# O Mistério dos Chocalhos

Cláudia Peixoto

IME-USP

O objetivo desta oficina é introduzir os conceitos de amostragem e estimação. Para tanto, iremos utilizar um objeto idealizado pela MATEMATECA (http://matemateca.ime.usp.br/). Trata-se de um chocalho com bolas de duas cores em seu interior. No cabo do chocalho é possível visualizar apenas duas bolas.

SERÁ POSSÍVEL SABER QUANTAS BOLAS DE CADA COR HÁ DENTRO DO CHOCALHO?

Como qualquer criança curiosa faria, o experimento que realizaremos é:

1. Misturar as bolas (chacoalhar o chocalho).
2. Anotar as duas bolas que apareceram no cabo do chocalho.
3. Repetir $N$ vezes esse procedimento.

Considerando-se que as bolas do chocalho formam a população sobre a qual temos interesse em descobrir algum parâmetro, temos que cada anotação refere-se a uma amostra aleatória de dois elementos sem reposição.

Assim, estamos sorteando $N$ amostras de dois elementos.

Note que o resultado da amostra $k$ (após chacoalhar $k$ vezes o chocalho) independe dos resultados das amostras $1$ a $k-1$, pois ao chacoalharmos o chocalho, teremos independência entre os resultados das amostras.

Considere um chocalho com bolas amarelas e verdes. Vamos denotar por $x$ o número de bolas amarelas e por $y$ o número de bolas verdes. Assim, o chocalho possui $x+y$ bolas.

Podemos observar, em uma amostra de dois elementos, duas bolas amarelas, duas bolas verdes ou uma bola verde e uma amarela.

A probabilidade de cada possível resultado da amostra pode ser calculada da seguinte forma:

$P(A,A)=\frac{x}{(x+y)}\frac{(x-1)}{(x+y-1)}$ -> probabilidade de duas bolas amarelas $p_A$

$P(A,V)=2\frac{x}{(x+y)}\frac{y}{(x+y-1)}$ -> probabilidade de uma bola amarela e uma verde ou uma verde e uma amarela $p$

$P(V,V)=\frac{y}{(x+y)}\frac{(y-1)}{(x+y-1)}$ -> probabilidade de duas bolas verdes $p_V$

A partir do que observamos nas $N$ amostras queremos estimar $x$ e $y$.

Podemos também estimar, por exemplo,

$$P(\textit{selecionar uma bola amarela})=\frac{x}{(x+y)},$$

que é a proporção de bolas amarelas no chocalho.

Agora, suponha que você realizou o procedimento proposto com $N=10$ e observou as seguintes amostras:

$$(\mathrm{A},\mathrm{A}),(\mathrm{A},\mathrm{V}),(\mathrm{A},\mathrm{A}),(\mathrm{V},\mathrm{A}),(\mathrm{V},\mathrm{V}),(\mathrm{A},\mathrm{A}),(\mathrm{V},\mathrm{A}),(\mathrm{A},\mathrm{A}),(\mathrm{V},\mathrm{V}),(\mathrm{V},\mathrm{A}).$$

A probabilidade de termos observado essas 10 amostras de dois elementos é, pela independência,

$$(p_A)^4(p_V)^2(1-p_A-p_V)^4. \qquad (1)$$

É razoável pensarmos que o que foi observado é o mais provável de ocorrer. Sendo assim, podemos pensar que $p_A$ e $p_V$ deveriam ser valores que maximizam a probabilidade do que foi observado, ou seja, tornam o valor de (1) máximo.

Derivando (1) em relação à $p_A$ temos

$$(p_V)^2\left[4(p_A)^3(1-p_A-p_V)^4-4(1-p_A-p_V)^3(p_A)^4\right]=0. \qquad (2)$$

Derivando em relação à $p_V$ temos

$$(p_A)^4\left[2p_V(1-p_A-p_V)^4-4(1-p_A-p_V)^3(p_V)^2\right]=0. \qquad (3)$$

Resolvendo o sistema formado por (2) e (3) temos que

$$p_A=\frac{2}{5}, \qquad p_V=\frac{1}{5}, \qquad p=\frac{2}{5}.$$

Observe que os valores que maximizam (1) são as frequências relativas observadas na amostra.

Para $N$ qualquer, onde $N$ são as repetições de nosso procedimento, teremos $N$ resultados e a probabilidade obtida em (1) pode ser escrita da seguinte maneira:

$(p_A)^{Na}(p_V)^{Nv}(1-p_A-p_V)^{N-Na-Nv}$, onde

$Na$ é o número de ocorrências de (A,A) e $Nv$ é o número de ocorrências de (V,V).

Derivando a expressão em relação à $p_A$ e em relação à $p_V$ e igualando a 0 temos:

$$(p_V)^{Nv}\left[Na(p_A)^{Na-1}(1-p_A-p_V)^{N-Na-Nv}-(N-Na-Nv)(1-p_A-p_V)^{N-Na-Nv-1}(p_A)^{Na}\right]=0.$$

$$(p_A)^{Na}\left[Nvp_V(1-p_A-p_V)^{N-Na-Nv}-(N-Na-Nv)(1-p_A-p_V)^{N-Na-Nv}(p_V)^{Nv}\right]=0.$$

Resolvendo o sistema chegaremos a

$$p_A=\frac{N_a}{N},$$

$$p_V=\frac{N_V}{N},$$

$$p=\frac{N-N_a-N_V}{N}.$$

Assim, temos um estimador para as proporções utilizando um método estatístico que tem um nome pomposo; Método de Máxima Verossimilhança. Escolhemos como estimativa o valor que torna as amostras observadas as mais prováveis de ocorrer.

Para uma outra ilustração do problema de estimar a composição dos chocalhos, considere dois chocalhos que possuem a mesma razão entre bolas amarelas e verdes. Por exemplo,

Chocalho I: 4 bolas amarelas e 4 bolas verdes

$$P(A,A)=\frac{4}{(4+4)}\frac{(4-1)}{(4+4-1)}=\frac{12}{56}$$

$$P(A,V) = 2\frac{4}{(4+4)}\frac{4}{(4+4-1)} = \frac{32}{56}$$

$$P(V,V) = \frac{4}{(4+4)}\frac{(4-1)}{(4+4-1)} = \frac{12}{56}$$

Chocalho II: 3 bolas amarelas e 3 bolas verdes

$$P(A,A) = \frac{3}{(3+3)}\frac{(3-1)}{(3+3-1)} = \frac{6}{30}$$

$$P(A,V) = 2\frac{3}{(3+3)}\frac{3}{(3+3-1)} = \frac{18}{30}$$

$$P(V,V) = \frac{3}{(3+3)}\frac{(3-1)}{(3+3-1)} = \frac{6}{30}$$

Observe que, apesar dos dois chocalhos terem a mesma razão entre os números de bolas verdes e amarelas, as probabilidades de cada resultado de uma amostra de dois elementos são diferentes.

É possível provar que a função

$$F(x,y) = (p_A, p_V),$$

que associa a composição do chocalho aos valores das probabilidades de serem observadas duas bolas amarelas ou duas bolas verdes é injetora. (Tente provar isto!)

Este fato garante que estimar as probabilidades nos levam a uma única composição do chocalho.