Tutorial 1: Probabilidad y máxima verosimilitud*

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

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

1. La probabilidad y su relación con los métodos de captura-recaptura

Una variable aleatoria es aquella que puede tomar más de un valor, el cual depende de una función de probabilidad determinada [2, 4]. Pensemos en el clásico experimento de lanzar una moneda al aire. La moneda puede tomar dos valores discretos y complementarios: águila (p) o sol ($1-p$). Por tal motivo, el resultado de lanzar una moneda al aire es una variable aleatoria con distribución binomial. Una moneda convencional ("no sesgada o cargada") tiene la misma probabilidad de caer águila o sol: ($p = 1-p = 0.5$). La variable aleatoria de la distribución binomial es $y \sim Bin(N, p)$, presenta un valor esperado de $E(y) = Np$, donde p es la probabilidad de caer águila y N el número de veces que lanzamos la moneda. La función de probabilidad de una variable aleatoria binomial es:

$$f(y|N,p) = \binom{N}{y} p^y (1-p)^{N-y} \tag{1}$$

Lancemos ocho veces la moneda al aire[1]:

```
> set.seed(123)
> volado <- rbinom(8, 1, 0.5)
> volado
```

```
[1] 0 1 0 1 1 0 1 1
```

Cuando lanzamos pocas veces la moneda al aire, el resultado no necesariamente refleja el valor esperado (cuatro águilas y cuatro soles). Sin embargo, al aumentar el número de tentativas, por ejemplo, modificando el número de volados a diez mil:

[1] El código de este y los siguientes tutoriales ha sido escrito en el lenguaje de programación R. Para mayor información sobre R ver: http://www.r-project.org.

```
> mat.vol <- matrix(rep(0, 80000), 10000, 8)
> for (i in 1:10000) {
+     mat.vol[i, ] <- rbinom(8, 1, 0.5)
+ }
```

El valor esperado tenderá asintóticamente a ser de cuatro águilas:

```
> suma <- apply(mat.vol, 1, sum)
> agui.esp <- mean(suma)
> agui.esp

[1] 3.965
```

Al ir a campo y colectar nuestra información para estimar la abundancia de una población, nosotros colectamos una historia de encuentros. Nuestra historia de encuentros recopilada en las sesiones de recapturas también están hechas de cadenas de 1 y 0s y su composición está determinada por la probabilidad de recaptura.

Diez individuos que conforman "una población" al tener una probabilidad de recaptura alta ($p = 0.7$) tendrán una mayor oportunidad de ser detectados al menos una vez que si tuvieran una probabilidad baja ($p = 0.2$).

Número de visitas $k = 6$.

```
> p1 = 0.7
> p2 = 0.2
> pob.a <- data.frame(matrix(rbinom(60, 1, p1), 10, 6))
> pob.b <- data.frame(matrix(rbinom(60, 1, p2), 10, 6))
> names(pob.a) <- paste("v", 1:6, sep = "")
> row.names(pob.a) <- paste("Ind", 1:10, sep = " ")
> names(pob.b) <- paste("v", 1:6, sep = "")
> row.names(pob.b) <- paste("Ind", 1:10, sep = " ")
> pob.a

       v1 v2 v3 v4 v5 v6
Ind 1   0  1  0  1  1  0
Ind 2   1  1  1  1  1  1
Ind 3   1  1  0  1  1  1
Ind 4   1  1  1  1  1  1
Ind 5   0  1  1  1  1  1
Ind 6   0  1  0  1  1  0
Ind 7   0  1  0  0  1  1
Ind 8   1  0  0  1  0  1
Ind 9   1  0  1  1  1  1
Ind 10  1  1  1  1  0  0

> pob.b
```

```
       v1 v2 v3 v4 v5 v6
Ind 1   1  1  0  0  0  1
Ind 2   0  0  0  0  1  0
Ind 3   0  0  0  0  0  0
Ind 4   0  0  1  0  0  0
Ind 5   0  0  0  0  0  0
Ind 6   0  1  0  0  0  1
Ind 7   1  0  0  0  0  1
Ind 8   0  0  1  0  0  0
Ind 9   1  0  1  1  0  0
Ind 10  0  0  0  0  0  1
```

En la población con probabilidad de recaptura alta todos los individuos fueron capturados al menos una vez. En la población de recaptura baja los individuos 3 y 5 no fueron capturados ni una sola vez en las seis oportunidades posibles.

2. Máxima verosimilitud de una variable aleatoria con distribución binomial

Ahora imaginemos que desconocemos el valor de probabilidad asociado a la obtención de águilas o soles cuando lanzamos la moneda al aire. Pensemos la posibilidad que la moneda pudiera estar "cargada" y que fuera posible obtener con mayor frecuencia más águilas que soles. En este caso, nosotros no conocemos el valor del parámetro de probabilidad p sino lo estimaremos a partir de la serie de datos que contamos a tráves del método de máxima verosimilitud [2, 4].

Se lanzó la moneda al aire y se obtuvieron dos águilas y ocho soles.

A principio de cuentas podríamos tener motivos para comparar las siguientes dos hipótesis:

- H_1 La moneda está cargada hacia los soles, tiene una $p = 0.3$
- H_2 La moneda no está cargada, tiene una $p = 0.5$

¿Cúal de esas hipótesis es más plausible?

La función de verosimilitud para una variable de distribución binomial es:

$$\mathscr{L} = \prod_{i=1}^{n} \binom{N}{k_i} p^{k_i} (1-p)^{N-k_i} \tag{2}$$

Calculamos los valores de verosimilitud para $p = 0.3$ y $p = 0.5$ y determinamos cual tiene mayor soporte de acuerdo a nuestros datos. Adicionalmente, podemos calcular la fuerza de evidencia hacia la hipótesis que encuentra mayor soporte de acuerdo a nuestros datos. La fuerza de evidencia tiene como valor el conciente resultante de los valores de verosimilitud de las dos hipótesis [1, 3]:

```
> h1 <- dbinom(2, 10, p = 0.3)
> h2 <- dbinom(2, 10, p = 0.5)
> h1
```

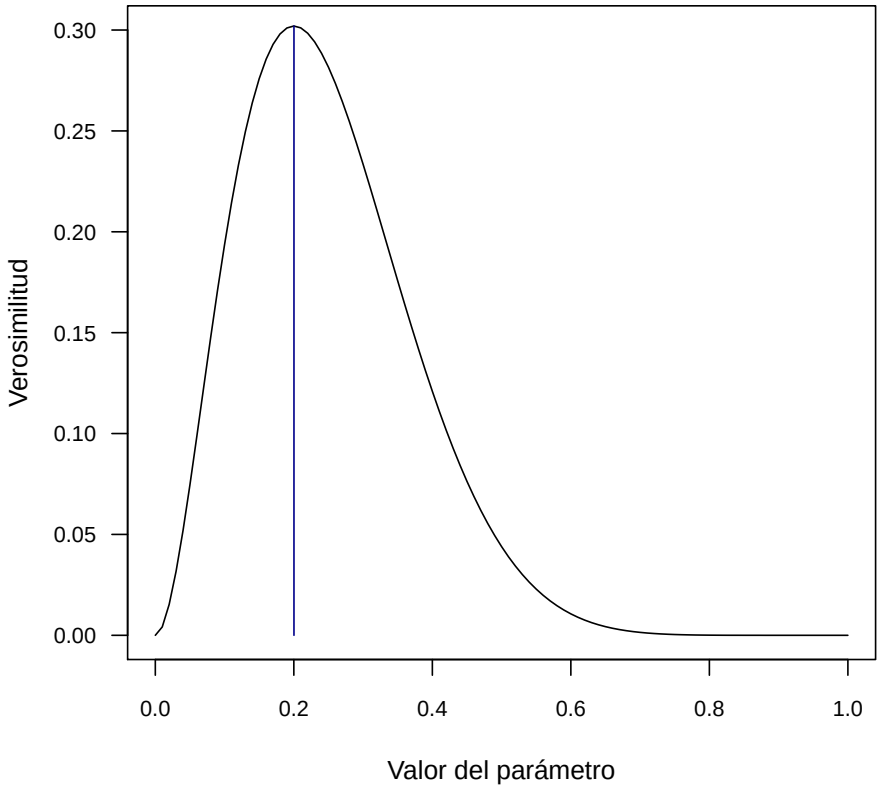
```
[1] 0.2335

> h2

[1] 0.04395

> f.ev <- h1/h2
> f.ev

[1] 5.313
```

La hipótesis de que la moneda está cargada H_1 con $p = 0.3$ es cinco veces más plausible que la hipótesis de la moneda no sesgada H_2. Sin embargo, en el ejemplo anterior empezamos con dos hipótesis *a priori*. Cabría preguntarse cuál es el valor del parámetro p más plausible, aquel que tiene una máxima verosimilitud de acuerdo a nuestros datos.

```
> curve(dbinom(2, 10, p = x), from = 0, to = 1)
```

Al procurar la máxima verosimilitud, por pura fuerza bruta, asignando a p diferentes valores (n = 1000) dentro del rango de 0 a 1 obtenemos:

```
> vpar <- seq(0, 1, length = 1000)
> y <- dbinom(2, 10, vpar)
> y.max <- max(y)
> vpar.mle <- vpar[y == max(y)]
> par(las = 1, cex.axis = 0.8)
> plot(c(seq(0, 1, length = 100)), c(seq(0, 0.3, length = 100)),
+     xlab = "Valor del parámetro", ylab = "Verosimilitud", type = "n")
> curve(dbinom(2, 10, x), 0, 1, add = T)
> lines(c(vpar.mle, vpar.mle), c(0, y.max), col = "darkblue", lty = 1,
+     lwd = 1)
```



La máxima verosimilitud de una variable binomial corresponde a la proporción de éxitos (águilas) dividido por el total de eventos, en nuestro caso, las veces que lanzamos la moneda al aire: $p = 0.2$.

3. Intervalos de confianza

La mayoría de las veces, con el fin de discutir los resultados y extraer conclusiones de ellos, nosotros precisamos obtener los valores de incertidumbre asociados al valor del parámetro estimado. Vamos a obtener tanto el valor del parámetro como los intervalos de confianza al 95 % de nuestro ejemplo anterior.

```
> a = 2
> n = 10
> logistica <- function(x) exp(x)/(1 + exp(x))
> logVerNeg <- function(psi) -dbinom(a, size = n, prob = logistica(psi),
+     log = T)
> ajust <- optim(par = 0.5, fn = logVerNeg, method = "BFGS", hessian = T)
> psi.mle <- ajust$par
> psi.se <- sqrt(1/ajust$hessian)
> z.crit <- qnorm(0.975)
> logistica(psi.mle)
```

```
[1] 0.2
```

```
> logistica(c(psi.mle - z.crit * psi.se, psi.mle + z.crit * psi.se))
```

```
[1] 0.05041 0.54071
```

Si hubiéramos lanzado 40 veces la moneda y hubiéramos obtenido ocho águilas, la misma proporción de águilas que el ejemplo anterior, el valor de máxima de verosimilitud sería el mismo, $p = 0.2$ pero el intervalo de confianza sería más estrecho. Realizamos la optimización númerica con otra alternativa, la función `mle2` del paquete `bbmle` escrito por Ben Bolker [2]:

```
> library(bbmle)
> logVerNeg <- function(psi) -dbinom(k, size = n, prob = logistica(psi),
+     log = T)
> parnames(logVerNeg) <- "psi"
> ajust2 <- mle2(logVerNeg, start = c(psi = 0.5), data = list(k = 8,
+     n = 40))
> logistica(ajust2@coef)
```

```
psi
0.2
```

```
> logistica(confint(ajust2))
```

```
  2.5 %  97.5 %
0.09684 0.34038
```

El intervalo de confianza se redujo, IC = (0.097, 0.340). Esto se debe a podemos extraer más información de nuestros datos, lo cual nos proporciona mayor certidumbre en el valor del parámetro. En el segundo caso podríamos afirmar que la moneda se encuentra sesgada dado que el intervalo de confianza al 95 %, dado que no abarca el valor de $p = 0.5$. La moraleja de la historia relativa a los métodos de captura-recaptura es que a medida que incrementemos el número de visitas (aumentemos la extensión de la historia de encuentros), los intervalos de confianza para la abundancia y la probabilidad de recaptura serán más angostos, la incertidumbre asociada a los valores de los parámetros disminuirá.

Referencias

[1] J.L.F. Batista. Inferência em Recursos Florestais e Ecologia: A Abordagem da Verossimilhança, 2008.

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

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

[4] 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.