

# Estimação do Tamanho de uma População Finita Rotulada via Simulações Monte Carlo

A.L. HENTGES, A. VIGO, Departamento de Estatística, UFRGS, Porto Alegre, RS, Brasil.

**Resumo.** Este artigo compara quatro diferentes estimadores do tamanho desconhecido de uma população finita rotulada de  $N$  elementos. Dois estimadores são sugeridos por seu apelo intuitivo e os outros dois são construídos segundo critérios teóricos de Estimação Estatística. Como a verdadeira distribuição de probabilidade destes mecanismos de estimação é difícil de ser obtida, recorreremos ao uso de simulações do tipo Monte Carlo para avaliar seus desempenhos.

## 1. Introdução

Considere uma população finita de  $N$  elementos, onde  $N$  é desconhecido e, portanto, é um parâmetro de interesse de certo pesquisador. Suponha ainda que as unidades populacionais  $U_1, U_2, \dots, U_N$  possam ser identificadas através de um rótulo indicando seu número. Em aplicações industriais, por exemplo, cada item produzido poderia receber o número de ordem na linha de produção. Se uma amostra aleatória de  $n$  elementos fornece o vetor de rótulos  $(X_1, X_2, \dots, X_n)$ , podemos estimar o verdadeiro valor desconhecido  $N$  a partir da informação amostral. Empiricamente, sob certas suposições e intuição, diferentes estimadores podem ser definidos para  $N$ , mas dificilmente teriam eficácia máxima. Ainda, pode ser difícil ou inviável identificar a verdadeira distribuição de probabilidade dos estimadores, impossibilitando a estimação de  $N$  através de intervalos de confiança exatos.

Intervalos de confiança construídos da forma tradicional geralmente assumem a normalidade assintótica do estimador, o que seria razoável admitir apenas para um grande tamanho amostral  $n$ .

Neste artigo apresentamos alguns estimadores empíricos para  $N$  e, mediante procedimentos de simulação Monte Carlo, a precisão destes estimadores é comparada com aquela do estimador não viciado e de variância uniformemente mínima. Nas simulações, observamos a eficácia dos estimadores para diferente tamanho amostral, baseado em um elevado número de replicações de vetores de tamanho fixado  $n$ .

## 2. Estimação do Tamanho de uma População Finita

Suponha que as unidades populacionais  $U_1, U_2, \dots, U_N$  possam ser identificadas através de um rótulo indicando seu número. Por simplicidade, assumamos que o vetor

de rótulos  $(1, 2, \dots, N)$  esteja associado a  $(U_1, \dots, U_N)$ . Um exemplo trivial seria a identificação dos carros de uma frota de táxis em certa cidade, através dos rótulos  $1, 2, \dots, N$ .

Usualmente é impraticável executar um Censo (devido ao custo, tempo, etc) e, assim, recorremos a uma amostra de  $n$  elementos  $(X_1, X_2, \dots, X_n)$  selecionados da população de forma independente.

Um estimador  $\hat{\theta}$  é uma função de dados amostrais  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  e tem como objetivo produzir uma estimativa para o verdadeiro valor desconhecido de um certo parâmetro  $\theta$ . Certas qualidades precisam estar presentes em um estimador, como por exemplo, ausência de vício e baixa variabilidade, o que evidenciariam boa precisão.

Um estimador  $\hat{\theta}$  é não-viciado quando

$$E[\hat{\theta}] = \theta ,$$

isto é, considerando amostras de tamanho  $n$ , a média de todas as possíveis estimativas produzidas por  $\hat{\theta}$  é igual ao verdadeiro valor  $\theta$ .

Define-se a variância do estimador  $\hat{\theta}$  por

$$Var(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2 ,$$

a qual indica a variabilidade das estimativas produzidas por  $\hat{\theta}$  em torno do alvo  $\theta$ . Quanto menor a variância de um estimador não-viciado melhor sua qualidade pois suas estimativas diferem pouco do parâmetro  $\theta$ ; veja, por exemplo, [3].

## 2.1. Técnicas de estimação

Suponha que uma amostra aleatória  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  tenha sido observada com reposição, caso em que uma unidade  $U_i$  qualquer pode ter sido selecionada mais de uma vez.

Basicamente, um estimador  $\hat{\theta} = f(X_1, X_2, \dots, X_n)$  é uma função dos dados amostrais e pode assumir uma forma qualquer. Pode-se escolher uma função  $f$  apropriada, de acordo com a conveniência ou, então, derivar a função que otimiza um certo critério, criando estimadores mais eficazes.

## 2.2. Estimadores intuitivos

Devido a seu apelo intuitivo definimos dois estimadores para estimar  $N$  (ou  $\theta$ , simbolicamente). O primeiro deles,

$$\hat{N}_1 = \hat{\theta}_1 = 2 \text{ med}(X_1, X_2, \dots, X_n) - 1 , \quad (2.1)$$

assume que, para um tamanho de amostra relativamente grande, a mediana amostral está próxima da verdadeira mediana populacional  $(N + 1)/2$ .

Um segundo estimador, definido de forma simplificada,

$$\hat{N}_2 = \hat{\theta}_2 = X_{(n)} + [X_{(1)} - 1] , \quad (2.2)$$

assume que, para  $n$  grande, o máximo amostral  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  guarda a mesma distância de seu limitante  $N$ , assim como o valor mínimo da amostra  $X_{(1)} = \min(X_1, X_2, \dots, X_n)$  está em relação ao limitante 1, o menor rótulo populacional possível.

Como ilustração, suponha que uma amostra de  $n = 8$  rótulos tenha sido observada como  $\mathbf{x} = (30, 41, 57, 80, 120, 137, 148, 156)$ , tal que  $\text{med}(\mathbf{x}) = 100$ ,  $x_{(1)} = 30$  e  $x_{(n)} = 156$ . Assim,  $\hat{N}_1 = \hat{\theta}_1 = 2(100) - 1 = 199$  e  $\hat{N}_2 = \hat{\theta}_2 = 156 + [30 - 1] = 185$ .

Frente aos mesmos dados amostrais produzidos por  $\mathbf{x}$  temos duas diferentes estimativas. Embora uma delas possa, eventualmente, ser bem mais adequada que a outra considerando a particular amostra observada, nosso objetivo será detectar o estimador mais eficiente. Para tanto, precisamos conhecer o comportamento deles frente à todas as amostras possíveis para um certo tamanho  $n$  fixado, ou seja, determinar suas distribuições de probabilidades.

### 2.3. Estimadores MV e de variância mínima

Estimadores mais eficientes são disponibilizados por métodos de estimação estatística que assumem um certo modelo probabilístico.

Um estimador de máxima verossimilhança (MV) é definido como aquele que maximiza a probabilidade de gerar aquela particular amostra que tenha sido efetivamente observada. A amostra aleatória de  $n$  elementos  $\mathbf{x} = (x_1, \dots, x_n)$ , observados ao acaso e com reposição, tem verossimilhança (ou probabilidade de ser observada)

$$L(x_1, \dots, x_n | N) = P(x_1, \dots, x_n | N) = \prod_{i=1}^n P[X_i = x_i | N] = \left(\frac{1}{N}\right)^n.$$

Note que quando a amostra já foi observada, os  $x_i$  são fixos e, portanto, a verossimilhança é uma função decrescente dependendo apenas de  $N$  e será maximizada pela escolha do menor  $N$  possível. Diante da restrição  $x_i \in \{1, 2, \dots, N\}$ ,  $i = \{1, \dots, n\}$ , o menor valor adequado para  $N$  é o maior valor de  $x_i$ , logo o estimador de máxima verossimilhança é  $\hat{\theta}_{MV} = X_{(n)} = \max(X_1, \dots, X_n)$ . Este estimador subestima  $\theta$  pois  $E(\hat{\theta}_{MV}) = N - \{\sum_{j=1}^N (j-1)^n\} / N^n$  ([1], página 194), mas para  $N$  grande  $E(\hat{\theta}_{MV}) \approx \frac{n}{n+1}N$ . Redefinimos então um estimador MV assintoticamente não-viciado como

$$\hat{N}_3 = \hat{\theta}_3 = \frac{n+1}{n} X_{(n)}. \quad (2.3)$$

Outro estimador utilizado, entre vários critérios possíveis, é o estimador não-viciado e de variância uniformemente mínima

$$\hat{N}_4 = \hat{\theta}_4 = \frac{[X_{(n)}]^{n+1} - [X_{(n)} - 1]^{n+1}}{[X_{(n)}]^n - [X_{(n)} - 1]^n}, \quad (2.4)$$

que é o estimador ótimo, pois estima  $\theta$  com a melhor precisão, sem apresentar vício ([3], p. 357).

### 3. Eficiência de Estimadores

Na Estatística, um requisito básico é que um estimador seja não-viciado, isto é  $E(\hat{\theta}) = \theta$ , ou razoavelmente próximo disso, apresentando vício desprezível, onde  $(\hat{\theta} - \theta) \rightarrow 0$  quando  $n \rightarrow \infty$ . Dois estimadores não-viciados  $\hat{\theta}_A$  e  $\hat{\theta}_B$  podem ser comparados através de suas variâncias, usualmente via seu quociente.

Recorrendo à simulações, sejam  $t_A$  and  $t_B$  as unidades de tempo necessárias para computar as estimativas definidas por  $\hat{\theta}_A$  e  $\hat{\theta}_B$ , respectivamente. Pode-se, então, afirmar que o estimador  $\hat{\theta}_A$  é mais eficiente que  $\hat{\theta}_B$  se

$$\epsilon_{AB} = \frac{t_A \text{Var}(\hat{\theta}_A)}{t_B \text{Var}(\hat{\theta}_B)} < 1. \quad (3.1)$$

Em certas situações é possível determinar a verdadeira variância de um estimador quando conhecemos sua distribuição de probabilidade, podendo-se então escolher aquele com menor variabilidade. Quando a distribuição de probabilidade do estimador é desconhecida ou de difícil manipulação, a simulação permite uma maneira alternativa de avaliar sua variabilidade.

Como usualmente as variâncias dos estimadores são desconhecidas, é comum utilizar estimativas das variâncias desconhecidas na definição de  $\epsilon_{AB}$  ([4]). Obtém-se assim uma estimativa da eficiência relativa dos mecanismos de estimação de  $\theta$ , tornando-se importante estimar tais variâncias com boa qualidade.

Como ilustração, considere os dois seguintes meios de estimação de  $\text{Var}(\hat{\theta})$ . O primeiro método utiliza a amostragem de  $k$  vetores do tipo  $(Y_1, \dots, Y_n)$ , distintos e independentes de tamanho  $n$ , produzindo  $k$  diferentes estimativas  $\hat{\theta}^{[1]}, \hat{\theta}^{[2]}, \dots, \hat{\theta}^{[k]}$ . A variância do estimador  $\hat{\theta}$  pode ser estimada por

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{\sum_{j=1}^k \{ \hat{\theta}^{[j]} - \bar{\hat{\theta}} \}^2}{k-1}, \quad (3.2)$$

onde  $\bar{\hat{\theta}} = (\sum_{j=1}^k \hat{\theta}^{[j]})/k$  indica a média amostral das  $k$  estimativas. Note que em amostragem com reposição existem  $K = N^n$  distintos vetores de tamanho  $n$  possíveis de serem selecionados. Na impossibilidade de gerar tal magnitude de vetores e, conseqüentemente, calcular o verdadeiro valor da variância  $\text{Var}(\hat{\theta})$ , recorreremos à amostragem de apenas  $k$  amostras via simulação, onde  $k$  deve ser grande para permitir uma boa estimativa.

Alternativamente, a técnica “jackknife” ([2]) trabalha com re-amostragem na amostra original. Para um certo estimador  $\hat{\theta}$ , seja  $\hat{\theta}_{(i)}$  a estimativa de  $\theta$  baseada no vetor observado  $(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ , onde a  $i$ -ésima observação  $x_i$  foi ignorada, produzindo-se um conjunto de  $n$  valores  $\{\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \dots, \hat{\theta}_{(n)}\}$ .

A estimativa da variância de  $\hat{\theta}$  é então

$$\widehat{\text{Var}}(\hat{\theta})_{jack} = \frac{n-1}{n} \sum_{i=1}^n \{ \hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)} \}^2,$$

onde  $\bar{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)} / n$ .

Embora a variância seja o critério usual para medir a precisão de um estimador, é interessante considerar sua raiz quadrada, definindo o desvio padrão como outra medida comum na Estatística. Sob certas suposições (como normalidade ou, pelo menos, simetria na distribuição de probabilidade de  $\hat{\theta}$ ), usamos o desvio padrão do estimador para definir intervalos de confiança para o verdadeiro valor de  $\theta$ , usualmente na forma

$$\hat{\theta} \pm z_{\alpha/2} [Var(\hat{\theta})]^{1/2}, \quad (3.3)$$

onde  $z_p$  é o percentil de ordem  $p$  de uma distribuição normal padrão. O intervalo acima, construído com uma estimativa da  $Var(\hat{\theta})$ , apresenta aproximadamente confiança  $(1 - \alpha) * 100\%$  de conter o verdadeiro valor  $\theta$ . Nem sempre um estimador tem distribuição normal, mas para  $n$  relativamente grande, costuma-se atingir uma distribuição assintoticamente normal.

## 4. Simulações

Neste exemplo de aplicação fixamos o verdadeiro valor  $\theta$  como  $\theta = N = 1000$ , para avaliar o desempenho dos estimadores propostos via simulação de amostras da população de rótulos  $(1, 2, \dots, N)$ .

Os estimadores são comparados através de diferentes medidas, entre elas  $\epsilon_{AB}$  (3.1), quando o tamanho da amostra  $n$  aumenta até 900. Cabe observar, no entanto, que devido ao custo, raramente um processo de amostragem selecionaria uma fração amostral  $n/N$  tão grande como a utilizada neste exercício computacional.

Para cada tamanho amostral  $n$  fixado, geramos  $k = 10000$  amostras aleatórias para obter estimativas de  $\theta$ , conforme (2.1)–(2.4). Tal número de vetores simulados é arbitrário, apenas para viabilizar uma melhor avaliação da eficiência dos quatro estimadores. A re-amostragem via *jackknife* também será desenvolvida.

Intervalos de confiança da forma (3.3) serão construídos sob a suposição de normalidade dos estimadores. Quando a distribuição de probabilidade de um estimador não for normal será necessário estimar sua distribuição a partir das amostras simuladas.

## 5. Resultados e Discussão

Inicialmente fica evidenciado que estimadores intuitivos do tipo (2.1) ou (2.2) dificilmente teriam alguma utilidade havendo a disponibilidade de estimadores mais sofisticados.

Exemplificando, a Figura 1 apresenta o comportamento de algumas estatísticas para as 10000 estimativas produzidas por  $\hat{\theta}_1$  e  $\hat{\theta}_4$ , conforme  $n$  aumenta. Os gráficos mostram em (a) os limites ( $\min\{\hat{\theta}\}, \max\{\hat{\theta}\}$ ) e, em (b), a média amostral  $\bar{\hat{\theta}}$  das  $k$  estimativas  $\{\hat{\theta}\}$  para os dois métodos usados.

Percebe-se nitidamente a superioridade da precisão do estimador  $\hat{\theta}_4$ . Para  $n = 100$ , por exemplo, as  $k$  estimativas produzidas por  $\hat{\theta}_1$  variavam entre 647 e 1356, com média amostral 1000.6; enquanto que  $\hat{\theta}_4$  apresentou média 999.8, para estimativas entre 906.5 e 1009.5. Conseqüentemente, tivemos  $\hat{Var}(\hat{\theta}_1) = 9642.74$  e  $\hat{Var}(\hat{\theta}_4) =$

101.87. Assumindo que os tempos computacionais  $t_1$  e  $t_4$  são aproximadamente iguais temos  $\epsilon_{\hat{\theta}_4} = 94.7$  conforme (3.1), isto é, o método  $\hat{\theta}_4$  é muito mais eficiente para estimar o verdadeiro valor de  $\theta$ .

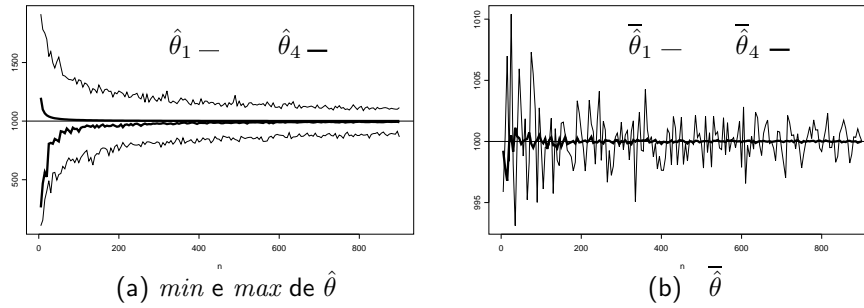


Figura 1: Resultados obtidos para 10 mil estimativas de  $\theta$ , produzidas pelos estimadores  $\hat{\theta}_1$  e  $\hat{\theta}_4$ , conforme  $n$

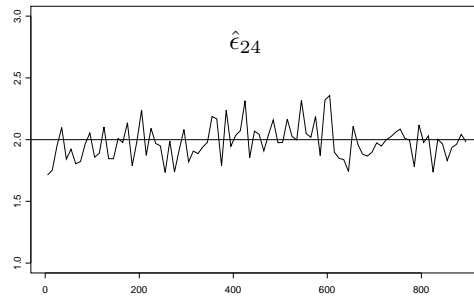


Figura 2: Evolução da eficiência relativa  $\hat{\epsilon}_{24} = \hat{Var}(\hat{\theta}_2)/\hat{Var}(\hat{\theta}_4)$  estimada por 10000 replicações, conforme  $n$

O estimador (2.2) também mostrou-se inferior ao estimador (2.4). A Figura 2 mostra que o estimador  $\hat{\theta}_4$  teve consistentemente cerca do dobro da eficiência de  $\hat{\theta}_2$ , independente do tamanho  $n$  da amostra. Sem generalizar, em nosso exemplo de aplicação tal fato evidencia que, mesmo com uma expressiva fração amostral, um estimador definido por meios intuitivos não possui a mesma precisão de um estimador definido por critérios mais rigorosos de otimalidade.

Diante dos resultados apresentados por  $k = 10000$  replicações para cada tamanho amostral  $n$  utilizado, constata-se claramente que a preferência por um estimador “ótimo” de  $\theta$  deve recair entre  $\hat{\theta}_3$  ou  $\hat{\theta}_4$ . O Quadro 1 sintetiza, para alguns valores de  $n$ , resultados já mostrados parcialmente nos gráficos anteriores.

As simulações permitiram constatar a normalidade da distribuição empírica dos estimadores intuitivos (2.1) e (2.2). Mesmo para  $n = 50$ , um tamanho amostral

modesto, tanto  $\hat{\theta}_1$  como  $\hat{\theta}_2$  apresentam nitidamente uma distribuição de Gauss, como mostra a Figura 3 e confirmado em testes estatísticos não explicitados aqui. Embora  $\hat{\theta}_1$  e  $\hat{\theta}_2$  não apresentem a mesma eficácia de  $\hat{\theta}_3$  e  $\hat{\theta}_4$ , a normalidade de suas distribuições possibilita a construção de intervalos de confiança do tipo (3.3).

Quadro 1: Estatísticas para  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$  e  $\hat{\theta}_4$   
para  $k = 10000$  replicações simuladas, para alguns valores de  $n$

| $n$ | estatística                | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $\hat{\theta}_4$ |
|-----|----------------------------|------------------|------------------|------------------|------------------|
| 50  | $\min(\hat{\theta})$       | 507              | 828              | 844.5            | 844.1            |
|     | $\max(\hat{\theta})$       | 1456             | 1171             | 1020             | 1019.5           |
|     | $\text{med}(\hat{\theta})$ | 1000             | 1000             | 1006.7           | 1006.2           |
|     | $\bar{\hat{\theta}}$       | 999.4            | 999.9            | 1000.4           | 999.9            |
|     | $\hat{V}ar(\hat{\theta})$  | 18068.77         | 757.12           | 393.14           | 393.13           |
| 100 | $\min(\hat{\theta})$       | 647              | 901              | 906.9            | 906.5            |
|     | $\max(\hat{\theta})$       | 1356             | 1088             | 1010             | 1009.5           |
|     | $\text{med}(\hat{\theta})$ | 1000             | 1000             | 1002.9           | 1002.4           |
|     | $\bar{\hat{\theta}}$       | 1000.6           | 999.7            | 1000.3           | 999.8            |
|     | $\hat{V}ar(\hat{\theta})$  | 9642.74          | 195.66           | 101.88           | 101.87           |
| 200 | $\min(\hat{\theta})$       | 748              | 956              | 958.8            | 958.3            |
|     | $\max(\hat{\theta})$       | 1273             | 1040             | 1005             | 1004.5           |
|     | $\text{med}(\hat{\theta})$ | 1000             | 1000             | 1001.9           | 1001.5           |
|     | $\bar{\hat{\theta}}$       | 1000.2           | 999.8            | 1000.4           | 999.9            |
|     | $\hat{V}ar(\hat{\theta})$  | 4792.24          | 49.85            | 25.98            | 25.97            |
| 500 | $\min(\hat{\theta})$       | 832              | 982              | 983.9            | 983.5            |
|     | $\max(\hat{\theta})$       | 1155             | 1017             | 1002             | 1001.5           |
|     | $\text{med}(\hat{\theta})$ | 999              | 1000             | 1001             | 1000.5           |
|     | $\bar{\hat{\theta}}$       | 999.8            | 999.9            | 1000.5           | 1000             |
|     | $\hat{V}ar(\hat{\theta})$  | 1959.98          | 7.85             | 3.97             | 3.97             |

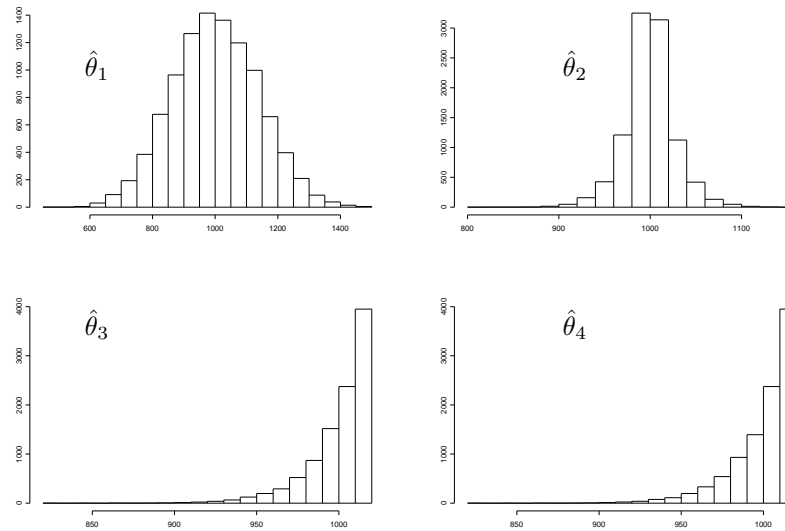


Figura 3: Histogramas obtidos para 10 mil estimativas de  $\theta$ , produzidas por  $\hat{\theta}_1$ ,  $\hat{\theta}_2$ ,  $\hat{\theta}_3$  e  $\hat{\theta}_4$ , quando  $n = 50$

Para  $n$  maior, o mesmo comportamento mostrado na Figura 3 repetiu-se de forma consistente. Como a ênfase na Estimação Estatística recai sobre estimativas via intervalo de confiança, ao invés de estimativas pontuais, a preocupação passa a ser com a qualidade que intervalos do tipo (3.3) teriam ao serem construídos via (2.3) ou (2.4). Frente ao comportamento extremamente assimétrico das distribuições de  $\hat{\theta}_3$  e  $\hat{\theta}_4$ , julgamos conveniente estudar a “taxa de cobertura”, ou a probabilidade que intervalos de confiança têm de conter o verdadeiro valor de  $\theta$ , que os estimadores apresentaram via simulações. Em outras palavras, queremos simplesmente avaliar em quantas amostras o intervalo de confiança contém o verdadeiro valor  $\theta = 1000$  entre as simulações realizadas. O Quadro 2 resume os resultados para os mesmos valores de  $n$  mencionados anteriormente, após estimação da variância dos estimadores mediante (3.2).

Quadro 2: Proporção de intervalos (3.3) com 95% de confiança que incluam o verdadeiro valor de  $\theta$  em  $k = 10000$  simulações, segundo o estimador utilizado, para alguns valores de  $n$

| $n$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $\hat{\theta}_4$ |
|-----|------------------|------------------|------------------|------------------|
| 50  | 95.21%           | 93.71%           | 94.64%           | 94.65%           |
| 100 | 94.97%           | 93.54%           | 94.72%           | 94.74%           |
| 200 | 94.95%           | 93.48%           | 95.33%           | 95.34%           |
| 500 | 95.21%           | 93.75%           | 95.08%           | 95.08%           |
| 900 | 95.12%           | 90.53%           | 93.52%           | 93.52%           |

A Figura 3 e os resultados mostrados no Quadro 2 são importantes. Embora a distribuição amostral dos estimadores  $\hat{\theta}_3$  e  $\hat{\theta}_4$  (já razoavelmente evidenciados como os melhores) sejam fortemente assimétricas, suas variâncias são tão pequenas que intervalos tradicionais da forma (3.3) ainda são plenamente confiáveis. De fato, o estimador  $\hat{\theta}_4$  assume um valor excedendo  $\theta$  para  $n = 50$ , por exemplo, desde que o máximo amostral  $x_{(n)}$  exceda 980, o que ocorre com probabilidade  $1 - 0.98^{50} = 0.64$ . Entretanto, ainda que a maior parte dos possíveis valores de  $\hat{\theta}_4$  superestimem  $\theta$ , a variabilidade deste estimador é quase inexpressiva, fazendo com que a grande maioria das estimativas girem muito próximas de  $\theta$ .

Diante disso, exploramos também a distribuição amostral dos quatro estimadores obtida a partir das  $k = 10000$  simulações para cada tamanho  $n$ , sem assumir a normalidade. Concentramos nossa atenção nos quantis  $Q_{0.025}$  e  $Q_{0.975}$ , que definiriam um intervalo de 95% de confiança para cada particular estimador, já que  $P[Q_{0.025} < \hat{\theta} < Q_{0.975}] = 0.95$ , isto é, o estimador  $\hat{\theta}$  teria 95% de probabilidade de gerar uma estimativa no intervalo  $(Q_{0.025}, Q_{0.975})$ .

Quadro 3: Quantis estimados da distribuição dos quatro estimadores para alguns valores de  $n$ .

| $n$ | Quantil           | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $\hat{\theta}_4$ |
|-----|-------------------|------------------|------------------|------------------|------------------|
| 50  | $\hat{Q}_{0.025}$ | 735              | 943              | 948.6            | 948.1            |
|     | $\hat{Q}_{0.975}$ | 1273             | 1059             | 1020             | 1019.5           |
| 100 | $\hat{Q}_{0.025}$ | 809              | 970              | 973.6            | 973.1            |
|     | $\hat{Q}_{0.975}$ | 1195             | 1029             | 1010             | 1009.5           |
| 200 | $\hat{Q}_{0.025}$ | 860              | 985              | 986.9            | 986.4            |
|     | $\hat{Q}_{0.975}$ | 1139             | 1015             | 1005             | 1004.5           |
| 500 | $\hat{Q}_{0.025}$ | 913              | 994              | 995              | 994.5            |
|     | $\hat{Q}_{0.975}$ | 1086             | 1006             | 1002             | 1001.5           |
| 900 | $\hat{Q}_{0.025}$ | 936              | 997              | 997.1            | 996.7            |
|     | $\hat{Q}_{0.975}$ | 1066             | 1003             | 1001.1           | 1000.7           |

Note no Quadro 3, por exemplo para  $n = 50$ , que cerca de 95% das estimativas produzidas por  $\hat{\theta}_1$  estão no intervalo (735, 1273), enquanto que  $\hat{\theta}_4$  produziu 95% de seus valores entre (948.1, 1019.5), uma vizinhança bem mais próxima do verdadeiro valor  $\theta = N = 1000$  fixado neste exercício de simulação. Mais uma vez fica evidenciado que  $\hat{\theta}_3$  e  $\hat{\theta}_4$  seriam os estimadores que melhor estimariam  $\theta$ .

Resultados via estimação *jackknife* mostraram-se sensíveis à natureza dos quatro estimadores, que apresentam dependência total em  $med(\mathbf{X})$ , caso de (2.1); em  $X_{(n)}$ , caso de (2.3) e (2.4); ou em  $X_{(1)}$  e  $X_{(n)}$ , quando (2.2) é definido. Como conseqüência, as variâncias dos estimadores foram subestimadas. Entretanto, a “taxa de cobertura” de  $\hat{\theta}_3$  e  $\hat{\theta}_4$  e seus quantis obtidos por re-amostragem *jackknife* evidenciaram novamente a preferência por estes dois estimadores. (Resultados não mostrados por limitação de espaço.)

As simulações foram realizadas com o software Splus versão 2000, em um com-

putador PC Pentium 133. Usualmente 10 mil replicações para  $n$  fixado como 100, por exemplo, exigiam cerca de 2 minutos, aumentando para 28 minutos caso a técnica “jackknife” fosse aplicada.

## 6. Conclusões

Neste trabalho constatamos a superioridade de métodos de estimação sustentados por teoria estatística mais rigorosa. Apesar do estimador ótimo  $\hat{\theta}_4$  ser conhecido, sua variância e sua distribuição de probabilidades são desconhecidas e apenas mediante as simulações foi possível construir um intervalo de confiança aproximado para o verdadeiro valor do parâmetro. Mesmo não possuindo normalidade assintótica, as taxas de cobertura para os estimadores  $\hat{\theta}_3$  e  $\hat{\theta}_4$  são muito próximas do nível de confiança especificado.

Embora a intuição seja importante, e particularmente valorizada na Estatística Bayesiana, dificilmente um método de estimação baseado apenas nesta qualidade consegue superar estimadores construídos sob critérios de otimalidade.

Em problemas mais realistas de estimação de uma população finita, como uma população animal vivendo em certa floresta, seria muito mais complexo elaborar adequadamente suposições relativas à taxas de migração, probabilidade de sobrevivência, probabilidade de captura-recaptura e outras variáveis. Um estimador seria então uma função de várias variáveis e parâmetros e a prioridade do pesquisador deveria recair sobre a busca de um estimador com sólidas propriedades estatísticas, seguindo princípios como aqueles usados na definição de (2.4).

As simulações como estas realizadas aqui, com cuidadosa programação para reproduzir fielmente um processo sofisticado, podem constituir uma poderosa ferramenta auxiliar, especialmente quando a distribuição de probabilidades do estimador não é conhecida.

**Abstract.** This paper compares four different estimators for the unknown size  $N$  of a finite population by simulating random samples via Monte Carlo methods. As expected from robust statistical theory it is found that more elaborate estimation methods have better performance.

## Referências

- [1] A. Feller, “Introdução à Teoria das Probabilidades e suas Aplicações”, Parte 1, Editora Edgard Blücher, 1976.
- [2] R. LePage e L. Billard, “Exploring the Limits of Bootstrap”, Wiley, 1992.
- [3] V.K. Rohatgi, “An Introduction to Probability Theory and Mathematical Statistics”, Wiley, 1976.
- [4] R.Y. Rubinstein, “Simulation and the Monte Carlo Method”, Wiley, 1981.