PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

INTERROGACIÓN 2 - Parte 2
Métodos Bayesianos - EYP2807 - EYP280I

Profesor	:	Johan Van Der Molen
Ayudante	:	Baltasar Díaz Robles
Fecha	:	20 de mayo de 2024
Entrega	:	27 de mayo de 2024

Contexto

Usualmente, al trabajar con una distribución Hipergeométrica(N, n, p), el parámetro de interés es p. Sin embargo, puede ocurrir que el tamaño poblacional, N, sea desconocido y, así, su estimación se vuelve necesaria. Esto sucede, generalmente, cuando el censo de la población es inviable (esto puede tener sus causas, por ejemplo, en temas monetarios o logísticos), es allí cuando se requiere la aplicación de métodos que permitan estimar su tamaño.

En cierta isla canadiense, un rebaño de renos vive sin la presencia de depredadores naturales. Para prevenir que los renos perturben el equilibrio ecológico de la isla, es necesario mantener el número de renos por debajo de 40. Esto se podría manejar fácilmente con un censo anual de la población pero las autoridades locales comentan que esto es muy costoso. Debido a ello, se prefiere estimar la población de renos con un **método de captura-recaptura**.

Wolter (1986) señala que la mayoría de los métodos de captura-recaptura pueden ser descritos por una distribución Multinomial con parámetros p_{ij}, $i, j \in \{1, 2\}$, donde 1 representa captura. La Tabla 1 describe las probabilidades de captura, con $p_{11} + p_{12} + p_{21} + p_{22} = 1$. Por ejemplo, p_{21} representa la probabilidad de que un individuo arbitrario sea capturado solamente en la segunda muestra. Después de dos experimentos de captura, la población se divide tal cual señala la Tabla 2, con $n_{11} + n_{12} + n_{21} + n_{22} = N$ (donde, *lógicamente*, el cuarto tamaño muestral n_{22} es desconocido).

Un modelo más simple de captura-recaptura, el **modelo de Darroch** (1958), trabaja con la distribución Hipergeométrica, en reemplazo de la Multinomial. Aquí, los dos tamaños muestrales $n_1 = n_{11} + n_{12}$ y $n_2 = n_{21} + n_{22}$ están fijos.

Para el modelo de Darroch, el estimador clásico de N es el **estimador de máxima verosimilitud**,

$$\hat{N} = \frac{n_1}{(n_{11}/n_2)}$$

que identifica tanto la proporción en la población (n_1/N) como la proporción en la muestra (n_{11}/n_2). Este estimador, es claro, no puede ser utilizado cuando $n_{11} = 0$ pero esto no es problema bajo el paradigma bayesiano ya que incluso cuando $n_{11} = 0$, dada una distribución a priori para N es relativamente sencillo encontrar la distribución a posteriori $p(N = n \mid n_{11})$ para realizar inferencia sobre el tamaño poblacional.

Estudios con procesos estocásticos de nacimiento y muerte han determinado que el tamaño poblacional de renos en la isla varía entre 36 y 50. Lamentablemente, tales estudios no permiten concluir acerca de una distribución a priori para N por lo que consideraremos una distribución a priori uniforme sobre $\{36, \ldots, 50\}$. Además, note que si $n_{11} \mid N, n_1, n_2 \sim$ Hipergeométrica(N, n_1, n_2), entonces

$$p(n_{11} = t \mid N, n_1, n_2) = \binom{n_1}{t}\binom{n_2}{n_2 - t} \Big/ \binom{N}{n_2}$$

Desde ahora en adelante, considere el modelo de Darroch con $n_1 = n_2 = 5$.

Note que esto implica que $n_{11} \in \{0, 1, 2, 3, 4, 5\}$

		Segunda Muestra	
		Capturado	No Capturado
Primera Muestra	Capturado	p_{11}	p_{12}
	No Capturado	p_{21}	p_{22}

Tabla 1

		Segunda Muestra	
		Capturado	No Capturado
Primera Muestra	Capturado	n_{11}	n_{12}
	No Capturado	n_{21}	n_{22}

Tabla 2

Ejercicios

1. **[0.5 ptos.]** Considere $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})$, el vector de probabilidades de captura-recaptura. Demuestre que si $\boldsymbol{p}$ sigue una distribución de Dirichlet, entonces $\boldsymbol{p} \mid n_{11}, n_{12}, n_{21}, n_{22}$ también sigue una distribución de Dirichlet y especifique sus parámetros. (De necesitarlo, defina hiperparámetros para la priori de $\boldsymbol{p}$ y detalle que condiciones deben de cumplir para que la distribución esté bien definada para las cuatro probabilidades).

2. **[0.5 ptos.]** Derive la **forma explícita** (esto es, incluyendo la constante de normalización) de la distribución a posteriori $p(N = n \mid n_{11})$.

3. **[1 pto.]** Defina en `R` la función `dDarroch(x,n_11)`, que permita obtener las probabilidades $p(N = x \mid n_{11})$ para todo $N \in \{36, \ldots, 50\}$ y $n_{11} \in \{0, 1, 2, 3, 4, 5\}$. Ingresar el comando `?Hypergeometric` en la consola de `R` podría ser de utilidad.

   Resuma las probabilidades en una tabla de 15×6 **en su informe**.

4. **[1 pto.]** Con ayuda de la función `dDarroch(x,n_11)` definida en el inciso anterior, simule una muestra de 5000 observaciones a posteriori para $N \mid n_{11}$. Reporte los seis histogramas correspondientes y **no olvide utilizar la semilla señalada en las instrucciones**. **Comente sus resultados**.

   Indicación: de utilizar la función `hist()`, **procure añadir** `hist(, breaks = 35.5:50.5)` para poder interpretar.

5. **[1 pto.]** Ocupando el **Método de Monte Carlo** y la función `dDarroch(x, n_11 = 1)`, estime la probabilidad de los renos no estén perturbando el equilibrio ecológico en la isla. Compare su resultado con la tabla del inciso 3. **Comente**.

6. **[1 pto.]** Encuentre el estimador bayesiano δ para las pérdidas

   a) $L(N, \delta) = (N - \delta)^2$

   b) $L(N, \delta) = \begin{cases} 10(\delta - N), & \text{si} \quad \delta > N, \\ N - \delta, & \text{si} \quad \delta \leq N \end{cases}$

   Reporte sus resultados en una o más tablas. **Comente**.

7. **[1 pto.]** ¿Es posible llevar a cabo una aproximación normal de la distribución a posteriori usando los resultados vistos en cátedras? **Argumente**. De ser posible, **explícite su forma analítica**.

8. **[1 pto.]** Suponga que a lo largo de un año, el tamaño de la población N de renos se ha mantenido constante. Cada semana se ha realizado un experimento de captura-recaptura y se ha registrado n_{11i}, la cantidad de individuos capturados y recapturados en el $i-$ésimo experimento semanal. Los S experimentos son independientes entre si. Cada n_{11i} se rige por el modelo de Darroch, esto es,

$$p(n_{11i} = t_i \mid N, n_1, n_2) = \binom{n_1}{t_i}\binom{n_2}{n_2 - t_i}\Big/\binom{N}{n_2} \quad i = 1, \ldots, S.$$

   Considerando la misma priori uniforme sobre $\{36 \ldots, 50\}$ para N, derive la forma explícita de $p(N = n \mid n_{111}, n_{112}, n_{113}, \ldots, n_{11S})$.

9. **[2 ptos.]** Defina en `R` la función `dSDarroch(x,n_11)` que permita obtener las probabilidades $p(N = n \mid n_{111}, n_{112}, n_{113}, \ldots, n_{11S})$.

   Note que, en esta función, `n_11` es el **vector de tamaño S** $n_{111}, n_{112}, n_{113}, \ldots, n_{11S}$.

10. **[3 ptos.]** En el Canvas del curso, usted encontrará el archivo `muestras_semanales.txt` que son las mediciones n_{11} en los experimentos semanales de captura-recaptura de la isla en el año 2014. Realice un **chequeo posteriori predictivo**, comparando el histograma de los datos con la distribución a posteriori predictiva. **Comente**.

    Indicación: de utilizar la función `hist()`, **procure añadir** `hist(, breaks = -0.5:5.5)` para poder interpretar.