite de un proceso Markoviano de saltos conocido como caminata o paseo aleatorio.

Sea $X_1, X_2, ..., X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuídas. A la sucesión $\{S_n\}$ definida por $S_n = \sum_{i=1}^{n} X_i$ se le conoce como caminata o paseo aleatorio. Decimos que este es simétrico si $E(S_n) = E(X_1) = 0$. En particular, podemos definir un paseo aleatorio sobre $\mathbb{Z}$ por:

$$p_{ij} = P(S_{n+1} = j \mid S_n = i) = \begin{cases} p, & \text{si } j = i+1 \\ 1-p, & \text{si } j = i-1 \\ 0, & \text{en otro caso} \end{cases}$$

donde $p \in (0, 1)$. Notar que de esta definición podemos deducir que $\{S_n\}$ es una cadena de Markov irreducible de periodo dos. El paseo será también simétrico si $p = 1/2$.

Ahora bien, asumamos que tenemos un paseo aleatorio simétrico $\{S_n\}$ tal que $E(X_1^2) = \sigma^2$, y consideremos el siguiente proceso re-escalado $S_n^h = hS_n$, donde $h = h(n) = 1/\sqrt{n}$. Así, notar que a medida que crece n los saltos de S_n^h son cada vez más pequeños. Cuando $n \to \infty$, tendremos que, por Teorema del Límite Central, S_n^h converge en distribución (o en ley) hacia una variable aleatoria normal de media cero y varianza σ^2. Pero podemos incluso ir un poco más allá, definiendo una dinámica al paseo, definiendo $S_n^h(t) = h \sum_{i=1}^{[nt]} X_i$, donde $[\cdot]$ es la función parte entera de un número

real. Luego, obtenemos un proceso $\{S_n^h(t)\}$ de saltos, a tiempo continuo, donde el tamaño del salto se reduce a medida que n crece. La pregunta que surge ahora es, ¿qué pasa con $\{S_n^h(t)\}$ cuando $n \to \infty$?

Del Teorema del Límite Central, podemos deducir también que para cada t fijo tendremos que $S_n^h(t)$ también convergerá en distribución a una variable aleatoria normal de media cero pero con varianza $\sigma^2 t$ cuando $n \to \infty$, ya que esta se trata de una variable aleatoria. Sin embargo, en caso de considerar el proceso $S_n^h(\cdot)$ necesitamos un resultado funcional límite, ya que este ahora se trata de la convergencia de S_n^h uniformemente hacia otro proceso sobre algún intervalo de tiempo $[0, T]$. Para ello, tenemos el principio de invarianza de Donsker, el cual nos dice que sobre un intervalo compacto $[0, T]$ el proceso $S_n^h(\cdot)$ converge en distribución hacia el proceso $\sigma W_\cdot$ cuando $n \to \infty$, donde $W_\cdot$ se conoce como proceso de Wiener o movimiento Browniano, el cual satisface lo siguiente:

- $W_0 = 0$, y para cada t, W_t tiene una distribución normal de media cero y varianza t, la que coincide con la distribución de $W_{t+s} - W_s$ (propiedad de incrementos estacionarios).
- Para cada $0 \leq s \leq t$, $W_t - W_s$ es independiente de W_s (propiedad de incrementos independientes).
- $W_\cdot$ tiene trayectorias continuas pero no diferenciables.

Además se tiene que $Cov(W_t, W_s) = \text{mín}\{t, s\}$. Otra notación usual de este proceso es $B_\cdot$ (podrán deducir el por qué). Este proceso presenta un comportamiento "centrado" pero muy errático, lo cual dota a este de una no diferenciabilidad siendo sin embargo este continuo. Este proceso es el fundamental para definir los procesos de difusión.

Dado lo anterior, podemos definir una primera estructura general de un proceso de difusión, que la definimos a través de una ecuación integral:

$$X_t = X_0 + \int_0^t b(X_s)ds + \sigma W_t, \tag{2.6}$$

o expresado en su forma abreviada "diferencial",

$$dX_t = b(X_t)dt + \sigma dW_t. \tag{2.7}$$

De esta última expresión, debemos tener presente lo siguiente:

- No es una expresión diferencial como usualmente lo entendemos en una ecuación diferencial ordinaria, ya que dW_t no es un diferencial. Luego, debemos entender (2.7) como una expresión abreviada de (2.6). (2.7) define un caso particular de lo que se conoce como una ecuación diferencial estocástica (EDE) de difusión.
- Dado que estamos describiendo un sistema estocástico, una solución de (2.6) se entiende como la solución de una realización; esto es, dado un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, una realización es por cada $\omega \in \Omega$,

  $$X_t(\omega) = X_0(\omega) + \int_0^t b(X_s(\omega))ds + \sigma W_t(\omega),$$

  aunque en la práctica se suele omitir ω, entendiendo que de esto se trata. Así, dado un $\omega \in \Omega$, la garantía de una única solución de (2.6) se da bajo la condición usual de Lipschitz para $b(x)$.

- A $b(X_\cdot)$ se le conoce como deriva de la ecuación (*drift* en inglés), y a σ como coeficiente de difusión.
- Dado que el proceso $X_\cdot$ es estocástico, existe asociado a este un proceso que describe su densidad de probabilidad en el tiempo. Esta se describe a través de una ecuación diferencial parcial, la cual vendría a ser su ecuación maestra, pero que en difusiones esta se define a través de la ecuación de Fokker-Planck, que veremos mas adelante.
- Cualquier función f de $X_\cdot$, $f(X_\cdot)$ se conoce como observable del sistema, mientras que la función de probabilidad del proceso describe el estado del sistema. Para ciertas funciones f, los observables del sistema se pueden describir también a través de una EDE, mediante la conocida fórmula de Itô, la cual también veremos más adelante.

Comenzaremos con el estudio de un caso particular simple de una EDE, donde podremos hallar una solución explícita fácilmente. Antes, debemos definir lo que es una integral estocástica con respecto al movimiento Browniano, y un poco de álgebra de integrales. Para poder llevar a cabo formalmente todo esto se requieren de varios aspectos técnicos, los cuales son tratados en un curso de cálculo estocástico. Así, acá simplemente daremos alguna noción, informal en algunos casos, de estos aspectos.

Sea $H_\cdot$ una función continua. Definimos la integral estocástica de $H_\cdot$ con respecto a $W_\cdot$, sobre un intervalo $[0,t]$, como al límite en probabilidad de sumas del estilo

$$\sum_{t_k, t_{k+1} \in \pi_t} H_{t_k}(W_{t_{k+1}} - W_{t_k}), \tag{2.8}$$

cuando $|| \ \pi_t \ ||\to 0$, donde π_t designa una partición del intervalo $[0,t]$ (es decir, el límite de la norma anterior se toma cuando máx $\mid t_{k+1} - t_k \mid$ en π_t tiende a cero). A este límite se le denota como $\int_0^t H_s dWs$, y que es la integral estocástica de $H_\cdot$ con respecto a $W_\cdot$ sobre $[0,t]$. Para operar con el producto de las diferenciales dt y dW_t, podemos ocupar la regla $dtdt = 0$, $dtdW_t = 0$ y $dW_t dW_t = dt$, lo cual se puede justificar formalmente utilizado también límites en probabilidad.

Consideremos ahora que $b(x) = \mu x$ en (2.7), con $\mu \in \mathbb{R}$. Al proceso resultante se le conoce como proceso de Ornstein-Uhlenbeck,

$$dX_t = \mu X_t dt + \sigma dW_t. \tag{2.9}$$

Este proceso se puede resolver análogamente al método de variación de parámetros en EDO. Sea $g(X_t, t) = X_t e^{-\mu t}$. Luego,

$$dg(X_t, t) = -\mu e^{-\mu t} X_t dt + e^{-\mu t} dX_t = \sigma e^{-\mu t} dW_t.$$

Luego, tomando la integral a ambos lados, obtenemos:

$$X_t e^{-\mu t} - X_0 = \int_0^t \sigma e^{-\mu s} dW_s \iff X_t = X_0 e^{\mu t} + \sigma \int_0^t e^{\mu(t-s)} dW_s.$$

Si suponemos que la condición inicial X_0 no es aleatoria sino fija $X_0 = x_0 \in \mathbb{R}$ (en algunos casos se suele considerar la condición inicial como una variable aleatoria independiente del movimiento

Browniano), tendremos que X_t se distribuirá normalmente debido a la normalidad del movimiento Browniano. Podemos justificar vía sumas del estilo (2.8) que $E\left(\int_0^t e^{\mu(t-s)}dW_s\right) = 0$. Así, la esperanza del proceso quedará dada por:

$$E(X_t) = x_0 e^{\mu t}.$$

Para el cálculo de la varianza, tendremos que tener la siguiente en cuenta la siguiente fórmula, la cual es un caso particular de la conocida isometría de Itô:

$$E\left[\left(\int_0^t H_s dW_s\right)^2\right] = \int_0^t H_s^2 ds,$$

para una función $H_\cdot$ continua, lo cual se puede también justificar vía sumas del estilo (2.8) y la toma de límites en probabilidad. Notar que en particular si $H_\cdot = 1$, entonces tenemos que $E(W_t^2) = V(W_t) = t$ y $Cov(X_t, X_s) = E[(X_t - \mu_X(t))\,(X_s - \mu_X(s))]$.

Podemos extender el proceso descrito por (2.7), considerando un coeficiente de difusión más general:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \tag{2.10}$$

donde $\sigma(x)$ es ahora una función positiva. La existencia y unicidad de una solución por trayectorias de (2.10) se garantiza adicionando la condición de Lipschitz para $\sigma(x)$. La integral estocástica de $\sigma(X_\cdot)$ con respecto al movimiento Browniano sobre un intervalo de tiempo $[0, t]$ se define análogamente como en (2.8); consideramos el límite en probabilidad de sumas como

$$\sum_{t_k, t_{k+1} \in \pi_t} \sigma(X_{t_k})(W_{t_{k+1}} - W_{t_k}),$$

cuando $||\ \pi_t\ || \to 0$. Este límite se expresa entonces como $\int_0^t \sigma(X_s)dW_s$, la cual tiene esperanza cero. Por otro lado, la isometría de Itô queda ahora expresada como:

$$E\left[\left(\int_0^t \sigma(X_s)dW_s\right)^2\right] = E\left(\int_0^t \sigma(X_s)^2 ds\right).$$

Pasaremos ahora a enunciar la fórmula de Itô.

Teorema 2.4.1 *Sea f una función en C^2 (conjunto de segundas derivadas parciales continuas). Entonces se tiene que $f(X_\cdot)$ satisface:*

$$df(X_t) = b(X_t)f'(X_t)dt + \sigma(X_t)f'(X_t)dW_t + \frac{1}{2}\sigma(X_t)^2 f''(X_t)dt$$

$$\iff f(X_t) = f(X_0) + \int_0^t b(X_s)f'(X_s)ds + \int_0^t \sigma(X_s)f'(X_s)dW_s + \frac{1}{2}\int_0^t \sigma(X_s)^2 f''(X_s)ds. \tag{2.11}$$

Consideremos por ejemplo simplemente el movimiento Browniano $W_\cdot$. Aplicando la fórmula anterior, obtenemos:

$$f(W_t) = f(0) + \int_0^t f'(W_s)dW_s + \frac{1}{2}\int_0^t f''(W_s)ds.$$

Esta fórmula también nos es útil para obtener en algunos casos puntuales soluciones explícitas, como en el caso del movimiento Browniano geométrico, descrito por:

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \tag{2.12}$$

con $X_0 > 0$. Apliquemos la fórmula de Itô con $f(x) = \ln(x)$,

$$d\ln(X_t) = \mu dt + \sigma dW_t - \frac{\sigma^2}{2} dt.$$

Así, tomando la integral y luego exponenciando, obtenemos la solución de la ecuación:

$$X_t = X_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right).$$

Se puede mostrar que X_t sigue un distribución log-normal con media $E(X_t) = X_0 e^{\mu t}$ y varianza $V(X_t) = X_0^2 e^{2\mu t}\left(e^{\sigma^2 t} - 1\right)$.

Es importante notar que en el proceso descrito por (2.12), dada una condición inicial positiva, sus trayectorias siempre se mantendrán en $(0, \infty)$, a diferencia del proceso (2.9) cuyas trayectorias se pueden mover por todo $\mathbb{R}$. El haber incluído un coeficiente de difusión proporcional al proceso mismo, $\sigma X_\cdot$, logró concebir un proceso con tales trayectorias positivas.

Dada la ecuación de Itô (2.11), a cada difusión descrita por (2.10) se le puede asociar un operador diferencial de segundo orden L a $X_\cdot$, el cual se define como:

$$Lf(x) = \lim_{t \downarrow 0} \frac{E(f(X_t)|X_0 = x) - f(x)}{t} = b(x)\frac{df(x)}{dx} + \frac{\sigma(x)^2}{2}\frac{d^2 f(x)}{dx^2}, \tag{2.13}$$

para todo $f \in C^2$. Es decir, L es el operador $L = b(\cdot)\frac{d}{dx} + \frac{\sigma(\cdot)^2}{2}\frac{d^2}{dx^2}$. A este se le conoce como "generador de una difusión de Itô" [ver "Stochastic Differential Equations" (Øksendal, 2003)], desde el cual se puede carecterizar a los observables de un proceso de difusión (2.10).

En general, para EDEs de la forma (2.10), en donde se hace complejo obtener soluciones explícitas, y por tanto, determinar su distribución de probabilidad, podemos plantear en general una ecuación diferencial parcial para hacer descripción de la evolución de la densidad de probabilidad del proceso en el tiempo. Esta ecuación es la conocida como la ecuación de Fokker-Planck (también conocida como ecuación *forward* de Kolmogórov) para la densidad de probabilidad del proceso $p(x,t)$, la cual es la única solución de

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial[b(x)p(x,t)]}{\partial x} + \frac{1}{2}\frac{\partial^2[\sigma(x)^2 p(x,t)]}{\partial x^2} \tag{2.14}$$

que satisface $\int_{\mathbb{R}} p(x,t)dx = 1$, para todo $t \in \mathbb{R}_+$. Esta ecuación corresponde a la ecuación de Chapman-Kolmogórov para la ley de un proceso de difusión de Itô [ver "Brownian Motion and Stochastic Calculus" (Karatzas y Shreve, 2012)].

Tratar de resolver explícitamente (2.14) puede ser también algo bastante complejo. Sin embargo, la distribución estacionaria del proceso, la cual surge como solución de $\frac{\partial p(x,t)}{\partial t} = 0$, $\rho(x)$, puede ser más fácil de obtener, siempre cuando ella exista. Por ejemplo, el movimiento Browniano no posee

una distribución estacionaria debido a que su varianza crece en proporción al tiempo (es decir, esta "no se estabiliza"); también se puede mostrar que en el proceso de Ornstein-Uhlenbeck (2.9), no contaremos tampoco con una distribución estacionaria cuando $\mu > 0$, ya que en este caso el proceso con probabilidad 1 tenderá a crecer indefinidamente a medida que t crece. Cuando no estamos bajo casos como estos, tenemos que la densidad de probabilidad invariante del proceso estará dada por:

$$\rho(x) = C \exp\left(\frac{2\int b(x)dx}{\sigma^2(x)}\right), \tag{2.15}$$

donde C es la constante de normalización $C^{-1} = \int_{\mathbb{R}} \exp\left(\frac{2\int b(x)dx}{\sigma^2(x)}\right) dx$ y es la única solución que satisface $-\frac{\partial[b(x)\rho(x)]}{\partial x} + \frac{1}{2}\frac{\partial^2[\sigma(x)^2\rho(x)]}{\partial x^2} = 0$.

Así por ejemplo, si en (2.9) tenemos $\mu < 0$, el cual podemos re-escribir como $\mu = -\nu$, con $\nu > 0$, la distribución estacionaria de este proceso estará dada por:

$$\rho(x) = \sqrt{\frac{\nu}{\pi\sigma^2}} e^{-\frac{\nu x^2}{\sigma^2}},$$

es decir, tendrá una distribución normal de media cero y varianza $\frac{\sigma^2}{2\mu}$.

En el capítulo siguiente se detallará la construcción de nuestro modelo, basado en los principales aspectos de modelos estocásticos usualmente utilizados en la dinámica de la riqueza de especies, como lo son los procesos de nacimiento y muerte, y los procesos de difusión.

Capítulo 3

El modelo propuesto

En este capítulo detallaremos la construcción matemática de nuestro modelo propuesto para la dinámica de la riqueza de especies. Esta construcción consta de tres etapas, a decir, (1) el planteamiento de la dinámica de la riqueza de especies vista como un proceso de nacimiento y muerte; luego, (2) la aproximación de este proceso a un proceso de difusión mediante un cambio de escala espacial y temporal, y finalmente (3) la obtención de una distribución de probabilidad estacionaria que nos permitirá modelar los datos de riqueza de especies.

3.1. El modelo

Definamos por $S_\cdot^N$ el proceso el proceso que describe la proporción de especies bajo un *pool* de N especies dentro de alguna comunidad local en el tiempo. Por lo tanto, tenemos que $S_t^N \in \{0, 1/N, 2/N, ..., 1\}$ para todo $t \in \mathbb{R}_+$.

Vamos a asumir que $S_\cdot^N$ sigue un proceso de nacimiento y muerte con tasas de transición dadas por:

$$Q_N(s, s') = \begin{cases} B_N(s) \text{ si } s' = s + 1/N,\ s \in \{0, 1/N, 2/N, ..., (N-1)/N\}\ , \\ D_N(s) \text{ si } s' = s - 1/N,\ s \in \{1/N, 2/N, ..., 1\}\ , \\ 1 - (B_N(s) + D_N(s)) \text{ si } s' = s,\ s \in \{0, 1/N, 2/N, ..., 1\}\ , \\ 0 \text{ en otro caso,} \end{cases} \tag{3.1}$$

donde $B_N(s)$ y $D_N(s)$ son las tasas de nacimiento y muerte cuando una proporción de s especies está presente. Notar que estas tasas están asociadas a los procesos de inmigración/colonización y emigración/extinción de especies. Además, debemos imponer las condiciones de frontera $B_N(1) = D_N(0) = 0$.

En Marquet, Espinoza, Abades, Ganz, y Rebolledo (2017) un modelo similar es propuesto, para la abundancia relativa de una especie. Allí, el objetivo es modelar la dinámica del número de individuos de una cierta especie, presente en una comunidad local. Para ello, tomaron este número relativo al número total de individuos de todas las especies presentes en tal comunidad local, de manera análoga a como hemos introducido nuestra proporción de riqueza de especies con respecto a su *pool*. Luego, utilizando un cambio de escala temporal, dilatando el tiempo proporcionalmente al número total considerado, se muestra que este proceso converge en distribución a un proceso

de difusión con valores en $[0,1]$. Este resultado da cuenta el siguiente teorema, y es el cual aplicaremos para pasar de nuestro proceso de nacimiento y muerte (3.1) a un proceso de difusión en $[0,1]$.

Teorema 3.1.1 *Sea $Z_N(t) = S^N_{Nt}$, para todo $N \geq 1$ y $t \geq 0$. Asumir $Z_N(0) = z \in [0,1]$ es fijo, y que existen dos funciones continuas $\theta, \sigma : [0,1] \to \mathbb{R}$, con $\sigma(x) > 0$, para todo $x \in]0,1[$, $\theta \in C^1(]0,1[)$, $\sigma \in C^2(]0,1[)$, de manera que satisfagan las dos hipótesis siguientes:*

(H1) Para todo $T > 0$, $\sup_{t\in[0,T]} |(B_N(Z_N(t-)) - D_N(Z_N(t-))) - \theta(Z_N(t-))| \to 0$ en probabilidad;

(H2) Para todo $T > 0$, $\sup_{t\in[0,T]} |\frac{1}{N}(B_N(Z_N(t-)) + D_N(Z_N(t-))) - \sigma^2(Z_N(t-))| \to 0$ en probabilidad, cuando $N \to \infty$.

Entonces, el proceso $Z_N(\cdot)$ converge en distribución hacia un proceso de difusión $Z(\cdot)$ que se puede representar como

$$Z(t) = Z(0) + \int_0^t \theta(Z(s))ds + \int_0^t \sigma(Z(s))dW_s, \ (t \geq 0). \tag{3.2}$$

$Z(t) \in [0,1]$ con probabilidad 1 para todo $t \geq 0$, el cual tiene un generador L dado por

$$Lf(x) = \frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2}f(x) + \theta(x)\frac{d}{dx}f(x), \ (x \in \mathbb{R}), \tag{3.3}$$

para cualquier $f \in C^2([0,1])$ tal que $f(0) = f(1) = 0$. Más aún, la densidad de probabilidad de (3.2) satisface la ecuación de Fokker-Planck ("ecuación maestra"),

$$\frac{\partial \rho_t(x)}{\partial t} = \frac{1}{2}\frac{d^2}{dx^2}\left(\sigma^2(x)\rho_t(x)\right) - \frac{d}{dx}\left(\theta(x)\rho_t(x)\right). \tag{3.4}$$

Comentario: Nos referiremos al proceso de límite $Z(\cdot)$ como "difusión asintótica de riqueza relativa". Básicamente, bajo un *pool* de N especies, tal proceso es estimado por $S^N_\cdot$, el cual es de naturaleza discreta. El inconveniente que trae modelar datos con esta escala discreta es que debemos contar la ocurrencia de cada estado de $S^N_\cdot$ en una muestra de riqueza relativa a un cierto *pool* de tamaño N durante el período de observación. Sin embargo, si aproximamos $S^N_\cdot$ por el proceso $Z(\cdot)$, y si asumimos que estamos bajo un estado estacionario, solo nos basta ajustar la distribución empírica de aquellas riquezas relativas observadas en una muestra a una curva apropiada de la distribución invariante de (3.4), cuya familia de densidades corresponderá a nuestro modelo estadístico.

Es posible mostrar que bajo ciertas funciones de tasas $B_N(s)$ y $D_N(s)$ en (3.1) podemos garantizar la convergencia hacia ciertas difusiones específicas de (3.2), tales que la distribución de probabilidad invariante de (3.4) queda explícitamente dada. La carecterización de tales tasas en nuestro contexto debe ir ligada a una interpretación ecológica razonable. Una de estas elecciones, la cual dará origen a un modelo estadístico, es la siguiente.

Para $s = k/N$, $k = 0, 1, 2, ..., N$, asumamos que las tasas de nacimiento y muerte en (3.1) tienen la siguiente estructura:

$$\begin{cases} B_N\left(\frac{k}{N}\right) = \lambda\left(1 - \frac{k}{N}\right) + \frac{\gamma}{2}\frac{k}{N}\left(1 - \frac{k}{N}\right) \\ D_N\left(\frac{k}{N}\right) = \mu\frac{k}{N} + \frac{\gamma}{2}\frac{k}{N}\left(1 - \frac{k}{N}\right), \end{cases} \tag{3.5}$$

donde $\lambda, \mu, \gamma > 0$ son constantes independientes del tamaño del *pool* N. En tal caso, los coeficientes del proceso de difusión asintótica $Z(\cdot)$ son $\theta(x) = \lambda(1-x) - \mu x$ y $\sigma^2(x) = \gamma x(1-x)$. El proceso resultante en (3.2) dará origen a una particular familia de densidades invariantes de (3.4). Este resultado se establecerá en el siguiente corolario.

Corolario 3.1.1.1 *Bajo la estructura dada en (3.5), el proceso de difusión asintótica $Z(\cdot)$ tiene una distribución de probabilidad invariante Beta con densidad $\rho_{\alpha,\beta}$ dada por:*

$$\rho_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \tag{3.6}$$

donde

$$\alpha = \frac{2\lambda}{\gamma},\ \beta = \frac{2\mu}{\gamma}.$$

Para mostrar este colorario, consideramos la ecuación invariante de (3.4):

$$0 = \frac{1}{2}\frac{d^2}{dx^2}\left(\sigma^2(x)\rho(x)\right) - \frac{d}{dx}\left(\theta(x)\rho(x)\right),$$

en donde se verifica que $\rho_{\alpha,\beta}(\cdot) = \rho(\cdot)$ es solución. Así, nuestro modelo estadístico estará caracterizado por (3.6). La interpretación ecológica del modelo se mostrará en la siguiente subsección.

3.1.1. Conexión de (3.6) con el modelo de MacArthur y Wilson

Como se vió en las secciones 2.3 y 2.4, la ecuación maestra de MacArthur y Wilson, cuando $\lambda_n = \lambda(N-n)$ y $\mu_n = \mu n$ con λ y μ constantes positivas, su distribución estacionaria, $P_n(\infty)$, sigue una distribución (2.5) de parámetros N (el *pool*) y $p = \lambda/(\lambda+\mu)$. En este sentido, tenemos una conexión con la distribución de probabilidad invariante de $Z(\cdot)$ (3.6): el parámetro p corresponde a la probabilidad de encontrar una especie en la comunidad focal, que coincide con la esperanza (3.6), $E_{\rho_{\alpha,\beta}}(Z(t)) = \alpha/(\alpha+\beta) = \lambda/(\lambda+\mu)$. Por lo tanto, los parámetros α y β de nuestro modelo están relacionados proporcionalmente a las tasas de colonización/inmigración y extinción/emigración, respectivamente. El parámetro γ corresponde a un parámetro que cuantifica la intensidad difusiva del proceso.

Adicionalmente, podemos aproximar la distribución de la riqueza absoluta bajo un *pool* N usando la distribución beta-binomial:

$$\pi(n \mid N, \alpha, \beta) = \frac{N!}{n!(N-n)!}\frac{B(n+\alpha, N-n+\beta)}{B(\alpha,\beta)}, \tag{3.7}$$

la cual es la distribución de muestreo de hallar un cierto número de especies en la comunidad focal. Notar además que este modelo podría verse como una versión de parámetro variante del modelo de Barton y David establecido en (1.1) y (1.2).

En el capítulo siguiente abordaremos nuestro modelo estadístico (3.6), sus propiedades y técnicas de estimación de parámetros, así como también procederemos al análisis estadístico de datos de riqueza reales previamente publicados, en donde evaluaremos la bondad de ajuste de nuestra propuesta.

Capítulo 4

Análisis estadístico de datos

En este capítulo comenzaremos estableciendo nuestro modelo estadístico para la riqueza relativa de especies a partir de los resultados obtenidos en el capítulo anterior. Posteriormente, daremos a conocer algunas de sus propiedades distribucionales, para posteriormente dar a conocer técnicas de estimación de parámetros y posterior la bondad de ajuste a datos reales previamente publicados. Finalizaremos este capítulo, así como el trabajo de esta investigación, con una evaluación y discusión de los resultados obtenidos, así como sus implicaciones en el contexto ecológico.

4.1. El modelo estadístico

Como bien se señaló al final del capítulo anterior, de los resultados acerca de la distribución límite invariante para la riqueza relativa, asintótica para un *pool* grande, el modelo estadístico corresponderá a la familia de distribuciones beta dada en la ecuación (3.6). Formalmente, nuestro modelo estadístico corresponderá a la familia paramétrica:

$$\mathfrak{F} = \left\{ f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\mathbf{1}_{\{x\in(0,1)\}} : (\alpha,\beta) \in (0,\infty)\times(0,\infty) \subset \mathbb{R}^2_+ \right\}. \tag{4.1}$$

En cuanto al rol del modelo estadístico en nuestro contexto, según lo derivado en el capítulo anterior, los parámetros α y β están proporcionalmente relacionados con las tasas de colonización/inmigración y extinción/emigración de especies en una comunidad local, respectivamente. Esto es, $\alpha \propto \lambda$ y $\beta \propto \mu$. La constante común de proporcionalidad depende de un parámetro difusivo desconocido γ, lo cual nos va a impedir identificar λ y μ a partir de nuestro modelo estadístico (4.1). Recordar que un modelo estadístico paramétrico dado por $\tilde{\mathfrak{F}} = \{\mathbb{P}_\theta \in \mathcal{P} : \theta \in \Theta\}$, donde $\mathcal{P}$ corresponde a una familia de medidas de probabilidad, se dice que es identificado si para cada $\theta_1, \theta_2 \in \Theta$ tales $\theta_1 \neq \theta_2$, se tiene que $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. En nuestro caso, nuestro modelo (4.1) es identificado bajo $\theta = (\alpha, \beta)$, pero sin embargo no es identificado para $(\lambda, \mu, \gamma) \in (0,\infty)\times(0,\infty)\times(0,\infty) \subset \mathbb{R}^3_+$. Por ejemplo, si tomamos $(1,1,2)$ y $(1/2,1/2,1)$, bajo la relación $\alpha = 2\lambda/\gamma$ y $\beta = 2\mu/\gamma$ dada en (3.6), tenemos que generamos dos distribuciones idénticas en (4.1), pero donde tenemos que $(1,1,2) \neq (1/2,1/2,1)$. Notar que este problema puede evitarse si se asume γ conocida. Sin embargo, no contamos con algún argumento teórico para asumir esto.

Por otro lado, algunas situaciones especiales en cuanto al esquema de muestreo utilizado y/o a los datos como fueron recolectados pueden provocar que tengamos que considerar variantes de nuestro modelo (4.1). Estas situaciones son:

- Existencia de ceros: esto ocurre debido al esquema de muestreo, en donde la división territorial de la metacomunidad es de tal "fineza" que existen territorios en donde no se encuentran especies dentro del *pool* considerado. En tal caso, debemos considerar la distribución beta "inflada en cero" (Ospina y Ferrari, 2010).

- No contar con la información del *pool*: esto es, contamos con los datos de riqueza (número de especies) en cada territorio muestreado, pero el tamaño del *pool* (o la lista de total de las especies consideradas) no se entrega. En tal caso, una alternativa consiste en considerar la distribución beta-binomial para la riqueza absoluta, en donde el parámetro N representa al *pool*, el cual es ahora desconocido, y por tanto, debe estimarse (ver Ecuación (3.7)).

La densidad de probabilidad de la beta "inflada en cero", o densidad beta cero-inflada, está dada por:

$$f_I(x;\delta,\alpha,\beta) = \begin{cases} \delta, & \text{si } x = 0; \\ (1-\delta)f(x;\alpha,\beta), & \text{si } x \in (0,1), \end{cases} \tag{4.2}$$

para $\delta \in (0,1)$ la cual representa la probabilidad de observar el valor 0. En nuestro contexto, este será nuestro modelo para la riqueza relativa en casos donde exista una probabilidad no nula de no observar especies en algunos de los sitios muestreados de la metacomunidad focal. Notar que también podríamos considerar esta distribución beta inflada en 1, es decir, donde pueden existir sitios dentro de la metacomunidad en donde todo el *pool* esté presente. Sin embargo, esta última situación es prácticamente inverosímil debido a la "capacidad de carga", la cual corresponde a la cantidad máxima de especies (o biomasa de especies) que puede albergar una comunidad local. Esta cantidad regula entonces la población de especies en donde se impide la sobrepoblación de estas. Por ello, solo consideraremos una posible inflación de la distribución con respecto a cero.

Por otro lado, el no contar con la información completa del *pool* hace que este ahora no sea un parámetro desconocido, y por ende, solo podemos trabajar con las riquezas absolutas. Luego, relacionado con nuestro modelo estadístico (4.1), la distribución de la riqueza absoluta podría aproximarse por la distribución beta-binomial antes mencionada (ver Ecuación (3.7)). Y a través de esta, podemos proceder a la estimación del *pool*.

A continuación procederemos a mostrar métodos de estimación de parámetros para nuestro modelo estadístico, y para los parámetros de los modelos alternativos recientemente mencionados.

4.2. Métodos de estimación de parámetros

A continuación mostraremos los métodos de estimación de momentos (EM) y de máxima verosimilitud (MV) para nuestro modelo estadístico (4.1), y para los modelos alternativos dados por (4.2) y (3.7), los cuales resultan ser métodos habituales de estimación paramétrica en inferencia frecuentista.

4.2.1. Estimación de parámetros en la distribución beta

Método de momentos

El método de estimación de momentos para los parámetros de la distribución beta implica igualar el primer y el segundo momento de la distribución con el primer y el segundo momento muestral, y resolver las ecuaciones para α y β.

Dada una muestra aleatoria $\mathbf{X} = (X_1, ..., X_n)$ los EM para los parámetros α y β de la distribución beta están dados, respectivamente, por :

$$\hat{\alpha} = \bar{X}\left(\frac{\bar{X}\left(1-\bar{X}\right)}{S^2} - 1\right) \tag{4.3}$$

y

$$\hat{\beta} = \left(1-\bar{X}\right)\left(\frac{\bar{X}\left(1-\bar{X}\right)}{S^2} - 1\right), \tag{4.4}$$

donde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_I - \bar{X}\right)^2$.
Este método de estimación resulta ser bastante práctico y sencillo de obtener. Sin embargo, en general se prefieren los EMV, dado que estos, bajo ciertas condiciones de regularidad, resultan ser asintóticamente normales, consistentes y asintóticamente eficientes, o *BAN* (abreviatura del inglés *Best Asymptotically Normal*) [ver, por ejemplo, Casella y Berger (2021)]. Aquellas propiedades son particularmente útiles para obtener criterios de contraste de hipótesis asintóticos en base a la distribución normal. Pasaremos entonces a ver este método a continuación.

Método de máxima verosimilitud

Las ecuaciones de máxima verosimilitud para la distribución Beta no tienen una solución en forma cerrada, por lo que las estimaciones se deben encontrar mediante el uso de un método iterativo, como por ejemplo el método de Newton-Raphson [(Ospina y Ferrari, 2010)]. Dada una muestra aleatoria de tamaño n, su función de verosimilitud está dada por:

$$L(\alpha, \beta|\mathbf{X}) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1} = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n \prod_{i=1}^{n}(x_i)^{\alpha-1}\prod_{i=1}^{n}(1-x_i)^{\beta-1} \tag{4.5}$$

Para maximizar con respecto a α y β, podemos considerar convenientemente la función de log verosimilitud y maximizar α y β con respecto a esta, la cual está dada por:

$$\begin{aligned}\log L(\alpha, \beta|\mathbf{X}) = {} & n\log(\Gamma(\alpha+\beta)) - n\log\Gamma(\alpha) - n\log\Gamma(\beta) \\ & + (\alpha-1)\sum_{i=1}^{n}\log(x_i) + (\beta-1)\sum_{i=1}^{n}\log(1-x_i)\end{aligned} \tag{4.6}$$

Por lo que para encontrar los EMV de α y β, debemos determinar las derivadas parciales respecto a cada parámetro de la función log verosimilitud y luego igualar estas a cero.

$$\frac{\partial}{\partial\alpha}\log L(\alpha, \beta|\mathbf{X}) = \frac{n\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\log(x_i) = 0 \tag{4.7}$$

$$\frac{\partial}{\partial\beta}\log L(\alpha, \beta|\mathbf{X}) = \frac{n\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{n\Gamma'(\beta)}{\Gamma(\beta)} + \sum_{i=1}^{n}\log(1-x_i) = 0 \tag{4.8}$$

Sin embargo, no existe una solución cerrada para este sistema de ecuaciones, por lo que para resolverlo se puede utilizar el método de Newton-Raphson. Dado que los estimadores de momentos se pueden obtener inmediatamente, podemos utilizar estos como punto de arranque del método iterativo. Por otro lado, este método puede encontrarse implementado en paquetes estadísticos de

uso popular, como R [(Ospina y Ferrari, 2010)].

En nuestro caso estimaremos $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ iterando:

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \mathbf{G}^{-1} g \tag{4.9}$$

donde $\mathbf{g}$ es el vector de ecuaciones normales:

$$\mathbf{g} = [g_1 \quad g_2] \tag{4.10}$$

con,

$$g_1 = \psi(\alpha) - \psi(\alpha + \beta) - \frac{1}{n}\sum_{i=1}^{n} \log(x_i) \tag{4.11}$$

y

$$g_2 = \psi(\beta) - \psi(\alpha + \beta) - \frac{1}{n}\sum_{i=1}^{n} log(1 - x_i) \tag{4.12}$$

y $\mathbf{G}$ es la matriz de la segunda derivadas:

$$\mathbf{G} = \begin{bmatrix} \frac{dg_1}{d\alpha} & \frac{dg_1}{d\beta} \\ \frac{dg_2}{d\alpha} & \frac{dg_2}{d\beta} \end{bmatrix} \tag{4.13}$$

con,

$$\frac{dg_1}{d\alpha} = \psi'(\alpha) - \psi'(\alpha + \beta) \tag{4.14}$$

$$\frac{dg_2}{d\beta} = \psi'(\beta - \psi'(\alpha + \beta) \tag{4.15}$$

$$\frac{dg_1}{d\beta} = \frac{dg_2}{d\alpha} = -\psi'(\alpha + \beta) \tag{4.16}$$

donde $\psi(\alpha)$ y $\psi'(\alpha)$ son llamadas funciones di-gamma y tri-gamma, definidas por:

$$\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \tag{4.17}$$

$$\psi'(\alpha) = \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \tag{4.18}$$

Es así como el algoritmo de Newton-Raphson converge, ya que las estimaciones de α y β con cada iteración sucesiva se aproximan a $\hat{\alpha}_{EMV}$ y $\hat{\beta}_{EMV}$.

4.2.2. Estimación de parámetros en la distribución beta cero-inflada

En Ospina y Ferrari (2010) se muestra que para el caso de la distribución beta cero-inflada de parámetros α, β y δ, dada una muestra aleatoria de tamaño n, su función de verosimilitud está dada por :

$$L(\delta, \alpha, \beta|\mathbf{X}) = \prod_{t=1}^{n} f_I\left(x_t; \delta; \alpha; \beta\right) = L_1\left(\delta\right) L_2\left(\alpha, \beta\right) \tag{4.19}$$

donde,

$$L_1(\delta) = \prod_{t=1}^{n} \delta^{\mathbf{1}_{\{0\}}(x_t)} (1-\delta)^{1-\mathbf{1}_{\{0\}}(x_t)} = \delta^{T_1} (1-\delta)^{n-T_1} \tag{4.20}$$

y,

$$L_2(\alpha, \beta) = \prod_{t=1}^{n} f(x_t; \alpha; \beta)^{1-\mathbf{1}_{\{0\}}} \tag{4.21}$$

con,

$$\begin{aligned} \sum_{t=1}^{n} T(x_t) &= (T_1, T_2, T_3) \\ T_1 &= \sum_{t=1}^{n} \mathbf{1}_{\{0\}}(x_t) \\ T_2 &= \sum_{t:x_t \in (0,1)} \log x_t \\ T_3 &= \sum_{t:x_t \in (0,1)} \log(1-x_t) \end{aligned} \tag{4.22}$$

La función de verosimilitud $L(\delta, \alpha, \beta|\mathbf{X})$ se puede factorizar en dos términos; el primer término depende solo de δ y el segundo, solo depende de (α, β). Por lo tanto, los parámetros son separables (Pace y Salvan 1997, p. 128) y la inferencia de máxima verosimilitud para (α, β) se puede realizar por separado de la de δ, como si se conociera el valor de δ, y viceversa.

Así, la función logarítmica de verosimilitud para la distribución beta inflada (4.2) está dada por:

$$l(\delta, \alpha, \beta|\mathbf{X}) = \log(L(\delta, \alpha, \beta|\mathbf{X})) = l_1(\delta) + l_2(\alpha, \beta)) \tag{4.23}$$

donde,

$$l_1 = T_1 \log \delta + (n - T_1 \log(1-\delta)) \tag{4.24}$$

y,

$$l_2(\alpha, \beta) = (n - T_1) \log \left\{ \frac{\Gamma(\beta)}{\Gamma(\alpha\beta)\Gamma((1-\alpha)\beta)} \right\} + T_2(\alpha\beta - 1) + T_3((1-\alpha)\beta - 1) \tag{4.25}$$

Para encontrar los EMVs de α y β y δ, debemos determinar las derivadas parciales respecto a cada parámetro de la función log de probabilidad,obteniendo así:

$$U_\delta(\delta) = \frac{T_1}{\delta - \frac{(n-T_1)}{(1-\delta)}} \tag{4.26}$$

$$U_\alpha(\alpha, \beta) = \beta(n - T_1)[\psi((1-\alpha)\beta) - \psi(\alpha\beta) + T_2 - T_3] \tag{4.27}$$

$$U_\beta(\alpha, \beta) = (n - T_1)[\psi((\beta) - \alpha\psi(\alpha\beta) - (1-\alpha)\psi((1-\alpha)\beta)] + T_2\alpha - T_3(1-\alpha) \tag{4.28}$$

Luego al igualar a cero la ecuación (4.26), se obtiene el EMV de δ que es $\hat{\delta} = \frac{T_1}{n}$ y represeta la proporción de ceros en la muestra, de la misma forma los EMVs para α y β se obtien igualando a cero las ecuaciones (4.27) y (4.28), sin embargo no existe una solución cerrada para este sistema de ecuaciones, por lo que para resolverlo se debe utilizar el método de Newton-Raphson.

4.2.3. Estimación de parámetros en la distribución beta-binomial

Cuando no se dispone de información del *pool* y solamente contamos con el número de especies por sitio, podemos utilizar la distribución Beta-Binomial para realizar los ajustes de los parámetros α, β y N conjutamente. Acá, tanto los métodos de momentos como los de máxima verosimilitud requieren aproximaciones númericas.

Sea $S_1, S_2, ..., S_n$ una muestra aleatoria de riqueza de especies (absoluta) en n sitios, ecológicamente similares, que comparten un grupo de N especies. Como estamos asumiendo que el proceso está en un estado estacionario, podemos considerar que la distribución de la riqueza de especies sigue una distribución Beta-Binomial (DBB).

$$\pi(s|N,\alpha,\beta) = \frac{N!}{s!(N-s)!} = \frac{B(s+\alpha, N-s+\beta)}{B(\alpha,\beta)} \tag{4.29}$$

Para estimar α y β cuando no se tiene una idea clara del tamaño del *pool*, se puede realizar primeramente la estimación del tamaño de este, en conjunto con los parámetros α y β , resolviendo las ecuaciones de momentos. Para ello, debe se tener en cuenta los tres primeros momentos de la DBB:

$$M_1 = \frac{N\alpha}{\alpha+\beta}, \tag{4.30}$$

$$M_2 = \frac{N\alpha[N(1+\alpha)+\beta]}{(\alpha+\beta)(1+\alpha+\beta)}, \tag{4.31}$$

$$M_3 = \frac{N\alpha[N^2(1+\alpha)(2+\alpha)+3N(1+\alpha)\beta+\beta(\beta-\alpha)]}{(\alpha+\beta)(1+\alpha+\beta)(2+\alpha+\beta)}. \tag{4.32}$$

Considerando los momentos muestrales en los lados izquierdos de las ecuaciones anteriores, $m_k = \frac{1}{n}\sum_{i=1}^{n} S_i^k$, con $k = 1, 2, 3$, en lugar de los poblacionales, tenemos que las estimaciones de momento α, β y N se pueden obtener resolviendo

$$m_1 = \frac{\hat{N}\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}, \tag{4.33}$$

$$m_2 = \frac{\hat{N}\hat{\alpha}[\hat{N}(1+\hat{\alpha})+\hat{\beta}]}{(\hat{\alpha}+\hat{\beta})(1+\hat{\alpha}+\hat{\beta})}, \tag{4.34}$$

$$m_3 = \frac{\hat{N}\hat{\alpha}[\hat{N}^2(1+\hat{\alpha})(2+\hat{\alpha})+3\hat{N}(1+\hat{\alpha})\hat{\beta}+\hat{\beta}(\hat{\beta}-\hat{\alpha})]}{(\hat{\alpha}+\hat{\beta})(1+\hat{\alpha}+\hat{\beta})(2+\hat{\alpha}+\hat{\beta})}, \tag{4.35}$$

para $\hat{\alpha}$, $\hat{\beta}$ y $\hat{N}$. Hay que tener en consideración que el sistema conjunto de ecuaciones es no lineal, y por tanto, se requerirá una solución numérica a estas. Dentro de las soluciones, deben considerarse solamente las que pertenezcan al espacio paramétrico, que en este caso sería $(0,\infty)\times(0,\infty)\times(0,\infty)$ (acá estamos considerando que N puede variar continuamente entre $(0,\infty)$, ya que para resolver el sistema anterior no se cuenta con la restricción que $N \in \mathbb{N}$; sin embargo en la práctica, la estimación resultante de N consistirá en su parte entera).

Al igual que en el caso de nuestro modelo beta, los valores resultantes pueden ser usados como entradas para resolver numéricamente los estimadores de máxima verosimilitud. En (Aldirawi, Yang, y Metwally, 2019) han tratado con esta distribución y su estimación numérica, la cual también ha sido implementada en R: https://rdrr.io/cran/iZID/man/bb.mle.html.

Sin embargo, podemos proveer de un método iterativo para obtener semejantes resultados, siempre en busca de maximizar la verosimilitud, pero en donde el *pool* ajustado vivirá en $\mathbb{N}$. Algorítmicamente procedemos como sigue: contamos con $\{\mathcal{S}(1), ..., \mathcal{S}(n)\}$ una muestra aleatoria de la riqueza (absoluta) en n sitios; luego, es claro que el *pool* debe satisfacer:

$$\max_{i=1,...,n} \mathcal{S}(i) < N \leq \sum_{i=1}^{n} \mathcal{S}(i).$$

A continuación, para $\tilde{N}$ de $\max_{i=1,...,n} \mathcal{S}(i) + 1$ a $\sum_{i=1}^{n} \mathcal{S}(i)$, considere los datos transformados $\{\mathcal{S}(1)/\tilde{N}, ..., \mathcal{S}(n)/\tilde{N}\}$, y se ajusta para cada caso la distribución beta. Finalmente, el *pool* estimado, $\hat{N}$, corresponderá al $\tilde{N}$ entre $\max_{i=1,...,n} \mathcal{S}(i) + 1$ y $\sum_{i=1}^{n} \mathcal{S}(i)$ tal que la verosimilitud beta obtuvo su máximo valor.

No obstante, debemos tener bastante cuidado con la interpretación de este resultado, en nuestro contexto ecológico. Si $\theta = (N, \alpha, \beta) \in \Theta \subset \mathbb{R}^3$, y $L_n^{B-B}(\theta)$ es la función de verosimilitud de una muestra aleatoria de tamaño n desde una población beta-binomial, tenemos que, si $\hat{\theta}$ el EMV de θ, tendremos que:

$$L_n^{B-B}(\hat{\theta}) = \sup_{\theta \in \Theta} L_n^{B-B}(\theta) \geq \sup_{\theta \in \Theta|_{N=N^*}} L_n^{B-B}(\theta),$$

donde $\Theta|_{N=N^*}$ es el espacio paramétrico Θ, restringido al conjunto $\{N = N^*\}$, el cual es ahora un subconjunto de $\mathbb{R}^2$. En otras palabras, el ajuste de la distribución beta bajo un *pool* ajustado dará una cota superior en cuanto a la calidad del ajuste de la distribución beta a los datos de riqueza relativa. Y claramente aquel *pool* ajustado no tiene por qué estar cercano al verdadero valor de este. Por lo tanto, bajo un *pool* ajustado, ya no podemos interpretar los parámetros estimados α y β en relación a tasas de colonización/inmigración y extinción/emigración, respectivamente, ya que estas son relativas al *pool* verdadero.

4.2.4. Prueba de bondad de ajuste de Kolmogórov-Smirnov

Para indicar la bondad de ajuste de nuestros modelos con los datos a considerar, se empleará la prueba de Kolmogórov-Smirnov, la cual se utiliza para decidir si una muestra proviene o no de una población con una distribución específica $F(\cdot)$, basándose en la función de distribución acumulada empírica. Dados los n puntos $x_1, x_2, \cdots, x_n$, esta se define como $F_n = n(i)/n$, donde $n(i)$ es el número de puntos menores a x_i y es una función que aumenta con n.

La prueba de Kolmogórov-Smirnov se basa en la distancia máxima entre $F(\cdot)$ y $F_n(\cdot)$. Más concretamente, esta prueba consiste en contrastar

$$H_0 : F(x) = F_0(x) \text{ versus } H_1 : F(x) \neq F_0(x),$$

donde $F_0(\cdot)$ es la distribución que se sospecha siguen los datos. La estadística de prueba está dada por

$$KS = \max_x |F_0(x) - F_n(x)|,$$

en donde la hipótesis se rechaza sí y solo sí KS es mayor que el percentil $1 - \alpha$ de la distribución de Kolmogórov-Smirnov, la cual viene implementada en la mayoría de los *softwares* estadísticos, como por ejemplo RStudio.

4.3. Análisis de datos

En esta sección estudiaremos la bondad de ajuste de nuestro(s) modelo(s) estadístico(s) mediante la aplicación de estos a datos reales publicados. Como caso preliminar, en Gaston y Blackburn (2000), Apéndice IV: *"Berkshire Breeding Bird Data"*, se muestra la presencia/ausencia de una lista de 118 especies, la cual asumimos como el *pool*, en 25 sitios diferentes de 2×2 km^2, en un tiempo determinado. Por cada uno de estos sitios, se cuenta el número de especies presentes (riqueza de especies). Si dividimos cada uno de estos valores por el *pool*, obtenemos la riqueza relativa muestral en cada uno de los 25 sitios. Sin embargo, si analizamos la incidencia de las 118 especies en estos sitios, solo 93 de las 118 especies del *pool* considerado están presentes en al menos uno de estos sitios. Esto se debe a que para el número de sitios en el estudio completo (Standley et al, 1996) se consideraron 391 sitios, que cubren completamente el condado de Berkshire (Inglaterra), donde las 118 especies de la lista estaban presentes en al menos uno de los sitios. Por lo tanto, podríamos considerar también en nuestra muestra reducida un *pool* de 93 especies. Esto nos abre una discusión acerca de cómo exactamente definir el *pool* de especies, bajo un hábitat común. Postergaremos esta discusión para más adelante (Sección 4.4), la cual no es precisamente un tema que haya sido "zanjado" definitivamente dentro del ámbito ecológico. Mostraremos, sin embargo, los resultados de los ajustes de nuestro modelo beta considerando ambos *pools*, asumiendo que la dinámica de especies se encuentra en un estado estacionario (Figura 4.1).

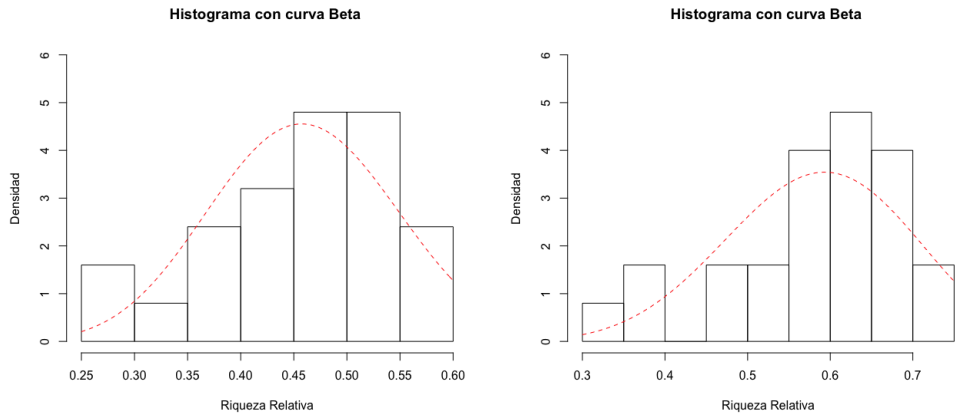


Figura 4.1: Histogramas de los datos tomados de Gaston y Blackburn (2000), Apéndice IV: *"Berkshire Breeding Bird Data"*. Obtuvimos la riqueza relativa muestral en cada uno de los 25 sitios, dividiendo la riqueza (absoluta) por el *pool* $= 118$ (izquierda), y por el *pool* $= 93$ (derecha). En rojo tenemos sobrepuesta la curva de densidad Beta teórica respectiva. Utilizando estimación de máxima verosimilitud, la estimación de los parámetros resultaron $\hat{\alpha} \approx 15,37$ y $\hat{\beta} \approx 18,10$ (izquierda), y $\hat{\alpha} \approx 12,34$ y $\hat{\beta} \approx 8,85$ (derecha). La prueba Kolmogórov-Smirnov dio un p-valor cercano a $0,72$ (izquierda), y cercano a $0,82$ (derecha).

4.3.1. Análisis datos distribución beta y beta cero-inflada

Para el ajuste de nuestro modelo beta y beta cero-inflada se utilizó cuatro bases de datos correspondiente a la riqueza de los grupos taxonómicos Líquenes, Fungis, Plantas y Briófitos obtenidas de *"Dataset on species incidence, species richness and forest characteristics in a Danish protected area"* (Mazziotta et al., 2016)

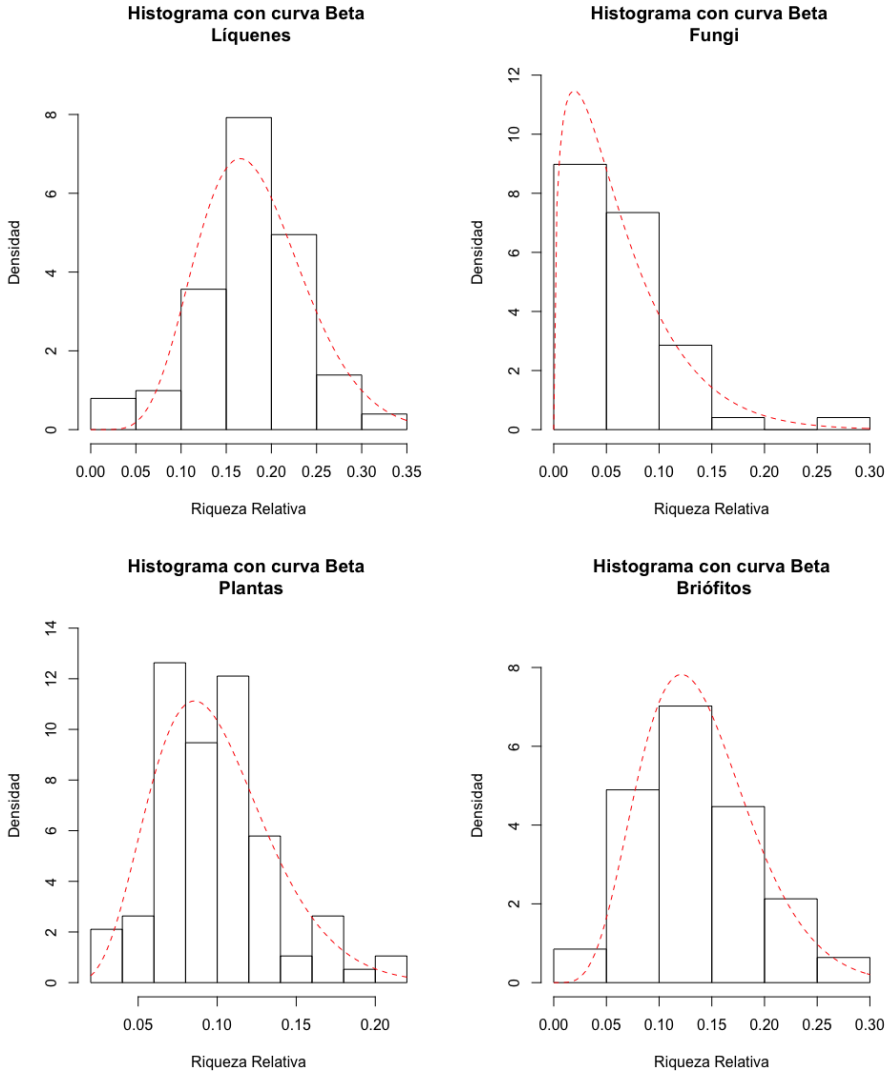


Figura 4.2: Histogramas de los datos tomados de *"Dataset on species incidence, species richness and forest characteristics in a Danish protected area"*. Para el caso de los Líquenes, obtuvimos la riqueza relativa muestral dividiendo la riqueza (absoluta) por el $pool = 120$, en el caso de la base de datos de los Fungi, obtuvimos la riqueza relativa muestral dividiendo la riqueza (absoluta) por el $pool = 193$, en la que la proporción de ceros corresponde a 0,016 aproximadamente. Para la base de datos de las Plantas se obtuvo la riqueza relativa muestral dividiendo la riqueza (absoluta) por el $pool = 215$, en la que la proporción de ceros corresponde a 0,028 aproximadamente y finalmente para la base de datos de los Briófitos la riqueza relativa muestral se obtuvo dividiendo la riqueza (absoluta) por el $pool = 61$, en la que la proporción de ceros corresponde a 0,115 aproximadamente. En rojo tenemos sobrepuesta la curva de densidad Beta teórica respectiva. Utilizando estimación de máxima verosimilitud, la estimación de los parámetros resultaron $\hat{\alpha} \approx 7{,}64$ y $\hat{\beta} \approx 34{,}48$ en el caso de los Líquenes $\hat{\alpha} \approx 1,40$ y $\hat{\beta} \approx 21,31$ para el caso de los Fungi, $\hat{\alpha} \approx 6{,}21$ y $\hat{\beta} \approx 56{,}64$, para las Plantas y $\hat{\alpha} \approx 5{,}89$ y $\hat{\beta} \approx 36{,}46$ para los Briófitos. La prueba Kolmogórov-Smirnov dio un p-valor cercano a $0,05$; $0,47$;0,75 y 0,42 respectivamente.

4.3.2. Análisis datos distribución beta binomial

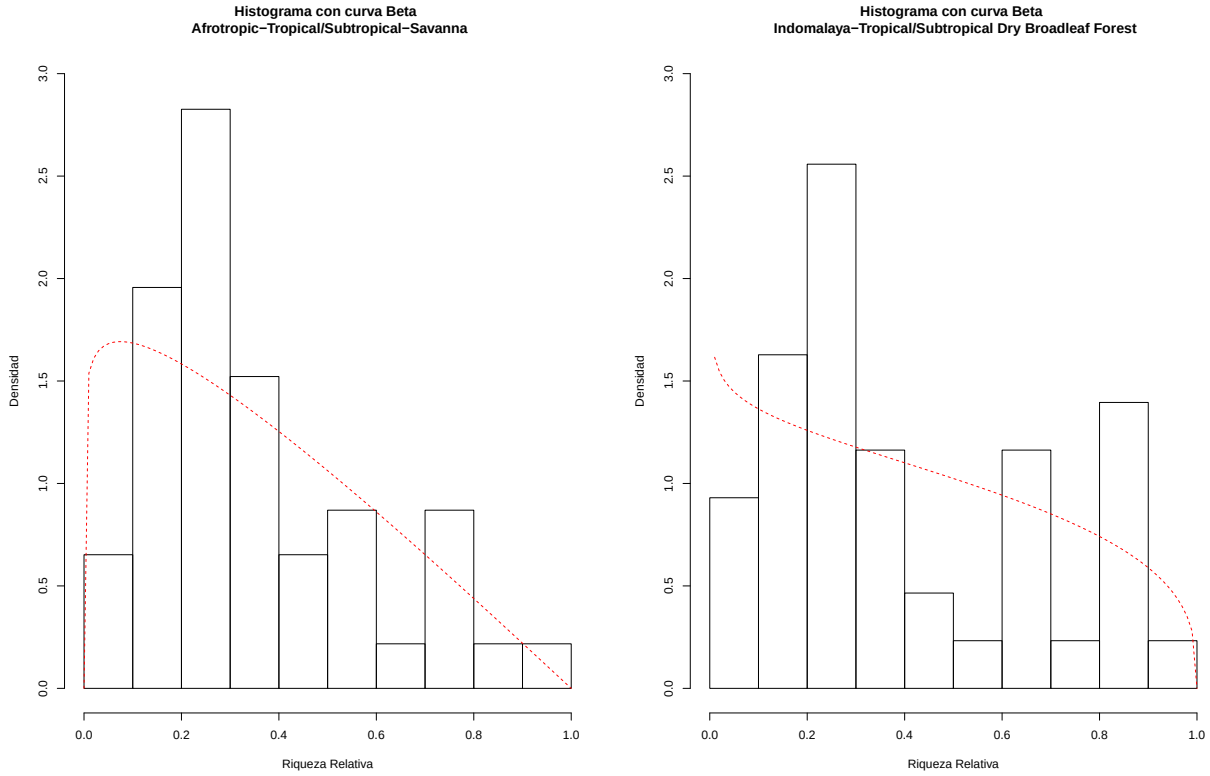
A continuación se muestran los resultados de los ajustes de nuestro modelo beta binomial, en los que se utilizó tres bases de datos correspondiente a la riqueza de mamíferos, obtenidas de *"The Ecological Register"* R: http://ecoregister.org. Las bases fueron clasificadas en tres grupos de acuerdo a su ecozone-hábitat:

(1) Afrotropic-tropical/subtropical savanna

(2) Indomalaya-tropical/subtropical dry broadleaf forest

(3) Indomalaya-tropical/subtropical moist broadleaf forest

Se consideró los mamíferos, porque habían más datos de estos que las otras taxas, y al hacer esta clasificación con las demás taxas, se reducían bastante los datos como para hacer una buena prueba de bondad de ajuste. Así, con estas tres bases de datos, se pudo realizar la estimación de la distribución de la riqueza relativa en conjunto con la estimación del *pool*, obteniendo así:



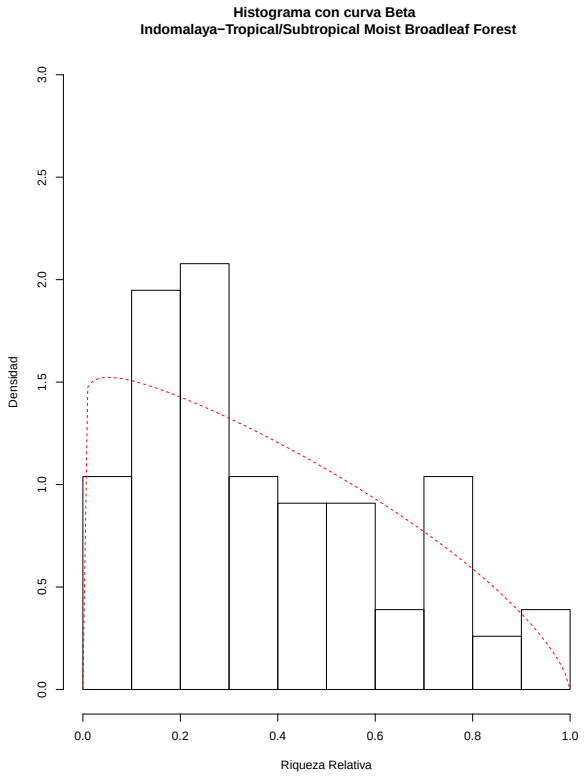


Figura 4.3: Histogramas de los datos tomados de "The Ecological Register". Para el caso de los mamíferos pertenecientes a la ecozone-hábitat (1) se obtuvo un $pool = 39$, para el (2) un $pool = 30$ y para (3) un $pool = 35$. Luego, con la estimación del *pool* se pudo dividir la riqueza (absoluta), obteniendo así la riqueza relativa. En rojo tenemos sobrepuesta la curva de densidad Beta teórica respectiva. Utilizando estimación de máxima verosimilitud, la estimación de los parámetros resultaron $\hat{\alpha} \approx 1,29$ y $\hat{\beta} \approx 2,18$ para la ecozone-hábitat (1), $\hat{\alpha} \approx 1,16$ y $\hat{\beta} \approx 1,52$ para la ecozone-hábitat (2), $\hat{\alpha} \approx 1,21$ y $\hat{\beta} \approx 1,84$ para la ecozone-hábitat (3). La prueba Kolmogórov-Smirnov dio un p-valor cercano a $0,17$; $0,12$ y $0,18$ respectivamente.

4.3.3. Conclusiones de los resultados obtenidos

Dados los resultados de bondad de ajuste obtenidos en las diferentes bases de datos y para los distintos grupos taxonómicos considerados, podemos concluir que los modelos estadísticos propuestos derivados de la distribución beta son bastante aceptables. Si bien para los grupos taxonómicos de los fungi, plantas y briófitos los p-valores (esto es, la probabilidades de observar un valor tan extremo en la prueba de Kolmogórov-Smirnov como el observado dado que H_0 [la distribución de los datos sigue una beta] es cierta) resultaron significativos al nivel del 5 %, para el caso del el grupo taxonómico de los líquenes y mamíferos, resultaron significativos al nivel del 1 %.Por ende, contamos con un sustento empírico para la teoría propuesta de la cual se dedujeron nuestros modelos.

Ahora bien, es importante mencionar el problema del muestreo subyacente, el cual dependiendo de la taxa objetivo, y por ende el territorio (metacomunidad) abarcado, el esquema de muestro puede variar, lo cual puede provocar que se den ajustes que "no sean tan buenos". Por lo que evaluar este impacto es también un tema abierto en este contexto.

4.4. Discusión y futuras consideraciones

Dado los resultados estadísticamente favorables observados en los análisis de datos, contamos con un sustento empírico de nuestro modelo teórico beta. Sin embargo, existen una serie de consideraciones relativos al concepto de *pool* en toda su extensión. Por su parte, MacArthur & Wilson en su teoría de biogeografía insular asumen la existencia de este en una metacomunidad focal como una constante fija en el tiempo. Este supuesto lo argumentan en parte debido a la diferencia de escalas temporales en ciertos procesos ecológicos. Por un lado, los tiempos evolutivos en los cuales pueden aparecer mutaciones. y por ende, nuevas especies, están en una escala temporal mayor a la de la estabilización de la metacomunidad. Y por otro lado, cambios drásticos en los factores abióticos que pueden interceder notablemente en cambios en la composición del *pool* de especies de una metacomunidad no son considerados. Este último punto es de crucial importancia, dado que el actual avance del calentamiento global ha provocado un gradiente de cambio climático, que se traduce en el desplazamiento geográfico de condiciones de hábitats para las distintas especies, y por tanto, en sus condiciones de nicho. Así, las especies tenderían en migrar en busca de sus condiciones de nicho, a menos que esta pueda adaptarse, proceso que suele ocurrir en una escala temporal mayor [ver *"Single species dynamics under climate change. Theoretical Ecology"*]. Por todo esto, el supuesto de la existencia del *pool* como parámetro constante, puede ser discutible.

En cuanto al problema de estimación de un *pool*, tenemos que ninguno de los métodos mencionados "distingue" una unidad (especie) de otra. Por tanto, a parte de dar una correcta definición y/o delimitación de este debido a lo mencionado en el párrafo anterior, tenemos el problema que no es posible identificar las distintas especies existentes dada la información cuantitativa de la riqueza. Así, que el modelo beta propuesto más bien nos proporciona un marco general por donde comenzar a modelar la riqueza de especies, tomando con precaución cualquier relación de causalidad, y siendo en ningún caso una "respuesta definitiva" a la interrogante planteada en Dobson, Holt, y Tilman (2020) ¿Cuál es la distribución de la riueza de especies? (*What is the species richness distribution?*)

Referencias

Aldirawi, H., Yang, J., y Metwally, A. A. (2019). Identifying appropriate probabilistic models for sparse discrete omics data. En *2019 ieee embs international conference on biomedical & health informatics (bhi)* (pp. 1–4).

Barton, D., y David, F. (1959). The dispersion of a number of species. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(1), 190–194.

Casella, G., y Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

De Vries, G., Hillen, T., Lewis, M., Müller, J., y Schönfisch, B. (2006). *A course in mathematical biology: quantitative modeling with mathematical and computational methods*. SIAM.

Dobson, A., Holt, R. D., y Tilman, D. (2020). *Unsolved problems in ecology*. Princeton University Press.

Gaston, K. J., y Blackburn, T. M. (2000). *Pattern and process in macroecology* (Vol. 414). Wiley Online Library.

Goel, N. S., y Richter-Dyn, N. (2016). *Stochastic models in biology*. Elsevier.

Hubbell, S. P. (2011). *The unified neutral theory of biodiversity and biogeography (mpb-32)*. Princeton University Press.

Karatzas, I., y Shreve, S. (2012). *Brownian motion and stochastic calculus* (Vol. 113). Springer Science & Business Media.

Kendall, D. G. (1948). On some modes of population growth leading to ra fisher's logarithmic series distribution. *Biometrika*, *35*(1/2), 6–15.

MacArthur, R. H., y Wilson, E. O. (1963). An equilibrium theory of insular zoogeography. *Evolution*, 373–387.

MacArthur, R. H., y Wilson, E. O. (1967). *The theory of island biogeography*. Princeton University Press.

Marquet, P. A., Espinoza, G., Abades, S. R., Ganz, A., y Rebolledo, R. (2017). On the proportional abundance of species: Integrating population genetics and community ecology. *Scientific reports*, *7*(1), 1–10.

Marquet, P. A., Tejo, M., y Rebolledo, R. (2020). What is the species richness distribution? En *Unsolved problems in ecology* (pp. 177–188). Princeton University Press.

Mazziotta, A., Heilmann-Clausen, J., Bruun, H. H., Fritz, Ö., Aude, E., y Tøttrup, A. P. (2016). Dataset on species incidence, species richness and forest characteristics in a danish protected area. *Data in brief*, *9*, 895–897.

Øksendal, B. (2003). Stochastic differential equations. En *Stochastic differential equations* (pp. 65–84). Springer.

Ospina, R., y Ferrari, S. L. (2010). Inflated beta distributions. *Statistical papers*, *51*(1), 111–126.

Pielou, D., y Pielou, E. (1967). The detection of different degrees of coexistence. *Journal of theoretical biology*, *16*(3), 427–437.

Pielou, E. (1975). Ecological diversity new york. *NY: Wiley [Google Scholar]*.

Strong Jr, D. R. (1982). Harmonious coexistence of hispine beetles on heliconia in experimental and natural communities: Ecological archives e063-003. *Ecology*, *63*(4), 1039–1049.

Volkov, I., Banavar, J. R., Hubbell, S. P., y Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, *424*(6952), 1035–1037.