Tutorial 2: Abundancia y la distribución multinomial*

Curso: Métodos de captura-recaptura, UNAM. Abril, 2010.

Roberto E. Munguía-Steyer
Dpto. Ecologia, IB-USP, Brasil.
rmunguia.steyer@gmail.com

1. Supuestos del modelo de abundancia clásico M_0

Los modelos más sencillos que tratan con la inferencia de la abundancia poblacional tienen los siguientes supuestos:

1. Durante el estudio, el tamaño de la población es constante. Se dice que la población se encuentra **cerrada** (no ocurren nacimientos, muertes, migraciones).
2. Todos los individuos tienen la misma probabilidad de ser capturados y es constante en el tiempo.
3. No existen diferencias entre la probabilidad de captura y recaptura (efectos comportamentales).
4. El marcar los individuos no afecta la probabilidad de recapturarlos.
5. La probabilidad de capturar un individuo es independiente de si otro individuo en particular fue capturado o no.
6. Las marcas colocadas en los individuos no se pierden (e.g. caen, borran).

En la literatura estos modelos son conocidos como Petersen ($k = 2$) o Schnabel ($k > 2$), siendo k el número de sesiones de captura-recaptura [1, 5].

2. Distribución binomial y las probabilidades de captura

En el tutorial pasado vimos que las historias de encuentro que construimos con los datos de ausencia presencia en nuestra población de interés forman las cadenas hechas de secuencias de ceros y unos. Adicionalmente, aprendimos que la composición de la historia de encuentro (cuántos unos) está determinada por la probabilidad de captura p.

Pensemos que nosotros, por algún pacto con la divinidad somos seres omniscientes y sabemos que la población de lagartijas de nuestro particular interés consta de 180 individuos. Vemos con pena a los herpetólogos rivales del mismo departamento que harán la estimación con métodos de captura-recaptura por desconfiar de nuestras conclusiones. Ellos construirán la historia de encuentros marcando los individuos y registrando la presencia-ausencia de los individuos marcados durante los tres días que durará el estudio. Como sabemos que la probabilidad de captura es de 0.5 sonreímos con malicia al conocer cuantos individuos en la población no serán atrapados y serán el objeto de esa obscura inferencia estadística para estimar la abundancia que de antemano conocemos.

```
> set.seed(111)
> N <- 180
> p.recap <- 0.5
> k = 3
> celdas <- N * k
> a <- matrix(rep(0, celdas), N, k)
> for (i in 1:N) a[i, ] <- rbinom(k, 1, p = p.recap)
> lagartija <- data.frame(a)
> names(lagartija) <- paste("c", 1:k, sep = "")
```

Veamos las primeras historias de encuentro de la población de lagartijas:

```
   c1 c2 c3
1   1  1  0
2   1  0  0
3   0  1  0
4   0  1  1
5   0  0  0
6   0  0  1
7   0  1  0
8   0  0  0
9   1  0  1
10  0  1  1
```

Existen 2^k posibles cadenas de presencia-ausencia siendo en este caso ocho ($2^3 = 8$).

Historia de Encuentros	Probabilidad p_i
100	$p_1(1-p_2)(1-p_3)$
010	$(1-p_1)p_2(1-p_3)$
001	$(1-p_1)(1-p_2)p_3$
110	$p_1p_2(1-p_3)$
101	$p_1(1-p_2)p_3$
011	$(1-p_1)p_2p_3$
111	$p_1p_2p_3$
000	$(1-p_1)(1-p_2)(1-p_3)$

Las primeras siete cadenas serán las posibles historias de encuentro que resultarán de nuestra toma de datos, la última cadena son los individuos que jamás capturamos, los individuos no detectados o f_0. Estos son los individuos que necesitamos estimar para conocer la abundancia. Para estimar f_0 necesitamos utilizar la distribución multinomial.

3. Sobre dados y la distribución multinomial

Al igual que en el primer tutorial donde demostramos que la historia de encuentros compuesta por ceros y unos, de nueva cuenta tenemos resultados complementarios (sumados presentan probabilidad = 1), discretos y mutuamente excluyentes. Solo que esta vez no tenemos dos posibles resultados (presencia o ausencia del individuo) propios de la variable binomial sino más de dos por lo cual usaremos la variable aleatoria con **distribución multinomial** [6].

En este caso, en vez de lanzar una moneda al aire y determinar la probabilidad de determinado evento, lanzaremos un dado con el número de lados fluctuante y dependiente de las posibles historias de encuentros. Recordemos que las posibles historias de encuentro $= 2^k$.

Empezemos con el dado tradicional de seis caras, la función de probabilidad de una variable multinomial es:

$$(f_i|n,y_i) = \binom{n}{y_i} p_1^{y1} p_2^{y2} p_3^{y3} p_4^{y4} p_5^{y5} p_6^{y6} \tag{1}$$

Donde el coeficiente multinomial representa:

$$\binom{n}{y_i} = \frac{n!}{y_1!y_2!\cdots y_k!} \tag{2}$$

Lanzamos 10 veces un dado convencional (seis caras y no cargados hacia ningún lado). ¿Cuál sería la probabilidad de sacar seis unos y cuatro dos?

$$f(6,4,0,0,0,0|10,1/6,1/6,1/6,1/6,1/6,1/6) = \frac{10!}{6!\,4!\,(0!)^8}\left(\frac{1}{6}\right)^{10} \tag{3}$$

La probabilidad es muy baja.

```
[1] 3.473e-06
```

¿y la probabilidad de sacar en 6 tentativas las seis distintas caras del dado?

$$f(1,1,1,1,1,1|6,1/6,1/6,1/6,1/6,1/6,1/6) = \frac{6!}{(1!)^6}\left(\frac{1}{6}\right)^6 \tag{4}$$

```
[1] 0.01543
```

4. Función de máxima verosimilitud y la distribución multinomial

Al analizar las historias de captura-recaptura, nosotros desconocemos el valor de probabilidad de captura, resultado de la frecuencia de las posibles historias de encuentro, los dados de nuestro dado en el ejemplo anterior. Más bien invertimos el proceso y estimamos los parámetros de captura a partir de nuestros datos.

La función de verosimilitud de la distribución multinomial es[1]:

$$\mathscr{L}(p_i|n_i,y_i) = \binom{n}{y_i}\prod p_i^{yi} \tag{5}$$

Podemos estimar la probabilidad de captura que usamos en la simulación de la historia de encuentros relativo a la población de lagartijas en la sección 2 del tutorial. Cabe recordar que en el modelo M_0 las probabilidades de captura no cambian en función del tiempo por lo cual sólo estimaremos un parámetro p. La analogía sería pensar que el dado no se encuentra cargado, por tanto todos los lados del dado tienen la misma probabilidad de caer hacia arriba.

```
> k.0 <- c(0, 1, 2, 3)
> num.cap <- apply(lagartija, 1, sum)
> nk <- table(num.cap)
```

- Número de individuos capturados n veces:

```
> nk
```

```
num.cap
 0  1  2  3
22 57 77 24
```

- Función de verosimilitud y optimización para encontrar la máxima verosimilitud del parámetro:

```
> llik <- function(param) {
+     p <- plogis(param[1])
+     -1 * sum(nk * log(dbinom(k.0, 3, p)))
+ }
> m.lag <- optim(par = 0, fn = llik, method = "BFGS", hessian = TRUE)
```

[1] Así como el símbolo $\sum$ se usa para la representación de la suma de los elementos de un vector, $\prod$ representa la multiplicación de los elementos indexados en el vector.

- Probabilidad de captura estimada e intervalos de confianza al 95 %:

```
> p.est <- plogis(m.lag$par)
> p.se <- sqrt(1/m.lag$hessian)
> z <- qnorm(0.975)
> ic.cru <- m.lag$par + c(-1, 1) * z * p.se
> ic <- plogis(ic.cru)
> p.est

[1] 0.5241

> ic

[1] 0.4819 0.5659
```

La probabilidad de captura estimada se encuentra muy próxima al valor que asignamos en la simulación de nuestra población de lagartijas $p = 0.5$, el cual se encuentra dentro de los intervalos de confianza al 95 %.

5. Estimación de la abundancia: Modelo M_0

En la vida real, cuando vamos al campo a colectar la información de nuestra especie animal de interés, nosotros no tenemos registro del número de individuos no capturados en la población representados por el símbolo f_0. Por lo tanto, tendremos que usar una función de verosimilitud que estime f_0 además del parámetro de captura. Aqui aprovechamos la complementaridad y exclusividad de los resultados posibles de una variable multinomial, conociendo que la estimación de los valores de probabilidad provenientes de la frecuencia de esa historia de encuentros f_0 es complementaria con la suma de las siete restantes. Por lo tanto, la función de verosimilitud es la siguiente:

$$\mathscr{L}(N, p|\mathbf{y}, n) = \frac{N!}{(N-n!)} p^{\Sigma_{i=1}^{n} y_i} (1-p)^{J \cdot N - \Sigma_{i=1}^{n} y_i} \tag{6}$$

Con una probabilidad de captura de 0.5, la probabilidad de capturar a un individuo al menos una vez durante las tres visitas es: $1-(1-0.5)^3 = 0.875$. Por lo tanto, por complementaridad la probabilidad de no capturar en las tres ocasiones a un individuo es de 0.125. Recordemos que nuestra población de lagartijas está conformada por 180 individuos, el numero esperado de individuos no capturados sería 180 * 0.125 = 22.5. Este valor se ajusta a nuestras expectativas, recordando que en la sección 4 descubrimos que 22 individuos se encuentran en esa condición.

Removamos a ese subconjunto de individuos pertenecientes a f_0 para hacer la estimación procediendo como si hubiéramos colectado la información de la población a partir de la historia de encuentros.

```
> rem0 <- which(num.cap == 0)
> num0 <- length(rem0)
```

```
> lag <- lagartija
> ifelse(num0 == 0, lag2 <- lag, lag2 <- lag[-rem0, ])

> dim(lag)

[1] 180   3

> dim(lag2)

[1] 158   3

> num0

[1] 22
```

Procedemos a estimar f_0 y comparar el valor estimado con el real (n=22) mediante la siguiente función de verosimilitud [3, 6].

```
> nvec <- c(57, 77, 24)
> pop.cerrada <- function(parms) {
+     p <- plogis(parms[1])
+     n0 <- exp(parms[2])
+     N <- sum(nvec) + n0
+     cpvec <- dbinom(0:3, 3, p)
+     -1 * (lgamma(N + 1) - lgamma(n0 + 1) + sum(c(n0, nvec) *
+         log(cpvec)))
+ }
```

Estimamos los valores de los parámetros p y f_0 por máxima verosimilitud usando el paquete `bbmle` desarrollado por Ben Bolker[2]:

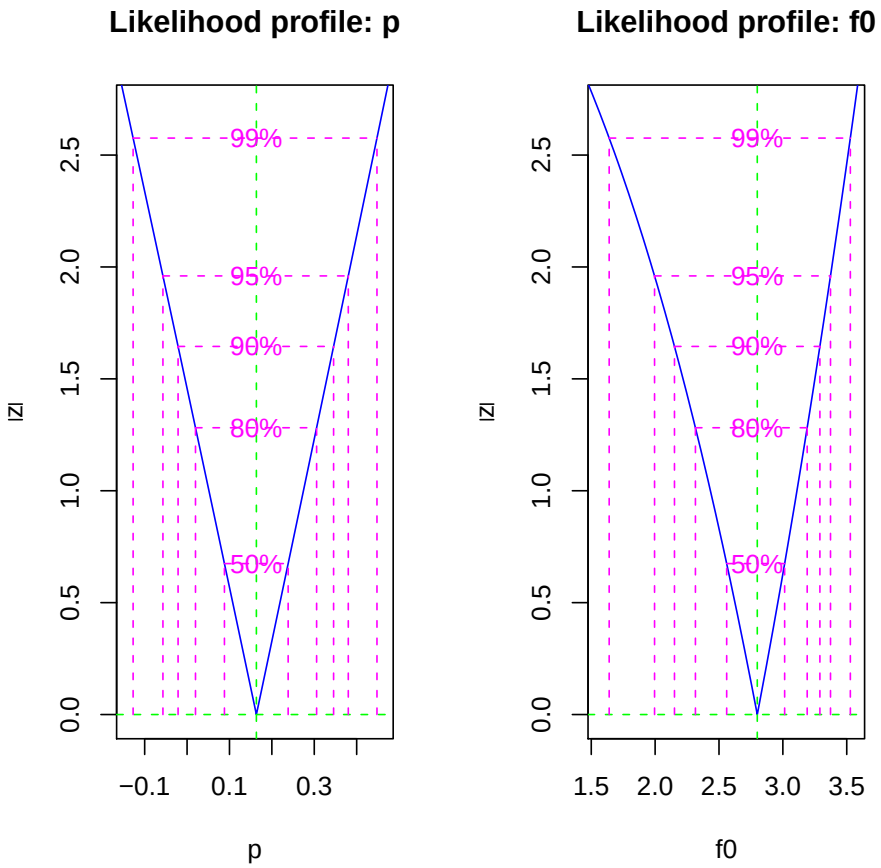
```
> library(bbmle)
> parnames(pop.cerrada) <- c("p", "f0")
> m1 <- mle2(pop.cerrada, start = c(p = 0, f0 = 3), data = list(nvec = nvec))
> intconf <- confint(m1)
> plogis(coef(m1)[1])

     p
0.5408

> plogis(intconf)[1, ]

 2.5 % 97.5 %
0.4858 0.5939
```

[2]Se pueden instalar los paquetes `bbmle` y `Rcapture` desde la consola con la función `install.packages()`.

```
> exp(coef(m1)[2])

   f0
16.44

> exp(intconf)[2, ]

 2.5 % 97.5 %
 7.353 29.160
```

Los intervalos de confianza al 95 % tanto de la probabilidad de captura p como el número de individuos no capturados f_0 abarcan los parámetros usados en la simulación de la población de lagartijas. La abundancia total estimada de la población sería el número de individuos capturados más $f_0 = 16 + 77 + 57 + 24 = 174$ individuos. Cabe hacer notar que con tres visitas y una probabilidad de captura = 0.5, el intervalo de confianza de los individuos no capturados es muy amplio y el valor estimado menos confiable dada la evidente asimetría que muestra el perfil de verosimilitud. Mientras más parecido a una V sea el perfil de verosimilitud, los intervalos de confianza por aproximación cuadrática de los parámetros son más confiables. La forma del perfil depende de la curvatura de la superficie de verosimilitud.

```
> plot(profile(m1))
```



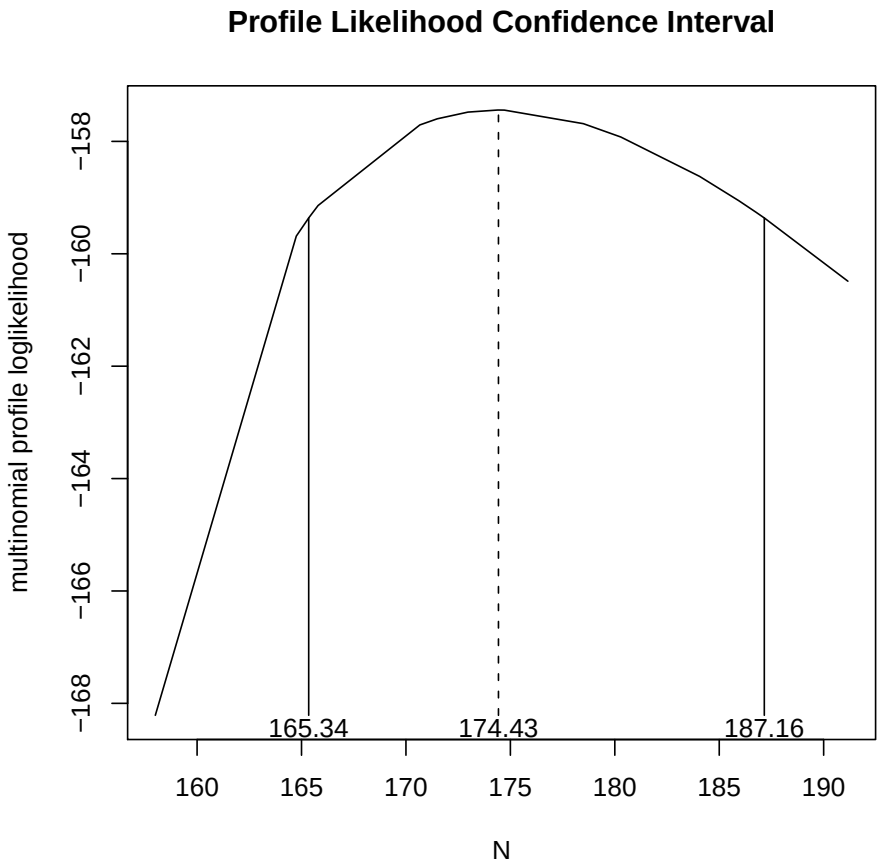
6. Cálculo de la abundancia del modelo M_0 usando el paquete `Rcapture`

Por último, usemos con fines de comparación el paquete `Rcapture` [2] también desarrollado para estimar abundancia:

```
> library(Rcapture)
> cpop <- closedpCI.0(lag2, h = "Poisson")
> cpop

Number of captured units: 158

Poisson estimation and model fit:
    abundance  stderr  deviance  df    AIC
M0      175.3     5.6     1.224   1  22.31

Multinomial estimation, 95% profile likelihood confidence interval:
    abundance  InfCL  SupCL
M0      174.4  165.3  187.2
```

Podemos observar que en este caso, el modelo de abundancia con distribución multinomial subestima la abundancia de la población. El modelo estima que la población cuenta con 174 individuos (exactamente el mismo número de nuestro código anterior) y el intervalo de confianza de la abundancia es muy amplio.

```
> plotCI(cpop)
```



`Rcapture` tiene una serie de funciones útiles que permiten una estimación rápida de la abundancia a partir de la selección de una serie de modelos mediante el uso del Criterio de Información de Akaike (AIC) [3, 4][3]. La desventaja de este paquete es su falta de flexibilidad para ajustar la misma cantidad de modelos que `MARK` o bien especificando las funciones de verosimilitud y optimizando "a mano" (por ejemplo escribiendo el código pertinente en R o en `MATLAB`). Otro punto a considerar es que no reporta las probabilidades de recaptura, información útil en muchos casos y sujeta a modelación con potenciales variables predictoras (tamaño, edad, etc).

Referencias

[1] S.C. Amstrup, T.L. McDonald, and B.F.J. Manly. *Handbook of capture-recapture analysis*. Princeton University Press, 2005.

[2] S. Baillargeon and L.P. Rivest. Rcapture: Loglinear Models for Capture-Recapture in R. *Journal of Statistical Software*, 19(5):1–31, 2007.

[3] B.M. Bolker. *Ecological models and data in R*. Princeton University Press, 2008.

[3] El tema será tratado con más detalle en el tutorial 3.

[4] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

[5] C.J. Krebs. *Ecological methodology*. Benjamin-Cummings, 1998.

[6] J.A. Royle and R.M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, 2008.