Marcação e recaptura

Inferência Bayesiana

Este texto introduz experimentos de marcação e recaptura, nos quais amostramos indivíduos de uma população, marcamos esses indivíduos de alguma forma e, em seguida, coletamos uma segunda amostra da mesma população. Observando quantos indivíduos na segunda amostra estão marcados, podemos estimar o tamanho da população. Experimentos desse tipo foram usados originalmente em ecologia, mas se mostraram úteis em muitas outras áreas.

O conteúdo desta aula faz parte do Capítulo 15 do livro Think bayes de Allen B. Downey.

1 O problema dos ursos pardos

Em 1996 e 1997, pesquisadores instalaram armadilhas para ursos em locais na Colúmbia Britânica e em Alberta, no Canadá, com o objetivo de estimar o tamanho da população de ursos pardos (grizzly bears). Eles descrevem o experimento em um artigo científico.

A "armadilha'' consiste em um atrativo (isca) e vários fios de arame farpado, destinados a capturar amostras de pelos dos ursos que visitam a isca. A partir das amostras de pelos, os pesquisadores utilizam análise de DNA para identificar individualmente cada urso.

Durante a primeira sessão, os pesquisadores instalaram armadilhas em 76 locais. Dez dias depois, eles retornaram e obtiveram 1.043 amostras de pelo, identificando 23 ursos diferentes. Durante uma segunda sessão de 10 dias, eles obtiveram 1.191 amostras, provenientes de 19 ursos diferentes, dos quais 4 já tinham sido identificados no primeiro conjunto de dados.

Para estimar o tamanho da população de ursos a partir desses dados, precisamos de um modelo para a probabilidade de que cada urso seja observado em cada sessão. Como ponto de partida, faremos a suposição mais simples: todo urso na população tem a mesma (desconhecida) probabilidade de ser amostrado em cada sessão.

Com essas suposições, podemos calcular a probabilidade dos dados para uma faixa de possíveis tamanhos de população.

Como exemplo, vamos supor que o tamanho real da população de ursos seja 100.

Após a primeira sessão, 23 dos 100 ursos foram identificados. Durante a segunda sessão, se escolhermos 19 ursos ao acaso, qual é a probabilidade de que 4 deles já tenham sido identificados anteriormente?

Vamos definir:

- N: tamanho real da população, $N = 100$;

- K: número de ursos identificados na primeira sessão, $K = 23$;
- n: número de ursos observados na segunda sessão, $n = 19$ neste exemplo;
- k: número de ursos na segunda sessão que já haviam sido identificados, $k = 4$.

Para valores dados de N, K e n, a probabilidade de encontrar k ursos previamente identificados é dada pela distribuição hipergeométrica:

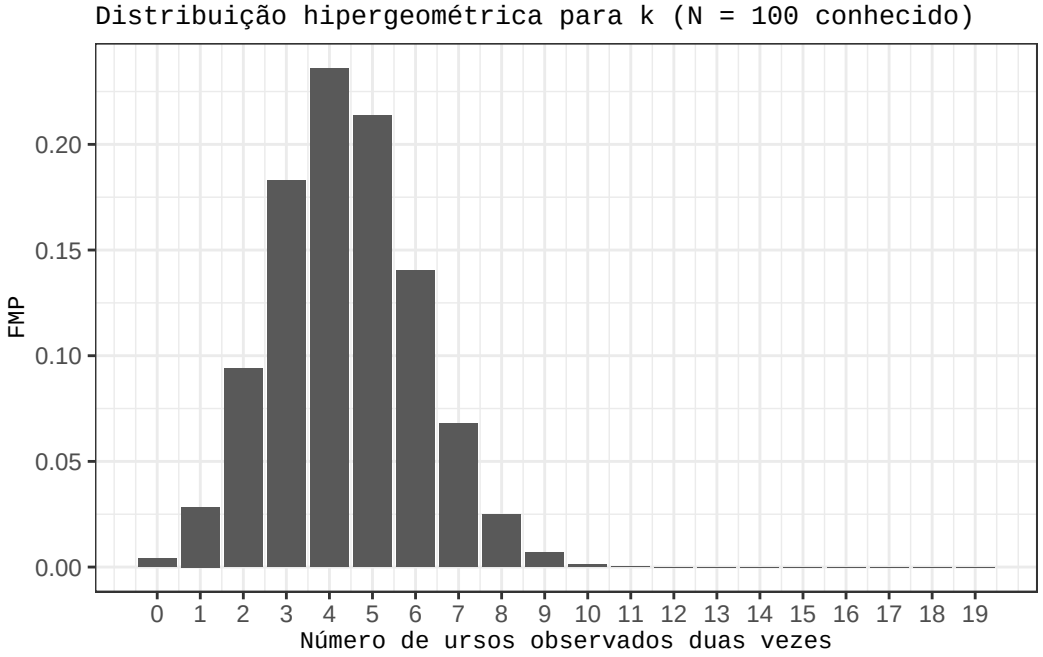
$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

em que o coeficiente binomial $\binom{K}{k}$ é o número de subconjuntos de tamanho k que podemos escolher de uma população de tamanho K.

Para entender essa expressão, observe que:

- o denominador, $\binom{N}{n}$, é o número de subconjuntos de tamanho n que poderíamos escolher de uma população de N ursos;
- o numerador é o número de subconjuntos que contêm k ursos dentre os K previamente identificados e $n-k$ dentre os $N-K$ ainda não observados.

Se $N = 100$, então como $K = 23$ e $n = 19$, temos que o valor mais provável de k é 4, que é justamente o valor observado no experimento. Isso sugere que $N = 100$ é uma estimativa razoável para o tamanho da população, dados esses dados.



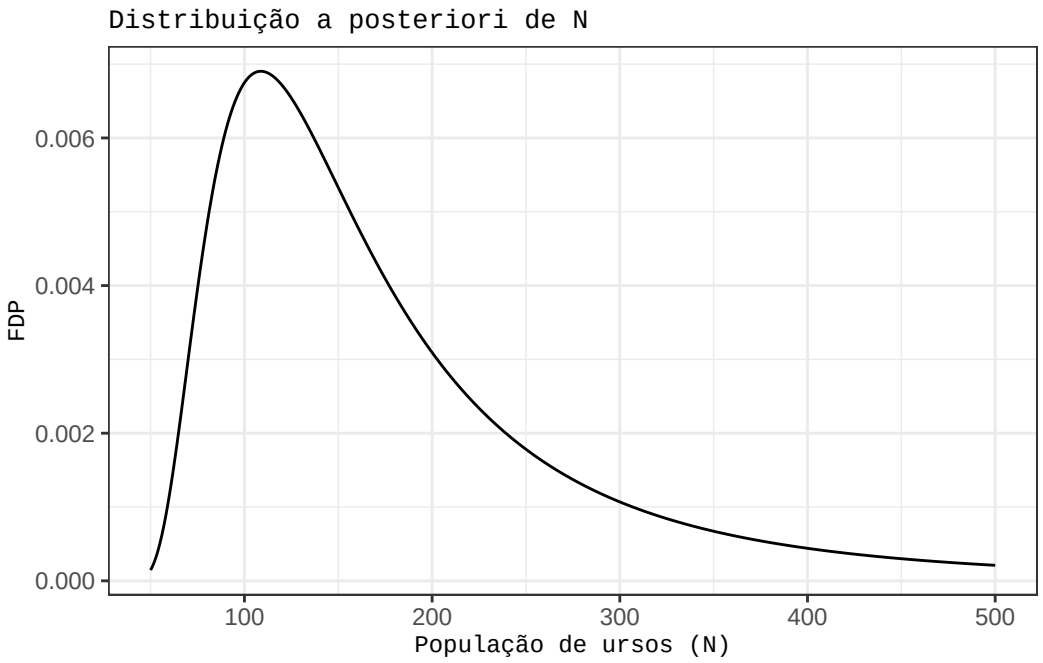
Até aqui, calculamos a distribuição de k dado N, K e n. Agora vamos no sentido inverso: dados K, n e k, como podemos estimar o tamanho total da população, N?

1.1 A atualização

Como ponto de partida, vamos supor que, antes deste estudo, uma pessoa especialista estima que a população local de ursos esteja entre 50 e 500 indivíduos, sendo igualmente provável qualquer valor inteiro nesse intervalo.

Para calcular a verossimilhança dos dados, podemos usar a distribuição hipergeométrica com K e n fixos, e um conjunto de valores possíveis para N.

A curva da posteriori pode ser vista na Figura abaixo.



Com o modelo de um único parâmetro, obtivemos uma distribuição a posteriori para N que tem modo (valor mais provável) em $N = 109$. No entanto, essa distribuição é assimétrica à direita, de forma que a média posterior é bem maior, em torno de $N \approx 174$. Além disso, o intervalo de credibilidade de 90% é bastante amplo, aproximadamente de 77 a 363 ursos, o que indica ainda muita incerteza sobre o tamanho real da população.

Essa solução é conceitualmente simples e já produz uma estimativa coerente, mas podemos melhorar o modelo incluindo explicitamente a probabilidade desconhecida de observar um urso, isto é, introduzindo um segundo parâmetro além de N. Isso nos leva ao modelo de dois parâmetros que veremos a seguir.

2 Modelo com dois parâmetros

Agora vamos tentar um modelo com dois parâmetros: o número de ursos, N, e a probabilidade de observar um urso, p. Vamos supor que a probabilidade p é a mesma nas duas rodadas, o que é razoável neste caso porque é o mesmo tipo de armadilha, no mesmo lugar. Também vamos supor que as probabilidades são independentes; isto é, a probabilidade de um urso ser observado na segunda rodada não depende de ele ter sido observado ou não na primeira rodada. Essa hipótese talvez seja menos realista, mas por enquanto é uma simplificação necessária.

Aqui estão, novamente, as contagens observadas:

- $K = 23$ (ursos observados na primeira rodada),
- $n = 19$ (ursos observados na segunda rodada),
- $k = 4$ (ursos observados nas duas rodadas).

Para este modelo, vamos escrever os dados em uma notação que facilita a generalização para mais de duas rodadas:

- k_{10} é o número de ursos observados na primeira rodada, mas não na segunda;
- k_{01} é o número de ursos observados na segunda rodada, mas não na primeira;
- k_{11} é o número de ursos observados em ambas as rodadas.

Com os valores do problema, temos:

- $k_{10} = 23 - 4$,
- $k_{01} = 19 - 4$,
- $k_{11} = 4$.

Agora, suponha que conhecemos os valores verdadeiros de N e p. Podemos usá-los para calcular a verossimilhança desses dados. Por exemplo, suponha que sabemos que $N = 100$ e $p = 0{,}2$. Podemos usar N para calcular k_{00}, que é o número de ursos não observados em nenhuma das duas rodadas:

$$\text{observado} = k_{01} + k_{10} + k_{11} = 15 + 19 + 4 = 38,$$

$$k_{00} = N - \text{observado} = 100 - 38 = 62.$$

Para organizar melhor os cálculos, é conveniente guardar os dados em um vetor que represente o número de ursos em cada categoria. No nosso exemplo, temos:

- k_{00}: ursos não observados em nenhuma das duas rodadas;
- k_{01}: ursos observados apenas na segunda rodada;
- k_{10}: ursos observados apenas na primeira rodada;
- k_{11}: ursos observados nas duas rodadas.

Para o caso $N = 100$ e $p = 0{,}2$, já vimos que $k_{00} = 62$, $k_{01} = 15$, $k_{10} = 19$ e $k_{11} = 4$. Podemos então escrever o vetor de contagens como

$$x = (k_{00}, k_{01}, k_{10}, k_{11}) = (62, 15, 19, 4).$$

Se conhecermos o valor de p, conseguimos calcular a probabilidade de um urso cair em cada uma dessas quatro categorias. Definindo $q = 1 - p$, temos:

- probabilidade de não ser observado em nenhuma rodada: q^2;
- probabilidade de ser observado apenas na segunda rodada: $q\,p$;
- probabilidade de ser observado apenas na primeira rodada: $p\,q$;
- probabilidade de ser observado nas duas rodadas: p^2.

Ou seja, o vetor de probabilidades para cada categoria é

$$y = (q^2,\ q\,p,\ p\,q,\ p^2).$$

No exemplo numérico com $p = 0{,}2$, obtemos $q = 0{,}8$ e, portanto,

$$y = (0{,}64,\ 0{,}16,\ 0{,}16,\ 0{,}04).$$

Dado um par (N, p), a probabilidade de observarmos exatamente as contagens x é fornecida pela distribuição multinomial:

$$P(x \mid N, p) = \frac{N!}{\prod_i x_i!} \prod_i y_i^{x_i},$$

em que N é o tamanho total da população, (x_i) são as contagens em cada categoria e (y_i) são as probabilidades correspondentes.

Esse valor é a verossimilhança dos dados para valores específicos de N e p. Na prática, porém, não conhecemos nem N nem p. Em vez disso, vamos atribuir distribuições a priori para N e para p, e usar essa verossimilhança para atualizar nossa incerteza e obter a distribuição a posteriori conjunta $P(N, p \mid x)$.

2.1 A priori

Para o parâmetro N, vamos reutilizar a mesma distribuição a priori do modelo de um parâmetro: uma distribuição uniforme discreta entre 50 e 500, isto é,

$$N \in \{50, 51, \ldots, 500\} \quad \text{e} \quad P(N) = \frac{1}{451}.$$

Para a probabilidade de observar um urso, p, adotaremos uma priori uniforme no intervalo $[0, 0{,}99]$. Para isso, aproximamos esse intervalo por uma grade de 100 pontos igualmente espaçados,

$$p_1, p_2, \ldots, p_{100} \in [0, 0{,}99],$$

atribuindo a cada ponto a mesma probabilidade a priori,

$$P(p_j) = \frac{1}{100}, \quad j = 1, \ldots, 100.$$

Assumiremos que as prioris de N e p são independentes. Assim, a distribuição a priori conjunta para o par de parâmetros (N, p) é dada por

$$P(N, p) = P(N)\, P(p),$$

definida sobre todos os pares (N, p_j) da grade. Como ambas as prioris são uniformes, cada combinação possível de N e p recebe a mesma probabilidade,

$$P(N, p_j) = \frac{1}{451} \cdot \frac{1}{100} = \frac{1}{45100}.$$

Em termos computacionais, podemos representar essa priori conjunta como uma tabela em que cada linha corresponde a um par (N, p) e há uma coluna com o valor da probabilidade $P(N, p)$. O índice dessa tabela pode ser pensado como um par ordenado (N, p), de modo que existe exatamente uma linha (e uma probabilidade a priori) para cada combinação possível dos dois parâmetros. O número total de linhas é, portanto, o produto do número de valores da grade de N pelo número de valores da grade de p.

A partir dessa distribuição a priori conjunta, vamos agora calcular a verossimilhança dos dados para cada par (N, p) e, em seguida, obter a distribuição a posteriori $P(N, p \mid x)$.

2.2 A atualização

Para calcular a verossimilhança para cada par (N, p), é conveniente aproveitar a tabela da priori conjunta e apenas acrescentar, para cada combinação de valores, a probabilidade dos dados.

Primeiro, lembramos o vetor de contagens

$$x = (k_{00}, k_{01}, k_{10}, k_{11}),$$

em que k_{00} é o número de ursos não observados em nenhuma das rodadas, k_{01} o número de ursos observados apenas na segunda, k_{10} apenas na primeira e k_{11} em ambas. Para cada valor de N, temos

$$k_{00} = N - k_{01} - k_{10} - k_{11}.$$

Dado um valor de p, definimos $q = 1 - p$ e construímos o vetor de probabilidades para as quatro categorias:

$$y = (q^2,\ q\,p,\ p\,q,\ p^2).$$

Assim, para cada par (N, p) da grade, a verossimilhança dos dados é dada pela distribuição multinomial:

$$L(N,p) = P(x \mid N,p) = \frac{N!}{\prod_i x_i!} \prod_i y_i^{x_i},$$

em que os x_i são as contagens em cada categoria e os y_i são as probabilidades correspondentes. Em termos computacionais, armazenamos esse valor de $L(N,p)$ na mesma estrutura em que guardamos a priori conjunta.

Aplicando a regra de Bayes ponto a ponto na grade, obtemos a distribuição a posteriori conjunta:

$$P(N,p \mid x) \propto P(N,p)\, L(N,p),$$

e em seguida normalizamos de modo que

$$\sum_N \sum_p P(N,p \mid x) = 1.$$

Podemos visualizar essa distribuição conjunta por meio de um gráfico de contorno em que o eixo horizontal representa p, o eixo vertical representa N e as curvas de nível indicam valores de $P(N,p \mid x)$. A figura de contorno mostra que os parâmetros são correlacionados: quando p está mais próximo do limite inferior da grade, os valores mais prováveis de N são maiores; quando p está mais próximo do limite superior, os valores mais prováveis de N são menores. Em outras palavras, combinações com N grande preferem p pequeno, e vice-versa.

Como agora dispomos da distribuição conjunta $P(N,p \mid x)$, podemos obter as distribuições marginais somando sobre o outro parâmetro. Em particular, a marginal de p é dada por

$$P(p \mid x) = \sum_N P(N,p \mid x),$$

e a marginal de N por

$$P(N \mid x) = \sum_p P(N,p \mid x).$$

A figura a seguir mostra a distribuição a posteriori marginal de N obtida com o modelo de dois parâmetros, juntamente com a posteriori de N do modelo de um parâmetro (hipergeométrico). Observa-se que, com o modelo de dois parâmetros, a média posterior de N é ligeiramente menor e o intervalo de credibilidade de 90% é um pouco mais estreito, indicando uma incerteza reduzida em relação ao modelo anterior.