# INTERVALOS DE CONFIANÇA VIA SIMULAÇÃO MONTE CARLO: O ESTADO DA ARTE

**Gean Carlos Feliciano de Almeida**[1], **Ivair Ramos Silva**[1,2]

**Resumo:** *A estimação por intervalos é uma das técnicas da inferência estatística mais utilizadas nas diversas áreas da ciência. O intervalo de confiança exato não é viável nos casos em que não se conhece a distribuição da estatística usada na estimação do parâmetro de interesse. Assim, uma alternativa é o uso de simulação Monte Carlo para construção do intervalo, os quais apresentam boa performance no que se refere à real probabilidade de cobertura comparativamente ao coeficiente de cobertura desejado. Este artigo se dedica a descrever três dos principais métodos Monte Carlo usados para este fim. Além de fazer um contra-ponto sobre prós e contras de cada método, fornecemos também exemplos de aplicaçoes envolvendo estimação de tamanho populacional via captura-recaptura, e estimação do risco relativo em conglomerados espaciais.*

**Palavras-chave:** *Probabilidade de Cobertura, Coeficiente de Confiança, Simulação Monte Carlo, Captura-recaptura, Conglomerados Espaciais.*

**Abstract:** *Interval estimation is one of the most used techniques of statistical inference in various areas of science. When the exact analytical solution for interval estimation is not computable, an appropriate alternative is to use a Monte Carlo method, and the main objective of this paper is to describe three of the main methodologies for obtaining confidence intervals through Monte Carlo simulation. We offer a theoretical discussion of the pros and cons of each method, focusing in their performance in which concerns the actual coverage probability in comparison to the target confidence coefficient. Numerical examples of application are used to illustrate each method involving the mark-recapture problem for population size estimation, and for the inference of the relative risk associated to spatial clusters.*

**Keywords:** *Coverage Probability, Confidence Coefficient, Monte Carlo Simulation, Mark-recapture, Spatial Clusters.*

## 1 Introdução

O 'intervalo de confiança' é um dos conceitos mais utilizados e bem sucedidos da Inferência Estatística frequentista. O método clássico para construção de intervalos de confiança é o método da 'inversão de um teste de hipóteses' que, em alguns casos, pode ser obtido facilmente pelo uso de uma 'quantidade pivotal'. Quando o intervalo de confiança não pode ser obtido analiticamente, métodos Monte Carlo podem ser usados como altenativas formais. Apesar dos métodos Monte Carlo serem amplamente utilizados em diversos problemas de grande relevância, ainda parece haver uma certa carência de material acadêmico, escrito na língua portuguesa, que aborde o tema de maneira detalhada e formal. O objetivo deste artigo, portanto, é oferecer uma descrição teórica dos métodos mais usados para construção de intervalos de confiança Monte Carlo, os quais podem ser divididos em três tipos: Método Percentil[9](também conhecido como 'bootstrap paramétrico'), Método Aleatorizado[5] e o Método Sequencial[18].

[1]Departamento de Estatística - Universidade Federal de Ouro Preto. e-mail: *geanfeliciano@yahoo.com.br*

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA.

Na oportunidade, apresentamos exemplos, teóricos e práticos, para cada método discutido ao longo do texto. Adicionalmente, descrevemos, passo a passo, a utilização do método sequencial para análise de dados reais voltada à estimação do risco relativo em conglomerados espaciais seguida da estatística Scan circular. Grande parte dos exemplos de aplicação aqui apresentados se utilizam do problema da estimação de tamanhos populacionais. Aproveitando a contextualização deste problema, oferecemos, como produto adicional, a descrição de um método exato para obtenção de intervalos de confiança para os contextos em que o parâmetro de interesse é o número de indivíduos de uma população, o que prova, portanto, não ser necessário fazer uso de métodos assintóticos, ou baseados em simulação Monte Carlo, para tratar esse problema em específico.

O conteúdo desse artigo está distribuído da seguinte forma: a próxima seção traz uma breve revisão sobre os métodos clássicos, exatos e assintóticos, usados na construção de intervalos de confiança. A Seção 3 traz o conteúdo principal desse texto, a saber, a definição, a motivação e a descrição de três dos diferentes tipos de métodos baseados em simulação Monte Carlo para obtenção de intervalos de confiança. A Seção 4 apresenta um método exato para estimação intervalar do tamanho populacional via captura-recaptura. A discussão teórica dos três diferentes métodos Monte Carlo citados no primeiro parágrafo deixa claro que os dois métodos de maior relevância, no sentido de serem de fato viáveis e válidos para um maior número de problemas práticos, são os métodos percentil e sequencial. Assim, a Seção 5 oferece uma comparação entre esses dois métodos em termos da verdadeira probabilidade de cobertura que cada um apresenta. Os resultados de tal comparação indicam que, além do método sequencial ser válido para um maior número de problemas práticos, é também mais confiável no que se refere ao real controle da probabilidade de cobertura comparativamente ao coeficiente de confiança pretendido pelo usuário. O método sequencial também se apresenta como única opção a propiciar uma unificação dos métodos Monte Carlo de modo a favorecer, simultaneamente, a aplicação do teste de hipóteses Monte Carlo convencional e a estimação intervalar Monte Carlo. Por fim, como o método sequencial apresenta a melhor performance do ponto de vista de sua generalidade e controle eficiente da probabilidade de cobertura, a Seção 6 traz uma aplicação deste método a dados reais. A referida aplicação se dedica à estimação do risco relativo associado a conglomerados espaciais e faz uso das contagens de câncer de cérebro no Novo México, Estados Unidos, observados nos anos de 1973 a 1991.

## 2 Intervalos de Confiança

Seja $X_1, \ldots, X_n$ uma sequência de observações aleatórias da variável aleatória $X$, e seja $\chi$ o espaço amostral associado ao vetor aleatório $\tilde{\mathbf{X}} = \{X_1, \ldots, X_n\}$. Denote a distribuição de probabilidades de $X$ por $F_X(x|\theta)$, onde $\theta$ é um parâmetro populacional desconhecido. Antes de abordarmos a definição de intervalo de confiança, será conveniente introduzir a noção de 'estimador intervalar'. Segundo [10], um estimador intervalar, para um parâmetro de valor real $\theta$, é qualquer par de funções de valores reais, $L(X_1, \ldots, X_n)$ e $U(X_1, \ldots, X_n)$, de uma amostra aleatória $X_1, \ldots, X_n$, que satisfaz $L(X_1, \ldots, X_n) < U(X_1, \ldots, X_n)$. Uma forma sucinta de se representar este intervalo aleatório, que pode ser interpretado como um estimador intervalar, é o uso da notação $IC(\tilde{\mathbf{X}}) = [L(\tilde{\mathbf{X}}); U(\tilde{\mathbf{X}})]$.

Naturalmente, é de interesse prático que um determinado estimador intervalar apresente uma alta probabilidade de conter $\theta$. Sob esta ótica, duas importantes medidas de performance precisam ser consideradas quando da construção de um estimador intervalar, as quais são a 'probabilidade de cobertura' e o 'coeficiente de confiança'. Dado um estimador intervalar $[L(\tilde{\mathbf{X}}); U(\tilde{\mathbf{X}})]$ para o parâmetro $\theta$, a probabilidade de cobertura, $\gamma_\theta$, de $[L(\tilde{\mathbf{X}}); U(\tilde{\mathbf{X}})]$, é a probabilidade do intervalo conter o verdadeiro parâmetro $\theta$, para um dado $\theta$. Isto é, $Pr(\theta \in [L(\tilde{\mathbf{X}}); U(\tilde{\mathbf{X}})]|\theta) = \gamma_\theta$. Com base na medida de performance $\gamma_\theta$, define-se a medida de performance que independe de $\theta$, que é o coeficiente de confiança. Dizemos que o coeficiente de confiança de $[L(\tilde{\mathbf{X}}); U(\tilde{\mathbf{X}})]$ é igual a $\gamma$ se $\inf_\theta \gamma_\theta = \inf_\theta Pr[(L(\tilde{\mathbf{X}}) < \theta < U(\tilde{\mathbf{X}})|\theta] \geq \gamma$. É fundamental ressaltar que as probabilidades

expressas nas duas últimas linhas se referem às variáveis aleatórias $L(\tilde{\mathbf{X}})$ e $U(\tilde{\mathbf{X}})$. O parâmetro $\theta$ é, por definição, uma constante fixa e desconhecida, e portanto a probabilidade de que pertença a um intervalo arbitrário da linha real é zero ou um. Note que, por construção, estimadores intervalares com coeficiente de confiança $\gamma$ sempre apresentam probabilidade de cobertura de no máximo $\gamma$. Estimadores intervalares que atendem a arbitrários coeficientes de confiança são conhecidos como intervalos de confiança.

## 2.1 Métodos Exatos

Os métodos para construção de intervalos de confiança exatos podem ser definidos em dois grandes grupos: 'método da inversão de um teste de hipóteses' e o 'método da quantidade pivotal'. Rigorosamente, o método da quantidade pivotal pode ser visto como caso particular do método da inversão de um teste de hipóteses, mas, como o primeiro pode ser descrito sem que se precise mencionar conceitos de testes de hipóteses, geralmente, o uso de uma quantidade pivotal para otenção do intervalo é descrito como um método particular.

### 2.1.1 Metodo da Inversão de um Teste de Hipóteses

Existe uma certa correspondência entre testes de hipóteses e estimação intervalar. Na verdade, podemos afirmar que, em geral, para cada teste de hipóteses de nível $\alpha \in (0,1)$, existe um intervalo de confiança de coeficientes $(1-\alpha)$, e vice-versa. Tanto para testes de hipóteses, quanto para intervalos de confiança, os procedimentos buscam por um intercâmbio entre estatísticas e parâmetros. No teste de hipóteses, fixa-se o parâmetro e pergunta-se para quais valores amostrais a hipótese nula será rejeitada. Em contra partida, o intervalo de confiança fixa o valor amostral e obtém um sub-conjunto do espaço paramétrico tal que, para cada ponto deste sub-conjunto, a probabilidade de se observar valores mais extremos que o de fato observado não seja maior que $(1-\gamma)/2$. Portanto, existe uma correspondência entre a região de aceitação de um teste e o intervalo de confiança.

**Teorema 1** *Para cada $\theta_0 \in \Theta$, seja $A(\theta_0)$ a região de aceitação de um teste de nível $\alpha \in (0,1)$ para testar $H_0 : \theta = \theta_0$. Para cada $\tilde{\mathbf{x}} \in \chi$, defina um conjunto $C(\tilde{\mathbf{x}})$ no espaço paramétrico tal que: $C(\tilde{\mathbf{x}}) = \{\theta_0 : \tilde{\mathbf{x}} \in A(\theta_0)\}$. Então, o conjunto aleatório $C(\tilde{\mathbf{X}})$ é um conjunto de confiança com coeficiente de confiança de $(1-\alpha)$. Para qualquer $\theta_0 \in \Theta$, definimos: $A(\theta_0) = \{\tilde{\mathbf{x}} : \theta_0 \in C(\tilde{\mathbf{x}})\}$, que é a região de aceitação de um teste de nível $\alpha$ para testar $H_0 : \theta = \theta_0$.*

Para elucidar, vamos considerar o seguinte exemplo: suponha que $x_1, \ldots, x_n$ seja uma amostra observada após a realização aleatória de $n$ experimentos independentes de uma população normal com média $\mu$, desconhecida, e variância $\sigma^2$ conhecida. Usando um nível de significância $\alpha \in (0,1)$, e para testar as hipóteses $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, considere o seguinte teste: a hipótese nula será rejeitada se $\bar{x} \in A(\mu_0)$, onde $A(\mu_0) = \{\bar{x} \in \Re : |\bar{x} - \mu_0| > Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}$. $H_0$ não é rejeitada para $\{|\bar{x} - \mu_0| \leq Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}$. Antes da realização de um experimento, isto é equivalente ao evento:

$$\{\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}. \tag{1}$$

Como este é um teste de nível $\alpha$, temos que $Pr(H_0 \text{ ser rejeitada} | \mu = \mu_0) = \alpha$, ou ainda, $Pr(H_0 \text{ não ser rejeitada} | \mu = \mu_0) = 1 - \alpha$. De (1), temos:

$$Pr\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} | \mu = \mu_0\right) = 1 - \alpha. \tag{2}$$

Veja que a expressão (2) é verdadeira para todo $\mu_0$ tomado do espaço paramétrico. Portanto, para $\mu$ arbitrário, temos:

$$Pr\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu\right) = 1 - \alpha. \quad (3)$$

Do que concluímos que o intervalo $\left[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$, obtido pela inversão da região de aceitação do teste de nível $\alpha$, é um intervalo de confiança para um coeficiente de confiança de $(1 - \alpha)$.

### 2.1.2 Método da Quantidade Pivotal

Segundo [4], uma quantidade pivotal é definida da seguinte forma: uma variável aleatória $Q(\tilde{\mathbf{X}}; \theta)$ é dita ser uma quantidade pivotal para o parâmetro $\theta$ se a sua distribuição não depende de $\theta$. Com o uso de uma quantidade pivotal, a obtenção do intervalo de confiança será geralmente simples, pois bastará aplicar os dois passos a seguir:

(a) Obtenha uma função da amostra que dependa de $\theta$, digamos $U = Q(\tilde{\mathbf{X}}; \theta)$, mas cuja distribuição não dependa de $\theta$.

(b) Encontrar constantes $a$ e $b$ tais que $Pr(a \leq U \leq b) \geq 1 - \alpha$.

Além disso, se para cada $\tilde{\mathbf{x}} \in \chi$ existirem funções de valor real $t_1(\tilde{\mathbf{x}})$ e $t_2(\tilde{\mathbf{x}})$ tais que $Pr(a \leq G(\tilde{\mathbf{x}}; \theta) \leq b)$ se e somente se $t_1(\tilde{\mathbf{x}}) \leq \theta \leq t_2(\tilde{\mathbf{x}})$, então:

$$Pr[t_1(\tilde{\mathbf{X}}) \leq \theta \leq t_2(\tilde{\mathbf{X}})|\theta] = \gamma. \quad (4)$$

Vejamos o seguinte exemplo. Sejam $X_1, \ldots, X_n$ uma amostra aleatória da variável aleatória $X$ com densidade $f_X(x|\theta) = \theta e^{-\theta x}$, para $\theta, x > 0$, e $f_X(x|\theta) = 0$, caso contrário. Seja $T(\tilde{\mathbf{X}}) = \sum_{i=1}^{n} X_i$. Observe que $T(\tilde{\mathbf{X}})$ não é uma quantidade pivotal, pois sua distribuição depende de $\theta$. É fato bem conhecido que $T(\tilde{\mathbf{X}}) \sim Gama(n, \theta)$. Denote a densidade de $T(\tilde{\mathbf{X}})$ por $f_T(t|\theta)$. Assim,

$$f_T(t|\theta) = \frac{1}{\Gamma(n)}\theta^n t^{n-1} e^{-\theta t}. \quad (5)$$

Defina $Q(\tilde{\mathbf{X}}, \theta) = 2\theta \times T(\tilde{\mathbf{X}})$, que por sua vez tem distribuição Qui-quadrado com $(2n)$ graus de liberdade, ou seja, a função densidade de probabilidade de $Q(\tilde{\mathbf{X}}, \theta)$ não depende de $\theta$ e é dada por:

$$f_Q(y) = \frac{1}{\Gamma(n)}\frac{y^{(n-1)}}{2n} e^{\frac{(-y)}{2}}. \quad (6)$$

Portanto, $Q(\tilde{\mathbf{X}}, \theta)$ é uma quantidade pivotal com respeito a $\theta$. A construção do intervalo de confiança é feita diretamente pelo uso da distribuição de $Q(\tilde{\mathbf{X}}, \theta)$ da seguinte forma:

$$Pr[a \leq Q(\tilde{\mathbf{X}}, \theta) \leq b] = Pr[a \leq 2\theta T(\tilde{\mathbf{X}}) \leq b] = Pr\left[\frac{a}{2T(\tilde{\mathbf{X}})} \leq \theta \leq \frac{b}{2T(\tilde{\mathbf{X}})}\right],$$

onde $a$ e $b$ são constantes que satisfazem $Pr[a \leq Q(\tilde{\mathbf{X}}, \theta) \leq b] = 1 - \alpha$. Suponha uma amostra com $n = 15$ observações e $\alpha = 0,05$. Assim, pela distribuição Qui-quadrado com 30 graus de liberdade, temos que os percentis $2,5$ e $97,5$ são $a = 16,791$ e $b = 46,979$, respectivamente. Para uma dada amostra observada, suponha que $\sum_{i=1}^{n} x_i = 8,5$. Portanto, o intervalo de confiança observado, com base em um coeficiente de confiança de $0,95$, será $I.C._{(0,95)}(\theta) = [16,791/(2 \times 8,5), 46,979/(2 \times 8,5)] = [0,9877, 2,7635]$.

## 2.2 Métodos Assintóticos

Em certos problemas não é possível encontrar o intervalo de confiança exato. Isto pode ocorrer por não se conhecer a distribuição da estatística utilizada ou pelo fato de tal distribuição ser de difícil manipulação algébrica. Neste caso, métodos assintóticos podem ser uma boa alternativa. A literatura de inferência estatística oferece um vasto acerto de métodos assintóticos para construção de intervalos de confiança. Para uma revisão mais minuciosa, sugerimos o livro de [10]. Aqui, descreveremos quatro dentre os métodos assintóticos mais usados.

### 2.2.1 Método Delta

Este método baseia-se na eficiência assintótica dos estimadores de máxima verossimilhança. Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória, e denote a função densidade de probabilidade (ou função de probabilidade no caso discreto) de $X_i$, para $i = 1, \ldots, n$, por $f_X(x|\theta)$. Seja $\hat{\theta}$ o estimador de máxima verossimilhança (EMV) para $\theta$, e seja $L(\theta|\tilde{\mathbf{X}} = \tilde{\mathbf{x}}) = L(\theta)$ a função verosssimilhança avaliada em $\theta$ dada a amostra observada $\tilde{\mathbf{x}}$. A variância da função de valor real $h(\hat{\theta})$ pode ser aproximada por:

$$Var(h|(\hat{\theta}) \approx \frac{\left[\frac{\partial}{\partial\theta}h(\theta)\right]^2|_{\theta=\hat{\theta}}}{\frac{-\partial^2}{\partial\theta^2}logL(\theta|x)|_{\theta=\hat{\theta}}}. \tag{7}$$

Assim, um intervalo de confiança aproximado para $h(\hat{\theta})$, com $100(1-\alpha)\%$ de confiança, pode ser obtido da seguinte forma:

$$I.C_{(1-\alpha)}(\theta) = \left[h(\hat{\theta}) - z_{\frac{\alpha}{2}}\sqrt{Var[h(\hat{\theta})|\theta]}, h(\hat{\theta}) + z_{\frac{\alpha}{2}}\sqrt{Var[h(\hat{\theta})|\theta]}\right], \tag{8}$$

onde $z_{(\frac{\alpha}{2})}$ é tal que $Pr(Z \leq z_{(\frac{\alpha}{2})}) = 1 - \alpha/2$, e $Z \sim N(0,1)$.

Vejamos um exemplo: sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória tomada de uma população seguindo uma distribuição exponencial de parâmetro $\lambda$, ou seja, a função densidade de $X_i$, $i = 1, \ldots, n$, é $f_X(x|\lambda) = \lambda e^{-\lambda x}$. Vamos construir um intervalo de confiança aproximado, de $100(1-\alpha/2)\%$ de confiança, para $\lambda$. Primeiramente, para uma hipotética amostra observada $x_1, x_2, \ldots, x_n$, vamos obter os termos necessários ao uso da expressão (7):

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n}\lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}, \\
&\Rightarrow \quad log[L(\lambda)] = nlog(\lambda) - \lambda\sum_{i=1}^{n}x_i, \\
&\Rightarrow \quad \frac{\partial}{\partial\lambda}log[L(\lambda)] = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i, \\
&\Rightarrow \quad \frac{\partial^2}{\partial\lambda^2}log[L(\lambda)] = -\frac{n}{\lambda^2}.
\end{aligned} \tag{9}$$

Fazendo $h(\theta) = \lambda$, temos que $\frac{\partial}{\partial\lambda}h(\theta) = 1$, e fazendo $\frac{\partial}{\partial\lambda}log[L(\lambda|x)] = 0 \Rightarrow \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n}x_i = 0 \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$, que é o EMV para $\lambda$, pois a segunda derivada de $log[L(\lambda)]$ em relação a $\lambda$ é negativa para todo $\lambda > 0$. De (7), a variância aproximada de $\hat{\lambda}$ é calculada com:

$$Var(\hat{\lambda}|\lambda) \approx \frac{1}{-\left(-\frac{n}{\lambda^2}\right)} = \frac{1}{-\left(-\frac{n}{(1/\bar{x})^2}\right)} = \frac{1}{\bar{x}n}. \tag{10}$$

Desta forma, um intervalo de confiança aproximado para $\lambda$, com $100(1-\alpha)\%$ de confiança, é dado por:

$$I.C_{(1-\alpha)}(\lambda) = \left[\frac{1}{\bar{x}} - Z_{(\frac{\alpha}{2})}\frac{1}{\sqrt{\bar{x}n}},\ \frac{1}{\bar{x}} + Z_{(\frac{\alpha}{2})}\frac{1}{\sqrt{\bar{x}n}}\right]. \tag{11}$$

### 2.2.2 Método Baseado na Estatística Escore

O método que discutiremos nesta seção nada mais é do que um caso particular do método Delta discutido na última seção, onde a função de valor real

$$Q(\tilde{\mathbf{X}}|\theta) = \frac{\frac{\partial}{\partial\theta}L(\theta|\tilde{\mathbf{x}})}{\sqrt{-E\frac{\partial^2}{\partial\theta^2}log[L(\theta|\tilde{\mathbf{x}})]}} \tag{12}$$

conhecida como 'estatística escore', é usada como uma espécie de quantidade pivotal assintótica no sentido de que sua distribuição de probabilidade praticamente não depende de $\theta$ quando $n$ é suficientemente grande. Mais ainda, $Q(\tilde{\mathbf{X}}|\theta)$ converge em distribuição, com respeito a $n$, para uma distribuição normal padrão. Para exemplificar, consideremos o caso de uma amostra aleatória, de tamanho $n$, tomada de uma população Bernoulli ($p$) (exemplo de [10], página 498). Veja que o EMV de $p$ é dado por $\hat{p} = \sum_{i=1}^{n} X_i/n$. Assim, temos:

$$Q(\tilde{\mathbf{X}}|\theta) = \frac{\frac{y}{p} - \frac{n-y}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}. \tag{13}$$

Com base nesta abordagem, um intervalo de confiança aproximado, com coeficiente de confiança de $100(1-\alpha)\%$, é dado por $[p_1,\ p_2]$, onde

$$\begin{aligned} p_1 &= \sup\left\{p : (\hat{p}-p)/\sqrt{p(1-p)n^{-1}} \geq z_{(1-\frac{\alpha}{2})}\right\}, \text{ e} \\ p_2 &= \inf\left\{p : (\hat{p}-p)/\sqrt{p(1-p)n^{-1}} \leq -z_{(1-\frac{\alpha}{2})}\right\}. \end{aligned} \tag{14}$$

onde $z_{(1-\frac{\alpha}{2})}$ é tal que $Pr(Z \leq z_{(1-\frac{\alpha}{2})}) = 1-\alpha/2$, e $Z \sim N(0,1)$.

### 2.2.3 Método Baseado na Estatística da Razão de Verossimilhanças

Um fato bem conhecido é que, se $\lambda(\tilde{\mathbf{X}}, \theta_0) = \left(\frac{L(\theta|\tilde{\mathbf{X}})}{L(\hat{\theta}|\tilde{\mathbf{X}})}\right)$ é a estatística da razão de verossimilhanças, com $\hat{\theta}$ o EMV de $\theta$, então a distribuição assintótica da tranformada $-2log[\lambda(\tilde{\mathbf{X}}, \theta_0)]$, sob certas condições, é Qui-quadrado com um grau de liberdade (ver [10], página 496), onde $\theta_0$ é um valor fixo e conhecido previamente especificado sob a hipótese nula em um teste de hipóteses bilateral. Portanto, novamente com inspiração no método da inversão de um teste, o intervalo de confiança para $\theta$ é dado por $[\theta_1, \theta_2]$, onde:

$$\begin{aligned} \theta_1 &= \sup\left\{0 \leq \theta < \hat{\theta} : -2log[\lambda(\tilde{\mathbf{X}}, \theta)] \geq q_{(1-\alpha)}\right\}, \text{ e} \\ \theta_2 &= \inf\left\{\theta > \hat{\theta} : -2log[\lambda(\tilde{\mathbf{X}}, \theta)] \geq q_{(1-\alpha)}\right\}, \end{aligned} \tag{15}$$

onde $q_{(1-\alpha)}$ é o percentil $100(1-\alpha)\%$ de uma distribuição Qui-quadrado com 1 grau de liberdade.

### 2.2.4 Métodos Baseados em Aproximações pela Distribuição Normal

Sejam $\hat{\theta}$ e $\hat{\sigma}$ estatísticas para o parâmetro $\theta$ tais que:

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}} \to N(0,1), \quad quando \quad n \to \infty, \tag{16}$$

então o intervalo de confiança aproximado para $\theta$, de coeficiente $\gamma$, pode ser obtido por:

$$I.C_\gamma(\theta) = \left[\hat{\theta} - \hat{\sigma}z_{(1-\gamma)/2}, \hat{\theta} + \hat{\sigma}z_{(1-\gamma)/2}\right], \tag{17}$$

onde $z_{(1-\gamma)/2}$ é o percentil $100\gamma\%$ da distribuição normal padrão. Como caso particular, se $X_1, X_2, \ldots, X_n$ são i.i.d com média $\mu$ e variância $\sigma^2$, então, pelo do Teorema Central do Limite, temos:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1). \tag{18}$$

Além disso, se $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1) \to \sigma^2$ em probabilidade, então:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to N(0,1). \tag{19}$$

Portanto, o intervalo de confiança aproximado de $100\gamma\%$ de confiança pode ser obtido por:

$$\left[\bar{X} - S\frac{z_{\left(\frac{1-\gamma}{2}\right)}}{\sqrt{n}}, \bar{X} + S\frac{z_{\left(\frac{1-\gamma}{2}\right)}}{\sqrt{n}}\right]. \tag{20}$$

# 3 Intervalos de Confiança Monte Carlo

Como vimos na Seção 2.1, o método exato convencionalmente usado para obtenção de intervalos de confiança é baseado na estratégia de inverter um teste de hipóteses. Lembremos que, se $T$ for uma estatística de teste, $t_0$ um valor realizado de $T$, e $F_T(t|\theta)$ denotar a distribuição de probabilidade de $T$ dado o parâmetro populacional $\theta$ de interesse, então um intervalo exato, com $100\times(1-2\alpha)\%$ de confiança, para $\theta$, digamos $\left[\hat{\theta}_l,\ \hat{\theta}_u\right]$, pode ser obtido resolvendo $F_T(t_0|\theta = \hat{\theta}_l) = (1-\alpha)$, e $F_T(t_0|\theta = \hat{\theta}_u) = \alpha$. Mas isso é válido se $F_T(t|\theta)$ for crescente em $\theta$. Se $F_T(t|\theta)$ for descrecente em $\theta$, então a solução é dada por se intercambiar as restrições $(1-\alpha)$ e $\alpha$ nas probabilidades $F_T(t_0|\theta = \hat{\theta}_l)$ e $F_T(t_0|\theta = \hat{\theta}_u)$, respectivamente . Se $F_T(t|\theta)$ não for monótona em $\theta$, então o que se obtém com o método da inversão de um teste não é um intervalo, mas sim um 'conjunto confiança', por sua vez formado por uniões de intervalos semi-abertos da reta real. Se a forma analítica da $F_T(t|\theta)$ é desconhecida, ou de difícil manipulação prática, mas amostras de $T$ podem, de alguma forma, ser geradas para valores fixos de $\theta$, então algum método Monte Carlo (MC) pode ser usado para construção de intervalos de confiança para $\theta$. Tais métodos serão referidos aqui apenas por 'intervalos de confiança Monte Carlo', ou simplesmente 'intervalos de confiança MC'.

Em geral, devido à variabilidade introduzida pela simulação Monte Carlo, a verdadeira probabilidade de cobertura dos intervalos de confiança MC não é uma constante, mas sim uma variável aleatória que oscila em torno do coeficiente de confiança do que seria o inviável método exato, mas tais oscilações tornan-se cada vez mais concentradas nas proximidades do coeficiente exato a medida com que o número de simulações de Monte Carlo aumenta. Um exemplo interessante de aplicação do intervalo de confiança MC é o problema de estimação de populações de animais por meio da técnica de amostragem de captura-recaptura [6, 7, 18]. Atualmente, pode-se dizer que, dentre os métodos Monte Carlo, três se destacam por sua generalidade e boa performance, sendo eles: (i) o método 'percentil', também conhecido por método 'Bootstrap paramétrico'; (ii) o método 'aleatorizado', inspirado no método da inversão de um teste de hipóteses aleatorizado, e; (iii) o método 'sequencial', que por sua vez é inspirado na inversão do teste Monte Carlo convencional. A descrição destes três métodos é o assunto deste artigo a partir daqui.

## 3.1 Método Percentil

O método percentil, descrito em detalhes por [9], pode ser visto como uma versão melhorada, no sentido de valer para uma gama maior de situações, do método Bootstrap paramétrico de [13]. Suponha que $T$ seja um estimador consistente de $\theta$ e que, apesar da forma analítica de $F_T(t|\theta)$ ser desconhecida, assuma que uma amostra Monte Carlo $T_1, \ldots, T_{M-1}$ de $T$ possa ser gerada

para $\theta := t_0$. Denote por $T_{(i)}$ a i-ésima observação da amostra ordenada, ou seja, $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(M-1)}$. O intervalo de confiança aproximado para $\theta$, de $100 \times (1-2\alpha)\%$ de confiança, baseado no método percentil, é dado por $[T_{(K_1)}, T_{(K_2)}]$, onde $K_1 = \alpha M$, e $K_2 = M(1-\alpha)$.

Para que a média do coeficiente de confiança verdadeiro do método percentil, com respeito à distribuição de $T$ e para $M$ próximo de infinito, seja igual ao coeficiente $(1-2\alpha)$ desejado, [8] mostra que a chamada 'condição de simetria' precisa ser atendida. A condição de simetria é satisfeita se $F_T(t|\theta = t_0) = 1 - F_T(t_0|\theta = t)$ para todo $t$ e $t_0$. Se a condição de simetria não se verifica, [9] mostra que o método percentil pode levar a intervalos de confiança extremamente assimétricos e, ainda mais grave, sua probabilidade de cobertura converge, tanto em $M$ quanto em $n$ (tamanho da amostra), para uma probabilidade de cobertura maior que o coeficiente de confiança desejado.

Para entendermos melhor a condição de simetria, vamos considerar o caso em que $F_T(t|\theta)$ segue uma distribuição Normal$(\mu, \sigma^2)$, onde $\mu$ é o parâmetro para o qual desejamos construir o intervalo de confiança. Veja que, como $[(t-\mu)]^2 = [(\mu - t)]^2$ para todo $t$ e $\mu$ reais, é trivial que

$$f_T(t_0|\mu = \mu_0) = f_T(\mu_0|\mu = t_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\mu_0 - t_0)^2}. \tag{21}$$

o que nos leva a concluir que a condição de simetria é atendida neste caso. Isto pode nos induzir à errônea conclusão de que todas as distribuições simétricas em torno da média atendem a condição de simetria. O caso em que $T$ segue uma distribuição t-student com $v$ graus de liberdade é um ótimo contra-exemplo. Neste caso, temos:

$$f_T(T|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{2}\right)^{-\frac{(v+1)}{2}}, \tag{22}$$

do que é fácil ver que, se $v \neq t$, então $f_T(t|v) \neq f_T(v|t)$, ou seja, apesar da distribuição t-student ser simétrica em torno da média, ela não atende a condição de simetria. Naturalmente, será também comum encontrarmos distribuições assimétricos em torno da média que também não atendem a condição de simetria. Tomemos o caso exponencial de média $1/\lambda$. Neste caso, a densidade de $T$ será dada por:

$$f_T(t|\lambda) = \lambda e^{-\lambda t}, t, \lambda > 0. \tag{23}$$

Veja que, para $\lambda = 2$ e $t = 1$, temos: $f_T(t = 1|\lambda = 2) = 2e^{-2} \neq f_T(t = 2|\lambda = 1) = e^{-2}$. Portanto, a condição de simetria não é satisfeita também no caso exponencial.

### 3.1.1 Exemplo de Aplicação do Método Percentil - Estimação de Tamanho Populacional por Captura-recaptura

Um ótimo exemplo de uso do método percentil foi oferecido por [9] para o problema de se estimar o tamanho de populações de animais via o mecanismo de amostragem baseado na captura, marcação, devolução e recaptura de indivíduos (captura-recaptura). Por exemplo, pode-se realizar estimação intervalar da população do mico leões dourados em seu habitat, a Mata Atlântica. Suponha que uma amostra de $n_1$ animais, de uma determinda espécie, é tomada de uma reserva ecológica onde reside um total de $N$ animais desta espécie, onde $N$ é o parâmetro desconhecido que se quer estimar. Estes $n_1$ animais são marcados e devolvidos à reserva. Depois de um certo período de tempo, uma segunda amostra, agora de tamanho $n_2$, é retirada desta mesma reserva. Seja $m$ o número de animais marcados (ou seja, indivíduos que estavam na primeira amostra de tamanho $n_1$) dentre os $n_2$ animais desta segunda amostra. Um estimador pontual para $N$, proposto por [9], é dado por:

$$\hat{N}^* = \left[\frac{(n_1+1)(n_2+1)}{m+1}\right] - 1, \tag{24}$$

Tabela 1: Estimativas para $N$, pontuais pelo método de captura e recaptura e intervalares ($\gamma = 0,95$) com base no método percentil, usando diferentes valores amostrais ($m$) observados e fixando $n_1 = n_2 = 300$ no problema.

| | $m$ | | | |
|---|---|---|---|---|
| | 20 | 30 | 50 | 100 |
| $\hat{N}^*$ | 4313,33 | 2921,61 | 1775,49 | 896,04 |
| $I.C_{0,95}(N)$ | $[2921,61, 6470,50]$ | $[2156,17, 4132,00]$ | $[1437,11, 2264,03]$ | $[793,75, 1028,56]$ |
| Amplitude | 3548,89 | 1975,83 | 826,92 | 234,81 |

que é não viciado se $n_1 + n_2 \leq N$, ou seja, $E(\hat{N}^*) = N$ se $n_1 + n_2 \leq N$. Vamos agora proceder à estimação intervalar. Como já enfatizado anteriormente, a validade do método percentil depende da veracidade da satisfação da condição de simetria para a distribuição de $\hat{N}^*$. Portanto, vamos analisar a situação em que $n_1 = n_2 = 300$, $t = 850$ e $\theta = N = 1000$. Para estes valores, temos que:

$$Pr(\hat{N}^* \leq 850|N = 1000) = Pr\left\{\left[\frac{(n_1+1)(n_2+1)}{m+1}\right] - 1 \leq 850|N = 1000\right\} = 0,55. \quad (25)$$

Por outro lado, se intercambiarmos $t$ e $N$, ou seja, fazendo $t = 1000$ e $N = 850$, temos:

$$Pr(\hat{N}^* \leq 1000|N = 850) = 1 - Pr\left\{\left[\frac{(n_1+1)(n_2+1)}{m+1}\right] - 1 \leq 1000|N = 850\right\} = 0,14. \quad (26)$$

Portanto, a condição de simetria não é satisfeita, pois $Pr(\hat{N}^* \leq 850|N = 1000) \neq 1 - Pr(\hat{N}^* \leq 1000|N = 850)$. Entretanto, vamos negligenciar a não verificação da condição de simetria por enquanto, e aplicar o método para o problema de captura-recaptura tal como sugerido por [9]. Ilustraremos os efeitos desta falha na verificação da condição de simetria na Seção 5, mais precisamente com a Tabela 5. A Tabela 1 mostra valores de intervalos de confiança obtidos pelo método percentil para cinco cenários fictícios distintos para o número de animais marcados dentre os da segunda amostra. Usou-se $M = 10000$, $\gamma = 0,95$ e um planejamento amostral com $n_1 = n_2 = 300$. Esta tabela revela um padrão bem marcante para o problema de estimação populacional, que é o fato do intervalo tornar-se cada vez mais concentrado em torno da estimativa pontual conforme o valor de $m$ aumenta.

Visando favorecer o entendimento sobre como aplicar o método percentil, consideremos um exemple em que o valor de $M$ seja bem pequeno, pois assim será possível mostrar, passo a passo, a obtenção dos limites inferior e superior do intervalo. Ainda para o problema de captura-recaptura, considere agora amostras $n_1 = n_2 = 100$, e suponha que foram observados 30 animais marcados dentre os 100 animais da segunda amostra. Assim, a estimativa pontual para $N$ será:

$$\hat{N}^* = \left[\frac{(n_1+1)(n_2+1)}{m+1}\right] - 1 = \left[\frac{(100+1)(100+1)}{30+1}\right] - 1 = 328,07. \quad (27)$$

Para obtermos uma estimativa intervalar, façamos $M = 20$. Seguem, já ordenados, os 20 valores Monte Carlo simulados:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 267,47 | 274,70 | 282,36 | 299,03 | 299,03 | 308,12 | 308,12 | 328,06 |
| 328,06 | 328,06 | 339,03 | 339,03 | 339,03 | 350,76 | 350,76 | 363,32 |
| 376,81 | 391,35 | 391,35. | | | | | |

Para gerar os valores acima, bastou gerar valores para a variável $m$ e aplicá-los à função $\hat{N}^*$. Note que $m$ segue uma distribuição hipergeométrica, e que neste exemplo está parametrizada por $(n_1 = 100, n_2 = 100, N = 328)$. Para $\alpha = 0,20$, temos $K_1 = \alpha M/2 = 0,20 \times 20/2 = 2$, e $K_2 = (1 - \alpha/2)M = (1 - 0,20/2) \times 20 = 18$. Ou seja, o limite inferior do intervalo será o valor de posição 2 da amostra acima, enquanto que o limite superior será o valor de posição 18, do que obtemos a estimativa intervalar $[274,70, 391,35]$.

### 3.1.2 Performance do Método Percentil

Agora mostraremos que, se a condição de simetria é satisfeita, então $E(\hat{\gamma}_M) = (1-2\alpha)$, onde $\hat{\gamma}_M$ é a probabilidade de cobertura do método percentil, e $(1-2\alpha)$ é a probabilidade de cobertura desejada, ou seja, queremos mostrar que o método percentil é exato se a condição de simetria é satisfeita. Sem perda de generalidade, isto será mostrado para o caso unilateral, onde $I.C. = (\hat{\theta}_1; \infty)$, e isto é suficiente, pois $\theta$ não pertencerá ao intervalo bilateral apenas por um dos dois limites a cada aplicação do método, ou seja, se $\theta \notin [T_{(K_1)}, T_{(K_2)}]$, e $T_{(K_1)} < T_{(K_2)}$ , então $\theta < T_{(K_1)}$ ou $\theta > T_{(K_2)}$. Assim, a probabilidade de não cobertura bilateral é a soma das probabilidades de não cobertura unilaterais. Portanto, bastará mostrar que o limite inferior do intervalo será menor do que $\theta$ com probabilidade de de no mínimo $(1-\alpha)$. Para uma estimativa $\hat{\theta}$ observada, temos que:

$$\hat{\gamma}_M = Pr(T_{(K_1)} \leq \theta | \hat{p}) = \sum_{x=h}^{M-1} \left( \begin{array}{c} M-1 \\ x \end{array} \right) \hat{p}^x (1-\hat{p})^{M-1-x}, \tag{28}$$

onde $K_1 = h = \alpha M$ e $\hat{p} = Pr(T_i \leq \theta | \hat{\theta})$, e seja $\hat{P}$ a correspondente variável aleatória assoaciada ao observado $\hat{p}$. Assim, a probabilidade de cobertura integrada para todo $\hat{\theta}$ será:

$$Pr(\hat{\theta}_1 \leq \theta) = \int_0^1 \left( \sum_{x=h}^{M-1} \left( \begin{array}{c} M-1 \\ x \end{array} \right) \hat{p}^x (1-\hat{p})^{M-1-x} f_{\hat{P}}(\hat{p}) \right) d\hat{p}, \tag{29}$$

onde $f_{\hat{P}}(\hat{p})$ é a função densidade de probabilidade da variável $\hat{P}$. Suponha que a condição de simetria se verifica, ou seja, $F_T(\theta|\hat{\theta}) = 1 - F_T(\hat{\theta}|\theta)$ para todo $\theta$ e $\hat{\theta}$. Por motivos didáticos, vamos usar agora a notação $\hat{\Theta}$ também para representar a variável aleatória $T$, ou seja, defina $T := \hat{\Theta}$. Observe que, como $\hat{\theta}$ é uma observação oriunda da distribuição de $T$ dado $\theta$, então $1 - F_T(\hat{\Theta}|\theta) \sim U(0,1)$, implicando que, pela condição de simetria, $\hat{P} = F_T(\theta|\hat{\Theta}) \sim U(0,1)$. Assim, $f_{\hat{P}}(\hat{p}) = \mathbf{1}_{\hat{p} \in (0,1)}(\hat{p})$. Desta forma, de (29), temos que:

$$\begin{aligned}
Pr(T_{(K_1)} \leq \theta) &= \sum_{x=h}^{M-1} \left[ \left( \begin{array}{c} M-1 \\ x \end{array} \right) \int_0^1 \hat{p}^x (1-\hat{p})^{M-1-x} d\hat{p} \right] \\
&= \sum_{x=h}^{M-1} \left[ \left( \begin{array}{c} M-1 \\ x \end{array} \right) \int_0^1 \hat{p}^{(x+1)-1} (1-\hat{p})^{(M-1-x+1)-1} d\hat{p} \right] \\
&= \sum_{x=h}^{M-1} \left[ \left( \begin{array}{c} M-1 \\ x \end{array} \right) \frac{x!(M-1-x)!}{M(M-1)!} \right] \\
&= \sum_{x=h}^{M-1} \left[ \frac{\left( \begin{array}{c} M-1 \\ x \end{array} \right)}{M \left( \begin{array}{c} M-1 \\ x \end{array} \right)} \right] \\
&= \sum_{x=h}^{M-1} \left( \frac{1}{M} \right) \\
&= \frac{M-1-h+1}{M} = \frac{M}{M} - \frac{h}{M} \\
&= 1 - \frac{\alpha M}{M} \\
&= 1 - \alpha.
\end{aligned} \tag{30}$$

Portanto, está provado que o método percentil é exato sob a condição de simetria.

### 3.1.3 Limitações do Método Percentil

Apesar do método percentil ser simples e intuitivo, seu uso exige cautela. Como $F_T(t|\theta)$ é desconhecida em aplicações práticas, a incerteza ao redor da atual probabilidade de cobertura é uma séria limitação devido à impossibilidade de se verificar a condição de simetria. [9] sugere o uso de uma correção de vício para atenuar o problema, mas [9] adverte que tal artifício pode produzir intervalos assimétricos quando as amostras são pequenas ou de magnitudes moderadas. O método percentil não é aplicável nos casos em que $T$ não é um estimador para $\theta$, o que não é raro, por exemplo, no contexto de teste de hipóteses.

Já vimos alguns casos em que a condição de simetria não é válida, e estes envolveram a distribuição t-student e exponencial. Vimos também um exemplo envolvendo um caso discreto, ilustrado pelo uso do problema de estimação populacional citado na seção 3.1.1 pelo estimador $\hat{N}^*$. Na prática, a condição de simetria para a distribuição de $T$ implica as seguintes restrições:

1. $T$ precisa ser um estimador de $\theta$;
2. O suporte de $T$ e o espaço paramétrico precisam coincidir;
3. A distribuição $F_T(t|\theta)$ deve ser não crescente em $\theta \; \forall \; t \in \mathbb{R}$.

Os itens 1 e 2 resultam diretamente da construção do método, do que resta demonstrar a restrição 3. Veja que para $t_1 < t_2 \in \mathbb{R}$, vale que $F_T(t_1|\theta = y) \leq F_T(t_2|\theta = y)$. Pela condição de simetria, temos que $1 - F_T(y|\theta = t_1) \leq 1 - F_T(y|\theta = t_2) \Rightarrow F_T(y|\theta = t_1) \geq F_T(y|\theta = t_2)$. Como esta conclusão vale para todo $(t_1 < t_2)$ e $y \in \mathbb{R}$, conclui-se que a retrição 3 está provado.

## 3.2 Método Aleatorizado

[5] propõem um método bem geral que utiliza do procedimento de aleatorização para obter intervalos de confiança Monte Carlo, o qual foi inspirado na ideia de se inverter o teste de MC analogamente ao método exato convencional. Primeiramente, é necessário supor que $F_T(t|\theta)$ seja monótona em $\theta$. O procedimento requer a existência e o conhecimento analítico por parte do usuário de uma nova variável aleatória $Z$, a qual é usada para criar uma nova variável $T^*$, que por sua vez precisa possuir a mesma tenha a mesma distribuição que $T$ para todo $\theta$. Com isso, calcula-se:

$$T^* = g(\theta|Z), \tag{31}$$

onde $g$ deve ser uma função de valor real. Além disso, $g$ deve ser monotonamente decrescente com $\theta$ se $F_T(t|\theta)$ é monótona decrescente em $\theta$, e vice-versa. É também imperativo que a função $g$ seja computável. Assuma que uma amostra Monte Carlo, $Z_1, \ldots, Z_{M-1}$ de $Z$, tenha sido gerada. Assim, a amostra $T^*_{(1)}, T^*_{(2)}, \ldots, T^*_{M-1}$ associada, de acordo com a função em (31), pode ser calculada e utilizada para obter a amostra ordenada $T^*_{(1)} < T^*_{(2)} < \cdots < T^*_{(M-1)}$. Deste modo, um intervalo de confiança, de aproximadamente $100 \times (1 - 2\alpha)\%$ de confiança, é dado por $[\theta^*_l, \theta^*_u]$, onde, se $F_T(t|\theta)$ é monotonamente decrescente em $\theta$, então $\theta^*_l = \sup\{\theta|T^*_{(k1)} \leq t_0\}$ e $\theta^*_u = \inf\{\theta|T^*_{(k2)} \leq t_0\}$, onde $K_1 = \alpha M$, e $K_2 = M(1 - \alpha)$. Se $F_T(t|\theta)$ é monotonamente crescence em $\theta$, o supremo e o ínfimo acima são intercambiados entre $\hat{\theta}^*_l$ e $\hat{\theta}^*_u$.

### 3.2.1 Performance do Método Aleatorizado

Segundo [5], o intervalo $[\hat{\theta}^*_l; \hat{\theta}^*_u]$ é exato quando $M \to \infty$. O método também tem a vantagem de não depender de suposições de difícil verificação, tal como a condição de simetria do método percentil, o que o torna bem mais geral neste sentido. Outra vantagem do método aleatorizado sobre o método percentil é que ele é válido para qualquer estatística cuja distribuição dependa do parâmetro $\theta$, e não apenas para estimadores de $\theta$, ou seja, o suporte de $T$ não precisa coincidir com o espaço paramétrico para que o método sejá aplicável.

### 3.2.2 Limitações do Método Aleatorizado

A principal desvantagem do método aleatorizado é que, para sua utilização, o usuário precisa descobrir uma nova variável $T^*$ que não dependa dos dados, mas que garantidamente tenha a mesma distribuição que $T$. Esta tarefa, evidentemente, exige muita criatividade por parte do usuário, o qual nem sempre possui treinamento técnico em teoria de probabilidades para estudar heurísticamente a possível distribuição de $T$ para cada $\theta$ e disto elaborar uma nova variável $T^*$ tendo distribuição equivalente. Uma transformação óbvia é fazer $T^* = F_T^{-1}(Z|\theta)$, com $Z \sim U(0,1)$. Esta escolha é trivial e ao mesmo tempo inútil, pois se conhecêssemos a forma analítica de $F_T^{-1}(Z|\theta)$ para calcular esta transformação, então também não precisaríamos de um método Monte Carlo para construir o intervalo, pois o método exato seria possível diretamente pelo uso da $F_T(t|\theta)$. [5] forneceu exemplos práticos de aplicação do método aleatorizado para distribuições da família locação-escala, e também mostrou como operar o método para alguns membros da família exponencial. No entanto, não foi fornecida uma regra geral para orientar o usuário na busca da transformação $g$ particular a cada possível aplicação. [5] também adverte que, se a $g$ não é monótona, o intervalo tenderá a ser conservador com respeito ao intervalo exato e, portanto, apresentando amplitudes bem maiores que as do respectivo método exato.

## 3.3 Método Sequencial

Aqui trataremos do método sequencial proposto por [18]. Segundo seu autor, o método é inspirado na inversão do teste Monte Carlo convencional. O teste Monte Carlo convencional é um método bem conhecido e amplamente usado para construir testes de hipóteses quando $F_T(t|\theta)$ é desconhecida ou não computável[12, 2, 3]. Para uma amostra obtida via simulação Monte Carlo, digamos $T_i, i = 1, \ldots, M-1$, o valor-p para o teste Monte Carlo é definido como $P_M = (Y+1)/M$, onde $Y = \sum_{i=1}^{M-1} I(T_i \geq T_0)$. A hipótese nula é rejeitada se $P_M \leq \alpha$, onde $\alpha \in (0,1)$ é o nível de significância desejado. Um fato bastante conhecido é que o teste Monte Carlo é sempre de nível $\alpha$, ou seja, é garantido que $Pr(P_M \leq \alpha|H_0) \leq \alpha$, e isso vale para o caso geral de qualquer estatística de teste e qualquer problema de teste de hipóteses. A igualdade estrita é válida quando $F_T(t|\theta)$ é contínua.

O método sequencial para obtenção do intervalo de confiança consiste na estratégia de gerar amostras Monte Carlo para uma grade de valores fixados para $\theta$, e então, com base nas amostras que apresentam um valor-p de Monte Carlo não significativo, tomar o respectivo conjunto de pontos da grade como a região de confiança para $\theta$. Esta técnica produz intervalos de confiança que são conservadores apenas em função da variabilidade introduzida pela simulação Monte Carlo. A condição suficiente para que o método seja válido é que $F_T(t|\theta)$ seja monótona (crescente ou decrescente) em $\theta$ para cada $t$ fixo, e que $\theta$ seja cotado, ou seja, $\theta \in (\hat{\theta}_1^*; \hat{\theta}_n^*)$ para $\hat{\theta}_1^*$ e $\hat{\theta}_n^*$ conhecidos. O algoritmo para aplicação do método sequencial será descrito para o caso em que a $F_T(t|\theta)$ é crescente em $\theta$, mas a construção para distribuições decrescentes é possível por aplicação de raciocínio análogo.

Admita que $\theta_1^* < \theta_2^* < \cdots < \theta_n^*$ seja uma sequência de constantes convenientemente escolhidas e ordenadas para atender $(\theta_j^* - \theta_{(j-1)}^*) \leq \delta$, $j = 1, 2, \ldots, n$, onde $\delta >$ é arbitrário e representa a precisão desejada, em termons de casas decimais, a ser usada no reporte dos limites do intervalo de confiança.

Seja $\tilde{T}_j$ um vetor de dimensão $(M-1)$ contendo $(M-1)$ cópias Monte Carlo geradas a partir da distribuição de $T$ para $\theta := \theta_j^*$. Seja $Y_j(M)$ a variável aleatória que conta o número de valores no vetor $\tilde{T}_j$ que são menores ou iguais que $t_0$, onde $t_0$ é o valor observado da estatística $T$ para uma amostra fixa. Assim, $Y_j(M)$ segue uma distribuição binomial gerada por $(M-1)$ realizações Bernoulli com probabilidade de sucesso dada por $p_1 = F_T(t_0|\theta = \theta_j^*)$. O limite superior do intervalo de confiança é obtido por se varrer os valores de $j$ de 1 até $n$ sequencialmente. O procedimento consiste em comparar $Y_j(M)$ com $h$, para cada $j$, até o primeiro $j$ tal que $Y_j(M) < h$, ou até ocorrer $j = n$, onde $h = \lfloor \bar{\alpha}(M+1) \rfloor$ e $0 < \bar{\alpha} < \alpha < 1$. Se o procedimento sequencial for interrompido em $j = u$, então o limite superior do intervalo de confiança é dado por $\theta_u^*$.

Para obter o limite inferior de confiança, defina a nova variável $X_j(M)$ para contar o número de valores em $\tilde{T}_j$ que são maiores ou iguais que $t_0$. Veja que $X_j(M)$ segue uma distribuição Binomial gerada por $(M-1)$ experimentos Bernoulli e com probabilidade de sucesso igual a $p_2 = 1 - F_T(t_0|\theta)$. Agora, a avaliação sequencial começa por $j = u$, e varre os valores $j-1, j-2...$ sequencialmente até ocorrer $X_j(M) < h$, ou até $j = 1$. Se o momento de interrupção desta nova varredura ocorrer em $j = l$, então o limite inferior de confiança é dado por $\theta_l^*$. Note que esta rotina garante $\theta_u^* \geq \theta_l^*$ desde que $n$ seja maior que 1. Para constantes arbitrárias $0 < \bar{\alpha} < \alpha_s < \alpha$, um intervalo de confiança conservador de $100 \times (1-2\alpha)\%$ de confiança, $IC_{1-2\alpha}(\theta) = [\theta_u^*, \theta_l^*]$, é obtido para qualquer escolha de $M \geq M_0$, com $M_0$ tal que:

$$Pr(Y_1(M_0) < h|p_1) = \alpha_s \leq 1 - \left[\frac{1-\alpha}{1-\alpha_s}\right]^{\frac{1}{n}}. \tag{32}$$

### 3.3.1 Performance do Método Sequencial

Por simplicidade notacional, algumas vezes ocultaremos aqui o termo $M$ de $Y_j(M)$ e $X_j(M)$, usando simplesmente $Y_j$ e $X_j$. Para um caso hipotético muito raro, quando $\theta_u^* = \theta_l^*$, defina $C(t_0) = [\theta_l, \theta_u)$. Isto certificará que apenas um lado do intervalo poderá falhar em cobrir o valor de $\theta$. Consequentemente, para provar que a probabilidade de cobertura é maior ou igual que $(1-2\alpha)$, é suficiente mostrar que $Pr(\theta_j^* > \theta) \leq (1-\alpha)$. Um raciocínio similar pode ser usado para provar que $Pr(\theta_l^* < \theta) \geq (1-\alpha)$. Defina $P_j = Pr(T \leq t_0|\theta = \theta_j)$, onde $0 < \alpha_s < \alpha$. Na j-ésima iteração, o método sequencial falhará em cobrir o valor $\hat{\theta}(\alpha_s)$ se $Y_j < h$ dado que $\theta_j^* < \hat{\theta}(\alpha_s)$. A probabilidade deste evento ocorrer pode ser cotada da seguinte forma:

$$\begin{aligned} Pr(Y_j < h|\theta_j^* < \hat{\theta}(\alpha_s)) &= Pr(Y_j < h|P_j > \alpha_s) \\ &\leq Pr(Y_j < h|P_j > \alpha_s) \\ &= \pi(h, \alpha_s, M). \end{aligned} \tag{33}$$

A inequação da penúltima linha é verdadeira porque $Pr(Y_j < h|P_j > \alpha_s)$ é decrescente em $P_j$. Assim, se $j^*$ é o maior $j$ tal que $\theta_j^* < \hat{\theta}(\alpha_s)$, então a probabilidade de termos o evento $\theta_u^* \geq \hat{\theta}(\alpha_s)$ é dada por:

$$\begin{aligned} Pr(\theta_u^* > \theta_{j^*}^*) &= Pr(\cap_{j=1}^{j^*} Y_j \geq h|\theta_j^* < \hat{\theta}(\alpha_s)) \\ &\geq [1 - \pi(h, \alpha_s, M)]^{j^*} \\ &\geq [1 - \pi(h, \alpha_s, M)]^n \end{aligned} \tag{34}$$

A última desigualdade é válida porque estamos assumindo que $\theta_n^*$ é maior ou igual $\hat{\theta}(\alpha_s)$ para cada $t_0$. Por construção, $\hat{\theta}(\alpha_s)$ cobrirá o verdadeiro parâmetro $\theta$ pelo menos $100 \times (1-\alpha_s)\%$ dos tempos. Em seguida, para a expressão (4), o limite superior, cobrirá o verdadeiro parâmetro $\theta$ com probabilidade de pelo menos $(1-\alpha_s)[1-\pi(h, \alpha_s, M)]^n$. Para uma probabilidade de cobertura unilateral de $(1-\alpha)$, façamos $(1-\alpha_s) = (1-\alpha_s)[1-\pi(h, \alpha_s, M)]^n$, onde $\alpha_s < \alpha$. Para um $\alpha_s$ fixado, e de acordo com a expressão (34), o limite superior do intervalo de confiança é obtido após encontrar-se o valor $M_0$ que satisfaça:

$$\pi(h, \alpha_s, M_0) \leq 1 - [(1-\alpha)/(1-\alpha_s)]^n. \tag{35}$$

Para $\alpha_s < \alpha$ fixos, o argumento $M_0$ é uma função crescente com respeito a $\bar{\alpha}$. Por exemplo, para um intervalo de $100 \times (1 - 2 \times 0,025)\%$ de confiança, se $\alpha_s = 0,02$ e $n = 100$, o lado direito da expressão (35) é igual a $5,11497 \times 10^{-5}$. Para $\bar{\alpha} = 0,01$, a solução $M_0$ é igual a 2243, o que leva a um valor de $h$ igual a 22. Para $\bar{\alpha}$ igual a $0,015$, o valor de $M_0$ resultante será igual a 10323 ($h = 154$). Naturalmente, a relação entre $M_0$ e $\bar{\alpha}$ depende do valor fixado para $\alpha$, mas, como regra geral, valores de $\bar{\alpha}$ e $\alpha_s$ muito menores que $\alpha$ implicarão valores baixos para $M_0$. Portanto, a fim de poupar esforço computacional na operacionalização da simulação

Monte Carlo, o óbvio deria escolher valores pequenos para $\bar{\alpha}$ e $\alpha_s$, pois isso levaria a pequenos valores para $M_0$. Porém, isso também implicaria em intervalos muito conservadores. Tal como explicado por [18], existe um intercâmbio entre $\bar{\alpha}$, $\alpha_s$ e a verdadeira probabilidade de cobertura.

Para um dado limite superior, $\gamma_{max}$, escolhido para cotar arbitrariamente a real probabilidade de cobertura, existe um valor mínimo para $\bar{\alpha}$ que deve ser respeitado para atender tal cota. Suponha $F_T(t|\theta)$ continua em $t$. Defina $\hat{\theta}_{(\alpha_{in})} = \inf\{\theta : Pr(T_0 \leq t|\theta)\}$, com $0 < \alpha_{(in)} < \hat{\alpha} < \alpha_s < \alpha$ arbitrários. Seja $j^*$ o maior $j$ tal que $\theta_j^* < \hat{\theta}_{\alpha_{(in)}}$. A probabilidade de ocorrer $\theta_u^* > \hat{\theta}_{\alpha_{(in)}}$ é:

$$\begin{aligned} Pr(\theta_u^* > \hat{\theta}_{(\alpha_{in})}) &= Pr(\cap_{j=1}^{j^*} Y_j \geq h) \\ &\leq Pr(Y_j > h|P_1 \leq \alpha_{in}) \\ &\leq Pr(Y_j > h|P_1 = \alpha_{in}) \end{aligned} \tag{36}$$

Assim, a probabilidade de cobertura associada a $\theta_u^*$ pode ser cotada como se segue:

$$\begin{aligned} Pr(\theta_u^* > \theta) &= Pr(\theta_u^* > \theta|\hat{\theta}_{(\alpha_{in})} \leq \theta) \times \\ &\times Pr(\hat{\theta}_{(\alpha_{in})} \leq \theta) + \\ &\times Pr(\theta_u^* > \theta|\hat{\theta}_{(\alpha_{in})} > \theta) \times \\ &\times Pr(\hat{\theta}_{(\alpha_{in})} > \theta) \\ &\leq Pr(\theta_u^* > \hat{\theta}_{(\alpha_{in})}|\hat{\theta}_{(\alpha_{in})} \leq \theta) \times \\ &\times Pr(\hat{\theta}_{(\alpha_{in})} \leq \theta) + (1 - \alpha_{in}) \\ &= Pr(\theta_u^* > \hat{\theta}_{(\alpha_{in})}, \hat{\theta}_{(\alpha_{in})} \leq \theta) + \\ &+ (1 - \alpha_{in}) \\ &\leq Pr(\theta_u^* > \hat{\theta}_{(\alpha_{in})}) + (1 - \alpha_{in}) \\ &\text{pela inequação (36)} \\ &\leq Pr(Y_1 > h|P_1 = \alpha_{in}) + (1 - \alpha_{in}). \end{aligned} \tag{37}$$

Portanto, para garantir uma cota superior, digamos $\gamma_{max}$, para a real probabilidade de cobertura, devemos fazer $Pr(Y_1 \geq h|P_1 = \alpha_{in}) + (1 + \alpha_{in}) = \gamma_{max}$, e resolver para $h$ e $M$. Por exemplo, considere construir um intervalo de confiança conservador aproximado de $100 \times (1 - 0,05)\%$, e suponha que não é permitido ter uma verdadeira probabilidade de cobertura maior que $\gamma_{max} = 0,96$. Para $n = 100$, uma solução é $\alpha_{in} = 0,0205$, $\bar{\alpha} = 0,022$, $\alpha_s = 0,024$ e $M = 100000$.

Observe que $Pr(Y_1 \geq |P_1 = \alpha_{in})$ é decrescente com $M$, pois $Y_1$ torna-se concentrada em torno de $E(Y_1) = \alpha_{in}\ (M + 1)$ quando $M$ aumenta. Para $M$ grande, $Pr(Y_1 \geq h|P_1 = \alpha_{in}) \approx 0$, portanto, será suficiente definir-se $\alpha_{in} = 1 - \gamma_{max}$.

Qualquer escolha $0 < \alpha_{in} < \overline{\alpha} < \alpha_s < \alpha$ levará a uma solução viável, mas algumas podem ser mais eficientes do que as outras no sentido de minimizar o requerido $M_0$, ou seja, o tempo de execução do método. [18] sugere a seguite regra de ouro para escolha dos parâmetros de ajuste do método sequencial: se $\gamma$ é o coeficiente de confiança desejado, e se $\gamma_{(max,dois)}$ é a cota superior admitida para a verdadeira probabilidade de cobertura, ambos arbitrários e fixados pelo usuário, este deverá fazer: $\alpha = \frac{(1-\gamma)}{2}$, $\gamma_{max} = \gamma_{(max,dois)} + \frac{(1-\gamma_{(max,dois)})}{2}$, $\alpha_{in} = 1 - \gamma_{max}$, $\alpha_s = \frac{(2\alpha+\alpha_{in})}{3}$ e $\bar{\alpha} = \frac{(\alpha_s+\alpha_{in})}{2}$. Depois encontrar $M_0$ tal que: $(i)$ satisfaça a expressão (35); e $(ii)$ forneça um valor satisfatoriamente pequeno para a medida $P_r(Y_1 > h|P_1 = \alpha_{in})$ de tal forma que $P_r(Y_1 > h|P_1 = \alpha_{in}) + (1 - \alpha_{in}) \leq \gamma_{max}$.

Para os casos em que a $F_T(t|\theta)$ não é continua em $t$, o limite superior para a probabilidade de cobertura já não será dado pela expressão (37), mas sim igual a $P_r(Y_1 > h|P_1 = \alpha_{in}) \leq \alpha_{in}$.

Os métodos percentil e aleatorizado são baseados em estratégias que não guardam relação com o teste Monte Carlo convencional. Com o método sequencial, sob a luz do que se faz no caso exato, será sempre possível garantir que as decisões tomadas com intervalo de confiança

Monte Carlo sempre concordem com a decisão obtida do teste Monte Carlo. Para mais detalhes, consulte [18].

O método sequencial é tão geral quanto o método aleatorizado, porém não exige a elaboração de uma transformação $g$ para que seja viável. Ademais, o método sequencial não apenas possibilita um controle real da confiança do intervalo, mas também favorece garantir uma cota superior, como uma função do número de simulações Monte Carlo, para a probabilidade de cobertura gerada pelo método. Esta possibilidade de obter, analiticamente, o valor do número de simulações que garanta cotar a probabilidade de cobertura é uma relevante vantagem sobre os métodos anteriores. Ademais, é também o único método que unifica os procedimento de teste e de estimação intervalar simultaneamente na mesma metodologia Monte Carlo.

### 3.3.2 Limitações do Método Sequencial

É notável que a restrição $\theta \in (\hat{\theta}_1^*, \hat{\theta}_n^*)$ pode gerar problemas na prática, pois, como não conhecemos $\theta$, como poderíamos então conhecer cotas para ele? [18] argumenta que isto não seria uma limitação em muitos problemas práticos. Geralmente, o parâmetro tem limites explicitamente definidos pelas próprias características intrínsecas ao fenômeno estudado. Por exemplo, o coeficiente de um modelo de regressão linear que relaciona peso (em quilogramas) e altura (em centímetros) de seres humanos é, em princípio, qualquer número positivo real, mas, por razões práticas, o verdadeiro coeficiente desconhecido não é admitido indicar mais de cem quilos por centímetro adicional na altura de uma pessoa, e portanto uma escolha intrínseca para os limites de $\theta$ seria o intervalo $(0, 100)$. Outra limitação do método é que, dependendo da elevada magnitude de $n$, o valor $M_0$ que garante as propriedades demonstradas poderá ser de magnitudes inviáveis na prática. Para tal problema, [18] sugere um refinamento do procedimento sequencial que quase sempre garantirá tempos de execução satisfatórios, e este é o assunto da próxima seção.

### 3.3.3 Um Refinamento para o Método Sequencial

Algumas aplicações do procedimento sequencial podem envolver amplitudes $(\theta_n^* - \theta_1^*)$ de elevada magnitude frente à desejada precisão decimal $\delta$. Por exemplo, se o parâmetro desconhecido é o tamanho de uma população de uma certa espécie de animal, então $n$ poderá ser da ordem dos milhares, e isto exigia um $M_0$ da ordem dos milhões. Para contornar este tipo de problema, a proposta de [18] é aplicar o algoritmo da bissecção. Para tanto, faça $\theta_{min} := \theta_1^*$ e $\theta_{max} := \theta_n^*$. Agora, o procedimento sequencial consiste em percorrer os valores de $\tilde{T}_j$ enquanto $(U_j - L_j) > \delta$. Para $j = 1$, defina $\theta_u^* := \theta_{max}$, $U_1 := \theta_{max}, L_1 := \theta_{min}$ e $\theta_1^* := (U_1 - L_1)/2$. Assim, $\theta_u^*$, $U_j$ e $L_j$ são atualizados dinamicamente após cada iteração Monte Carlo. Se $Y_j < h$, atualize $\theta_u^* := \theta_j^*$, $U_{(j+1)} := \theta_j^*$ e $L_{(j+1)} = L_j$. Mas, se $Y_j \geq h$, mantenha o valor atual de $\theta_u^*$, faça $U_{(j+1)} := U_j$ e $L_{(j+1)} = \theta_j^*$.

Todas as propriedades já vistas anteriormente também são válidas para este refinamento, mas com a vantagem de promover uma melhoria substancial em termos do número máximo, $n$, de iterações de $(M - 1)$ simulações Monte Carlo.

Mesmo quando a relação $(\theta_n^* - \theta_1^*)/\delta$ se encontra na escala dos bilhões, a magnitude de $n$ sequer extrapolará a escala das dezenas. A expressão para o número de iterações sob este refinamento é $n = \lceil [(\ln 2)^{-1} \ln[(\theta_n^* - \theta_1^*)/\delta)] \rceil$. Por exemplo, para $(\theta_n^* - \theta_1^*)/\delta = 1000.000.000$, teremos $n = 30$. Para $n = 30$, e usando a regra para escolha dos parâmetros de ajuste do método sequencial descritas na Seção 3.3.1, a Tabela 4 apresenta valores de $M_0$ que garantem uma cota superior igual a $\gamma_{(max,dois)} = \gamma + \epsilon$, com coeficientes de confiança $\gamma = 0,8$**,** $0,9$**,** $0,95$**,** $0,99$, e $\epsilon = 0,001$**,** $0,005$**,** $0,01$**,** $0,02$**,** $0,03$**,** $0,04$**,** $0,05$. Os dados desta tabela para as colunas correspondentes a $\gamma = 0,9$**,** $0,95$**,** $0,99$ foram retirados do artigo de [18], e as demais colunas foram especialmente calculadas para este artigo. Naturalmente, ela também é válida para $n < 30$.

Tabela 2: Valores mínimos de $M$ ($M_0$) para uma cota superior de $(\gamma + \epsilon) < 1$ para a real probabilidade de cobertura bilateral, e para uma cota inferior (coeficiente de confiança) de $\gamma$. As céluas com 'na' representam os cenários que não fazem sentido prático.

| | $\gamma$ | | | | |
|---|---|---|---|---|---|
| $\epsilon$ | 0.8 | 0.9 | 0.95 | 0.98 | 0.99 |
| 0.05 | 13,520 | 6,300 | na | na | na |
| 0.04 | 22,380 | 10,480 | 4,220 | na | na |
| 0.03 | 42,460 | 20,760 | 8,860 | na | na |
| 0.02 | 103,880 | 52,520 | 24,000 | na | na |
| 0.01 | 466,700 | 241,660 | 117,800 | 40,960 | na |
| 0.005 | 2,056,520 | 1,079,040 | 541,000 | 205,080 | 88,620 |
| 0.001 | 61,776,920 | 32,699,440 | 16,741,940 | 6,715,500 | 3,288,000 |

# 4 Método Exato para Estimação Intervalar do Tamanho Populacional

Apesar de não mencionado por [9], na realidade é possível obter o intervalo de confiança exato para $N$ no problema de captura-recaptura baseado na estatística $\hat{N}^*$. Aproveitamos que a contextualização do problema já está dada para oferecer, como contribuição adicional deste artigo, a descrição sobre como obter o intervalo de confiança exato para $N$ com base na estatística $\hat{N}^*$. Seja $\hat{N}$ uma estimativa obtida pela aplicação da estatística $\hat{N}^*$ a uma dada amostra observada. Assim, o limite inferior do intervalo de confiança exato é:

$$\hat{N}_L = \max\{n^* : Pr(\hat{N}^* \geq \hat{N}|N = n^*) \leq \alpha\}. \tag{38}$$

Similarmente, o limite superior será da forma:

$$\hat{N}_S = \min\{n^* : Pr(\hat{N}^* \leq \hat{N}|N = n^*) \leq \alpha\}. \tag{39}$$

Ou seja, o intervalo de $100 \times (1 - 2\alpha)\%$ de confiança para $N$ é:

$$IC_{1-2\alpha}(N) = (\hat{N}_L; \hat{N}_S) \tag{40}$$

Como a distribuição de $\hat{N}^*$ é discreta, sabemos que o intervalo em (39) é conservador. A título de ilustração, vamos verificar a verdadeira probabilidade de cobertura deste sob a parametrização $N = 1000$, $n_1 = 100$, $n_2 = 100$ e $\alpha = 0,025$. Temos que

$$Pr(\hat{N}_L \leq N \leq \hat{N}_S) = 1 - Pr(\hat{N}_L > N \cup \hat{N}_S < N) = 1 - Pr(\hat{N}_L > N) - Pr(\hat{N}_S < N). \tag{41}$$

Vamos investigar os valores de $m$ que implicam $\hat{N}_L > N$. Para tanto, oferecemos a Tabela 3 que oferecem a estimativa pontual, e o limite inferior observado, para os seis possíveis menores valores de $m$. Para obtermos a terceira coluna da Tabela 2, que contém os limites inferiores, fazemos $P(\hat{N}^* > \hat{N}|n^*)$ para diferentes valores de $n^*$ que satisfaça $P(\hat{N}^* > \hat{N}|n^* \leq \alpha)$. Por exemplo, para $m = 0$, temos $\hat{N} = 10200$, o que leva à solução $\hat{N}_L = 2810$. Veja que para $n^*$ fixo, temos:

$$\begin{aligned} Pr(\hat{N}^* \geq \hat{N}|n^*) &= Pr\left(\frac{(n_1+1)(n_2+1)}{(m+1)} - 1 \geq \hat{N}|n^*\right) \\ &= Pr((n_1+1)(n_2+1) \geq (\hat{N}+1)(m+1)|n^*) \\ &= \sum_{x=0}^{c} \frac{\left(\begin{array}{c} n_1 \\ x \end{array}\right)\left(\begin{array}{c} n^* - n_1 \\ n_2 - x \end{array}\right)}{\left(\begin{array}{c} n^* \\ n_2 \end{array}\right)}, \end{aligned} \tag{42}$$

Tabela 3: Limites inferiores do intervalo de confiança exato para $N$ com base no estimador $\hat{N}^*$ para os seis menores possíveis valores de $m$ no caso de $n_1 = n_1 = 100$.

| $m$ | $\hat{N}$ | $\hat{N}_L$ |
|---|---|---|
| 0 | 10200 | 2810 |
| 1 | 5099 | 1876 |
| 2 | 3399 | 1456 |
| 3 | 2549 | 1206 |
| 4 | 2039 | 1038 |
| 5 | 1699 | 914 |

Tabela 4: Limites superiores do intervalo de confiança exato para $N$ com base no estimador $\hat{N}^*$ para dois valores intermediários de $m$ e para os dois maiores possíveis valores de $m$ no caso de $n_1 = n_1 = 100$.

| $m$ | $\hat{N}$ | $\hat{N}_S$ |
|---|---|---|
| 100 | 100 | 100 |
| 99 | 101 | 101 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| 17 | 565 | 947 |
| 16 | 599 | 1028 |

onde $c = (n_1+1)(n_2+1)/(\hat{N}+1) - 1$. Para $m = 0$, temos $\hat{N} = 10200$, e por avaliação numérica de uma sequência crescente de valores para $n^*$, encontra-se que o maior valor que faz com que a probabilidade em (42) seja menor que $0,025$ é $n^* = 2810$, o qual fornece:

$$Pr(\hat{N}^* \leq 10200 | N = 2810) \approx 0,02497. \tag{43}$$

Veja que esta é uma situação em que $\hat{N}_L > N$ , ou seja, se $m = 0$, o intervalo não cobrira $N$ inferiormente. Se $m = 1$, teremos $\hat{N}_L = 1876$. Na realidade, $\hat{N}_L$ é uma função decrescente com respeito a $m$. Para $m \geq 5$, teremos sempre um limite inferior cobrindo $N = 1000$. Assim, a probabilidade de não cobertura inferior é $Pr(m \leq 4 | N = 1000) \approx 0,0188$. Vamos agora avaliar o limite superior, para $n^*$ fixo, temos:

$$\begin{aligned} Pr(\hat{N}^* \leq \hat{N} | n^*) &= Pr\left( m \geq \frac{(n_1+1)(n_2+2)}{(\hat{N}+1)} - 1 | n^* \right) \\ &= 1 - Pr\left( m \leq \frac{(n_1+1)(n_2+2)}{(\hat{N}+1)} - 2 | n^* \right) \end{aligned} \tag{44}$$

Para $M = 100$, teríamos $\hat{N}_S = 100$, o que não cobriria $N$ superiormente. Como $\hat{N}_S$ é decrescente com $m$, teremos uma região de cobertura para baixos valores de $m$.

Assim, $Pr(\text{cobertura} | N = 1000) \approx 1 - 0,0188 - 0,015 \approx 0,965$. Naturalmente, a probabilidade de cobertura varia de acordo com o verdadeiro $N$, mas será sempre conservador tendo em vista a natureza discreta de $\hat{N}^*$.

# 5 Percentil Versus Sequencial

Já vimos, Seção 4, que é possível construir o intervalo de confiança exato para o problema de captura-recaptura, e que portanto não é necessário usar métodos Monte Carlo. Mas a possibili-

dade do cálculo exato é útil para que se possa fazer a comparação de diferentes métodos Monte Carlo. Nesta seção usaremos esta conveniência para comparar as performances dos métodos percentil e sequential. O método aleatorizado não foi incluído nesta comparação pois, além da trivial escolha pela inversa da função distribuição de $\hat{N}^*$, os autores deste artigo não conseguiram encontrar uma função $g$ requerida para a utilização do método tal como descrito na Seção 3.2.

As expressões para o cálculo exato das probabilidades de cobertura associadas ao método sequencial já estão dadas na seção 3.3.1. Precisaremos deduzir a expressão analítica da probabilidade de cobertura do método percentil. Para tanto, defina $P = Pr(\hat{N}^* \leq N|\hat{N})$, onde $\hat{N}$ é o valor de $\hat{N}^*$ observado com uma amostra particular. Sejam $p_1 < p_2 < \cdots < p_{k_2-k_1+1}$ os possíveis valores de $p$, com $k_1 = \lfloor (n_1+1)(n_2+1)/n_2 + 1 - 1 \rfloor = n_1$ e $k_2 = (n_1+1)(n_2+1) - 1$. Veja que $p$ é decrescente com respeito a $\hat{N}$. Assim:

$$p_i = Pr(\hat{N}^* \leq N|\hat{N} = k_{2-i+1}), i = 1, 2, \ldots, k_2 - k_1 + 1. \tag{45}$$

A probabilidade de cobertura inferior do método percentil fica assim explicitada:

$$\Rightarrow Pr(\hat{N}_L^* \leq N|N) = \sum_{i=1}^{k_2-k_1+1} Pr(S_i \geq h) Pr(P = p_i), \tag{46}$$

onde $S_i$ é o número de valores simulados menores que ou iguais a $N$, ou seja, $S_i \sim Bin(M-1, p_i)$, e $h = \lfloor \alpha M \rfloor$. Para favorecer o cálculo em (46), precisamos explicitar $p_i$ e $Pr(P = p_i)$. Começando por $p_i$, temos:

$$\begin{aligned} p_i &= Pr\left( \left[ \frac{(n_1+1)(n_2+1)}{m+1} \right] - 1 \leq N|\hat{N}_i \right) \\ &= Pr(n_1+1)(n_2+1) \leq (m+1)(N-1)|\hat{N}_i \\ &= Pr\left( \left[ \frac{(n_1+1)(n_2+1)}{N+1} \right] - 1 \leq m|\hat{N}_i \right) \\ &= 1 - \sum_{l=0}^{k_3} \frac{\left( \begin{array}{c} n_1 \\ l \end{array} \right) \left( \begin{array}{c} \hat{N}_i - n_1 \\ n_2 - l \end{array} \right)}{\left( \begin{array}{c} \hat{N}_i \\ n_2 \end{array} \right)}, \end{aligned} \tag{47}$$

onde $k_3 = \lfloor (n_1+1)(n_2+1)/(N+1) \rfloor$ e $\hat{N}_i = k_{2-i-1}$. Vamos agora desenvolver o termo $Pr(P = p_i)$:

$$\begin{aligned} Pr(P = p_i) &= Pr(\hat{N}^* = \hat{N}_i^*) = Pr(\hat{N}^* = (n_1+1)(n_2+1) - i|N) \\ &= Pr\left( \left[ \frac{(n_1+1)(n_2+1)}{N+1} \right] - 1 = (n_1+1)(n_2+1) - i|N \right) \\ &= Pr(m = i-1|N) = \frac{\left( \begin{array}{c} n_1 \\ i-1 \end{array} \right) \left( \begin{array}{c} N - n_1 \\ n_{2-i} + 1 \end{array} \right)}{\left( \begin{array}{c} N \\ n_2 \end{array} \right)}. \end{aligned} \tag{48}$$

Por analogia, defina $Q = Pr(\hat{N}^* \geq N|\hat{N})$, e sejam $q_1 < q_2 < \cdots < q_{k_2-k_{1+1}}$ os possíveis valores de $q$. Veja que Q é crescente com $\hat{N}$. Para $i$ fixo, temos:

$$q_i = Pr(\hat{N}^* \geq N|\hat{N} = k_{1-i+1}), i = 1, 2, \ldots, (n_2+1). \tag{49}$$

A probabilidade definida em (49) pode ser usada no cálculo da probabilidade de cobertura superior do método percentil da seguinte forma:

$$Pr(\hat{N}_s^* \geq N|N) = \sum_{i=1}^{k_2-k_1+1} Pr(G_i \geq h) Pr(Q = q_i), \tag{50}$$

Tabela 5: Verdadeiras Probabilidades de Cobertura para os Métodos Exato, Percentil e Sequential ($\delta = 1$) para o Problema de Captura-recaptura.

| $\gamma(100)\%$ | $N$ | $n_1$ | Exato ($\gamma_e$) | Percentil | Sequencial |
|---|---|---|---|---|---|
| 90% | 500 | 100 | 0.947796 | 0.844218 | 0.947838 |
| ($m = 241,660$) | | 250 | 0.925313 | 0.890391 | 0.925355 |
| | 1000 | 100 | 0.949170 | 0.806956 | 0.949213 |
| | | 250 | 0.922606 | 0.860378 | 0.922648 |
| 95% | 500 | 100 | 0.973209 | 0.928896 | 0.973236 |
| ($m = 117,800$) | | 250 | 0.957259 | 0.928122 | 0.957286 |
| | 1000 | 100 | 0.978961 | 0.882125 | 0.978988 |
| | | 250 | 0.964410 | 0.926206 | 0.964436 |
| 98% | 500 | 100 | 0.987046 | 0.960240 | 0.987048 |
| ($m = 40,960$) | | 250 | 0.987462 | 0.967136 | 0.987464 |
| | 1000 | 100 | 0.991922 | 0.966692 | 0.991924 |
| | | 250 | 0.985169 | 0.968002 | 0.985171 |

onde $G_i \sim Bin(M-1, q_i)$, e:

$$\begin{aligned} q_i &= Pr(\hat{N}^* \geq N|\hat{N} \\ &= = Pr\left[m \geq \frac{(n_1+1)(n_2+1)}{N+1} - 1|\hat{M}_i\right] \\ &= \sum_{l=0}^{k_3} \frac{\left(\begin{array}{c} n_1 \\ l \end{array}\right)\left(\begin{array}{c} \hat{M}_i - n_1 \\ n_2 - 1 \end{array}\right)}{\left(\begin{array}{c} \hat{M}_i \\ n_2 \end{array}\right)}, \end{aligned} \tag{51}$$

com $\hat{M}_i = k_{1-i} + 1$. Para completar a descrição dos intens necessários para o cálculo expresso em (51), temos: $Pr(Q = q_i) = Pr(m = n_2 - i + 1|N) = \left(\begin{array}{c} n_1 \\ n_2 - i + 1 \end{array}\right)\left(\begin{array}{c} N - n_1 \\ i - 1 \end{array}\right)$. No caso de empate, ou seja, quando $\hat{N}_L = \hat{N}_S$, ambos os limites, $\hat{N}_L$ e $\hat{N}_S$, falharão em captar $N$, portanto, esta possibilidade deve ser considerada no cálculo da exata probabilidade de cobertura do método percentil. Desta forma, a probabilidade de cobertura global do método percentil, para o problema de captura-recaptura, pode ser obtida com a seguinte expressão:

$$Pr(\hat{N}_L \leq N \leq \hat{N}_S) = 1 - Pr(\hat{N}_L > N) + Pr(\hat{N}_S < N) - Pr(\hat{N}_L = \hat{N}_S), \tag{52}$$

onde $Pr(\hat{N}_L = \hat{N}_S) = \sum_{l=k_1}^{k_2}[Pr(\hat{N}^*)^m] = \sum_{l=k_1}^{k_2}\left[\frac{\left(\begin{array}{c} n_1 \\ l \end{array}\right)\left(\begin{array}{c} N - N_1 \\ n_2 - 1 \end{array}\right)}{\left(\begin{array}{c} N \\ n_2 \end{array}\right)}\right]$.

A Tabela 5 apresenta a exata probabilidade de cobertura na estimação intervalar do tamanho populacional via estatística $\hat{N}^*$ para os três métodos, a saber: exato, percentil, e sequencial. Foram usados dois valores distintos, 100 e 250, para os parâmetros $n_1 = n_2$. Foram usados também dois cenário para o verdadeiro $N$, fixados em 500 e 1000. Lembremos que a distribuição de $\hat{N}^*$ é discreta, portanto nenhum dos três métodos apresentará uma probabilidade de cobertura de exatamente $\gamma$ para nenhum valor verdadeiro de $N$, mas o que se espera é que eles sejam, no mínimo, conservadores, pois isso indica o atendimento ao coeficiente de confiança. Já vimos que no método sequencial é possível escolher $M_0$ tal que garanta uma cota superior para a probabilidade de cobertura. Para tanto, todos os cenários desta tabela foram construídos sob

a restrição de $\gamma_e + 0.01$, onde $\gamma_e$ é a probabilidade de cobertura do método exato. Esta cota superior, naturamente, é garantida para o procedimento sequencial mas não para o percentil. Para o procedimento sequencial, a escolha da parametrização para $M$ foi feita por se tomar os valores $M_0$ de acordo com a Tabela 2. Para uma comparação justa, os mesmos valores para $M$ foram usados nos cálculos exatos para o método percentil, por sua vez possíveis graças à expressão (52). Vemos que o método percentil tende a funcionar bem para valores de $n_1$ próximos de $N$. No entanto, em todos os cenários a probabilidade de cobertura (segunda coluna da Tabela 5) é liberal com respeito a $\gamma$, ou seja, a probabilidade de cobertura é menor que o coeficiente de confiança desejado. Isto ocorre, por exemplo, para $\gamma = 0.9$, $N = 1000$ e $n = 100$, caso em que a probabilidade de cobertura verdadeira do método percentil é de aproximadamente 0.807, bem menor que o 0.9 desejado. Sobre o método sequencial, e nesta tabela em que fixamos $\epsilon = 0.01$, mais do que corroborando a demonstração analítica da Seção 3.3.1 de que a probabilidade de cobertura, digamos $\gamma_s$, sempre atenderá $\gamma \leq \gamma_s \leq (\gamma_e + 0.01)$, observa-se que a cota é de fato atendida. Para além, a cota é extremamente elevada frente aos valores verdadeiros de cobertura, pois, em todos os cenários, a probabilidade de cobertura do método sequencial (terceira coluna da Tabela 5) coincide com a probabilidade de cobertura exata (primeira coluna da Tabela 5) em pelo menos três casas decimais.

# 6 Aplicação do Método Sequencial para Estimação do Risco Relativo em Conglomerados Espaciais

Nesta seção, descreveremos a aplicação do método sequencial apresentada no artigo de [18]. A descrição desta aplicação é motivadora por ser tratar de um problema de extrema relavância prática, pois envolve métodos de análise de estatística para garantia da segurança da saúde pública. Este exemplo de aplicação envolve o desafiador objetivo de se fazer inferências sobre o risco relativo associado a conglomerados espaciais.

Segundo [11], um conglomerado espacial é um conjunto de áreas vizinhas de uma determinada região (mapa), as quais apresentam um elevado risco para a ocorrência de um determinado evento. Suponha que um determinado mapa seja composto por $K$ áreas, as quais estão rotuladas com os algarismos de 1 a $K$. Um método para detecção e localização de conglomerados, bastante difundido e de excelente performance, é o chamado teste Scan, proposto por [15]. O teste Scan consiste na realização de um espécia de 'varredura' de um grande número de sub-regiões do mapa, cada uma destas contendo uma ou mais áreas, e, a cada sub-conjunto selecionado, usa-se a estatística da razão de verossimilhanças, aplicada aos número de eventos dentro e fora de tal sub-conjunto, para medir o quão verossímil a conglomerado sta seria. As sub-áreas são criadas por círculos, cada um com um determinado raio e centrado no centroide de cada área. Para um dado raio e centroide, seja $z$ o conjunto de todas as áreas com centroides dentro de um certo círculo. É importante destacar que cada um dos possíveis círculos contém uma parcela do total da população e uma porção do número total de eventos observados em todo o mapa. O raio máximo a ser utilizado em cada centroide é geralmente definido como um percentual do da população total do mapa, pois não faz sentido procurar conglomerados que representem a maioria da populacão. Para a aplicação mostrada por [18] o percentual utilizado foi de 50%, que é a escolha mais corriqueira.

Seja $C_i \sim Pois(R_i \lambda p_i)$, a variável aleatória que conta o número de casos (eventos) dentro do círculo $z_i$, onde $p_i$ é a população de $z_i$, $\lambda$ é o valor esperado para o número de casos dentro de $z_i$ sob a hipótese de não existência de conglomerados, e $R_i$ é o risco relativo associado à ocorrência de um caso dentro do círculo com respeito à possibilidade de ocorrência deste caso fora do círculo. Assim, para um total de $B$ círculos, as hipóteses a serem testadas são $H_0 : R_i = 1$ para todo $i = 1, \ldots, B$, contra $H_a : R_i > 1$ para algun $i$. A estatística Scan é então definida como o máximo sob todos os círculos, denotado aqui por $\Lambda$, da log razão de verossimilhanças, que para

o círculo $k$ é dada por:

$$\begin{aligned} LLR_k &= \ln\left[\left(\frac{C_k}{p_k}\right)^{C_i}\left(\frac{p_k - C_k}{p_k}\right)^{p_k - C_k}\right] + \\ &+ \ln\left[\left(\frac{C - C_k}{p - p_k}\right)^{C - C_k}\left(\frac{(p - p_i) - (C - C_k)}{p - p_k}\right)^{(p-p_k)-(C-C_k)}\right] - \\ &- \ln\left[\frac{C^C(p - C)^{p-C}}{p^p}\right], \end{aligned} \tag{53}$$

se $\frac{C_i}{p_i} > \frac{C-C_i}{p-p_i}$, e é igual a zero caso contrário.

O teste de hipótese exato não é viável pelo fato da distribuição de $\Lambda$ não ser ainda conhecida. O recurso alternativo amplamente usado para obtenção do valor-p no teste Scan é o teste Monte Carlo. Dado o número total de casos, digamos $C$, a distribuição conjunta dos eventos no mapa é uma multinominal. Assim, a probabilidade de se observar um evento na $k-$ésima área é igual a $(R_k p_k)/\sum_{l=1}^{K}(R_l p_l)$. Portanto, amostras de $\Lambda$, sob $H_0$, podem ser geradas facilmente por se simular eventos no mapa provindos de uma distribuição multinominal com $R = 1$. Vamos denotar o círculo de maior $LLR$ por $z_{\hat{k}}$ , isto é, $\Lambda = LLR_{\hat{k}}$.

Para aplicação do método sequencial após o teste Scan, pode-se simplesmente operar o algoritmo sequencial por se gerar casos de uma distribuição multinomial dentro de $z_{\hat{k}}$, e com isto encontrar o intervalo de confiança para $R$ associado ao circulo $z_{\hat{k}}$. É importante destacar que o parâmetro genérico $\theta$ usando ao longo do artigo é o parâmetro de risco relativo $R$.

A validade do método sequencial depende da veracidade da suposição de que a distribuição de $\Lambda$ é monótona em $R$. Primeiramente, notemos que a distribuição de $\Lambda$ pode ser escrita da seguinte forma:

$$Pr(\Lambda \leq \lambda | R, z_{\hat{k}}) = Pr(\cap_{i=1}^{B} C_i \leq c_i | R, z_{\hat{k}}), \tag{54}$$

onde $c_i$ é o valor máximo para o número de eventos no i-ésimo círculo tal que o valor de $\Lambda$ seja menor ou igual que $\lambda$. Como $C_{\hat{k}}$ segue uma distribuição Poisson de parâmetro $E(C_{\hat{k}})$ monótonamente decrescente em $R$, conclui-se que a distribuição em (54) é descendente com $R$, ou seja, o método sequencial poderá ser usado corretamente para estimação intervalar do risco relativo do conglomerado encontrado

## 6.1 Descrição dos Dados

Os dados usados neste exemplo são reais, e tratam-se dos casos de câncer de cérebro observados no Novo México, Estados Unidos, no período de 1973 a 1991. Estes dados estão disponíveis gratuitamente em 'http;//www.satscan.org/datasets'. As informações desta base de dados são longitudinais, portanto seria possível explorar a existência não apenas de conglomerados espaciais, mas também temporais. [18] optou por concatenar as informações de todo o período de observações em uma só informação para cada área. Isto simplifica o entendimento desta aplicação que, [18] buscou reforçar, não visou analisar de fato os dados, mas apenas ilustrar a aplicabilidade do método sequencial. O total de casos de câncer observados no mapa, no período descrito, foi de 1175, enquanto que o total da população exposta no mapa no ano de 1991 era de 1548640. A localização do centroide geográfico (latitude e longitude) de cada município também está disponível no mesmo site. Uma descrição detalhada da analise final e dedicada a esta base de dados foi oferecida por [16].

## 6.2 Definindo os parâmetros de Ajuste

A menor população do ano de 1991 foi de 987, e foi observada na cidade de Harding. Mesmo para esta pequena população, se o risco relativo associado fosse da ordem de 1000.000, ainda

assim a probabilidade de observarmos menos de 987 casos nesta área seria de aproximadamente zero, portanto, não há nenhuma razão para se definir o valor de $\theta_{max}$ maior do que 1000000 porque, se tal risco fosse uma realidade neste município, então não existiria nenhuma necessidade de se usar uma inferência estatística uma vez que bastaria olhar para os dados brutos para identificar que aquela área se trataria de um conglomerado. Com alta probabilidade ela conteria, se não todos, quase todos o casos do mapa. Qualquer outro sub-conjunto de áreas terá população maior que 987, e portanto o argumento anterior seria ainda mais válido. Assim, será suficiente definir $\theta_{min} = 0$, e $\theta_max = 1000000$.

Uma precisão de três casas decimais foi usada para os limites do intervalo ($\delta = 0,001$). Com isto, $n = \lceil \ln(1000000/0,001)/\ln(2) \rceil$. Para um coeficiente de confiança bilateral de $\gamma = 0,9$, temos $\alpha = (1-\gamma)/2 = 0,05$. Vamos também estabecer uma cota superior para a probabilidade de cobertura bilateral de $\gamma_{(max,dois)} = 0,93$, implicando $\gamma_{max} = 0,93 + (1-0,93)/2 = 0,965$. Aplicando a regra sugerida na Seção 3.3.1, a parametrização de ajuste estará completa fazendo-se $\alpha_{in} = 1-0,965 = 0,035$, $\alpha_s = (2\times 0.05+0.035)/3 \approx 0.0283$ e $\hat{\alpha} = (0.0283+0.035)/2 = 0.03165$. Conforme a Tabela 4, o valor para o número de simulações Monte Carlo a ser usado em cada iteração será de $20,800$.

### 6.3 Resultados da Análise de Dados

Após aplicada ao banco de dados originais, o valor observado da estatística Scan foi igual a 5,4587. O conglomerado indentificado é formado pelos seguintes municípios: Chaves, Colfax, Curry, Debaca, Guadelupe, Harding, Mora, Quay, Roosevelt, San Miguel, and Union. Na primeira iteração Monte Carlo de 20799 simulações para $\Lambda$ sob $H_0$, observou-se 914 valores valores maiores que ou iguais a 5,4587, o que resultado em um valor-p de $(1+914)/20800 = 0,04399$. Portanto, a hipótese nula é rejeitada a um nível de 5%. Com isto, o valor de $\theta_{min}$ é fixado inicialmente no valor 1. Para operar o método da bissecção, o valor de $R$ para a segunda iteração Monte Carlo foi fixado em $\theta_1^* = (1+10000000)/2 = 5000000,5$, que foi usado para gerar clusters artificiais no conjunto de municípios identificados acima. Após atualiar os valores de $R_j^*$ para $i = 2,\ldots,30$, sequencialmente, chegou-se a um limite superior observado para o intervalo de confiança igual a $R_u = \theta_u^* = 1,475$. Para encontrarmos o limiter inferior, iniciou-se com $\theta_{max} = 1,475$, e $\theta_{min} = 1$. Após 30 iterações de Monte Carlo, o limite inferior encontrado foi de $R_l = \theta_l^* = 1,026$, ou seja, o intervalo de confiança observado para o risco relativo no interior do cluster observado é: $[1,026; 1,475]$.

## Referências


[1] Base de dados de incidência de câncer no cérebro, *http;//www.satscan.org/datasets*. Acessado em 2 de novembro de 2014.

[2] BERNARD, G. A. Discussion of Professor Bartletts papers, **J. R. Statist. Soc.** . v. 25(B), 1963.

[3] BIRNBAUM, Z. W. Computers and unconventional test-statistics, *In: F. Proschan and R. J. Serfling*, **Realiability and Biometry**, pp. 441-458, 1974.

[4] BOLFARINE, H.; SANDOVAL, M. C. Introdução a Inferência Estatística, Coleção Matemática Aplicada, SBM, 2010.

[5] BOLVIKEN, E.; SKOVLUND, E. Confidence Intervals from Monte Carlo Test, **Journal of the American Statistical Association**. v. 91, pp. 1071-1078, 1996.

[6] BUCKLAND, S. T., A modified analysis of the Jolly-Saber capture-recapture model. **Biometrics**. v. 36, pp. 419-435, 1980.

[7] BUCKLAND, S. T. A mark-recapture survival analysis. **Journal of the Animal Ecology**. v. 51, pp. 833-847, 1982.

[8] BUCKLAND, S. T. Monte Carlo methods for confidence interval estimation using the bootstrap techinique. **BIAS**. v. 10, pp. 194-212, 1983.

[9] BUCKLAND, S. T. Monte Carlo Confidencial Interval. **Biometrics**. v. 40, pp. 811-817, 1984.

[10] CASELLA, G.; BERGER, R. L. Inferência Estatística, Cengage Learning, 2010.

[11] CRESSIE, N. Statistics for Spatial Data. John Willey and Sons, 1993.

[12] DWASS, M. Modified randomization test for nonparametric hypotheses, **Annal of Mathematical Statistics**. v. 28, pp. 181-187, 1957.

[13] EFROM, B. Bootstrap methods: another look at jackknife. **Annal of Statistics**. v. 7, pp. 1-26, 1979.

[14] HOPE, A. A simplified Monte Carlo Significance Test Procedure, **Journal of the Royal Statistical Society**. v. 30(B), pp. 582-598, 1968.

[15] KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: Detection and Inference, **Statistic and Medicine**, v. 14, pp. 799-810, 1995.

[16] KULLDORFF, M.;ATHAS, W. F.; FEUER, E. J.; MILLER, B. A.; KEY, C R. Evaluating cluster alarms: A space-time scan statics and brain cancer in los Alamos, **American Journal of Public Health**, v. 88, pp. 1377-1380, 1998.

[17] LEHMANN, E. L. Testing Statistical Hypothesis, 2nd ed., John Wiley and Sons, New York, 1986.

[18] SILVA, I. R. A Simplified Method for Finding Confidence Intervals Through Sequential Monte Carlo Simulation. Publisher: WSEAS Press, *Recent Advances in Applied Mathematics, Modelling and Simulation - Proceedings of the 8th Conference on Applied Mathematics, Simulation, Modelling (ASM/14)*. WSEAS Press, Athens, 2014, Greece. Series 34. ISBN: 978-960-474-398-8.