"El saber de mis hijos hará mi grandeza"

UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Programa de Posgrado en Matemáticas

Estimación de abundancia de animales a través de un modelo Binomial jerárquico con covariables

T E S I S

Que para obtener el título de:

Maestra en Ciencias
(Matemáticas)

Presenta:

Arling Vázquez Alcántar

Director de Tesis: Dr. José Arturo Montoya Laos

Hermosillo, Sonora, México, 23 de agosto, 2016.

SINODALES

Dr. Daniel Olmos Liceaga
Universidad de Sonora

Dra. Gudelia Figueroa Preciado
Universidad de Sonora

Dr. José Arturo Montoya Laos
Universidad de Sonora

Dr. Juan Manuel Preciado Rodríguez
Centro de Investigación en Alimentación y Desarrollo, A. C.

Agradecimientos

Por el apoyo económico brindado durante la realización de este proyecto agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT).

A mi director de tesis Dr. José Arturo Montoya Laos por sus ideas y el tiempo invertido en este trabajo.

Al jurado conformado por Dra. Gudelia Figueroa Preciado, Dr. Daniel Olmos Liceaga y Dr. Juan Manuel Preciado Rodríguez por aceptar ser parte de este proyecto y sus aportaciones al mismo.

A mi familia por su apoyo. A Alicia Franco por su paciencia.

A mis colegas y amigas: Alma Padilla, Angélica Moreno, Carol Corral y Daniela Romero.

A la familia Romero Robles, en especial a María del Rosario Robles.

A mi compañeros: Marla Arcega, Alejandro Orozco y Dante Terán.

A Datos y Cifras y al equipo Dayci: Juan Pablo Piñeda, Fernanda Miranda, Jesús Rentería, Verónica Andrade, Mayra Valencia, Carolina Granados, Lemuel Anduaga, Ricardo Wendlandt, Iván Chairez, Karla Vieyra, Celia Beauregard, David Fimbres, Luis López, Dulce Valencia y Finita.

...

Contenido

Prefacio **7**

1 Motivación e introducción al problema **9**

1.1 Introducción 9

1.2 Protocolos de muestreo y modelo estadístico 10

1.3 Problemáticas generales 11

1.4 Problema de tesis 13

2 El problema de estimar abundancia de animales usando el modelo Binomial **15**

2.1 Introducción 15

2.2 Problema de identificabilidad 19

2.3 Uso de información *a priori* para estimar la abundancia de animales . . 23

2.3.1 Función de verosimilitud integrada 23

2.3.2 Densidad marginal posterior Bayesiana 24

2.3.3 Ejemplo 25

3 Modelo Binomial jerárquico con covariables **29**

3.1 Introducción 29

3.2 Modelo Binomial jerárquico 30

3.2.1 Distribución marginal 31

3.2.2 Uso de covariables 32
3.2.3 Reparametrización $(a_i, b_i, \lambda_i) \leftrightarrow (\alpha, \beta, \phi)$ 33
3.3 Heurística del modelo 34

4 Inferencia estadística **37**
4.1 Estimación puntual de parámetros 37
4.2 Regiones de confianza 38
4.3 Regiones de incertidumbre 39
4.4 Caso sintético 41
4.5 Caso real: *Ambystoma ordinarium* 45

5 Conclusiones **55**

Índice de figuras

2.1 Función de verosimilitud perfil relativa de n. 22

2.2 Función de verosimilitud perfil y función de verosimilitud integrada. 26

3.1 Distribución verdadera de la abundancia. 35

3.2 Distribución verdadera de la probabilidad de detección. 36

3.3 Probabilidad marginal verdadera. 36

4.1 Verosimilitud perfil de α: Caso sintético. 42

4.2 Verosimilitud perfil de β: Caso sintético. 42

4.3 Verosimilitud perfil de ϕ: Caso sintético. 43

4.4 Función de probabilidad condicional Región 1: Caso sintético. 44

4.5 Función de probabilidad condicional Región 2: Caso sintético. 44

4.6 Función de probabilidad condicional Región 3: Caso sintético. 45

4.7 Verosimilitud perfil de α: *Ambystoma ordinarium*. 48

4.8 Verosimilitud perfil de β: *Ambystoma ordinarium*. 48

4.9 Verosimilitud perfil de ϕ: *Ambystoma ordinarium*. 49

4.10 Verosimilitud perfil de α: *Ambystoma ordinarium*, $\phi = 0$. 50

4.11 Verosimilitud perfil de β: *Ambystoma ordinarium*, $\phi = 0$. 51

4.12 Contornos de la función de verosimilitud relativa de α y β: *Ambystoma ordinarium*, $\phi = 0$. 51

4.13 Función de densidad de probabilidad de N: *Ambystoma ordinarium*, $\phi = 0$. . 53

4.14 Función de densidad de probabilidad de P: *Ambystoma ordinarium*, $\phi = 0$. . 53
4.15 Función de probabilidad Marginal: *Ambystoma ordinarium*, $\phi = 0$. 54

Índice de tablas

2.1 Datos simulados de una variable aleatoria Binomial. 21

2.2 EMV para la muestra original ($\hat{n}_1$) y para la muestra perturbada ($\hat{n}_2$). . . . 22

3.1 Escenario de simulación y datos simulados. 34

4.1 Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: Caso sintético. . . 41

4.2 Regiones de incertidumbre para el valor espeado de N dada la muestra $X = x_{\text{máx}}$: Caso sintético. 43

4.3 Datos de conteo de Salamandra Michoacana de arroyo. 47

4.4 Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: *Ambystoma ordinarium*. 47

4.5 Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: *Ambystoma ordinarium*, $\phi = 0$. 50

4.6 Regiones de incertidumbre para el valor esperado N dada la muestra $X_{\text{máx}}$: Caso *Ambystoma ordinarium*, $\phi = 0$. 52

Prefacio

El problema de estimar abundancia de especies es un tema de interés en estudios ecológicos. Conocer el número de individuos de cierto taxón proporciona información valiosa para la clasificación de especies según su riesgo de extinción; además, estimadores de la abundancia pueden ayudar a detectar cambios de la población en el tiempo o espacio y a detectar hábitats.

Conocer el número exacto de individuos de una población podría ser imposible; por lo general, se recurre a estudios donde se lleva a cabo un muestreo que conlleva sus propias dificultades teóricas y prácticas como captura de individuos, marcas adecuadas, natalidad, mortalidad, migración, entre otras.

La distribución Binomial (n, p) es una elección natural para modelar datos de conteos en estudios de estimación abundancia de especies; en este trabajo n es el número de individuos de la especie de interés y p es la probabilidad de detectar o capturar un ejemplar de la especie, ambos parámetros se consideran desconocidos. Realizar estimaciones del modelo Binomial cuando ambos parámetros son desconocidos se considera una tarea difícil según la literatura estadística y ha sido abordada desde 1968.

En esta tesis se enfatiza que muchos de los problemas asociados con la estimación del parámetro n son causados por la falta de identificabilidad del modelo y para abordar este problema se sugiere incorporar covariables, información adicional a los datos de conteo, por medio de un modelo jerárquico. En particular, se propone un modelo Binomial jerárquico donde las covariables están ligadas a los valores esperados de

n y p, consideradas como variables independientes con distribución Poisson y Beta, respectivamente. Este modelo es analizado y utilizado para propósitos de inferencia considerando un escenario simulado. También, el modelo propuesto es aplicado a datos de conteos de salamandras *(Ambystoma ordinarium)* con el objetivo de mostrar la utilidad del modelo en un caso real y el uso de la metodología de inferencia estadística desarrollada en este trabajo.

La tesis esta estructurada de la siguiente forma. En el Capítulo 1 se expone la importancia de estimar abundancia de especies desde un punto de vista ecológico; además, se describen algunos protocolos de muestreo junto con algunos problemas que conlleva la ejecución de los mismos. Finalmente, se describe el problema de tesis. En el Capítulo 2 se exhibe el modelo Binomial (n, p) como una elección natural para modelar datos de conteos, se expone que realizar estimaciones cuando ambos parámetros son desconocidos es una tarea difícil y que se debe a la falta de identificabilidad del modelo; también se muestra que el método bayesiano y la verosimilitud integrada son dos enfoques estadísticos que permiten abordar el problema de identificabilidad a través del uso de información *a priori* sobre los parámetros del modelo. Se ilustra con un caso sintético el uso de la verosimilitud integrada. En el Capítulo 3 se define un modelo Binomial jerárquico y se sugiere el uso de covariables como una opción para agregar información externa que permita hacer inferencias útiles sobre parámetros de interés. Además, se presenta una reparametrización apropiada y se explora el modelo propuesto mediante procedimientos heurísticos. En el Capítulo 4 se presenta la metodología estadística de verosimilitud para hacer inferencia sobre los parámetros de interés, se exponen los estimadores puntuales, regiones de confianza y regiones de incertidumbre para el modelo jerárquico propuesto. Finalmente, este modelo es utilizado para propósitos de inferencia en dos casos: El primero con datos simulados y el segundo con datos reales de conteos de salamandras *(Ambystoma ordinarium)*. Por último, en el Capítulo 5 se exponen las conclusiones de este trabajo de tesis.

Capítulo 1

Motivación e introducción al problema

1.1 Introducción

La Lista Roja de Especies Amenazadas es un inventario del estado de conservación de animales y plantas a nivel mundial que es elaborada por la Unión Internacional para la Conservación de la Naturaleza (UICN). Su objetivo es dar a conocer la urgencia de los problemas de conservación y ayudar a la comunidad a reducir la extinción de especies. Cuenta con un conjunto de criterios, aplicables a prácticamente todos los taxones del planeta, para evaluar el riesgo de extinción de miles de especies y subespecies.

Actualmente, la Lista Roja considera nueve categorías estructuradas de la siguiente manera, de mayor a menor riesgo: Extinta (EX), extinta en estado silvestre (EW), en peligro crítico (CR), en peligro (EN), vulnerable (VU), casi amenazada (NT), preocupación menor (LC), datos insuficientes (DD) y no evaluado (NE). Esta última corresponde a las especies no evaluadas en ninguna de las otras categorías. Para asignar a una especie a cierta categoría es necesario evaluar una serie de criterios, entre ellos se encuentra la reducción de la población observada, estimada, inferida o sospechada en

algún periodo establecido. En este contexto, conocer o estimar el número de individuos de una especie es un dato relevante para valorar su estado de conservación.

Por otro lado, estimadores de la abundancia se pueden usar para detectar cambios de la población entre sitios o en el tiempo; así como para identificar el ambiente que ocupa una población biológica (hábitat).

1.2 Protocolos de muestreo y modelo estadístico

Con el propósito de conocer el número de individuos en cierta población podría llevarse a cabo un conteo completo o censo; sin embargo, en la mayoría de los casos esto no es posible y se realiza un estudio cuyo objetivo es obtener una estimación del tamaño de la población con base en una muestra. Si la región es muy amplia se pueden seleccionar ciertos sitios para muestrear. El criterio de selección para tales sitios o zonas geográficas es generalmente su representatividad y por ser de fácil acceso.

Existen distintos métodos de muestreo para obtener información acerca de una población y estimar su abundancia o densidad, algunos de ellos son: Muestreo simple, captura-recaptura, muestreo por remoción, entre muchos otros. Por ejemplo, la técnica de captura-recaptura (en dos etapas) consiste en visitar un área de estudio y capturar individuos vivos de la especie de interés. Los elementos capturados se marcan de forma individual para su futura identificación y posteriormente son liberados. Una vez que el grupo se ha distribuido entre los individuos no marcados, el investigador realiza una segunda visita al área para atrapar a otro grupo; entre éstos puede haber elementos marcados y no marcados. Por otro lado, la técnica de muestreo por remoción consiste en capturar individuos de la especie bajo estudio en una o varias ocasiones sin regresar a los elementos a su entorno.

El protocolo de muestreo da lugar a un modelo estadístico, que se denominará modelo protocolar por la naturaleza de su origen. En la literatura dicho modelo recibe

otros nombres como modelo observacional (Royle y Dorazio, 2006), modelo de los datos (Dorazio *et al.*, 2008) o modelo de conteo (Dorazio *et al.*, 2005).

Usualmente el modelo protocolar es una distribución multinomial o producto de binomiales. En su forma mas sencilla el modelo protocolar depende de dos parámetros desconocidos, n y p, el número de individuos de la especie bajo estudio y la probabilidad de que un individuo sea detectado. Sin embargo, algunos protocolos de muestreo como observadores múltiples, captura-recaptura y muestreo por remoción pueden conducir a un modelo protocolar multinomial mucho más complejo (más de dos parámetros); por ejemplo, con más estructura matemática en la probabilidad de detección y el uso de covariables para describir la variación en la abundancia entre sitios (Royle y Dorazio, 2006; Williams *et al.*, 2002).

1.3 Problemáticas generales

La aplicación de protocolos de muestreo y métodos de estimación con base en datos obtenidos bajo estos protocolos conlleva dificultades. Por ejemplo, el protocolo de captura-recaptura puede proveer información adecuada para calcular estimadores de abundancia; sin embargo, existen dificultades técnicas y prácticas para su adecuada realización tales como: La captura del animal, una marca adecuada, entre otras. Aunado a esto los métodos de estimación de tamaños de población, basados en datos obtenidos bajo el protocolo de captura y recaptura, exigen suposiciones matemáticas que muchas veces no ocurren ni de manera aproximada en la realidad: Todos los individuos de la población tienen la misma probabilidad de ser capturados, la proporción de animales marcados respecto a los no marcados se mantiene constante a lo largo del tiempo (desde el momento de la captura hasta el momento de la recaptura), los individuos marcados, una vez liberados, se redistribuyen de manera homogénea entre la población de individuos no marcados, los animales marcados no pierden sus marcas, la

población es cerrada. Además, si la población es pequeña o la probabilidad de captura es baja posiblemente se obtengan una cantidad insuficiente de datos para llegar a una buena estimación del tamaño de la población (Royle, 2004).

Cabe mencionar aquí que la probabilidad de detección de un animal en un área de muestreo durante un estudio consiste de dos componentes: La probabilidad de que un animal este diponible para su detección, la cual puede ser altamente variable en ambientes heterogeneos, y la probabilidad de que un animal sea detectado, condicionado a que esté disponible para su detección. Muchos estudios sólo estiman la probabilidad anterior porque modelar la disponibilidad requiere información externa al estudio (Pollock *et al.*, 2006).

La baja detectabilidad es otra de las dificultades que se presentan en estudios ecológicos; específicamente, en aquellos sobre estimación de abundancia de especies. Con frecuencia el número de animales presentes y disponibles a ser detectados en un sitio específico es bajo, lo que casi siempre acarrea que en el conteo se observarán pocos o cero individuos. Existe ambigüedad en los ceros observados ya que pueden significar que en esa área no hay animales o que su detectabilidad es tan baja que no fue posible observarlos. Esta situación es común cuando la especie bajo estudio está en peligro de extinción (Dorazio, 2007). Es importante considerar que los individuos pueden variar su detectabilidad en el espacio y en el tiempo; esto último puede darse por el camuflaje, la técnica o capacidad de ciertos taxones de pasar inadvertidas ante los sentidos de otra especie.

Realizar estimaciones con datos obtenidos en estudios con baja detectabilidad, pueden ocasionar sesgo negativo o subestimación. Este sesgo puede ser ignorado si es razonable suponer que la tasa de captura o detección es la misma en todas las áreas y todos los tiempos, o si las covariables (humedad, temperatura, vegetación, altura, entre otras) que pueden causar la variación en la detectabilidad pueden identificarse y sus efectos pueden ser modelados.

El problema de hacer estimaciones adecuadas de la abundancia requiere de un modelo que ligue parámetros de abundancia y detección de diferentes unidades muestrales. En estadística, una clase de modelos que cumplen con esta característica son los modelos jerárquicos; un modelo jerárquico toma en cuenta la variación en la abundancia local y la variabilidad en la detección a través del uso de covariables.

1.4 Problema de tesis

Estimar la abundancia de especies es un problema relevante en áreas como ecología, especialmente cuando se requiere evaluar el estado de conservación una especie, para definir estrategias adecuadas de manejo de las mismas o identificar su hábitat. El modelo Binomial (n, p) es un ejemplo de modelo estadístico comunmente utilizado para estimar abundancia de especies a través de datos de conteo. Bajo este modelo se supone que la abundancia n de la región es constante durante el estudio y que la probabilidad de detección p es la misma para cada individuo dentro de la región de interés.

Realizar estimaciones del modelo Binomial cuando ambos parámetros son desconocidos es considerado un problema difícil en la literatura estadística y ha sido abordado por Feldman y Fox (1968), Draper y Guttman (1971), Olkin *et al.* (1981), Carroll y Lombard (1985), entre otros. En Montoya (2008) se puede consultar la descripción y explicación de un conjunto de referencias bibliográficas relacionadas con el problema de estimación del parámetro n del modelo Binomial (n, p). Recientemente el problema ha sido estudiado por Bayoud (2011).

En esta tesis se enfatiza que las dificultades asociadas con la estimación del parámetro n, con base en una muestra Binomial (n, p), son causadas por un problema de falta de identificabilidad del modelo Binomial. Para abordar este nuevo problema se propone incorporar datos de covariables, información adicional a los datos de conteos, a través

de una propuesta de modelo jerárquico. Además, se usa el enfoque de verosimilitud para hacer inferencia sobre los parámetros de este modelo. Se usan datos simulados del modelo jerárquico y datos reales de conteos de salamandras *(Ambystoma ordinarium)* para mostrar el uso del modelo propuesto y la metodología de inferencia.

Capítulo 2

El problema de estimar abundancia de animales usando el modelo Binomial

En este capítulo se presenta la distribución Binomial de parámetros n y p, una de las distribuciones más simples e importantes en Probabilidad y Estadística y elección natural para modelar datos de conteos en estudios de estimación de abundancia de animales. Además, se exhiben problemas que surgen a la hora de querer estimar el parámetro n cuando p es desconocida. Se mostrará que muchos de estos problemas son causados por la falta de identificabilidad del modelo Binomial. Por último, se enfatizará la necesidad de incorporar información adicional a la suministrada por los datos de conteos, en pro de eliminar problemas de estimación del modelo Binomial.

2.1 Introducción

Sea X el número de éxitos en una sucesión de n ensayos independientes de Bernoulli con probabilidad de éxito p. Entonces, X es una variable aleatoria Binomial con función

de densidad de probabilidad

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} I_{\mathcal{X}}(x), \tag{2.1}$$

donde $I_{\mathcal{X}}(\cdot)$ es la función indicadora del conjunto $\mathcal{X} = \{0, 1 \ldots, n\}$ y $p \in [0, 1]$. En lo que sigue, se usará la notación $X \sim \text{Binomial}(n, p)$ para indicar que X es una variable aleatoria Binomial de parámetros n y p.

La distribución Binomial es quizás una de las distribuciones de probabilidad más simples e importantes en Probabilidad y Estadística. La Binomial es útil para modelar diversos fenómenos aleatorios de la vida real; en particular, una variable aleatoria Binomial se puede usar para modelar datos de conteos que surgen en el problema de estimación de abundancia de animales, (Carroll y Lombard, 1985; Olkin *et al.*, 1981; Royle y Dorazio, 2006). Por ejemplo, cuando se usa el modelo Binomial en el problema de estimación de abundancia de animales, el parámetro n representa el tamaño de la población en la región geográfica bajo estudio y p la probabilidad de que un individuo sea detectado. Nótese que, en teoría, la probabilidad p de detección es la misma para cada individuo y el tamaño n de la población es constante durante el periodo de estudio.

El modelo Binomial está indexado por dos parámetros, n y p, lo que da lugar a distintos problemas de estimación. Uno de ellos consiste en suponer que se conoce el tamaño de la población $n = n_0$ y seguir el propósito de estimar la probabilidad de detección p. Realizar estimaciones del parámetro p cuando n es conocida es un problema estándar asociado a la distribución Binomial y resulta ser una tarea sencilla. Sin embargo, el problema que se estudia en este trabajo consiste en estimar el tamaño de la población, evidentemente desconocida. Así, considerando a n como un parámetro desconocido y de interés, se tienen dos posibles situaciones o casos.

Caso 1: Parámetro p conocido. El caso donde la probabilidad de detección es conocida, $p = p_0$, es artificial y difícilmente se da en la vida real. Aún así, si se diera este caso, las inferencias sobre n no representan problema alguno (Montoya, 2004).

Caso 2: Parámetro p desconocido. Este caso es el que más se asemeja a la vida real y será objeto de estudio en este trabajo.

Obsérvese que bajo el modelo Binomial, el problema real de estimar el tamaño de una población consiste en estimar el parámetro n, involucrado directamente en este modelo de probabilidad. Al parecer, esta parametrización hace que la distribución Binomial sea, dentro de las distribuciones de probabilidad para datos de conteos, una elección natural cuando se desea estimar abundancia de animales.

El problema de estimar el parámetro n de la ditribución Binomial ha sido abordado desde distintos puntos de vista y en muchos de ellos se utiliza la función de verosimilitud. Fisher (1921) define la verosimilitud como una cantidad proporcional a la probabilidad de obtener la muestra observada; pero vista como función de los parámetros del modelo. Así, si $x = (x_1, \ldots, x_k)$ es una muestra observada de un vector $X = (X_1, \ldots, X_k)$ de variables aleatorias independientes con distribución Binomial de parámetros desconocidos n y p, entonces la función de verosimilitud de los parámetros n y p es

$$L(n, p; x) \propto \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^t (1-p)^{nk-t} I_{[x_{\text{máx}}, \infty)}(n)\, I_{[0,1]}(p)\,, \tag{2.2}$$

donde $x_{\text{máx}} = \text{máx}\{x_1, \ldots, x_k\}$ y $t = \sum_{i=1}^{k} x_i$.

La función de verosimilitud juega un papel fundamental en la Inferencia Estadística. En el caso Binomial, su rol principal es inferir qué valores de n y p de la función de densidad de probabilidad $f(x; n, p)$, dada en (2.1), son razonables a la luz de la muestra observada (datos). Nótese que esto es particularmente relevante después de que el protocolo de muestreo fue llevado a cabo y la muestra ya fue observada.

En problemas donde se desea estimar n, el tamaño de una población de animales, la probabilidad de detección p se puede considerar como un parámetro de estorbo. Es importante aclarar que el concepto parámetro de estorbo es relativo ya que en ocasiones lo que es parámetro de estorbo para alguien puede ser de interés para otro. Una manera de realizar estimaciones sobre n considerando a la probabilidad de detección p como

un parámetro de estorbo es a través de la función de verosimilitud perfil.

La función de verosimilitud maximizada o perfil es un método estadístico muy simple que sirve para estimar por separado un parámetro de interés en presencia de parámetros de estorbo. Fue presentada formalmente como un método para tal propósito en Kalbfleisch y Sprott (1970). La verosimilitud perfil de un parámetro de interés se obtiene maximizando la función de verosimilitud sobre el parámetro de estorbo; pero manteniendo fijo el parámetro de interés. Es decir, los parámetros de estorbo son eliminados a través de un proceso de maximización. En el caso Binomial, la función de verosimilitud perfil del parámetro de interés n se obtiene maximizando la función de verosimilitud $L(n, p; x)$, dada en (2.2), sobre p; pero manteniendo fijo n. Es fácil demostrar que el valor de p que maximiza $L(n, p; x)$ para cada valor fijo de n es $\hat{p}\,(n) = t/nk$. Así, la función de verosimilitud perfil de n es

$$\begin{aligned} L_P\,(n; x) &= L\,[n, p = \hat{p}\,(n)\,; x] \\ &\propto \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] \left(\frac{t}{nk}\right)^{t} \left(1 - \frac{t}{nk}\right)^{nk-t} I_{[x_{\text{máx}},\infty)}\,(n)\,. \end{aligned} \tag{2.3}$$

Como se mencionó anteriormente, el problema de estimar el parámetro n de la ditribución Binomial ha sido abordado con distintos métodos de estimación, entre ellos el de verosimilitud perfil, verosimilitud integrada, el método de momentos y el Bayesiano. Este problema ha recibido atención especial puesto que inferencias sobre n puden resultar absurdas y ser consideradas inestables frente a ligeros cambios en los datos de conteos. Montoya (2004) hace una revisión histórica de este problema y da evidencia de que la génesis del problema de estimación es la falta de identificabilidad del modelo Binomial, la cual será explicada en la siguiente sección.

2.2 Problema de identificabilidad

Considérese un vector de variables aleatorias X con función de distribución $F(x;\theta)$ que depende de un vector de parámetros desconocido θ. El vector θ es identificable por la observación de X si distintos valores de θ (θ_1 y θ_2) dan lugar a distintas distribuciones de X; es decir, si $\theta_1 \neq \theta_2$ entonces $F(x;\theta_1) \neq F(x;\theta_2)$ para algún x, (Bickel y Doksum, 1977).

Existen otras definiciones de identificabilidad menos generales que la mencionada en el párrarro anterior. Por ejemplo, el uso de la función de densidad o la esperanza de X en lugar de la función de distribución de X conducen a dos formas diferentes de definir identificabilidad; véase *International Encyclopedia of Statistical Science*. Con el objetivo de ilustrar de manera sencilla el problema de identificabilidad del modelo Binomial y por conveniencia matemática, se presenta la siguiente definición particular de identificabilidad.

Definición 1 *Considérese una variable aleatoria X con valor esperado $E(X;\theta)$ que depende de un vector de parámetros desconocido θ. Sean θ_1 y θ_2 dos valores de θ. Si $\theta_1 \neq \theta_2$ implica que $E(X;\theta_1) \neq E(X;\theta_2)$ para algún x, entonces el vector θ es identificable por la observación de X.*

A continuación se presentan dos ejemplos que ilustran el concepto de identificabilidad de parámetros; el primero corresponde a un modelo exponencial y el segundo al modelo Binomial.

Ejemplo 2 *Considérese una variable aleatoria X con función de distribución distribución acumulada dada por*

$$F(x;\lambda) = 1 - \exp\left(-\frac{x}{\lambda}\right),$$

donde $x \geq 0$ y $\lambda > 0$. Entonces, el valor esperado de X es $E(X;\lambda) = \lambda$. Es claro e inmediato que si λ_1 y λ_2 son dos valores distintos de λ entonces $E(X;\lambda_1) \neq E(X;\lambda_2)$.

Por lo tanto, de acuerdo a la Definición 1, el modelo exponencial es un ejemplo de un modelo estadístico cuyo parámetro λ es identificable.

Ejemplo 3 *Supóngase que $X \sim Binomial(n, p)$ con función de probabilidad dada en (2.1). Entonces, el valor esperado de X es $E(X; n, p) = np$. Sean (n_1, p_1) y (n_2, p_2) dos valores del vector de parámetros (n, p) tales que $p_1 = v/n_1$ y $p_2 = v/n_2$, con $0 < v < \min\{n_1, n_2\}$. Supóngase que $n_1 \neq n_2$. Entonces, (n_1, p_1) y (n_2, p_2) son dos valores diferentes del vector de parámetros (n, p) que satisfacen que $E(X; n_1, p_1) = E(X; n_2, p_2) = v$, para todo $0 < v < \min\{n_1, n_2\}$. Por lo tanto, de acuerdo a la Definición 1, el modelo Binomial es un ejemplo de un modelo estadístico cuyo vector de parámetros (n, p) es no identificable.*

Cuando el modelo es no identificable suelen existir complicaciones en el proceso de inferencia. Por ejemplo, si se cuenta con un juego de datos x_{obs}, los cuales se suponen observaciones de un vector aleatorio X con función de probabilidad conjunta $f(x; \theta)$, donde θ es un vector de parámetros en el espacio paramétrico Θ, entonces el hecho de que el modelo sea no identificable implica que puede existir θ_1 y θ_2 en el espacio paramétrico Θ, con $\theta_1 \neq \theta_2$, tales que $f(x_{obs}; \theta_1) = f(x_{obs}; \theta_2)$. Así, θ_1 hace a la muestra observada tan probable como la hace θ_2, y por lo tanto no es posible determinar, con base en los datos, cuál de los dos valores de θ es más razonable para el fenómeno aleatorio modelado a través de $f(x; \theta)$.

A continuación se usa la función de verosimilitud perfil y datos simulados para ilustrar de manera clara algunos de los problemas de estimación asociados al modelo Binomial. Considérese una variable aleatoria $X \sim$ Binomial(n, p). En la Tabla 2.1 se presentan muestras (de tamaño $k = 5$) simuladas de X con diferentes valores de n y p que satisfacen que $E(X; n, p) = 20$. Claramente se observa que las muestras simuladas correspondientes al caso (a), (b) y (c) son similares a pesar de que provienen de parámetros muy diferentes. La Figura 2.2 presenta, en la misma gráfica, la función de verosimilitud perfil de n correspondiente a cada caso. En esta figura se observa

que todas las perfiles de n son casi planas en un subconjunto muy grande del espacio parametral. Además, se observa que el valor de $n = \infty$ es altamente plausible o creible a la luz de la muestra observada. Claramente, esta estimación de n es inapropiada para los tres escenarios de simulación, donde el valor verdadero de n es finito; en particular, para el caso (a) donde $n = 100$. Nótese que no sería raro obtener este tipo de funciones de verosimilitud perfil (planas) con datos reales de conteos de animales; véase Montoya (2004). En estos casos, inferir que $n = \infty$ es altamente plausible sería seguramente absurdo.

Caso	N	p	$E(X; n, p)$	x_{obs}
(a)	100	0.2	20	(17, 18, 18, 24, 26)
(b)	1000	0.02	20	(16, 17, 20, 21, 25)
(c)	10000	0.002	20	(15, 17, 21, 21, 26)

Tabla 2.1: Datos simulados de una variable aleatoria Binomial.

Otro problema de estimación asociado al modelo Binomial es la inestabilidad en la estimación puntual de n. Olkin *et al.* (1981) ilustran este problema de la siguente manera. Primero, consideran una muestra original (datos simulados de una Binomial) y una muestra perturbada (se suma 1 a la observación más grande de la muestra original). Luego, tanto para la muestra original como para la muestra perturbada, calculan el valor numérico de un estimador puntual de n. Por último, ellos comparan ambas estimaciones y hacen notar que la estimación puntual de n (obtenida con la muestra original) puede cambiar dramáticamente con una pequeña perturbación de la muestra. Por ejemplo, la Tabla 2.2 presenta para cada escenario considerado en la Tabla 2.1, la muestra original, la muestra perturbada y el estimador de máxima verosimilitud de n correspondiente a cada caso, $\hat{n}_1$ para la muestra original y $\hat{n}_2$ para la muestra perturbada. Claramente se observa que $\hat{n}_1$ es "muy diferente" a $\hat{n}_2$ en particular en el caso (c), donde el estimador de máxima verosimilitud (EMV) de n

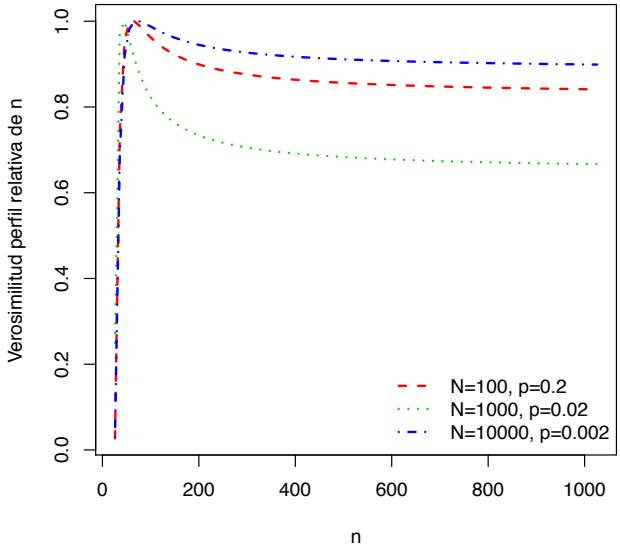

Figura 2.1: Función de verosimilitud perfil relativa de n.

cambia casi al doble al perturbar la muestra. Aquí es importante notar que la falta de estabilidad en la estimación puntual de n es causada por la forma casi plana de la función de verosimilitud de n.

Es relevante mencionar que problemas de estimación asociados al modelo Binomial no son causados, generalmente, por los métodos de estimación, sino por la poca información que proporciona la muestra para poder discernir qué valores de los parámetros

Caso	N	p	Muestra original	$\hat{n}_1$	Muestra perturbada	$\hat{n}_2$
(a)	100	0.2	(17, 18, 18, 24, 26)	61	(17, 18, 18, 24, 27)	90
(b)	1,000	0.02	(16, 17, 20, 21, 25)	41	(16, 17, 20, 21, 26)	56
(c)	10,000	0.002	(15, 17, 21, 21, 26)	71	(15, 17, 21, 21, 27)	131

Tabla 2.2: EMV para la muestra original ($\hat{n}_1$) y para la muestra perturbada ($\hat{n}_2$).

n y p son creibles. A continuación se presentan dos enfoques estadísticos de estimación que incorporan información adicional (que no está en los datos de conteos observados) sobre los parámetros y permiten eliminar problemas de estimación causados por la falta de identificabilidad del modelo Binomial. También se presentan algunas críticas que han recibido estos enfoques.

2.3 Uso de información *a priori* para estimar la abundancia de animales

La verosimilitud integrada y el método Bayesiano son dos enfoques estadísticos que permiten hacer inferencia sobre parámetros de interés en presencia de otros parámetros considerados de estorbo. Ambos enfoques incorporan infomación externa sobre los parámetros y serán descritos a continuación.

2.3.1 Función de verosimilitud integrada

La función de verosimilitud integrada es un método para eliminar parámetros de estorbo a través de integración y hacer inferencia sobre el parámetro de interés; véase Kalbfleisch y Sprott (1970) y Berger *et al.* (1999). En el caso Binomial, si el parámetro de estorbo p tiene especificada una función de densidad de probabilidad $\pi(p)$ entonces esta densidad previa puede multiplicarse con la función de verosimilitud $L(n, p; x)$ dada en (2.2) e integrarse respecto a p. Como resultado de este proceso se obtiene la función de verosimilitud integrada de n,

$$\begin{aligned} L_I(n; x) &\propto \int_0^1 L(n, p; x)\pi(p)dp \\ &\propto \int_0^1 \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^t(1-p)^{nk-t}\pi(p)I_{[x_{\text{máx}},\infty]}(x)dp, \end{aligned} \tag{2.4}$$

que sólo depende del parámetro de interés n ya que $\pi(p)$ es conocida. Nótese que sólo se incorpora información externa para el parámetro de estorbo a través de $\pi(p)$.

Carroll y Lombard (1985) proponen utilizar el método de verosimilitud integrada para estimar n. Suponen que $\pi(p)$ es una densidad Beta de parámetros a y b fijos y conocidos, tal que

$$L_I(n;x) \propto \left[\prod_{i=1}^{k}\binom{n}{x_i}\right]\left[\binom{nk+a+b+1}{t+a+1}\right]^{-1} I_{[x_{\text{máx}},\infty]}(x). \tag{2.5}$$

Ellos utilizan el valor que maximiza la función de verosimilitud integrada dada en (2.5) como un estimador puntual de n.

2.3.2 Densidad marginal posterior Bayesiana

En el enfoque estadístico Bayesiano se considera que el modelo observacional es condicionado sobre los parámetros y que todos ellos, tanto los de interés como los de estorbo, son variables aleatorias. En este enfoque de estimación se requiere una función de densidad conjunta inicial, previa o *a priori* para todos los parámetros. Luego, a través del teorema de Bayes, se calcula la probabilidad posterior conjunta de los parámetros dada la muestra observada. Las inferencias acerca de parámetros de interés se hacen a través de su densidad marginal posterior, la cual se obtiene integrando la probabilidad posterior conjunta de todos parámetros con respecto a los parámetros considerados de estorbo.

Por ejemplo, para el caso Binomial, se debe especificar una densidad conjunta inicial para los parámetros n y p, $\pi(n,p)$. Así, por el teorema de Bayes se tiene que la probabilidad posterior conjunta de n y p dada la muestra observada $x = (x_1, \ldots, x_k)$ es

$$\begin{aligned} P(n,p|x) &\propto L(n,p;x)\pi(n,p) \\ &\propto \left[\prod_{i=1}^{k}\binom{n}{x_i}\right] p^t(1-p)^{nk-t}\pi(n,p)I_{[x_{\text{máx}},\infty]}(x). \end{aligned}$$

Las inferencias acerca de el parámetro de interés n se hacen a través de la densidad marginal posterior de n dada la muestra observada. Esta densidad se calcula integrando la densidad posterior $P(n, p|x)$ con respecto a p, el parámetro considerado de estorbo,

$$\begin{aligned} P(n|x) &\propto \int_0^1 P(n,p|x)dp \\ &\propto \int_0^1 \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^t(1-p)^{nk-t}\pi(n,p)I_{[x_{\text{máx}},\infty]}(x)dp. \end{aligned} \tag{2.6}$$

En este contexto, Draper y Guttman (1971) suponen que n y p son independientes con función de densidad conjunta $\pi(n, p) = (1/N)p^a(1-p)^b$, donde $1 \leq n \leq N$ y N es un número entero preseleccionado que representa una cota superior para n. Luego, calculan la distribución posterior $P(n|x)$ dada en (2.6),

$$P(n,p|x) \propto \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] \left[\binom{nk+a+b+1}{t+a+1}\right]^{-1} I_{[x_{\text{máx}},N]}(x). \tag{2.7}$$

Nótese que esta posterior es matemáticamente idéntica a la función de verosimilitud integrada de Carroll y Lombard (1985) dada en (2.5), excepto por el dominio de la función. Ellos utilizan la moda de la distribución posterior de n como un estimador puntual de n.

2.3.3 Ejemplo

En este ejemplo se ilustra, a través de la verosimilitud integrada dada en (2.5), el impacto que puede tener en las inferencias sobre n la incorporación de información adicional hecha a través de la densidad de probabilidad inicial o previa de p. Se considerará el escenario (a) de la Tabla 2.1; es decir, la muestra observada es $x = (17, 18, 18, 24, 26)$ y fue simulada con $n = 100$ y $p = 0.2$. Los valores de a y b, de la distribución *a priori* Beta de p son 6 y 20, respectivamente. Nótese que estos valores fueron seleccionados de forma que la esperanza de la $Beta\,(a, b)$ sea 0.2, el valor verdadero de p. La Figura 2.2 muestra la función de verosimilitud integrada

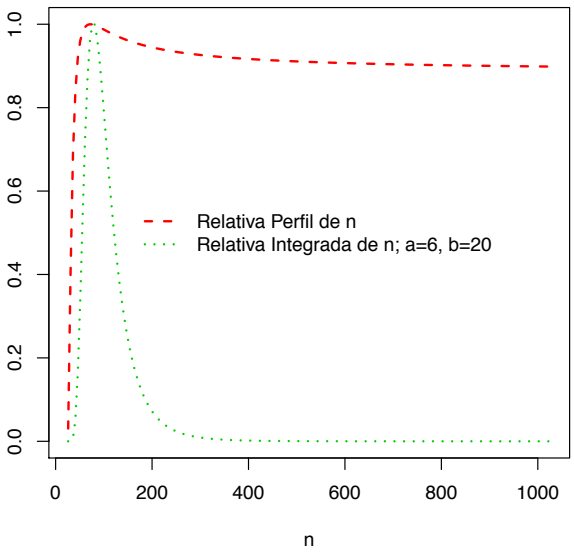

Figura 2.2: Función de verosimilitud perfil y función de verosimilitud integrada.

de n (estandarizada a uno en su máximo) y la función de verosimilitud perfil de n. Se observa que la forma de la función de verosimilitud integrada no es plana como la de la perfil de n, crece hasta alcanzar su máximo y luego decrece hasta alcanzar valores cercanos a cero, no presenta problemas para hacer inferencias razonables sobre n y el valor verdadero de $n = 100$ tiene alta credibilidad. Cabe mencionar que esta verosimilitud es matemáticamente idéntica a la posterior Bayesiana de manera que ésta tampoco tiene problemas de estimación.

La verosimilitud integrada y la densidad marginal posterior Bayesiana son dos herramientas estadísticas que tienen el potencial de eliminar los problemas de estimación del modelo Binomial; sin embargo, la elección de una distribución previa ha sido objeto de críticas ya que con frecuencia se elige por conveniencia matemática más que basándose en algún fundamento físico o ecológico relacionado con el problema real. Más aún, proponer una distribución inicial apropiada parece ser una cuestión de fe. Todo

esto es relevante pues la elección de la distribución previa puede tener gran impacto en las inferencias sobre n; véase Montoya (2004).

Por el problema de identificabilidad (cuando p es desconocida), los datos por si solos no proporcionan información suficiente para hacer inferencias razonables y prácticas acerca del parámetro n. Por lo tanto, es necesario incorporar información que ayude a discernir qué valores de los parámetros son más creíbles a la luz de una muestra observada. En el siguiente capítulo, se propone usar covariables con el propósito de incorporar información objetiva al problema de estimación de abundancia bajo el modelo Binomial. Se supondrá que dichas covariables proporcionan información acerca de la relación que existe entre la abundancia y características ambientales de la región de estudio; es decir, describen hábitats.

Capítulo 3

Modelo Binomial jerárquico con covariables

En este capítulo se presenta un modelo Binomial jerárquico para estimar abundancia de especies; se describe la jerarquía y distribución de las variables que constituyen el modelo y se calcula la distribución marginal relevante para el modelado de los datos de conteos. Además, se propone una forma de incorporar covariables en el modelado y una reparametrización en términos de parámetros considerados de interés para el problema de estimación de abundancia.

3.1 Introducción

El punto de partida de un proceso de inferencia estadística es, generalmente, el modelado estadístico paramétrico. Un modelo estadístico paramétrico $\mathcal{M}$ se define como un conjunto de funciones de densidad de probabilidad $f(x;\theta)$ indexadas por un parámetro θ que toma valores en un conjunto $\Theta \subset \mathbb{R}^d, d \in \mathbb{N}$.

Si en un modelo estadístico paramétrico $\mathcal{M}_1 = \{f(x;\theta_1)|\theta_1 \in \Theta_1, \Theta_1 \subset \mathbb{R}^{d_1}\}$ el parámetro θ_1 se considera una variable aleatoria con función de densidad de probabili-

dad que pertenece a un modelo estadístico paramétrico $\mathcal{M}_2 = \{f(\theta_1;\theta_2)|\theta_2 \in \Theta_2, \Theta_2 \subset \mathbb{R}^{d_2}\}$, entonces se dice que la densidad de probabilidad condicional $f_1(x|\,\theta_1)$ y $f_2(\theta_1;\theta_2)$ constituyen un modelo llamado jerárquico. Pueden agregarse tantos niveles como sea necesario; sin embargo, la mayoría de los problemas suelen involucrar dos o tres niveles (Lovric, 2011).

En la siguiente sección se propone un modelo Binomial jerarquico para estimar la abundancia de cierta especie dentro de un área de interés dividida en k regiones. El área de cada región debe tener las dimensiones adecuadas para que la población sea demográficamente cerrada durante el periodo de muestreo; es decir, se mantiene constante y no es afectada por los procesos naturales de la emigración, inmigración, nacimientos, muertes o reclutamiento de individuos.

3.2 Modelo Binomial jerárquico

Sean N_i y P_i variables aleatorias independientes que representan el número de individuos y la probabilidad de detección de la especie bajo estudio en la i-ésima área. Se supondrá que N_i sigue una distribución Poisson con parámetro λ_i mientras que la probabilidad de detección, P_i, tiene función de densidad de probabilidad Beta con parámetros a_i y b_i. Supóngase que el número de observaciones o conteos en la i-ésima región, dado N_i y P_i, tiene función de densidad de probabilidad Binomial con parámetros $N_i = n$ y $P_i = p$. Así, el modelo jerárquico esta constituido por:

$$\left.\begin{aligned} f_{X_i|N_i,P_i}(x|N_i = n, P_i = p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ f_{P_i}(p; a_i, b_i) &= \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} p^{a_i - 1}(1-p)^{b_i - 1} \\ f_{N_i}(n;\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^n}{n!} \end{aligned}\right\}, \tag{3.1}$$

donde $\Gamma(z)$ es la función Gamma definida por $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\ \mathrm{dt}$.

El modelo de $X_i|N_i, P_i$ suele ser llamado modelo protocolar puesto que su elección

depende del protocolo de muestreo. En este tipo de problemas, la función de probabilidad relevante para modelar los conteos en la i-esima área es la distribución marginal de X_i, la cual se presenta a continuación.

3.2.1 Distribución marginal

Bajo el modelo jerárquico presentado en (3.1) la distribución marginal de X_i es:

$$\begin{aligned}
f_{X_i}(x; a_i, b_i, \lambda_i) &= \int_0^1 \left\{ \sum_{n=x}^{\infty} f_{X_i,N_i,P_i}(x, n, p; a_i, b_i, \lambda_i) \right\} dp \\
&= \int_0^1 \left\{ \sum_{n=x}^{\infty} f_{X_i|N_i,P_i}(x|n, p; a_i, b_i, \lambda_i) f_{N_i}(n; \lambda_i) f_{P_i}(p; a_i, b_i) \right\} dp \\
&= \int_0^1 \left\{ \sum_{n=x}^{\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \frac{e^{-\lambda_i} \lambda_i^n}{n!} \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)} p^{a_i-1} (1-p)^{b_i-1} \right\} dp \\
&= \frac{1}{B(a_i, b_i) x!} \int_0^1 \left\{ \sum_{n=x}^{\infty} \frac{e^{-\lambda_i} \lambda_i^n}{(n-x)!} p^{a_i+x-1} (1-p)^{b_i+n-x-1} \right\} dp,
\end{aligned}$$

donde $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

Tomando $z = n - x$,

$$\begin{aligned}
f_{X_i}(x; a_i, b_i, \lambda_i) &= \frac{1}{B(a_i, b_i) x!} \int_0^1 \left\{ e^{-\lambda_i} \sum_{z=0}^{\infty} \frac{1}{z!} p^{a_i+x-1} (1-p)^{b_i+z-1} \lambda_i^{z+x} \right\} dp \\
&= \frac{\lambda_i^x}{B(a_i, b_i) x!} \int_0^1 \left\{ e^{-\lambda_i} p^{a_i+x-1} (1-p)^{b_i-1} \sum_{z=0}^{\infty} \frac{[\lambda(1-p)]^z}{z!} \right\} dp \\
&= \frac{\lambda_i^x}{B(a_i, b_i) x!} \int_0^1 \left\{ e^{-\lambda_i} p^{a_i+x-1} (1-p)^{b_i-1} e^{[\lambda(1-p)]} \right\} dp \\
&= \frac{\lambda^x}{B(a_i, b_i) x!} \int_0^1 \left\{ e^{-\lambda_i p} p^{a_i+x-1} (1-p)^{b_i-1} \right\} dp \\
&= \frac{B(a_i + x, b_i) \lambda_i^x}{B(a_i, b_i) x!} \int_0^1 \left\{ e^{-\lambda_i p} \frac{1}{B(a_i + x, b_i)} p^{a_i+x-1} (1-p)^{b_i-1} \right\} dp \\
&= \frac{B(a_i + x, b_i) \lambda_i^x}{B(a_i, b_i) x!} M_{P_i^*}(-\lambda_i),
\end{aligned}$$

donde

$$\begin{aligned} M_{P_i^*}(t) &= E(e^{tP_i^*}) \\ &= \int_0^1 \left\{ e^{tp} \frac{1}{B(a_i+x, b_i)} p^{a_i+x-1}(1-p)^{b_i-1} \right\} dp, t \in \mathbb{R}, \end{aligned} \tag{3.2}$$

es la función generadora de momentos de una variable aleatoria P_i^* con función de densidad de probabilidad Beta con parámetros $(a_i + x, b_i)$. Para enfatizar que $M_{P_i^*}$ depende de x, a_i y b_i, de aquí en adelante se escribirá $M_{P_i^*}(t; a_i, b_i, x)$.

Entonces, para el modelo dado en (3.1), la distribución marginal de X_i está dada por

$$f_{X_i}(x; a_i, b_i, \lambda_i) = \frac{B(a_i+x, b_i)}{B(a_i, b_i)} \frac{\lambda_i^x}{x!} M_{P_i^*}(-\lambda_i; a_i, b_i, x). \tag{3.3}$$

3.2.2 Uso de covariables

En esta tesis se supondrá que la probabilidad de detección y la abundancia media se ven afectadas por ciertas características ambientales y/o espaciales. A continuación se propone una forma muy simple de incorporar esta información en el modelo.

Supóngase que el valor esperado de P_i esta dado por

$$\mu_i = E(P_i) = Logit(v_i\beta + \phi), \tag{3.4}$$

donde β y ϕ son parámetros desconocidos que gobiernan el comportamiento de la distribución de P_i y en particular de la su media. El valor de la covariable en la *i-ésima* región se representa por v_i.

De manera similar, la variación de la abundancia media en cada región se especifica como función del valor de una covariable w_i,

$$\lambda_i = E(N_i) = e^{\alpha w_i}, \tag{3.5}$$

donde α es un parámetro desconocido.

En general, la variación en la probabilidad de detección en la i-ésima región puede modelarse como función del valor observado de un vector de covariables $\boldsymbol{v_i} = (1, v_{i1}, \ldots, v_{ir})'$ y un vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)$, de tal forma que $Logit(P_i) = \boldsymbol{v_i}'\boldsymbol{\beta} + \boldsymbol{\phi}$. Para el modelo de abundancia, puede considerarse que $\log(\lambda_i) = \boldsymbol{w_i'}\boldsymbol{\alpha}$, donde $\boldsymbol{w_i} = (1, w_{i1}, \ldots, w_{is})'$ y $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_s)$. Debe tenerse en cuenta que utilizar parámetros de manera excesiva puede llegar a causar falta de identificabilidad debido a una sobreparametrización del modelo.

3.2.3 Reparametrización $(a_i, b_i, \lambda_i) \leftrightarrow (\alpha, \beta, \phi)$

Nótese que las funciones liga (3.4) y (3.5), utilizadas para relacionar el valor esperado de N_i y P_i con las covariables v_i y w_i dependen de los parámetros metapoblacionales α, β y ϕ, que pretenden describir el cambio en la detección y la abundancia debido al hábitat de la especie bajo estudio. Obsérvese también que en la distribución marginal de X_i dada en (3.3), no están involucrados explícitamente estos parámetros y tampoco las covariables. Debido al nivel lógico e importancia de los parámetros α, β y ϕ, en esta tesis se propone efectuar la siguiente reparametrización:

$$a_i = a_i(\beta) = e^{v_i\beta}, \tag{3.6}$$

$$b_i = b_i(\phi) = e^{-\phi}, \tag{3.7}$$

$$\lambda_i = \lambda_i(\alpha) = e^{\alpha w_i}. \tag{3.8}$$

Así, la función de densidad de probabilidad marginal de X_i se expresa de la siguiente manera:

$$f_{X_i}(x; \alpha, \beta, \phi) = f_{X_i}(x; a_i = a_i(\beta), b_i = b_i(\phi), \lambda_i = \lambda_i(\alpha)). \tag{3.9}$$

3.3 Heurística del modelo

En esta sección se explora el comportamiento de la distribución marginal para datos de conteos dada en (3.3). El objetivo principal es mostrar el efecto que tienen los parámetros sobre esta densidad de probabilidad y valorar la flexibilidad de esta distribución para describir datos de conteos.

Considérese una área dividida en k regiones. Supóngase que dos covariables son registradas en cada región, w_i y v_i con $i = 1, \ldots, k$. La primera (w_i) explica el cambio en la abundancia esperada a través de la ecuación (3.5) mientras que la segunda (v_i) explica el cambio en la detectabilidad esperada a través de la ecuación (3.4). En la Tabla 3.1 se muestra un escenario simulado con $k = 3$ y 12 réplicas, $(\alpha, \beta, \phi) = (1, 2, -1)$ y valores especificados para estas covariables. Además, en esta tabla se presentan los correspondientes valores esperados de abundancia y detectabilidad (columna 4 y 5). Cabe mencionar aquí que este escenario simulado fue elegido simplemente con la intención de provocar una heterogeneidad en las regiones.

Región	w_i	v_i	λ_i	μ_i	Datos simulados
R1	3	0.0	20.09	0.27	0, 0, 1, 1, 1, 1, 2, 2, 7, 11, 16, 18
R2	4	0.5	54.60	0.50	8, 12, 17, 18, 21, 23, 26, 31, 33, 37, 41, 42
R3	5	1.0	148.41	0.73	72, 77, 92, 100, 100, 108, 123, 128, 128, 128, 135, 143

Tabla 3.1: Escenario de simulación y datos simulados.

La Figura 3.1 muestra la distribución verdadera de la abundancia en la región 1, 2 y 3, denotadas como R1, R2 y R3 respectivamente. Nótese que cuando w_i se incrementa entonces la media de la distribución se incrementa (la distribución se localiza cada vez más a la derecha) y aumenta la varianza. Por otro lado, la Figura 3.2 muestra la distribución verdadera de la probabilidad de detección en cada región. Nótese que cuando v_i aumenta entonces la media de la distribución se incrementa y la forma de la distribución cambia drásticamente; pasa de tener cola pesada a la derecha (región

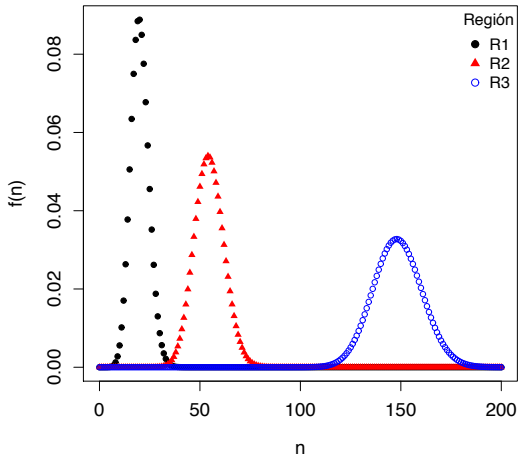

Figura 3.1: Distribución verdadera de la abundancia.

1) a tener cola pesada a la izquierda (región 3). Por último, la Figura 3.3 muestra la distribución marginal verdadera para cada región. Esta distribución incorpora la información de las covariables de forma directa y logra capturar la esencia de la variabilidad en la abundancia y detectabilidad.

En este capítulo se presentó un modelo para estimar abundancia de especies, obtenido a partir del marco de un modelo Binomial jerárquico. Además, se mostró una forma de incorporar información de covariables en este modelo y se exploró su comportamiento con base en un escenario simulado. En el siguiente capítulo, se considerará este modelo y se realizarán inferencias sobre parámetros relacionados a la abundancia. Se usarán los datos simulados de la Tabla 3.1 (columna 6); así como datos reales de conteos de salamandras *(Ambystoma ordinarium)*.

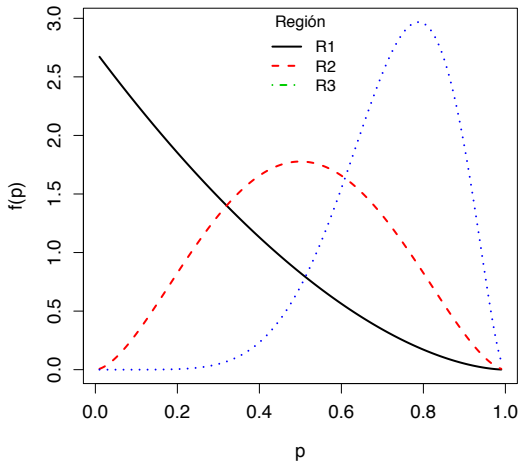

Figura 3.2: Distribución verdadera de la probabilidad de detección.

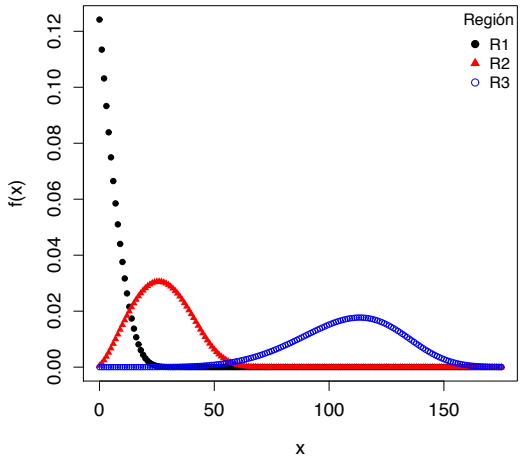

Figura 3.3: Probabilidad marginal verdadera.

Capítulo 4

Inferencia estadística

En este capítulo se presentará la metodología estadística de verosimilitud para hacer inferencia sobre los parámetros involucrados en el modelo presentado en el capítulo anterior. Además, se utilizarán las propiedades asintóticas de la verosimilitud para definir una región de incertidumbre sobre parámetros de interés asociados al modelo. Las herramientas de inferencia desarrolladas en este capítulo serán ejemplificadas con datos simulados del modelo y datos reales de conteos de una salamandra en peligro de extinción: La salamandra mexicana de arroyo, *Ambystoma ordinarium*.

4.1 Estimación puntual de parámetros

Sea $x = (x_1, \ldots, x_k)$ una muestra observada de $X = (X_1, \ldots, X_k)$; donde $X_1, \ldots, X_k$ son variables aleatorias independientes con función de densidad de probabilidad $f_{X_i}(x_i; \alpha, \beta, \phi)$ dada en (3.9). La función de verosimilitud del vector de parámetros (α, β, ϕ) es pro-

porcional a la función de densidad de probabilidad conjunta $f_X(x;\alpha,\beta,\phi)$; esto es,

$$\begin{aligned} L(\alpha,\beta,\phi;x) &\propto f_X(x;\alpha,\beta,\phi) \\ &= \prod_{i=1}^{k} f_{X_i}(x_i;\alpha,\beta,\phi) \\ &= \prod_{i=1}^{k} \frac{B(a_i+x_i,b_i)}{B(a_i,b_i)} \frac{\lambda_i^{x_i}}{x_i!} M_{P_i^*}(-\lambda_i;a_i,b_i,x_i), \end{aligned}$$

donde $a_i = e^{(v_i\beta)}$, $b_i = e^{-\phi}$, $\lambda_i = e^{\alpha w_i}$ y $M_{P_i^*}(\cdot)$ es la función generadora de momentos definida en (3.2). Así, la función de log-verosimilitud es

$$\begin{aligned} l(\alpha,\beta,\phi;x) &= \log\left(L(\alpha,\beta,\phi;x)\right) \\ &= \sum_{i=1}^{k} \{\log(B(a_i+x,b_i)) - \log(B(a_i,b_i)) + x_i\log(\lambda_i) - \log(x_i!) \\ &\quad + \log(M_{P_i^*}(-\lambda_i;a_i,b_i,x_i))\}. \end{aligned}$$

Los estimadores de máxima verosimilitud de α, β y ϕ son aquellos que maximizan la función de log-verosimilitud; es decir,

$$(\hat{\alpha},\hat{\beta},\hat{\phi}) = \text{argmáx}_{\alpha,\beta,\phi} l(\alpha,\beta,\phi;x). \tag{4.1}$$

4.2 Regiones de confianza

En un proceso de inferencia estadístico además obtener estimadores puntuales es importante cuantificar el grado de incertidumbre en las estimaciones, tanto de los parámetros como de funciones de estos; las regiones de versimilitud-confianza son una herramienta muy útil para este propósito.

Teorema 4 *Sea $X = (X_1,\ldots,X_n)$ un vector de variables aleatorias con función de densidad de probabilidad $f(x;\theta)$, $\theta \in \Theta \subset \mathbb{R}^k$. Para todo $\theta_0 \in \Theta$, bajo ciertas condiciones de regularidad, la estadística de la razón de verosimilitud $-2\log(R(\theta_0;X))$ converge a una distribución Ji-cuadrada con k grados de libertad, denotada por χ^2_k, donde*

k es la dimensión del vector θ y $R(\theta_0; X) = L(\theta_0; X)/L(\hat{\theta}; X)$ es la función de verosimilitud relativa evaluada en θ_0.

El Teorema 4 implica que si q_γ es un cuantil tal que $P[\chi^2_k \leq q_\gamma] = 1 - \gamma$, $\gamma \in (0, 1)$, entonces

$$P[-2 \log R(\theta; x) \leq q_\gamma] \approx P[\chi^2_k \leq q_\gamma] = 1 - \gamma.$$

Esto es equivalente a,

$$\begin{aligned} P[-2 \log R(\theta; x) \leq q_\gamma] &= P[-2 \log L(\theta; x) + 2 \log L(\hat{\theta}; x) \leq q_\gamma] \\ &= P\left[l(\theta; x) \geq l(\hat{\theta}; x) - \frac{q_\gamma}{2}\right] \approx 1 - \gamma. \end{aligned}$$

Así, todos los valores de θ que satisfacen la siguiente desigualdad $l(\theta; x) \geq l(\hat{\theta}; x) - q_{\gamma/2}$ pertenecen a un conjunto llamado región de verosimilitud-confianza del $(1 - \gamma)\%$ denotado por $RC_\theta(\gamma)$. Una demostración del Teorema 4 puede consultarse en Serfling (1980); ver también Cox y Hinkley (1976).

Cabe mencionar aquí que estos resultados también son válidos cuando se usa la función de verosimilitud perfil relativa. Por ejemplo, cuando el parámetro de interés es de dimensión 1 entonces q_γ es el cuantil $1 - \gamma$ de una Ji-cuadrada con un grado de libertad, $k = 1$. En caso de que el parámetro de interés sea de dimensión 2, entonces q_γ sería el cuantil $1 - \gamma$ de una Ji-cuadrada con $k = 2$ grados de libertad. Estos resultados serán usados en las Secciones 4.4 y 4.5 para calcular intervalos y regiones de verosimilitud-confianza para los parámetros de nuestro modelo.

4.3 Regiones de incertidumbre

En esta sección se presenta en concepto de regiones de incertidumbre, que será utilizado más adelante para hacer inferencia sobre el valor esperado de la abundancia. Se utilizará esta herramienta estadística puesto que, como se verá a continuación,

este parámetro depende de los otros parámetros del modelo a través de una expresión matemática poco simple.

Una distribución relevante para inferencias sobre la abundancia N dada la muestra X es la distribución condicional de N dado X, con función de densidad dada por

$$f_{N|X}(n;\alpha,\beta,\phi) = \frac{1}{(n-x)!}\lambda_i^{n-x}e^{-\lambda_i}\frac{B(a_i+x,b_i+n-x)}{B(a_i+x,b_i)}\frac{1}{M_{P_i^*}(-\lambda_i;a_i,b_i,x_i)}, \quad (4.2)$$

donde a_i, b_i y λ_i tienen la expresión dada en (3.6), (3.7) y (3.8), respectivamente. Así, una cantidad que puede ser de interés en el contexto ecológico y se relaciona con esta distribución es el valor esperado,

$$E[N|X = x_{\text{máx}}] = \sum_{n=x_{\text{máx}}}^{\infty} n f_{N|X}(n;\alpha,\beta,\phi).$$

Nótese que esta es una función de los parámeros involucrados en el modelo. En este caso, para hacer inferencia sobre este parámetro se usarán regiones de incertidumbre.

Proposición 1 *Sea $\theta \in \mathbb{R}^k$ y $G : \mathbb{R}^k \leftarrow \mathbb{R}^j$. Como $\theta \in RC_\theta(\gamma)$ implica que $G(\theta) \in RI_{G(\theta)}(\gamma)$, entonces*

$$(1-\gamma) \approx P\left[\theta \in RC_\theta(\gamma)\right] \leq P\left[G(\theta) \in RI_{G(\theta)}(\gamma)\right].$$

La demostración es inmediata del Teorema 4 y de la propiedad de la medida de probabilidad que dice que $A \subseteq B$ entonces $P(A) \leq P(B)$.

El conjunto $RI_{G(\theta)}(\gamma)$ no es una región de confianza puesto que en el únicamente se tiene control de al menos cierta cobertura. A esta región conservadora se le ha llamado Región de Incertidumbre ya que aporta información sobre la incertidumbre para $G(\theta)$, Basurto (2008).

A continuación se presentarán dos ejemplos donde se usa el modelo presentado en esta tesis y se hace inferencia a través del enfoque de verosimilitud; es decir, se muestra la función de verosimilitud perfil de los parámetros y se calcula el EMV; así

como intervalos y regiones de verosimilitud-confianza, y regiones de incertidumbre para la abundancia esperada. El primer ejemplo corresponde a datos simulados del modelo (caso sintético). En contraste, en el segundo ejemplo (caso real) se analizan datos reales de conteos de una salamandra en peligro de extinción: La salamandra mexicana de arroyo, *Ambystoma ordinarium*.

4.4 Caso sintético

Considérese el conjunto de datos simulados en la Tabla 3.1. La función de densidad de probabilidad para la abundancia, para la probabilidad de detección y la distribución marginal verdadera se muestran en las Figuras 3.1, 3.2 y 3.3, respectivamente. En este caso, el estimador de máxima verosimilitud del vector de parámetros (α, β, ϕ) es $(\hat{\alpha}, \hat{\beta}, \hat{\phi}) = (1.0055, 2.1224, -1.1487)$. La función de verosimilitud perfil relativa de cada parámetro así como sus correspondientes regiones de verosimilitud-confianza (intervalos) del 95% se muestran en las Figuras 4.1, 4.2 y 4.3. En cada caso, el estimador de máxima verosimilitud se marca con un asterisco. El resumen de las inferencias se muestra en la Tabla 4.1. En esta tabla también se presentan los valores verdaderos de los parámetros con los que fue simulada la muestra.

Parámetro	Valor verdadero	EMV	LI	LS
α	1	1.0055	0.975	1.059
β	2	2.1224	1.44	2.74
ϕ	-1	-1.1487	-1.74	0.5

Tabla 4.1: Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: Caso sintético.

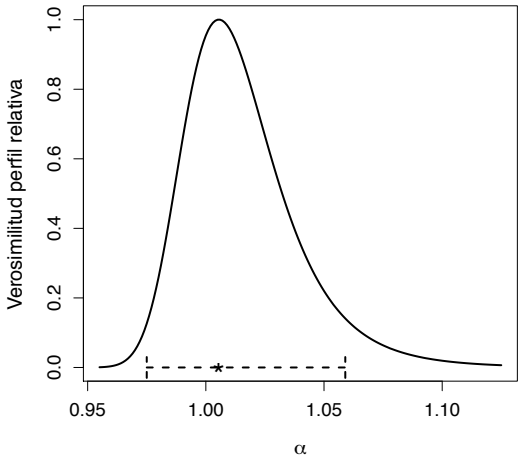

Figura 4.1: Verosimilitud perfil de α: Caso sintético.

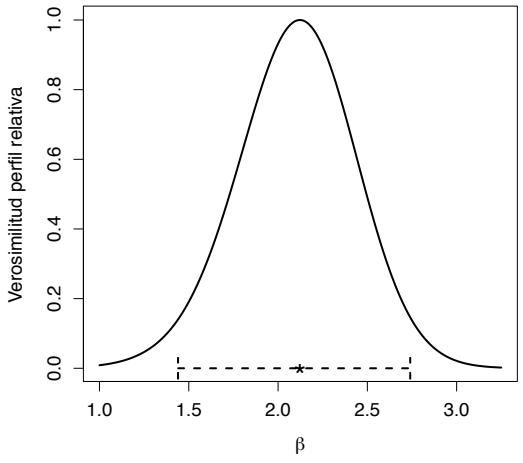

Figura 4.2: Verosimilitud perfil de β: Caso sintético.

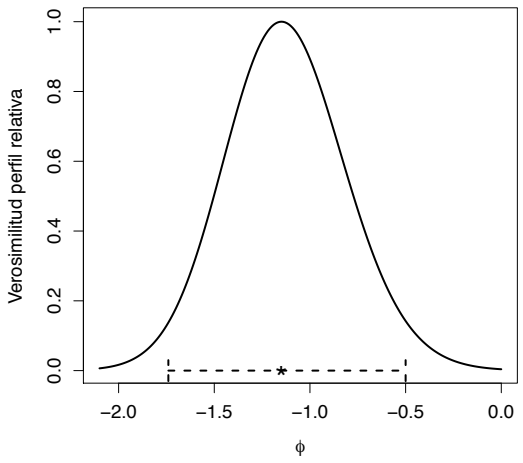

Figura 4.3: Verosimilitud perfil de ϕ: Caso sintético.

En las Figuras 4.4, 4.5 y 4.6 se comparan las funciones de probabilidad condicional de la abundancia dada en (4.2), verdadera y estimada, para cada una de las regiones de estudio. Nótese que en los tres casos, estas funciones de probabilidad son muy similares. En la Tabla 4.2 se presenta el resumen de las inferencias para el valor esperado de N dada la muestra $X = x_{\text{máx}}$.

Región	EMV	LI	LS
R1	24.23	21.62	31.51
R2	57.28	50.35	83.45
R3	159.55	148.69	243.88

Tabla 4.2: Regiones de incertidumbre para el valor espeado de N dada la muestra $X = x_{\text{máx}}$: Caso sintético.

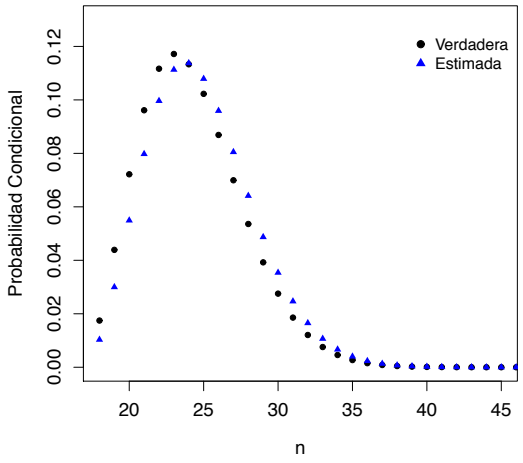

Figura 4.4: Función de probabilidad condicional Región 1: Caso sintético.

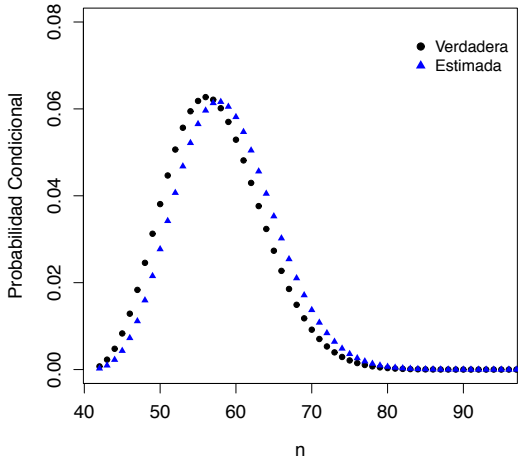

Figura 4.5: Función de probabilidad condicional Región 2: Caso sintético.

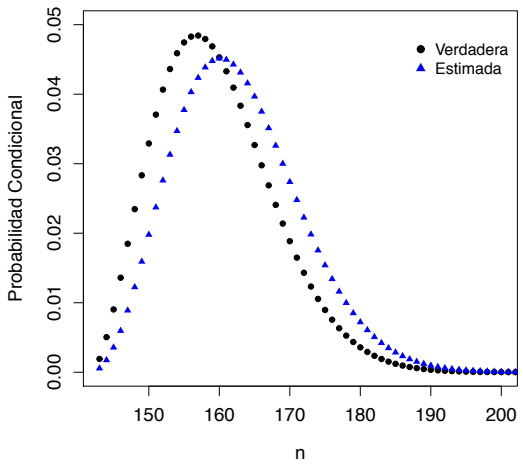

Figura 4.6: Función de probabilidad condicional Región 3: Caso sintético.

4.5 Caso real: *Ambystoma ordinarium*

En la actualidad, los anfibios representan el foco rojo de todas las especies de vertebrados a nivel mundial, ya que varias especies se han extinto y existe un declive global de muchas poblaciones según Green (2003), Stuart *et al.* (2004), Baillie *et al.* (2010) y Lista Roja UICN, accesada en junio de 2015. Las causas son diversas; pero se ha reportado que la pérdida y modificación del hábitat tienen un papel preponderante en el declive de anfibios a nivel mundial (Anciães y Marini, 2000; Baillie *et al.*, 2010). Esto se debe a que directa o indirectamente se pueden modificar las características bióticas y abióticas en los hábitats como: La disponibilidad de alimento; así como las interacciones entre depredador y presa (Davies y Nelson, 1994; Hamer y McDonnell, 2008; Kiesecker, 1996).

Las alteraciones del hábitat terrestre pueden tener consecuencias profundas sobre los sistemas acuáticos ya que existe una dependencia de las características fisi-

coquímicas y estructurales al paisaje (Ficetola *et al.*, 2011). Por ejemplo, en los arroyos la pérdida de la cubierta forestal y vegetal del área aledaña hace que la escorrentía permita un aporte de contaminantes y de material del suelo al fondo de los arroyos, disminuyendo la profundidad, aumentando la temperatura del agua y alterando los patrones del flujo del agua (Davies y Nelson, 1994). La salamandra de arroyo, *Ambystoma ordinarium*, es un anfibio metamórfico facultativo; es decir, que puede retener las características juveniles aun siendo adulto y vivir dentro del agua permanentemente o transformase y salir de los arroyos. Actualmente, la salamandra Michoacana de arroyo es considerada en riesgo de extinción y catalogada como "Amenazada" según la UICN, debido a su limitada área de ocupación y su distribución fragmentada. Más aún, fue enlistada como "Sujeto a Protección Especial" por la NOM-059-ECOL 2001, debido a su endemicidad y a la degradación de sus hábitats a nivel nacional.

En la Tabla 4.3 se muestran datos reales de conteos de salamandra Michoacana en tres sitios (arroyos) diferentes y en tres ocasiones consecutivas de muestreo durante el periodo de marzo a julio del 2014. Además, se presenta la información registrada sobre dos covariables que pueden afectar los parámetros poblacionales: El índice de Barbour (w) y la profundidad máxima en escala logarítmica (v). El índice de Barbour es un indicador del hábitat; valores grandes significan buena calidad del mismo y valores muy pequeños se interpretan como un hábitat sumamente alterado o modificado.

Cabe mencionar que la información mostrada en la Tabla 4.3, que será usada para mostrar las herramientas estadísticas desarrolladas en esta tesis, fue obtenida de un proyecto de investigación doctoral sobre el efecto de estresores ambientales en parámetros poblacionales y morfológicos sobre *Ambystoma ordinarium*, del Instituto de Investigaciones sobre los Recursos Naturales (INIRENA), dependencia de la Universidad Michoacana de San Nicolás de Hidalgo.

Sitio	w	v	Ocasión 1	Ocasión 2	Ocasión 3
Agua Zarca (AGZAR)	153	-1.35	27	31	27
Río Bello Conservado (RBC)	178	0.26	6	15	16
Lienzo (LNZ)	75	-0.16	2	2	6

Tabla 4.3: Datos de conteo de Salamandra Michoacana de arroyo.

El estimador de máxima verosimilitud del vector de parámetros (α, β, ϕ), según la expresión dada en (4.1), es $(\hat{\alpha}, \hat{\beta}, \hat{\phi}) = (0.02, -2.59, -0.28)$. La función de verosimilitud perfil relativa de cada parámetro; así como sus correspondientes regiones de verosimilitud-confianza (intervalos) del 95% se muestran en las Figuras 4.7, 4.8 y 4.9. En cada caso el estimador de máxima verosimilitud se marca con un asterisco. El resumen de las inferencias se presenta en la Tabla 4.4.

Parámetro	EMV	LI	LS
α	0.0220	0.0204	0.0250
β	-2.5890	-5.59	-0.73
ϕ	-0.2783	-1.385	0.940

Tabla 4.4: Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: *Ambystoma ordinarium*.

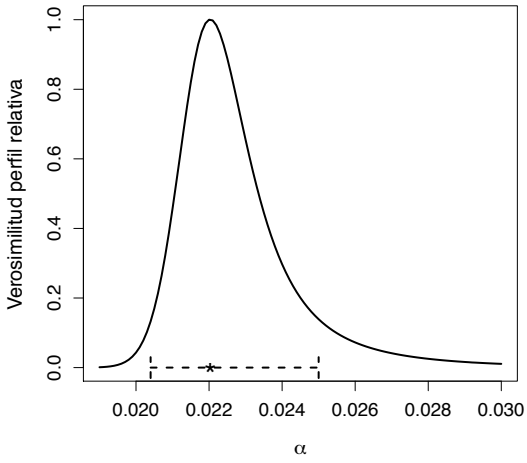

Figura 4.7: Verosimilitud perfil de α: *Ambystoma ordinarium*.

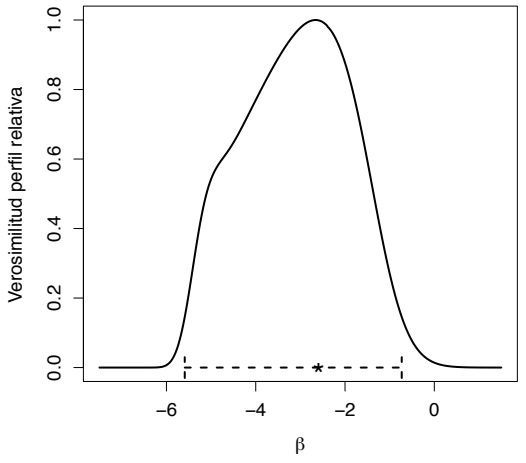

Figura 4.8: Verosimilitud perfil de β: *Ambystoma ordinarium*.

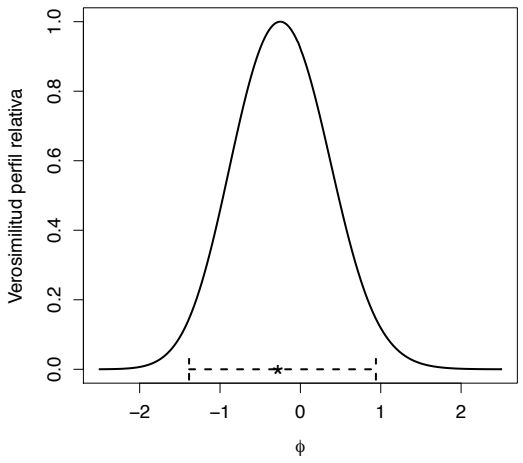

Figura 4.9: Verosimilitud perfil de ϕ: *Ambystoma ordinarium*.

Nótese que la región de verosimilitud-confianza de ϕ captura al valor 0. Es decir, una prueba de hipótesis de nivel 0.05, no rechazaría la hipótesis $\phi = 0$. En lo que sigue, se supondrá que $\phi = 0$ con el propósito de mostrar de manera sencilla las inferencias sobre los parámetros α y β a través de la verosimilitud perfil y contornos de verosimilitud para ambos parámetros (regiones de verosimilitud-confianza). Además, se obtendrá regiones de incertidumbre para la abundancia esperada bajo esta suposición.

El estimador de máxima verosimilitud del vector de parámetros (α, β), fijando $\phi = 0$, es $(\hat{\alpha}, \hat{\beta}) = (0.0219, -2.6593)$. La función de verosimilitud perfil relativa de cada parámetro; así como sus correspondientes regiones de verosimilitud-confianza (intervalos) del 95% se muestran en las Figuras 4.10 y 4.11. En cada caso el estimador de máxima verosimilitud se marca con un asterisco. El resumen de las inferencias se presenta en la Tabla 4.5.

Parámetro	EMV	LI	LS
α	0.0219	0.0204	0.0250
β	-2.6593	-5.59	-0.73

Tabla 4.5: Estimador de máxima verosimilitud (EMV) y límite inferior (LI) y superior (LS) de los intervalos de verosimilitud-confianza del 95%: *Ambystoma ordinarium*, $\phi = 0$.

En la Figura 4.12 se muestran diferentes contornos de la función de verosimilitud relativa de (α, β). En particular, el contorno de nivel 0.05 corresponde a una región de verosimilitud-confianza del 95% para ambos parámetros. Además, se marca el EMV con un asterisco.

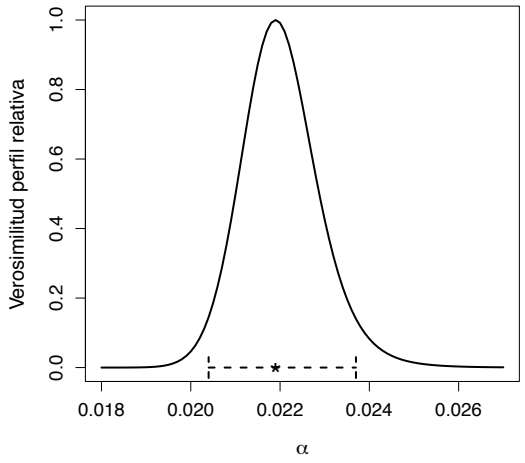

Figura 4.10: Verosimilitud perfil de α: *Ambystoma ordinarium*, $\phi = 0$.

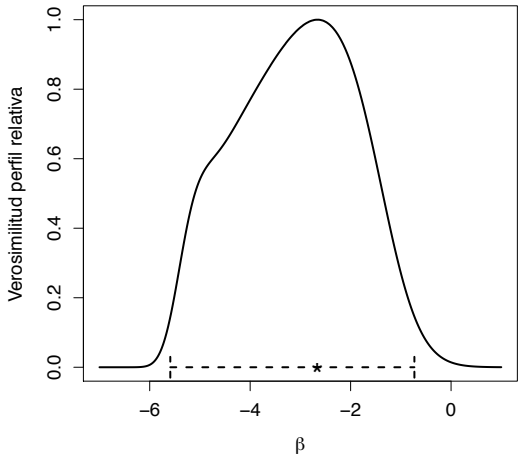

Figura 4.11: Verosimilitud perfil de β: *Ambystoma ordinarium*, $\phi = 0$.

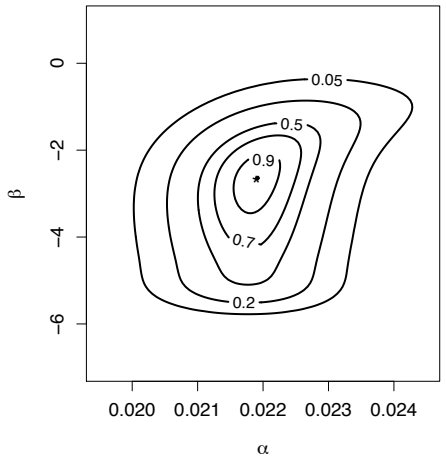

Figura 4.12: Contornos de la función de verosimilitud relativa de α y β: *Ambystoma ordinarium*, $\phi = 0$.

En la Tabla 4.6 se resumen las inferencias para N dada la muestra $X = x_{\text{máx}}$.

Región	EMV	LI	LS
AGZAR	31.70	31.02	39.17
RBC	48.81	35.04	74.16
LNZ	7.04	6.76	7.48

Tabla 4.6: Regiones de incertidumbre para el valor esperado N dada la muestra $X_{\text{máx}}$: Caso *Ambystoma ordinarium*, $\phi = 0$.

Las Figuras 4.13, 4.14 y 4.15 muestran la estimación de la función de probabilidad de N (la abundancia), la estimación de la función de p (la detección) y la estimación de la función de probabilidad marginal de X (conteos), respectivamente. En la Figura 4.13 se observa el efecto de la covariable llamada índice de Barbour sobre la media de la función de probabilidad Poisson de N; es decir, los valores esperados de las funciones de probabilidad de N, en cada sitio, están ordenados de menor a mayor según el valor del índice Barbour. Nótese que el sitio RBC tiene el mayor índice Barbour. Por otro lado, en la Figura 4.14 se observa el efecto de la covariable profundidad (escala logarítmica); es decir, los valores esperados de las funciones de probabilidad de p, en cada sitio, están ordenados de menor a mayor según el valor de la profundidad. Nótese que el arroyo RBC es el más profundo de los tres.

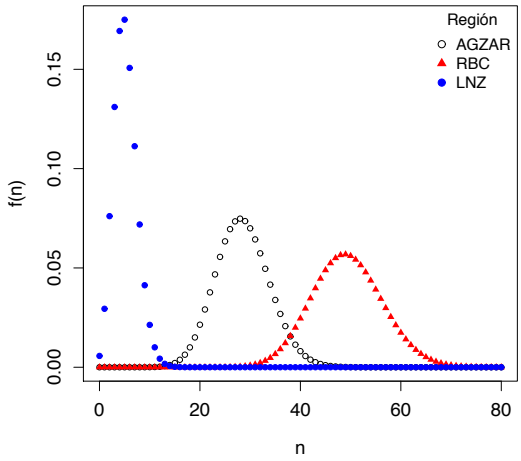

Figura 4.13: Función de densidad de probabilidad de N: *Ambystoma ordinarium*, $\phi = 0$.

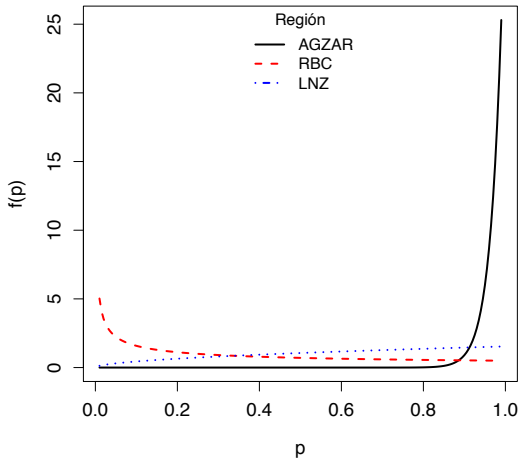

Figura 4.14: Función de densidad de probabilidad de P: *Ambystoma ordinarium*, $\phi = 0$.

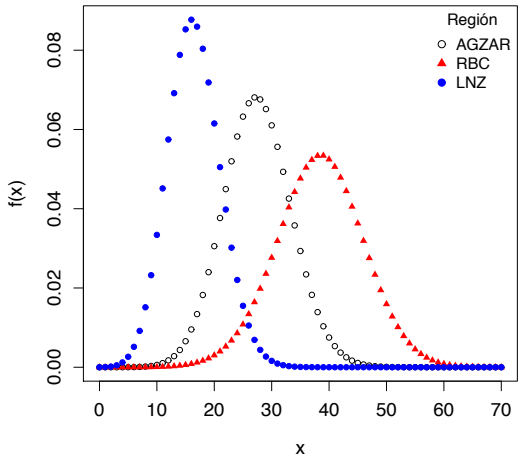

Figura 4.15: Función de probabilidad Marginal: *Ambystoma ordinarium*, $\phi = 0$.

Capítulo 5

Conclusiones

En este trabajo de tesis se abordó el problema de estimar abundancia de especies. Se señaló que la distribución Binomial (n, p) es una elección natural para modelar datos de conteos y que estimar ambos parámetros es considerado un problema difícil en la literatura estadística.

Los problemas de estimación del modelo Binomial (n, p) suelen ser causados por su falta de identificabilidad. Incorporar conocimiento *a priori* sobre los parámetros del modelo es una forma de reprimir dicha carencia de identificabilidad y solucionar problemas de estimación; sin embargo, el uso de este enfoque agrega subjetividad al análisis estadístico y las inferencias pueden ser extremadamente dependientes de esta información *a priori*. Incluso muchas veces la información previa sobre los parámetros, dada en términos de funciones de densidad de probabilidad, suele ser especificada más por conveniencia matemática que por tener algún sentido en el contexto del estudio.

En esta tesis se propone hacer inferencia sobre la abundancia a través de datos de conteos e información adicional proveniente de covariables observadas junto dichos datos. Para ejemplificar la viabilidad de esta propuesta se desarrolló un modelo Binomial jerárquico, donde las covariables están ligadas a los valores esperados de n y p, consideradas como variables aleatorias independientes con distribución Poisson y Beta,

respectivamente. Además, se presentaron dos casos de estudio, uno con datos simulados y otro con datos reales de conteos de salamandras *(Ambystoma ordinarium)*, donde se utilizó este modelo junto con un enfoque estadístico de verosimilitud para hacer inferencia sobre parámetros de interés.

El modelo Binomial jerárquico propuesto en este trabajo de tesis es básico porque contiene una sola jerarquía; además, se considera simple debido a la forma en que se ligan dos covariables a los valores esperados de n y p. Sin embargo, este modelo matemático básico y simple permitió ejemplificar de una forma clara, a través de escenarios simulados y datos reales, el efecto que pueden tener las características ambientales de los ecosistemas en la densidad de conteos; así como en la estimación de la abundancia.

Por último, es importante mencionar que el marco matemático de modelos jerárquicos permite incorporar mayor complejidad al modelado estadístico y diferentes métodos de inferencia se pueden utilizar con base en este modelo. Sin embargo, siempre es necesario vigilar si se está en una situación de identificabilidad de parámetros. Nótese que la carencia de identificabilidad, como la exhibida en esta tesis para el caso de estimar el parámetro n de una Binomial, no se soluciona con métodos estadísticos; es necesario que los datos contengan suficiente información para discernir las relaciones entre los parámetros del modelo.

Bibliografía

Anciães, M. y Marini, M. A. (2000). The effects of fragmentation on fluctuating asymmetry in passerine birds of brazilian tropical forests. *Journal of Applied Ecology*, 37(6):1013–1028.

Baillie, J., Griffiths, J., Turvey, S., Loh, J., Collen, B., Mace, G., y Stuart, S. (2010). *Evolution Lost: Status and Trends of the World's Vertebrates*. London: Zoological Society of London.

Basurto, M. (2008). *Modelación de Tamano de Nidada de Loro Corona Lila*. Centro de Investigacion en Matematicas A.C.

Bayoud, H. A. (2011). Bayes and empirical bayes estimation of the parameter n in a binomial distribution. *Communications in Statistics - Simulation and Computation*, 40(9):1422–1433.

Berger, J. O., Liseo, B., y Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14:1–28.

Bickel, P. J. y Doksum, K. A. (1977). Mathematical statistics: Basic ideas and selected topics.

Carroll, R. J. y Lombard, F. (1985). A note on n estimators for the binomial distribution. *Journal of the American Statistical Association*, 80(390):423–426.

Davies, P. y Nelson, M. (1994). Relationship between riparian buffer width and the effects of logging on stream habitat, invertebrate community composition and fish abundance. *Australian Journal of Marine and Freshwater Research*, (44):289–1305.

Dorazio, R. M. (2007). On the occurrence of statistical models for estimatin occurrence and extinction from animal surveys. *Ecology*, 88(11):2773–2782.

Dorazio, R. M., Jelks, H. L., y Jordan, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics*, 61:1093–1101.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., y Jordan, F. (2008). Modelling unobserved sources of heterogeneity in animal abundance using a dirichlet process prior. *Biometrics*, 64:635–644.

Draper, N. y Guttman, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics*, 13:667–673.

Feldman, D. y Fox, M. (1968). Estimation of the parameter n in the binomial distribution. *Journal of the American Statistical Association*, 63(321):150–158.

Ficetola, G. F., Marziali, L., Rossaro, B., De Bernardi, F., y Padoa-Schioppa, E. (2011). Landscape-stream interactions and habitat conservation for amphibians. *Ecological Applications*, 21(4):1272–1282.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced form a small sample. *Metron*, 1:3–32.

Green, D. M. (2003). The ecology of extinction: population fluctuation and decline in amphibians. *Biological conservation*, 111(3):331–343.

Hamer, A. J. y McDonnell, M. J. (2008). Amphibian ecology and conservation in the urbanising world: a review. *Biological conservation*, 141(10):2432–2449.

Kalbfleisch y Sprott (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*, 32:175–208.

Kiesecker, J. (1996). Ph-mediated predator-prey interactions between ambystoma tigrinum and pseudacris triseriata. *Ecological Applications*, 6(4):1325–1331.

Lovric, M., editor (2011). *International Encyclopedia of Statistical Science*. Springer.

Montoya, J. A. (2004). *El modelo Binomial (n, p) para estimar abundancia de animales*. Centro de Investigacion en Matematicas A.C.

Montoya, J. A. (2008). *La verosimilitud perfil en la Inferencia Estadistica*. Centro de Investigacion en Matematicas A.C.

Olkin, I., Petkau, J., y Zidek, J. (1981). A comparison of n-estimators for the binomial distribution. *Journal of the American Statistical Association*.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60:108–115.

Royle, J. A. y Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological and Enviroment Statistics*, 11(3):249–263.

Stuart, S. N., Chanson, J. S., Cox, N. A., Young, B. E., Rodrigues, A., Fischman, D. L., y Waller, R. W. (2004). Status and trends of amphibian declines and extinctions worldwide. *Science*, 306(5702):1783–1786.

Williams, B. K., Nichols, J. D., y Conroy, M. J. (2002). *Analysis and management of animal population*. London: Academic Press.