CIMAT

Centro de Investigación en Matemáticas, A.C.

Modelos estadísticos para describir la detectabilidad de especies cuando se muestrea por cuadrantes

T E S I S

Que para obtener el título de
Maestro en Ciencias con Especialidad en Probabilidad y Estadística

Presenta

Laura Jiménez Jiménez

Director de Tesis:

Dra. Eloísa Díaz-Francés Murguía

Guanajuato, Gto. Febrero de 2013

CIMAT
CENTRO DE INVESTIGACION EN MATEMATICAS A. C.

Centro de Investigación en Matemáticas, A.C.

Acta de Examen de Grado

Acta No.: 055

Libro No.: 002

Foja No.: 055

En la Ciudad de Guanajuato, Gto., siendo las 16:00 horas del día 11 de febrero del año 2013, se reunieron los miembros del jurado integrado por los señores:

DR. JOSÉ IGNACIO BARRADAS BRIBIESCA (CIMAT)
DR. VÍCTOR MANUEL RIVERO MERCADO (CIMAT)
DRA. ELOÍSA DÍAZ-FRANCÉS MURGUÍA (CIMAT)

bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

sustenta

LAURA JIMÉNEZ JIMÉNEZ

en cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"MODELOS ESTADÍSTICOS PARA DESCRIBIR LA DETECTABILIDAD DE ESPECIES CUANDO SE MUESTREA POR CUADRANTES".

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a):

APROBADA

DR. JOSÉ IGNACIO BARRADAS BRIBIESCA
Presidente

DR. VÍCTOR MANUEL RIVERO MERCADO
Secretario

DRA. ELOÍSA DÍAZ-FRANCÉS MURGUÍA
Vocal

CIMAT
DIRECCION GENERAL

DR. JOSÉ ANTONIO STEPHAN DE LA PEÑA MENA
Director General

A quienes siempre han creído en mi:

Coco, Tía Sil y Tío Pepe.

Agradecimientos

Quiero agradecer a todas las personas e instituciones que me apoyaron a lo largo de mis estudios de maestría por todo el apoyo y las enseñanzas, y por ofrecerme una formación de calidad.

Al **Centro de Investigación en Matemáticas, A.C. (CIMAT)**, particularmente a su **Departamento de Probabilidad y Estadística**, donde encontré excelentes profesores y guías.

Al **Consejo Nacional de Ciencia y Tecnología (CONACyT),** por su apoyo de beca de maestría con la que me permitió desarrollar y concluir una parte importante de mi formación profesional. Además, el desarrollo de este trabajo fue posible gracias al financiamiento de CONACyT, Proyecto Ciencia Básica 2007-83441-R: "Análisis de la Vulnerabilidad del Socio-Ecosistema de Bosque Tropical Seco al Cambio Global en la Región de Chamela, Jalisco".

A la **Dra. Eloísa Díaz-Francés Murguía** por el apoyo y consejos, y por guiarme en todo este proceso con paciencia, entrega y entusiasmo.

Al **Dr. J. Ignacio Barradas Bribiesca** y al **Dr. Víctor M. Rivero Mercado**, por sus valiosas aportaciones en este trabajo, dada su experiencia en el contexto ecológico y probabilístico, respectivamente.

Al **Dr. Andrés García Aguayo** (UNAM), por compartir la base de datos de reptiles y anfibios de la Reserva de Chamela, producto de su esfuerzo de colecta y base importante de esta tesis. A la **Dra. Ireri Suazo Ortuño** (Universidad Michoacana) y el **Dr. Enrique Martínez Meyer** (UNAM), por sus aportaciones desde el punto de vista ecológico.

A mi familia, **Norma, Iván y Marily**, y mis amigos, **Noemí y Harold**, por apoyarme incondicionalmente, escucharme y darme ánimos para continuar con nuevos proyectos y seguir creciendo.

Laura Jiménez.

Febrero, 2013.

Contenido

Prefacio **1**

1 Preliminares **5**

1.1 Contexto ecológico del problema 5

1.2 Conteos de individuos y diseño de muestro por cuadrantes 7

1.2.1 Algunas recomendaciones para obtener una muestra representativa . . 10

1.3 Conceptos estadísticos básicos 12

1.3.1 Proceso de modelar estadísticamente de un fenómeno aleatorio 13

1.3.2 Estimación de parámetros de un modelo estadístico con base en la función de verosimilitud 14

1.3.3 Intervalos de verosimilitud-confianza 18

1.3.4 Calibración de porcentajes de cobertura de intervalos de verosimilitud para muestras pequeñas 24

1.3.5 Función de verosimilitud perfil 27

1.3.6 Modelos de mezcla infinita o modelos jerárquicos 32

1.3.7 Gráficas cuantil-cuantil para datos censurados por intervalo 34

1.3.8 Razón de verosimilitudes para la comparación de modelos 37

2 Planteamiento del modelo estadístico general **39**

2.1 Datos observados: conteos de individuos y especies 40

2.2 Supuestos importantes y distribuciones de variables aleatorias de interés . . . 41
2.2.1 Distribución Poisson del número de individuos de una especie detectable . . . 42
2.2.2 Cota inferior para los parámetros de intensidad Poisson . . . 43
2.2.3 Distribución truncada de los parámetros de intensidad Poisson . . . 45
2.2.4 Estimación de intervalos para los parámetros de intensidad Poisson . 51
2.2.5 Probabilidad conjunta de $T_{1r}, ..., T_{mr}$ . . . 52
2.2.6 Dos resultados importantes . . . 54
2.2.7 Distribución del número de especies no observadas . . . 55
2.3 Modelos estadísticos para estimar el número de especies detectables . . . 57
2.4 Comparación de la definición tradicional en Ecología de *detectabilidad de una especie* con la que se da en esta tesis para *especie detectable* . . . 60

3 Simulaciones **63**
3.1 Calibración de porcentajes de cobertura de intervalos de verosimilitud para los parámetros del modelo . . . 64
3.1.1 Modelo Gama-Binomial . . . 64
3.1.2 Modelo Lognormal-Binomial . . . 70
3.2 Contraste de modelos Gama y Lognormal . . . 76
3.3 Efecto del aumento de cuadrantes en la muestra sobre el porcentaje de cobertura de intervalos de verosimilitud . . . 79

4 Aplicaciones **81**
4.1 Reptiles de Chamela . . . 82
4.1.1 Características generales de los datos y del área de estudio . . . 85
4.2 Arroyo 1995 (La Niña) . . . 91
4.2.1 Intervalos de verosimilitud-confianza para las intensidades Poisson . . 92

4.2.2 Modelo Gama-Binomial 93
4.2.3 Modelo Lognormal-Binomial 98
4.3 Arroyo 1996 (Normal) 103
4.3.1 Intervalos de verosimilitud-confianza para las intensidades Poisson . . 104
4.3.2 Modelo Gama-Binomial 104
4.3.3 Modelo Lognormal-Binomial 108
4.4 Arroyo 1997 (El Niño) 113
4.4.1 Intervalos de verosimilitud-confianza para las intensidades Poisson . . 114
4.4.2 Modelo Gama-Binomial 114
4.4.3 Modelo Lognormal-Binomial 118
4.5 Discusión 122
4.5.1 Razón de verosimilitudes para los tres años 122
4.5.2 Densidades estimadas para las intensidades Poisson 124
4.5.3 Comparación de perfiles para k 126

5 Conclusiones generales **133**

Apéndice **137**

A Intervalos de verosimilitud-confianza para parámetros de intensidad Poisson **137**

B Demostración de la Proposición 1 **140**

Bibliografía **145**

Prefacio

El número de especies detectables en una comunidad en un periodo de tiempo dado y bajo ciertas condiciones ambientales es una variable difícil de medir y estimar. Desde el punto de vista estadístico, es interesante analizar modelos matemáticos planteados con el objetivo de estimar esta variable.

En esta tesis se proponen modelos estadísticos para estimar el número de especies detectables k, en una región de interés cerrada y homogénea, cuando se muestrea por cuadrantes. Las especies pertenecen al mismo grupo taxonómico y los individuos son fácilmente identificados con su especie. Se define que una especie es detectable si la probabilidad de observar al menos un individuo de ella en la región de interés es alta.

Se considerarán dos modelos para representar la distribución de los valores esperados de los conteos de los individuos dentro de cuadrantes para las especies detectables. Estos modelos son la distribución Gama y la Lognormal. El ajuste de estos modelos a una muestra observada y la comparación de ambos para ver si alguno describe mejor a los datos, se hará a través de razones de verosimilitudes de ambos modelos, de gráficas Q-Q y de gráficas de las funciones de densidad estimadas.

Se propone un modelo general que depende de dos distribuciones. La primera asociada a la variable aleatoria que representa el número de especies no observadas en los cuadrantes. La segunda corresponde a los promedios de conteos de individuos para cada especie observada en al menos un cuadrante.

La estimación de los parámetros del modelo estadístico se llevó a cabo mediante la función de verosimilitud. Se presentan los estimadores puntuales y estimaciones por intervalo, donde

los intervalos se obtienen a partir de la función de verosimilitud perfil relativa de cada parámetro usando un nivel de verosimilitud que se calibra mediante simulaciones para cada juego de datos.

Se calibraron mediante simulaciones los niveles de verosimilitud que se requerían para obtener intervalos de confianza alta. Cuando se tienen muestras suficientemente grandes, generalmente se pueden usar directamente resultados asintóticos que permiten asociar un nivel de confianza aproximado a un intervalo de verosimilitud. Desafortunadamente en el contexto de estimación de especies detectables, suele usarse un número pequeño de cuadrantes (que toma el rol de tamaño de muestra) y además, la proporción de especies con pocos individuos suele ser alta (la cual como veremos, puede afectar la calidad de las estimaciones). Por tanto, no siempre son válidos los resultados asintóticos en estos casos.

En los ejemplos presentados en este trabajo fue necesario realizar simulaciones para elegir niveles de verosimilitud que permitieran obtener intervalos de estimación tales que su probabilidad de cobertura fuera alta.

Se usan gráficas Q-Q para evaluar el ajuste del modelo planteado para los datos observados, en particular, se considera una versión de estas gráficas que contempla datos censurados por intervalo.

Mediante la simulación de muestras similares a las observadas bajo diferentes escenarios, se evaluará el método estadístico aquí propuesto para estimar el número de especies detectables en una región.

Los ejemplos presentados en esta tesis se obtuvieron de la base de datos sobre reptiles y anfibios proporcionada amablemente por el Dr. Andrés García Aguayo investigador del Instituto de Biología de la UNAM (Unidad Occidente, Universidad de Colima, Facultad de Ciencias). Para la realización de esta tesis se tuvo el apoyo del grupo de investigadores que trabajan en el proyecto "Análisis de la Vulnerabilidad del Socio-Ecosistema de Bosque Tropical Seco al Cambio Global en la Región de Chamela, Jalisco" (CONACyT, Proyecto Ciencia Básica 2007-83441-R).

En Chamela, el efecto principal de los fenómenos Niño y Niña es la ocurrencia de can-

tidades mayores de lluvia en distintas partes del año, según el tipo de año. Si bien el total de lluvia acumulado anual es similar, hay meses donde llueve más. Por ejemplo, en los años Niña se acumula la lluvia anual en el verano y el invierno suele ser seco. Mientras que, en los años Niño, llueve menos en el verano pero se tienen inviernos lluviosos.

Se han seleccionado los datos de años en los que se presentan estos fenómenos meteorológicos y un año más donde no están presentes para estimar en cada uno de ellos el número de especies de reptiles detectables bajo condiciones climáticas y geográficas particulares. Los datos corresponden a conteos de especies de reptiles realizados durante las estaciones de lluvia (se analizaron observaciones realizadas durante los meses de Julio, Agosto, Septiembre y Octubre, de cada año). En esta tesis se muestra cómo los métodos estadísticos propuestos resultaron ser muy informativos para comparar años distintos en la época de lluvias y para valorar si existe algún efecto de estos fenómenos meteorológicos sobre el número de especies detectables de reptiles de la región cercana al arroyo en Chamela.

En el primer capítulo se presentan los conceptos y antecedentes básicos de ecología básicos para comprender el contexto en el cual se aplicaron los métodos estadísticos propuestos. En el segundo capítulo se presentan y explican los conceptos relevantes de estadística necesarios para entender el modelo estadístico que se plantea para estimar el númeo de especies detectables en una región. En el tercer capítulo se presentan simulaciones para evaluar las inferencias que se proponen, al aplicarlas a situaciones con muestras pequeñas. En el cuarto capítulo, se aplican las ideas anteriores a datos de reptiles observados en la Estación de Biología Chamela en la época de lluvias de tres años distintos con respecto a la presencia del fenómeno de la Oscilación del Sur, El Niño. Finalmente, como último capítulo se presentan las conclusiones generales.

Los programas requeridos para la tesis se elaboraron en el lenguaje de cómputo R, pueden ser solicitados vía correo electrónico a la dirección de la autora de este trabajo: jimenez@cimat.mx.

Capítulo 1

Preliminares

En este capítulo se proporciona el material de estadística y ecología básico que será útil para abordar el problema de la estimación del número de especies detectables que pertenecen a cierto grupo taxonómico en una región geográfica de interés.

Inicialmente se describirá el contexto ecológico del problema, dando algunas definiciones que serán usadas para establecer los supuestos del modelo estadístico planteado en el Capítulo 2. En seguida, se presentan los conceptos de estadística utilizados en esta tesis junto con ejemplos sencillos muy útiles para entender los conceptos básicos.

1.1 Contexto ecológico del problema

Medir la **riqueza de especies**, es decir, el número de especies que habitan en un área geográfica dada, es un objetivo esencial para muchos ecólogos que estudian comunidades y biólogos dedicados a la conservación de especies (Pielou, 1969; Magurran, 2011). Sin embargo, el número de especies que residen en una región difícilmente es una cantidad conocida, por lo que su estimación es un problema fundamental en ecología (Fisher et al., 1943).

Otro objetivo importante por el que se estudia la riqueza de especies es para medir la biodiversidad en una región. Las mediciones de diversidad frecuentemente aparecen como indicadores del buen funcionamiento de los ecosistemas ya que muchas veces el desarrollo de un ecosistema implica el incremento de la diversidad, estructura y organización.

Históricamente, la **biodiversidad** se ha medido a través de un rango desconcertante de índices que frecuentemente consisten de dos componentes: el número de especies que habitan la región de interés y la distribución[1] relativa de sus abundancias (Patil y Taillie, 1982; Fisher et al., 1943; Hubbell, 2001; Magurran, 2011). Recientemente, ambos componentes se han medido y estimado por separado (Soberón y Llorente, 1993; Brose et al., 2003; Boulinier et al., 1998; Mao y Colwell, 2005), lo cual ha llevado a que la riqueza de especies sea la medida de biodiversidad más frecuente.

En el presente contexto, se define **comunidad** como el conjunto de organismos que viven en un espacio seleccionado (área o volumen) que pertenecen al mismo taxón o grupo de especies. Así, se deben realizar dos elecciones en la identificación de una comunidad: 1) el taxón o grupo de especies; y 2) el área dentro de la cual se considerará que es razonable decir que tales especies habitan en ella.

Al definir el grupo de animales o plantas que van a constituir la comunidad no podemos sólo tomar en cuenta todos los entes vivientes en el área especificada. Sería poco práctico considerar cada tipo de organismo viviente y habría que considerar una fuerte heterogeneidad debida a la variedad de especies existentes. Por ejemplo, en una hectárea de bosque podemos encontrar: mamíferos, aves, reptiles, anfibios y microfauna, junto con árboles, arbustos, hiervas, helechos, musgos y bacterias. Así que usualmente se escoge un **grupo taxonómico** que el ecologista considera como una entidad y cuyos individuos pueden ser clasificados de acuerdo a su especie de una forma fácil.

[1] Cabe señalar que en este caso el término distribución se refiere al arreglo espacial de los especímenes bajo estudio, aunque más adelante se usará también su connotación estadística para referirse a una distribución de probabilidad.

En cuanto a la delimitación de la región geográfica de interés, se supondrá que **la población es cerrada**. Esto es, que la comunidad no cambia durante el periodo de muestreo o entre diferentes sitios dentro del área de interés. Además, se debe buscar que las características ambientales (e.g., temperatura, altitud, humedad) sean lo más **homogéneas** posibles dentro del área de estudio. Así pues, considerando una comunidad cerrada y con características ambientales homogéneas, el número total de especies se considera constante en el instante de tiempo en que se toma la muestra.

Ahora bien, para que la cantidad de especies en la comunidad esté bien definida, debemos asegurar de algún modo que cada especie está representada en la comunidad y que al realizar un muestreo dentro de la región de interés es posible observar algún individuo en la muestra con cierta probabilidad. Entenderemos por **especie detectable** a una especie que está presente en la región de estudio tal que su probabilidad de ser observada es suficientemente alta para ser considerada como una especie que habita en esta comunidad. Nótese que este supuesto hace que la cantidad k, el número de especies detectables, esté bien definida en el periodo de tiempo y el escenario ecológico al que hace referencia.

Es importante resaltar que este trabajo se centra en estimar el número de especies detectables de una comunidad ecológica en un periodo de tiempo dado y bajo ciertas condiciones ambientales sobre la región geográfica de interés, de aquí en adelante esta cantidad será denotada por k.

1.2 Conteos de individuos y diseño de muestro por cuadrantes

Como se ha discutido en la Sección 1.1, el ecólogo debe comenzar determinando la composición de la comunidad en la que se centrará el estudio para proceder a la delimitación del área que ocupa y entonces planear la colecta de datos. Esta planeación de colecta constituye el diseño de muestreo y se deben considerar varios puntos como el tiempo de duración de

la colecta, el tipo de unidades de muestreo, las variables que se medirán, el tamaño de la muestra, etc.

El diseño de muestreo que se elija para obtener datos depende de la naturaleza de la población bajo estudio, del tipo de información que provee y del proceso de colecta empleado. Un diseño de muestreo simple está basado en conteos de individuos o conteos de sus marcas (modelos de captura-recaptura) sobre unidades de muestreo escogidas al azar dentro de un área de interés.

Un diseño de muestreo muy común consiste en el conteo de individuos sobre una muestra de regiones elegidas al azar dentro del área geográfica de interés. Estas regiones son llamadas cuadrantes o sitios. El término "cuadrante" se refiere a regiones rectangulares, circulares, cuadradas, o también a las áreas de influencia en el caso de que los conteos de individuos se lleven a cabo por medio de trampas (como veremos en los ejemplos presentados en el Capítulo 3).

Un supuesto importante es que los organismos bajo estudio deben poder identificarse como individuos distintos, de otra forma no se podría registrar la abundancia (es decir, el número de individuos) de cada especie, solamente podríamos registrar si la especie está presente o ausente en un cuadrante y esto llevaría a considerar un modelo estadístico diferente.

Dado que el objetivo principal de esta tesis es estimar el número de especies detectables en un periodo de tiempo y en una región homogénea de interés, proponemos el siguiente diseño de muestreo:

1. Establecer el grupo taxonómico de interés y determinar las condiciones climáticas, el tipo de terreno y vegetación donde habita.

2. Verificar que la región de interés cuya superficie total es A, es cerrada y homogénea durante el periodo de tiempo establecido para el estudio.

3. Escoger al azar un número de cuadrantes r, de un mismo tamaño dentro de la región de interés. Se denotará con h al área de un cuadrante.

4. En cada uno de los cuadrantes, mediante la misma metodología de colecta de especies del grupo taxonómico de interés, se registra la abundancia e identidad (de qué especie es cada individuo) de los individuos detectados.

5. Con este procedimiento se obtienen conteos de individuos etiquetados con dos características: 1) el cuadrante del que proviene el conteo, y 2) la especie a la que pertenecen los individuos. La información obtenida se resume en una tabla de datos tal que los renglones representan cada especie observada y las columnas corresponden a cada cuadrante que conforma la muestra (ver Tabla 1.2.1).

Observemos que los puntos 1. y 2. corresponden a trabajo puramente ecológico, por lo que deben ser establecidos por el especialista en cuestión mediante su conocimiento sobre el grupo taxonómico de interés; la forma y tamaño de los cuadrantes dependerán de ello. Por ejemplo, el ecólogo elegirá usar trampas si es de interés colectar animales terrestres pequeños; en tal caso un cuadrante estará conformado por el área de influencia de la trampa. También podrá elegir contar individuos de las diferentes especies dentro de pequeñas parcelas que puedas ser completamente recorridas, si el grupo taxonómico de interés está conformado por arbustos.

Supóngase que se realiza un muestreo por cuadrantes en una región homogénea de superficie A donde existen $k = 12$ especies detectables con alta probabilidad. La información obtenida de este muestreo está representada en la Tabla 1.2.1. La última columna representa el número total de individuos observados en los $r = 8$ cuadrantes que conformaron la muestra, las especies han sido ordenadas de mayor a menor abundancia en los renglones de esta tabla. Se recomienda hacer este acomodo para fines de distinguir a las especies comunes de las raras.

En este ejemplo, el total de especies detectadas fue ocho. Nótese que en realidad el valor de k es desconocido y se desea estimar. Necesariamente k debe ser mayor o igual al número de especies observadas en la muestra, de tal forma que al estimar el parámetro k se debe tomar en cuenta esta restricción.

Cuadrantes:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Individuos por especie**
Especie 1	2	0	1	1	0	2	2	3	11
Especie 2	0	0	2	2	1	1	2	1	9
Especie 3	0	1	3	0	1	2	1	1	9
Especie 4	2	0	1	0	3	2	1	0	9
Especie 5	1	2	0	0	2	0	1	2	8
Especie 6	1	0	1	0	0	0	3	0	5
Especie 7	0	2	0	2	0	0	1	0	5
Especie 8	0	0	1	0	0	1	1	2	5

(1.2.1)

Tabla 1.2.1: Ejemplo de tabla de conteos obtenida de un muestreo en ocho cuadrantes..Se observaron ocho especies distintas en la muestra.

1.2.1 Algunas recomendaciones para obtener una muestra representativa

En la literatura de ecología relacionada a la estimación de modelos estadísticos, donde es de interés estimar parámetros poblacionales (como el tamaño poblacional, las tasas de crecimiento y sobrevivencia, entre otros) los trabajos de George A. F. Seber han sido de gran relevancia (1982, 1986, 1992). Particularmente, Seber menciona en sus artículos que uno de los métodos más utilizados para estimar la abundancia de una especie en cierta región, es el muestreo por cuadrantes, sin embargo, no menciona aplicaciones de muestreos por cuadrantes donde el objetivo sea estimar el número de especies de un grupo taxonómico como es el caso de esta tesis. Sin embargo, en sus trabajos podemos destacar dos recomendaciones que son válidas en general para este esquema de muestreo:

1. La eficiencia de cualquier esquema de muestreo puede incrementarse haciendo uso de cualquier variable auxiliar altamente correlacionada con la probabilidad de observar

a las especies. Ejemplos de estas variables auxiliares son: a) la época del año en la que se lleva a cabo la colecta de individuos, ya que la presencia o ausencia de algunas especies puede verse afectada por las condiciones climáticas; y b) el aprendizaje que pueden adquirir las personas que llevan a cabo la colecta de individuos para mejorar sus técnicas de búsqueda.

2. Si la población no está esparcida uniformemente sobre el área de interés (lo cual indica que no es homogénea, en términos de su distribución espacial) y esta área es muy grande, entonces es difícil asignar un esquema de muestreo adecuado debido al alto coeficiente de variación. En este caso, escoger los cuadrantes al azar no es lo más conveniente pues se pueden tener unidades de muestreo vacías. Un diseño de muestreo más apropiado podría ser un esquema adaptativo en el cual la elección de unidades de muestreo depende de la densidad de especies encontradas en áreas previamente estudiadas. De esta forma, el muestreo estará concentrado en áreas de alta densidad de especies.

La determinación de la diversidad y abundancia de especies en una comunidad ecológica está fuertemente influenciada por la estacionalidad, la época del año en que fue tomada la muestra, y posiblemente por la edad reproductiva (si se trata de una comunidad de animales) o la temporada de crecimiento (para el caso de plantas). El proceso de muestreo debe ejecutarse en una temporada tal que la mayoría de las especies de interés sean suceptibles de ser detectadas.

Los tamaños de muestra adecuados (en este caso, el número de cuadrantes) para estimar los parámetros de un modelo con cierto grado de confianza, se obtienen de diferentes formas dependiendo de los objetivos del estudio y tomando en cuenta lo siguiente:

a) La heterogeneidad entre las especies de interés sobre el área de estudio A, se captura de manera más adecuada cuando las unidades de muestreo (cuadrantes) se encuentran esparcidas uniformemente dentro de A. Por tanto, es más adecuado considerar una muestra donde se considere más de una unidad de muestreo.

b) Entre más muestras sean tomadas, mejor será la posibilidad de detectar especies raras, y por lo tanto, de tener más información sobre el número total de especies detectables. Así que colectar varias muestras pequeñas usualmente es mejor que colectar sólo una muestra de gran tamaño.

c) El tamaño de muestra adecuado puede calcularse con base en el grado de confianza o en la variabilidad de los estimadores obtenidos en estudios previos o pilotos. Además, estará fuertemente relacionado con la cobertura de muestreo, es decir, la propoción del área total A que representa la muestra de r cuadrantes. Supóngase que el área de un cuadrante es h y que la superficie A de la región de interés es de dimensión Wh, con W un número entero:

$$A = W \cdot h. \tag{1.2.2}$$

Si la muestra está conformada por r cuadrantes elegidos al azar dentro de A, entonces, la superficie total muestreada es

$$s = h \cdot r = \frac{A}{W} \cdot r. \tag{1.2.3}$$

En general, a mayor superficie muestreada, mayor será la precisión de los estimadores y menor variabilidad.

1.3 Conceptos estadísticos básicos

En las siguientes secciones se presentan algunos conceptos básicos de estadística que serán utilizados en los capítulos posteriores. El planteamiento del modelo estadístico y las inferencias sobre los parámetros del mismo se basarán en los resultados aquí expuestos.

1.3.1 Proceso de modelar estadísticamente de un fenómeno aleatorio

Uno de los problemas fundamentales en la inferencia estadística consiste en encontrar un modelo adecuado que explique el proceso de generación de un conjunto de datos. Poder delimitar un conjunto de modelos estadísticos para un problema particular requiere entrenamiento y conocimiento de la disciplina pues deben argumentarse las razones por las cuales se incluye cierto modelo, entre los posibles candidatos a elegir, así como una justificación sobre el porqué descartar algún otro. La especificación del modelo seleccionado no es arbitraria, está basada en las características del fenómeno aleatorio bajo estudio y la manera en que los datos fueron obtenidos. Por ello, la selección de un modelo es, en ocasiones, más difícil que la estimación de los parámetros del modelo mismo y para elegir el más apropiado se debe contar con experiencia.

Con base en los trabajos de Sprott (2000) y Figueroa (2012), los siguientes pasos describen el proceso de modelar estadísticamente un fenómeno aleatorio:

1. Obtener observaciones del fenómeno aleatorio de interés.

2. Plantear un modelo estadístico $f(x;\theta)$ que sea razonable, dadas las características del fenómeno.

3. Verificar si es posible combinar experimentos realizados para obtener información sobre el comportamiento del fenómeno. Estos experimentos deberán ser homogéneos, en el sentido de que proporcionen información sobre el mismo parámetro de interés.

4. Estimar el vector de parámetros θ, por medio de intervalos de estimación, así como de manera puntual.

5. Validar el modelo estadístico estimado. En caso de identificar que el modelo no es el más adecado, modificarlo o incluso cambiarlo, regresando nuevamente al punto 2.

6. Comparar las inferencias sobre los parámetros de interés y las conclusiones obtenidas, con las que resultan de otros modelos estadísticos razonables que también se considere describen bien al fenómeno de interés.

En general, se preferirán modelos simples para describir al fenómeno aleatorio bajo estudio. Sin embargo, el modelo que describa bien a ciertos datos no tiene porqué ser único, ya que todo modelo es tan solo una aproximación a la realidad compleja y siempre habrá modelos más adecuados que otros para explicar al fenómeno aleatorio de interés.

Es importante considerar el principio de parsimonia[2], esto es, el número de entidades requeridas para explicar algo no debe incrementarse más alla de lo necesario. Entre más simple sea el modelo estadístico, siempre que sea razonable y explique bien a los datos, será más fácil de entender, interpretar y usar para hacer predicciones sobre el fenómeno de interés.

1.3.2 Estimación de parámetros de un modelo estadístico con base en la función de verosimilitud

Supóngase que se ha planteado un modelo estadístico para describir un fenómeno aleatorio de interés, el cual toma en cuenta el diseño del experimento con el cual se obtienen datos. Este modelo depende de un vector de parámetros desconocido $\boldsymbol{\theta}$. Se desea usar los datos obtenidos del experimento para estimar el valor de $\boldsymbol{\theta}$, para así determinar cuáles de los valores posibles de $\boldsymbol{\theta}$ hacen o más probables a los datos observados.

Ronald A. Fisher (1921) definió por primera vez a la verosimilitud como una función de los parámetros $\boldsymbol{\theta}$ que es proporcional a la probabilidad de los datos observados, en el caso de variables aleatorias discretas. A continuación se dan las definiciones que serán útiles para fines de este trabajo.

[2]El principio de parsimonia, también conocido como Navaja de Occam, es atribuido al fraile franciscano inglés del siglo XIV Guillermo Occam. Es un principio metodológico y filosófico, según el cual "en igualdad de condiciones, la explicación más sencilla es preferible".

Definición 1 (Función de verosimilitud para variables aleatorias discretas) *Sea* $\mathbf{X} = (X_1, ..., X_n)$ *una muestra de n variables aleatorias discretas independientes e idénticamente distribuidas (vaiid) con función de distribución* $F(x; \boldsymbol{\theta})$ *donde* $\boldsymbol{\theta}$ *es un vector de parámetros desconocidos de dimensión d,* $\boldsymbol{\theta} = (\theta_1, ..., \theta_d) \in \Theta \subseteq \mathbb{R}^d$*. Dada una realización de la muestra* $\mathbf{x} = (x_1, ..., x_n)$*, la función de verosimilitud de* $\boldsymbol{\theta}$*,* $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$*, es una función real valuada con dominio en* Θ *definida como*

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = C(\mathbf{x}) \cdot \prod_{i=1}^{n} \mathbf{P}[X_i = x_i],$$

donde $C(\mathbf{x})$ *es una función positiva que depende sólo de los datos y no de* θ*.*

Así, $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ es proporcional a la probabilidad conjunta de la muestra observada.

La función de verosimilitud se definió en términos de v.a. discretas. Sin embargo, esto no involucra una pérdida de generalidad ya que en realidad los datos observados $\mathbf{x}$ siempre son discretos puesto que todo instrumento de medición tiene precisión finita y sólo pueden registrarse mediciones de la variable aleatoria con un número finito de decimales. Cuando X es una variable aleatoria continua, la observación $X = x$ debe interpretarse como $X \in \left[x - \frac{1}{2}\delta, x + \frac{1}{2}\delta\right]$, donde δ es un número positivo fijo que representa la precisión del instrumento de medición (ver Kalbfleisch, 1985, p.12).

Definición 2 (Función de verosimilitud exacta para variables aleatorias continuas) *Sea* $\mathbf{X} = (X_1, ..., X_n)$ *una muestra de n variables aleatorias continuas independientes e idénticamente distribuidas (vaiid) con función de distribución* $F(x; \boldsymbol{\theta})$ *donde* $\boldsymbol{\theta}$ *es un vector de parámetros desconocidos de dimensión d,* $\boldsymbol{\theta} = (\theta_1, ..., \theta_d) \in \Theta \subseteq \mathbb{R}^d$*. Dada una realización de la muestra* $\mathbf{x} = (x_1, ..., x_n)$*, la función de verosimilitud de* $\boldsymbol{\theta}$ *es proporcional a la probabilidad conjunta de la muestra observada,*

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) &= \prod_{i=1}^{n} \mathbf{P}\left[x_i - \frac{1}{2}\delta \leq X_i \leq x_i + \frac{1}{2}\delta\right] \\ &= \prod_{i=1}^{n} \int_{x_i - \frac{1}{2}\delta}^{x_i + \frac{1}{2}\delta} f(x; \boldsymbol{\theta})\, dx. \end{aligned}$$

Si la precisión del intrumento de medición o el diseño de muestreo empleado para obtener los datos pueden provocar que en una muestra haya observaciones repetidas siendo que estas provienen de una distribución continua. En este caso se sabe que la variable aleatoria continua tomó valores dentro de un intervalo alrededor del valor puntual observado. Los extremos de estos intervalos dependen de la precisión del instrumento (2δ) y de los valores registrados por el mismo que usualmente son los puntos medios. La muestra consistirá entonces de observaciones censuradas por intervalo.

La función de verosimilitud para datos censurados por intervalo se puede establecer de forma similar a la función de verosimilitud exacta pues cada observación contribuye a la verosimilitud con la probabilidad asociada al intervalo en el que se cree que está contenida.

Definición 3 (Función de verosimilitud para datos censurados por intervalo) *Sea* $\mathbf{X} = (X_1, ..., X_n)$ *una muestra de vaiid continuas con función de densidad común* $f(x; \boldsymbol{\theta})$ *que depende de un vector de parámetros desconocidos* $\boldsymbol{\theta}$ *de dimensión* d*. Supóngase que cuando se registra la observación* x_i*, se sabe que la variable aleatoria correspondiente* X_i *tomó algún valor del intervalo* $[A(x_i), B(x_i)]$*. Entonces, se tienen datos censurados por intervalo y la función de verosimilitud asociada a estos datos es proporcional al producto de las probabilidades de que la variable aleatoria caiga en el intervalo asociado a cada observación*

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \propto \prod_{i=1}^{n} \mathbf{P}\left[A(x_i) \leq X_i \leq B(x_i)\right].$$

Nótese que la verosimilitud exacta es un caso particular de la verosimilitud para datos censurados por intervalo.

De la definición anterior vemos que usualemente $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ será un producto de varios términos, así que será más conveniente trabajar con el logaritmo de esta función. Para ello, se define la función log-verosimilitud de θ, $\ell(\boldsymbol{\theta}; \mathbf{x})$, como el logaritmo natural de $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$,

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}). \tag{1.3.1}$$

Definición 4 (Función Score) *La función Score* $Sc(\boldsymbol{\theta}; \mathbf{x})$ *de un modelo estadístico es el vector de primeras derivadas parciales de su función log-verosimilitud con respecto a* $\boldsymbol{\theta}$*,*

$$Sc(\boldsymbol{\theta}; \mathbf{x}) = \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}}.$$

El estimador de máxima verosimlitud (EMV) de $\boldsymbol{\theta}$ es el valor de $\boldsymbol{\theta}$ que que maximiza la probabilidad de observar $\mathbf{x}$ bajo el modelo dado.

Definición 5 (EMV) *El EMV usualmente denotado por $\hat{\boldsymbol{\theta}}$, es el valor de $\boldsymbol{\theta} \in \Theta$ que maximiza $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$, o equivalentemente, $\ell(\boldsymbol{\theta}; \mathbf{x})$*

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \left(\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \right) \\ &= \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \left(\ell(\boldsymbol{\theta}; \mathbf{x}) \right). \end{aligned} \tag{1.3.2}$$

Definición 6 (Información observada de Fisher) *La información observada de Fisher $I_{\hat{\theta}}$ es el negativo de la matriz de segundas derivadas parciales de $\ell(\boldsymbol{\theta}; \mathbf{x})$ con respecto a $\boldsymbol{\theta}$, evaluada en el EMV $\hat{\boldsymbol{\theta}}$,*

$$I_{\hat{\theta}} = - \left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}^2} \bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right\}.$$

Nótese que el EMV se obtiene resolviendo el sistema de ecuaciones

$$Sc(\boldsymbol{\theta}; \mathbf{x}) = \overrightarrow{\mathbf{0}}$$

con respecto a $\boldsymbol{\theta}$, y para asegurar que el vector $\hat{\boldsymbol{\theta}}$ obtenido es un máximo basta verificar que $I_{\hat{\theta}} \geq 0$.

Para facilitar la interpretación de la función de verosimilitud y para poder comparar las funciones asociadas a distintos experimentos. conviene estandarizar la función de verosimilitud con respecto a su máximo, (ver Sprott, 2000, p.9)

Definición 7 (Función de verosimilitud relativa) *La función de verosimilitud relativa de $\boldsymbol{\theta}$ se define como*

$$R(\boldsymbol{\theta}; \mathbf{x}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})}{\mathcal{L}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right)}, \tag{1.3.3}$$

donde $\hat{\boldsymbol{\theta}}$ es el EMV de $\boldsymbol{\theta}$.

Nótese que $0 \leq R(\boldsymbol{\theta}; \mathbf{x}) \leq 1$ para cualquier valor de $\boldsymbol{\theta}$.

La función de verosimilitud relativa proporciona la plausibilidad de cualquier valor especificado de $\boldsymbol{\theta}$ relativo al máximo verosimil $\hat{\boldsymbol{\theta}}$, basada en la muestra observada $\mathbf{x}$. De tal

forma que $R\left(\hat{\boldsymbol{\theta}};\mathbf{x}\right)=1$ pues $\hat{\boldsymbol{\theta}}$ es el valor más plausible para $\boldsymbol{\theta}$. Valores de $\boldsymbol{\theta}$ con $R\left(\boldsymbol{\theta};\mathbf{x}\right)$ cercanos a cero son poco creíbles a la luz de los datos, mientras que valores cerca de uno son muy plausibles.

Las funciones de verosimilitud permiten combinar de manera simple experimentos diferentes que den información sobre el parámetro $\boldsymbol{\theta}$ (Kalbfleisch, 1985). Dado que la densidad conjunta de variables aleatorias independientes es el producto de sus densidades marginales, entonces la función de verosimilitud de $\boldsymbol{\theta}$, basada en diferentes conjuntos de datos que provienen de experimentos independientes es el producto de las verosimilitudes individuales basadas en cada uno de los conjuntos de datos. Esto significa que la forma apropiada de combinar información de experimentos diferentes que involucran un parámetro común $\boldsymbol{\theta}$, es a través de la suma de los respectivos logaritmos de las verosimilitudes individuales, la función resultante se maximiza para encontrar $\hat{\boldsymbol{\theta}}$ el EMV.

1.3.3 Intervalos de verosimilitud-confianza

Una forma usual de hacer inferencia sobre un parámetro es a través de intervalos o regiones de estimación. Los intervalos de verosimilitud, o en general, las regiones de verosimilitud (en caso de representar valores de un parámetro con dimensión mayor o igual a 2), indican los valores más plausibles del parámetro dados los datos.

Definición 8 (Intervalo de verosimilitud) *Un intervalo o región de verosimilitud de nivel c para θ, $IV\left(c\right)$, se define como*

$$IV\left(c\right)=\left\{\theta|R\left(\theta;\mathbf{x}\right)\geq c\right\},c\in\left[0,1\right]. \tag{1.3.4}$$

Todos los valores del conjunto $IV\left(c\right)\subset\Theta$ tienen una verosimilitud relativa mayor o igual a c, y los valores del espacio parametral Θ, que no están en este conjunto tienen verosimilitud relativa menor que c. Por tanto, el $IV\left(c\right)$ separa los valores más plausibles de θ de los no plausibles a un nivel c de verosimilitud (ver Sprott, 2000, p.14). El EMV de θ está contenido en todos los intervalos de verosimilitud puesto que $R\left(\hat{\theta};\mathbf{x}\right)=1\geq c,\forall c\in\left[0,1\right]$.

Un intervalo de verosimilitud por sí solo no es muy informativo. Al menos debe estar acompañado del valor del EMV $\hat{\boldsymbol{\theta}}$ para dar una idea de la simetría de la función de verosimilitud con respecto a $\hat{\boldsymbol{\theta}}$ y de cómo cambia la plausibilidad en el interior del intervalo. En lo posible también se debe graficar y analizar la función de verosimilitud relativa completa.

Dada la muestra observada $\mathbf{x} = (x_1, ..., x_n)$, se puede calcular un intervalo de verosimilitud de nivel $c \in (0, 1)$. Los extremos del intervalo $\theta_1(\mathbf{x})$ y $\theta_2(\mathbf{x})$, son aleatorios y cambian para muestras distintas. Nótese que este intervalo puede o no incluir al verdadero valor θ_0, así que resulta importante saber con qué probabilidad un intervalo para θ contiene al verdadero valor θ_0.

Definición 9 (Probabilidad de cobertura) *La probabilidad de cobertura de un intervalo aleatorio* $[\theta_1(\mathbf{x}), \theta_2(\mathbf{x})]$ *para* θ*, es la probabilidad de que el intervalo incluya, o cubra, al verdadero valor del parámetro* θ_0 :

$$PC(\theta_0) = \mathbf{P}[\theta_1(\mathbf{x}) \leq \theta_0 \leq \theta_2(\mathbf{x}); \theta = \theta_0]. \tag{1.3.5}$$

La probabilidad de cobertura $PC(\theta_0)$ es la proporción de veces que el intervalo $[\theta_1(\mathbf{x}), \theta_2(\mathbf{x})]$ incluye a θ_0 cuando el número de repeticiones del experimento con θ fijo en θ_0, tiende a infinito.

Definición 10 (Intervalo de confianza) *Un intervalo aleatorio* $[\theta_1(\mathbf{x}), \theta_2(\mathbf{x})]$ *se llama intervalo de confianza para* θ *si su probabilidad de cobertura no depende de* θ_0*. Es decir, cuando el valor de* $PC(\theta_0)$ *es el mismo para todo valor del parámetro* θ_0*.*

Para un tamaño de muestra fijo, la distribución de probabilidad de $\theta_1(\mathbf{x})$ y $\theta_2(\mathbf{x})$ se puede calcular a partir de la distribución de $\mathbf{X}$. Sin embargo, existen otras formas de asociar un nivel de confianza aproximado a un intervalo de estimación para θ. Particularmente, la probabilidad de cobertura de un $IV(c)$ se puede calcular a través de la distribución asintótica de la estadística de razón de verosimilitudes para θ fijo en θ_0,

$$\Lambda_n \equiv -2 \log R(\theta_0; \mathbf{x}). \tag{1.3.6}$$

Se sabe que Λ_n converge en distribución a una Ji-cuadrada con d grados de libertad, donde $d = \dim(\theta)$. Por lo tanto, para n suficientemente grande,

$$\begin{aligned} PC(\theta_0) &= \mathbf{P}[IV(c) \supset \theta_0 | \theta = \theta_0] \\ &= \mathbf{P}[R(\theta_0; \mathbf{x}) \geq c | \theta = \theta_0] \\ &= \mathbf{P}[\Lambda_n \leq -2\log(c) | \theta = \theta_0] \\ &\approx \mathbf{P}[\Lambda_n \leq Q_\pi | \theta = \theta_0] = \pi, \end{aligned} \tag{1.3.7}$$

donde Q_π es el cuantil de probabilidad π de una distribución Ji-cuadrada con un grado de libertad.

La distribución asintótica de Λ_n y la relación $Q_\pi = -2\log(c)$, o equivalentemente, $c = \exp(-Q_\pi/2)$ nos permiten dos cosas: 1) si se desea asignar una probabilidad a un intervalo de verosimilitud de nivel c, se calcula el cuantil Q_π (a partir de la primera expresión) y luego se averigua a cuál probabilidad corresponde; 2) si se desea encontrar el nivel de verosimilitud c para un intervalo que tenga probabilidad π, entonces se calcula c usando la segunda expresión. En la Tabla 1.3.8, que aparece a continuación se muestran los valores más comunes de c y π asociados:

Nivel de verosimilitud c	Nivel de confianza π	Cuantil Q_π
0.7965	0.50	0.45
0.2585	0.90	2.71
0.1465	0.95	3.86
0.0362	0.99	6.63

(1.3.8)

Tabla 1.3.8: Niveles de confianza aproximados de los intervalos de verosimilitud cuando θ tiene dimensión 1.

Ejemplo 1: Inferencia para el parámetro de intensidad Poisson

Consideremos una muestra de vaiid $\mathbf{X} = (X_1, ..., X_n)$ con distribución común Poisson de parámetro λ desconocido. La función de probabilidad de cada v.a. es

$$\mathbf{P}[X = x; \lambda] = f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\mathbf{1}_{\{0,1,2,...\}}(x).$$

La muestra observada es $\mathbf{x} = (x_1, ..., x_n)$. Así, la función de verosimilitud correspondiente a la muestra $\mathbf{x}$ es

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}; \mathbf{x}) &= C(\mathbf{x}) \cdot \prod_{i=1}^{n} \mathbf{P}[X_i = x_i] && (1.3.9) \\ &= C(\mathbf{x}) \left[e^{-n\lambda}\lambda^t \prod_{i=1}^{n} \mathbf{1}_{(0,\infty)}(x_i) \right] && (1.3.10) \\ &= e^{-n\lambda}\lambda^t, && (1.3.11) \end{aligned}$$

donde $t = \sum_{i=1}^{n} x_i$ y $C(\mathbf{x}) = \left[\prod_{i=1}^{n} \mathbf{1}_{(0,\infty)}(x_i) \right]^{-1}$. Luego, la expresión de la log-verosimilitud es

$$\ell(\boldsymbol{\lambda}; \mathbf{x}) = -n\lambda + \ln(\lambda) \sum_{i=1}^{n} x_i.$$

De esta expresión, el EMV de λ se obtiene derivando con respecto a λ e igualando a cero:

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}. \tag{1.3.12}$$

Dado que el EMV de λ es igual a la media muestral $\overline{x}$, es fácil probar que $\widehat{\lambda}$ es un estimador consistente fuerte (por la Ley Fuerte de los Grandes Números).

De la Definición 4 (1.3.3), la verosimilitud relativa de λ es

$$R(\lambda; \mathbf{x}) = \frac{\mathcal{L}(\boldsymbol{\lambda}; \mathbf{x})}{\mathcal{L}\left(\widehat{\lambda}; \mathbf{x}\right)} = \exp\left[-n\left(\lambda - \widehat{\lambda}\right)\right] \left(\frac{\lambda}{\widehat{\lambda}}\right)^t. \tag{1.3.13}$$

La log-verosimilitud relativa está dada por

$$r(\lambda; \mathbf{x}) = \sum_{i=1}^{n} x_i \left[\log(\lambda) - \log\left(\widehat{\lambda}\right)\right] - n\left(\lambda - \widehat{\lambda}\right). \tag{1.3.14}$$

El intervalo de verosimilitud de nivel $c \in (0,1)$ está dado por el conjunto de valores

$$\begin{aligned} IV(c) &= \{\lambda : R(\lambda;\mathbf{x}) \geq c\} = \{\lambda : r(\lambda;\mathbf{x}) \geq \log c\} \\ &= \left\{\lambda : \sum_{i=1}^{n} x_i \left[\log(\lambda) - \log\left(\widehat{\lambda}\right)\right] - n\left(\lambda - \widehat{\lambda}\right) \geq \log c\right\} \\ &= (\lambda_1(\mathbf{x}), \lambda_2(\mathbf{x})). \end{aligned} \tag{1.3.15}$$

En este caso, dada la forma de la log-verosimilitud, no se puede obtener una expresión algebraica cerrada para λ_1 y λ_2. Los extremos del $IV(c)$ deben obtenerse numéricamente y equivalen a obtener las raíces de la de ecuación

$$\sum_{i=1}^{n} x_i \left[\log(\lambda) - \log\left(\widehat{\lambda}\right)\right] - n\left(\lambda - \widehat{\lambda}\right) = \log c, \tag{1.3.16}$$

con respecto a λ. Así, el extremo izquierdo λ_1 corresponde a la raíz de menor magnitud, y el extremo derecho λ_2 corresponde a la raíz de mayor magnitud.

En seguida se da un ejemplo. Supóngase que una muestra de tamaño $n = 5, \mathbf{x} = (x_1, ..., x_n)$, es tal que $x_i = 1$, para algún $i \in \{1, ..., n\}$ y $x_j = 0$,para todo $j \neq i$. Esto quiere decir que si la v.a. X_i cuenta el número de individuos observados de cierta especie de animales o plantas, en n sitios, entonces sólo en uno de los sitios se observó la presencia de esta especie y el resto de los sitios la especie no fue detectada en la zona de muestreo.

En este caso, el EMV de λ es

$$\widehat{\lambda} = \overline{x} = \frac{1}{5} = 0.2.$$

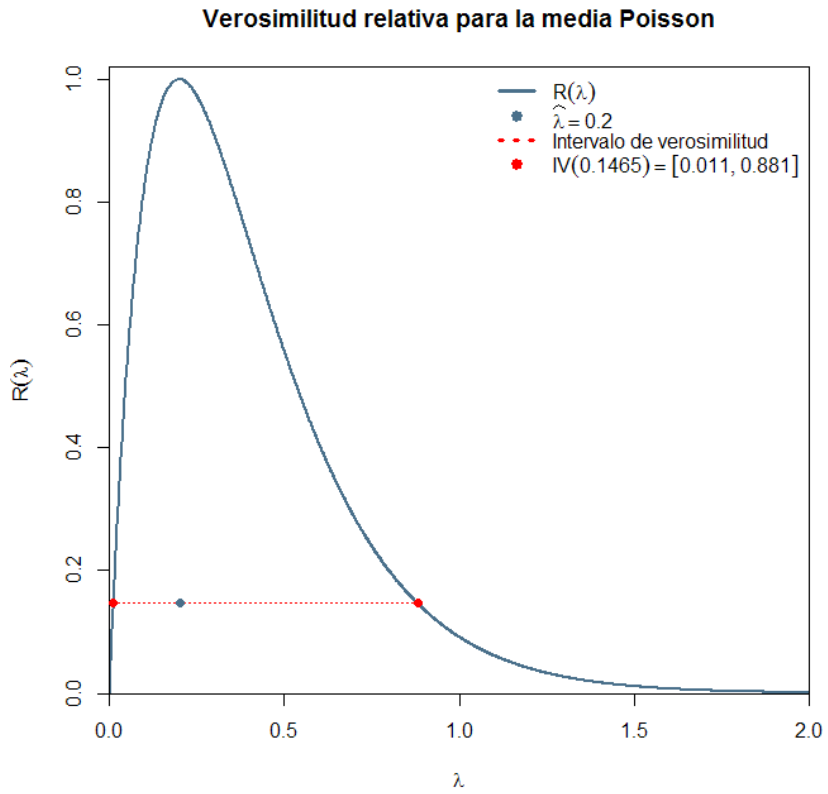
La función de verosimilitud relativa

$$R(\lambda;\mathbf{x}) = 5\lambda \exp[-5\lambda + 1],$$

se muestra en la Figura 1.3.17. En el caso unidimensional, el $IV(c)$ se obtiene trazando un línea horizontal a distancia c, paralela al eje λ en la gráfica de $R(\lambda;\mathbf{x})$. Considerando un nivel $c = 0.1465$, se marca con una línea punteada roja el intervalo de verosimilitud para λ que se obtuvo numéricamente (con ayuda de un programa elaborado en R, basado en la ecuación **??**, ver Apéndice A). Los puntos rojos indican los extremos del intervalo:

$$IV(0.1465) = (0.0114, 0.8806).$$

Se marca con un con un punto el EMV sobre el intervalo de verosimilitud de nivel $c = 0.1465$.



(1.3.17)

Figura 1.3.17: Gráfica de la verosimilitud relativa de λ, para una muestra tal que $\widehat{\lambda} = \overline{x} = 1/n$, con intervalo de verosimilitud de nivel $c = 0.1465$.

Nótese que la verosimilitud relativa de λ es bastante asimétrica con respecto al valor del EMV $\widehat{\lambda}$. Esto puede ser una consecuencia de tener un tamaño de muestra pequeño $n = 5$. Al aumentar el tamaño de muestra, por el Teorema de Máxima Verosimilitud (Pawitan, 2001), se espera que la verosimilitud relativa se vuelva más angosta y se simetrice, de tal forma que se aproximará a una campana de Gauss.

El intervalo de verosimilitud obtenido $(\lambda_1, \lambda_2) = (0.0114, 0.8806)$ está asociado a un nivel de confianza aproximada de $\alpha = 0.05$, lo cual quiere decir que la proporción de veces que el intervalo (λ_1, λ_2) contiene al verdadero valor del parámetro λ_0 (en una cantidad grande de repeticiones del experimento de conteo), es aproximadamente 0.95. Dicho de otra forma,

si repetimos 100 veces el experimento de conteo, esperamos que aproximadamente 95 de las veces el intervalo (λ_1, λ_2) contenga al verdadero valor λ_0.

1.3.4 Calibración de porcentajes de cobertura de intervalos de verosimilitud para muestras pequeñas

Una forma alternativa de asociar una probabilidad de cobertura a un intervalo de estimación $(\theta_1(\mathbf{x}), \theta_2(\mathbf{x}))$ para un parámetro θ, se obtiene mediante simulaciones de muestras aleatorias[3] que provienen del modelo estadístico estimado $f\left(x; \hat{\theta}\right)$, a partir de la muestra original.

Para cada muestra simulada se calcula el intervalo de verosimilitud de nivel $c = 0.1465$ y se verifica si este contiene o no al verdadero valor del parámetro que se supondrá es $\hat{\theta}$, el EMV de la muestra original observada, ya que con este valor se simularon las muestras. Así, se puede obtener la cobertura empírica de este intervalo de verosimiliud. Este porcentaje de cobertura empírico refleja de cierta forma qué tan bueno es el intervalo de verosimilitud obtenido con la muestra original bajo el modelo estimado $f\left(x; \hat{\theta}_n\right)$.

Por otro lado, también mediante simulaciones, se puede obtener un nivel de verosimilitud c, tal que el 95% de los intervalos asociados a las muestras simuladas contengan al verdadero valor del parámetro, en este caso $\hat{\theta}$ para las muestras simuladas. Para ello, basta poner en práctica el siguiente algoritmo:

1. Simular una cantidad suficientemente grande N_s de muestras aleatorias de tamaño n del modelo $f\left(x; \hat{\theta}_n\right)$.

2. Para cada muestra simulada $\mathbf{x}_i = (x_{i,1}, ..., x_{i,n}), i \in \{1, ..., N_s\}$ calcular el EMV de θ, $\hat{\theta}_{i,n}$.

[3]El término *muestra aleatoria* se usará como sinónimo de *muestra de variables aleatorias independientes e idénticamente distrbuidas.*

3. Calcular el valor de la verosimilitud relativa evaluada en el verdadero valor del parámetro $\theta = \hat{\theta}_n$, para cada una de las muestras,

$$R_i = R_i\left(\hat{\theta}_n; \mathbf{x}\right) = \frac{\mathcal{L}\left(\hat{\theta}_n; \mathbf{x}\right)}{\mathcal{L}\left(\hat{\theta}_{i,n}; \mathbf{x}\right)}, i \in \{1, ..., N_s\}.$$

4. Con los valores de $R_i\left(\hat{\theta}_n; \mathbf{x}\right) : \{R_1, ..., R_{N_s}\}$ identificar el nivel de verosimilitud $c \in (0, 1)$ tal que el 95% de los valores de $R_i\left(\hat{\theta}_n; \mathbf{x}\right)$ sean mayores a tal nivel de verosimilitud. Esto se logra calculando el cuantil empírico de probabilidad $\alpha = 0.05$ de la muestra $\{R_1, ..., R_{N_s}\}$. Dicho cuantil q_α representa el nivel de verosimilitud de un intervalo para que su probabilidad de cobertura sea 0.95.

Entonces, el nivel de verosimilitud obtenido mediante este procedimiento $c = q_\alpha$, logrará que el intervalo de verosimilitud que se obtiene a partir de la ecuación

$$R(\theta; \mathbf{x}) \geq q_\alpha,$$

tenga una probabilidad de cobertura aproximada de 0.95.

Ejemplo 2: Calibración de porcentaje de cobertura para el intervalo de verosimilitud de una muestra Poisson de tamaño $n = 5$.

Recordemos que en el Ejemplo 1.*a* se tenía una muestra de tamaño $n = 5, \mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, tal que $x_i = 1$, para algún $i \in \{1, ..., n\}$ y $x_j = 0$,para todo $j \neq i$. Con esta muestra se obtuvo un EMV para λ

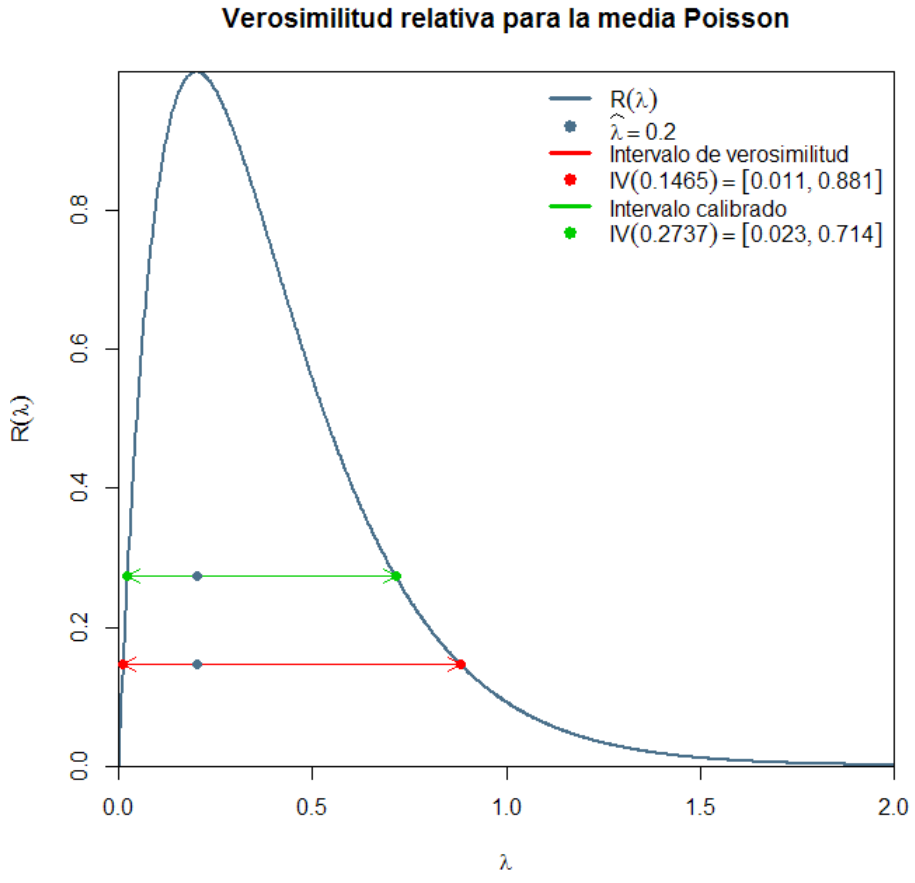
$$\widehat{\lambda}_n = \overline{x}_n = 0.2,$$

y un intervalo de verosimilitud de nivel $c = 0.1465$

$$\begin{aligned} IV(0.1465) &= (\lambda_1, \lambda_2) \\ &= (0.0114, 0.8806). \end{aligned}$$

Siguiendo el algoritmo descrito, se simularon $N_s = 1000$ muestras de tamaño $n = 5$. En este caso, el porcentaje de cobertura empírico obtenido de las N_s muestras fue igual a 97.81%. es decir, un poco mayor al esperado (95%) . De tal forma que existe un nivel $c' \in (0, 1)$ tal que $c' > c = 0.1465$ para el cual se obtiene una probabilidad de cobertura de 0.95.

Aplicando el paso 4 del algoritmo, el cuantil empírico de probabilidad $\alpha = 0.05$ es $q_a = 0.2737$. Esto quiere decir que si se corta la verosimilitud relativa a un nivel $c' = 0.2737$, entonces el intervalo de verosimilitud obtenido tendrá una probabilidad de cobertura aproximada de 0.95. Este intervalo se muestra en la Figura 1.3.18, resulta ser un intervalo más angosto debido a que $c' > c$.



(1.3.18)

Figura 1.3.18: Gráfica de la verosimilitud relativa de λ, para una muestra Poisson tal que $\widehat{\lambda} = 0.2$ con intervalos de verosimilitud de nivel $c = 0.1465$, y de nivel calibrado $c' = 0.2737$ para obtener una cobertura del 95%.

Cabe señalar que este algoritmo será utilizado en este trabajo en un caso en el que el porcentaje de cobertura empírico resulta ser mucho menor que el teórico, de tal forma que los niveles de verosimilitud calibrados a través de simulaciones estarán por debajo de $c = 0.1465$ en ese caso.

1.3.5 Función de verosimilitud perfil

Frecuentemente se tienen modelos estadísticos con varios parámetros desconocidos donde solo interesa estimar uno de ellos. El problema de la estimación por separado de parámetros de interés en presencia del resto de los parámetros, conocidos como parámetros de estorbo, es relevante en estadística puesto que los parámetros de estorbo pueden tener un impacto significativo en las inferencias del parámetro de interés.

Supongamos que el vector de parámetros del modelo $\boldsymbol{\theta} = (\theta_1, ..., \theta_j)$, puede separarse en dos subconjuntos de parámetros

$$\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta \subseteq \mathbb{R}^j,$$

donde θ_1 es el parámetro de interés y tiene dimensión 1 $(\theta_1 \in \Theta_1 \subseteq \mathbb{R})$, y θ_2 contiene a los parámetros de estorbo y es de dimensión $j-1, (\theta_2 \in \Theta_2 \subseteq \mathbb{R}^{j-1})$. Un método muy útil para estimar θ_1 por separado en presencia de del vector de parámetros de estorbo θ_2, está basado en el uso de la función de verosimilitud perfil o maximizada (Pawitan, 2001).

Definición 11 (Función de verosimilitud perfil) *Dada* $\mathbf{x} = (x_1, ..., x_n)$ *una realización de una muestra de tamaño* n *de vaiid, la función de verosimilitud perfil o maximizada del parámetro de interés* θ_1, $\mathcal{L}_P(\theta_1; \mathbf{x})$, *se define como*

$$\mathcal{L}_P(\theta_1; \mathbf{x}) = \max_{\theta_2|\theta_1} \mathcal{L}(\theta_1, \theta_2; \mathbf{x}) = \mathcal{L}\left(\theta_1, \widehat{\theta_2}(\theta_1; \mathbf{x}); \mathbf{x}\right),$$

donde $\widehat{\theta_2}(\theta_1; \mathbf{x})$ *es el estimador de máxima verosimilitud restringido (emvr) de* θ_2 *para un valor fijo de* θ_1.

El emvr $\widehat{\theta_2}(\theta_1; \mathbf{x})$ es el valor de θ_2 que tiene mayor plausibilidad para un valor fijo de θ_1 dada la muestra observada $\mathbf{X} = \mathbf{x}$. Es decir, la verosimilitud perfil de θ_1 se obtiene maximizando la función de verosimilitud global $\mathcal{L}(\theta_1, \theta_2; \mathbf{x})$ sobre θ_2 pero fijando θ_1. De esta manera se reduce la dimensión de la verosimilitud j, a una sola dimensión, la del parámetro de interés θ_1.

Cuando θ_1 y θ_2 son parámetros unidimensionales, la función de verosimilitud global , $\mathcal{L}(\theta_1, \theta_2; \mathbf{x})$, es una superficie en $\mathbb{R}^3$ cuyo dominio es el plano cartesiano correspondiente al espacio parametral $\Theta_1 \times \Theta_2$. Es decir, la verosimilitud perfil está asociada a una trayectoria sobre esta superficie y será una función real valuada que corresponde a una curva en $\mathbb{R}^2$. Así, al colocarse en un punto sobre el eje real muy distante del parámetro de estorbo θ_2, la silueta o perfil que se observa de esta verosimilitud global $\mathcal{L}(\theta_1, \theta_2; \mathbf{x})$ es justamente la función de verosimilitud maximizada de θ. Por esta razón también recibe el nombre de verosimilitud perfil.

Algunas propiedades importantes de la función de verosimilitud perfil

1. El emv perfil de θ es igual al emv global o no restringido $\widehat{\theta}$,

$$\widehat{\theta_2} = \widehat{\theta_2}\left(\widehat{\theta_1}; \mathbf{x}\right).$$

2. La función de verosimilitud perfil relativa (estandarizada para tomar el valor de uno en su máximo):

$$R_P(\theta; \mathbf{x}) = \frac{\mathcal{L}_P(\theta; \mathbf{x})}{\mathcal{L}_P\left(\widehat{\theta}; \mathbf{x}\right)},$$

sigue la teoría asintótica usual (Pawitan, 2001). Más precisamente, bajo condiciones de regularidad, se tiene que

$$W = -2 \log R_P(\theta_0; \mathbf{x}) \xrightarrow{d} \chi^2_{(1)},$$

donde θ_0 es el verdadero valor del parámetro. Este resultado si bien es asintótico, generalmente se cumple para muestras de tamaño moderado e incluso pequeño.

3. Un intervalo o región de verosimilitud de nivel c, obtenido a partir de la función de verosimilitud perfil de θ es

$$R_c = \{\theta | R_P(\theta; \mathbf{x}) \geq c\}, 0 \leq c \leq 1.$$

Además, R_c también es un intervalo de confianza aproximada para el parámetro de interés θ (Pawitan, 2001).

Ejemplo 3: Verosimilitud perfil de N en un modelo Binomial (N, p), cuando N y p son desconocidos.

Considérese una realización $\mathbf{x} = (x_1, ..., x_n)$ de una muestra de n v.a.i.i.d $\mathbf{X} = (X_1, ..., X_n)$ con distribución común Binomial de parámetros N y p, desconocidos. Nos interesa hacer inferencia sobre el tamaño poblacional N, por tanto, el parámetro p que también es desconocido, es considerado como parámetro de estorbo.

La probabilidad conjunta de esta muestra es

$$\begin{aligned} \mathbf{P}[\mathbf{X} = \mathbf{x}; N, p] &= \prod_{i=1}^{n} \mathbf{P}[X_i = x_i; N, p] \\ &= p^t (1-p)^{Nn-t} \prod_{i=1}^{n} \binom{N}{x_i}, \end{aligned} \tag{1.3.18}$$

donde $t = \sum_{i=1}^{n} x_i$.

De la Definición 1, la función de verosimilitud global de N y p es proporcional a la probabilidad de observar la muestra dada en 1.3.18, y es

$$\mathcal{L}(N, p; \mathbf{x}) = p^t (1-p)^{Nn-t} \prod_{i=1}^{n} \frac{N!}{(N-x_i)!}, \ 0 < p < 1, N \in \mathbb{N}. \tag{1.3.19}$$

La función de verosimilitud perfil de N se obtiene maximizando sobre p la función de verosimilitud dada en 1.3.19, fijando el valor de N. Para un valor fijo de N la logverosimilitud asociada es

$$\ell(N, p; \mathbf{x}) = t \log(p) + (Nn - t) \log(1-p) + n \log N! - \sum_{i=1}^{n} \log(N - x_i)!. \tag{1.3.20}$$

Al derivar con respecto a p, se tiene

$$\frac{\partial \ell\left(N,p;\mathbf{x}\right)}{\partial p}=\frac{t}{p}-\frac{Nn-t}{1-p}.$$

Si se resuelve la ecuación $\partial \ell\left(N,p;\mathbf{x}\right)/\partial p=0$ para luego despejar p, se obtiene el estimador de máxima verosimilitud de p restringido a N,

$$\widehat{p}\left(N\right)=\frac{t}{Nn}.$$

Así, la función de verosimilitud perfil de $N,\mathcal{L}_{P}\left(N;\mathbf{x}\right)$, se obtiene al reemplazar el emvr de p en 1.3.19, es decir,

$$\begin{aligned}\mathcal{L}_{P}\left(N;\mathbf{x}\right) &= \mathcal{L}\left(N,\widehat{p}\left(N\right);\mathbf{x}\right) \\ &= \left(\frac{t}{Nn}\right)^{t}\left(1-\frac{t}{Nn}\right)^{Nn-t}\prod_{i=1}^{n}\binom{N}{x_{i}}.\end{aligned} \tag{1.3.21}$$

Se ilustrará lo anterior con un ejemplo de datos de abundancias de impalas[4], anteriormente analizado en la literatura de modelos de abundancias (Montoya, 2008), para ilustrar la forma de la verosimilitud perfil de N. Un avión sobrevoló una zona geográfica de interés en el Parque Nacional de Kruger en Sudáfrica durante cinco días consecutivos, realizando conteos de manadas con más de 25 impalas. Es razonable suponer que la probabilidad p, de observar a una manada, es la misma para todos los días. También es razonable suponer que la variable X_i que representa el número de manadas que se observan en un día dado, $i=1,...,5$, sigue una distribución Binomial con parámetros N y p. El parámetro de interés es N, el número total de manadas de al menos 25 impalas en esta zona del parque Kruger. El tamaño de la muestra observada es $n=5$ (pequeña) y los valores observados fueron:

$$\mathbf{x}=(15,20,21,23,26)$$

donde los valores del estadístico $t=\sum_{i=1}^{n}x_{i}$ y la media muestral son

$$t=105,\overline{x}=21.$$

[4]Un impala es un antílope africano de tamaño medio, su nombre proviene de la lengua africana Zulu y significa gacela.

La verosimilitud perfil de N para esta muestra es entonces

$$\mathcal{L}_P(N;\mathbf{x}) = \left(\frac{105}{5N}\right)^{105}\left(1-\frac{105}{5N}\right)^{5N-105}\prod_{i=1}^{5}\binom{N}{x_i}.$$

A partir de esta expresión se puede obtener el EMV perfil de N. Este coincide con el EMV global de N., obtenido a partir de la verosimilitud global dada en 1.3.19. Para ello, es necesario maximizar numéricamente $\mathcal{L}_P(N;\mathbf{x})$ con respecto a N, o equivalentemente, maximizar el logaritmo $\ell_P(N;\mathbf{x}) = \ln\left(\mathcal{L}_P(N;\mathbf{x})\right)$. Con ayuda de un programa sencillo de R se obtiene numéricamente el EMV para N,

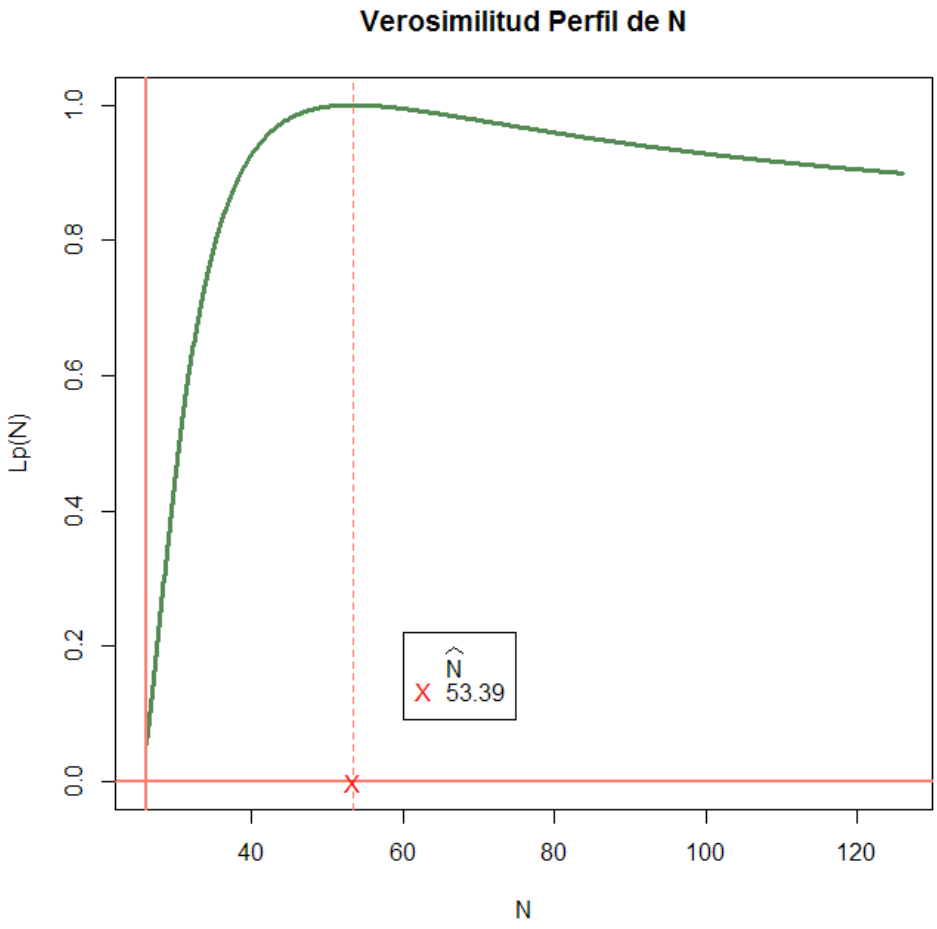
$$\widehat{N} = 53.95.$$

El valor del EMV restringido de p, evaluado en $\widehat{N}$, coincide con el EMV global de p

$$\widehat{p} = \widehat{p}\left(\widehat{N}\right) = \frac{105}{5\widehat{N}} = \frac{105}{5\,(53.95)} = 0.39,$$

que resulta ser un valor moderado para p.

En la Figura 1.3.22 se presenta la gráfica de la función de verosimilitud perfil relativa de N, donde se señala $\widehat{N}$ (en el eje X) con un símbolo "X". La línea vertical en guiones indica el mínimo valor que puede tomar N. Nótese que N tiene que ser mayor o igual al máximo de los valores observados. En este caso ocurre que $N \geq x_{(5)}$, donde $x\,(5) = máx\,\{x_1, ..., x_5\} = 26$. Se observa que la curva de la verosimilitud perfil de N es muy plana y por tanto no es muy informativa; solamente indica que N debe ser mayor a 26 pero no da una cota superior razonable. Incluso, los intervalos de verosimilitud de niveles menores al 25% no tienen extemo derecho finito. Cabe señalar que esto no es una falla de la verosimilitud sino de tener pocas observaciones que no permiten identificar bien al modelo Binomial (Montoya, 2008).



(1.3.22)

Figura 1.3.22. Verosimilitud perfil relativa de N. Ejemplo de muestra chica de abundancias de impalas. La curva de la perfil relativa no es informativa para N pues no da alguna cota superior razonable. Los EMV de N y p son: $\widehat{N} = 53.3, \widehat{p} = 0.39$.

1.3.6 Modelos de mezcla infinita o modelos jerárquicos

Fisher propuso una mezcla infinita Poisson-Gamma para describir las probabilidades de conteos de individuos de una especie arbitraria (Fisher et al., 1943). En su ejemplo, consideró

que la variable aleatoria X_j que cuenta individuos de la especie j, sigue una ley Poisson con media λ_j. Como esta media y la de otras especies son desconocidas, y además diferentes entre sí debido a la heterogeneidad del área geográfica que habitan, es razonable suponer que a su vez $\Lambda = \lambda_j$ es una variable aleatoria. Por ejemplo, supongamos que Λ sigue una ley Gama que dependerá sólo de dos parámetros (α, β):

$$\Lambda \sim Gama\,(\alpha, \beta)\,.$$

Así, los conteos originales de la especie j se pueden describir con la probabilidad condicional

$$\mathbf{P}\left[X_j = x | \Lambda = \lambda_j\right] = \frac{\lambda_j^x e^{-\lambda_j}}{x!}, \tag{1.3.23}$$

que corresponde a una probabilidad Poisson. De tal forma que cuando el número total de especies es suficientemente grande, los conteos de una especie arbitraria se pueden concebir como provenientes de una mezcla infinita Poisson-Gamma

$$\mathbf{P}\left[X = x\right] = \int_0^\infty \mathbf{P}\left[X = x | \Lambda = \lambda\right] f_\Lambda\left(\lambda\right) d\lambda,$$

donde $\mathbf{P}\left[X = x | \Lambda = \lambda\right]$ es la probabilidad asociada a cada una de las especies, como en la ecuación 1.3.23.

Este tipo de mezclas son llamadas también densidades compuestas y son de la forma

$$f\left(x\right) = \int g\left(x; \theta\right) dH\left(\theta\right),$$

donde H es una medida de probabilidad sobre el espacio parametral Θ.

La ventaja principal de un modelo de mezcla infinita de distribuciones sobre una de mezcla finita, es que se logra una reducción importante del número de parámetros desconocidos. En el caso de la mezcla infinita Poisson-Gama, originalmente se tenían tantos parámetros desconocidos como número total de especies k en la zona de interés: $\lambda_1, ..., \lambda_k$. Luego, al considerar que $\Lambda = \lambda_j$ es una v.a. Gama de parámetros (α, β) desconocidos, el número de parámetros desconocidos se reduce de k a 2.

En general, en estadística Bayesiana, estos modelos son conocidos como modelos jerárquicos (Jiménez, 2011), mientras que en un contexto no Bayesiano considerar que los parámetros

de una distribución de probabilidad son a su vez variables aleatorias es equivalente a trabajar con distribuciones compuestas o mezclas de distribuciones. En ambos casos se pueden incluir tantos niveles de estocasticidad como sean necesarios, sin olvidar que debemos tratar de conservar el equilibrio entre la complejidad de los modelos y la aproximación a la realidad (principio de parsimonia).

De esta manera, se considera que los datos son generados por un proceso compuesto o proceso jerárquico, donde el parámetro λ_j correspondiente a la especie j, es muestreado de una densidad a priori f_Λ, con parámetro θ en el nivel 2 de jerarquía:

$$\Lambda \sim f_\Lambda(\lambda;\theta).$$

A $f_\Lambda(\lambda;\theta)$ se le conoce como *hiperdensidad* y a los parámetros θ se les llama *hiperparámetros.* Mientras que en el primer nivel de jerarquía las observaciones son muestreadas de una distribución condicional dados los parámetros de cada especie

$$(X_{ij}|\Lambda = \lambda_j) \sim f(x_{ij}|\lambda_j), j = 1, ..., k.$$

En el estudio de riqueza de especies de una comunidad ecológica resulta conveniente plantear modelos jerárquicos basados en conteos Poisson que nos permitan incluir heterogeneidad sobre las diferentes especies que habitan la comunidad. Consideraremos el modelo jerárquico sugerido por Fisher para plantear un modelo estadístico que nos permita estimar el número de especies detectables en una zona geográfica de interés.

1.3.7 Gráficas cuantil-cuantil para datos censurados por intervalo

Definición 12 (Función cuantil) *La función cuantil de una distribución de probabilidad* $F : \mathbb{R} \longrightarrow (0,1)$ *se define como la inversa generalizada de la función de distribución*

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : \alpha \leq F(x)\},$$

para una probabilidad $\alpha \in (0,1)$. *Así, la función cuantil devuelve el valor mínimo de* x *tal que* $\mathbf{P}[X \leq x] = \alpha$.

Definición 13 (Cuantil de probabilidad α) *Sean $\alpha \in (0,1)$ y $F : \mathbb{R} \longrightarrow (0,1)$ una función de distribución. El cuantil de probabilidad α, q_α, se define como*

$$q_\alpha = Q(\alpha).$$

En general, una gráfica cuantil-cuantil (o gráfica Q-Q) es un método gráfico que se usa para comparar dos distribuciones graficando puntos cuyas coordenadas son los cuantiles de las dos observaciones asociados a diferentes probabilidades.. En este trabajo se usarán gráficas Q-Q para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria $\mathbf{x} = (x_1, ..., x_n)$, y una distribución teórica estimada a partir de la muestra $\mathbf{x}$. Es decir, compararemos la distribución empírica que se obtiene con la muestra $\mathbf{x}$ contra cierta distribución estimada cuyos parámetros serán los EMV obtenidos con la muestra $\mathbf{x}$.

Cuando la muestra no contiene datos censurados, el $i-$ésimo punto en la gráfica Q-Q corresponde al $i-$ésimo cuantil de la distribución teórica estimada (eje X), $\widehat{q}_{\frac{i}{n+1}}$, asociado a la probabilidad $\frac{i}{n+1}$, contra el $i-$ésimo cuantil de la función de distribución empírica (eje Y), el cual corresponde a la $i-$ésima estadística de orden, $x_{(i)}$. Esto es, $(x_i, y_i) = \left(\widehat{q}_{\frac{i}{n+1}}, x_{(i)}\right), i = 1, ..., n$. Sin embargo, no es inmediato construir una gráfica Q-Q cuando los datos poseen algún tipo de censura. Si no se conoce el valor preciso de la observación x_i no es claro cuál es su cuantil empírico asociado. En el caso de tener datos censurados por intervalo, sólo se sabe cuántas observaciones ocurren dentro de cada uno de esos intervalos. A continuación se propone una forma de construir gráficas Q-Q en este caso particular.

Sea $\mathbf{x} = (x_1, ..., x_n)$ una muestra aleatoria de observaciones censuradas por intervalo. Supóngase que el modelo estimado que se cree describe a los datos tiene función de distribución $F\left(x; \widehat{\boldsymbol{\theta}}\right)$ donde $\widehat{\boldsymbol{\theta}}$ es el EMV del parámetro desconocido. Denotaremos por I_{cen} al conjunto de intervalos en los que se sabe ocurren las observaciones de la muestra, $I_{cen} = \{[A_l, B_l]\}_{l=1}^{s}$, de tal forma que $\forall i \in \{1, ..., n\}\ \exists l \in \{1, ..., s\}$ tal que $x_i \in [A_l, B_l]$. Es decir, toda observación pertenece a alguno de los intervalos del conjunto I_{cen}.

Ahora bien, dado que los datos son censurados por intervalo, el valor preciso de cada x_i es desconocido pero se sabe cuántas observaciones pertenecen a cada uno de los in-

tervalos $[A_l, B_l]$. Sea f_l la frecuencia de ocurrencia de las observaciones en el intervalo $[A_l, B_l], l = 1, ..., m$. Nótese que $n = \sum_{l=1}^{s} f_l$. Se supondrá que las f_l observaciones ocurrieron uniformemente dentro del intervalo $[A_l, B_l]$.

Entonces, se propone el siguiente algoritmo para construir una gráfica Q-Q que permita contrastar la información que se tiene de la muestra contra el modelo especificado para los datos $F(x; \boldsymbol{\theta})$:

1. A partir del modelo estimado $F\left(x; \widehat{\boldsymbol{\theta}}\right)$, las coordenadas en el eje X se obtienen de manera usual. Dadas las probabilidades $\alpha_i = \frac{i}{n+1}, i = 1, ..., n$, se calculan los cuantiles teóricos correspondientes al modelo estimado $F\left(x; \widehat{\boldsymbol{\theta}}\right) : \widehat{q}_{\alpha_1}, ..., \widehat{q}_{\alpha_n}$ (según la Definición 13).

2. Para obtener los cuantiles empíricos correspondientes a las observaciones que ocurrieron en el intervalo $[A_j, B_j]$, se simulan f_j v.a. uniformes en $[A_j, B_j] : u_{j,1}, ..., u_{j,f_j}$. Este paso se repite para cada una de las frecuencias $f_l > 1$. Para aquellas con $f_l = 1$, se tomará como cuantil empírico asociado a la observación registrada en ese intervalo.

3. Basta con ordenar las observaciones simuladas.en el paso anterior para obtener los cuatiles empíricos correspondientes. El $i-$ésimo cuantil empírico es $u_{(i)}$, la $i-$ésima estadística de orden de la muestra $\{u_{l,1}, ..., u_{l,f_l}\}_{l=1}^{s}$.

4. Graficamos los puntos $\left(u_{(i)}, \widehat{q}_{\alpha_i}\right)$.

Las simulaciones realizadas en el paso 3 representarán a las observaciones censuradas.

Como es usual, se espera que los puntos graficados tengan un arreglo lineal alrededor de la recta identidad. En caso contrario, el modelo especificado para describir la muestra no es el más adecuado.

Ahora bien, los puntos graficados por sí solos no son tan fáciles de interpretar, se recomienda comparar el conjunto de puntos asociados a la muestra original con los que se obtienen de distintas muestras simuladas bajo la distribución estimada $F\left(x; \widehat{\boldsymbol{\theta}}\right)$. Con esto se podrá observar la variabilidad inherente a muestras provenientes de tal distribución. Por

lo tanto, si el conjunto de puntos asociados a la muestra original se observa dentro de la nube de puntos de las simulaciones, y además presentan un arreglo lineal alrededor de la recta identidad, entonces se dirá que el modelo $F\left(x;\widehat{\boldsymbol{\theta}}\right)$ es razonable para los datos.

1.3.8 Razón de verosimilitudes para la comparación de modelos

Bajo el paradigma de máxima verosimilitud, cuando se desea comparar dos modelos A y B que ajusten bien a los datos y elegir cuál de estos es el modelo que haga más probable a los datos observados, se usan cocientes de funciones de verosimilitud evaluadas en sus correspondientes estimadores de máxima verosimilitud.

Definición 14 (Cociente o razón de verosimilitudes) *Dada* $\mathbf{x} = (x_1, ..., x_n)$ *una muestra de tamaño* n *de vaiid, la estadística*

$$\Lambda(A, B) = \frac{L_A\left(\hat{\theta}_A; \mathbf{x}\right)}{L_B\left(\hat{\theta}_B; \mathbf{x}\right)},$$

es llamada cociente o razón de verosimilitudes, la cual compara las plausibilidades de los modelos A *y* B *a la luz de los datos* $\mathbf{x}$*, donde* $L_A\left(\hat{\theta}_A; \mathbf{x}\right)$ *es la verosimilitud bajo el modelo* A*, evaluada en su respectivo EMV,* $\hat{\theta}_A$*; y* $L_B\left(\hat{\theta}_B; \mathbf{x}\right)$ *es la verosimilitud bajo el modelo* B*, evaluada en* $\hat{\theta}_B$*.*

El valor de $\Lambda(A, B)$ indica cuantas veces más probable hace el modelo A a los datos observados que el modelo B. Por ejemplo, el cociente $\Lambda(A, B) = d > 1$ significa que el modelo A es d veces más plausible que el modelo B.

Otro criterio de selección de modelos muy utilizado es el Criterio de Akaike (Pawitan, 2001). A diferencia del cociente de verosimilitudes, este criterio es utilizado cuando los modelos que se desean comparar no son anidados y tienen diferente número de parámetros. Se recomienda considerar también este criterio cuando se trabaja con ese tipo de modelos.

Capítulo 2

Planteamiento del modelo estadístico general

El número k de especies detectables de un grupo taxonómico de interés que habitan en una región homogénea delimitada, en un periodo de tiempo dado y bajo ciertas condiciones ambientales, es el parámetro principal que se desea estimar. Este parámetro resulta ser de gran interés para el monitoreo y comparación de comunidades como se ha descrito en la Sección 1.1. En este capítulo se presentará una metodología estadística adecuada para estimar k.

2.1 Datos observados: conteos de individuos y especies

En la región de interés que se elige lo más homogénea posible en cuanto a características físicas y ambientales, se eligen al azar r cuadrantes de área h. El número r se elige según el presupuesto que se tenga para el estudio y con base en la precisión de las inferencias sobre k que se desean obtener.

Los datos consisten en conteos de individuos X_{ij} de la especie j que fueron observadas en el cuadrante i, donde $i = 1, ..., r$ y $j = 1, ..., m$. El número M de especies detectadas en la muestra de r cuadrantes es a su vez una variable aleatoria discreta observable. El valor que toma la variable M para un conjunto de datos particular se denotará por m, donde $m \in \{1, ..., k\}$ siendo k el número de especies detectables, esto es, que tienen al menos un individuo con alta probabilidad en la región de interés. Es claro que $k \geq M > 0$.

En la Tabla 2.1.1 se resume la información colectada en un muestreo por cuadrantes de tamaño r cuando se observan m especies. Las variables X_j denotan el número total de individuos de la especie j observados en todos los cuadrantes. Así,

$$X_j = \sum_{i=1}^{r} X_{ij}, j = 1, ..., m.$$

Cuadrantes:	**1**	**2**	$\cdots$	**r**	**Individuos por especie**
Especie 1	X_{11}	X_{21}	$\cdots$	X_{r1}	X_1
Especie 2	X_{12}	X_{22}	$\cdots$	X_{r2}	X_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Especie m	X_{1m}	X_{2m}	$\cdots$	X_{rm}	X_m

(2.1.1)

Tabla 2.1.1: Tabla de conteos obtenida de un muestreo por cuadrantes. Se han observado m especies distintas en r cuadrantes y X_j es el número de individuos observados por especie en todos los cuadrantes.

Los datos relevantes se resumen en un vector aleatorio de dimensión $m+1$. Las primeras m componentes de este vector son las variables aleatorias discretas $X_1, ..., X_m$, correspondientes al número total de individuos de cada una de las especies observadas en la muestra, y la componente $m+1$ corresponde al número observado de especies detectables $M = m$ en los r cuadrantes:

$$(X_1, ..., X_m, M)\,. \tag{2.1.2}$$

Sin pérdida de generalidad, las observaciones de las v.a. de las primeras m componentes serán presentadas en orden decreciente, de tal forma que $X_1 \geq ... \geq X_m$.

Para fines del planteamiento del modelo, será conveniente considerar el siguiente vector aleatorio en lugar del vector dado en 2.1.2, el cual contiene información equivalente. Al reemplazar cada X_j por su promedio correspondiente sobre los r cuadrantes, $T_{rj} = X_j/r, j = 1, ..., k$, el vector resultante es

$$(T_{r1}, ..., T_{rm}, M)\,, \tag{2.1.3}$$

este es el vector que usaremos.

2.2 Supuestos importantes y distribuciones de variables aleatorias de interés

Supóngase que el número de especies detectables en la región de interés es k. Es decir, hay k especies tales que la probabilidad de que haya al menos un individuo de cada una de ellas en toda la región de interés es alta. La región de interés se considera homogénea en cuanto a condiciones climáticas y físicas. Los supuestos importantes de los cuales depende el modelo estadístico con el que se propone estimar el parámetro de interés k, se listan en las siguientes secciones.

2.2.1 Distribución Poisson del número de individuos de una especie detectable

Se supondrá que los individuos de cada especie son asignados independientemente y al azar sobre la región homogénea de interés. Se dice que tienen un patrón aleatorio o que están dispersos aleatoriamente. Bajo el supuesto de distribución uniforme de los individuos, el número de individuos de una especie dada detectados por unidad de área sigue una ley Poisson (Jiménez, 2011). Esto es, si Y_j es la variable aleatoria que cuenta el número de individuos de la especie j presentes en la región de interés de superficie A, entonces

$$Y_j \sim Poisson\left(\varphi_j\right), \tag{2.2.1}$$

donde φ_j es el parámetro de intensidad Poisson y es el valor esperado de individuos de la especie j presentes en A. A veces este parámetro se interpreta como la tasa de detección de la especie j (Mao y Colwell, 2005).

Típicamente el área total de estudio A, consiste de decenas, cientos o miles de hectáreas. De tal forma que no es posible recorrer cada metro cuadrado dentro de A y además contar los individuos de cada especie. Además de significar un esfuerzo de colecta enorme y costoso, el tiempo invertido en este recorrido puede necesitar de muchos años invertidos. Comúnmente, cuando se realizan estudios para investigar el patrón de agregación o la abundancia en poblaciones, se usan muestreos con cuadrantes escogidos aleatoriamente dentro del área de interés (Pielou, 1969; Magurran, 2011; Seber, 1982). El muestreo por cuadrantes aleatorios será central en esta tesis para obtener información sobre k.

Para una muestra observada en r cuadrantes se observa el núméto total de especies distintas m. Para cada cuadrante $i \in \{1, ..., r\}$ y para cada especie $j \in \{1, ..., m\}$ se tiene una variable aleatoria X_{ij} que denota el número de individuos de la especie j detectados en el cuadrante i. Dado que la región de interés de área A se ha supuesto homogénea, al fijarse en cada uno de los cuadrantes se conservan las características de la región total. Por tanto, es razonable suponer que las v.a. X_{ij} son independientes (por tratarse de conteos sobre cuadrantes que no se traslapan) y siguen una ley Poisson.

Supóngase que cada uno de los cuadrantes de la muestra tiene la misma área h y que este tamaño ha sido elegido de tal forma que dentro de la superficie total A se pueden acomodar W cuadrantes. Entonces, se tiene la siguiente relación entre el área total y el área de un cuadrante:

$$A = Wh. \tag{2.2.2}$$

Nótese que el área total muestreada de r cuadrantes de área k es rh. Por lo que se da la relación

$$\frac{rh}{A} = \frac{rh}{Wh} = \frac{r}{W}, \tag{2.2.3}$$

que representa la proporción de área muestreada del total A.

Por otro lado, existe una relación entre los parámetros de cada una de las v.a. Y_j y los parámetros de la v.a. X_{ij}, por tratarse de conteos sobre áreas proporcionales. Dicha relación es

$$\varphi_j = W\lambda_j, \tag{2.2.4}$$

por lo que para cada $j = 1, ..., m$, se tiene que

$$\lambda_j = \frac{\varphi_j}{W}. \tag{2.2.5}$$

Así, el número esperado λ_j, de individuos de la especie j en un cuadrante de área h, se obtiene a partir de φ_j al conocer W (el número de cuadrantes de área h que caben dentro de A).

En resumen, los conteos de las diferentes especies en un cuadrato de área h, siguen una ley Poisson de parámetro λ_j,

$$X_{ij} \sim Poisson\left(\lambda_j\right), j \in \{1, ..., m\}. \tag{2.2.6}$$

2.2.2 Cota inferior para los parámetros de intensidad Poisson

Para que el número de especies detectables k, en una región de superficie A esté bien definido, se requiere que cada especie esté representada en A por al menos un individuo con probabilidad alta. Sin embargo, estas k probabilidades, si bien todas son altas, usualmente son

diferentes entre sí debido a las respuestas distintas de cada especie frente a factores ambientales. Se supondrá aquí además que el esfuerzo de muestreo es el mismo en todos los cuadrantes, para que esto no sea un factor adicional que influya en la posibilidad de detectar una especie dada.

El requisito anterior implica la siguiente restricción en términos de las variables aleatorias $Y_1, ..., Y_k$ que representan el número de individuos de cada una de las k especies y que se supusieron siguen distribuciones Poisson con parámeros φ_j para $j = 1, ..., k$. Las funciones de probabilidad de estas variables son

$$P\left[Y_j = y; \varphi_j\right] = \frac{\varphi_j^y e^{-\varphi_j}}{y!}, y = 0, 1, ...$$

De esta forma, la probabilidad de ver al menos un individuo de la especie j en todo A está dada en términos de su parámetros de intensidad λ_j, como

$$P\left[Y_j \geq 1; \varphi_j\right] = 1 - P\left[Y_j = 0; \varphi_j\right] = 1 - e^{-\varphi_j}.$$

Se supondrá entonces que estas probabilidades son todas mayores que un cierto valor p tal que $0.9 \leq p \leq 1$. Con ello se tienen las siguientes restricciones que deben cumplir los parámetros de intensidad Poisson,

$$1 - e^{-\varphi_j} \geq p \Leftrightarrow \varphi_j \geq -\ln(1-p) \tag{2.2.7}$$

Esto es, los valores esperados Poisson deben ser razonablemente grandes para que con alta probabilidad haya al menos un individuo de la especie j en la región. Si esto no fuese así, entonces casi imposible ver a esa especie en alguno de los cuadrantes. Nótese que no se tiene la certeza absoluta que haya al menos un individuo de la especie j en A; se pide una condición más débil, que la probabilidad de que estén presentes uno o más individuos en A sea alta. Al cumplirse estas condiciones, el parámetro k queda bien definido[1].

[1]Nótese que el valor del parámetro p debe ser fijado desde el planteamiento del modelo estadístico. Su valor no debe ser elegido con base en las caracterísitcas de los conteos observados, sino con base en los conocimientos previos de la zona geográfica y las especies de interés.

Dada la relación que guardan los parámetros φ_j y λ_j (ver ec. 2.2.4), se tiene que la restricción 2.2.7 impone las siguientes condiciones para los parámetros Poisson asociados a los cuadrantes,

$$\lambda_j \geq \frac{-\ln(1-p)}{W}, j = 1, ..., k. \tag{2.2.8}$$

En resumen, para que el parámetro k, esté bien definido, se debe pedir que sea posible detectar a la especie en la región de interés. Esto implica que la probabilidad de observar al menos un individuo de cada especie en un cuadrante debe ser más alta que una cierta cota inferior p. Todas estas condiciones, imponen una cota inferior λ_0, para las intensidades Poisson asociadas a cada cuadrante y para todas las k especies detectables, donde

$$\lambda_0 = \frac{-\ln(1-p)}{W}, \tag{2.2.9}$$

la cual depende de dos cantidades: p y la razón de áreas de la región de interés sobre la de un cuadrante, $W = A/h$. Es decir, se debe cumplir que

$$\lambda_j \geq \lambda_0, \text{ para } j = 1, ..., k. \tag{2.2.10}$$

2.2.3 Distribución truncada de los parámetros de intensidad Poisson

En general, la heterogeneidad de las abundancias de especies distintas se debe a diversos factores ambientales y a las propiedades particulares de cada especie. Cada especie puede desarrollarse de forma diferente dependiendo de la disponibilidad de recursos, la competencia con otras especies, la presencia de predadores, las condiciones climáticas, etc. Estos hechos se resumen en contar con valores distintos de los parámetros de intensidad Poisson correspondientes a las especies distintas.

Para describir la heterogeneidad de las especies, R. A. Fisher (1943) sugirió suponer que los parámetros de intensidad Poisson eran variables aleatorias independientes entre sí, que

seguían una distribución común Gama, cuyo soporte considera solamente valores estrictamente positivos. Esta distribución juega el rol de una hiperdensidad (ver Sección 1.3.6) para los parámetros λ_j que serán entonces considerados como variables aleatorias (Jiménez, 2011). Sin embargo, para que se cumpla la restricción 2.2.10, conviene considerar distribuciones truncadas por la izquierda en λ_0. En esta tesis se considerará no solamente una distribución Gama sino que también una Lognormal para describir la distribución aleatoria de los parámetros de intensidad Poisson. Más adelante se definirá la versión truncada de ambas.

A este tipo de modelos que tienen dos o más niveles de aleatoriedad se les llama jerárquicos. En un primer nivel está el comportamiento aleatorio de las abundancias de cada especie, que se ha sugerido siguen distribuciones Poisson con distintos parámetros de intensidad. En el segundo nivel se describe el comportamiento aleatorio de estos parámetros Poisson, suponiendo que siguen una distribución truncada Gama o Lognormal. Esta distribución caracteriza de manera global la comunidad en la región de interés. Por ello va a ser de interés poderlas estimar a partir de observaciones, para posteriormente comparar las distribuciones estimadas para distintas regiones, o para distintos momentos en el tiempo.

Considérese entonces una muestra de variables aleatorias $\Lambda_1, ...\Lambda_k$ que son independientes e idénticamente distribuidas como $G(\lambda;\theta)$, donde θ denota al vector de parámetros desconocidos. La distribución G es la versión truncada por la izquierda en λ_o de una distribución F, que será Gama o Lognormal. La distribución y densidad truncadas correspondientes son

$$G(\lambda;\theta) = \frac{F(\lambda;\lambda_0)}{1 - F(\lambda_0;\theta)}, \tag{2.2.11}$$

$$g(\lambda;\theta) = \frac{f(\lambda;\theta)}{1 - F(\lambda_0;\theta)}, \tag{2.2.12}$$

respectivamente, para toda $\lambda \geq \lambda_0 \geq 0$.

En resumen, se considerará el siguiente modelo jerárquico. En el primer nivel se consideran distribuciones Poisson para los conteos de invididuos X_{ij} de la especie j en el cuadrante i de área h,

$$(X_{ij}|\Lambda = \lambda_j) \sim Poisson(\lambda_j), j = 1, ..., k.$$

En el segundo nivel se considera una distribución truncada para los parámetros de intensidad Poisson correspondientes,

$$\Lambda \sim G\left(\lambda; \theta, \lambda_0\right), j = 1, ..., k.$$

Cabe señalar que valores pequeños de λ_j harán que la especie j sea difícil de observar, pues un valor pequeño está asociado típicamente a especies raras o poco abundantes. Mientras que un valor grande de λ_j hará que la especie j sea observada con mayor facilidad en uno o más cuadrantes de la muestra. Así pues, a través de la forma de la densidad de Λ se podrá inferir si en el grupo taxonómico bajo estudio hay muchas o pocas especies raras. Esto se exhibirá en los ejemplos del Capítulo 4.

Densidad Gama de parámetros (α, β) :

Supóngase que la variable aleatoria Λ se distribuye como una Gama de parámetros (α, β). La función de densidad de Λ es

$$f\left(\lambda; \alpha, \beta\right) = \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} e^{-\beta\lambda} \mathbf{1}_{[0,\infty)}\left(\lambda\right), \tag{2.2.13}$$

para valores estrictamente positivos de α y β. Donde $\Gamma\left(x\right)$ es la función Gama que se define como:

$$\Gamma\left(x\right) = \int_0^{\infty} t^{x-1} e^{-t} dt, x > 0,$$

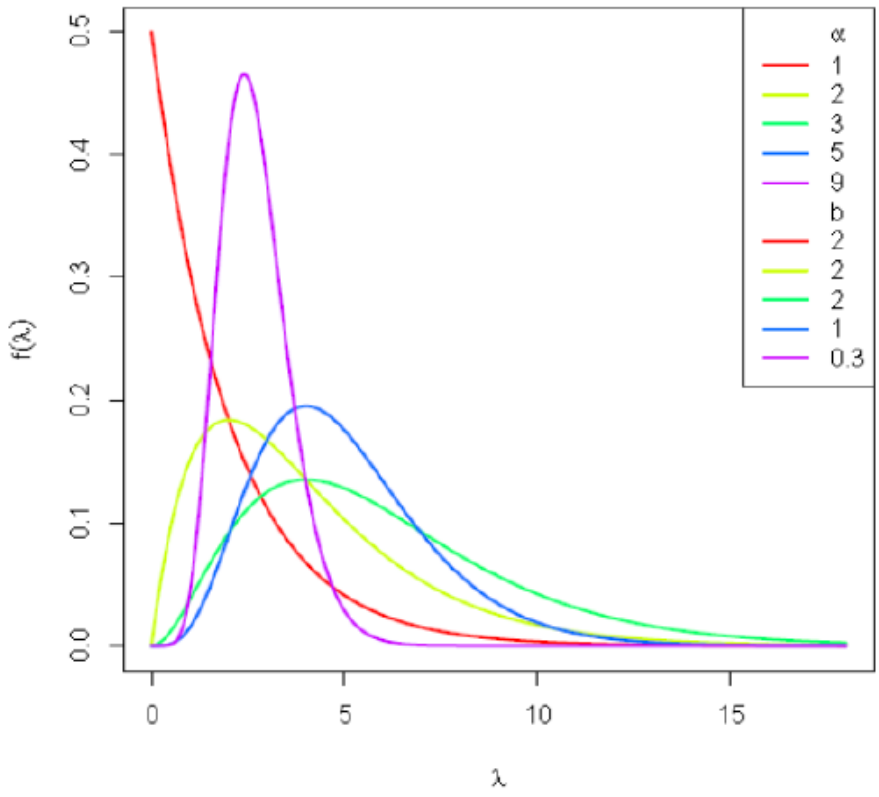
y posee la siguiente propiedad que se utilizará en los programas realizados para evaluar la función factorial en números naturales,

$$\Gamma\left(n+1\right) = n\Gamma\left(n\right) = n!.$$

Una reparametrización conveniente (porque simetriza la función de verosimilitud para estimar los parámetros) se obtiene al tomar $\beta = \frac{1}{b}$, con $b > 0$; de tal forma que α es el parámetro de forma y b el parámetro de escala en la familia de distribuciones Gama. Cuando el parámetro α es mayor que uno, la densidad es unimodal; en contraste, cuando $\alpha \leq 1$, la densidad es decreciente y de forma convexa.

El valor esperado y la varianza de Λ tienen las siguientes expresiones

$$\begin{aligned} E(\Lambda) &= \frac{\alpha}{\beta}, \\ Var(\Lambda) &= \frac{\alpha}{\beta^2}. \end{aligned}$$



(2.2.14)

Figura 2.2.14: Densidad Gama para diferentes valores de los parámetros de forma y escala. Esta gráfica ilustra la variedad de formas que puede tomar la función de densidad de una distribución Gama.

En la Figura 2.2.14 se observan algunas densidades Gama asociadas a pares de parámetros diferentes. Se puede observar la variedad de formas de las densidades, propiedad que hace a tal familia muy flexible. Esta característica apoya la afirmación de Fisher de que la distribución Gama es adecuada para modelar la heterogeneidad de las medias en los conteos

Poisson; sobre todo Fisher hace notar que las formas observadas comúnmente son aquellas donde el parámetro de forma α es menor que uno. En esos casos las densidades asociadas son decrecientes y de forma convexa, favoreciendo una mayoría de especies raras con pocos individuos presentes en la región y muy pocas especies abundantes.

Por el contrario, cuando el parámetro de forma $\alpha > 1$, la densidad es unimodal y los parámetros de intensidad Poisson de las especies distintas se concentran alrededor de esa moda. En esos casos habrá una proporción reducida de especies raras y la mayoría de las especies serán fáciles de detectar.

Densidad Lognormal de parámetros (μ, σ) :

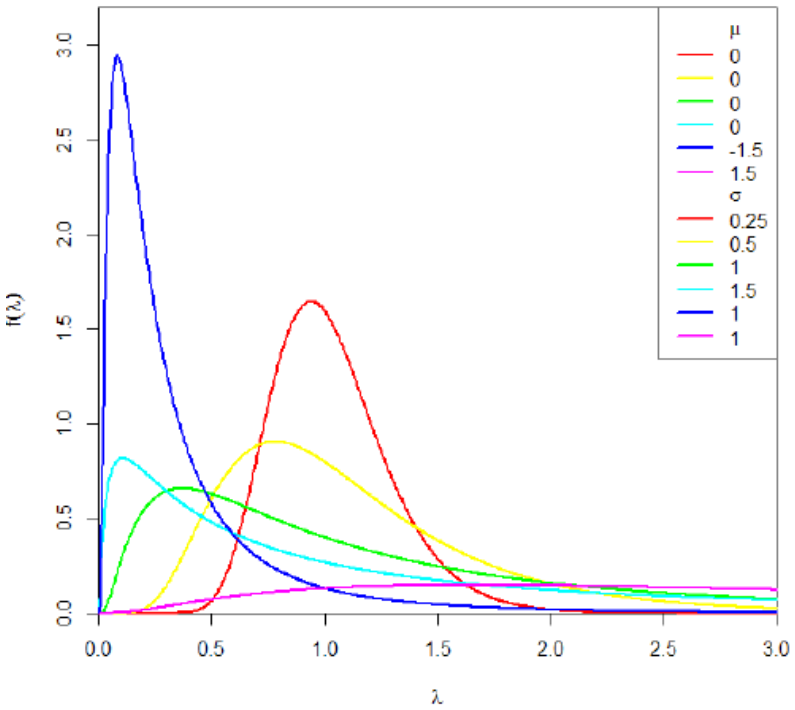
Supóngase que la variable aleatoria Λ tiene una distribución Lognormal de parámetros (μ, σ). La función de densidad de Λ es

$$f(\lambda; \mu, \sigma) = \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right] \mathbf{1}_{[0,\infty)}(\lambda), \sigma > 0, \mu \in \mathbb{R}. \tag{2.2.15}$$

En este caso, el valor esperado y la varianza de Λ tienen las siguientes expresiones

$$\begin{aligned} E(\Lambda) &= \exp\left(\frac{1}{2}\sigma^2 + \mu\right), \\ Var(\Lambda) &= \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]. \end{aligned}$$

La Figura 2.2.16 muestra una gráfica donde cada curva representa la densidad Lognormal asociada a un par de parámetros diferentes. A diferencia de las densidades Gama donde para ciertos valores del parámetro de forma α, la curva es cóncava, en las densidades Lognormales siempre hay una moda.

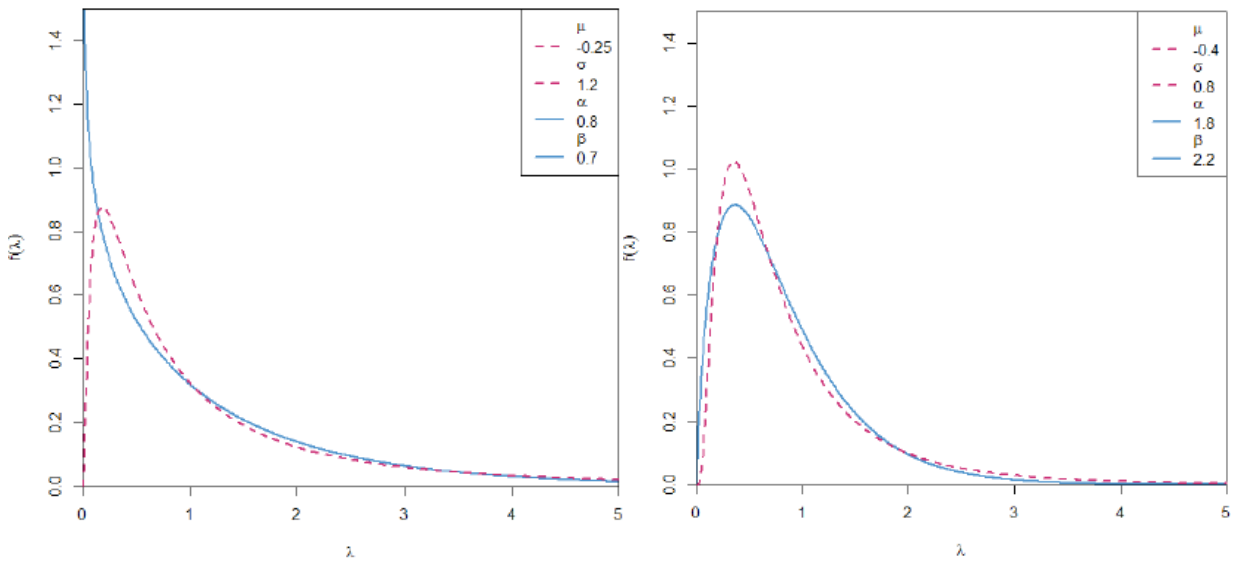


(2.2.16)

Figura 2.2.16: Esta gráfica ilustra las formas de la función de densidad Log-normal para diferentes valores de los parámetros μ y σ, siempre hay una moda pero la curva puede ser asimétrica o casi simétrica alrededor de esta moda.

La distribución Lognormal puede ser muy parecida a una distribución Gama, incluso aunque el parámetro de forma de esta última sea menor a la unidad. En la Figura 2.2.17 se muestran dos casos de distribuciones Gama y Lognormal cuyas densidades son muy parecidas.

En esta tesis se considerarán y compararán ambos modelos. Sin embargo, en las Conclusiones generales se recomendará usar al modelo Gama y se darán las razones explícitas de ello.



(2.2.17)

Figura 2.2.17**:** Comparación de densidades Gama y Lognormal para dos pares distintos de parámetros. En la figura de la izquierda, la densidad Gama es convexa, aún así existe una densidad Lognormal similar a ella. A la derecha se muestra una densidad Gama cuyo parámetro de forma es mayor que uno junto con una densidad Lognormal similar.

2.2.4 Estimación de intervalos para los parámetros de intensidad Poisson

Considérense ahora la muestra de variables $T_{r1}, ..., T_{rm}$, donde $T_{rj} = X_j/r$, para $j = 1, ..., m$. Para cada una de estas variables, la distribución condicional de $T_{rj} \mid \Lambda = \lambda_j$, dado el valor del parámetro de intensidad Poisson λ_j correspondiente, es una distribución discreta que se obtiene de la distribución Poisson de la variable X_j. Recuérdese que $X_j = \sum_{i=1}^{r} X_{ij}$ es la suma de variables aleatorias independientes Poisson con parámetro de intensidad λ_j, por lo

que el parámetro de intensidad de X_j es $r\lambda_j$. Por tanto, el valor esperado condicional de T_{rj} es $E\left(T_{rj} \mid \Lambda = \lambda_j\right) = \lambda_j$.

Los valores observados $t_{1r}, ..., t_{rm}$ de las variables aleatorias $T_{rj} = X_j/r$, para $j = 1, ..., m$, son estadísticas suficientes minimales para los parámetros de intensidad Poisson λ_j. Por ello, el intervalo de verosimilitud de nivel c depende de ellos y se define como

$$\begin{aligned} IV\left(c\right) &= \left\{\lambda : \left(\sum_{i=1}^{r} X_{ij}\right)\left[\log\lambda - \log T_{rj}\right] - r\left(\lambda - T_{rj}\right) \geq \log c\right\} \\ &= \left[\lambda_{j1}, \lambda_{j2}\right], \end{aligned} \tag{2.2.18}$$

donde $X_j = \sum_{i=1}^{r} X_{ij}$.

El intervalo de verosimilitud $[\lambda_{j1}, \lambda_{j2}]$ es aleatorio (ya que cambia para muestras distintas) y cuando el nivel de verosimilitud es $c = 0.1465$, la probabilidad de que incluya al verdadero valor del parámetro de intensidad Poisson es aproximadamente 0.95. Nótese que se deben usar métodos numéricos para obtener los extremos del intervalo, en el Apéndice A se dan más detalles sobre ello.

2.2.5 Probabilidad conjunta de $T_{1r}, ..., T_{mr}$

Considérese nuevamente la muestra de variables aleatorias independientes, $T_{r1}, ..., t_{rm}$, que se definieron en la sección anterior y que representan los promedios de conteos de las m especies observadas en r cuadrantes. La distribución condicional de cada una de ellas, dado el parámetro Poisson correspondiente λ_j, para $j = 1, ..., m$, es discreta. Por tanto, en la muestra puede haber valores repetidos con probabilidades no despreciables. En particular esto va a ocurrir, por ejemplo, para las especies raras de las que se observó solamente un individuo en alguno de los cuadrantes.

Parecería que estos valores repetidos están contradiciendo la afirmación que se hará más adelante de que $T_{r1}, ..., T_{rm}$ es un conjunto de variables aleatorias continuas, independientes e idénticamente distribuidas (de manera no condicional) como Gama o Lognormal. Recuérdese que para una variable continua la probabilidad de tomar un valor particular es

cero, entonces la probabilidad de ver valores repetidos en una muestra de dichas variables continuas independientes también es cero.

En muchos contextos en los que se consideran variables aleatorias continuas, cuando se observan valores repetidos en los datos observados, se puede explicar esta situación debido a la imprecisión que se tiene al medirlos. Todo instrumento de medición tiene una precisión finita y por ello induce una censura por intervalo en la observación. Es decir, cuando se dice que una variable aleatoria continua $X = x_0$, en realidad se está afirmando que cayó en un intervalo determinado por la precisión $2h$ del instrumento de medición,

$$x_0 - h \leq X < x_0 + h.$$

Así, la probabilidad de ver el valor registrado x_0 no es cero, sino que la probabilidad de que la variable X caiga en el intervalo mencionado es positiva,

$$P\left[x_0 - h < X \leq x_0 + h\right] = F_X\left(x_0 + h\right) - F_X\left(x_0 - h\right) > 0,$$

donde F_X es la función de distribución continua de X.

Esta misma idea se aplicará para la muestra de variables contínuas $T_{r1}, ..., T_{rm}$. Cuando se diga que $T_{rj} = T_{rj}$, se considerará que la variable T_{rj} cayó en el intervalo de verosimilitud $[\lambda_{j1}, \lambda_{j2}]$ de nivel $c = 0.1465$, para $j = 1, ..., m$. Estos intervalos son los de estimación para el parámetro de intensidad Poisson que se presentaron en la Sección 2.2.4. Esta convención es razonable, en vista de la Proposición que se establecerá en una sección más adelante.

Con esta consideración se resuelve la aparente contradicción de usar al valor observado T_{rj} de manera dual, primero como el resultado de una variable aleatoria discreta (cuando se considera la distribución condicional discreta de $T_{rj} \mid \Lambda = \lambda_j$), o como el resultado de una variable continua (cuando se considera la distribución no condicional Gama o Lognormal de T_{rj}).

Por lo anterior, la probabilidad conjunta (no condicional) de la muestra $T_{r1}, ..., T_{rm}$, toma

la siguiente expresión

$$\begin{aligned} v\left(T_{r1},...,t_{rm};\theta\right) &= \prod_{j=1}^{m}\mathbf{P}\left[T_{rj}\in\left[\lambda_{j1},\lambda_{j2}\right]\right] \\ &= \prod_{j=1}^{m}\left[G\left(\lambda_{j2};\theta\right)-G\left(\lambda_{j1};\theta\right)\right], \end{aligned} \tag{2.2.19}$$

donde $G\left(\lambda;\theta\right)$ es la distribución truncada Gama o Lognormal que se describió en la Sección 2.2.3.

2.2.6 Dos resultados importantes

Hay dos distribuciones relevantes asociadas a las variables $T_{r1},...,T_{rm}$, las cuales se usarán para plantear el modelo estadístico propuesto para estimar el total de especies detectables k en una región de interés.

La primera es la distribución condicional discreta de T_{rj}, condicionada en el parámetro Poisson λ_j que se describió en la Sección 2.2.4.

Debido a que las variables T_{rj} son promedios de variables aleatorias independientes Poisson, $X_{1j},...,X_{rj}$,

$$T_{rj}=\frac{1}{r}\sum_{i=1}^{r}X_{ij},$$

al aplicar la Ley Fuerte de los Grandes Números, se sabe que al aumentar el número de cuadrantes r, la distribución condicional de T_{rj} dada la media correspondiente λ_j, converge fuertemente al valor esperado λ_j,

$$T_{rj}|\Lambda=\lambda_j \underset{r\to\infty}{\overset{c.p.1}{\to}} \lambda_j, \tag{2.2.20}$$

para $j=1,...,m$. Este es el primer resultado importante a considerar.

La segunda distribución relevante de las variables independientes e idénticamente distribuidas $T_{r1},...,T_{rm}$, es una distribución (no condicional) continua truncada Gama o Lognormal. Fisher (1943) había afirmado que los parámetros de intensidad Poisson seguían

una distribución Gama; sin embargo, él nunca mencionó a las variables $T_{r1}, ..., T_{rm}$. La justificación de la afirmación que se da en la tesis para la distribución no condicional de las $T_{r1}, ..., T_{rm}$ se sustenta en la siguiente Proposición, cuya demostración se da en el Apéndice B. La proposición misma es el segundo resultado importante en esta sección.

Proposición 1 *Sean Λ una variable aleatoria continua que toma valores positivos y X_{rj} variables aleatorias, $r, j \in \{1, .2, ...\}$.*

Supongamos que para cada $j \in \{1, .2, ...\}$ y para cualquier $r \in \{1, .2, ...\}$, condicionalmente a $\Lambda = \lambda_j$, las v.a. $X_{1j}, ..., X_{rj}$ son independientes e idénticamente distribuidas como Poisson con media λ_j, $\mathbf{E}[X_{1j}|\Lambda = \lambda_j] = \lambda_j$, donde $\lambda_j \in (0, \infty)$.

Definamos

$$T_{rj} := \frac{1}{r} \sum_{i=1}^{r} X_{ij}, r, j \in \{1, .2, ...\}.$$

Entonces, de forma no condicional y para cualquier $j \in \{1, .2, ...\}$, T_{rj} converge en distribución a Λ cuando $r \to \infty$, esto es,

$$T_{rj} \underset{r \to \infty}{\overset{d}{\to}} \Lambda.$$

Este resultado es crucial para el planteamiento del modelo estadístico que se propondrá en la siguiente sección para estimar el parámetro de interés k. Aunque el resultado es asintótico, a través de simulaciones que se realizaron y no se presentan aquí por brevedad, se vio que la distribución de los T_{rj} llega a ser muy cercana a la distribución G considerada, incluso para valores pequeños de r (por ejemplo, $r = 7$) bajo ciertas condiciones, las cuales se cumplen para los ejemplos que se darán en el Capítulo 4.

2.2.7 Distribución del número de especies no observadas

Sea Z la variable aleatoria que representa el número de especies detectables de un total k, que no fueron observadas en ninguno de los r cuadrantes. De estas especies no se vio individuo

alguno en los r cuadrantes. Se tiene que $Z = k - M$, donde M es la variable aleatoria que representa el número de especies detectables que sí fueron observadas en la muestra de r cuadrantes.

Es importante notar que Z no es observable puesto que depende del parámetro desconocido k. La variable aleatoria que sí se observa es M. Resulta razonable suponer que para las especies que no fueron observadas en los cuadrantes, su parámetro de intensidad Poisson haya sido muy pequeño, porque para tal valor, la probabilidad de ver cero individuos en todos los cuadrantes es más grande que para parámetros Poisson mayores.

A manera de ilustrar esta idea, considérense dos parámetros de intensidad Poisson asociados a dos especies: $\lambda_1 = 1/r$ y $\lambda_2 > \lambda_1$. Consecuentemente, las probabilidades de ver cero individuos en los r cuadrantes para individuos de estas dos especies guardan la siguiente relación de orden,

$$\left\{\mathbf{P}\left[X_1 = 0; \lambda_1\right]^r = \exp\left(-r\lambda_1\right)\right\} > \left\{\mathbf{P}\left[X_2 = 0; \lambda_1\right]^r = \exp\left(-r\lambda_2\right)\right\}.$$

Nótese que es más probable ver cero individuos de la especie con parámetro Poisson menor, que con la que tiene uno mayor.

Por esta razón, se supondrá que los parámetros Poisson de las especies no vistas en los r cuadrantes son pequeños y se encuentran en el intervalo $[\lambda_0, \lambda_1]$, donde

$$\lambda_1 = \frac{1}{r}.$$

El valor de λ_1 se eligió con base en que este es el valor que toma la variable T_{rj} de una especie j cuando solamente se observa un individuo en alguno de los cuadrantes y ninguno en los restantes. Es razonable suponer que en ese caso el parámetro Poisson asociado estaría cercano a λ_1. Por tanto es razonable tomar este valor como cota superior para los parámetros Poisson de las $k - m$ especies no observadas en los cuadrantes.

La constante λ_0, definida en 2.2.9, siempre debe ser menor que la cota superior λ_1. Para que esta condición se cumpla, una vez que se elige el número de cuadrantes (generalmente por razones presupuestales) se debe encontrar el valor de p en **??** lo más grande posible que cumpla esta condición y evaluar si es razonable en cuanto al contexto del problema. En

general esto no representa un problema, porque usualmente el número de cuadrantes no es muy grande.

Bajo el supuesto de que las k especies detectables tienen parámetros de intensidad Poisson que siguen una distribución truncada en λ_0 como se estableció en la Sección 2.2.3, y que los parámetros de intensidad Poisson de las especies no observadas en la muestra pertenecen al intervalo $[\lambda_0, \lambda_1]$, es razonable considerar que Z es una variable aleatoria con distribución Binomial de parámetros desconocidos k y q^*. Es decir, se tienen k ensayos Bernoulli independientes donde el "éxito" es que la especie presente cero individuos en todos los cuadrantes. Se propone aquí que la probabilidad de que ocurra un tal éxito está dada por la proporción de especies que tienen un parámetro Poisson en el intervalo $[\lambda_0, \lambda_1]$, bajo la hiperdensidad truncada supuesta (que se supondrá es Gama o Lognormal en esta tesis),

$$\begin{aligned} q^* &= q^*(\theta) = \int_{\lambda_0}^{\lambda_1} g(\lambda;\theta)\, d\lambda \\ &= \frac{F(\lambda_1;\theta) - F(\lambda_0;\theta)}{1 - F(\lambda_0;\theta)}. \end{aligned} \tag{2.2.21}$$

En resumen, la función de probabilidad de Z, el número de especies no observadas en los cuadrantes, condicionando en que fueron observadas $M = m$ especies en r cuadrantes, es una binomial con parámetros desconocidos k y θ. Dicha probabilidad tiene la siguiente expresión,

$$\mathbf{P}\left[Z = k - m | T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}; k, \theta\right] = \binom{k}{k-m} \left[q^*(\theta)\right]^{k-m} \left[1 - q^*(\theta)\right]^m, \tag{2.2.22}$$

para $m = 0, 1, ..., k$.

2.3 Modelos estadísticos para estimar el número de especies detectables

Los datos que se obtienen del muestreo por cuadrantes de tamaño r se resumen en una observación, $(t_{r1}, ..., t_{rm}, m)$, del vector aleatorio $(T_{r1}, ..., T_{rm}, M)$ (definido en la Sección

2.1). Las primeras m componentes corresponden al promedio de los conteos observados de individuos de cada una de las especies observadas en los r cuadrantes, T_{rj}, para $j = 1, ..., m$. La última componente del vector contiene al número de especies distintas observadas en la muestra, m. De hecho, nótese que la dimensión del vector observado dependerá precisamente de esta última componente observada.

El vector de datos observados $(t_{r1}, ..., t_{rm}, m)$ contiene información importante para estimar tanto el parámetro de interés k, como los parámetros θ de la distribución truncada $G(\lambda; \theta)$ para las intensidades Poisson (dada en 2.2.11). Por ello la funcion de verosimilitud que se usará para estimar los parámetros k y θ es proporcional a la probabilidad conjunta del vector aleatorio mencionado,

$$\mathbf{P}\left[M = m, T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}\right]. \tag{2.3.1}$$

Para encontrar una expresión para esta probabilidad conjunta, nótese que la variable aleatoria Z del número de especies detectables que no se vieron en r cuadrantes, toma el valor $Z = k - m$ sí y sólo sí $M = m$, para $m = 0, 1, ..., k$. Por tanto, la probabilidad conjunta (2.3.1) es igual a la densidad conjunta del vector aleatorio siguiente, evaluado en los valores que se indican,

$$\mathbf{P}\left[Z = k - m, T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}\right]. \tag{2.3.2}$$

Esta última probabilidad conjunta resulta ser igual al producto de dos probabilidades que ya se han calculado en secciones anteriores de este capítulo: a) la probabilidad condicional $\mathbf{P}\left[Z = k - m | T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}; k, \theta\right]$, dada en 2.2.22, y b) la probabilidad conjunta $v(t_{r1}, ..., t_{rm}; \theta)$ dada en 2.2.19. Este hecho resulta evidente de la definición de la probabilidad condicional de $Z = k - m$ dado $\{T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}\}$,

$$\mathbf{P}\left[Z = k - m | T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}; k, \theta\right] = \frac{\mathbf{P}\left[Z = k - m, T_{r1} = t_{r1}, ..., T_{rm} = t_{rm}; k, \theta\right]}{v(t_{r1}, ..., t_{rm}; \theta)}. \tag{2.3.3}$$

Dado que las probabilidades 2.3.1 y 2.3.2 son iguales, reemplazando la última por la primera en 2.3.3, y despejando esta primera, se obtiene una expresión para la probabilidad deseada,

$$\begin{aligned} & \mathbf{P}\left[M=m, T_{r1}=t_{r1}, ..., T_{rm}=t_{rm}; k, \theta\right] && (2.3.4) \\ = \; & \mathbf{P}\left[Z=k-m | T_{r1}=t_{r1}, ..., T_{rm}=t_{rm}; k, \theta\right] \cdot v\left(t_{r1}, ..., t_{rm}; \theta\right). && (2.3.5) \end{aligned}$$

Con esto, la verosimilitud de los parámetros desconocidos k y θ se define como proporcional a esta última probabilidad,

$$\begin{aligned} & \mathcal{L}\left(k, \theta; t_{r1}, ..., t_{rm}, m\right) \propto \mathbf{P}\left[Z=k-m | T_{r1}=t_{r1}, ..., T_{rm}=t_{rm}; k, \theta\right] \cdot v\left(t_{r1}, ..., t_{rm}; \theta\right) && (2.3.6) \\ & = \left\{ \binom{k}{k-m} \left[q^*\left(\theta\right)\right]^{k-m} \left[1-q^*\left(\theta\right)\right]^m \right\} \left\{ \prod_{j=1}^{m} \left[G\left(\lambda_{j2}; \theta\right) - G\left(\lambda_{j1}; \theta\right)\right] \right\} I_{\{0,1,...,k\}}\left(m\right), \end{aligned}$$

donde λ_{j1} y λ_{j2} son todas mayores o iguales que λ_0, para $j = 1, ..., m$. Nótese que la verosimilitud vale cero si $k < m$ puesto que k debe ser necesariamente mayor o igual al número m de especies observadas en la muestra.

Para encontrar los EMV de k y de θ, se debe maximizar esta función. Para facilitar este proceso, se puede considerar que el parámetro k es continuo y tal que está en el intervalo (m, ∞). Para maximizar la verosimilitud, conviene usar métodos numéricos de búsqueda, como el de Nelder-Mead (1965), y éstos requieren que se den valores iniciales para los parámetros. Como valor inicial para k se recomienda dar $m+1$; para θ conviene dar en el caso de la distribución Gama los estimadores de momentos de α y de β, que están dados por

$$\begin{aligned} \widetilde{\beta} &= \frac{\frac{1}{m}\sum_{i=1}^{m} t_{ir}}{\frac{1}{m}\sum_{i=1}^{m} t_{ir}^2 - \left(\frac{1}{m}\sum_{i=1}^{m} t_{ir}\right)^2}, \\ \widetilde{\alpha} &= \widetilde{\beta}\left(\frac{1}{m}\sum_{i=1}^{m} t_{ir}\right). \end{aligned}$$

Para el caso Lognormal, dar como valores iniciales para μ y σ los siguientes valores,

$$\begin{aligned} \widetilde{\mu} &= \frac{1}{m}\sum_{i=1}^{m} t_{ir}, \\ \widetilde{\sigma} &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(t_{ir}-\mu_0)^2}. \end{aligned}$$

Se calcularán y graficarán las verosimilitudes perfiles de cada uno de los parámetros desconocidos, para dar una estimación puntual y por intervalo de los parámetros k y θ. Finalmente, se sugiere encontrar el nivel adecuado de verosimilitud que esté asociado a un nivel de confianza deseado, por ejemplo del 95% de confianza, a la manera que se describió en la Sección 1.3.4. Estas ideas se aplicarán a datos de conteos de reptiles en el Capítulo 4 para ilustrarlas claramente.

2.4 Comparación de la definición tradicional en Ecología de *detectabilidad de una especie* con la que se da en esta tesis para *especie detectable*

Los biólogos Boulinier et al. (1998) definen el concepto de *detectabilidad* de la siguiente manera:

> "La detectabilidad de una especie se define como la probabilidad de detectar al menos un individuo de esa especie en una unidad de muestreo (o esfuerzo de colecta), dado que hay individuos de esa especie presentes en el área de interés durante la recolección de la muestra".

En esta tesis se define una *especie detectable* como una especie para la cual la probabilidad de que haya al menos un individuo en toda la región de interés es alta. En cuanto a la presencia de individuos de una especie de interés en la región considerada, Boulinier et al

exigen que haya dos o más individuos presentes con certeza en la región. En contraste, en esta tesis se pide una condición más débil sobre la presencia de individuos de una especie: solamente se pide que haya al menos un individuo de la especie con probabilidad alta (mayor que cierta p, donde $0.9 \leq p \leq 1$) en toda la región.

Por otra parte, en este trabajo solamente se define y se usa el concepto de *especie detectable*, más no el de *detectabilidad*. El concepto de especie detectable sirvió en esta tesis solamente para definir, sin ambigüedades, el parámetro de interés k que representa el total de *especies detectables* en la región homogénea de interés.

Capítulo 3

Simulaciones

En este capítulo se muestra un resumen de los resultados obtenidos mediante simulaciones de muestras repetidas similares a las que se analizarán en el Capítulo 4.

Primero se discribe cómo elegir niveles de verosimilitud tales que el intervalo de verosimilitud obtenido de cortar a la verosimilitud perfil relativa de un parámetro en ese nivel calibrado, tenga asociada una probabilidad de cobertura aproximada de 95%.

En la segunda sección, se muestra el resultado de calcular razones de verosimilitudes del modelo estadístico presentado en el Capítulo 2, considerando los modelos Gama y Lognormal como posibles distribuciones de las intensidades Poisson. Se mostrarán la proporción de veces que el cociente de verosimilitudes favorece al modelo del que realmente provienen los datos y la proporción de veces en que las muestras provienen de cierto modelo pero aún así el otro modelo es razonable para explicarlas.

Finalmente, se realizaron simulaciones de muestras bajo el modelo Gama para analizar el efecto del aumento del número de cuadrantes de la muestra. Se espera que a mayor número de cuadrantes en la muestra, mejor sea la cobertura de los intervalos de verosimilitud de un nivel dado de verosimilitud.

3.1 Calibración de porcentajes de cobertura de intervalos de verosimilitud para los parámetros del modelo

Considérese la verosimilitud planteada en el Capítulo 2 (dada en la ecuación 2.3.4). Aquí se analizará la cobertura de los intervalos de verosimilitud que se obtienen a partir de las funciones de verosimilitud perfil relativas para cada uno de los parámetros del modelo (k, θ), en muestras simuladas repetidas y bajo un modelo.

Se comprobó mediante simulaciones que la cobertura de los intervalos de nivel $c = 0.1465$ en realidad era mucho menor a 95%, cobertura que debieran poseer bajo condiciones regulares para muestras grandes (ver Sección 1.3.3). Como el número de cuadrantes no suele ser muy grande, conviene calibrar los intervalos de verosimilitud para obtener la cobertura deseada (ver Sec. 1.3.4). Es decir, mediante simulaciones se tuvo que determinar el nivel $c' \in (0, 1)$ adecuado para que el intervalo tuviera asociada una confianza del 95%.

El procedimiento mediante el cual se obtuvieron los intervalos calibrados se describe con detalle en la Sección 1.3.4. Para ello, primero se determinaron los valores de los parámetros del modelo pues para realizar simulaciones se necesita tomar todos estos parámetros como conocidos para así generar muestras aleatorias Los parámetros con los que se simularon las muestras fueron los parámetros estimados bajo los dos modelos considerados, primero el Gama y luego el Lognormal, con los datos de reptiles presentados en el Capítulo 4. Una vez generadas las muestras aleatorias, se supone que los valores de (k, θ) son desconocidos y se desean estimar a partir de las muestras aleatorias simuladas.

3.1.1 Modelo Gama-Binomial

Los juegos de parámetros elegidos para ilustrar los resultados de este capítulo están inspirados en los parámetros estimados para los juegos de datos presentados en el Capítulo 4. A continuación se especifican los valores de los parámetros que forman parte del escenario

general de simulaciones, es decir, aquellos parámetros que estarán fijos para generar todas las muestras simuladas. Se especifican también tres juegos de parámetros particulares, asociados a la distribución de las intensidades Poisson, para comparar los resultados de las coberturas de intervalos bajo estos tres escenarios distintos.

Escenario general	
k	35
p	0.99
λ_0	0.00194
λ_1	0.143
N_{sim}	1000

(3.1.1)

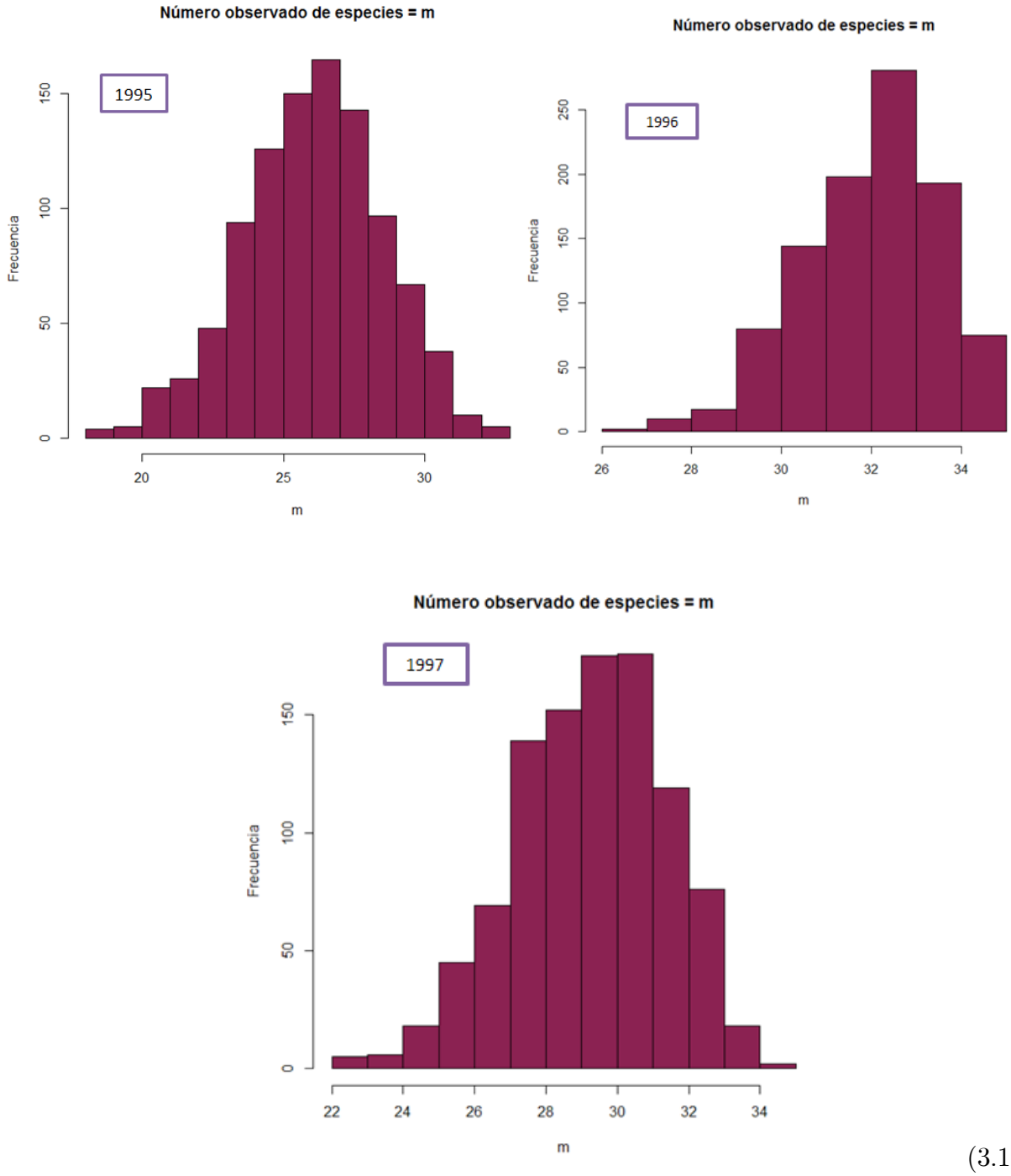
Escenario 1 (1995)		Escenario 2 (1996)		Escenario 3 (1997)	
α_1	0.33	α_2	1.4	α_3	0.86
β_1	0.2	β_2	1.27	β_3	0.65
r_1	7	r_2	7	r_3	5

(3.1.2)

Tablas 3.1.1 **y** 3.1.2**:** Parámetros comunes de las muestras simuladas bajo el modelo Gama (arriba). Tamaño de muestra y parámetros de la distribución Gama considerados para simular diferentes conjuntos de muestras repetidas (abajo).

El número de especies detectables se ha fijado en $k = 35$. Dados los parámetros (α, β) y el número de cuadrantes en una muestra, se procedió a simular $N_{sim} = 1000$ muestras aleatorias, esto es, N_{sim} tablas de conteos como la que se muestra en 2.1.1. Estas muestras corresponden a conteos Poisson que siguen una distribución Gama truncada en λ_0.

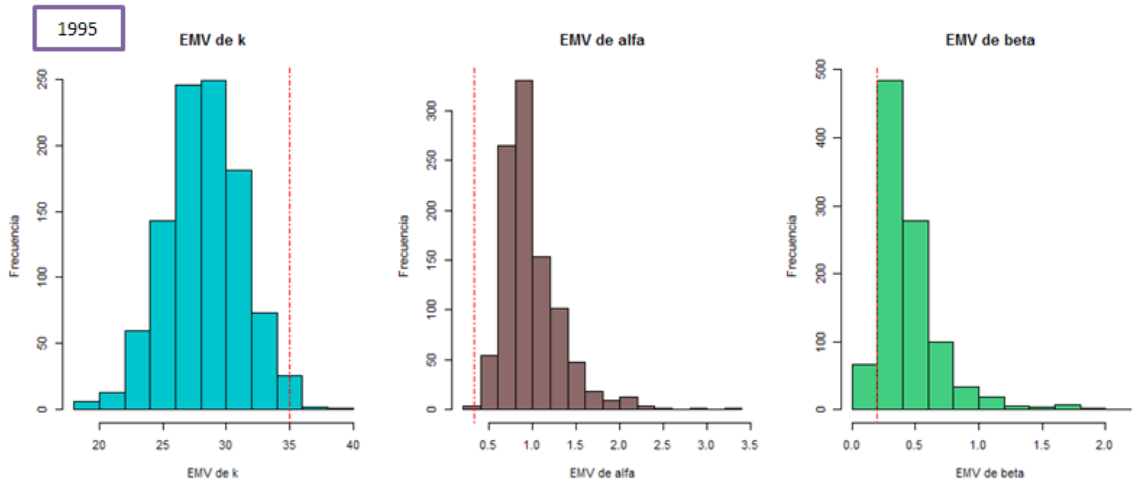
Los siguientes histogramas muestran el número de especies detectadas m, que se observaron en las muestras simuladas.



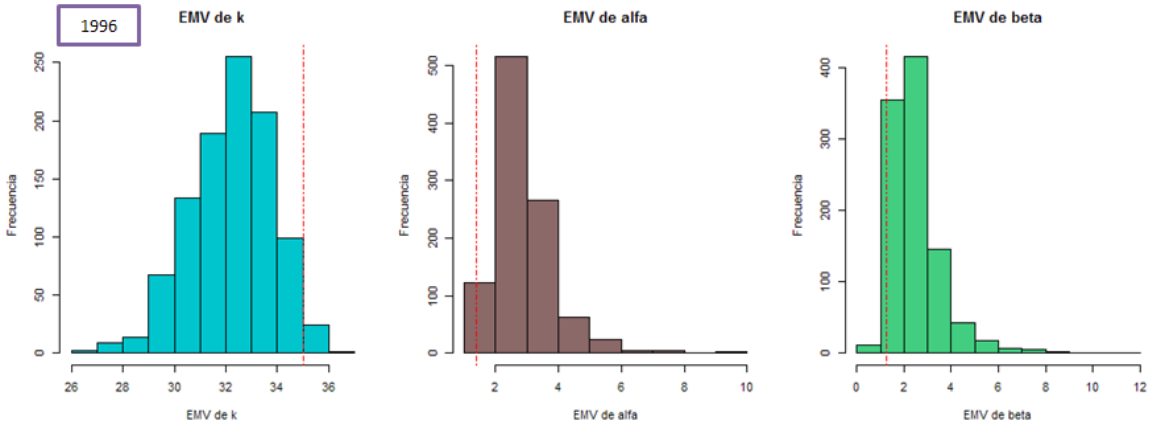
(3.1.3)

Figura 3.1.3**:** Histogramas de los valores de m obtenidos en las simulaciones correspondientes a los escenarios $1, 2, 3$, bajo el modelo Gama. El rango de valores observados para m en los tres escenarios es entre 18 y 35 especies.

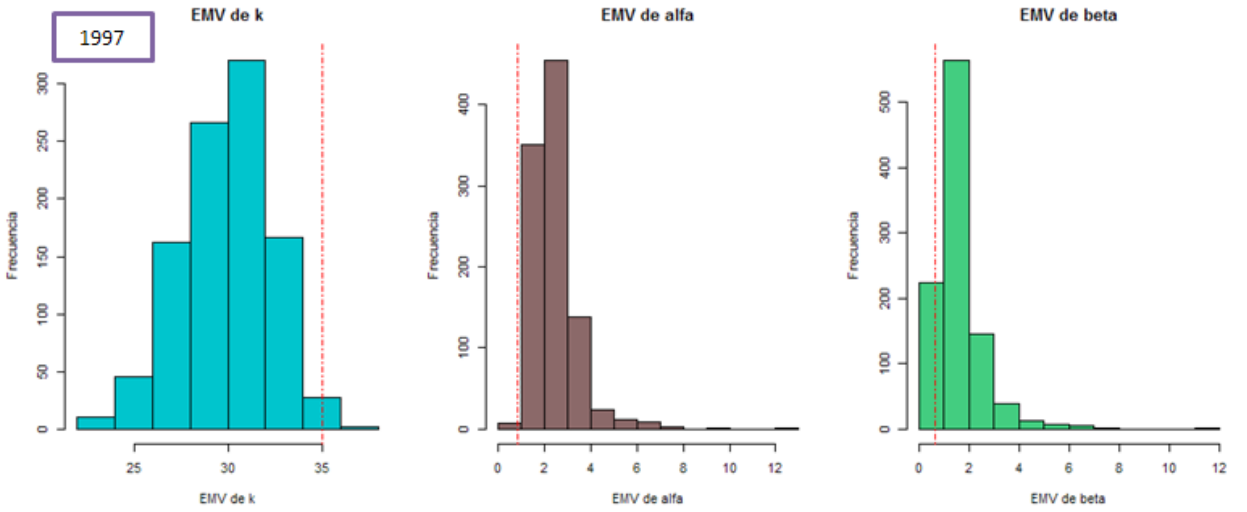
Una vez generadas las muestras aleatorias, se obtuvieron $N_{sim} = 1000$ tablas de conteos y se procedió a estimar los parámetros del modelo Gama-Binomial. Todos los valores de los $EMV's$ obtenidos están concentrados en los histogramas que se muestran a continuación. En cada uno de ellos, se indica con una línea vertical roja el valor verdadero del parámetro (aquel que se usó para generar la muestra). Esta línea sirve como referencia evaluar visualmente las proporciones de los estimadores de las muestras simuladas que subestiman o sobrestiman los parámetros teóricos.



(3.1.4)



(3.1.5)



(3.1.6)

Figuras 3.1.4, 3.1.5 y 3.1.6: Histogramas de los valores de los EMV obtenidos para las muestras simuladas para los tres modelos estadísticos asociados a los escenarios $1, 2, 3$ y la distribución Lognormal para las intensidades Poisson. Las líneas verticales punteadas indican el verdadero valor del parámetro en cada caso.

Una vez obtenidos los $EMV's$ para cada una de las muestras, se obtuvo el valor de la verosimilitud perfil relativa de cada uno de los parámetros evaluada en los verdaderos valores de los parámetros. Con esto se consegue obtiene la verosimilitud relativa evaluada en los parámetros de los que provienen las muestras. Nótese que si este valor resulta mayor que $c = 0.1465$, esto quiere decir que el verdadero valor del parámetro está contenido en el intervalo de verosimilitud de nivel $c = 0.1465$. Entonces, al contar cuántos valores de la verosimilitud perfil evaluada en el verdadero valor del parámetro, son mayores que este valor de c, se obtiene el número de intervalos de nivel c que contienen al verdadero parámetro. Si se divide esta cantidad por el número total de simulaciones N_{sim}, se obtiene la proporción de veces que un intervalo de nivel c cubre al verdadero valor, es decir, el nivel de confianza del intervalo obtenido a partir de estas simulaciones.

Los porcentajes de cobertura asociados al nivel $c = 0.1465$ de verosimilitud que se obtuvieron para cada parámetro y cada escenario se muestran en la siguiente tabla:

Porcentajes de cobertura	k	α	β
Escenario 1	49.2	12.3	64.1
Escenario 2	54.7	44.7	63.5
Escenario 3	42.2	22.1	51.5

(3.1.7)

Tabla 3.1.7**:** Porcentajes de cobertura de los intervalos de verosimilitud de nivel $c = 0.1465$ para muestras simuladas bajo los escenarios $1, 2, 3$. Se observa que todos son menores a los esperados para muestras grandes, por tanto, se recomienda calibrar los niveles de verosimilitud.

Como se puede observar, los porcentajes de cobertura son mucho menores a los esperados para muestras grandes. Los resultados asintóticos indican que para muestras grandes el nivel $c = 0.1465$ de verosimilitud suele estar asociado a un porcentaje de cobertura aproximado de 95%. Se recomienda calibrar dichos niveles para alcanzar la cobertura deseada, para el tamaño de muestras observadas, esto se logra usando las simulaciones anteriores. Dados los valores de cada verosimilitud perfil evaluada en el verdadero valor del parámetro, se busca el cuantil de probabilidad 0.05 de entre todos estos valores, esto es, el valor de c tal que solo 5% de las simulaciones tenga asociado un valor de la verosimilitud perfil menor al nivel c. Siguiendo este procedimiento se obtienen los siguientes niveles de verosimilitud calibrados, mismo que serán utilizados en el Capítulo 4 para dar un intervalo de estimación para los parámetros del modelo.

Para corroborar que los niveles de verosimilitud calibrados tienen una confianza aproximada de 95%, se realizó una réplica del experimento, esto es, se simularon N_{sim} muestras nuevas a las que se les calculó el intervalo de verosimilitud del nivel calibrado correspondiente (mostrado en la Tabla 3.1.8) y se calculó la proporción de intervalos que contenían el ver-

dadero valor de los parámetros. Comprobando así que el porcentaje de cobertura obtenido con las muestras nuevas era aproximado a 95%.

Niveles de verosimilitud calibrados	k	α	β
Escenario 1	0.00265	0.00015	0.00363
Escenario 2	0.00682	0.00441	0.00898
Escenario 3	0.00190	0.00034	0.00223

(3.1.8)

Tabla 3.1.8**:** Niveles de verosimlitud calibrados mediante muestras simuladas asociadas a intensidades Poisson con distribución Gama bajo los escenarios $1, 2, 3$; para obtener un porcentaje de cobertura aproximado de 95%.

3.1.2 Modelo Lognormal-Binomial

En esta sección se realizará un análisis análogo al de las simulaciones presentadas en la sección anterior pero considerando que el modelo Gama es el que se elige como distribución de las intensidades Poisson. A continuación se especifican los valores de los parámetros que forman parte del escenario general de simulaciones, además de tres juegos de parámetros particulares para la distribución Lognormal, para comparar los resultados de las coberturas de intervalos bajo estos tres escenarios distintos.

Escenario general	
k	35
p	0.99
λ_0	0.00194
λ_1	0.143
N_{sim}	1000

(3.1.9)

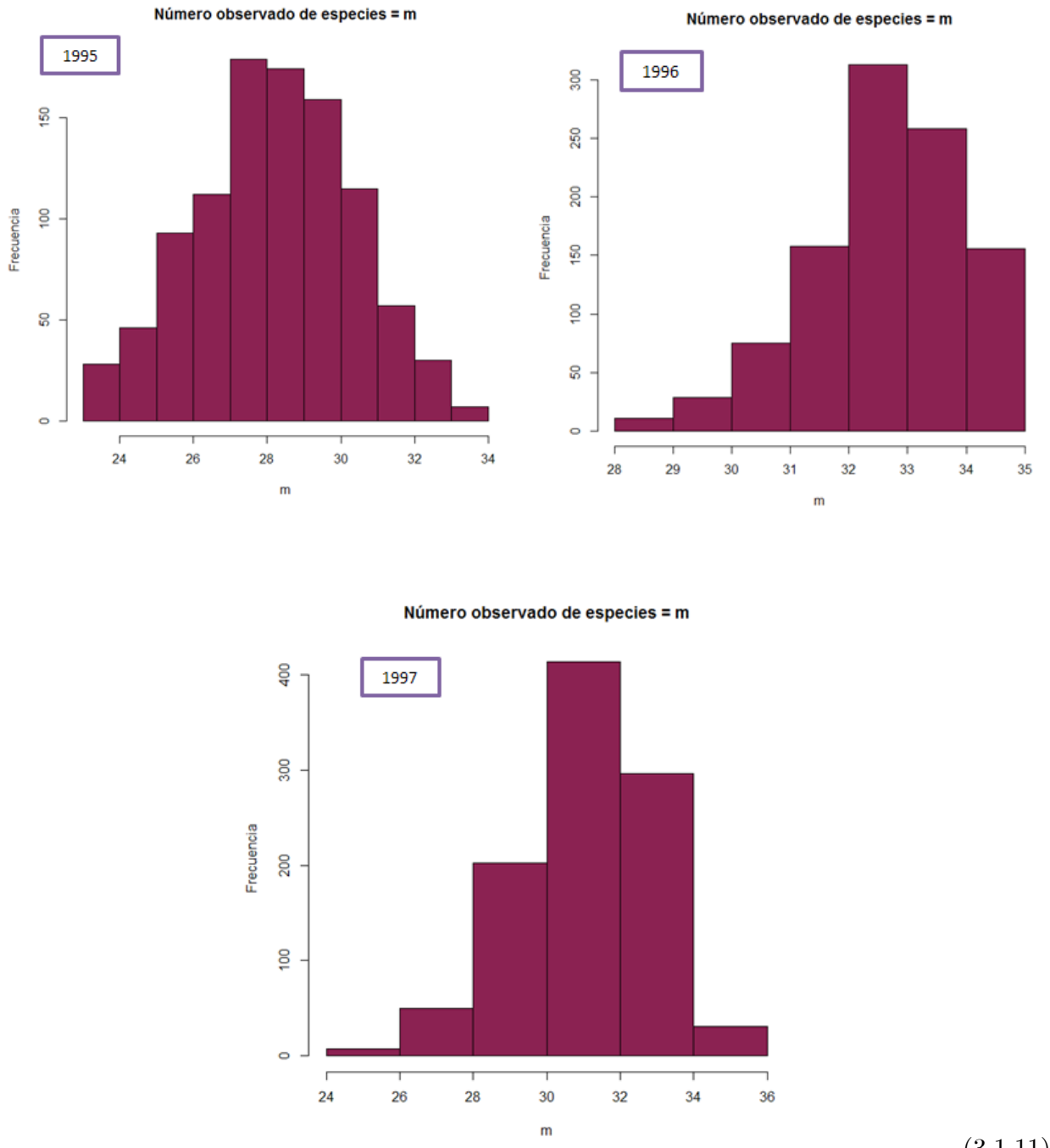
Escenario 1 (1995)		Escenario 2 (1996)		Escenario 3 (1997)	
μ_1	-0.7	μ_2	-0.3	μ_3	-0.22
σ_1	1.55	σ_2	0.86	σ_3	1
r_1	7	r_2	7	r_3	5

(3.1.10)

Tablas 3.1.9 **y** 3.1.10**:** Parámetros comunes de las muestras simuladas bajo el modelo Gama (arriba). Tamaño de muestra y parámetros de la distribución Gama considerados para simular diferentes conjuntos de muestras repetidas (abajo).

Al igual que en el caso Gama, el número de especies detectables se ha fijado en $k = 35$ y se simularon $N_{sim} = 1000$ muestras aleatorias similares pero en este caso de la distribución Lognormal, para cada uno de los escenarios. Así, estas muestras corresponden a conteos Poisson que siguen una distribución Lognormal truncada en λ_0 y tales que a la izquierda de $\lambda_1 = 1/r$ se encuentran las intensidades Poisson de las especies no observadas.

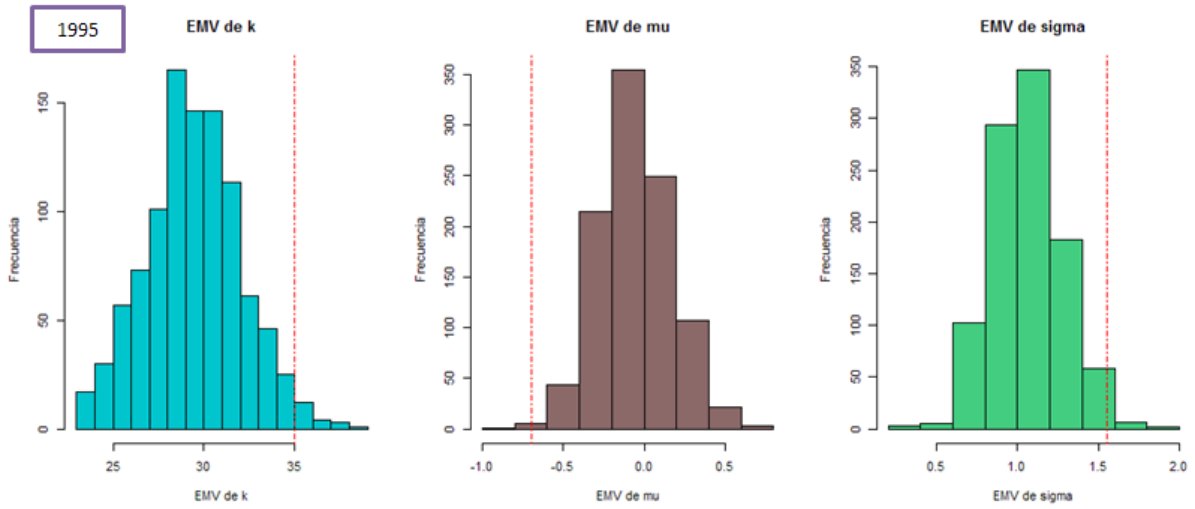
Los siguientes histogramas muestran el número de especies detectadas m, que se observaron en las muestras simuladas.



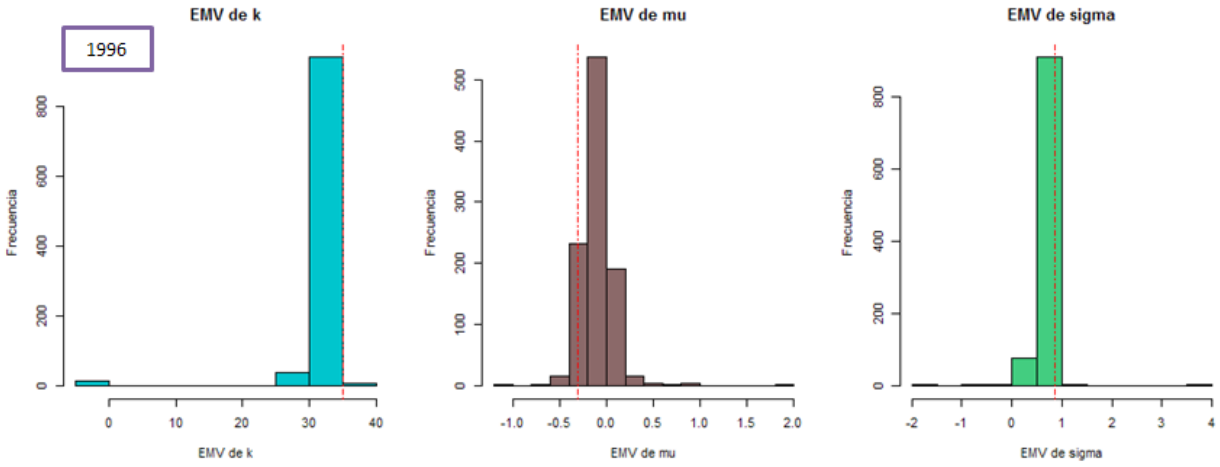
(3.1.11)

Figura 3.1.11**:** Histogramas de los valores de m obtenidos en las simulaciones correspondientes a los escenarios $1, 2, 3$, bajo el modelo Lognormal. El rango de valores observados en los tres casos es entre 22 y 35 especies.

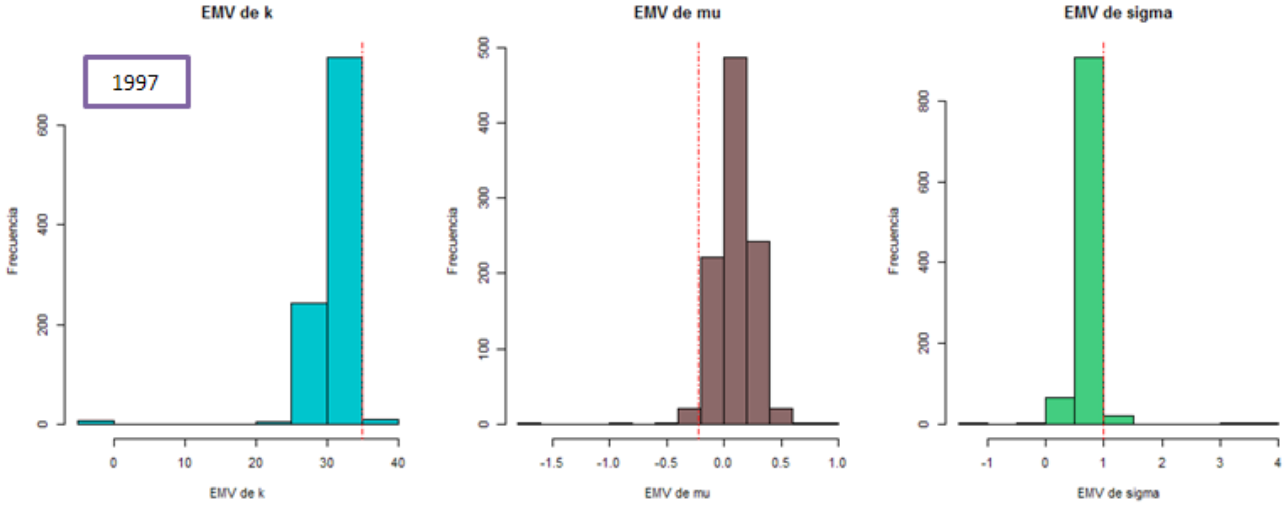
Una vez generadas las muestras aleatorias, se obtuvieron $N_{sim} = 1000$ tablas de conteos y se procedió a estimar los parámetros del modelo Lognormal-Binomial. Todos los valores de los $EMV's$ obtenidos están concentrados en los histogramas que se muestran a continuación. En cada uno de ellos, se indica con una línea vertical roja el valor verdadero del parámetro (aquel que se usó para generar la muestra). Esta línea sirve como referencia para determinar si los parámetros son subestimados o sobrestimados.



(3.1.12)



(3.1.13)



(3.1.14)

Figuras 3.1.12, 3.1.13 y 3.1.14: Histogramas de los valores de los EMV obtenidos para las muestras simuladas para los tres modelos estadísticos asociados a los escenarios $1, 2, 3$ y la distribución Lognormal para las intensidades Poisson. Las líneas verticales punteadas indican el verdadero valor del parámetro en cada caso.

Una vez obtenidos los $EMV's$ para cada una de las muestras, se obtuvo el valor de la verosimilitud perfil relativa de cada uno de los parámetros evaluada en los verdaderos valores de los parámetros, para así obtener la proporción de veces que un intervalo de nivel c cubre al verdadero valor, es decir, el nivel de confianza del intervalo obtenido a partir de estas simulaciones.

Los porcentajes de cobertura asociados al nivel $c = 0.1465$ de verosimilitud que se obtuvieron para cada parámetro y cada escenario se muestran en la siguiente tabla:

Porcentajes de cobertura	k	μ	σ
Escenario 1	48.1	25.6	44.6
Escenario 2	44.1	70.5	67.5
Escenario 3	39.3	31.1	48.5

(3.1.15)

Tabla 3.1.15**:** Porcentajes de cobertura de los intervalos de verosimilitud de nivel $c = 0.1465$ para muestras simuladas bajo los escenarios $1, 2, 3$. Se observa que todos son menores a los esperados para muestras grandes, por tanto, se recomienda calibrar los niveles de verosimilitud.

También en este caso, los porcentajes de cobertura son mucho menores a los esperados para muestras grandes. Los niveles de verosimilitud calibrados para muestras Lognormales, mismos que serán utilizados en el Capítulo 4 para dar un intervalo de estimación para los parámetros del modelo, se muestran en la siguiente tabla.

Niveles de verosimilitud calibrados	k	α	β
Escenario 1	0.00268	0.00077	0.00094
Escenario 2	0.00140	0.02199	0.01132
Escenario 3	0.00203	0.00444	0.00314

(3.1.16)

Tabla 3.1.16**:** Niveles de verosimlitud calibrados mediante muestras simuladas asociadas a intensidades Poisson con distribución Gama bajo los escenarios $1, 2, 3$; para obtener un porcentaje de cobertura aproximado de 95%.

Para corroborar que los niveles de verosimilitud calibrados tienen una confianza aproximada de 95%, se realizó una réplica del experimento, esto es, se simularon N_{sim} muestras nuevas a las que se les calculó el intervalo de verosimilitud del nivel calibrado correspondiente (mostrado en la Tabla 3.1.16) y se calculó la proporción de intervalos que contenían el verdadero valor de los parámetros. Comprobando así que el porcentaje de cobertura obtenido con las muestras nuevas era aproximado a 95%.

3.2 Contraste de modelos Gama y Lognormal

En esta sección se presenta una forma de contrastar el ajuste de los modelos estadísticos para estimar k, obtenidos de considerar los modelos Gama y Lognormal para las intensidades Poisson, al vector de obseravciones $(t_{1r}, ..., t_{mr}, m)$. Para ello, se realizaron simulaciones de muestras similares bajo el Escenario 1, tanto para el caso Gama como para el Lognormal (ver Tablas 3.1.2 y 3.1.10).

El primer ejercicio de simulación consistió en los siguientes pasos:

1. Generar muestras repetidas considerando que las intensidades Poisson de las especies siguen un modelo Gama cuyos parámetros están dados por el Escenario 1.

2. Para cada muestra simulada:

   (a) Se calularon los EMV de los parámetros (k, θ) del modelo estadístico presentado en el Capítulo 2; primero considerando la distribución Gama misma y luego considerando la distribución Lognormal.

   (b) Se evaluó la verosimilitud relativa correspondiente a cada modelo en los valores verdaderos de los parámetros (k, θ).

   (c) Se calculó el cociente de verosimilitudes obtenido de dividir el valor asociado a la distribución Gama por el asociado a la Lognormal.

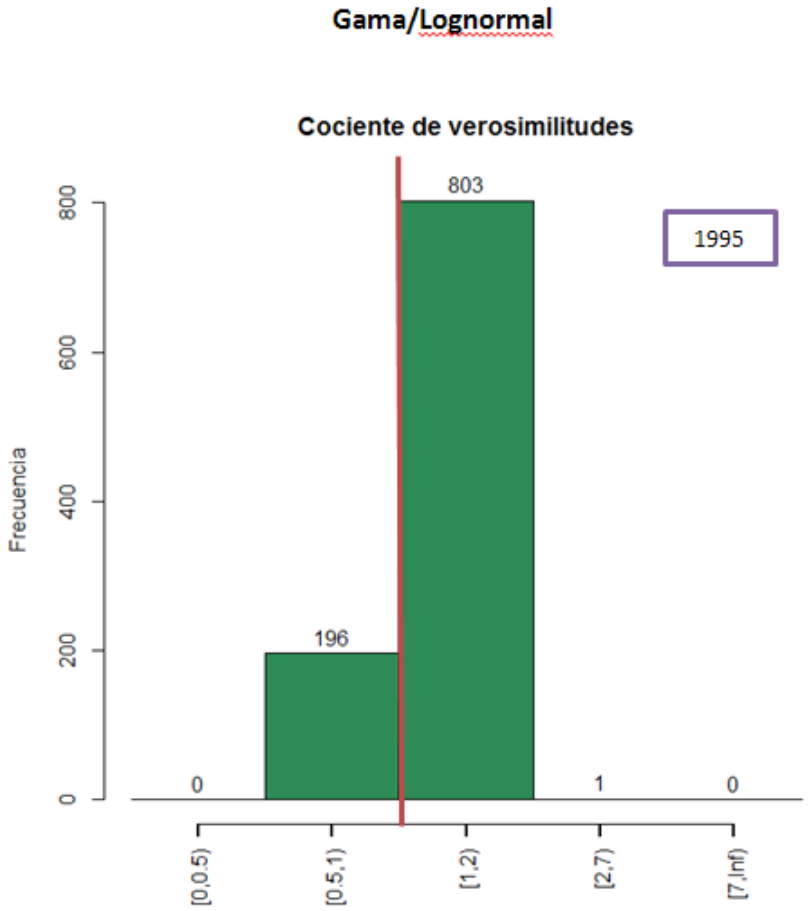
3. Se agruparon los valores de los cocientes de verosimilitud en las clases $[0, 0.5)$, $[0.5, 1)$, $[1, 2)$, $[2, 7)$, $[7, \infty)$.

El segundo grupo de simulaciones se obtuvo de manera análoga, tomando el modelo Lognormal en lugar del Gama en los tres pasos anteriores.

Si los modelos Gama y Lognormal dan lugar a modelos estadísticos equivalestes o indistinguibles para estimar k, entonces la mayoría de los valores de las razones de verosimlitudes calculadas estarán en los intervalos $[0.5, 1)$ y $[1, 2)$, es decir, son valores pequeños y por tanto indican que los modelos contrastados son indistinguibles.

Simulando del modelo Gama

Para el caso de simular bajo el Escenario 1 y la distribución Gama, la mayoría de las razones de verosimilitudes están en el intervalo $[1, 2)$, nuevamente indicando que ambos modelos son equivalentes para las muestras simuladas. Aunque realmente dado el valor obtenido para el resto de las muestras, es razonable concluir que los modelos son indistinguibles.

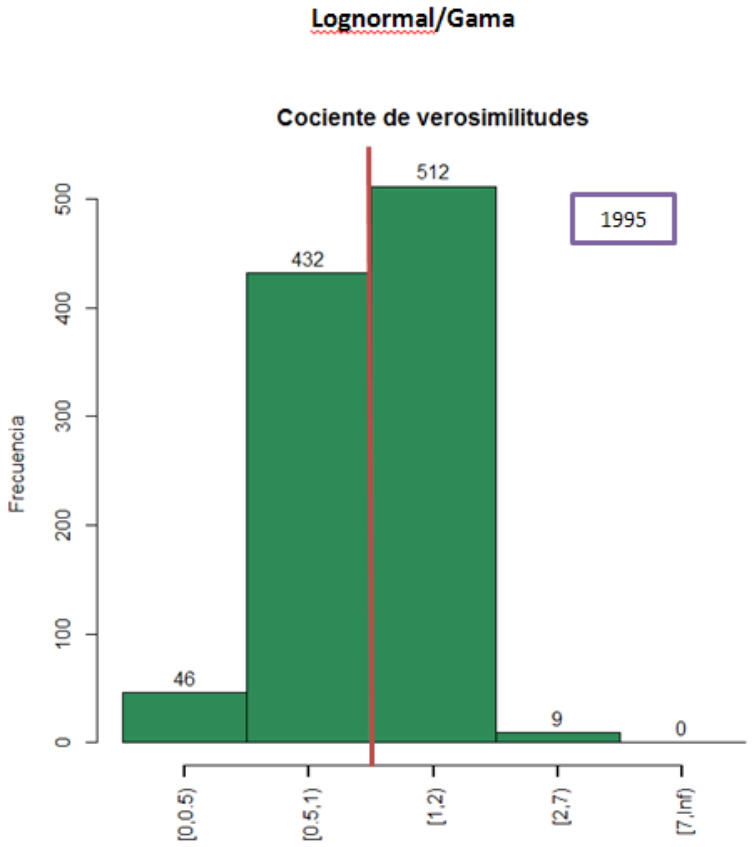


(3.2.1)

Figura 3.2.1**:** Cocientes de verosimilitudes calculados para muestras asociadas al modelo Gama para las intensidades Poisson, donde el denominador corresponde a evaluar el modelo estadístico del Capítulo 2 bajo la distribución Lognormal, y el numerador bajo la distribución Gama misma.

Simulando del modelo Lognormal

Para el caso de simular bajo el Escenario 1 y la distribución Lognormal, hay una proporción similar de las razones de verosimilitudes que están en el intervalo $[0, 1)$ y el intervalo $[1, 2)$. Si bien en la mayoría de los casos se prefiere el modelo Gama sobre el Lognormal, el valor obtenido para el resto de las muestras sugiere que los modelos son indistinguibles.



(3.2.2)

Figura 3.2.2**:** Cocientes de verosimilitudes calculados para muestras asociadas al modelo Lognormal para las intensidades Poisson, donde el denominador corresponde a evaluar el modelo estadístico del Capítulo 2 bajo la distribución Gama, y el numerador bajo la distribución Lognormal misma.

A diferencia de la comparación de modelos anterior, aquí se observa que hay más casos de confusión entre un modelo y otro, esto es, a pesar de que en este caso las muestras están asociadas a una distribución Lognormal, hay muchos casos donde se prefiere el modelo Gama. Este fenómeno también se presenta en el caso anterior, sin embargo, aquí ocurre con mayor frecuencia.

3.3 Efecto del aumento de cuadrantes en la muestra sobre el porcentaje de cobertura de intervalos de verosimilitud

Con el objetivo de analizar el efecto de aumentar el número de cuadrantes donde se lleva a cabo la colecta de reptiles en Chamela, se realizaron simulaciones de muestras repetidas donde en lugar de considerar muestras de $r = 7$ cuadrantes, se tomara una muestra sobre $r = 25$ cuadrantes. Para ello se mostrarán a continuación resultados de muestras simuladas bajo el modelo Gama y el Escenario 1 (ver Tabla 3.1.2), donde los conteos de individuos se supone fueron realizados sobre $r = 25$ cuadrantes. Para ambos tamaños de muestra se considera un nivel de verosimilitud $c = 0.1465$.

Porcentajes de cobertura	k	α	β
$r = 7$	49.2	12.3	64.1
$r = 25$	79.1	50.3	78.4

(3.3.1)

Tabla 3.3.1**:** Efecto del aumento del número de cuadrantes r sobre la cobertura de intervalos de verosimilitud de nivel $c = 0.1465$, para un caso particular donde se considera el modelo Gama para las intensidades Poisson.

Es claro que al aumentar el número de cuadrantes en la muestra, aumenta también el porcentaje de cobertura de los intervalos de nivel $c = 0.1465$. Esto se cumple para cada uno de

los parámetros del modelo estadístico k, α, β. Se mostrará en la Sección 4.1.1 que una muestra de $r = 7$ cuadrantes, para el caso de los datos presentados en ese capítulo, la proporcion de área muestreada es muy pequeña, de alrededor del 0.3% ($(\delta_1 = 0.00295) \times 100 \approx 0.3\%$). Por lo que se recomienda aumentar el número de cuadrantes muestreados a $r = 25$ cuadrantes, pues con este número de unidades de muestreo se consigue obtener una muestra del 1% del área total de la región de interés.

Por brevedad solo se muestra un ejemplo donde se observa el efecto del aumento de cuadrantes en la muestra, sin embargo, durante la elaboración de esta tesis se realizaron más ejemplos de estas simulaciones, obteniendo la misma conclusión sobre el efecto de aumentar el número de cuadrantes para mejorar el porcentaje de cobertura asociada los intervalos de verosimilitud: a mayor número de cuadrantes mayor porcentaje de cobertura asociado a un intervalo de verosimilitud.

La velocidad a la cual aumenta esta probabilidad de cobertura usualmente, en el caso Gama, depende del valor del parámetro de forma de esta distribución α. Cuando éste es mayor a uno, el porcentaje aumenta en mayor proporción conforme aumenta el número de cuadrantes. En contraste, para valores pequeños de α, se necesitan tamaños de muestra muy grandes para obtener porcentajes de cobertura altos.

Capítulo 4

Aplicaciones

En este capítulo, se aplicarán las ideas y resultados de los capítulos anteriores a datos de reptiles observados en la Estación de Biología Chamela en la época de lluvias de tres años distintos con respecto a la presencia del fenómeno meteorológico de la Oscilación del Sur, El Niño. Para cada año, se desea estimar el número de especies detectables de reptiles que habitan la región cercana al arroyo.

Se han elegido tres grupos de datos de años distintos de los cuales dos se clasificaron de acuerdo a la presencia de los fenómenos meteorológicos El Niño y La Niña que forman parte del ciclo de la Oscilación del Sur, y el tercero como año normal.

Se estimará el modelo estadístico propuesto en el Capítulo 2, suponiendo que los parámetros Poisson siguen una distribución Gama o Lognormal. A manera de comparación, se verán las consecuencias de adoptar cada modelo sobre las inferencias del parámetro k para cada uno de los años de interés.

4.1 Reptiles de Chamela

Los ejemplos presentados en este trabajo se obtuvieron de la base de datos sobre reptiles y anfibios de la Estación de Biología Chamela (EBCh). La EBCh es una dependencia del Instituto de Biología de la Universidad Nacional Autónoma de México (UNAM) que desde su fundación en 1971, se ha dedicado a la investigación de campo y el apoyo a la enseñanza y divulgación de la Biología; esfuerzos que han permitido conocer una parte importante de la diversidad biológica existente en la región y estudiar el funcionamiento de los ecosistemas que la habitan.

La base de datos original, de donde se obtuvieron los ejemplos que se presentan en este capítulo, fue proporcionada por el Dr. Andrés García Aguayo, investigador del Departamento de Biología Evolutiva de la UNAM, quien estudia los reptiles y anfibios que habitan en la EBCh.

(4.1.1)

Figura 4.1.1: Ubicación geográfica dentro de la República Mexicana de la Estación de Biología Chamela, del Instituto de Biología de la UNAM.

La EBCh está ubicada en el Estado de Jalisco, México (ver Figura 4.1.1) a 2 km al oeste de la costa y 6 km al sureste de la Bahía de Chamela. La reserva cuenta con un total de 3319 Ha de terreno y su topografía presenta lomas bajas y pequeñas cañadas, que confluyen en arroyos donde no hay corrientes de agua permanentes. El clima de Chamela es tropical, cálido subhúmedo, con una marcada estacionalidad. La sequía se presenta de noviembre a junio, algunas veces interrumpida por lluvias ligeras o fuertes en diciembre o enero; mientras que la estación de lluvias se presenta de julio a octubre. El promedio de días con lluvia apreciable es 52.

(4.1.2)

Figura 4.1.2: Fotografía del cambio de paisaje en un mismo sitio de Chamela entre la época de lluvias y la de sequía en la región de selva baja caducifolia en la reserva de Chamela. En la época de lluvia se observa una vegetación espesa, mientras que en la época de sequía el paisaje es muy diferente, muy pocas plantas tienen hojas.

La vegetación dominante de la región es la selva baja caducifolia o bosque tropical caducifolio[1], alcanza una altura de alrededor de 10 m y la mayoría de las plantas se quedan sin hojas en la sequía. Esta característica permite observar paisajes muy diferentes entre las estaciones de lluvia y de sequía (ver Figura 4.1.2). La región cercana al arroyo sostiene una selva más alta con árboles de talla mayor. Hasta el momento se han registrado más de 1100 especies de plantas en la región.

La fauna de la región está representada principalmente por aves, aunque también abundan diferentes especies de reptiles, anfibios y felinos. Dentro de la reserva de Chamela se conocen bien los diferentes grupos de vertebrados y el total de las especies de cada grupo taxonómico registradas históricamente (ver Tabla 4.1.3). De los invertebrados sólo se conocen bien algunos grupos de insectos, como las abejas, de las cuales se han registrado alrededor de 230 especies.

VERTEBRADOS	
TAXA	No. de Spp.
ANFIBIOS	18
REPTILES	67
AVES	270
MAMIFEROS	70

(4.1.3)

Tabla 4.1.3: La reserva ecológica de Chamela está habitada principalmente por cuatro grupos de vertebrados. En la segunda columna se muestra el número total de especies registradas históricamente para cada grupo taxonómico.

[1] El adjetivo caducifolio hace referencia a los árboles o arbustos que pierden su follaje durante una parte del año, la cual coincide en la mayoría de los casos con la llegada de la época más fría, en los climas templados, o de la más seca, en los climas cálidos y áridos.

4.1.1 Características generales de los datos y del área de estudio

La base de datos original proporcionada por el Dr. Andrés García contiene información de conteos de reptiles y anfibios registrados entre los años 1995 y 2000. Tales conteos se llevaron a cabo siguiendo un diseño de muestreo por cuadrantes.

La información contenida en la base de datos es muy específica para cada individuo, sin embargo, la que fue útil para este estudio se menciona a continuación:

Estaciones. Los meses del año se clasifican en tres estaciones de acuerdo a la cantidad de lluvias registradas:

Estación	**Meses**
Lluvias	Junio, Julio, Agosto, Septiembre, Octubre
Transición	Noviembre, Diciembre, Enero, Febrero
Sequía	Marzo, Abril, Mayo,

Se cuenta con información adicional sobre el comportamiento de las lluvias en los diferentes años. En particular, cada año se clasifica según la presencia de los fenómenos climáticos del Ciclo de la Oscilación del Sur conocidos como El Niño[2] y La Niña[3], los cuales causan cambios notables en las temperaturas y en los regímenes de lluvias.de un lugar. La principal diferencia entre un año designado como El Niño o La niña es el periodo del año donde se acumulan más lluvias. Particularmente, en la EBCh, a los años tipo El Niño se les identifica

[2]El Niño es un fenómeno meteorológico cuyo nombre científico es Oscilación del Sur;,está caracterizado por la aparición de corrientes oceánicas cálidas en las costas de América y la alteración del sistema global océano-atmósfera que se origina en el Océano Pacífico Ecuatorial, generalmente durante un periodo comprendido entre diciembre y marzo. (http://elnino.cicese.mx/nino.htm)

[3]La Niña es un fenómeno meteorológico que forma parte del ciclo natural global del clima conocido como Oscilación del Sur El Niño. El Niño es conocido como el periodo seco y La Niña como el frío. En los años identificados con La Niña se ha observado que en el periodo seco (noviembre-abril) la precipitación es superior a la histórica.

como secos porque llueve mucho menos en la época de lluvias, mientras que los inviernos suelen ser muy lluviosos.

Se han seleccionado los datos de años en los que se presentan estos fenómenos meteorológicos y un año más donde no están presentes para estimar en cada uno de ellos el número de especies detectables bajo condiciones climáticas y geográficas particulares. Así, los datos que serán analizados en este capítulo corresponden a los años:

1995	La Niña
1996	Normal
1997	El Niño

Debido a que uno de los efectos principales de los fenómenos Niño y Niña es la variación en la cantidad de lluvias a lo largo del año, consideraremos solamente los conteos de especies de reptiles realizados durante las estaciones de lluvia, esto es, únicamente aquellas observaciones realizadas durante los meses de Julio, Agosto, Septiembre y Octubre, de cada año. Esto con el objetivo de analizar si existe algún efecto de estos fenómenos sobre el número de especies detectables. La cantidad de lluvia registrada en cada año se muestra en la siguiente tabla:

1995 (Niña)			**1996 (Normal)**			**1997 (Niño)**		
Meses	Días	Lluvia	Meses	Días	Lluvia	Meses	Días	Lluvia
Jul(7)	5	139.5	Jul(7)	5	79.5	Jul(7)	4	149.0
Ago(8)	5	263.0	Ago(8)	4	278.0	Ago(8)	4	33.5
Sept(9)	3	296.6	Sept(9)	3	7.0	Sept(9)	3	156.0
Oct(10)	5	0	Oct(10)	0	266	Oct(10)	3	177.5

(4.1.4)

Tabla 4.1.4**:** Lluvia total por mes en milímetros, registrada en los meses correspondientes a la estación de lluvia de los años $1995, 1996, 1997$.

Cuadrantes. Un cuadrante está representado por una trampa y su área de influencia, es decir, la superficie alrededor de la trampa en la que es muy probable que un individuo pueda ser capturado. Cada cuadrante muestreado a lo largo de un año está etiquetado con un número y cada individuo detectado en un cuadrante se etiquetó también con este número. Para cada uno de los meses del año se eligieron ciertos días en los que se revisaron las trampas para contabilizar a los individuos de cada especie, de tal forma que los datos considerados para cada año son conteos registrados en los cuatro meses de la época de lluvias (julio-octubre). Hay un total de 7 cuadrantes muestreados, sin embargo, en el año 1997 sólo se tienen conteos registrados en 5 de estos cuadrantes; no se sabe si no fueron visitados o si no hubo reptiles capturados en ellos durante este año.

En la Figura 4.1.8 se muestra un esquema que representa el sistema de captura de reptiles a través de trampas y el radio de influencia (ρ) de las mismas. Se considerará que $\rho = 15m$, por lo que el área de un cuadrante se calcula de manera aproximada como el área de un círculo de radio ρ, esto es, el área h de un cuadrante es

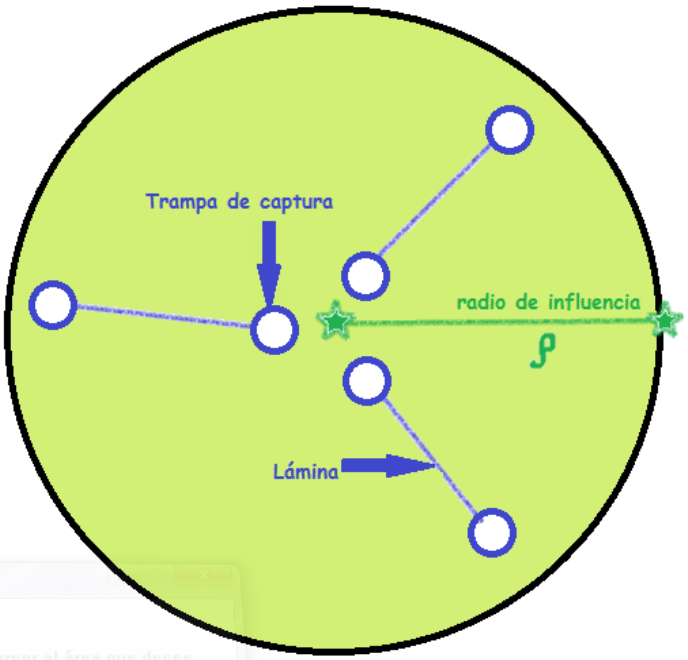
$$h = \pi\rho^2 = \pi\left(15m\right)^2 = 706.86m^2. \tag{4.1.5}$$

Esto quiere decir que para los años 1995 y 1996, la superficie muestreada s_1, está dada por

$$s_1 = h \cdot r = \left(706.86m^2\right) \cdot 7 = 4948.02m^2, \tag{4.1.6}$$

mientras que en el año 1997 se muestreó una superficie s_2 igual a

$$s_2 = h \cdot r = \left(706.86m^2\right) \cdot 5 = 3534.3m^2 \tag{4.1.7}$$



(4.1.8)

Figura 4.1.8**:** Sistema de trampas empleado para capturar y contar reptiles en Chamela. Consiste de 6 trampas unidas de par en par por una lámina, la cual sirve para atraer a los animales hacia las trampas. El radio de influencia ρ, determina el tamaño del cuadrante.

Grupo taxonómico. Los grupos taxonómicos de especies registrados en la base de datos original son dos, reptiles y anfibios. Trabajaremos solamente con los datos asociados a reptiles, pues según los especialistas no es conveniente considerarlos junto con los anfibios como un solo grupo taxonómico. Si se desea trabajar con los datos de anfibios el análisis se llevaría a cabo de manera análoga al que se presenta en este trabajo para los datos de reptiles.

Especie. Cada individuo observado (o capturado) ha sido identificado con su especie y se le asignó una etiqueta a cada una de las especies distintas observadas. En total, en los años bajo estudio $1995, 1996, 1997$ se observó un total de $m_1 = 16, m_2 = 11, m_3 = 13$ especies, respectivamente.

Vegetación. Los siete cuadrantes se eligieron dentro de la zona cercana al arroyo en Chamela. Een cuanto a vegetación, esta zona es claramente distinta a la región que conforma la selva baja y el resto del terreno de la EBCh. Uno de los supuestos especificados en la Sección 1.1, establece que la comunidad ecológica bajo estudio debe elegirse de tal forma que pueda considerarse cerrada y homogénea al momento del estudio. Atendiendo a este requisito, se estableció que el área de estudio está determinada por la zona cercana al arroyo, la cual corresponde a siete tramos de un kilómetro de largo que rodean al arroyo ($120m$ hacia cada lado del arroyo). Luego, el área de interés ocupa una superficie total de

$$\begin{aligned} A &= 7km \times 240m = 1680000m^2 \\ &= 1.68 \times 10^6 m^2. \end{aligned} \tag{4.1.9}$$

La razón del área total sobre el área de un cuadrante es

$$W = \frac{A}{h} = \frac{1.68 \times 10^6 m^2}{706.86m^2} \approx 2376. \tag{4.1.10}$$

En resumen, los datos analizados corresponderán a observaciones de reptiles registradas en sitios ubicados dentro de la vegetación cercana al arroyo, durante la estación de lluvias de tres años distintos, un año normal, uno donde se presentó el fenómeno Niño y otro donde se presentó el fenómeno Niña. Se determinó la superficie total de estudio A, así como el tamaño de un cuadrante h, y la razón de áreas W. Estos valores son constantes e iguales para los datos asociados a los tres años bajo estudio

$$\begin{aligned} A &= 1.68 \times 10^6 m^2, \\ h &= 706.86m^2, \\ W &\approx 2376. \end{aligned}$$

Otro valor común para los tres juegos de datos es el parámetro λ_0, la cota inferior de la distribución de las intensidades Poisson. Si se supone que las k especies detectables tienen

al menos un individuo presente en la región con probabilidad mayor o igual a $p = 0.99$, entonces

$$\lambda_0 = \frac{-\ln(1-p)}{W} = \frac{-\ln(1-0.99)}{2376} = 0.00194. \tag{4.1.11}$$

A partir de las cantidades, A, W y h, es posible determinar en cada caso cuál es la proporción de área muestreada δ. Es decir, de la superficie total A el área representada por la muestra está dada por rh, de tal forma que al dividir por A se obtiene la proporción de superficie cubierta por la muestra

$$\delta = \frac{rh}{A} = \frac{rh}{Wh} = \frac{r}{W}. \tag{4.1.12}$$

De aquí en adelante se usarán los subíndices $1, 2$ y 3 para referirse a los años $1995, 1996$ y 1997, respectivamente. Así, los valores de δ para los tres años fueron

$$\begin{aligned} \delta_1 &= \delta_2 = \frac{7}{2376} = 0.00295, \\ \delta_3 &= \frac{5}{2376} = 0.0021. \end{aligned}$$

En el año 1995, la proporción de área muestreada es menor debido a que sólo se consideraron cinco cuadrantes, en lugar de siete.

En las siguientes secciones, se muestran los juegos de datos y los modelos estadísticos propuestos para estimar del número de especies detectables. Para cada año, primero se determinará el vector de observaciones de interés $(t_{r1}, ..., t_{m_1 r}, m_1)$ y los intervalos de censura de los conteos promedio observados $t_{jr}, j = 1, ..., m$, bajo el modelo Poisson. Se presentarán gráficas Q-Q que permitan determinar si los modelos Gama y Lognormal son razonables para describir a los datos asociados a las intensidades Poisson.

A manera de comparación, se estimará el modelo estadístico propuesto suponiendo que los parámetros Poisson siguen una distribución Gama, pero también se considerará la distribución Lognormal. Finalmente, se verán las consecuencias de adoptar cada modelo en las inferencias sobre el parámetro k.

4.2 Arroyo 1995 (La Niña)

Parámetros fijos que definen el modelo estadístico:

W	r_1	δ_1	p	λ_0	λ_1
2376	7	0.00295	0.99	0.00194	0.143

1995 (La Niña)	Cuadrante							$m_1 = 16$
Reptiles	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Individuos por especie
Especie 1	1	35	3	45	1	10	0	95
Especie 2	1	13	5	6	9	2	4	40
Especie 3	0	0	4	6	3	1	4	18
Especie 4	0	1	2	12	1	0	0	16
Especie 5	1	3	0	0	0	0	0	4
Especie 6	2	1	0	0	0	0	0	3
Especie 7	0	1	1	0	1	0	0	3
Especie 8	0	0	1	0	1	0	0	2
Especie 9	0	1	0	0	0	0	0	1
Especie 10	0	0	1	0	0	0	0	1
Especie 11	1	0	0	0	0	0	0	1
Especie 12	0	1	0	0	0	0	0	1
Especie 13	0	0	0	0	0	1	0	1
Especie 14	0	1	0	0	0	0	0	1
Especie 15	0	0	0	0	0	1	0	1
Especie 16	1	0	0	0	0	0	0	1

(4.2.1)

Tabla 4.2.1: Conteos observados en una muestra de 7 cuadrantes durante los meses de julio a octubre de 1995. Se observaron en total $m_1 = 16$ especies distintas.

4.2.1 Intervalos de verosimilitud-confianza para las intensidades Poisson

De acuerdo a sus abundancias, las especies observadas en 1995 se pueden clasificar según el número de individuos observados en todos los cuadrantes, como se muestra en la Tabla 4.2.2. Están registrados los conteos promedio distintos t_{rl}, observados en los $r_1 = 7$ cuadrantes de la muestra, las frecuencias de cada promedio f_l, y los extremos del intervalo de censura asociados a los intervalos verosimilitud-confianza del 95% correspondientes a cada intensidad Poisson λ_l distinta de la muestra observada, $[\lambda_{l1}, \lambda_{l2}]$. Nótese que el subíndice l sólo toma valores en el conjunto $\{1, ..., s\}$ donde s es el número de conteos promedio distintos observados en la muestra, por tanto, s es el número de frecuencias distintas f_l. En este ejemplo $s = 8$.

1995 **(La Niña)**			
t_{lr}	f_l	λ_{l1}	λ_{l2}
13.57	1	11.02	16.48
5.71	1	4.12	7.67
2.57	1	1.55	3.94
2.28	1	1.34	3.59
0.57	1	0.18	1.32
0.42	2	0.106	1.11
0.28	1	0.047	0.88
0.14	8	0.008	0.63

(4.2.2)

Tabla 4.2.2**:** Hay $s = 8$ conteos promedio distintos t_{lr}, en la muestra de $r = 7$ cuadrantes del año 1995. En la tabla se muestran sus frecuencias e intervalos de censura.

Nótese que $m_1 = \sum_{l=1}^{s} f_l = 16$ y $\lambda_0 < \lambda_{l1}, l = 1, ..., s$, lo cual quiere decir que no hace falta truncar por la izquierda a los intervalos de censura pues todos contienen valores a la derecha de la cota inferior λ_0, de la distribución de intensidades Poisson.

Es importante notar que toda la información contenida en la Tabla 4.2.2 será utilizada para definir la función de verosimilitud de cada modelo estadístico.

4.2.2 Modelo Gama-Binomial

La función de verosimilitud para (k_1, θ_1) basada en la muestra $(t_{r1}, ..., t_{r_1 m_1}, m_1)$ se considerará como en el Capítulo 2 (en 2.3.4). La distribución truncada G_T y los parámetros $q^*(\theta_1)$ y $p^*(\theta_1)$ se calculan a partir de una distribución Gama.

Se procederá a implementar este modelo para estimar los parámetros k_1 y θ_1 primero bajo el modelo Gama, y en la siguiente sección bajo el modelo Lognormal. Se repetirá lo mismo para los años 1996 y 1997.

La primer columna de la Tabla 4.2.2 determina al vector de observaciones

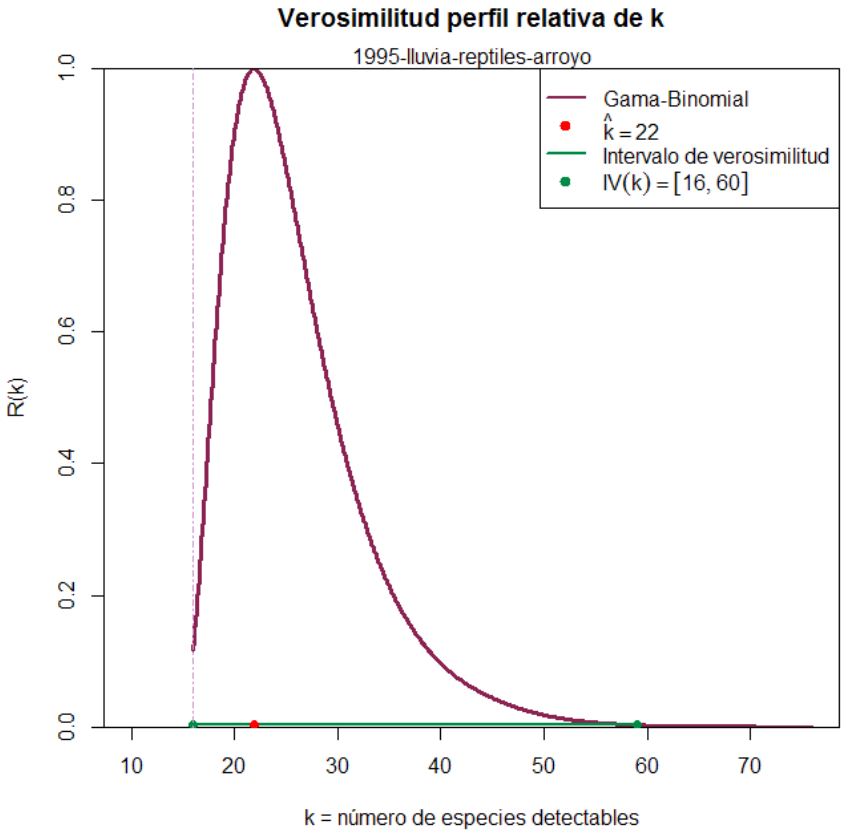
$$(t_{r1}, ..., t_{r_1 m_1}, m_1) = (13.57, 5.71, 2.57, 2.28, 0.57, 0.42, 0.28, 0.14, 16) \tag{4.2.3}$$

Maximizando la verosimilitud de esta muestra observada con respecto a (k_1, α_1, β_1) se obtienen los estimadores e intervalos de verosimilitud de la Tabla 4.2.4. Los primeros para un nivel usual $c = 0.1465$ y los segundos calibrados mediante simulaciones para obtener una probabilidad de cobertura de 0.95, como se indicó en la Sección 1.3.4.

1995	EMV	IV (0.1465)	IV (95%)
k_1	21.89	$[16, 38]$	$[16, 60]$
α_1	0.333	$(0, 0.8)$	$(0, 1.69)$
β_1	0.209	$(0.03, 0.57)$	$(0, 0.97)$

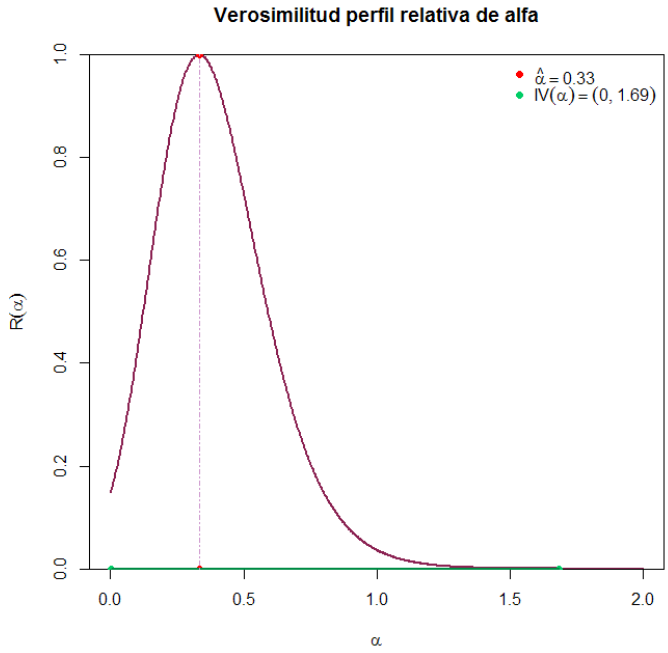
(4.2.4)

Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos calibrados para obtener el 95% de confianza para cada parámetro. Se muestran también los contornos de nivel para los parámetros (α_1, β_1).

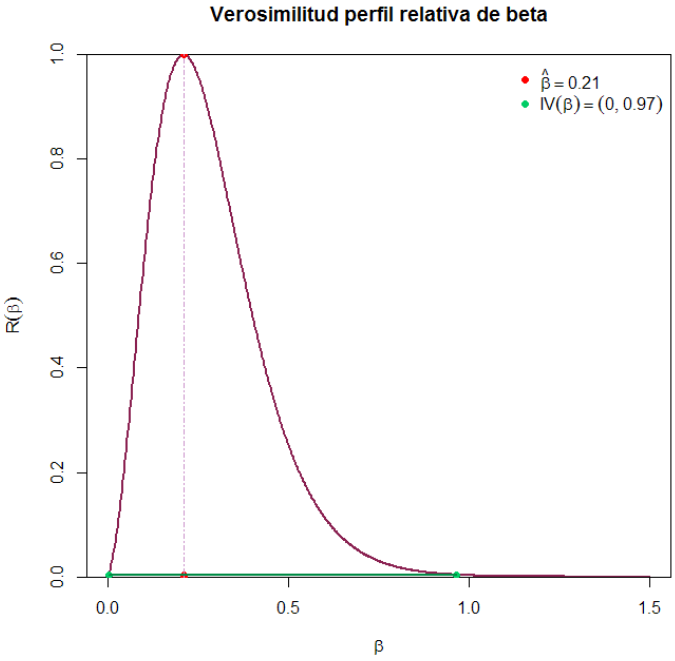


(4.2.5)

Figura 4.2.5**:** Función de verosimilitud perfil relativa para k e intervalo de confianza del 95% para los datos del año 1995. Se observaron $m_1 = 16$ especies en $r_1 = 7$ cuadrantes, este valor está marcado con una línea vertical en la gráfica. El número de especies detectables estimado bajo el modelo Gama es $\widehat{k}_1 = 22$, indicado con un punto rojo sobre el intervalo.

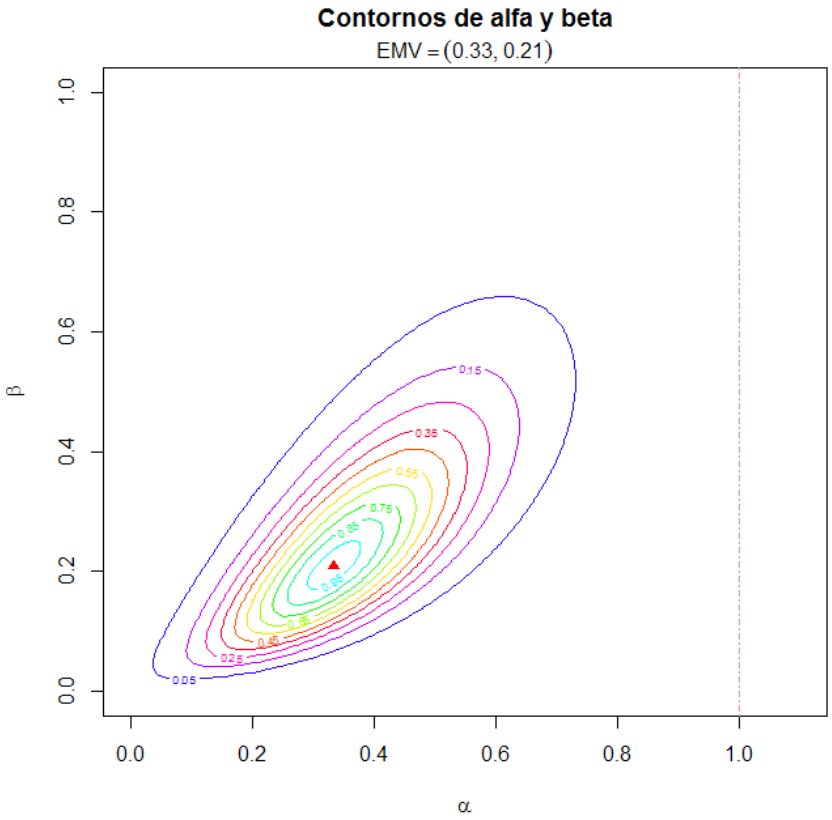


(4.2.6)



(4.2.7)

Figuras 4.2.6 y 4.2.7**:** Funciones de verosimilitud perfil relativa para α y β, e intervalos calibrados de confianza 95% para los datos del año 1995. Los EMV se marcan con líneas verticales punteadas.



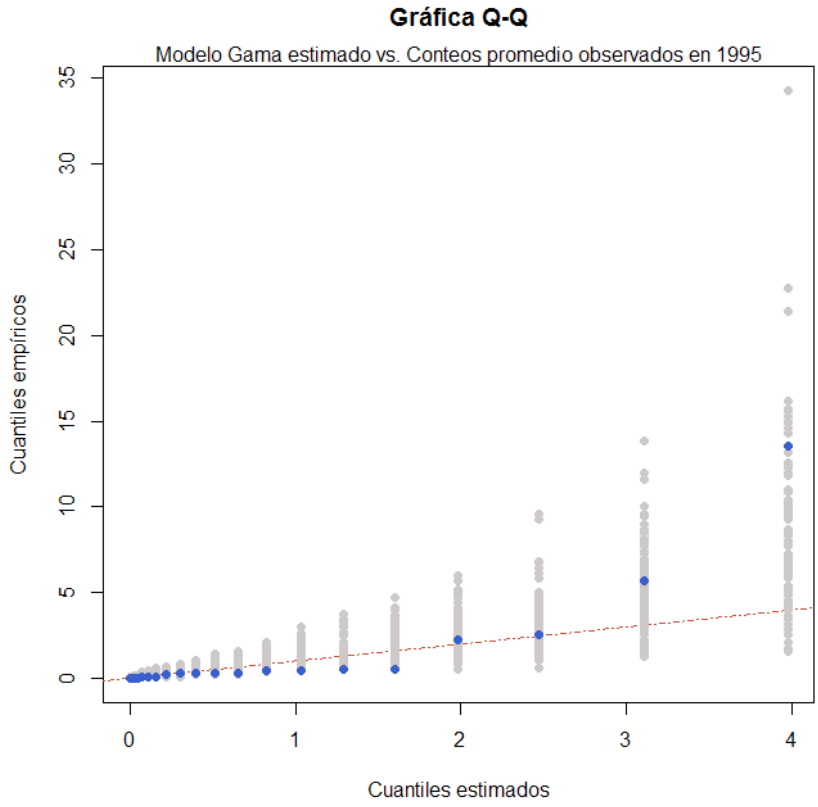
(4.2.8)

Figura 4.2.8**:** Curvas de nivel de la función de verosimilitud relativa de α y β. Se observa que el valor $\alpha = 1$ no es un valor razonable para este parámetro dados los datos observados.

Finalmente, se elaboró una gráfica Q-Q a partir de las observaciones de los conteos promedio $(t_{r1}, ..., t_{r_1 m_1})$ y la distribución Gama estimada (cuyos parámetros son $\widehat{\alpha}_1$ y $\widehat{\beta}_1$), siguiendo el esquema sugerido en la Sección 1.3.7. Los conteos promedio son considerados como datos censurados por intervalo en el caso haber observaciones repetidas en el vector $(t_{r1}, ..., t_{r_1 m_1})$, tales intervalos están dados por $[\lambda_{j1}, \lambda_{j2}]$ en la Tabla 4.2.2. Para incluir en la gráfica Q-Q las intensidades Poisson de especies no observadas, se consideró como censuradas las $k_1 - m_1$ especies no observadas en la muestra se supone se encuentran en el intervalo $[\lambda_0, \lambda_1)$. Se incluyó además una nube de puntos obtenida de simulaciones del modelo estimado

Gama para comparar la variabilidad natural del modelo propuesto contra la de los datos observados.

Se observa que la mayoría de los puntos correspondientes a los datos (puntos azules de la Figura 4.2.9) se encuentran cerca de la recta de 45°, más aún, todos se encuentran dentro de la nube de puntos, lo cual indica que el modelo Gama es razonable para este conjunto de datos.



(4.2.9)

Figura 4.2.9**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Gama estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1995.

4.2.3 Modelo Lognormal-Binomial

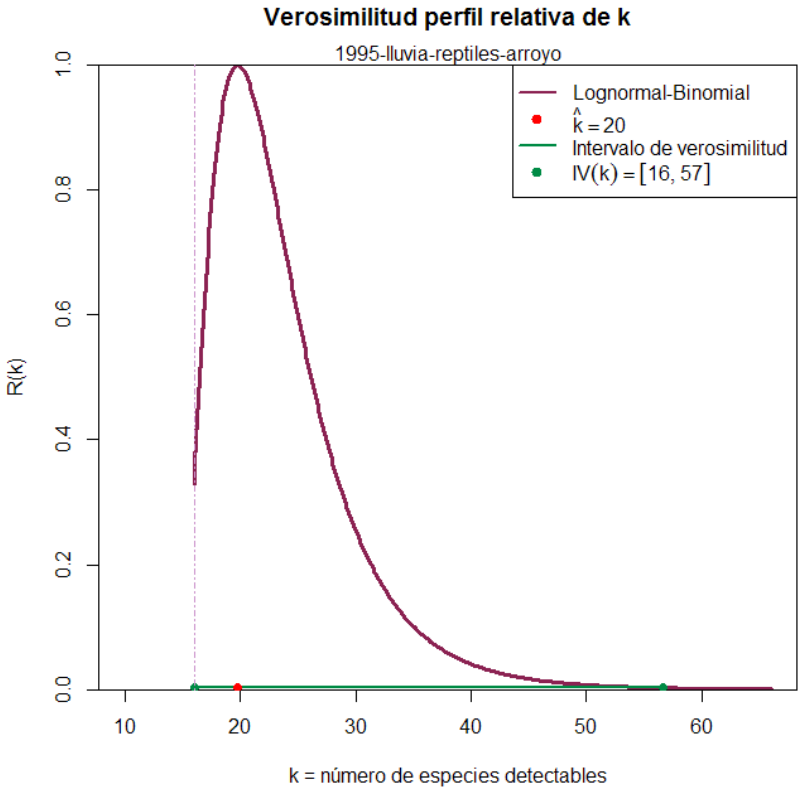
Dada la función de verosimilitud para (k_1, θ_1) basada en la muestra 4.2.3 definida en (2.3.4) , supóngase ahora que $G_T(\lambda; \theta_1)$ corresponde a una ley Lognormal truncada por la izquierda en λ_0 con parámetros $\theta_1 = (\mu_1, \sigma_1)$.

La verosimilitud correspondiente a la combinación de los modelos Lognormal y Binomial dada la muestra 4.2.3 se ha maximizado con respecto a (k_1, μ_1, σ_1) para obtener los siguientes estimadores e intervalos de verosimilitud calibrados para obtener una probabilidad de cobertura de 0.95.

1995	EMV	$IV\,(0.1465)$	$IV\,(95\%)$
k_1	19.77	$[16, 34]$	$[16, 57]$
μ_1	-0.7	$(-1.87, 0.17)$	$(-7.89, 1.24)$
σ_1	1.55	$(0.99, 2.62)$	$(0.71, 6.25)$

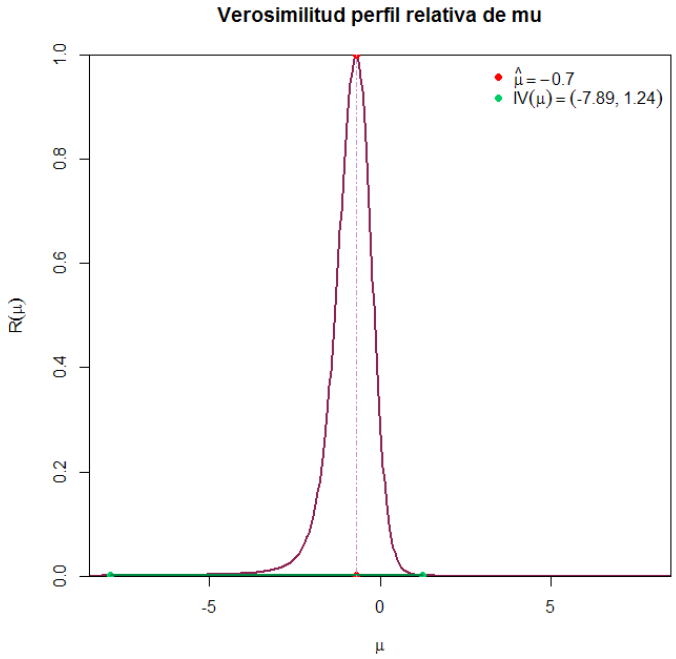
(4.2.10)

Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos del 95% de confianza calibrados de cada parámetro. Se muestran también los contornos de nivel para los parámetros (μ_1, σ_1) .

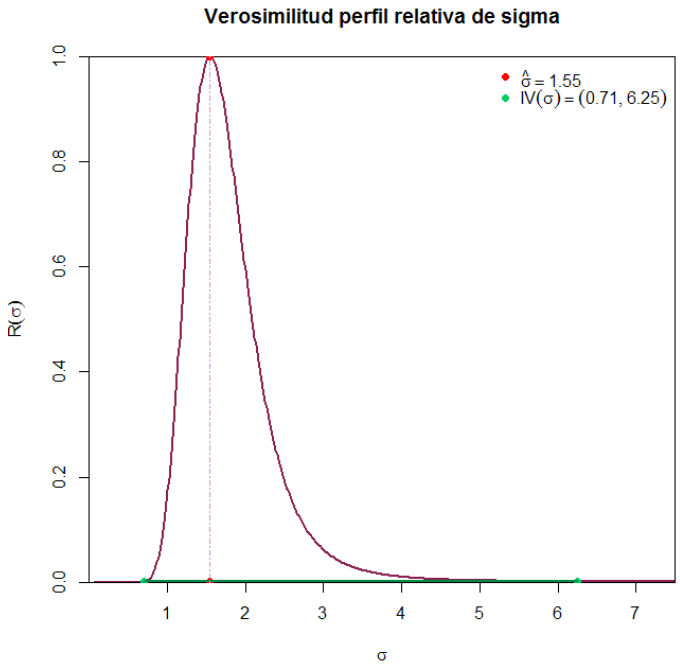


(4.2.11)

Figura 4.2.11**:** Función de verosimilitud perfil relativa para k e intervalo de confianza del 95% para los datos del año 1995. Se observaron $m_1 = 16$ especies en $r_1 = 7$ cuadrantes (se marca con una línea vertical punteada este valor). El número de especies detectables estimado bajo el modelo Lognormal es $\widehat{k}_1 = 20$ (indicado con un punto rojo sobre el intervalo).

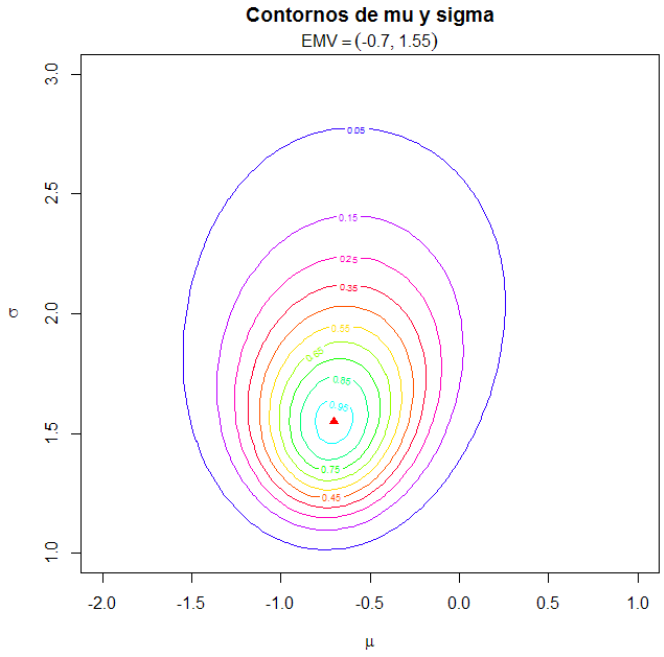


(4.2.12)



(4.2.13)

Figuras 4.2.12 y 4.2.13**:** Funciones de verosimilitud perfil relativa para μ y σ. Se muestran los intervalos calibrados de confianza del 95% para los datos del año 1995. Las líneas verticales señalan el valor del EMV.

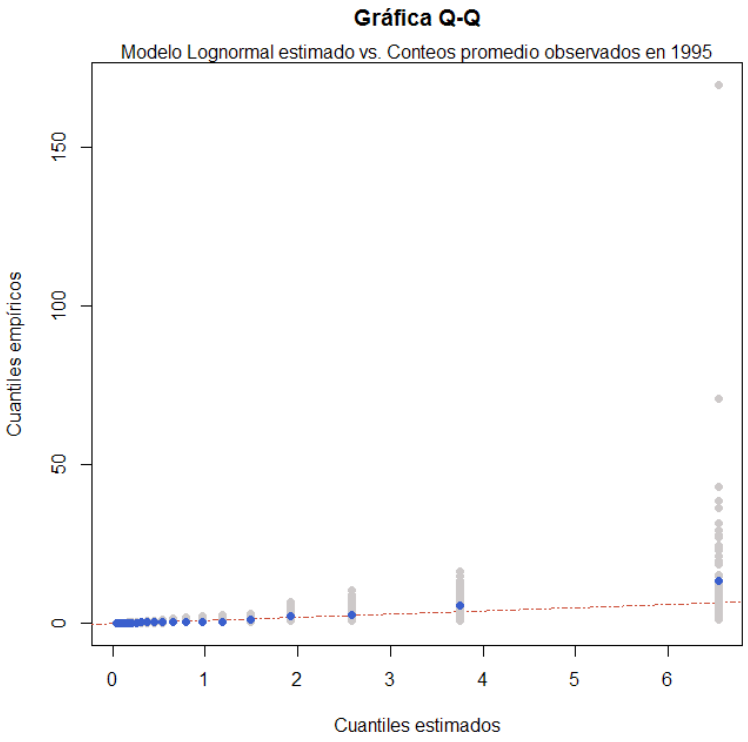


(4.2.14)

Figura 4.2.14**:** Curvas de nivel de la función de verosimilitud relativa de μ y σ.

En la gráfica Q-Q se observa una relación lineal entre los puntos correspondientes a los datos observados. La nube de puntos simulados cubre razonablemente a las observaciones. Así el modelo Lognormal parece adecuado para modelar las intensidades Poisson. De hecho, en comparación con el modelo Gama mostrado anteriormente (ver Figura 4.2.9), esta gráfica Q-Q nos muestra que el modelo Lognormal es también un buen candidato.

Además, la razón de verosimilitudes del modelo Gama al Lognormal fue de $\Lambda_r = 0.169$, lo cual indica que ambos modelos son razonables para los datos.



(4.2.15)

Figura 4.2.15**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Lognormal estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1995.

4.3 Arroyo 1996 (Normal)

Parámetros fijos que definen el modelo estadístico:

W	r_2	δ_2	p	λ_0	λ_1
2376	7	0.00295	0.99	0.00194	0.143

1996 (Normal)	Cuadrante							$m_2 = 11$
Reptiles	**1**	**2**	**3**	**4**	**5**	**6**	**7**	Individuos por especie
Especie 1	2	5	0	10	1	10	0	28
Especie 2	1	5	3	2	1	2	4	18
Especie 3	1	0	3	3	2	0	4	13
Especie 4	0	3	0	1	0	1	0	5
Especie 5	0	0	0	5	0	0	0	5
Especie 6	0	0	0	0	0	0	2	2
Especie 7	2	0	0	0	0	0	0	2
Especie 8	0	0	1	0	0	1	0	2
Especie 9	0	0	0	2	0	0	0	2
Especie 10	0	0	0	0	1	0	0	1
Especie 11	0	0	0	0	1	0	0	1

(4.3.1)

Tabla 4.3.1: Conteos observados en una muestra de $r = 7$ cuadrantes durante los meses de julio a octubre de 1996. Se observaron en total $m_2 = 11$ especies distintas.

4.3.1 Intervalos de verosimilitud-confianza para las intensidades Poisson

En la Tabla 4.3.2 se muestran los conteos promedio distintos observados en los r_2 cuadrantes de la muestra, t_{rl}, las frecuencias de cada promedio f_l, y los extremos del intervalo de verosimilitud-confianza del 95% para cada conteo promedio, $[\lambda_{l1}, \lambda_{l2}]$. Nótese que $m_2 = \sum_{l=1}^{6} f_l = 11$, además, $\lambda_0 < \lambda_{l1}, l = 1, ..., 6$ lo cual indica que todos los intervalos están dentro del dominio de la distribución truncada de las intensidades Poisson.

1996 **(Normal)**			
t_{lr}	f_l	λ_{l1}	λ_{l2}
4.0	1	2.69	5.67
2.57	1	1.55	3.95
1.85	1	1.02	3.05
0.71	2	0.25	1.53
0.28	4	0.047	0.88
0.14	2	0.008	0.63

(4.3.2)

Tabla 4.3.2**:** Hay $s = 6$ conteos promedio distintos en la muestra de $r_2 = 7$ cuadrantes del año 1996. En la tabla se muestran sus frecuencias e intervalos de verosimilitud-confianza del 95%.

4.3.2 Modelo Gama-Binomial

La primer columna de la Tabla 4.3.2 determina el vector de observaciones de este ejemplo

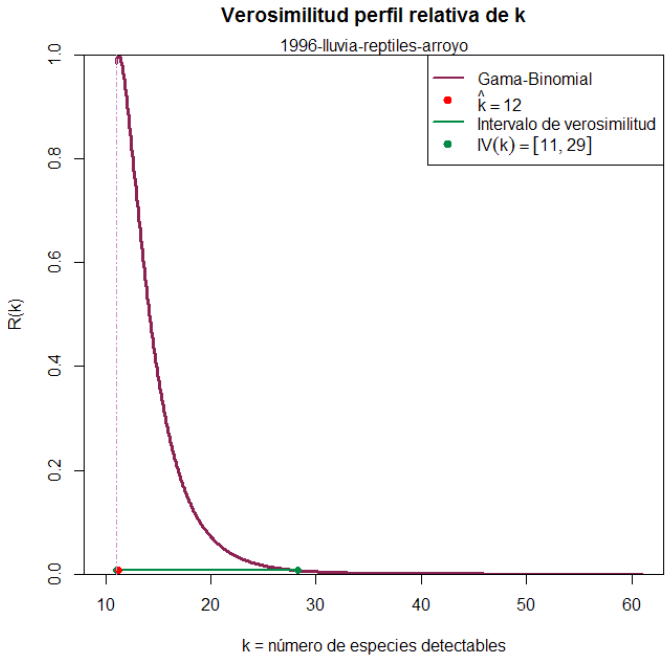
$$(t_{r1}, ..., t_{r_2 m_2}, m_2) = (4.0, 2.57, 1.85, 0.71, 0.28, 0.14, 11)\,. \tag{4.3.3}$$

Evaluando la verosimilitud en esta muestra y maximizando con respecto a (k_2, α_2, β_2) se obtienen los estimadores e intervalos de verosimilitud (Tabla 4.3.4) calibrados mediante simulaciones para obtener una probabilidad de cobertura de 0.95.

1996	EMV	IV (0.1465)	IV (95%)
k_2	11.25	$[11, 18]$	$[11, 33]$
α_2	1.39	$(0.39, 3.49)$	$(0.04, 6.22)$
β_2	1.27	$(0..28, 3.56)$	$(0.07, 5.76)$

(4.3.4)

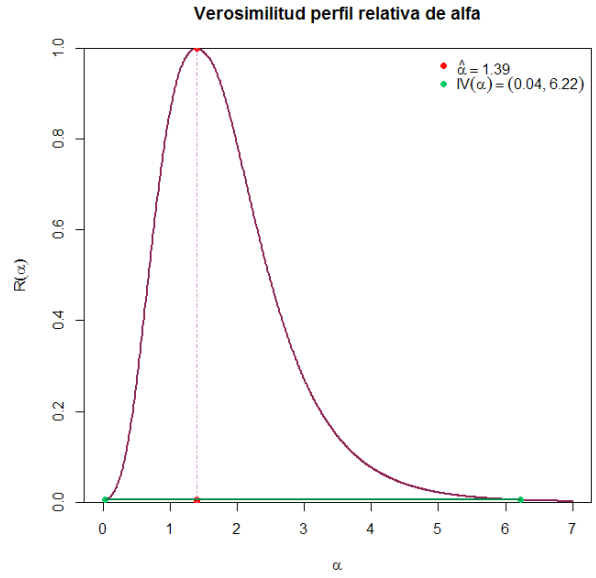
Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos calibrados para obtener el 95% de confianza para cada parámetro. Se muestran también los contornos de nivel para los parámetros (α, β).



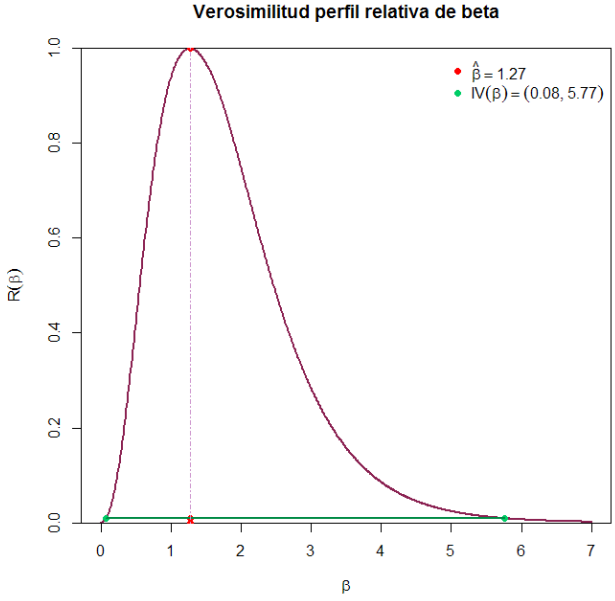
(4.3.5)

Figura 4.3.5**:** Función de verosimilitud perfil relativa de k e intervalo de confianza del 95% para el año 1996. Se observaron $m_2 = 11$ especies en $r_2 = 7$ cuadrantes (este valor se marca con una línea vertical punteada). El número de especies detectables estimado bajo el modelo Gama es $\widehat{k}_2 = 12$ (se indica con un punto sobre el intervalo).

.

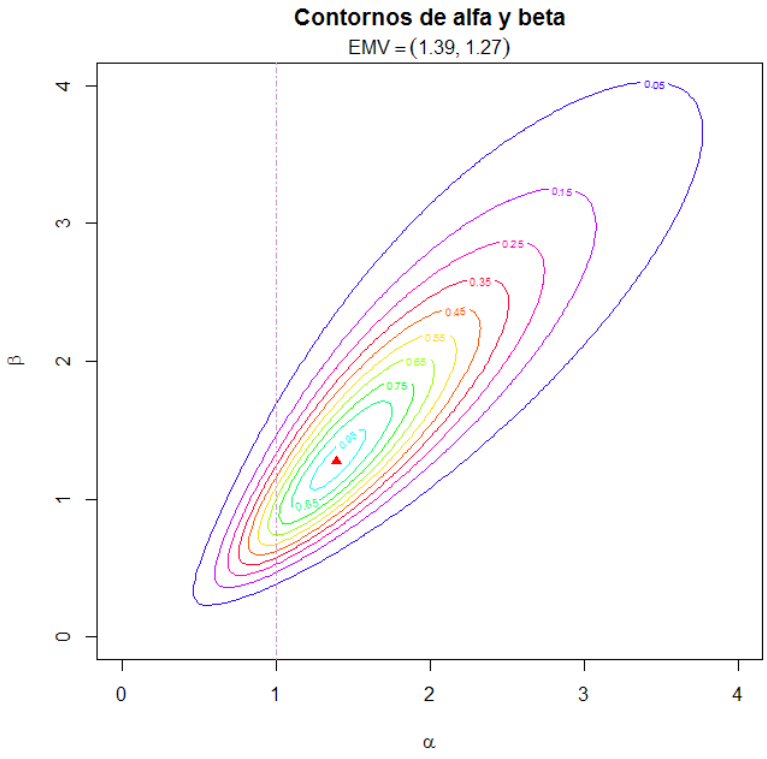


(4.3.6)



(4.3.7)

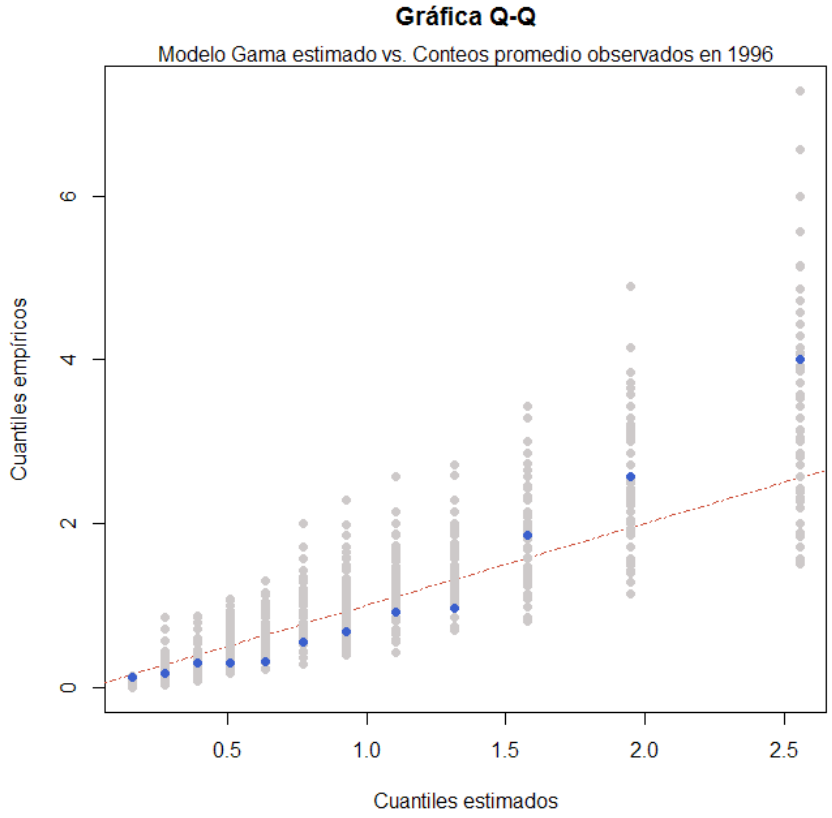
Figuras 4.3.6 y 4.3.7**:** Funciones de verosimilitud perfil relativa para α y β, e intervalos calibrados de confianza 95% para los datos del año 1996. Las líneas verticales señalan el valor del EMV.



(4.3.8)

Figura 4.3.8**:** Curvas de nivel de la función de verosimilitud relativa de α y β. Se observa que el valor $\alpha = 1$ es un valor razonable para este parámetro dados los datos observados, aunque las curvas de nivel se concentran a la derecha de este valor.

En la gráfica Q-Q asociada a este ejemplo bajo el modelo Gama, se observa que en su mayoría los puntos correspondientes a los datos se encuentran cerca de la recta de 45°, y todos se encuentran dentro de la nube de puntos. Así, el arreglo de puntos obtenido indica que el modelo Gama es razonable para este conjunto de datos.



(4.3.9)

Figura 4.3.9**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Gama estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1996.

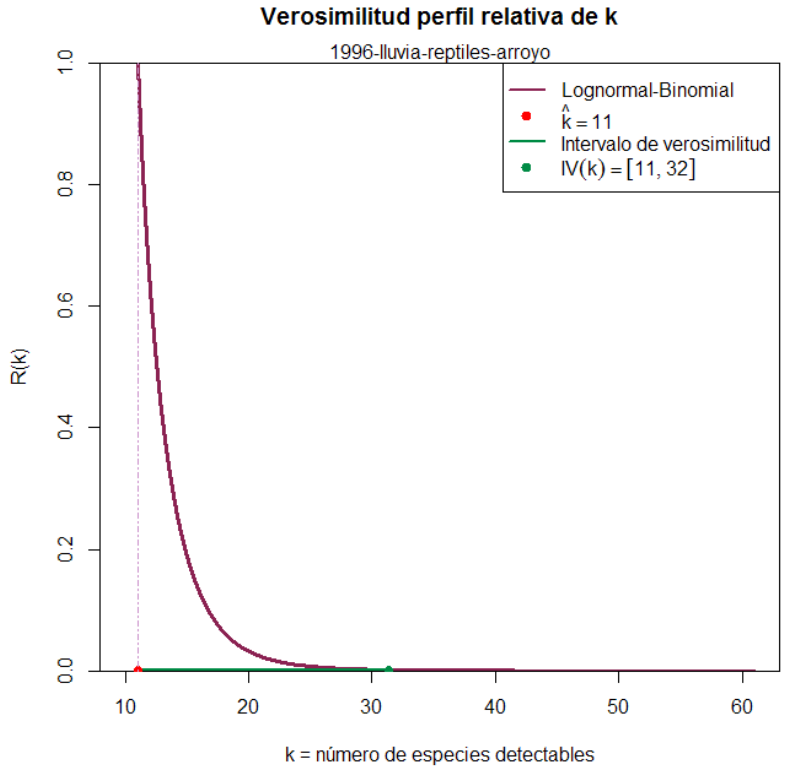
4.3.3 Modelo Lognormal-Binomial

La verosimilitud correspondiente a la combinación de los modelos Lognormal y Binomial dada la muestra 4.3.3 se ha maximizado con respecto a (k_2, μ_2, σ_2) para obtener los siguientes estimadores e intervalos de verosimilitud calibrados para obtener una probabilidad de cobertura de 0.95.

1996	EMV	$IV\,(0.1465)$	$IV\,(95\%)$
k_2	11.0	$[11, 16]$	$[11, 32]$
μ_2	-0.308	$(-1.1, 0.28)$	$(-1.63, 0.62)$
σ_2	0.861	$(0.53, 1.64)$	$(0.42, 2.47)$

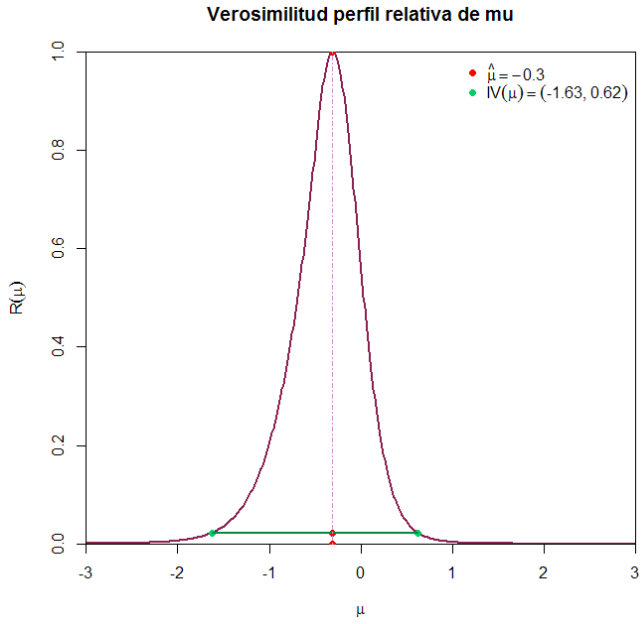
(4.3.10)

Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos del 95% de confianza calibrados de cada parámetro. Se muestran también los contornos de nivel para los parámetros (μ_2, σ_2).

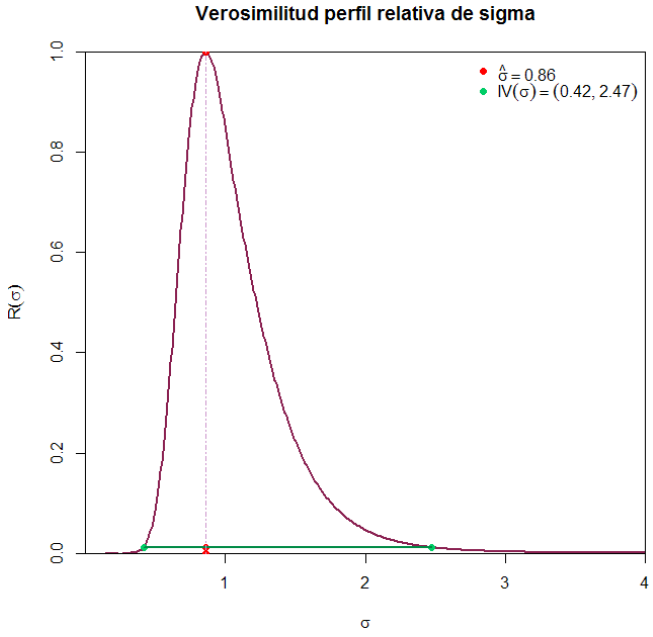


(4.3.11)

Figura 4.3.11**:** Función de verosimilitud perfil relativa para k e intervalo de confianza del 95% para los datos del año 1995. Se observaron $m_2 = 16$ especies en $r_2 = 7$ cuadrantes (este valor se marca con una línea vertical punteada en la gráfica). El número de especies detectables estimado bajo el modelo Lognormal es $\widehat{k}_2 = 20$ (en la gráfica se indica con un ponto rojo sobre el intervalo).

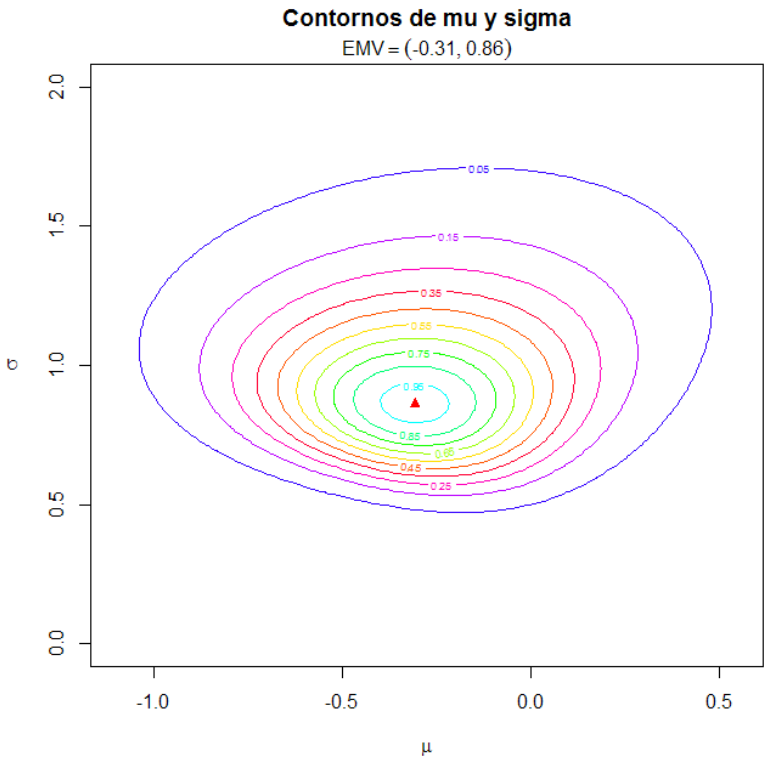


(4.3.12)



(4.3.13)

Figuras 4.3.12 y 4.3.13**:** Funciones de verosimilitud perfil relativa para μ y σ. Se muestran los intervalos calibrados de confianza del 95% para los datos del año 1996. Las líneas verticales señalan el valor del EMV.

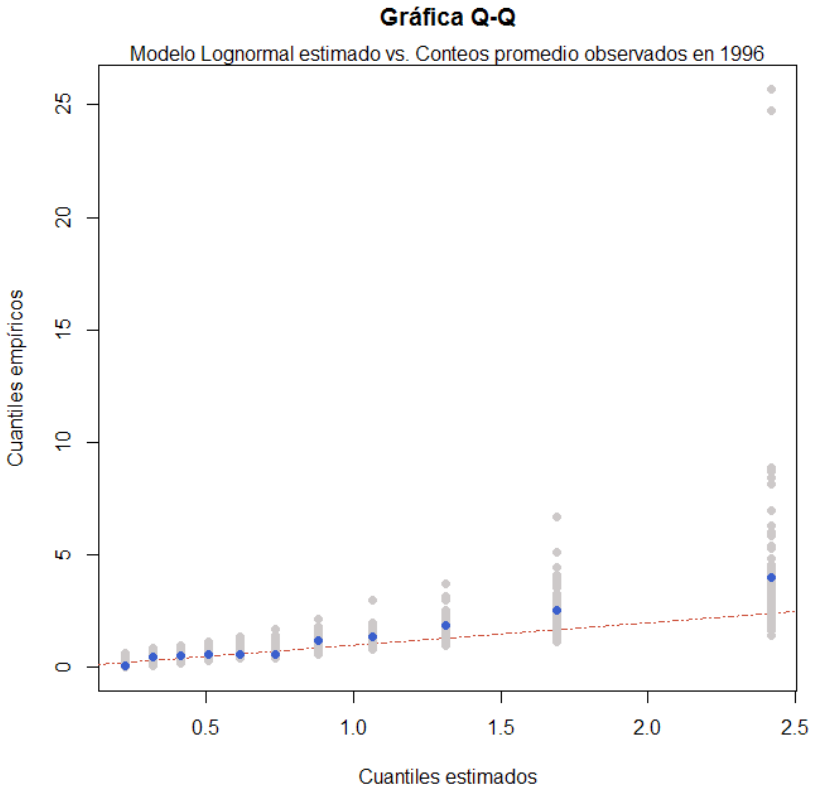


(4.3.14)

Figura 4.3.14**:** Curvas de nivel de la función de verosimilitud relativa de μ y σ.

La gráfica Q-Q para los datos de 1996 bajo el modelo Lognormal se muestra en la Figura 4.3.15. Ahí se puede observar que el arreglo de puntos correpondientes a los datos no se desvía significativamente de la recta de 45° ni sale de la nube de puntos simulada, por tanto, el modelo Lognormal parece representar bien a los datos. Comparando con la gráfica Q-Q obtenida bajo el modelo Gama (ver Figura 4.3.9), no se observa gran diferencia entre un modelo, ambos parecen explicar razonablemente a los datos.

Además, la razón de verosimilitudes del modelo Gama al Lognormal fue de $\Lambda_r = 0.417$, lo cual indica que ambos modelos son razonables para los datos.



(4.3.15)

Figura 4.3.15**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Lognormal estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1995.

4.4 Arroyo 1997 (El Niño)

Parámetros fijos que definen el modelo estadístico:

W	r_3	δ_3	p	λ_0	λ_1
2376	5	0.0021	0.99	0.00194	0.2

1997 (El Niño)	Cuadrante					$m_3 = 13$
Reptiles	**1**	**2**	**3**	**4**	**5**	Individuos por especie
Especie 1	9	3	4	8	6	30
Especie 2	1	6	0	9	2	18
Especie 3	2	4	0	1	7	14
Especie 4	0	3	1	0	1	5
Especie 5	1	1	1	0	2	5
Especie 6	0	0	0	2	0	2
Especie 7	0	1	1	0	0	2
Especie 8	1	0	0	0	0	1
Especie 9	1	0	0	0	0	1
Especie 10	0	0	0	1	0	1
Especie 11	0	0	1	0	0	1
Especie 12	0	0	0	0	1	1
Especie 13	0	0	0	1	0	1

(4.4.1)

Tabla 4.4.1 : Conteos observados en una muestra de $r_3 = 7$ cuadrantes durante los meses de julio a octubre de 1997. Se observaron en total $m_3 = 13$ especies distintas.

4.4.1 Intervalos de verosimilitud-confianza para las intensidades Poisson

La Tabla 4.4.2 contiene los conteos promedio distintos observados en los r_3 cuadrantes de la muestra, t_{rl}, las frecuencias de cada promedio f_l, y los extremos del intervalo de verosimilitud-confianza del 95% para cada $t_{rl}, [\lambda_{l1}, \lambda_{l2}]$. Nótese que $m_3 = \sum_{l=1}^{6} f_l$, además, $\lambda_0 < \lambda_{l1}, l = 1, ..., 6$.

1997 **(La Niña)**			
t_{lr}	f_l	λ_{l1}	λ_{l2}
6.0	1	4.10	8.4
3.6	1	2.18	5.52
2.8	1	1.57	4.53
1.0	2	0.35	2.14
0.4	2	0.066	1.23
0.2	6	0.011	0.88

(4.4.2)

Tabla 4.4.2**:** Hay 6 conteos promedio distintos en la muestra de 5 cuadrantes del año 1997. En la tabla se muestran sus frecuencias e intervalos de verosimilitud-confianza del 95%.

4.4.2 Modelo Gama-Binomial

La primer columna de la Tabla 4.4.2 determina el vector de observaciones de este ejemplo

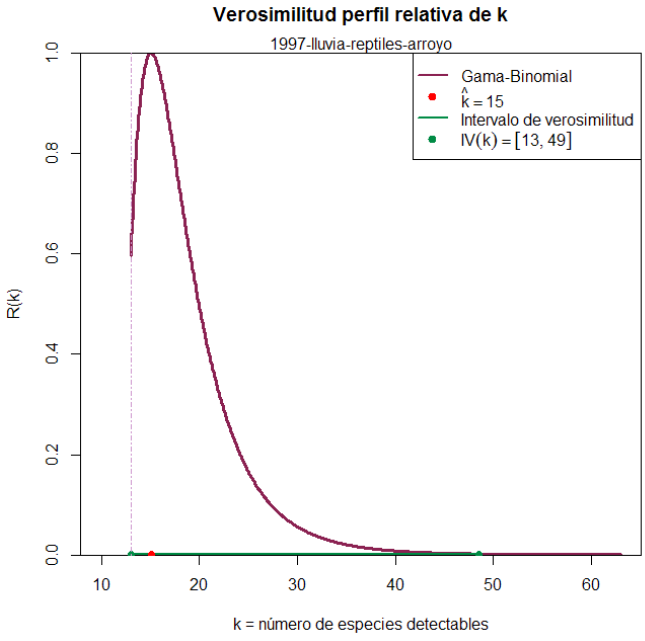
$$(t_{r1}, ..., t_{r_3 m_3}, m_3) = (6.0, 3.6, 2.8, 1.0, 0.4, 0.2, 13)\,. \tag{4.4.3}$$

Evaluando la verosimilitud en esta muestra y maximizando con respecto a (k_2, α_2, β_2) se obtienen los estimadores e intervalos de verosimilitud (Tabla 4.3.4) calibrados mediante simulaciones para obtener una probabilidad de cobertura de 0.95.

1997	EMV	IV (0.1465)	IV (95%)
k_3	15.06	$[13, 26]$	$[13, 49]$
α_3	0.87	$(0.23, 2.39)$	$(0, 5.4)$
β_3	0.66	$(0.15, 1.95)$	$(0.01, 3.76)$

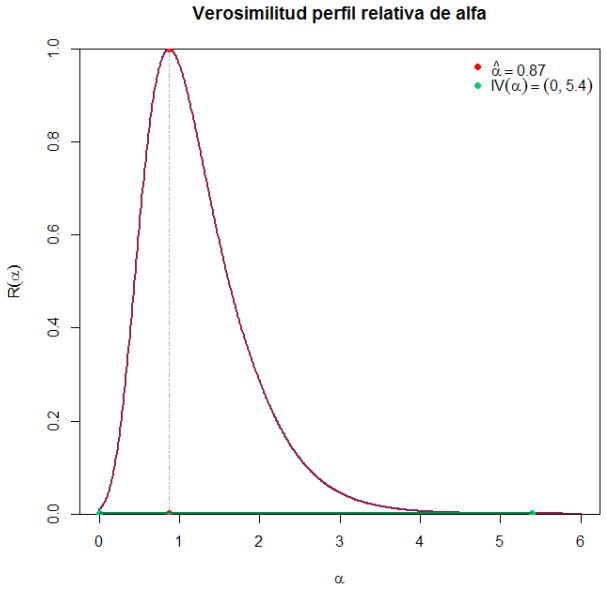
(4.4.4)

Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos calibrados para obtener el 95% de confianza para cada parámetro. Se muestran también los contornos de nivel para los parámetros (α, β).

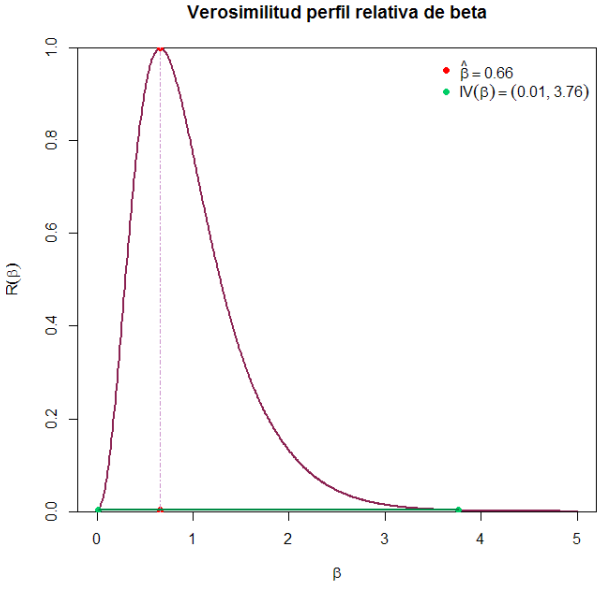


(4.4.5)

Figura 4.4.5**:** Función de verosimilitud perfil relativa para k e intervalo de confianza del 95% para los datos del año 1997. Se observaron $m_3 = 13$ especies en $r_3 = 5$ cuadrantes (línea vertical punteada). El número de especies detectables estimado bajo el modelo Gama es $\widehat{k_3} = 15$ (punto rojo sobre el intervalo).

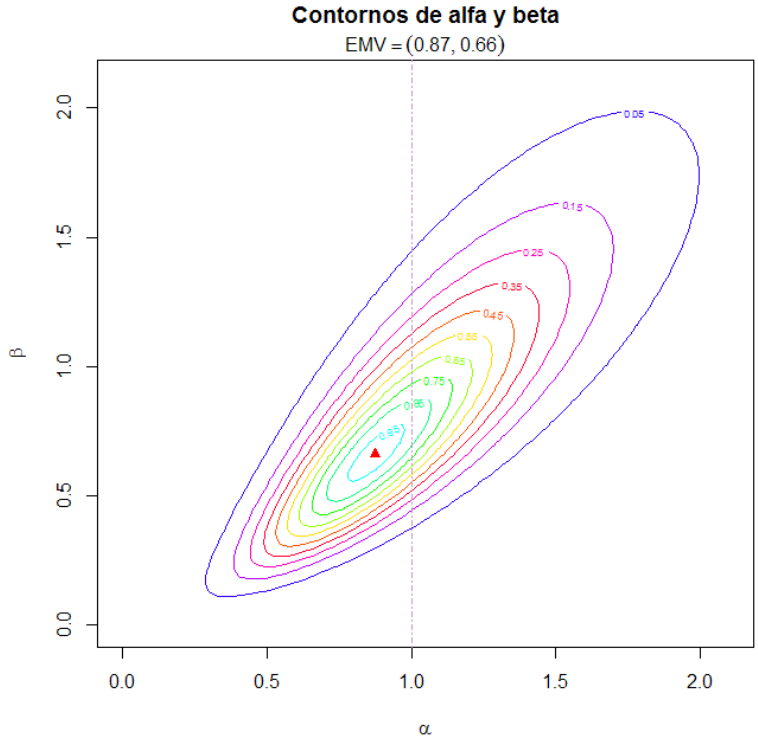


(4.4.6)



(4.4.7)

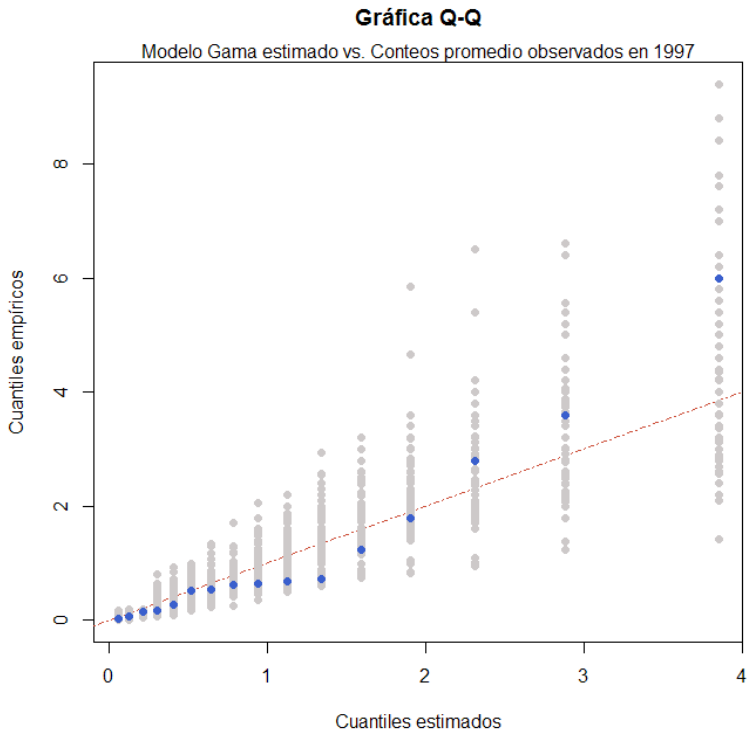
Figuras 4.4.6 y 4.4.7**:** Funciones de verosimilitud perfil relativa para α y β, e intervalos calibrados de confianza 95% para los datos del año 1997. Las líneas verticales señalan el valor del EMV.



(4.4.8)

Figura 4.3.8**:** Curvas de nivel de la función de verosimilitud relativa de α y β. Se observa que el valor $\alpha = 1$ es un valor muy plausible para este parámetro dados los datos observados.

La Figura 4.4.9 corresponde a la gráfica Q-Q bajo el modelo Gama, los puntos correspondientes a los datos se encuentran cerca de la recta de 45°, sobre todo aquellos de magnitud pequeña. Además, todos los puntos se encuentran dentro de la nube que representa la variabilidad inherente del modelo. Así, el arreglo de puntos obtenido indica que el modelo Gama es razonable para este conjunto de datos.



(4.4.9)

Figura 4.4.9**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Gama estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1997.

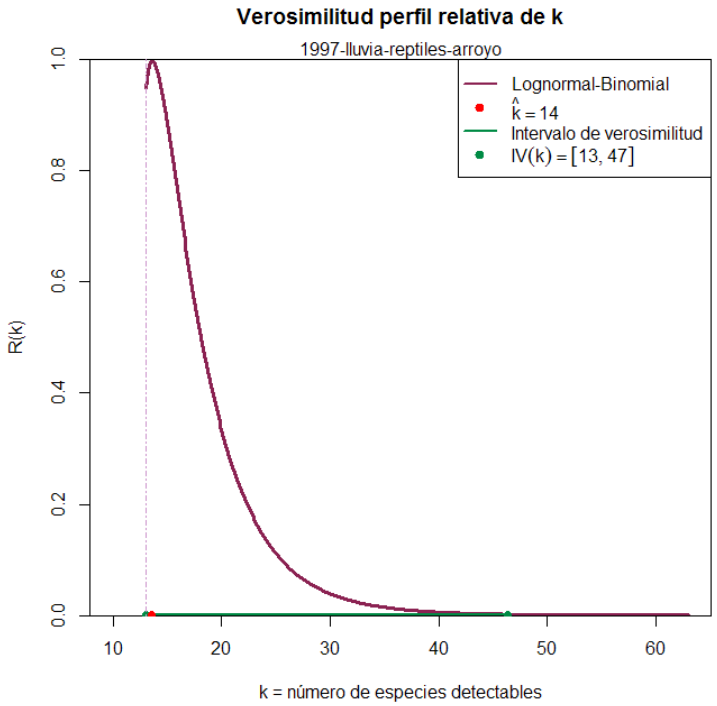
4.4.3 Modelo Lognormal-Binomial

Dada la muestra 4.4.3, la verosimilitud correspondiente a la combinación de los modelos Lognormal y Binomial se ha maximizado con respecto a (k_3, μ_3, σ_3) para obtener los siguientes estimadores e intervalos de verosimilitud calibrados con probabilidad de cobertura de 0.95.

1997	EMV	$IV\,(0.1465)$	$IV\,(95\%)$
k_3	13.62	$[13, 24]$	$[13, 47]$
μ_3	-0.21	$(-1.27, 0.43)$	$(-2.49, 1.03)$
σ_3	0.99	$(0.6, 1.99)$	$(0.44, 3.58)$

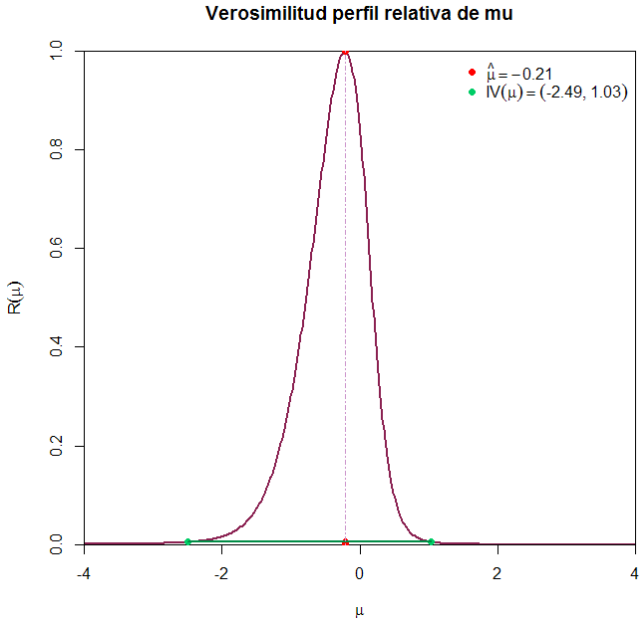
(4.4.10)

Las verosimilitudes perfil relativas se muestran a continuación junto con los intervalos del 95% de confianza calibrados de cada parámetro. Se muestran también los contornos de nivel para los parámetros (μ_3, σ_3).

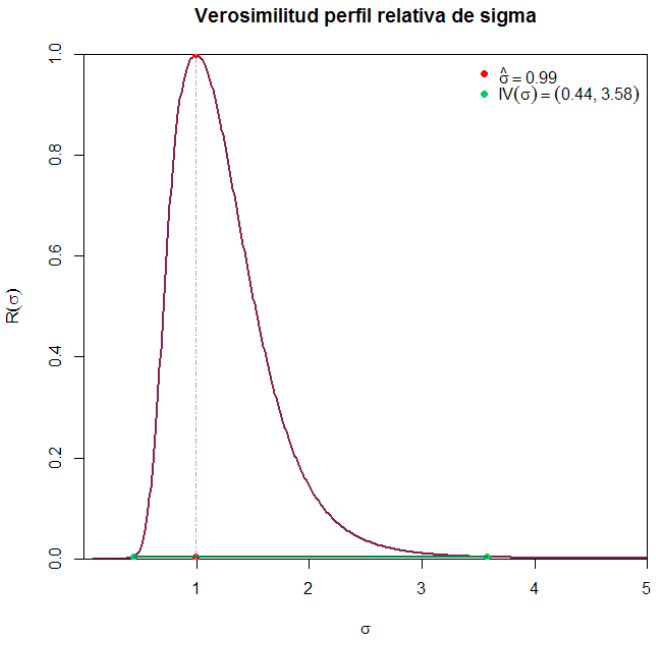


(4.4.11)

Figura 4.4.11**:** Función de verosimilitud perfil relativa de k e intervalo de confianza del 95% para el año 1997. Se observaron $m_3 = 13$ especies en $r_3 = 5$ cuadrantes (este valor se marca con una línea vertical punteada). El número de especies detectables estimado bajo el modelo Lognormal es $\widehat{k}_3 = 14$ (se indica con un ponto rojo sobre el intervalo).

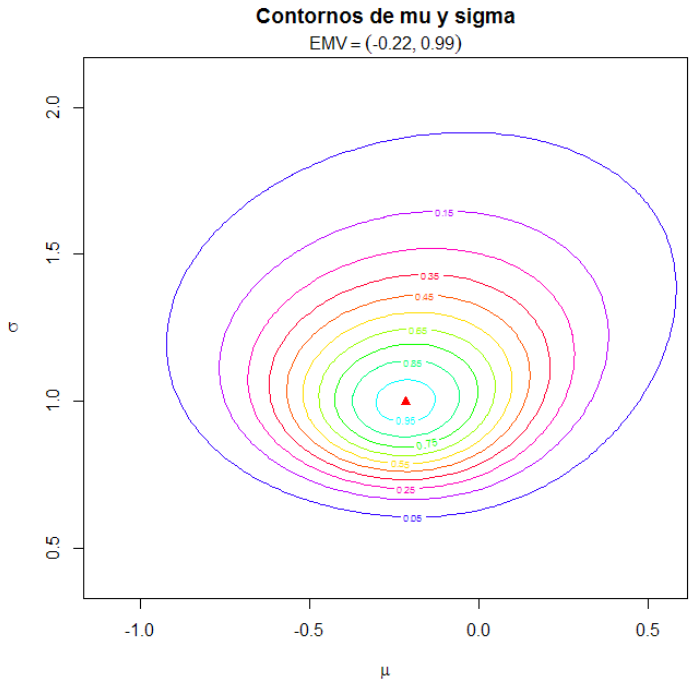


(4.4.12)



(4.4.13)

Figuras 4.4.12 y 4.4.13**:** Funciones de verosimilitud perfil relativa para μ y σ. Se muestran los intervalos calibrados de confianza del 95% para los datos del año 1997. Las líneas verticales señalan el valor del EMV.

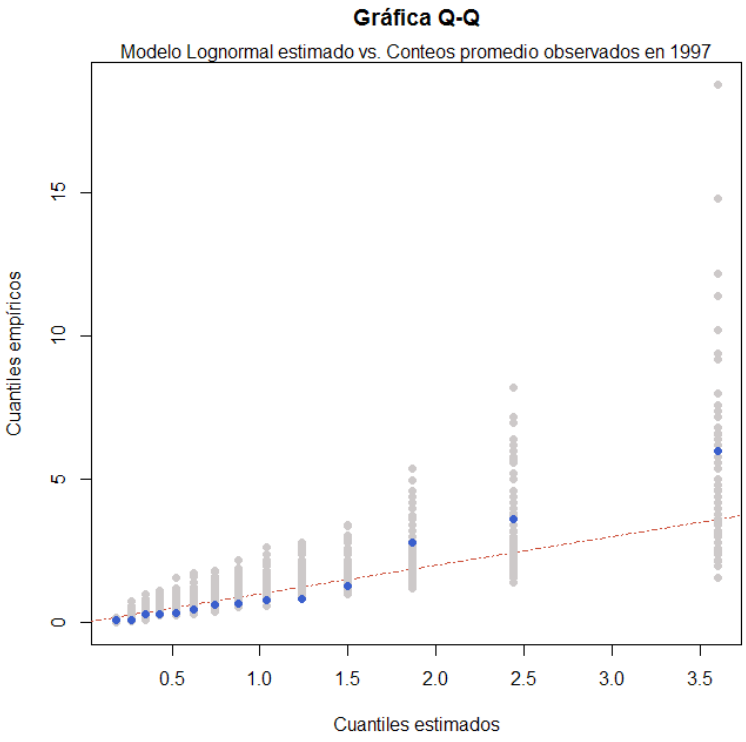


(4.4.14)

Figura 4.4.14**:** Curvas de nivel de la función de verosimilitud relativa de μ y σ.

La gráfica Q-Q para los datos de 1996 cuando se considera el modelo Lognormal para las intensidades Poisson se muestra en la Figura 4.4.15. Se puede observar que el arreglo de puntos correpondientes a los datos no se desvía significativamente de la recta de 45°, además la nube de puntos simulada cubre todos los puntos de la muestra. Por tanto, el modelo Lognormal parece representar bien a los datos. Comparando con la gráfica Q-Q obtenida bajo el modelo Gama (ver Figura 4.4.9), no se observa gran diferencia entre un modelo y otro, ambos parecen explicar razonablemente a los datos.

Además, la razón de verosimilitudes del modelo Gama al Lognormal fue de $\Lambda_r = 0.413$, lo cual indica que ambos modelos son razonables para los datos.



(4.4.15)

Figura 4.4.15**:** Gráfica Q-Q correspondiente a los cuantiles del modelo Lognormal estimado (eje X) y los cuantiles empíricos (eje Y) obtenidos de los datos del año 1997.

4.5 Discusión

4.5.1 Razón de verosimilitudes para los tres años

La Tabla 4.5.1 contiene los valores de la razón de verosimilitudes del modelo Gama al Lognormal, en la segunda columna, y del modelo Lognormal al Gama, en la tercera columna. Fijándose en el primer cociente, en los tres años se obtuvieron valores de la razón de verosimilitud cercanos a uno, lo cual indica que ambos modelos son razonables para los datos. En

muchos contextos se suele tomar la convención de que si la razón de verosimilitudes se encuentra entre 0.1465 y 6.82 (el recíproco de 0.1465), los dos modelos dan un ajuste similar a los datos. La validación de cualquiera de los modelos dependerá entonces de otros métodos como las gráficas Q-Q que se han descrito previamente.

En el caso particular de los datos de reptiles de Chamela, las gráficas Q-Q indicaron un buen ajuste para los dos modelos y los valores de la razón de verosimilitudes están dentro del intervalo mencionado, por lo que los dos modelos son razonables y ajustan bien a los datos. Como se mencionará más adelante, se prefiere al modelo Gama, debido a que da respuestas más conservadoras sobre la estimación de k y los intervalos de estimación tienen mejores cobertura que los del modelo Lognormal. A continuación se dan más detalles.

Como se comprobó en el Capítulo 3 a partir de simulaciones, los modelos Gama y Lognormal suelen ser indistinguibles para modelar los valores esperados de individuos de especies detectables. Así que los valores de la razón de verosimilitudes obtenidos para los ejemplos de este capítulo concuerdan con el resultado obtenido de simulaciones.

	$\Lambda_r\left(\mathcal{L}_{gama}, \mathcal{L}_{\log normal}\right)$	$\Lambda_r\left(\mathcal{L}_{\log normal}, \mathcal{L}_{gama}\right)$
1995	0.169	5.92
1996	0.417	2.39
1997	0.413	2.42

(4.5.1)

Figura 4.5.1**:** Razones de verosimilitudes de los modelos Gama y Lognormal obtenidos para los datos de los años bajo estudio, 1995, 1996 y 1997. Los valores obtenidos indican que los modelos Gama y Lognormal son ambos razonables para los datos.

Aun para el año 1995, donde el modelo Lognormal es casi seis veces más preferible que el modelo Gama, ambos modelos son razonables. Sin embargo, el modelo Gama es más conservador para k y se preferirá en los tres casos.

4.5.2 Densidades estimadas para las intensidades Poisson

En las Figuras 4.5.3, 4.5.4 y 4.5.5, se grafican las densidades Gama y Lognormal estimadas para los años 1995, 1996 y 1997, respectivamente. En general, se observa que las formas de las densidades son similares sobre todo en las colas derechas. Sin embargo, discrepan un poco para valores pequeños de λ. La densidad Gama sustenta que hay una proporcion mayor de valores pequeños de λ; esto implica que la Gama sustenta que hay mas especies raras en contraste con la Lognormal. Estas especies raras tienen probabilidades altas de no ser observadas en los cuadrantes por tener pocos individuos en la region.

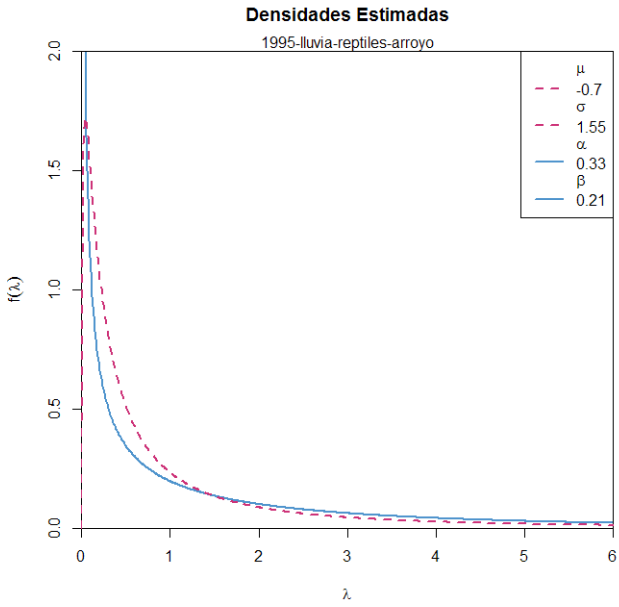
Los diferentes valores estimados para los parámetros de la hiperdensidad asociada a los parámetros de intensidad Poisson, se muestran en la Tabla 4.5.2, tanto para el modelo Gama $\left(\widehat{\alpha}, \widehat{\beta}\right)$ como para el Lognormal $(\widehat{\mu}, \widehat{\sigma})$ y los tres años de interés.

Para el año 1996, las densidades estimadas coinciden en tener una moda; mientras que en los años 1995 y 1996 el parámetro de forma estimado para el modelo Gama es menor que uno, lo cual provoca que la densidad sea cóncava y no hay una moda. Así la discrepancia entre estas densidades es notable para valores pequeños de λ y puede implicar una diferencia significativa en la cantidad de especies detectables raras (de las que se observan pocos individuos en la muestra).

	Gama		Lgnormal	
	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\mu}$	$\widehat{\sigma}$
1995	0.33	0.21	-0.7	1.55
1996	1.4	1.27	-0.3	0.86
1997	0.87	0.66	-0.2	0.99

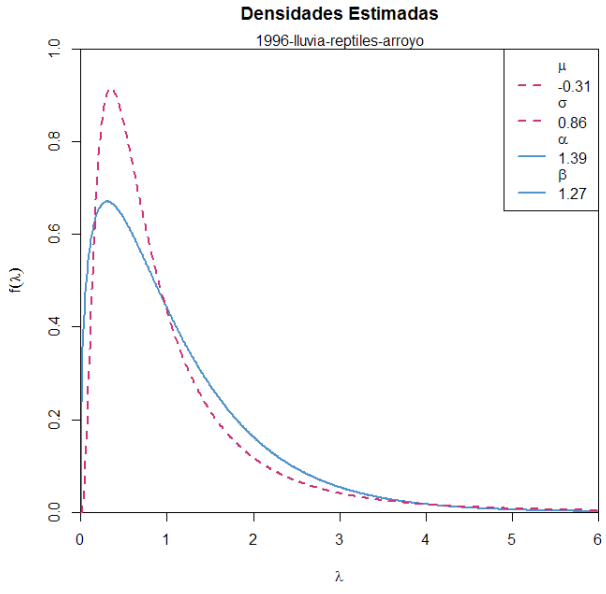
(4.5.2)

Tabla 4.5.2**:** Estimadores de máxima verosimilitud de los parámetros de los modelos Gama y Lognormal asociados a las intensidades Poisson $\lambda_j, j = 1, ..., m$.



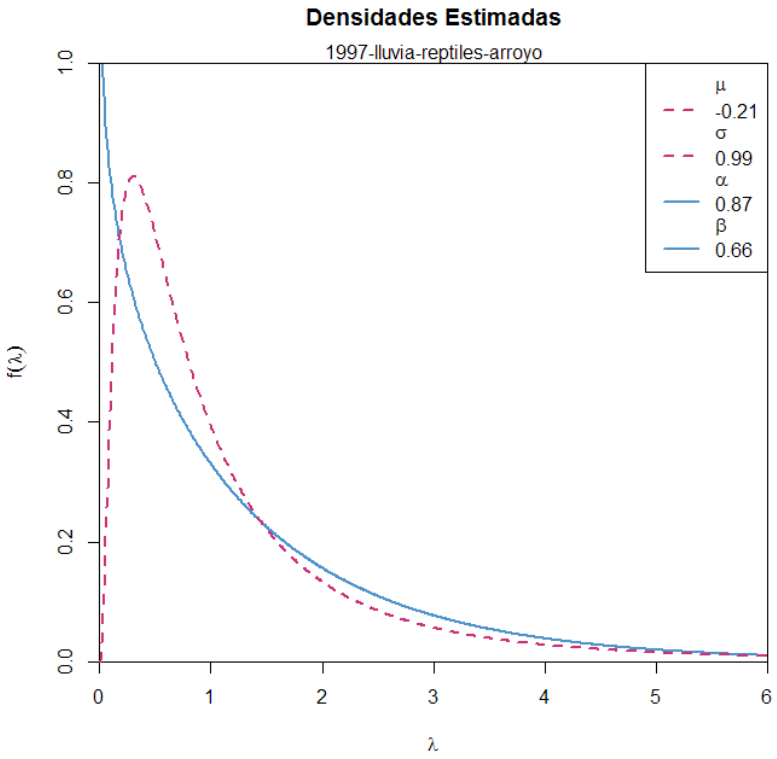
(4.5.3)

Figura 4.5.3 : Densidades Gama (línea sólida) y Lognormal (línea punteada) estimadas a partir de los datos de 1995.



(4.5.4)

Figura 4.5.4 : Densidades Gama (línea sólida) y Lognormal (línea punteada) estimadas a partir de los datos de 1996.



(4.5.5)

Figura 4.5.5 : Densidades Gama (línea sólida) y Lognormal (línea punteada) estimadas a partir de los datos de 1997.

4.5.3 Comparación de perfiles para k

Las funciones de verosimilitud perfil de k asociadas a los modelos Gama o Lognormal para las intensidades Poisson conllevan a inferencias sobre k un tanto distintas. En las Figuras 4.5.8, 4.5.9 y 4.5.10 se muestran ambas verosimilitudes perfiles para cada uno de los años de interés. En los tres casos, la verosimilitud perfil obtenida bajo el modelo Gama proporciona inferencias más conservadoras para k, en el sentido de que el intervalo de verosimilitud calibrado para obtener un 95% de confianza para k incluye valores mayores, mismos que son descartados bajo el modelo Lognormal.

Los niveles de verosimilitud calibrados mediante simulaciones de muestras similares a la original para cada año, se muestran en la Tabla 4.5.6. Los intervalos de verosimilitud obtenidos de cortar la verosimilitud perfil relativa de k en estos niveles calibrados tienen una probabilidad de cobertura de aproximadamente 0.95.

No se observa un comportamiento particular en los valores de estos niveles de verosimilitud calibrados. Sólo se nota que para el año 1996 el nivel de verosimilitud del modelo Gama es mayor que el obtenido para el modelo Lognormal. Sin embargo, en todos los casos los intervalos asociados al modelo Gama son más conservadores. Por tanto, se recomienda trabajar con el modelo Gama para hacer inferencias sobre el número de especies detectables k.

	Niveles de verosimilitud calibrados para k	
Año	Modelo Gama	Modelo Lognormal
1995	0.00215	0.00268
1996	0.00682	0.0014
1997	0.00223	0.00314

(4.5.6)

Tabla 4.5.6**:** Niveles de verosimlitud calibrados mediantes simulaciones para obtener intervalos de verosimilitud de confianza aproximada de 95%, para los modelos Gama y Lognormal propuestos para modelar los valores esperados de individuos de una especie detectable, en tres años distintos.

Finalmente, algunas comentarios respecto a los efectos de los fenómenos meteorológicos de la Oscilación del Sur, que comprende los años bajo estudio 1995, tipo Niña y 1997 tipo Niño. El año 1996 fue declarado normal.

Como se observa en la Figura 4.5.9, para el año normal el número de especies detectables fue marcadamente menor, a comparación de los años 1995 y 1997. El extremo derecho del intervalo de verosimilitud para k es aproximadamente la mitad que en los años anterior y posterior.

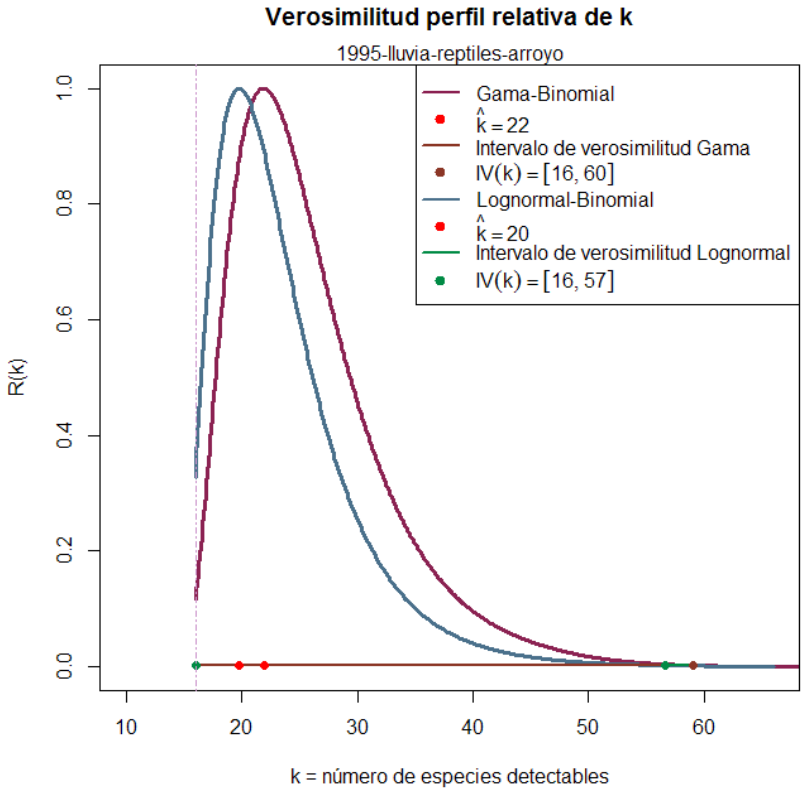
El año anterior al año normal 1995, fue de tipo Niña. Durante este año el invierno fue notoriamente seco (solo 28 mm de lluvia) al igual que el resto del año hidrológico.donde no llovió. Posiblemente el tener muchos meses consecutivos anteriores tan secos haya influido en la poca presencia de especies de reptiles para el verano de 1996.

Habría que ratificar si para otros años se cumple este patrón, de tener pocas especies detectables en el verano posterior a un año Niña, donde el invierno haya sido en general seco (de tal forma que se acumulen varios meses consecutivos con pocas lluvias).

	Lluvias				
Tipo de Año Hidrológico	Jun-Sep	Oct-Ene	Feb-May	Total anual	Intervalo para k
1994, Niño	275.5	151.5	0	427	—
1995, Niña	752.6	28	0	780.6	$[16, 60]$
1996, Normal	595.5	330	25	950.5	$[11, 32]$
1997, Niño	346.6	253.5	4	604.1	$[13, 49]$

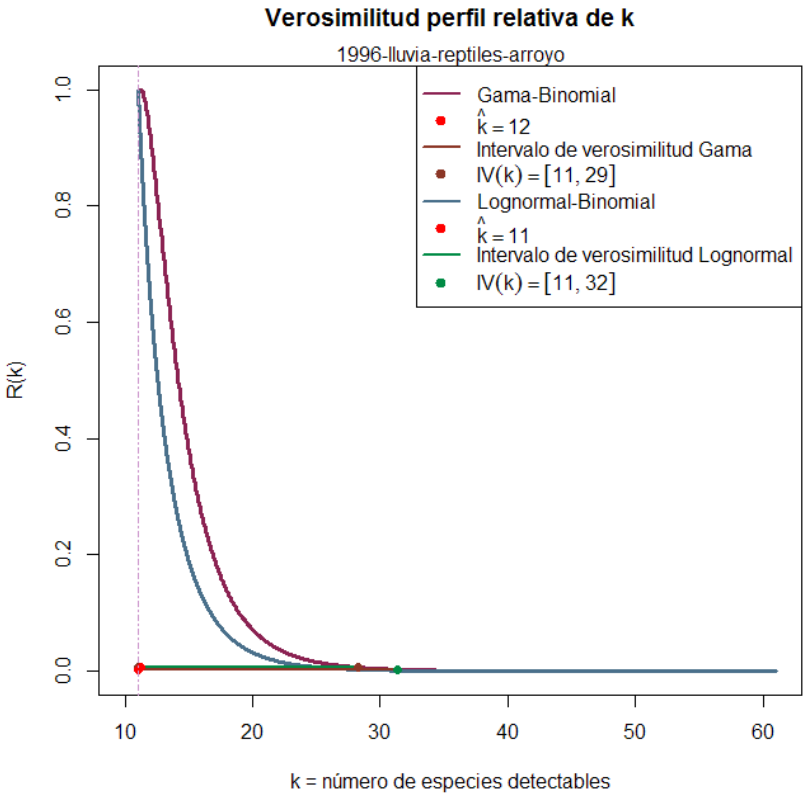
(4.5.7)

Tabla 4.5.7: Lluvia acumulada en mm por periodos del año hidrológico.
La última columna muestra los intervalos de estimación que se proponen para k.



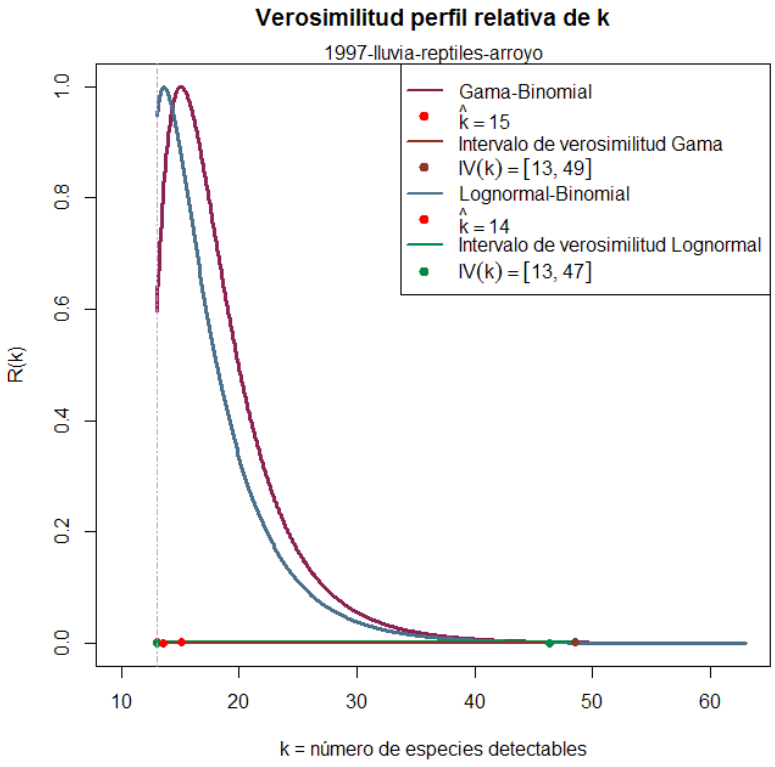
(4.5.8)

Figura 4.5.8**:** Funciones de verosimilitud perfil relativas para k obtenidas bajo el modelo Gama (línea azul) y el modelo Lognormal (línea violeta) para los datos de 1995.



(4.5.9)

Figura 4.5.9**:** Funciones de verosimilitud perfil relativas para k obtenidas bajo el modelo Gama (línea azul) y el modelo Lognormal (línea violeta) para los datos de 1996.



(4.5.10)

Figura 4.5.10**:** Funciones de verosimilitud perfil relativas para k obtenidas bajo el modelo Gama (línea azul) y el modelo Lognormal (línea violeta) para los datos de 1997.

Capítulo 5

Conclusiones generales

Mediante simulaciones se comprobó que al aumentar el número de cuadrantes en la muestra se mejora la cobertura de los intervalos de verosimilitud. Además, resulta más adecuado considerar una muestra donde se considere más de un cuadrante para compensar que usualmente no se cumple estrictamente el supuesto de homogeneidad sobre toda la región de interés.

En general, se recomienda que el número de cuadrantes represente al menos el 1% de la superficie total de la región de interés. En el caso de Chamela, esto implica considerar al menos 25 cuadrantes en la muestra. Con este tamaño de muestra se observó que los porcentajes de cobertura de los intervalos para estimar los parámetros del modelo estadístico son mejores y por tanto, el nivel de verosimilitud calibrado no es tan pequeño.

Se sugiere usar el método aquí presentado para construir gráficas Q-Q para datos con censura por intervalo provenientes de una variable aleatoria continua. En las tres aplicaciones, las gráficas Q-Q mostraron arreglos de puntos alrededor de la recta de 45° y dentro de las nubes de puntos que representan la variabilidad natural del modelo estimado; por lo que ambos modelos Gama y Lognormal resultaron razonables para describir a los datos.

A través de simulaciones con diferentes valores de los parámetros del modelo estadístico se observó lo siguiente:

a) En el caso de la distribución Gama, los estimadores de máxima verosimilitud de α suelen

sobre estimar a este parámetro. Por tanto, si en una muestra dada se obtiene un estimador de alfa menor que uno, entonces es muy probable que el verdadero valor de este parámetro sea pequeño. En consecuencia, la densidad Gama será cóncava. Una forma cóncava en la densidad Gama estimada indica la existencia de muchas especies raras (aquellas representadas por pocos individuos en la población), mientras que una densidad con moda indica que es posible observar especies con abundancias altas.

b) Los modelos Gama y Lognormal considerados para modelar las intensidades Poisson resultaron indistinguibles, en el sentido de que no se prefirió más uno que otro en las muestras simuladas. Esto se comprobó mediante el cálculo de la razón de verosimilitudes para cada muestra simulada bajo cada uno de los modelos. Los valores de la razón de verosimilitudes son en su mayoría pequeños y un modelo se prefiere en proporciones similares al otro.

c) Las coberturas de intervalos de verosimilitud que se obtienen al considerar el modelo Gama suelen ser mejores que las obtenidas bajo el modelo Lognormal, en especial para el parámetro de interés k. Además, como se observó en el Capítulo 4, el intervalo obtenido bajo el modelo Gama siempre incluye valores mayores para k que se descartan bajo el model Lognormal. Así, se sugiere usar el modelo Gama, pues el intervalo de estimación será más conservador e incluirá valores de k razonables.

En los tres años considerados en las aplicaciones, las formas de las densidades Gama y Lognormal obtenidas para un mismo juego de datos fueron similares en las colas pero discreparon en las probabilidades que asignan a valores pequeños de los parámetros de intensidad Poisson. Cuando el parámetro de forma de la distribución Gama es menor que uno, hay más discrepancia con el modelo Lognormal correspondiente (como es el caso de los años 1995 y 1997), sin embargo, las colas de las ambas densidades fueron similares.

En todos los casos, el modelo Gama dio mayor probabilidad a especies raras, en contraste con el modelo Lognormal. Esto explica porqué las verosimilitudes perfiles Gama favorecen también valores mas grandes del total de especies detectables en la region, y por tanto, porqué

los intervalos de estimación son mas amplios. Frente a la incertidumbre natural que se suele tener, será preferible el modelo Gama sobre el Lognormal debido a que da inferencias más conservadoras, además de que en estos ejemplos ajustó muy bien a los datos considerados. Es decir, la recomendación que se hace aquí a favor del modelo Gama está condicionada a que dicho modelo ajuste bien a los datos observados t_{jr}, para $j = 1, ..., m$.

Finalmente, es importante señalar que los modelos estadísticos propuestos en esta tesis no se limitan a la estimación del número de especies detectables y a su uso en problemas de ecología. En general, si se está interesado en estimar el número de clases (especies) que existen en una población en la que puede llevarse a cabo un muestreo por cuadrantes, de tal forma que se colecte información sobre la identidad y abundancia de los individuos (a qué clase pertenecen y cuántos individuos de la misma clase se observan en cada cuadrante); entonces la metodología estadística aquí propuesta puede utilizarse para estimar el número de clases, siempre que se cumplan los supuestos establecidos en el planteamiento del modelo estadístico de manera equivalente según la naturaleza del problema de interés.

Apéndice A

Intervalos de verosimilitud-confianza para parámetros de intensidad Poisson

Para una especie observada $j = 1, ..., m$, en un muestreo por cuadrantes de tamaño r, estimamos su intensidad Poisson λ_j usando el EMV que se obtiene con los r conteos Poisson observados en un cuadrante X_{ij}, para $i = 1, ..., r$,

$$T_{rj} = \widehat{\lambda_j} = \frac{1}{r}\sum_{i=1}^{r} X_{ij} = \overline{X}_j. \tag{A.0.1}$$

Dada la muestra observada de conteos para la especie j en los r cuadrantes, $x_{1j}, x_{2j}, ..., x_{rj}$, la verosimilitud para λ_j es

$$\begin{aligned}
\mathcal{L}(\lambda_j; x_{1j}, x_{2j}, ..., x_{rj}) &= C(x_{1j}, ..., x_{rj})\,\mathbf{P}[X_{1j,} = x_{1j}, X_{2j} = x_{2j}, ..., X_{rj} = x_{rj}] \quad \text{(A.0.2)} \\
&= C(x_{1j}, ..., x_{rj}) \prod_{i=1}^{r} \mathbf{P}[X_{ij,} = x_{ij}] \\
&= C(x_{1j}, ..., x_{rj}) \prod_{i=1}^{r} \frac{\lambda_j^{x_{ij}} e^{-\lambda_j}}{x_{ij!}} \\
&= \lambda_j^t e^{-r\lambda_j},
\end{aligned}$$

para $\lambda_j > 0$. Donde $t = \sum_{i=1}^{r} x_{ij}$ es una estadística suficiente-minimal para λ_j y la función

C que solo depende de la muestra es

$$C\left(x_{1j},...,x_{rj}\right)=\prod_{i=1}^{r}\left(x_{ij!}\right)^{-1}.$$

Luego, la log-verosimilitud para λ_j y la función Score, respectivamente, tienen las siguientes expresiones

$$\ell\left(\lambda_j;x_{1j},x_{2j},...,x_{rj}\right)=t\log\lambda_j-r\lambda_j, \tag{A.0.3}$$

$$S_c\left(\lambda_j\right)=\frac{\partial\ell\left(\lambda_j\right)}{\partial\lambda_j}=\frac{t}{r}-r, \tag{A.0.4}$$

Resolviendo la ecuación

$$S_c\left(\lambda_j\right)=0,$$

se obtiene el EMV de λ_j, tal y como se expresó en $A.0.1$, $\widehat{\lambda_j}=\overline{X}_j$.

La función de verosimilitud relativa para λ_j es entonces

$$\begin{aligned}R\left(\lambda_j\right) &= \frac{\mathcal{L}\left(\lambda_j;x_{1j},x_{2j},...,x_{rj}\right)}{\mathcal{L}\left(\widehat{\lambda_j};x_{1j},x_{2j},...,x_{rj}\right)} \\ &= \frac{\lambda_j^t e^{-r\lambda_j}}{\left(\overline{X}_j\right)^t e^{-r\overline{X}_j}}=\left(\frac{\lambda_j}{\overline{X}_j}\right)^t e^{-r\left(\lambda_j-\overline{X}_j\right)},\lambda_j>0.\end{aligned} \tag{A.0.5}$$

Así, la log-verosimilitud relativa está dada por

$$r\left(\lambda_j\right)=t\left(\log\lambda_j-\log\overline{X}_j\right)-r\left(\lambda_j-\overline{X}_j\right),\lambda_j>0. \tag{A.0.6}$$

Dado un nivel de verosimilitud $c\in(0,1)$, el intervalo de verosimilitud de nivel c está dado por el conjunto

$$\begin{aligned}IV\left(c\right) &= \left\{\lambda:R\left(\lambda\right)\geq c\right\} \\ &= \left\{\lambda:r\left(\lambda\right)\geq\log c\right\} \\ &= \left\{\lambda:t\left(\log\lambda_j-\log\overline{X}_j\right)-r\left(\lambda_j-\overline{X}_j\right)\geq\log c\right\} \\ &= \left[\lambda_1,\lambda_2\right].\end{aligned} \tag{A.0.7}$$

Los extremos de tal intervalo deben obtenerse numéricamente pues dada la expresión de la verosimilitud relativa $(A.0.5)$ no existe una expresión algebraica cerrada para expresarlos en términos de la muestra. Para ello, basta restar el logaritmo natural del nivel c de

verosimilitud a la log-verosimilitud relativa y calcular numéricamente las raíces de la función resultante. Esto es, se buscan los valores de λ que satisfagan

$$r(\lambda) - \log c = 0. \tag{A.0.8}$$

Apéndice B

Demostración de la Proposición 1

Proposición 1 *Sean Λ una variable aleatoria continua que toma valores positivos y X_{rj} variables aleatorias, $r, j \in \{1, .2, ...\}$.*

Supongamos que para cada $j \in \{1, .2, ...\}$ y para cualquier $r \in \{1, .2, ...\}$, condicionalmente a $\Lambda = \lambda_j$, las v.a. $X_{1j}, ..., X_{rj}$ son independientes e idénticamente distribuidas como Poisson con media λ_j, $\mathbf{E}[X_{1j}|\Lambda = \lambda_j] = \lambda_j$, donde $\lambda_j \in (0, \infty)$.

Definamos

$$T_{rj} := \frac{1}{r} \sum_{i=1}^{r} X_{ij}, r, j \in \{1, .2, ...\}.$$

Entonces, de forma no condicional y para cualquier $j \in \{1, .2, ...\}$, T_{rj} converge en distribución a Λ cuando $r \to \infty$, esto es,

$$T_{rj} \underset{r \to \infty}{\overset{d}{\to}} \Lambda.$$

Demostración Sea $F_{T_{rj}}(t)$ la función de distribución no condicional de $T_{rj}, r \in \{1, .2, ...\}$.

Bajo los supuestos de la proposición, el objetivo es probar que

$$F_{T_{rj}}(t) \underset{r \to \infty}{\to} F_{\Lambda}(t; \theta), \forall t \in \mathbb{R},$$

donde $F_{\Lambda}(t; \theta)$ es la función de distribución de la v.a. Λ.

Por la Ley Débil de los Grandes Números, el promedio T_{rj}, condicionado a $\Lambda = \lambda_j$, converge en probabilidad a $\mathbf{E}[X_{1j}|\Lambda = \lambda_j] = \lambda_j$,

$$T_{rj}|\Lambda = \lambda_j \xrightarrow[r\to\infty]{p} \lambda_j. \tag{B.0.1}$$

Se sigue de este resultado que $T_{rj}|\Lambda = \lambda_j \xrightarrow[r\to\infty]{d} \lambda_j$,ya que convergencia en probabilidad implica convergencia en distribución.

Así, la distribución límite corresponde a una v.a. degenerada en λ_j, y se tiene la siguiente relación

$$\begin{aligned} \lim_{r\to\infty} F_{T_{rj}|\Lambda}(t|\lambda_j) &= \lim_{r\to\infty} \mathbf{P}[T_{rj} \leq t|\Lambda = \lambda_j] \\ &= \mathbf{1}_{[\lambda_j,\infty)}(t) = \begin{cases} 1, t \geq \lambda_j \\ 0, t < \lambda_j \end{cases}. \end{aligned} \tag{B.0.2}$$

Sean $t \in \mathbb{R}$ fija y $r \in \{1, .2, ...\}$. Considérese la v.a. Bernoulli $I_{(-\infty,t]}(T_{rj})$, la cual toma el valor 1 con probabilidad $\mathbf{P}[T_{rj} \leq t] = F_{T_{rj}}(t)$, y toma el valor 0 con probabilidad $\mathbf{P}[T_{rj} > t] = 1 - F_{T_{rj}}(t)$. De esta manera, la distribución no condicional deseada, $F_{T_{rj}}(t)$, es equivalente al valor esperado de la variable aleatoria $I_{(-\infty,t]}(T_{rj})$,

$$F_{T_{rj}}(t) = \mathbf{P}[T_{rj} \leq t] = \mathbf{E}\left[I_{(-\infty,t]}(T_{rj})\right]. \tag{B.0.3}$$

Además, el valor esperado condicional de la v.a. Bernoulli $I_{(-\infty,t]}(T_{rj})$, dado $\Lambda = \lambda_j$, es equivalente a la distribución condicional de T_{rj}, dado $\Lambda = \lambda_j$. Esto es,

$$F_{T_{rj}|\lambda_{jj}}(t|\lambda_j) = \mathbf{P}[T_{rj} \leq t|\Lambda = \lambda_j] = \mathbf{E}\left[I_{(-\infty,t]}(T_{rj})|\Lambda = \lambda_j\right] \tag{B.0.4}$$

Ahora bien, usando la propiedad de la esperanza condicional conocida como la Ley de la Esperanza Total (LET), se tiene que

$$\mathbf{E}\left[I_{(-\infty,t]}(T_{rj})\right] = \mathbf{E}_\Lambda\left[\mathbf{E}\left[I_{(-\infty,t]}(T_{rj})|\Lambda = \lambda_j\right]\right]. \tag{B.0.5}$$

De esta manera,

$$\begin{aligned} F_{T_{rj}}(t) &= \mathbf{P}\left[T_{rj} \leq t\right] \\ &\overset{(B.0.3)}{=} \mathbf{E}\left[I_{(-\infty,t]}\left(T_{rj}\right)\right] \\ &\overset{(B.0.5)}{=} \mathbf{E}_{\Lambda}\left[\mathbf{E}\left[I_{(-\infty,t]}\left(T_{rj}\right)|\Lambda = \lambda_j\right]\right] \\ &\overset{(B.0.4)}{=} \mathbf{E}\left[F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right)\right]. \end{aligned} \tag{B.0.6}$$

Nótese que la justificación de cada igualdad se debe a la ecuación indicada sobre el símbolo " $=$ ".

Dado que el límite cuando $r \to \infty$ de $F_{T_{rj}|\Lambda}\left(t|\lambda_j\right)$ es $\mathbf{1}_{[\lambda_j,\infty)}\left(t\right)$, por $(B.0.2)$, y la función $F_{T_{rj}|\Lambda}$ es acotada por ser función de distribución, es decir, $|F_{T_{rj}|\Lambda}\left(t|\lambda_j\right)| \leq 1$; aplicando el Teorema de Convergencia Dominada para Esperanzas se tiene lo siguiente

$$\begin{aligned} \lim_{r\to\infty} \mathbf{E}_{\Lambda}\left[F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right)\right] &= \int_{-\infty}^{\infty} F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right) f\left(\lambda_j;\theta\right) d\lambda_j \\ &= \int_{-\infty}^{\infty} \mathbf{1}_{[\lambda_j,\infty)}\left(t\right) f\left(\lambda_j;\theta\right) d\lambda_j = \mathbf{E}_{\Lambda}\left[\lim_{r\to\infty} F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right)\right]. \end{aligned} \tag{B.0.7}$$

Así, al poder intercambiar el límite con la esperanza, el resultado práctico es el siguiente

$$\lim_{r\to\infty} \mathbf{E}_{\Lambda}\left[F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right)\right] = \int_{-\infty}^{\infty} \mathbf{1}_{[\lambda_j,\infty)}\left(t\right) f\left(\lambda_j;\theta\right) d\lambda_j. \tag{B.0.8}$$

Entonces, tomando el límite cuando $r \to \infty$ en $(B.0.6)$ y usando $(B.0.7)$, obtenemos

$$\lim_{r\to\infty} F_{T_{rj}}\left(t\right) \overset{(B.0.6)}{=} \lim_{r\to\infty} \mathbf{E}_{\Lambda}\left[F_{T_{rj}|\lambda_{jj}}\left(t|\lambda_j\right)\right] \overset{(B.0.8)}{=} \int_{-\infty}^{\infty} \mathbf{1}_{[\lambda_j,\infty)}\left(t\right) f\left(\lambda_j;\theta\right) d\lambda_j$$

$$= \int_{-\infty}^{\infty} \mathbf{1}_{(-\infty,t]}\left(\lambda_j\right) f\left(\lambda_j;\theta\right) d\lambda_j = \int_{-\infty}^{t} f\left(\lambda_j;\theta\right) d\lambda_j = F_{\Lambda}\left(t;\theta\right), t \in \mathbb{R},$$

donde la tercera igualdad se obtiene de expresar la indicadora $\mathbf{1}_{[\lambda_j,\infty)}\left(t\right)$ como función de λ_j.

Por lo tanto,

$$\lim_{r\to\infty} F_{T_{rj}}(t) = F_{\Lambda}(t;\theta), \forall t \in \mathbb{R},$$

como se quería demostrar.

■

Observaciones

O1 Es importante señalar que el resultado práctico que se usa para plantear el modelo estadístico presentado en la sección 2.3, es que para r suficientemente grande, se tiene la relación

$$F_{T_{rj}}(t) \approx F_{\Lambda}(t;\theta), t \in \mathbb{R}.$$

En particular, esta aproximación fue usada para expresar cada término de la verosimilitud asociado a un promedio de individuos observados de una especie detectada en la muestra.

De acuerdo a las simualciones realizadas en este trabajo, la relación anterior se satisface para $r \geq 5$. Es decir, con un número pequeño de cuadrantes la aproximación ya es razonable. Este valor de r pequeño puede deberse al supuesto de que las variables aleatorias T_{rj} son promedios de v.a. Poisson ya que esta distribución posee ciertas propiedades de regularidad que hacen que algunos resultados asintóticos se satisfagan con muestras pequeñas. Si omitimos este supuesto necesitaríamos averiguar (a través de simulaciones o resultados teóricos) cuál es la r suficientemente grande tal que el resultado de la proposición se satisface.

O2 La importancia de tener a las v.a. T_{rj} definidas como promedios de v.a.i.i.d. consiste en que de esta manera la Ley Débil de los Grandes Números nos proporciona el resultado $(B.0.1)$. Nótese que no es necesario que las v.a. $X_{ij}|\Lambda = \lambda_j$ sigan una distribución

Poisson, en general, podrían tener otra distribución. Se trabajó con la distribución Poisson porque es la distribución que surge naturalmente en el contexto de este trabajo.

Por tanto, en general, basta pedir que las v.a. T_{rj} satisfagan lo siguiente

$$T_{rj}|\Lambda = \lambda_j \underset{r\to\infty}{\overset{d}{\to}} \lambda_j,$$

donde $\mathbf{E}\left[X_{1j}|\Lambda = \lambda_j\right] = \lambda_j$, para que el resultado de la Proposición 1 siga siendo válido. Más aún, la prueba de este resultado es la misma que la presentada anteriormente.

Bibliografía

Boulinier, T., et al; 1998. *Estimating species richness: The importance of heterogeneity in species detectability. Ecology,* Vol. 79 : 1018 − 1028.

Brose, U.; Martínez, N.D. y Williams, R.J.; 2003. *Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. Ecology, Vol.* 84 : 2364 − 2377.

Hubbell, S.P.; 2001. *The unified neutral theory of biodiversity and biogeography.* Princeton University Press.

Figueroa, G.P.; 2012. *Las funciones de verosimilitud discretizada y restringida perfil en la inferencia científica.* Tesis de Doctorado. Programa de Doctorado en Ciencias de la Universidad de Sonora, División de Ciencias Exactas y Naturales. Hermosillo, Sonora, México.

Fisher, R.A.; Corbet, A. S. y Williams, C.B.; 1943. *The relation between the number of species and the number of individuals in a random sample of animal population.* Journal of Animal Ecology, Vol. 12 : 42 − 58.

Jiménez, L.J.; 2011. *Modelos estadísticos para la riqueza y abundancia de especies.* Tesis de Licenciatura. Programa de Licenciatura en Matemáticas, Depto. de Matemáticas, Universidad de Guanajuato. Guanajuato, Gto. México.

Kalbfleisch, J.G.; 1985. *Probability and Statistical Inference.* Volumen 2: Statistical inference. Segunda Edición. Springer-Verlag.

Magurran, A. E.; 2011. *Biological Diversity: frontiers in measurement* and assessment. Oxford University Press.

Mao, C.X. y Colwell, R.K.; 2005. *Estimating species richness: Mixture models, the roles of rare species, and inferential challenges.* Ecology, Vol. 86 : 1143 − 1153.

Montoya, J.A.; 2008. *La verosimilitud perfil en la inferencia estadística.* Tesis de Doctorado. Programa de Doctorado en Ciencias con Especialidad en Probabilidad y Estadística, CIMAT. Guanajuato, Gto. México.

Patil, G.P. y Taillie, C.; 1982. *Diversity as a concept and its measurement.* Journal of the American Statistical Association, Vol. 77 : 548 − 561.

Pawitan, Y.; 2001.*In all likelihood: Statistical modeling and inference using likelihood.* Oxford University Press.

Pielou, E.C.; 1969. *An introduction to mathematical ecology.* John Wiley and Sons, New York.

Seber, G.A.F.; 1982. *The estimation of animal abundance and related parameters.* London: Charles Griffin.

Seber G.A.F.; 1986. *A review of estimating animal abundance.* Biometrics. Vol.42 : 267−292.

Seber G.A.F.; 1992. *A review of estimating animal abundance II.* International Statistical Review. Vol. 60, No. 2 : 129 − 166.

Soberón, J. y Llorente, J.; 1993. *The use of species accumulation functions for the prediction of species richness.* Conservation Biology, Vol. 7 : 480 − 488.

Sprott, D.A.; 2000. *Statistical inference in science.* Springer series in Statistics.