FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado:

Muestreo de captura-recaptura: Diseño, estimación y análisis de librerías en R

Autor:

Manuel Pozo Barbero

Dirigido por:

Ana María Muñoz Reyes

2017 - 2018

Resumen

En esta memoria se estudian distintos métodos clásicos de captura-recaptura, válidos para la estimación del tamaño de poblaciones y algunas características de las mismas, como las tasas de supervivencia o de mortalidad. Para clasificar los diferentes métodos, se tiene en cuenta básicamente el tipo de población que estamos considerando; esta puede ser cerrada, caso en el que asumimos que el tamaño de la población no cambia en el tiempo de estudio, o bien puede ser abierta, caso en el que se acepta que dicho tamaño sí puede variar. Para finalizar el trabajo, se proponen librerías del software R en las que podemos encontrar funciones para trabajar con estos métodos. Además, de entre estas librerías se toma la librería Rcapture y se ejecutan en ella varias funciones, utilizando distintos conjuntos de datos ya implementados en dicha librería.

Abstract

This memory studies several capture-recapture methods, which can be used to estimate the size of a population and some of its characteristics, like the survival or the mortality rate. To classify the different methods, it is basically taken into account the type of population that is being considered; it could be a closed population or an open population. In the first case, it is assumed that the population size does not change during the time of the study while, in the second case, the population size could modify. Finally, it is proposed some libraries in software R where different functions to work with these methods can be found. Moreover, it is chosen the Rcapture library among the R libraries where several functions are executed, using distinct datasets implemented in this library.

Índice

1. Introducción

Los métodos de captura-recaptura son técnicas de muestreo estadístico que sirven para estimar el tamaño de una población, así como para el estudio de ciertos aspectos de la misma como pueden ser las tasas de natalidad, mortalidad, inmigración o emigración. Esto supone una gran utilidad para poblaciones en las que el tamaño poblacional es muy grande y no se puede hacer un censo total de los individuos, así como para aquellas en las que hay movimiento constante. Así, por ejemplo, mediante los métodos que aquí se plantean se puede estimar tanto el tamaño poblacional de una ciudad, en la que los efectos de migración en un corto período de tiempo puede ser despreciable, como que el tamaño de una pequeña población de aves, en las que posiblemente haya continuas idas y venidas de nuevos individuos.

El muestreo de captura-recaptura consiste, básicamente, en extraer una muestra aleatoria de una población de interés, marcar mediante algún método apropiado a cada miembro de la muestra, y volver a dejar libres a los individuos de esa muestra en la población. Pasado un tiempo, se vuelve a tomar una muestra aleatoria de la misma población y se observa cuántos individuos de esta nueva muestra están marcados, es decir, nos fijamos en qué fracción de los individuos en esta segunda muestra ya fueron también seleccionados la primera vez. Mediante la realización de este experimento una o varias veces, se puede estimar el tamaño de la población sobre la que estemos trabajando. Además, a partir de los modelos más simples, se pueden construir modelos algo más sofisticados que sirvan para explicar no sólo la abundancia de la población, si no que nos permitan estudiar cómo varía nuestra población en el tiempo a partir de las estimaciones que podemos hacer de las tasas de natalidad y mortalidad y emigración e inmigración.

La primera vez que se conoce que se haya usado estos métodos es en 1662, por

John Graunt, para estimar el tamaño de la población londinense. De la misma manera, también se usó por C.G.Petersen en 1896 para estimar el tamaño de una población de peces y medir tasas de mortalidad. En 1930, F.C.Lincoln usó también este método para estimar una población de aves silvestres en USA. También el conocido matemático Laplace, en 1793, usó esta misma idea para estimar el tamaño de la población francesa.

En estos cuatro casos se usó el mismo procedimiento, que sigue una idea bastante simple:

En primer lugar, se toma una muestra (captura) de M elementos que son marcados mediante algún procedimiento y, a continuación, son devueltos a la población. Seguidamente, se toma una nueva muestra (recaptura) de tamaño n en la que se obtienen m elementos marcados. Así, mediante argumentos lógicos de proporcionalidad ("si hay m elementos marcados en una población de tamaño m, entonces M elementos marcados corresponderán a una población de tamaño N), podríamos decir que, si N es el tamaño total de la población,

$$\frac{N}{M} \approx \frac{n}{m}$$

Por lo tanto, mediante un simple despeje de N, podemos aproximar el tamaño poblacional a través del siguiente estimador:

$$\widehat{N} = M\frac{n}{m}$$

Esto se conoce como ecuación de Lincoln-Petersen, la cual estudiaremos con más detenimiento más adelante.

Estos métodos han sido fuertemente potenciados a lo largo de la segunda mitad del siglo XX gracias, sobretodo, a la importancia que han tenido en muchos campos

científicos como pueden ser la biología, la ecología, las ciencias sociales, la medicina o la epidemiología.

Para el tratamiento de este tipo de métodos, una característica de suma importancia a considerar es si la población es abierta o cerrada. Cuando hacemos el estudio considerando que la población es cerrada, estamos asumiendo que en el tiempo de estudio no se altera en ningún momento el tamaño de la población. Es decir, no hay ni nacimientos, ni muertes, ni migración o, en caso de haberlos, son despreciables. Este supuesto, aunque nos facilita mucho el estudio y desarrollo del método, es poco realista, ya que estos métodos requieren de mucho tiempo y esfuerzo, por lo que en el período de tiempo que dure nuestra experiencia, que puede no ser corto, no es complicado imaginar que el tamaño de la población pueda variar. Los métodos cerrados, por tanto, son más apropiados para tamaños poblacionales que sean pequeños y en los que la población en cuestión no tenga muchos efectos migratorios. Si, por el contrario, nuestra población tiene un tamaño considerablemente grande o tiene una buena proporción de efectos migratorios, entonces nos podemos plantear el desarrollo de métodos abiertos, en los cuales aceptamos que pueda haber cambios en el tamaño de la población durante el tiempo de estudio. Como es lógico, estudiar el tamaño poblacional considerando que nuestra población es abierta, nos lleva a un caso bastante más realista que el anterior, aunque, por otro lado, tienen el incoveniente de que hace que nuestro estudio y desarrollo sean, por regla general, considerablemente más complejos. No obstante, ambos métodos requieren de varios requisitos, en algunas ocasiones muy estrictos, así como de una gran inversión de tiempo y esfuerzo.

En este trabajo se abarcará el estudio de varios modelos para ambos tipos de métodos. En primer lugar, se estudiará el modelo de Petersen o modelo hipergeométrico, que es el caso más sencillo de todos los que se verán a lo largo de esta memoria. Este modelo se aplica en poblaciones cerradas y hace uso de

un único marcaje. El primer estimador que se propone en este método es el que ya ha sido expuesto en líneas precedentes. Dicho estimador, desafortunadamente, es insesgado, por lo que en el estudio que se haga del método se propondrá uno que no difiera mucho numéricamente del ya propuesto pero que sí sea insesgado. El hecho de hacer un único marcaje hará que su estudio no sea complicado, pero debemos tener en cuenta que también influirá negativamente en cómo se ajuste nuestro estimador a la realidad. Es por eso que, acto seguido, como segundo método de poblaciones cerradas, se estudiará el modelo de Schnabel, el cual difiere del primero en el hecho de que se toman varias recapturas en lugar de tan solo una (aunque todas ellas serán marcadas con el mismo procedimiento), y esto puede ser conveniente para tener una aproximación más ajustada. Este método consistirá en tomar una captura y marcar los elementos; en la primera recaptura, observaremos cuántos elementos hay marcados, y marcaremos los que no lo estén; en la tercera recaptura, observaremos los elementos que están marcados (es decir, que ya se han tomado en la primera y/o en la segunda muestra) y marcaremos los que no estén marcados. Procederemos así hasta las s capturas que hagamos de la población. Schnabel nos proporcionará un estimador que es una extensión del primer estimador propuesto en el método de Petersen-Lincoln. Este estimador es

$$\widehat{N} = \frac{\sum_{t=1}^{s} C_t M_t}{\sum_{t=1}^{s} R_t}$$

Se ve fácilmente que es una extensión del estimador antes mencionado, ya que en el caso en el que $s = 1$ (se toma una única recaptura), $\widehat{N}$ coincide con el estimador de Petersen-Lincoln. En el método de Schnabel veremos también dos modelos distintos en función de cómo sea la probabilidad de captura en las distintas recapturas: se puede suponer que la probabilidad de captura es la misma en todas las recapturas,

lo que nos llevará al modelo $\mathbf{M}_0$, o suponer por el contrario que esta probabilidad es distinta en cada recaptura, lo que nos llevará al modelo $\mathbf{M}_t$.

Como ejemplo de modelo para poblaciones abiertas, se estudiará el modelo de Jolly-Seber, en el que se necesitará tomar muestras de recaptura en tres o más ocasiones. Es decir, realmente se verá este método como una extensión del método de Schnabel. En este método, será necesario definir bastantes parámetros, ya que en este punto nuestra población comienza a considerarse como población abierta y, por tanto, se tendrán en cuenta más factores para el desarrollo de los modelos. Una vez visto el modelo de Jolly-Seber, veremos dos aplicaciones directas de este modelo, que corresponderán a la migración en la población y a las tasas de natalidad. Se verá cómo las estimaciones hechas en el modelo de Jolly-Seber se pueden utilizar para hacer estimaciones de otros parámetros que pueden resultar de interés. No obstante, tanto en los métodos cerrados como en los abiertos, se propondrán variantes y nuevos modelos que se obtengan de los que ya se hayan propuesto.

Tras haber visto ejemplos de modelos para poblaciones cerradas y abiertas, se nos queda una pregunta abierta: ¿De entre todos los modelos que puedo elegir (los aquí propuestos y todos los demás existentes)para realizar un estudio, con cuál de ellos me quedo? Esta pregunta la trataremos brevemente tras haber expuesto el modelo de Jolly-Seber. Se verá que existen dos métodos distintos para seleccionar el modelo que mejor se ajusta a la realidad. Este apartado será de utilidad para la última parte de esta memoria.

Para finalizar, se proponen varias librerías de R que sirven para hacer estudios de captura-recaptura, cada una de ellas con un enfoque distinto. Tomaremos la librería Rcapture, que se centra en modelos log-lineales de captura-recaptura. Veremos que dicha librería contiene dos funciones específicas que nos valen para realizar nuestras estimaciones: closedp y openp, en función de si estamos considerando que nuestra población es abierta o cerrada. Veremos también que estas funciones pueden tomar

como argumento conjuntos de datos con distinta estructura, que serán estudiados. Se utilizarán distintos conjuntos de datos para ver cómo se debe actuar con esta librería frente a distintas estructuras del conjunto de datos.

2. Métodos para poblaciones cerradas

2.1. Modelo hipergeométrico o de Petersen-Lincoln

2.1.1. Descripción general del modelo. Hipótesis necesarias.

En este apartado se estudiará el modelo de Petersen-Lincoln o hipergeométrico, el cual, como ya hemos dicho previamente, es el más sencillo de todos los que se tratarán. Este método implica una única captura y una única recaptura, lo que explica la facilidad de su estudio.
Para el análisis de este modelo es necesario partir de las siguientes hipótesis:

- **H1**.La población es cerrada. Durante el tiempo de estudio asumimos que el tamaño poblacional N permanece constante o, si se produce una variación en dicho tamaño, esta es insignificante respecto al verdadero tamaño poblacional.
- **H2.** Todos los individuos de la población tienen la misma probabilidad de ser capturados en la primera muestra.
- **H3.** La segunda muestra es también una muestra aleatoria. Para la validez de esta hipótesis, necesitamos que los métodos con los que se marquen a los individuos no afecten a la capturabilidad, de modo que en la recaptura, un individuo que no está marcado tenga la misma probabilidad de ser capturado que uno que no lo está.
- **H4.** Todos los individuos marcados observados en la 2^{a} muestra deben ser incluidos en la misma.

Bajo estas hipótesis, se puede comenzar a plantear la estimación más simple que se nos pudiera ocurrir, la cual ya fue propuesta en la introducción de este mismo artículo. Considérese que:

M= Número de individuos seleccionados en la primera muestra.

C= Número de individuos tomados en la segunda muestra.

R= Número de individuos marcados que han sido seleccionados en la segunda muestra.

Entonces, si el tamaño total de la población es N, podemos hacer una estimación suponiendo que el cociente del número de individuos marcados en la primera muestra entre el tamaño total de la población es proporcional al cociente del número de individuos que están marcados en la segunda muestra entre el tamaño total de esa segunda muestra. Por tanto, se puede estimar el tamaño poblacional N por:

$$\widehat{N} = C\frac{M}{R}$$

Desafortunadamente, este estimador es sesgado y tiende a sobrestimar el tamaño real de la población, siendo mayor el sesgo cuanto más pequeño es dicho tamaño. La sesgadez se puede tratar, gracias a Seber (1982), en ciertos casos. Bajo una cierta condición, que se verá poco más adelante, se nos propone el siguiente estimador, que aunque numéricamente no difiere mucho del primero, sí que es insesgado:

$$\widehat{N} = \frac{(M+1)(C+1)}{R+1} - 1$$

2.1.2. Desarrollo del modelo. Estudio de la insesgadez y la varianza de los estimadores

Supongamos que se satisfacen las hipótesis **H1**, **H3** y **H4**. Sea X una variable aleatoria que se distribuye según una distribución hipergeométrica $H(N,M,C)$ donde:
X= número de elementos marcados en la segunda muestra, $0 \leq X \leq min(M, C)$
La distribución de probabilidad de esta distribución es:

$$\mathbb{P}[X = x|M, C] = f(x|M, C) = \frac{\binom{N-M}{C-x}\binom{M}{x}}{\binom{N}{C}}$$

Este hecho nos facilita el estudio de este modelo. Tenemos, en particular, las dos siguientes proposiciones:

Proposición 1. *En la distribución hipergeométrica H(N,M,C), conocidos el par (M, C), el estimador de máxima verosimilitud del tamaño poblacional N viene dado por la siguiente expresión:*

$$\widehat{N} = \left[M\frac{C}{x}\right], \ \ x \neq 0$$

donde [-] es la función parte entera.

Este estimador, aunque es bastante lógico y sencillo de deducir, tiene el inconveniente de que es sesgado. Para abarcar este asunto, se utilizará la siguiente proposición, que nos proporcionará un estimador insesgado en algunos casos
Proposición 2. *Se verifica que, si $X \sim H(N, M, C)$, entonces*

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \begin{cases} \frac{N+1}{(C+1)(M+1)}\left[1 - \frac{(N-M)!(N-C)!}{(N+1)!(N-M-C-1)!}\right] & si\ N \geq C+M+1 \\ \frac{N+1}{(C+1)(M+1)} & si\ N < C+M+1 \end{cases}$$

Notemos que de esta proposición se puede obtener un estimador insesgado en el caso en el que $N < C + M + 1$, que viene dado por la siguiente expresión

$$\widehat{N^*} = \frac{(C+1)(M+1)}{x+1} - 1$$

La insesgadez de este estimador se deduce fácilmente gracias a la anterior proposición y al hecho de que C y M son constantes:

$$\mathbb{E}\left[\widehat{N^*}\right] = (C+1)(M+1)\mathbb{E}\left[\frac{1}{X+1}\right] - 1 = (C+1)(M+1)\frac{N+1}{(C+1)(M+1)} - 1 = N$$

En el caso en el que $N \geq C + M + 1$, se podría aproximar a la insesgadez si el término $\frac{(N-M)!(N-C)!}{(N+1)!(N-M-C-1)!}$ fuese suficientemente pequeño, caso en el que $\mathbb{E}\left[\frac{1}{X+1}\right]$ tiende a ser siempre $\frac{N+1}{(C+1)(M+1)}$, sin distinción de casos.

Nos interesa tener también un intervalo de confianza que nos de información sobre los valores entre los que está comprendido el parámetro de interés. Para el cálculo de intervalo de confianza, primeramente se necesita conocer la varianza, algo que, en este caso, no es sencillo. Utilizando los desarrollos en series de Chapman(1951), se llega a que

$$var\left[\widehat{N^*}\right] \simeq N^2\left[\frac{N}{nM} + 2\left(\frac{N}{nM}\right)^2 + 6\left(\frac{N}{nM}\right)^3\right]^{1/3}$$

Una estimación de esta varianza viene dada por

$$\widehat{var}\left[\widehat{N^*}\right] = \frac{(n+1)(M+1)}{(x+1)^2}\frac{(n-x)(M-x)}{(x+2)}$$

Usando esta estimación, se tiene que un intervalo de confianza asintótico es

$$\widehat{N^*} \pm Z_{1-\alpha/2}\sqrt{\widehat{var}\left[\widehat{N^*}\right]}$$

donde $Z_{1-\alpha/2}$ denota el cuantil $1-\alpha/2$ de la distribución normal.

2.1.3. Análisis de las hipótesis iniciales

Hipótesis H1: El tamaño poblacional permanece constante en el tiempo de estudio.

Para que esta hipótesis sea válida, se debe realizar la experiencia en un tiempo que sea corto, ya que a largo plazo el tamaño poblacional puede variar considerablemente y puede que no sean útiles o se ajusten muy poco a la realidad los resultados que obtengamos. Hay varios imprevistos, aún así, que pueden presentarse:

- **1. Muertes accidentales**

  Si en el período de estudio se presenta alguna muerte accidental, el desarrollo teórico permanecería inalterable.

- **2. Muertes naturales**

  El caso que realmente afecta al estudio es que se produzcan muertes entre la toma de la primera y de la segunda muestra. Si se cumplen las hipótesis **H3** y **H4** se podría considerar que las muertes naturales son una muestra aleatoria de la población de la cual no conocemos el tamaño. Asimismo, las "no muertes" también se podrían considerar como otra muestra aleatoria de la cual tampoco se conoce el tamaño. Por tanto, la recaptura es una submuestra de una muestra aleatoria (las "no muertes"), y, por consiguiente, una muestra aleatoria de la población, por lo que las estimaciones hechas son válidas.

  Notemos que si las muertes no se considerasen una muestra aleatoria, se puede suponer que la probabilidad de supervivencia es ϕ en ese período de tiempo. Entonces, como se puede ver de manera intuitiva en la siguiente expresión, las estimaciones se pueden seguir considerando válidas:

$$\mathbb{E}\left[\frac{X}{n}|M\right] \backsimeq \frac{\phi M}{\phi N} = \frac{M}{N}$$

- **3. Población capturable**

Cuando se da una estimación de N, la estimación se está dando del total de la población capturable, no de toda la población completa. Si se tuviese una estimación $\widehat{p}$ de la proporción de la población capturable, entonces se podría estimar el total mediante

$$\widehat{N_T} = \frac{\widetilde{N}}{\widehat{p}}$$

Si las estimaciones $\widehat{N}$ y $\widehat{p}$ son independientes (es decir, están basadas en distintos experimentos), entonces se puede usar el método delta de aproximaciones en las estimaciones.

Para $f(\widehat{p}) = \frac{1}{\widehat{p}}$, se puede considerar el desarrollo en serie

$$\sum_{k=1}^{\infty} \frac{f^{k)}(t_0)}{k!}(t-t_0)^k \Rightarrow (t_0 = p) : f(\widehat{p}) = \frac{1}{p} + \sum_{k=1}^{\infty} (-1)^k \frac{1}{p^{k+1}}(\widehat{p}-p)^k$$

Así, se puede calcular la esperanza y la varianza. Debido a que $\widetilde{N}$ y $\frac{1}{\widehat{p}}$ son independientes y que $\widetilde{N}$ es insesgado de N,

$$\mathbb{E}\left[\widehat{N_T}\right] = \mathbb{E}\left[\widetilde{N}\right]\mathbb{E}\left[\frac{1}{\widehat{p}}\right] = N\mathbb{E}\left[\frac{1}{p} + \sum_{k=1}^{\infty}(-1)^k \frac{1}{p^{k+1}}(\widehat{p}-p)^k\right] \backsimeq$$

$$\backsimeq \frac{N}{p} - \frac{N}{p^2}\mathbb{E}\left[\widehat{p}-p\right] + \frac{N}{p^3}\mathbb{E}\left[(\widehat{p}-p)^2\right] = N\frac{1}{p} - N\frac{1}{p^2}0 + N\frac{1}{p^3}Var(\widehat{p}) =$$

$$= N_T + \frac{N_T}{p^2}Var(\widehat{p})$$

Análogamente para la varianza:

$$Var\left[\hat{N}_T\right] \backsimeq \frac{1}{p^2}Var(\hat{N}) + \frac{1}{p^2}N_T^2 Var(\hat{p})$$

- **4. Incremento poblacional**

  Si durante el transcurso del estudio se produce una incorporación de individuos,

bien sea por nacimientos, bien sea por inmigración, entonces:

$$N_{FIN} = N + K \Rightarrow \frac{M}{N+K} \backsimeq \frac{x}{n} \Rightarrow N + K \backsimeq \frac{CM}{x}$$

Por tanto, $\frac{nm_1}{x}$ es un estimador de $N+K$, y, en consecuencia, es una sobrestimación de N. Esto indica que, bien se produzcan solamente incrementos, bien se produzcan incrementos y decrementos simultáneamente, esta situación se debe modelar como un población abierta.

Hipótesis H2 y H3: Captura equiprobable.

La equiprobabilidad en la selección de las muestras, obviamente, no es siempre posible. Es necesario, por tanto, la realización de estudios previos que permitan estudiar la capturabilidad o detectabilidad y corregir sus defectos en los procesos de selección y/o estimación.

En el tiempo de estudio, puede ser que los elementos pierdan sus marcas, por lo que las recapturas marcadas pueden ser inferiores a las esperadas y, en consecuencia, esto llevará a una sobrestimación del tamaño total de la población.

Si se pierden en el tiempo de estudio k marcas:

$$\frac{M-k}{N} \backsimeq \frac{x}{n} \Rightarrow N \backsimeq \frac{n(M-k)}{x} = \frac{nM}{x} - \frac{nk}{x}$$

Si se perdiesen, en total, un porcentaje de $\alpha 100\,\%$ marcas

$$\frac{(1-\alpha)M}{N} \backsimeq \frac{x}{N} \Rightarrow N \backsimeq (1-\alpha)\frac{nM}{x}$$

Para evitar este problema, es muy importante hacer una experimentación previa para determinar qué marca debemos usar para no tener pérdidas de elementos marcados. Otra forma de actuar ante este problema es hacer una estimación del número de marcas perdidas. En este sentido, un procedimiento sencillo a seguir es

el siguiente:

En el proceso de captura se toma una muestra de tamaño M y se realiza una doble marca sobre cada individuo: marca A y marca B.

En el proceso de recaptura, se obtiene una muestra de tamaño n, y se identifican a los individuos según las marcas:

π_A: probabilidad de pérdida de la marca A

π_B: probabilidad de pérdida de la marca B

π_{AB}: probabilidad de pérdida de ambas marcas simultáneamente.

x_A: número de elementos marcados solo con A en la recaptura

x_B: número de elementos marcados solo con B en la recaptura

x_{AB}: número de elementos marcados con A y B en la recaptura

x: número de elementos de los M marcados inicialmente obtenidos en la recaptura

Si suponemos que hay independencia entre las marcas: $\pi_{AB} = \pi_A \pi_B$, entonces el modelo resultante será multinomial, siendo su función de probabilidad:

$$f(x_A, x_B, x_{AB}|n) = \frac{n!}{x_A! x_B! x_{AB}! x_0!} \left[(1-\pi_A)\pi_B\right]^{x_A} \left[\pi_A(1-\pi_B)\right]^{x_B} \left[(1-\pi_A)(1-\pi_B)\right]^{x_{AB}} \left[\pi_A \pi_B\right]^{x_0}$$

donde $x_0 = x - x_A - x_B - x_{AB}$

En un modelo multinomial, los estimadores de máxima verosimilitud coinciden con los estimadores por el método de los momentos, y dado que los parámetros desconocidos son $\{x, \pi_A, \pi_B\}$, entonces:

$\mathbb{E}\left[X_A|X = x\right] = x(1-\pi_A)\pi_B \Rightarrow x_A = \widehat{x}(1-\widehat{\pi}_A)\widehat{\pi}_B$

$\mathbb{E}\left[X_B|X = x\right] = x(1-\pi_B)\pi_A \Rightarrow x_B = \widehat{x}(1-\widehat{\pi}_B)\widehat{\pi}_A$

$\mathbb{E}\left[X_{AB}|X = x\right] = x(1-\pi_A)(1-\pi_B) \Rightarrow x_{AB} = \widehat{x}(1-\widehat{\pi}_A)(1-\widehat{\pi}_B)$

Con esto, tenemos un sistema de 3 ecuaciones con 3 incógnitas:

$$\left.\begin{array}{l} x_A = \widehat{x}\,\widehat{\pi}_B(1-\widehat{\pi}_A) \\ x_B = \widehat{x}\,\widehat{\pi}_A(1-\widehat{\pi}_B) \\ x_{AB} = \widehat{x}\,(1-\widehat{\pi}_A)(1-\widehat{\pi}_B) \end{array}\right\}$$

Despejando $\widehat{x}$ en la última ecuación, se tiene que

$$\widehat{x} = \frac{x_{AB}}{(1-\widehat{\pi}_A)(1-\widehat{\pi}_B)}$$

Sustituyendo esta expresión en las dos primeras ecuaciones se tiene que

$$x_A = \frac{x_{AB}}{(1-\widehat{\pi}_A)(1-\widehat{\pi}_B)}\widehat{\pi}_B(1-\widehat{\pi}_A) \Rightarrow x_A = \frac{x_{AB}}{1-\widehat{\pi}_B}\widehat{\pi}_B \Rightarrow \widehat{\pi}_B = \frac{x_A}{x_A + x_{AB}}$$
$$x_B = \frac{x_{AB}}{(1-\widehat{\pi}_A)(1-\widehat{\pi}_B)}\widehat{\pi}_A(1-\widehat{\pi}_B) \Rightarrow x_A = \frac{x_{AB}}{1-\widehat{\pi}_A}\widehat{\pi}_A \Rightarrow \widehat{\pi}_A = \frac{x_B}{x_B + x_{AB}}$$

Por último, si se sustituyen nuestras soluciones de $\widehat{\pi}_A$ y $\widehat{\pi}_B$ en la expresión de $\widehat{x}$, entonces

$$\widehat{x} = \frac{x_{AB}}{(1-\frac{x_B}{x_B+x_{AB}})(1-\frac{x_A}{x_A+x_{AB}})} \Longrightarrow \widehat{x} = \frac{x_{AB}}{\frac{x_{AB}x_{AB}}{(x_B+x_{AB})(x_A+x_{AB})}} \Longrightarrow$$

$$\Longrightarrow \widehat{x} = \frac{(x_B+x_{AB})(x_A+x_{AB})}{x_{AB}}$$

Por lo que la solución del sistema sería

$$\begin{cases} \widehat{\pi}_A = \dfrac{x_B}{x_B+x_{AB}} \\ \widehat{\pi}_B = \dfrac{x_A}{x_A+x_{AB}} \\ \widehat{x} = \dfrac{(x_B+x_{AB})(x_A+x_{AB})}{x_{AB}} \end{cases}$$

2.1.4. Variantes del modelo hipergeométrico

Se pueden encontrar pequeñas variantes del modelo hipergeómetrico que siguen teniendo cálculos y desarrollos sencillos pero que igualmente no nos proporcionan estimadores que se ajusten bien a los parámetros reales, así como tampoco nos proporcionan intervalos de confianza de pequeña longitud para niveles de confianza altos (90 % en adelante). No entraremos en los desarrollos de ninguna de estas variantes ya que son parecidos al modelo hipergeométrico. Una de estas principales variantes es el método positivo de Jackson. Para este método, se necesitan los mismos supuestos que para el modelo hipergeométrico. Básicamente, el método positivo de Jackson consiste en realizar una captura, marcar los elementos, y después realizar varias recapturas, sin marcar ninguna de ellas, y, tras cada recaptura, hacer una estimación mediante el modelo antes desarrollado. El beneficio de este método respecto del anterior, es que se realizan s recapturas y se pueden considerar entonces s experimentos distintos e independientes del modelo hipergeómetrico, lo que hará que, en base a los distintos resultados, se pueda llegar a un estimador más ajustado. Sin embargo, presenta también un gran inconveniente, y es que se necesita un tiempo de estudio considerablemente más largo, (este aumentará conforme aumente s) por lo que mantener algunas de nuestras hipótesis, como las relacionadas con la supervivencia en el tiempo de estudio o la equiprobabilidad de captura de la muestra, es bastante difícil de conseguir.

Como caso particular de este modelo, se encuentra el método de triple captura de Bailey, que se aplica en el caso concreto en el que tenemos el método de Jackson con s=3.

2.2. Método de Schnabel

2.2.1. Introducción. Modelo de Schnabel como extensión del modelo de Petersen Lincoln

Schnabel (1938) y Darroch (1958) extendieron el modelo de Petersen a una sucesiva serie de recapturas, ya que un modelo que trata tan solo una captura y una recaptura puede que no nos de un resultado fehaciente. Como consecuencia de esto, para el desarrollo de este segundo método cerrado, seguirá siendo necesario suponer que la población es cerrada, es decir, que durante el período de estudio el tamaño poblacional total N permanece constante. La idea básica de este método es hacer una captura de individuos y marcarlos mediante algún procedimiento. En las sucesivas recapturas, observaremos el número de individuos que ya hayan sido marcados y, los miembros de esa muestra que no estén marcados, los marcamos mediante el mismo procedimiento que en todas las muestras anteriores. Es decir, utilizamos un único tipo de marcaje. Este método, por tanto, difiere del método positivo de Jackson en el hecho de que aquí, los elementos que se tomen en las recapturas y no estén marcados se marcarán, mientras que en el método positivo de Jackson tan solo los observábamos porque el interés estaba en hacer s experimentos del modelo hipergeométrico. Con este nuevo método se sigue, no obstante, necesitando un único tipo de marca porque realmente nuestro interés está en distinguir dos tipos de individuos en la población:

- Marcados: Son aquellos individuos que ya han sido extraídos en alguna de las muestras anteriores.

- No marcados: Es la primera vez que ese individuo ha sido capturado.

Con esto se tendrá que, en la t-ésima recaptura

$$C_t = R_t + U_t$$

donde

• C_t= Número total de individuos capturados en la muestra t

• R_t= Número de individuos que están marcados en la muestra t

• U_t= Número de individuos que son marcados por primera vez en la muestra t

Por tanto, si se denota por M_t a los individuos de la población que son marcados antes de la muestra t, entonces, como la población es cerrada, $\dfrac{M_t}{N}$ será la probabilidad de capturar en la muestra t un individuo que ya está marcado, es decir, que ya ha sido seleccionado previamente. Además,

$$M_t = \sum_{k=1}^{t-1} U_k$$

es decir, el número de individuos marcados en la población es acumulativo a medida que se van extrayendo las demás muestras.

Con esto, Schnabel proporciona la siguiente ecuación de verosimilitud para estimar el tamaño poblacional. Si tomamos un total de s recapturas, entonces

$$\sum_{t=1}^{s} \frac{C_t M_t - R_t N}{N - M_t} = 0$$

Esta ecuación se puede resolver de forma iterada mediante distintos métodos (por ejemplo, a través del método de Newton-Raphson). Si suponemos que $\frac{M_t}{N} \ll 1$, es decir, si M_t es insignificante comparado con el tamaño total de la población (el tamaño poblacional es suficientemente grande), entonces una solución aproximada de la ecuación anterior es

$$\widehat{N} = \frac{\sum_{t=1}^{s} C_t M_t}{\sum_{t=1}^{s} R_t}$$

Al comienzo de este punto dijimos que este método es una extensión del método de Petersen. Notemos que la solución obtenida de la ecuación anterior tiene sentido en esta extensión, pues si tomamos una única recaptura, que correspondería con el método de Petersen, s=1, entonces nuestro estimador es

$$\hat{N} = \frac{CM}{R}$$

que coincide con el primer estimador que propusimos en el modelo hipergeométrico. Es decir, este estimador que hemos propuesto para el método de Schnabel, consiste, básicamente, en hacer la media ponderada de los resultados que se obtienen de hacer s experimentos distintos del método de Petersen, y es, principalmente en este hecho, donde radica la diferencia de este método con el método positivo de Jackson.

Ahora bien, en este punto se han de distinguir dos casos distintos; la probabilidad de captura puede ser igual o no a lo largo de las s recapturas que se realizarán a lo largo del experimento. Es en esto donde ahora pondremos nuestra atención. Cuando supongamos, en este mismo apartado, que las probabilidades de captura son iguales en todas las muestras, hablaremos del modelo $\mathbf{M}_0$ (el subíndice 0 indica que no hay variación en la probabilidad). Por el contrario, cuando queramos suponer que la población puede tener distintas probabilidades de captura en cada ocasión de muestreo, hablaremos del modelo $\mathbf{M}_t$ (t se referirá a cada ocasión de muestreo). En lo que resta de apartado, se hará un pequeño análisis y desarrollo de cada uno de estos modelos.

2.2.2. Modelo $\mathbf{M}_0$

Este modelo, como ya hemos comentado previamente, lo llamaremos modelo $\mathbf{M}_0$. Es fácil darse cuenta intuitivamente de la distribución de probabilidad que tendrá este modelo. Supongamos que $s = 6$. Entonces la probabilidad de que un animal presente el historial de captura 111111 (que haya sido capturado las 6 veces) es p^6, siendo p la probabilidad de captura en una muestra. Análogamente, la probabilidad de que presente el historial 110101 será $p^4(1-p)^2$ y la de que no presente el historial 000000 será $(1-p)^6$. Hay que resaltar el hecho de que desde el punto de vista probabilístico, los historiales con el mismo número de unos y ceros son iguales, ya que la probabilidad de que suceda es la misma. Llegados aquí, se tiene que la distribución que seguirá el modelo $\mathbf{M}_0$ será

$$L(N,p) = \frac{N!}{(N - M_{s+1})! \prod_h f_h!} p^C (1-p)^{sN-C}$$

donde f_h denota la frecuencia para el historial de captura observable h y $C = \sum_{i=t}^{s} C_t$ es la suma de las capturas de las s tomas muestrales. El historial de captura observable es el número de animales que han sido recapturados i veces $\forall i = 1, \ldots, s$ En el modelo $\mathbf{M}_0$, tan solo hay dos parámetros: N y p, y para poder estimarlos necesitamos conocer M_{s+1} ($s+1$ es el total de capturas; una captura más s recapturas) y C. Para obtener el estimador de máxima verosimilitud de N y p se pueden usar distintos métodos numéricos que maximizan la función de probabilidad escrita anteriormente. Una varianza de N para tamaños muestrales suficientemente grandes (Darroch, 1958) viene dado por la fórmula siguiente

$$\mathrm{Var}(\widehat{N}) = \frac{N}{(1-p)^{-s} + (s-1) - s(1-p)^{-1}}$$

2.2.3. Modelo $\mathbf{M}_t$

Este modelo que ahora daremos se aplica en el caso en el que se supone que la probabilidad de captura es distinta en cada ocasión muestral. Este es el modelo $\mathbf{M}_t$. Necesitamos, por tanto, definir p_k, que será la probabilidad de captura en la muestra k. Notemos que ahora tenemos s parámetros de probabilidad; la probabilidad de presentar un historial de captura 111111 cuando $s = 6$ será $p_1p_2p_3p_4p_5p_6 = \prod_{i=1}^{6} p_i$, la probabilidad de no presentar un historial de captura 000000 será $\prod_{i=1}^{6}(1-p_i)$ y la probabilidad de presentar un historial de captura 110101 será $p_1p_2(1-p_3)p_4(1-p_5)p_6$. En este caso, sí importa el orden en el que aparezcan los ceros y los unos en el historial de captura. En este caso, por tanto, la distribución de probabilidad será

$$L(N, p_1, \ldots, p_k) = \frac{N!}{(N - M_{s+1})! \prod_h f_h!} \prod_{t=1}^{s} p_t^{C_t}(1 - p_t)^{N-C_t}$$

En total hay $s + 1$ parámetros: $N, p_1, \ldots, p_s$ y para estimarlos necesitamos los estadísticos $n_1, \ldots, n_s, M_{s+1}$. El procedimiento iterativo para determinar el estimador de máxima verosimilitud se describe en el libro [10] (pág. 106) y la varianza viene dada por

$$\mathrm{Var}(\widehat{N}) = \frac{N}{\prod_{t=1}^{s}(1 - p_t)^{-1} + (s - 1) - \sum_{t=1}^{s}(1 - p_t)^{-1}}$$

Con el modelo $\mathbf{M}_t$, una suposición básica es que todos los animales tienen la misma probabilidad de captura en cada ocasión de muestreo. La igualdad de capturabilidad puede ser un ideal inalcanzable cuando se trabaja con poblaciones salvajes. En algunas circunstancias, sin embargo, los efectos de violar esta suposición se pueden evaluar y los sesgos resultantes pueden no ser demasiado grandes (Carothers 1973). Si solo hay dos ocasiones de captura, la suposición de igualdad en la capturabilidad no es verificable porque con dos ocasiones hay tres parámetros y tres frecuencias de capturas observables con el modelo multinomial. Por lo tanto, no quedarían grados de libertad para la prueba de bondad de ajuste. Para más de dos ocasiones,

Seber (1982, p.157) describió el test de bondad de ajuste basado en la prueba chi-cuadrado para todas las frecuencias de captura observadas. La puesta en común a gran escala puede ser necesaria para realizar esta prueba de chi-cuadrado cuando los recuentos son escasos.

3. Métodos para poblaciones abiertas

3.1. Método de Jolly-Seber

3.1.1. Descripción del modelo. Hipótesis y parámetros necesarios.

En este apartado se estudiará el modelo de Jolly-Seber, que será el único método que desarrollaremos considerando que nuestra población es abierta. Es decir, en este método ya se comenzará a suponer que el tamaño de la población con la que estemos trabajando puede variar en el tiempo en el que realizamos el estudio, por lo que, ya aquí, se trabajará con una situación que será más práctica y realista que las que hemos estando considerando anteriormente. En este punto cabe mencionar que no es éste el método más simple que se puede crear para poblaciones cerradas. El modelo de Cormack-Jolly-Seber precede a este método pero tan sólo vale para estimar la supervivencia y las probabilidades de captura. Es por eso que tomamos el modelo Jolly-Seber, que es una ampliación del anterior, para hacer una descripción de un método abierto. Para poder usar este método, se necesita tomar, al menos, 3 recapturas, por lo que el método de Jolly-Seber se puede ver como una extensión del modelo de Schnabel, como ya indicábamos en la introducción. Los individuos son marcados mediante alguna etiqueta o marcaje específico en un tiempo de muestreo. Nuestro interés estará en saber en qué momento exactamente fue marcado nuestro individuo de la muestra. Las muestras se obtienen de forma puntual en períodos de igual duración y son separadas por una larga duración hasta la próxima muestra. Aquí, a diferencia de los métodos cerrados, no nos supone un inconveniente tener una experiencia de estudio larga, ya que consideramos que la población es abierta y no nos afecta el hecho de que se produzcan cambios en el tamaño poblacional durante el tiempo de estudio. El intervalo de tiempo entre las distintas muestras no tiene por qué ser necesariamente constante.

Se necesita hacer varias suposiciones antes de comenzar con este estudio, así como también es necesario definir todos los parámetros y estadísticos que vamos a utilizar. Nuestras hipótesis serán:

HP1 Todo individuo de la población, marcado o no, tiene la misma probabilidad, $p_i = 1 - q_i$, de ser seleccionado en la i-ésima muestra i=1,...,s, supuesto que dicho individuo esté vivo cuando se realice la selección muestral.

HP2 Todo elemento marcado tiene la misma probabilidad, ϕ_i, de sobrevivir entre la selección muestral i-ésima y la (i+1)-ésima, dado que pertenece a la población inmediatamente después de realizarse la devolución de la i-ésima muestra i=1,2,...,s-1.

HP3 Todo elemento observado en la i-ésima muestra tiene la misma probabilidad v_i (i=1,2,...,s-1) de ser devuelto a la población. Es decir, $1 - v_i$ es la probabilidad de desaparición o de muerte accidental mientras que se está observando a un individuo que ha sido capturado en la i-ésima selección muestral.

HP4 Los elementos marcados no pierden su marca y todos los marcados son registrados al recapturarlos.

HP5 Todas las muestras son instantáneas de forma que el tiempo en observarlas es despreciable y la devolución se realiza inmediatamente después de la observación.

HP6 Toda emigración de la población es permanente.

Se utilizará la siguiente notación para el método de Jolly-Seber:

- t_i: instante en el que es tomada la i-ésima recptura (i=1,2,...,s).
- N_i: Número total de individuos de la población antes del instante de tiempo t_i (i=1,...,s).

- M_i: Número total de individuos marcados antes del instante de tiempo t_i (i=1,...,s). Obsérvese que M_1=0.

- U_i: Número de individuos que en el instante de tiempo t_i aún no han sido marcados en ningún momento. $U_i = N_i - M_i$ (i=1,...,s)

- B_i: Número de elementos que se incorporan a la población en el tiempo que transcurre entre los instantes de tiempo t_i y t_{i+1}, y que además pertenecen a la población en el instante de tiempo t_{i+1} (i=0,...,s-1). Notemos el hecho de que, por definición, $B_0 = N_1$.

- ϕ_i: Probabilidad de sobrevivir en el intervalo de tiempo $(t_i, t_{i+1}]$, (i=1,. . . ,s-1).

- $p_i = 1 - q_i$: Probabilidad de ser capturado en la i-ésima muestra (i=1,...,s).

- ρ_i: Proporción de elementos marcados en la población en el instante de tiempo t_i. $\rho_i = \frac{M_i}{N_i}$ (i=1,...,s)

- α_i= $\mathbb{P}$ [un individuo no sea seleccionado en la $(i+1)$-ésima extracción y pertenezca a la población entre la i-ésima y la $(i+1)$-ésima extracción]= =$\mathbb{P}$[sobrevivir entre la i-ésima y la (i+1)-ésima extracción]×$\mathbb{P}$[no ser capturado en la $(i+1)$-esima extracción]=$\phi_i q_{i+1}$.

- β_i= $\mathbb{P}$ [un individuo sea seleccionado en la (i+1)-ésima extracción y pertenezca a la población entre la i-ésima y la $(i+1)$-ésima extracción]= =$\mathbb{P}$[sobrevivir entre la i-ésima y la $(i+1)$-ésima extracción]×$\mathbb{P}$ [ser capturado en la $(i+1)$-ésima extracción]=$\phi_i p_{i+1}$.

- n_i: Número de individuos capturados en la i-ésima muestra (i=1,2,...,s).

- m_i: Número de individuos marcados observados en la i-ésima muestra (i=1,2,...,s).

- u_i=$n_i - m_i$ (i=1,..,s).

- R_i: Número de elementos devueltos después de marcarlos a la población en la i-ésima muestra (i=1,...,s).
- r_i: Número de elementos de los R_i devueltos a la población tras la i-ésima muestra y que vuelven a ser recapturados nuevamente más tarde (i=1,...,s).
- z_i: Número de elementos diferentes observados antes de la i-ésima extracción que no son observados en la i-ésima muestra, pero que sí que lo son en alguna de las muestras posteriores.

3.1.2. Estimación de los parámetros

De modo similar a lo que se hacía el método hipergeométrico, se puede estimar fácilmente la proporción de la población total mediante un muestreo aleatorio. Si suponemos que la muestra es aleatoria, entonces ésta contendrá la misma proporción de animales marcados que la población total. Por tanto, intuitivamente, como se hacía en el caso del método de Petersen-Lincoln, podemos construir el siguiente estimador:

$$\frac{m_i}{n_i} \simeq \frac{M_i}{N_i} = \rho_i \Longrightarrow \widehat{N_i} = \frac{n_i}{m_i}\widehat{M_i}$$

Notemos que tiene sentido hablar de $\widehat{M_i}$, ya que, aunque por definición pueda parecer que conocemos M_i (son individuos que se han marcado durante la experimentación, por lo que a priori conocemos el número), debemos tener en cuenta que nuestra población es abierta y podemos perder individuos marcados a lo largo de nuestra experiencia; la población marcada puede disminuir por efecto de la muerte y de la emigración de los individuos marcados. Por tanto, lo que se nos plantea ahora es encontrar un estimador $\widehat{M_i}$ de M_i para estimar el tamaño total de la población.
Si nos centramos únicamente en la población marcada, M_i, se puede hacer una división de individuos. Considerando por un lado los individuos marcados que en el instante de tiempo t_i (muestra i) han sido recapturados (m_i) y, por otro lado los individuos marcados que no han sido seleccionados en el instante de tiempo t_i ($M_i - m_i$, podemos suponer que las tasas de recaptura futuras de los dos grupos de animales marcados en el instante t_i serán aproximadamente iguales, es decir,

$$\frac{z_i}{M_i - m_i} \simeq \frac{r_i}{R_i}$$

ya que:

- $M_i - m_i$: elementos marcados no recapturados en i-ésima muestra
- z_i: elementos marcados previamente que no son observados en el instante t_i, pero que son recapturados en alguna de las muestras sucesivas.

- R_i: elementos marcados y después devueltos a la población en la i-ésima muestra
- r_i: elementos de los R_i devueltos a la población tras la i-ésima muestra y que son recapturados en ocasiones posteriores.

Así, se construye el siguiente estimador para M_i

$$\frac{z_i}{M_i - m_i} \simeq \frac{r_i}{R_i} \Longrightarrow \widehat{M}_i = z_i \frac{R_i}{r_i} + m_i \quad i = 2, \ldots, s-1$$

Con el objetivo de obtener un estimador insesgado, podemos modificar levemente el que nosotros hemos obtenido y considerar el siguiente estimador, que sí es insesgado

$$\widetilde{M}_i = m_i + \frac{R_i + 1}{r_i + 1} z_i$$

En consecuencia, si seguimos añadiendo el término $+1$ para evitar el sesgo, tendremos entonces que un estimador insesgado para N es

$$\widetilde{N}_i = \frac{n_i + 1}{m_i + 1} \widetilde{M}_i$$

Es decir, se considera la siguiente proporción de marcados

$$\widehat{\rho}_i = \frac{m_i + 1}{n_i + 1}$$

Generalmente, el interés no está en hacer una única estimación del tamaño de una población, por lo que efectuaremos un censo múltiple, es decir, pondremos en práctica este mismo proceso durante varias veces. El interés de esto es que nos puede llevar a comenzar el estudio de la dinámica de una población.

También nos interesa estimar las tasas de supervivencia. El estimador de ésta se obtiene a partir de la proporción de animales marcados presentes en el tiempo $i+1$ con respecto a los presentes en el tiempo i:

$$\widehat{\phi}_i = \frac{\widehat{M}_{i+1}}{\widehat{M}_i + R_i - m_i}$$

En esta expresión, el término $R_i - m_i$ representa el número de individuos de la población recién marcados en el instante i. Nótese que el estimador de la tasa de

supervivencia no distingue entre las pérdidas debidas a las muertes y las debidas a la emigración permanente. Por tanto, esta cantidad es normalmente denominada 'aparente supervivencia' debido a este motivo.

La probabilidad de captura, p_i, se puede estimar como el cociente entre los animales marcados en el tiempo i y el número total de animales presentes en la población en dicho instante. Como dicho número no lo conocemos, para estimar p_i, usamos la estimación de M_i

$$\widehat{p}_i = \frac{m_i}{\widehat{M}_i} \quad i = 2, 3, \ldots, s-1$$

Para estimar el número de nacimientos, debemos considerar la diferencia entre el tamaño poblacional en el instante $i+1$ y en el instante i, teniendo en cuenta los individuos que han fallecido por causas de muerte natural $(1-\phi_i)$ y los que han fallecido a causa de la captura $(n_i - R_i)$:

$$\widehat{B}_i = \widehat{N}_{i+1} - \widehat{\phi}_i[\widehat{N}_i - (n_i - R_i)] \quad i = 2, \ldots, s-2$$

Los tres últimos estimadores que se han propuesto tienen la ventaja de que son insesgados, a diferencia de lo que pasaba originalmente con los estimadores de M_i y N_i.

Las varianzas y covarianzas de estos estimadores, se encuentran en Seber (1982, p.202) y Pollock (1990). Aquí se presentan solamente la varianza del estimador de la abundancia y de la tasa de supervivencia

$$\mathrm{var}(\widehat{N}_i) = N_i[N_i - \mathbb{E}(n_i)]\left[\frac{M_i - \mathbb{E}(m_i) + R_i}{M_i}\left(\frac{1}{\mathbb{E}(r_i)} - \frac{1}{R_i}\right) + \frac{N_i - M_i}{N_i \mathbb{E}(m_i)}\right]$$

$$\mathrm{var}(\widehat{\phi}_i) = \phi_i^2\Big\{\frac{[M_{i+1} - \mathbb{E}(m_{i+1})][M_{i+1} - \mathbb{E}(m_{i+1}) + R_{i+1}]}{M_{i+1}^2}\left[\frac{1}{\mathbb{E}(r_{i+1})} - \frac{1}{R_{i+1}}\right] + \\ + \frac{M_i - \mathbb{E}(m_i)}{M_i - \mathbb{E}(m_i) + R_i}\left[\frac{1}{\mathbb{E}(r_i)} - \frac{1}{R_i}\right]\Big\} + \left[\frac{\phi_i(1-\phi_i)}{M_i - \mathbb{E}(m_i) + R_i}\right]$$

3.1.3. Dos aplicaciones del modelo de Jolly-Seber

En este apartado se verá cómo el modelo de Jolly-Seber nos vale para modelar situaciones que estimen parámetros más allá de la abundancia.
Para el desarollo de esta sección este, es necesario definir algunos parámetros más que se van a utilizar. Utilizaremos dos subíndices para los parámetros, ya que en este apartado se dividirá a la población en distintos grupos, para estudiar cuánto difieren las estimaciones de un mismo parámetro en distintos grupos. Se utilizará el subíndice i para indicar la recaptura (el instante) en la que estamos y el subíndice g para indicar que el elemento pertenece al grupo g. Consideraremos, como hemos estado haciendo hasta ahora, que el experimento tiene s recapturas. Los parámetros que utilizaremos son:

- p_{gi}: Probabilidad de capturar a un individuo perteneciente al grupo g en la recaptura i
- ϕ_{gi}: Probabilidad de que un individuo que se encuentra vivo en la recaptura i del grupo g, sobreviva entre las muestras i e i+1
- N_{gi}: Tamaño de la población del grupo g en el momento de la recaptura i. En este punto, es importante distinguir entre N_{gi}^{-}, que es el número de estos individuos justo antes de realizar la recaptura, y N_{gi}^{+}, que es el número de individuos justo después de realizar la recaptura. Puede ocurrir que $N_{gi}^{-} - N_{gi}^{+} > 0$ a causa de las pérdidas que se puedan tener debido a la experimentación.
- N_g: Número total de animales en el grupo g que se introducen en la población y permanecen vivos hasta la siguiente muestra.
- η_{gi}: Fracción del total de nuevos nacimientos que se introducen en la población entre las recapturas i e i+1. Aquí, cuando hablamos de nacimientos,

nos estamos refiriendo a cualquier mecanismo por el cual haya nuevos individuos que ingresen en la población capturable (nacimientos biológicos, inmigracción,etc.)

- B_{gi}: Número de animales del grupo g que entran en la población después de la recaptura i y sobreviven hasta la recaptura i+1. B_{g0} se define como el número de individuos vivos justo antes de la toma de la primera captura.

- U_{gi}: Número de animales del grupo g que en la recaptura i aún no han sido seleccionados en ningún momento.

- λ_{gi}: Tasa de crecimiento de la población del grupo g entre las muestras i e i+1.

- f_{gi}: Tasa de fecundidad del grupo g entre las recapturas i e i+1.

- α_{gi}: Probabilidad de que un individuo que se encontraba presente en el grupo g justo antes de la recaptura i, también estuviese presente justo después de la recaptura i-1.

- ψ_{gi}: Probabilidad de que un animal se adentre en la población, permanezca vivo, y no sea capturado antes de la recaptura i.

Estos parámetros presentan algunas relaciones entre ellos:

$$N_g = \sum_{i=0}^{s-1} B_{gi} \qquad \eta_{gi} = \frac{B_{gi}}{N_g} \Rightarrow \eta_i = \frac{B_i}{N}$$

$$N_{g1} = B_{g0}$$

$$U_{g1} = N_{g1}; \qquad U_{g(i+1)} = U_{gi}(1 - p_{gi})\phi_{gi} + B_{gi}$$

$$\alpha_{gi} = \frac{N^{+}_{g(i-1)}\phi_{g(i-1)}}{N^{-}_{gi}}$$

$$\psi_{g1} = \eta_{g0}, \qquad \psi_{g(i+1)} = \psi_{gi}(1 - p_{gi})\phi_{gi} + \eta_{gi}$$

Habiendo definido nuestros nuevos parámetros, se puede hacer ya el estudio de las dos situaciones que se van a modelar:

• **A1: Modelado del patrón de la entrada y salida de individuos en la población**

En las situaciones que hasta ahora hemos visto, el estudio solo se ha centrado en estimar el número de individuos que habitan en una determinada población. Ahora que estamos trabajando con poblaciones abiertas, podemos suponer que el cambio del número de animales en el tiempo de estudio sigue un determinado patrón. El objetivo de este apartado es estudiar ese patrón, que puede ser de suma importancia, ya que nos puede aportar información, por ejemplo, de si el número de animales macho y animales hembra cambian mediante el mismo patrón o no largo del tiempo.

Se verá primero un estimador que fue propuesto por Jolly y Seber en 1965. Para el comienzo de este estudio, necesitamos estudiar la expresión que tiene la probabilidad de ser capturado en todas las muestras en el modelo de Jolly-Seber. Esta probabilidad, L, la podemos dividir en tres partes de la siguiente manera:

$$L = L_1 \times L_2 \times L_3$$

donde

• L_1=$\mathbb{P}$[ser capturado en la primera muestra]

• L_2=$\mathbb{P}$[ser liberado en la muestra i | ha sido capturado en la muestra i], $\forall i = 1, \ldots, s-1$

• L_3=$\mathbb{P}$[ser recapturado en la muestra i | haya sido capturado en cada muestra previa a i]

L_2 y L_3 se pueden modelar, como se muestra en [13], mediante productos de binomiales condicionadas. Con L_1 no se puede hacer esto. Se necesitará, por tanto,

suponer también que los U_{gi} son parámetros fijos. Así, teniendo en cuenta que B_{gi}=$U_{gi} - \tau_{gi}(U_{gi} - u_{gi})$, entonces L_1 lo podemos escribir como un producto de distribuciones binomiales:

$$\prod_{g=1}^{G}\prod_{i=1}^{s}\binom{U_{gi}}{u_{gi}}p_{gi}^{u_{gi}}(1-p_{gi})^{U_{gi}-u_{gi}}$$

En esta expresión, el estimador de máxima verosimilitud viene dado por $\widehat{U}_{gi} = \dfrac{u_{gi}}{\widehat{p}_{gi}}$. Sin embargo, esta aproximación supone algunos problemas. Por un lado, no se están teniendo en cuenta los nacimientos, lo que complica el poder imponer restricciones a los B_{gi}, como por ejemplo que puedan ser cero en algún momento o que puedan ser iguales en dos grupos distintos en algún momento concreto. Por otro lado, la probabilidad modela el tamaño de la población en bruto. No es posible trasladar esta probabilidad al patrón que estamos buscando.

Estos incovenientes los trataron Schwarz y Arnason en 1996. Tomaron $B_{gi}, i = 1, .., s-1$ como variables aleatorias condicionadas a N_g. Entonces estas variables siguen una distribución multinomial, lo que nos lleva a que

$$L_1 = \prod_{g=1}^{G}\binom{N_g}{u_g}\left(\sum_{i=1}^{k}\psi_{gi}p_{gi}\right)^{u_g}\left(1-\sum_{i=1}^{k}\psi_{gi}p_{gi}\right)^{N_g-u_g}\binom{u_g}{u_{g1}...u_{gk}}\prod_{i=1}^{k}\left(\frac{\psi_{gi}p_{gi}}{\sum_{i=1}^{k}\psi_{gi}p_{gi}}\right)^{u_{gi}}$$

Esta nueva formulación responde mejor a nuestros intereses ya que se ha hecho uso de los parámetros que están definidos sobre las tasas de nacimiento, y también resuelve los problemas anteriormente planteados.

Si hiciésemos el modelo en función de los animales que entran en la población entre dos muestras consecutivas, también resultaría ventajoso. En primer lugar, si realizásemos el mismo experimento a 2 grupos de animales con un tamaño de población parecido, no nos esperaríamos que el patrón fuese el mismo para los dos grupos, aunque, teóricamente, este debería ser igual. Por otro lado, la estimación de β_{gi} es relativamente insesgada debido a la heterogeneidad en la capturabilidad.

Carothers probó que el sesgo relativo asintótico es función de los α_{gi}. El probó que $\mathbb{E}\left[\widehat{N}_{gi}\right] \approx \frac{N_{gi}}{1+\alpha_{gi}^2}$, pero sin embargo a las estimaciones de supervivencia no les afecta la heterogeneidad. Si el coeficiente de variación de capturabilidad es relativamente constante a lo largo del tiempo, entonces $\widehat{B}_{gi} = \widehat{N}_{g(i+1)} - \widehat{N}_{gi}\widehat{\phi}_{gi}$ y $\widehat{N}_g = \sum_{i=0}^{k-1} \widehat{B}_{gi}$ tienen los mismos sesgos relativos, pero, sin embargo, el estimador $\widehat{\eta}_{gi} = \frac{\widehat{B}_{gi}}{\widehat{N}_g}$ será relativamente insesgado. Por tanto, no es necesario hacer correcciones en estos casos.

A2: Modelado del crecimiento y fecundidad de la población

El modelo de Jolly-Seber se desarrolló para estimar la el tamaño total de una población. Sin embargo, en muchas ocasiones tiene más interés en el campo práctico conocer cómo crece o decrece la población. La tasa de crecimiento de una población está definida como

$$\lambda_i = \frac{N_{i+1}^-}{N_i^+} = \frac{N_i^+\tau_i + B_i}{N_i^+} = \phi_i + \frac{B_i}{N_i^+} = \phi i + f_i$$

donde f_i es el crecimiento neto en el intervalo de tiempo comprendido entre i e i+1 por individuo inicialmente vivo al comienzo del experimento. El modelo de Jolly-Seber se puede usar para estimar la probabilidad de precedencia y el crecimiento de la población. La probabilidad de precedencia se define matemáticamente como:

$$\alpha_{i+1} = \frac{N_i^+\tau_i}{N_{i+1}^-} = \frac{N_{i+1}^- - B_i}{N_{i+1}^-} = 1 - \frac{B_i}{N_{i+1}^-}$$

Todos estos parámetros se pueden definir cambiando la notación, escribiéndolos en función de los parámetros del modelo de Jolly-Seber, de los cuales conocemos ya los estimadores. Esta redefinición es la siguiente

$$\lambda_i = \phi_i + \frac{\eta_i}{\eta_0\phi_1\phi_2...\phi_{i-1} + \eta_1\phi_2\phi_3...\phi_{i-1} + ... + \eta_{i-1}}$$

$$\alpha_{i+1} = 1 - \frac{\eta_i}{\eta_0\phi_1\phi_2...\phi_i + \eta_1\phi_2\phi_3...\phi_i + ... + \eta_i}$$

$$f_i = \lambda_i - \phi_i = \phi_i\left(\frac{1}{\alpha_{i+1}} - 1\right) = \frac{\eta_i}{\eta_0\tau_1\phi_2...\phi_{i-1} + \eta_1\phi_2\phi_3...\phi_{i-1} + ... + \eta_{i-1}}$$

En cada uno de estos tres casos, las estimaciones se pueden obtener mediante simple sustitución, ya que el estimador de ϕ_i se ha dado en el modelo de Jolly-Seber y $\eta_i = \dfrac{B_i}{N}$, por lo que también se puede estimar. Las varianzas se pueden hallar mediante las expansiones en serie de Taylor.

Hay varias ventajas de obtener las estimaciones como función de parámetros fundamentales en lugar de parámetros intrínsecos dentro de una nueva probabilidad. En primer lugar, se muestra claramente que estos parámetros dependen tanto de las tasas de supervivencia como de los nuevos individuos en la población. Por otro lado, todas las estimaciones están construidas automáticamente para ser consistente respecto de las otras. Por ejemplo, $\widehat{\lambda}_i$ nunca puede estar por debajo de la tasa de supervivencia estimada, $\widehat{\alpha}_i \leq 1$, y $\widehat{f}_i$ debe ser siempre positivo. Por último, el estimador de máxima verosimilitud difiere en función de los parámetros que definamos.

La mayor dificultad al usar la aproximación de Jolly-Seber es ajustar los modelos donde los parámetros definidos sean iguales en distintos tiempos o distintos grupos. Debido a que hemos obtenido funciones no lineales, habrá técnicas, como el diseño de matrices, que no funcionarán. Sin embargo, podemos imponer restricciones (lineales o no) usando el método de los multiplicadores de Lagrange, para resolver este problema.

4. Procedimientos de selección de modelos

Es claro que aparte de los modelos que aquí se han desarrollado, existen muchísimos más modelos para poblaciones tanto abiertas como cerradas que se han desarrollado en la segunda mitad del siglo XX y comienzos del siglo XXI, por lo que a la hora de realizar un estudio sobre una determinada población, el abanico de modelos en los que podemos basar nuestro estudio es demasiado amplio. Por ejemplo, al considerar una población abierta, podría ser que las probabilidades de captura y supervivencia varíen en el tiempo, en función del sexo del individuo, las condiciones climatológicas y ambientales, etc. El problema que se plantea es por tanto elegir un modelo que represente adecuadamente el conjunto de datos sin necesidad de tener que elegir más parámetros de los que realmente necesitemos.

Con respecto a esto, hay dos resultados que nos pueden resultar particularmente útiles.

Para el primer resultado, supongamos que debemos elegir entre dos modelos distintos para trabajar con un determinado conjunto de datos. El primer modelo diremos que tiene I parámetros para estimar, una función de verosimilitud $\log(L_1)$ para hallar el estimador de máxima verosimilitud de los parámetros. El segundo modelo, que es una generalización del primero, debe estimar los I parámetros del primer modelo además de otros J más, y su función de verosimilitud para la estimación de los parámetros por máxima verosimilitud es $\log(L_2)$. Debido a que el segundo modelo es más general que el primero, se tendrá que $\log(L_2)$ será menor o igual que $\log(L_1)$. Sin embargo, si los J parámetros extra del segundo modelo no son necesarios, entonces valdrán cero. Entonces el estadístico

$$\mathrm{D} = 2[\log(L_1) - \log(L_2)]$$

aproximará un valor aleatorio de una distribución chi-cuadrado con J grados de libertad. Consecuentemente, si D es significativamente grande con los valores de

la distribución chi-cuadrado, entonces esto sugiere que el modelo 2 es mejor para describir los datos. Por el contrario, si D no es muy grande, entonces será más apropiado escoger el modelo primero para describir los datos. Este método tiene una limitación importante, y es que tan solo vale para modelos "encajados", es decir, en los que un modelo es generalización del otro. Como alterativa a este método se puede considerar otro que supera esta limitación, basado en criterio de información de Akaike (AIC; Akaike's information criterion).

En su forma más simple, la selección del modelo mediante el AIC involucra definir de un conjunto de modelos 'candidatos' a representar nuestro conjunto de datos. Cada modelo se ajustará a los datos, y se le asigna un valor correspondiente:

$$\text{AIC} = -2\log(L) + 2P$$

donde L es la máxima verosimilitud del modelo y P es el número de parámetros estimados. El modelo que tenga menor AIC será considerado el mejor para representar los datos, teniendo en cuenta la relación entre el ajuste del modelo y el número de parámetros a estimar. La diferencia que haya entre el ajuste del modelo y los parámetros a estimar es importante para la precisión de las estimaciones. Una comparación complementaria entre modelos se basa en el cáculo de los pesos de Akaike. Si hay M 'candidatos', entonces el peso del modelo i es

$$w_i = \frac{\exp(-\Delta_i/2)}{\prod_{j=1}^{M} \exp(\Delta_i/2)}$$

donde Δ_i es la diferencia entre el valor AIC del modelo i y el valor más pequeño de entre los AIC de todos los modelos. Los pesos de Akaike, de este modo, sirven para medir la fuerza de la evidencia de cada uno de los modelos, de modo que un gran peso indicará una gran evidencia (esto es, que puede servir mejor para ajustar los datos).

5. Análisis de librerías en R

5.1. Principales librerías para métodos de captura-recaptura

Para los métodos de captura-recaptura, existen, a junio de 2018, varias librerías en R a las que hacemos mención a continuación, señalando y describiendo algunas de sus funciones más importantes en algunas de ellas.

1 : Librería Rcapture. Esta librería utiliza modelos log-lineales para métodos de captura-recaptura. Harremos un análisis más detallado de esta librería en la siguiente sección.

2 : Librería Marked. Esta librería analiza las estimaciones de supervivencia y de abundancia.

3 : Librería unmarked. Aquí se analizan modelos para conjuntos de datos de animales que no han sido marcados. Esta librería tiene varias funciones de detección de muestreo de distancia y funciones de densidad asociadas.

4 : Librería Rmark. Esta librería contiene un código R para el análisis de marcas.

5 : Librería multimark. Esta librería tiene las funciones necesarias par realizar un análisis de captura-marca-recaptura utilizando un marcaje múltiple con marcas no invasivas. Las funciones más importantes de esta librería son

5.1 markCJS: Ajusta los modelos de supervivencia de poblaciones abiertas para conjuntos de datos de captura, marca y recaptura que constan de un único tipo de marca

5.2 multimarkCJS: Ajusta los modelos de supervivencia de poblaciones

abiertas para conjuntos de datos de captura, marca y recaptura que constan de múltiples marcas no invasivas.

4.3 markClosed: Ajusta los modelos de abundancia en poblaciones cerradas para conjuntos de datos de captura, marca y recaptura que constan de un único tipo de marca

5.4 multimarkClosed: Ajusta los modelos de abundancia en poblaciones cerradas para conjuntos de datos de captura, marca y recaptura que constan de múltiples marcas no invasivas

Aunque a todas estas funciones se le pueden dar varios argumentos en función del estudio que queramos hacer, al darle como argumento tan solo el conjunto de datos, ya nos proporcionan el ajuste deseado.

6 : Librería CARE1. Este paquete también es válido para estimar la abundancia de una determinada población

7 : Otras librerías con menos funciones que estas son: secr, CARE1, mra, R2UCARE y mrds.

5.2. Análisis de la librería Rcapture

En este apartado, estudiaremos en primer lugar cuáles son los distintos tipos de estructuras que nos podemos encontrar en el conjunto de datos y analizaremos las funciones que utlizaremos, examinando qué argumentos debe tomar cada función. Hay básicamente 4 formatos distintos de conjunto de datos que nos podemos encontrar. Son los siguientes:

- **FORMATO 1**: Este formato de conjunto de datos consta de una matriz que tan solo tiene ceros y unos. En cada captura, el individuo i-ésimo tendrá un uno en la matriz del conjunto de datos si ha sido capturado y un cero en caso contrario. El número de filas de la tabla indicará el número total de individuos que han sido seleccionados en al menos una ocasión y el número de columnas el número total de muestras que hemos tomado en el experimento.En este conjunto de datos se debe tomar dfreq=FALSE (no se han considerado las frecuencias) y dtype="hist". Los datos aparecen de la siguiente manera

$$\begin{matrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{matrix}$$

- **FORMATO 2**: En este tipo de formato, tendremos una tabla similar a la del formato 1 con la diferencia de que aquí aparecerá una columna más, que mostrará la frecuencia con la que ese individuo ha sido capturado. En este conjunto de datos se debe tomar dfreq=TRUE

y dtype="hist". Los datos aparecen de la siguiente manera

$$\begin{matrix} 1 & 1 & 2 \\ 1 & 0 & 4 \\ 0 & 1 & 1 \end{matrix}$$

- **FORMATO 3**: Este formato contiene un vector que para cada elemento que ha sido capturado alguna vez, dice cuántas veces lo ha sido. Por lo tanto, este formato no contiene historiales de captura completos.En este conjunto de datos se debe tomar dfreq=FALSE y dtype="nbcap". En este formato, el conjunto de datos se ve así

  2 2 1 1 1 1 1

- **FORMATO 4**: En este tipo de formato tendremos una matriz de dos columnas. La primera columna contiene el número observado de capturas, las segundas columnas contienen sus frecuencias. En este conjunto de datos se debe tomar dfreq=TRUE y dtype="nbcap". El conjunto de datos en este formato se ve así

$$\begin{matrix} 2 & 2 \\ 1 & 5 \end{matrix}$$

Comenzamamos ya nuestro estudio. En primer lugar consideraremos poblaciones cerradas. Veremos dos ejemplos para dos conjuntos de datos distintos (con estructura distinta). A la hora de realizar un análisis en R para poblaciones cerradas, hay un procedimiento general a seguir:

1. Análisis descriptivo de los datos. En la librería que nosotros estamos usando, se pueden utilizar para ello las funciones descriptive y plot

2. Ajustes de los modelos. Para ello, haremos uso de tres funciones ya definidas en la librería que usamos. Estas funciones son closedp, closedp.h y closedp.mX

3. Selección del modelo. Para esta parte, utilizaremos básicamente las funciones boxplot.closedp y uifit

4. Estimación de la abundancia. Para este último punto, se usarán las funciones closedp.bc y profileCI, para calcular intervalos de confianza.

Estamos ya en las condiciones para comenzar nuestro análisis de esta librería. Comenzamos por instalar el paquete y cargar la librería.

```
# install.packages("Rcapture")
library(Rcapture)
```

El primer conjunto de datos que analizaremos contiene información sobre un determinado tipo de liebres. Este conjunto de datos contiene una matriz 68x6 con los historiales de captura para seis ocasiones muestrales. El conjunto de datos es del formato 1.

```
data("hare")
```

Ahora haremos un análisis descriptivo del conjunto de datos. Nos aparecerá una tabla descriptiva, en la que se incluyen los siguientes parámetros:

f_i es el número de individuos que se capturan i veces.

u_i es el número de individuos que se capturan por primera vez en la ocasión i.

v_i es el número de individuos que se capturan por última vez en la ocasión i.

n_i es el número de individuos que se capturan en total en la ocasión i.

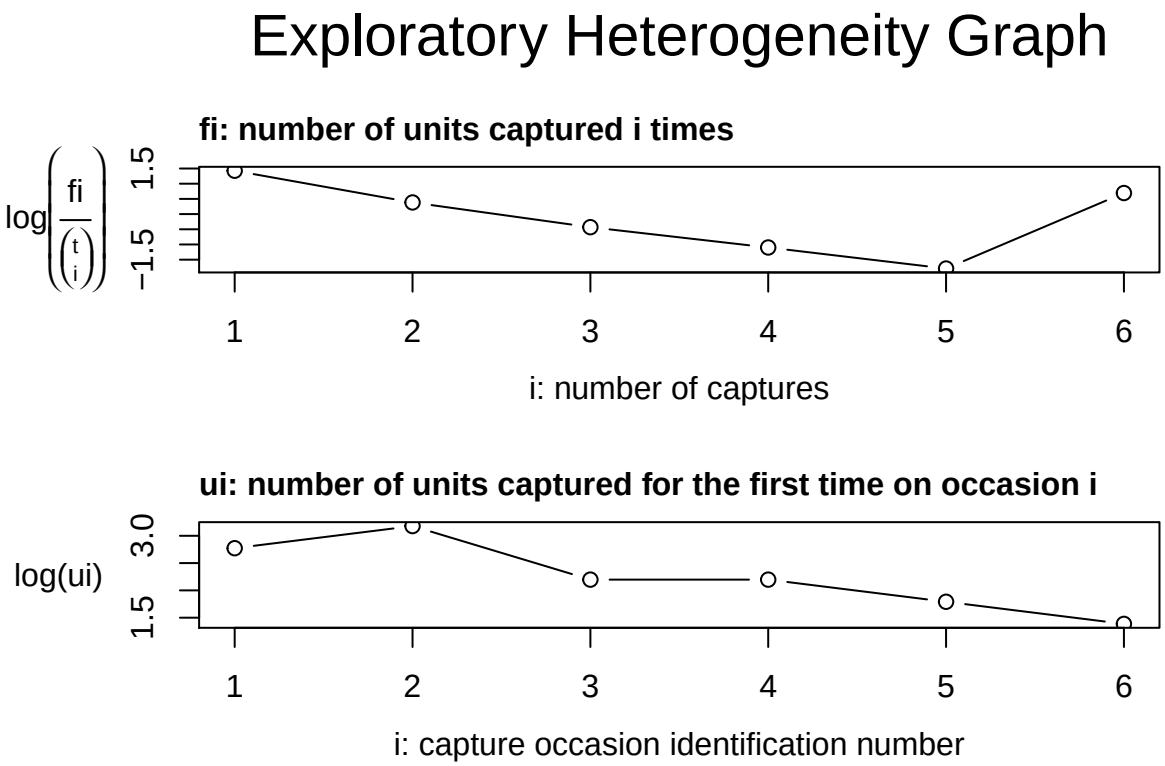
```
desc<-descriptive(hare) #Descripción estadística de los datos
desc
```

```
##
## Number of captured units: 68
```

```
##
## Frequency statistics:
##        fi  ui  vi  ni
## i = 1  25  16   3  16
## i = 2  22  24   6  28
## i = 3  13   9   6  20
## i = 4   5   9   9  26
## i = 5   1   6  12  23
## i = 6   2   4  32  32
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
```

Se estudiará ahora la heterogeneidad. La función plot.descriptive genera dos gráficos para analizar la heterogeneidad de la captura. En ausencia de heterogeneidad, el gráfico debe presentar una relación lineal o casi-lineal. Funciones convexas indicarían heterogeneidad.

```
plot(desc)
```



Hacemos un ajuste log-lineal del modelo mediante la función closedp. A esta función solo debemos darle como argumento la matriz del conjunto de datos porque ya tiene el valor dfreq=FALSE por defecto.

```
closedp(hare)
```

```
##
## Number of captured units: 68
##
## Abundance estimations and model fits:
##               abundance stderr deviance df     AIC     BIC infoFit
## M0                 75.4    3.5   68.516 61 154.707 159.146      OK
## Mt                 75.1    3.4   58.314 56 154.505 170.041      OK
```

```
## Mh Chao (LB)       79.8    6.4   58.023 58 150.214 161.311      OK
## Mh Poisson2        81.5    5.7   59.107 60 147.298 153.956      OK
## Mh Darroch         90.4   11.6   61.600 60 149.791 156.449      OK
## Mh Gamma3.5       100.6   21.7   62.771 60 150.961 157.619      OK
## Mth Chao (LB)      79.6    6.3   47.115 52 151.305 175.720      OK
## Mth Poisson2       81.1    5.6   48.137 55 146.327 164.083      OK
## Mth Darroch        90.5   11.7   50.706 55 148.896 166.652      OK
## Mth Gamma3.5      101.6   22.4   51.956 55 150.147 167.903      OK
## Mb                 81.1    8.3   67.027 60 155.217 161.876      OK
## Mbh                74.2   14.6   63.257 59 153.447 162.325      OK
```

En esta tabla aparecen distintos modelos y la estimación de la abundancia utilizando cada uno de ellos. En particular, los dos primeros modelos que aparecen son el modelo M_0 y el modelo M_t desarrollados en teoría. Observamos también que aparece el AIC para cada modelo, lo que nos puede ayudar a elegir el modelo que mejor se ajuste a los datos. A priori, parece que el mejor modelo que podemos tomar es el modelo *Mth Poisson*. Aparte de la estimación de la abundancia y del AIC, los elementos que aparecen en la tabla para describir el modelo son

- stderr: Error estándar de la estimación
- deviance: Desviación del modelo
- df: Número de grados de libertad
- BIC: Criterio de información bayesiano
- infoFit: Código que nos indica si ha habido errores en el ajuste del modelo.

Teniendo en cuenta que hay dos individuos que han sido seleccionados todas las veces, hacemos un estudio eliminando estas observaciones porque nos pueden estar llevando a sobrestimar el tamaño poblacional.

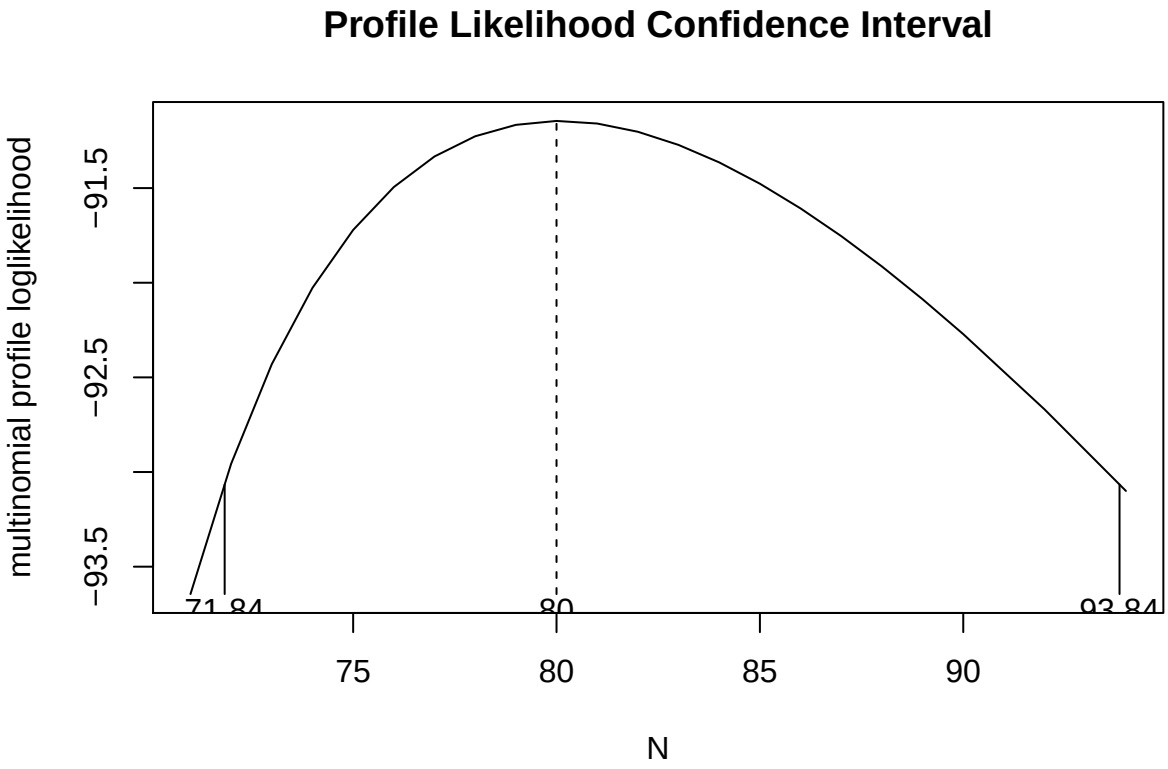
```r
col<-rep(0,2^6-1)
mat<-histpos.t(6)
col[apply(mat,1,sum)==6]<-1
cp.m2<-closedp.mX(hare,mX=cbind(mat,col),mname="Mt sin 111111")
cp.m2$results
```

```
##                    abundance   stderr deviance df      AIC
## Mt sin 111111  76.77761 3.911153 47.89417 55 146.0846
```

Nos da una estimación del tamaño poblacional de 76,7. Observamos que el AIC es menor que todos los anteriores. Esto nos indica que ajustar los datos sin estas experimentaciones puede que nos de mejores resultados.

Hallamos por último un intervalo de confianza. A la función profileCI necesitamos darle el conjunto de datos, el modelo del que queremos que nos de el IC (Mth, por ejemplo) y el método para hallar el intervalo (Poisson, por ejemplo). Tomamos *Mth Poisson* porque es el modelo que menor AIC tenía.

```r
CI1<-profileCI(hare,m="Mth",h="Poisson",a=2)
```



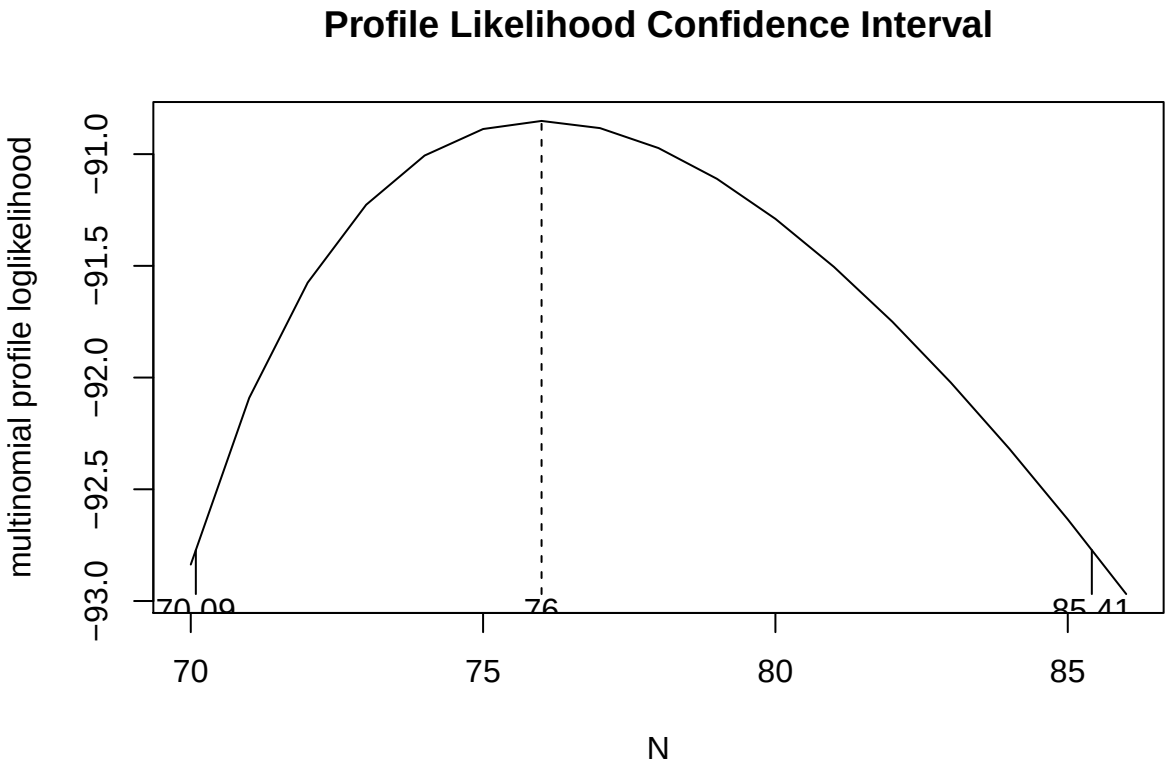

```
CI1$results
```

```
##               abundance    InfCL    SupCL
## Mth Poisson2         80 71.84073 93.84254
```

Utilizando este modelo, el intervalo de confianza que se obtiene es [71,94].

Ahora pedimos un intervalo de confianza eliminando las observaciones que han sido seleccionadas las seis veces y comparamos

```
CI2<-profileCI(hare,mX=cbind(mat,col),mname="Mt sin 111111")
```




```
CI2$results
```

```
##                    abundance    InfCL    SupCL
## Mt sin 111111         76 70.08663 85.41181
```

En este caso, la estimación por intervalos de la abundancia es [70,86]. Este segundo modelo nos proporciona un intervalo de confianza de amplitud más reducida que el anterior.

Ahora vamos a trabajar con otro conjunto de datos. Utlizaremos el conjunto de datos HIV, que contiene información sobre datos epidemiológicos de captura y recaptura del VIH de cuatro centros de información en Roma, Italia. Haremos un análisis básico de este conjunto de datos. Este ejemplo se expone para ver que estos

métodos no solo valen para trabajar con animales, si no que también tiene utilidad en medicina y en ciencias sociales. Cargamos el conjunto de datos:

```
data("HIV")
HIV
```

```
##       c1 c2 c3 c4 freq
##  [1,]  1  1  1  1    0
##  [2,]  1  1  1  0    3
##  [3,]  1  1  0  1    1
##  [4,]  1  1  0  0   33
##  [5,]  1  0  1  1    0
##  [6,]  1  0  1  0   20
##  [7,]  1  0  0  1    6
##  [8,]  1  0  0  0  403
##  [9,]  0  1  1  1    3
## [10,]  0  1  1  0   35
## [11,]  0  1  0  1   10
## [12,]  0  1  0  0  545
## [13,]  0  0  1  1   11
## [14,]  0  0  1  0  621
## [15,]  0  0  0  1  205
```

Vemos que la estructura de este conjunto de datos es del tipo 2. Contiene la matriz de ceros y unos y la tabla de frecuencias. Por tanto, a lo largo de nuestro análisis, debemos tener en cuenta que se debe considerar dfreq=TRUE

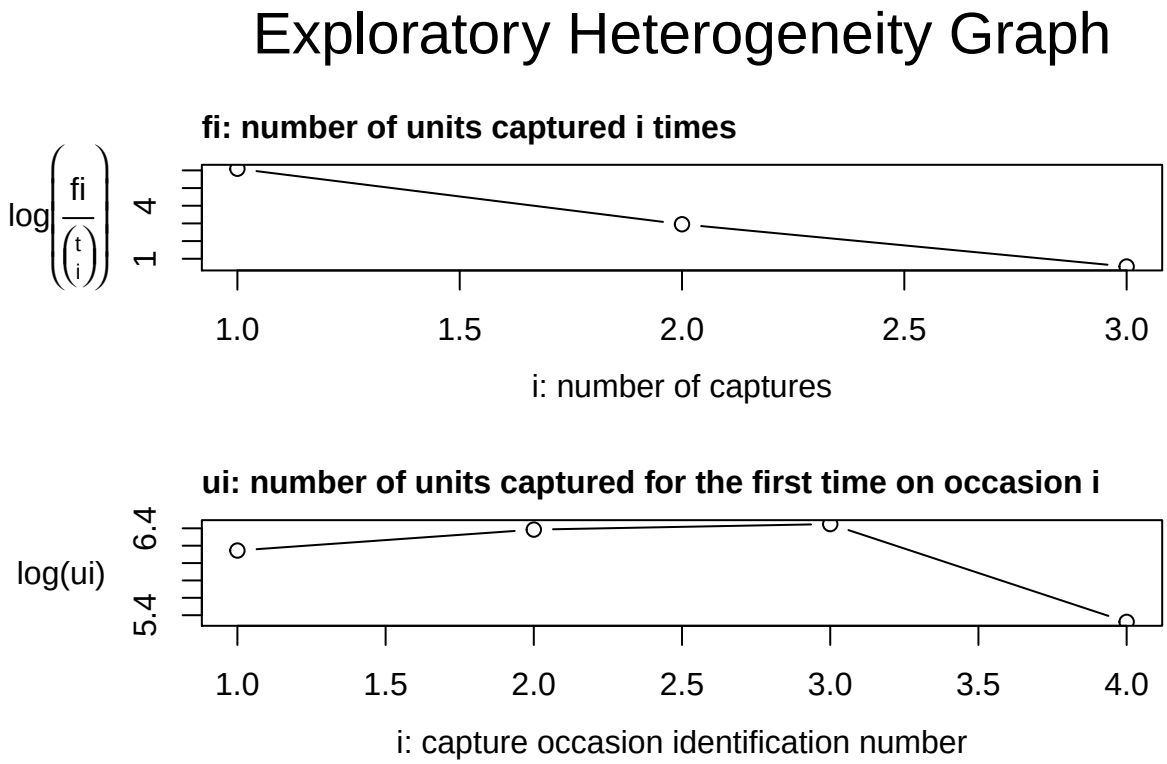
Realizamos el análisis descriptivo

```
descriptive(HIV,dfreq=TRUE)
```

```
##
## Number of captured units: 1896
##
## Frequency statistics:
##         fi    ui    vi    ni
## i = 1  1774   466   403   466
## i = 2   115   593   578   630
## i = 3     7   632   679   693
## i = 4     0   205   236   236
```

Veamos el gráfico

```
desc<-descriptive(HIV,dfreq=TRUE)
plot(desc)
```




```
mat<-histpos.t(4) #t=4 porque es el número de capturas
mX1<-cbind(mat,mat[,1]*mat[,2],mat[,1]*mat[,3],mat[,1]
*mat[,4],mat[,2]*mat[,3],mat[,2]*mat[,4],mat[,3]*mat[,4])
```

Para el ajuste del modelo, utilizaremos closedp.mX para usar nuestra matriz de diseño. Los argumentos que necesita esta función son el conjunto de datos, dfreq=TRUE, ya que nuestro conjunto de datos utiliza las frecuencias y la matriz de diseño

```
cp.m1<-closedp.mX(HIV,dfreq=TRUE,mX=mX1,mname=
"Interacciones dobles Mt")
str(cp.m1)
```

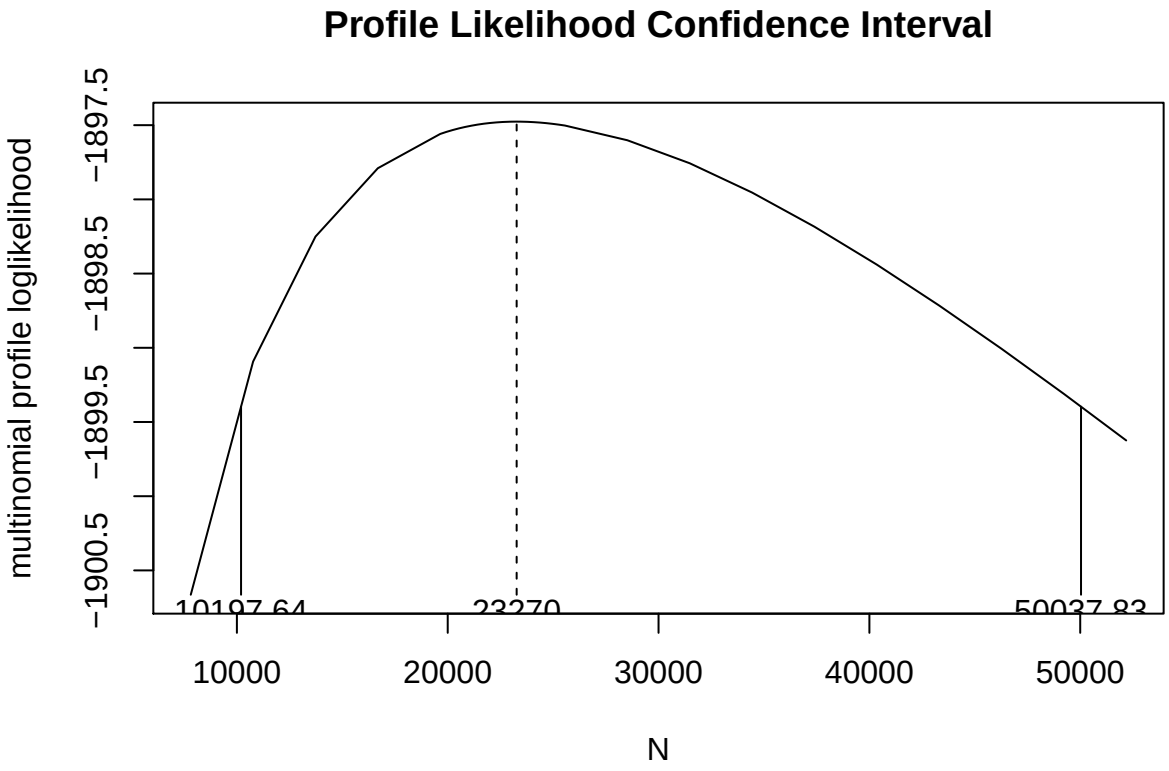
Queremos conocer los resultados

```
cp.m1$results
```

```
##                          abundance    stderr deviance df      AIC
## Interacciones dobles Mt  23443.54 9594.879 3.036804  4 92.07266
```

Se estima que los enfermos de VIH en la ciudad en cuestión es de 23443. Calculamos ahora un intervalo de confianza

```
CI<-profileCI(HIV,dfreq=TRUE,mX=mX1)
```




```
CI$results
```

```
##                    abundance     InfCL     SupCL
```

```
## Customized model     23270 10197.64 50037.83
```

El intervalo de confianza es [10197, 50038]

Ahora realizaremos algunos estudios con distintos conjuntos de datos para poblaciones abiertas. Al igual que ocurría en el caso de poblaciones cerradas, hay un procedimiento general a seguir a la hora de hacer un análisis de un conjunto de datos en los que se supone que la población es abierta.

1-Realizamos un análisis descriptivo de los datos. Realizamos también gráficos que nos ayuden a visualizar los resultados

2-Ajuste del modelo y estimación de los parámetros. Para ello, considerando que ahora vamos a tratar poblaciones abiertas, utilizaremos la función openp, que viene ya implementada en el paquete Rcapture

3-Estudio del ajuste del modelo

4-Realizamos un análisis más extenso utilizando la normalidad multivariante.

El primer conjunto de datos que utilizaremos es lesbian. Este conjunto de datos contiene datos epidemiológicos de captura y recaptura sobre la población lesbiana de cuatro organizaciones que sirven a la población gay y lesbiana del condado de Allegheny, Pennsylvania, Estados Unidos

Cargamos el conjunto de datos y vemos qué formato tiene

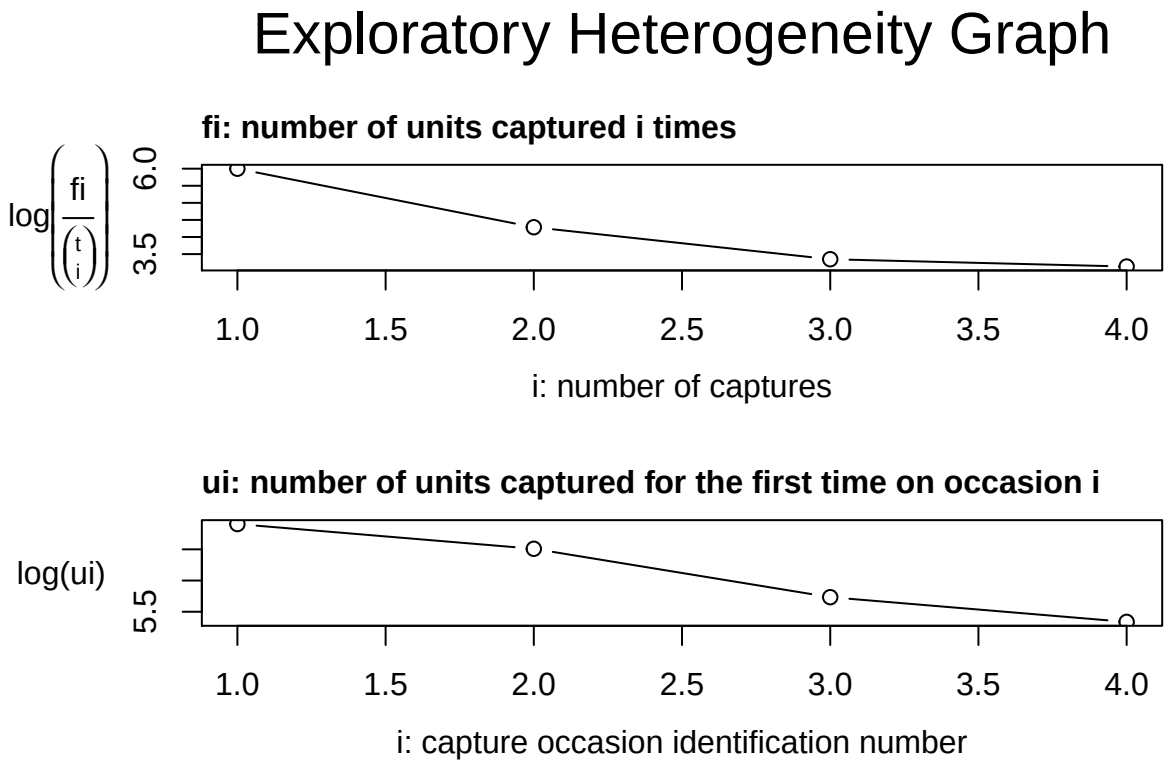
```
data("lesbian")
lesbian #El formato de este conjunto es de tipo 2
```

Realizamos el análisis descriptivo. Como tenemos el formato 2, debemos añadir la orden dfreq=TRUE.

```r
desc<-descriptive(lesbian,dfreq=TRUE)
desc
```

```
##
## Number of captured units: 2185
##
## Frequency statistics:
##         fi    ui    vi    ni
## i = 1  1612   997   589   997
## i = 2   436   671   638   873
## i = 3   114   309   416   506
## i = 4    23   208   542   542
```

```r
plot(desc)
```



Realizamos ahora las estimaciones. A la función openp debemos de darle como argumentos el conjunto de datos y dfreq=TRUE para indicarle que en nuestro conjunto de datos ya van incluidas las frecuencias

```
op<- openp(lesbian,dfreq=TRUE)
op
```

```
##
## Model fit:
##                deviance      df        AIC
## fitted model    139.439       6    249.792
##
## Test for trap effect:
```

```
##                                     deviance      df        AIC
## model with homogenous trap effect    107.908       5    220.261
##
## Capture probabilities:
##           estimate  stderr
## period 1        --      --
## period 2    0.2088  0.0188
## period 3    0.1256  0.0149
## period 4        --      --
##
## Survival probabilities:
##                estimate  stderr
## period 1 -> 2    0.9702  0.0629
## period 2 -> 3    0.9576  0.1006
## period 3 -> 4        --      --
##
## Abundances:
##           estimate  stderr
## period 1        --      --
## period 2    4180.3   354.2
## period 3    4029.6   446.4
## period 4        --      --
##
## Number of new arrivals:
##                estimate  stderr
## period 1 -> 2        --      --
## period 2 -> 3      26.4   322.1
```

```
## period 3 -> 4        --      --
##
## Total number of units who ever inhabited the survey area:
##               estimate  stderr
## all periods    4061.8   202.9
##
## Total number of captured units: 2185
```

Obtenemos estimaciones no solo de la abundancia, si no también de la probabilidad de captura (que lógicamente es desconocida), de la probabilidad de supervivencia y de los individuos que se han introducido en la población (imigración)

Este análisis ha sido bastante básico. Consideremos ahora otro conjunto de datos con la misma estructura que el anterior. Utlizamos el conjunto de datos duck, que contiene información sobre datos de captura y recaptura de una determinada especie de pato. En primer lugar, como siempre, caragamos el conjunto de datos y comprobamos qué estructura tiene.

```
data("duck")
duck #Este conjunto de datos corresponde al formato 2
```

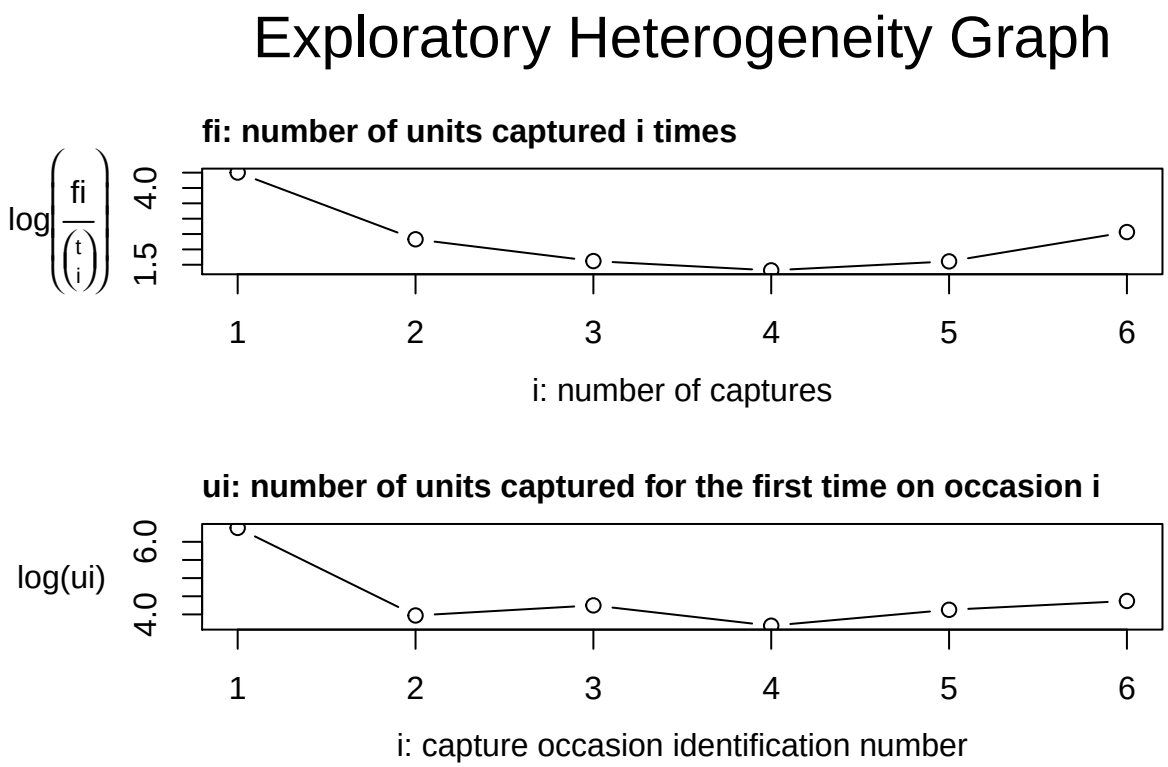
Realizamos un análisis descriptivo de los datos

```
desc<-descriptive(duck,dfreq=TRUE)
desc
```

```
##
## Number of captured units: 896
##
## Frequency statistics:
##        fi   ui   vi   ni
```

```
## i = 1  542  592  362  592
## i = 2  154   53   43  176
## i = 3  101   70   83  211
## i = 4   56   40   41  159
## i = 5   30   62   90  190
## i = 6   13   79  277  277
```

```
plot(desc)
```



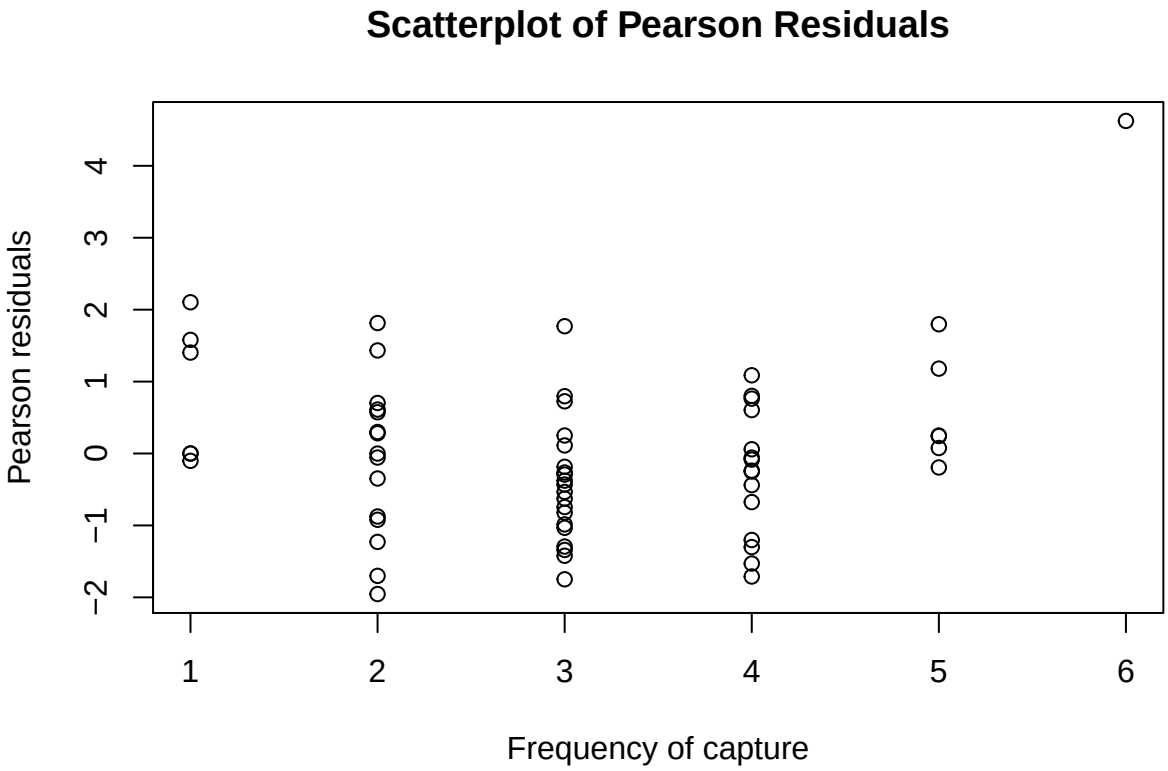

```
op.m1<-openp(duck,dfreq=TRUE)
op.m1
```

```
##
```

```
## Model fit:
##               deviance      df        AIC
## fitted model     83.36      49     328.83
##
## Test for trap effect:
##                                     deviance      df        AIC
## model with homogenous trap effect    82.526      48    329.996
## model with trap effect               78.049      46    329.519
##
## Capture probabilities:
##           estimate  stderr
## period 1        --      --
## period 2    0.4649  0.0346
## period 3    0.4635  0.0354
## period 4    0.4457  0.0338
## period 5    0.4170  0.0306
## period 6        --      --
##
## Survival probabilities:
##                estimate  stderr
## period 1 -> 2    0.4469  0.0256
## period 2 -> 3    0.9578  0.0523
## period 3 -> 4    0.7135  0.0423
## period 4 -> 5    1.0000  0.0000
## period 5 -> 6        --      --
##
## Abundances:
```

```
##           estimate  stderr
## period 1        --      --
## period 2     378.6    18.9
## period 3     455.2    26.1
## period 4     356.7    17.0
## period 5     455.7    21.9
## period 6        --      --
##
## Number of new arrivals:
##                estimate  stderr
## period 1 -> 2        --      --
## period 2 -> 3      92.6    23.5
## period 3 -> 4      31.9    18.5
## period 4 -> 5      99.0    23.6
## period 5 -> 6        --      --
##
## Total number of units who ever inhabited the survey area:
##              estimate  stderr
## all periods       962      12
##
## Total number of captured units: 896
```

```
plot(op.m1)
```



Calculamos el p-valor de bondad de ajuste basado en la desviación

```
1-pchisq(op.m1$model.fit[1,1],df=49) #df=49 grados de libertad
```

p-valor=0.001592682. Este p-valor es demasiado bajo, hacemos un estudio nuevo eliminando las observaciones que han sido capturadas las 6 veces, que son 13 en total según nuestro análisis descriptivo.

```
desc2<-apply(histpos.t(6),1,sum)!=6
op.m2<-openp(duck,dfreq=TRUE,keep=desc2)
                                    op.m2
```

```
##
## Model fit:
```

```
##                deviance      df        AIC
## fitted model     67.311      48    308.366
##
## Test for trap effect:
##                                      deviance      df        AIC
## model with homogenous trap effect      67.304      47    310.359
## model with trap effect                 63.038      45    310.092
##
## Capture probabilities:
##           estimate  stderr
## period 1        --      --
## period 2    0.4385  0.0354
## period 3    0.4367  0.0362
## period 4    0.4152  0.0346
## period 5    0.3900  0.0311
## period 6        --      --
##
## Survival probabilities:
##                 estimate  stderr
## period 1 -> 2     0.4422  0.0265
## period 2 -> 3     0.9655  0.0574
## period 3 -> 4     0.7092  0.0454
## period 4 -> 5     1.0000  0.0000
## period 5 -> 6         --      --
##
## Abundances:
##           estimate  stderr
```

```
## period 1        --      --
## period 2     388.3    20.8
## period 3     470.0    28.9
## period 4     368.6    19.1
## period 5     471.2    24.5
## period 6        --      --
##
## Number of new arrivals:
##                estimate  stderr
## period 1 -> 2        --      --
## period 2 -> 3      95.1    25.5
## period 3 -> 4      35.3    20.3
## period 4 -> 5     102.6    25.9
## period 5 -> 6        --      --
##
## Total number of units who ever inhabited the survey area:
##              estimate  stderr
## all periods     971.5    13.1
##
## Total number of captured units: 896
```

```
1-pchisq(op.m2$model.fit[1,1],df=48)
```

```
## [1] 0.03427131
```

p-valor=0.03427131. Aunque hemos mejorado bastante, sigue sin ser del todo satisfactorio. Quitamos también las observaciones que han sido seleccionadas 5 de las 6 veces

```r
desc3<-apply(histpos.t(6),1,sum)<5
op.m3<-openp(duck,dfreq=TRUE,keep=desc3)
op.m3
```

```
##
## Model fit:
##                deviance      df         AIC
## fitted model     56.831      42     277.201
##
## Test for trap effect:
##                                       deviance      df         AIC
## model with homogenous trap effect       56.499      41     278.870
## model with trap effect                  52.337      39     278.708
##
## Capture probabilities:
##           estimate  stderr
## period 1        --      --
## period 2    0.4111  0.0382
## period 3    0.4131  0.0394
## period 4    0.3728  0.0373
## period 5    0.3552  0.0333
## period 6        --      --
##
## Survival probabilities:
##                 estimate  stderr
## period 1 -> 2     0.4360  0.0282
## period 2 -> 3     0.9659  0.0650
```

```
## period 3 -> 4    0.7076  0.0508
## period 4 -> 5    1.0000  0.0000
## period 5 -> 6        --      --
##
## Abundances:
##           estimate  stderr
## period 1        --      --
## period 2     395.8    23.0
## period 3     483.6    32.4
## period 4     386.9    22.8
## period 5     494.2    29.0
## period 6        --      --
##
## Number of new arrivals:
##                estimate  stderr
## period 1 -> 2        --      --
## period 2 -> 3     101.3    27.7
## period 3 -> 4      44.7    23.1
## period 4 -> 5     107.3    29.8
## period 5 -> 6        --      --
##
## Total number of units who ever inhabited the survey area:
##              estimate  stderr
## all periods     985.4      15
##
## Total number of captured units: 896
```

```
1-pchisq(op.m3$model.fit[1,1],df=42)
```

```
## [1] 0.06298297
```

En este caso el p-valor es 0.06298297. Aunque es mejor aún que el anterior, sigue siendo bajo. Para investigar si las probabilidades de captura son homogéneas, ajustamos un modelo con igual probabilidad de captura. Introducimos la instrucción m=ep para la homogeneidad de las probabilidades.

```
op.m4<-openp(duck,dfreq=TRUE,keep=desc3,m="ep")


op.m4$model.fit[1,]
```

```
##  deviance        df       AIC
##  117.9115   47.0000  328.2822
```

La desviación es mucho más grande, por lo que rechazamos la hipótesis. El modelo con el que nos quedaremos es en el que eliminamos las observaciones que han sido observadas cinco o seis veces.

Por último vamos a calcular las tasas de crecimiento, que las definimos tal y como están definidas en la teoría.

```
growth<-op.m3$N[3:5,1]/op.m3$N[2:4,1]
partial<-matrix(c(-op.m3$N[3,1]/op.m3$N[2,1]^2,1/op.m3$N[2,1],0,0,
0,-op.m3$N[4,1]/op.m3$N[3,1]^2,1/op.m3$N[3,1],0,
0,0,-op.m3$N[5,1]/op.m3$N[4,1]^2,1/op.m3$N[4,1]),3,4,byrow=TRUE)
sig<-partial%*%op.m3$cov[9:12,9:12]%*%t(partial)
cbind(estimate=growth,stderr=sqrt(diag(sig)))
```

```
##           estimate     stderr
## period 3 1.2218081 0.11281160
```

```
## period 4 0.8000063 0.07563975
## period 5 1.2771890 0.08512487
```

```
#####################################################
siginv<-solve(sig)
growth.e<-t(rep(1,3))%*%siginv%*%growth/(t(rep(1,3))
%*%siginv%*%rep(1,3))
se<-1/sqrt(t(rep(1,3))%*%siginv%*%rep(1,3))
data.frame(estimate=growth.e,stderr=se,row.names=
"Tasa de crecimiento común: ")
```

```
##                              estimate     stderr
## Tasa de crecimiento común:  1.037558 0.03187539
```

```
data.frame(stat=chisq2,pvalue=1-pchisq(chisq2,df=2),row.names=
"Test chi-cuadrado: ")
```

```
##                         stat      pvalue
## Test chi-cuadrado:  13.53338 0.001151498
```

Conclusión

Como conclusión a este TFG diremos que estos métodos, gracias a su avance y a que han sido fuertemente potenciados a lo largo del siglo XX, resultan de gran utilidad en un buen porcentaje de estudios estadísticos en campos tan diversos como la epidemiología, la biología o las ciencias sociales. No siempre es posible conocer el tamaño de una población, pero sin embargo en muchos estudios es imprescindible conocerlo para poder realizar inferencia en ella. Es ahí donde radica la importancia de estos métodos. Interesa, por tanto, tener distintos modelos que nos proporcionen estimadores diferentes en función de las características que tenga nuestra población y del tiempo que vaya a tardar nuestro estudio.

Además, resulta interesante saber que a partir de los modelos más básicos, se pueden construir modelos que estimen datos poblacionales como el crecimiento de la población o las tasas de natalidad y mortalidad. El interés de esto último es que a partir de ahí se puede comenzar a estudiar la dinámica de una población, por lo que a partir de estos métodos se puede llegar a puntos que combinan estadística y ecuaciones diferenciales, algo que ayuda a realizar estudios que son enriquecedores para el mundo de las matemáticas.

Referencias

[1] Badii, M.H., Guillen, A., Landeros, J.,Cerna, E., Ochoa, Y. y Valenzuela, J., *Muestreo por métodos de Captura-Recaptura (Sampling via Capture-Recapture Methods)*, Saltillo Coah, México, 2012.

[2] Baillargeon,S. y Rivest,L.P., *Rcapture: Loglinear Models for Capture-Recapture in R*, Université Laval, Québec, volumen 19, 2007

[3] Brownie,C., Anderson,D.R., Burnham, K.P., Robson, D.S, *Statistical inference from band recovery data: a handbook*, Washington D.C., 1985

[4] Bryan F., Manly J., Trent L.McDonald y Amstrup,S.C., *Handbook of Capture-Recapture Analysis*, Princeton University Press, 2005

[5] Carothers,A.D., *Capture-Recapture Methods Applied to a Population with Known Parameters*, Universidad de Edimburgo, 1973

[6] Chapman, D.G., *Some properties of the hypergeometric distribution with applications to zoological sample censuses.*,Berkeley, University of California Press, 1951

[7] Darroch,J.N, *The multiple recapture census: I*, págs 343-359, 1958

[8] Jolly, G.M., «Explicit Estimates from Capture-Recapture Data with Both Death and Inmigration-Stochastic Model», *Biometrika* volumen 52,**1** págs 225-247, 1965.

[9] Otis,D.L., Burnham,K.P., White, G.C., Anderson, D.R., *Statistical Inference from Capture Data on Closed Animal Populations*, 1978

[10] Pollock,K.H., Nichols,J.D., Brownie, C., Hines,J.E., textitStatistical Inference for Capture-Recapture Experiments, 1990

[11] Schwarz, C.J., «The Jolly-Seber Model: More than just abundance», *Journal of Agricultural, Biological and Environmental Statistics*, volumen 6 **2**, págs. 195–205, 2001.

[12] Seber,G.A.F., «A Note on the Multiple-Recapture Census», *Biometrika* volumen 52,**1** págs 249-259, 1965.

[13] Seber,G.A.F., *The estimation of animal abundance and related parameters*, 1982