Determinação do Tamanho Amostral em Populações Finitas Dicotômicas; Perspectiva Bayesiana

Cléber C. Figueiredo, Daniela C. Ramires, Marcos S. Oliveira e Carlos A. de B. Pereira

Março, 2003

1 Introdução

O objetivo do presente trabalho é a determinação do tamanho amostral, n, para a estimação do número, θ, de itens com determinada característica C de uma população finita de tamanho N conhecido. A perspectiva é a Bayesiana e o modelo é o apresentado em Basu & Pereira (1982).

Notação: $Be(a,b)$, $Bebi(n;a,b)$, representam, respectivamente, as distribuições Beta e Beta-binomial, com parâmetro (a,b) e tamanho de amostra n. Além disso, consideremos $n_0 = a+b$, $A = a+x$, $B = b+(n-x)$, $\tilde{n} = n + n_0$, $\tilde{N} = N + n_0$, $m = \frac{A}{\tilde{n}}$, $E(x)$ = média de x e $V(x)$ = variância de x.

2 Modelo

Itens são produzidos segundo um processo de Bernoulli com taxa de falha π. Os itens são armazenados em lotes de tamanho N. Para um determinado lote deseja-se estimar a quantidade θ de itens defeituosos e para isso seleciona-se uma amostra de tamanho n ($< N$), a ser determinado. Considere x e $(\theta - x)$ os números de itens defeituosos na amostra, de tamanho n, e na parte remanescente, de tamanho $N - n$, respectivamente. De acordo com Basu & Pereira (1982), temos:

Se $\pi \sim Be(a,b)$, então: (i) $\pi|x \sim Be(A,B)$; (ii) $\theta \sim Bebi(N;a,b)$; (iii) $x \sim Bebi(n;a,b)$; (iv) $(\theta - x) \sim Bebi(N-n;a,b)$; (v) x e θ são condicionalmente independentes dado π e finalmente (vi) $(\theta - x)|x \sim Bebi(N-n;A,B)$.

Também das propriedades da Beta-Binomial segue que:

$$E(\theta - x|x) = (N-n)m \qquad e \qquad V(\theta - x|x) = \frac{(N-n)\tilde{N}m(1-m)}{(\tilde{n}+1)} \leq \frac{(N-n)\tilde{N}}{4(\tilde{n}+1)}.$$

Note que o limite superior da variância é atingido quando $A = B$, ou equivalentemente, quando $m = 1-m = \frac{1}{2}$. Neste caso a distribuição de $(\theta - x)|x$ é simétrica em torno de m.

3 Determinação do Tamanho Amostral

No processo inferencial descrito acima, os valores a, b e N são especificados e x será observado após a determinação de n. O parâmetro de interesse é θ, ou equivalentemente $\rho = \frac{\theta}{N}$. Note também que se $r_0 = \frac{n_0}{N}$

e $r = \frac{n}{N}$ são as razões amostrais, a priori e a posteriori, então

$$E = E(\rho|x) = (1-r)m + \frac{x}{N} \qquad e \qquad V = V(\rho|x) = \frac{(1-r)(1+r_0)m(1-m)}{(\tilde{n}+1)} \leq \frac{(1-r)(1+r_0)}{4(\tilde{n}+1)}.$$

Para determinar o tamanho amostral, vamos nos concentrar nesta proporção ρ, de itens defeituosos. Para a inferência final, o objetivo é obter o menor intervalo $[I_1, I_2]$, de tal modo que ρ pertença a este intervalo com probabilidade de ao menos $1-\alpha$, isto é, $Pr\{I_1 < \rho < I_2|x\} \geq 1-\alpha$.

Por simplicidade, fixamos $\alpha = 0,0455$. O caso menos favorável é aquele que apresenta a maior variância a posteriori, isto é, quando $m = 1-m = \frac{1}{2}$. Neste caso, temos uma distribuição a posteriori simétrica em torno de $(1-r)m + \frac{x}{N}$. Usando o método padrão de aproximação pela distribuição normal, o comprimento do intervalo será aproximadamente $4D = 0,1$, onde $D = V^{\frac{1}{2}}$. Considerando o caso em que a distribuição a priori é uniforme, isto é, $a = b = 1$, e para uma população finita de $N = 5000$ unidades teríamos $n = 365$. A tabela 1 apresenta os valores de n para diferentes situações onde variamos o valor de D (ou $I_2 - I_1$) e de N, mantendo $\alpha = 0,0455$. O software S-Plus (veja Krause & Olson, 1997) foi utilizado para esses cálculos.

Tabela 1:
Tamanhos amostrais obtidos para valores do desvio padrão e tamanhos populacionais fixados baseados em uma distribuição posteriori $Beta - Binomial$ não simétrica.

Population	$D = 0,01$	$D = 0,025$	$D = 0,05$	$D = 0,07$	$D = 0,1$
5000	1655	**365**	93	48	22
2000	1108	329	90	47	22
1000	712	**284**	86	46	22
500	416	221	80	44	21
100	96	83	49	33	18
50	49	44	33	24	15

4 Comentários Finais

Apresentamos uma alternativa Bayesiana para a determinação do tamanho amostral. Note que a credibilidade de no mínimo $95,45\%$ foi usada apenas como ilustração. Entretanto, aliada a simplicidade, temos o fato de que, se mantivermos a credibilidade fixa, nos casos não simétricos ($A < B$ or $A > B$) a precisão aumenta ($I_2 - I_1$ diminui) consideravelmente quando m está próximo de zero ou da unidade. Por outro lado, mantendo o comprimento do intervalo fixo, a credibilidade aumenta em conformidade com a assimetria, o que não ocorre com a teoria padrão de amostragem. Por exemplo, seja $N = 1000$, $n = 284$ e $x = 142$ um caso simétrico. O intervalo para ρ com $95,55\%$ de credibilidade é $[0,4997; 0,5022]$. Por outro lado, se $x = 80$, um caso não-simétrico, o intervalo seria $[0,2816; 0,2836]$, com comprimento menor do que o anterior.

Referências

[1] Basu, D & Pereira, CAB (1982) *On the Bayesian analysis of categorical data: The problem of nonresponse.* Journal of Planning and Inference 6: 345 − 362.

[2] Krause, A & Olson, M (1997) *The Basics of S and S-Plus.* Springer.