3 Método de Reamostragem

No contexto de Inferência Estatística, fazemos a suposição de que uma amostra aleatória, denotada por $\mathbf{Y} = (Y_1, \ldots, Y_n)$, é proveniente de uma população com função de probabilidade ou densidade f e função de distribuição acumulada F que dependem de um parâmetro desconhecido $\theta \in \Theta$ (também podemos supor θ como um vetor de parâmetros). No entanto, em algumas situações, essa suposição nem sempre é satisfeita.

Nesses casos, uma alternativa consiste em fazer uma reamostragem, isto é, a partir dos valores observados da amostra, retirar repetidas amostras na qual a inferência será baseada. Dessa forma, os métodos de reamostragem são procedimento, em geral, não-paramétricos para extrair as amostras do conjunto de dados original com o intuito de estimar a precisão, bem como a distribuição amostral de uma estatística de interesse.

Na literatura existem vários tipos de reamostragem, a seguir apresentaremos os métodos de jackknife e *bootstrap* que podem ser utilizados para determinar a precisão de estimadores em populações finitas.

3.1 Método de jackknife

O método jackknife foi introduzido por Quenouille (1949), cujo o objetivo era melhorar uma estimativa, corrigindo seu viés. Posteriormente, Tukey (1958) popularizou o método utilizando-a para estimar erros padrão de uma estimativa. O método jackknife baseia-se na remoção de um elemento da amostra, recalculando a estimativa do parâmetro de interesse a partir dos valores restantes.

Dessa forma, vamos supor que $\mathbf{y} = (y_1, \ldots, y_n)$ seja o valor observado de uma amostra aleatória de tamanho n e $\hat{\theta} = t(\mathbf{y})$ a estimativa do parâmetro θ, em que $t(\mathbf{y})$ é o valor observado da estatística. Então, a i-ésima amostra jackknife é denotada por

$$\mathbf{y}_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n), \quad i = 1, \ldots, n,$$

que corresponde a um conjunto de dados sem o i-ésimo elemento da amostra original. Como consequência, o número total de réplicas jackknife possíveis é igual ao tamanho da amostra original. Assim, para cada amostra jackknife,

$$\hat{\theta}_{(i)} = t(\mathbf{y}_{(i)})$$

representa a i-ésima réplica jackknife de $\hat{\theta}$.

Portanto, a estimativa jackknife do erro padrão é dada por

$$\widehat{\text{ep}}_{jack} = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left[\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right]^2}, \tag{1}$$

em que $\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$, também conhecida como estimativa jackknife. Enquanto que a estimativa jackknife do viés é definida por

$$\widehat{\text{vies}}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

Por outro lado, uma outra forma de pensar sobre o método jackknife é em termos de pseudo-valores. Dessa forma, o i-ésimo pseudo-valor, definido por Tukey (1958), é dado por

$$\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}.$$

Logo, com base nos pseudo-valores, a estimativa jackknife do erro padrão (1) pode reescrita como

$$\widehat{\text{ep}}_{jack} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_i - \tilde{\theta}\right)^2},$$

em que $\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\theta}_i$.

3.2 Método *bootstrap*

Método *bootstrap* foi proposto por Efron (1979) e utilizado para obter alguma informação sobre a distribuição de probabilidade de uma estatística, como por exemplo o seu viés, o seu desvio padrão ou ainda a forma como se pode calcular os limites de confiança para um determinado parâmetro. Podemos considerar o método *bootstrap* uma estratégia mais abrangente que o método jackknife por permitir um maior número de réplicas, uma vez que é baseada na geração de uma "nova amostra" de mesmo tamanho da amostra original.

O elemento básico do método *bootstrap* é a função de distribuição acumulada empírica. Então, dada uma amostra com valores observados $\mathbf{y} = (y_1, \ldots, y_n)$, proveniente de uma população com função de probabilidade ou densidade f e função de distribuição acumulada F que dependem de um parâmetro desconhecido $\theta \in \Theta$, a função distribuição acumulada empírica é definida como

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I\{y_i \leq y\}, \tag{2}$$

em que $I\{y_i \leq y\}$ é a função indicadora do evento $\{y_i \leq y\}$ para $y_{\min} \leq y \leq y_{\max}$. Além disso, podemos observar que $F_n(y)$ corresponde à proporção de valores menores ou iguais a y. Assim, de forma alternativa, a função em (2) pode ser reescrita como

$$F_n(y) = \begin{cases} 0 & y < y_{(1)} \\ \frac{k}{n} & y_{(k)} \leq y < y_{(k+1)} \\ 1 & y_{(n)} < y \end{cases} \tag{3}$$

em que $y_{(i)}$ para $i = 1, \ldots, n$ são as estatísticas de ordem e $k = \sum_{i=1}^{n} I\{y_i \leq y\}$. Em outras palavras, a função de distribuição acumulada empírica é uma distribuição discreta que atribui peso igual a cada ponto amostral, ou seja, atribui probabilidade $1/n$ a cada uma das n observações originais. Se o tamanho da amostra for suficientemente grande, a lei dos grandes números nos diz que F_n deve se aproximar muito bem de F, ou seja, F_n deve ser um estimador não viciado de F.

Nesse contexto, de acordo com Chernick e LaBudde (2011), o "princípio *bootstrap*" baseia-se na ideia de que F_n desempenha o papel de F e por sua vez F_n^*, a função de distribuição *bootstrap*, desempenha o papel de F_n no processo de reamostragem. Isto é, dada uma amostra aleatória

$$\mathbf{Y} = (Y_1, \ldots, Y_n) \quad \text{de uma população} \quad Y \sim F \tag{4}$$

e θ um parâmetro de interesse, então o estimador de θ é definido como

$$\hat{\theta} = T(F_n).$$

Analogamente, dada uma amostra *bootstrap* denotada por

$$\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*) \quad \text{de uma "população"} \quad Y \sim F_n, \tag{5}$$

podemos definir o estimador de θ como sendo

$$\hat{\theta} = T(F_n^*).$$

O método *bootstrap* pode ser classificado como:

► Não-paramétrico

Denominamos o procedimento de coleta de amostras com reposição a partir da função de distribuição acumulada empírica F_n como *bootstrap* não-paramétrico. O método é não-paramétrico, pois utilizamos F_n definida em (3), uma vez que não conhecemos a distribuição dos dados F.

Uma amostra *bootstrap* (5) é uma amostra de tamanho n retirada com reposição da amostra original (4), por essa razão, alguns elementos podem aparecer zeros vezes, alguns aparecendo uma vez, outros aparecendo duas vezes e assim por diante. Utilizamos a notação $*$ para indicar que $\mathbf{y}^*$ não é o conjunto de dados real $\mathbf{y}$.

Portanto, dada uma amostra *bootstrap*, a réplica *bootstrap* de θ é definida como

$$\hat{\theta}^* = T(\mathbf{y}^*). \tag{6}$$

Agora, para calcular a incerteza em torno de $\hat{\theta}$, retiramos B amostras *bootstrap* de F_n e calculamos $\hat{\theta}^* = T(y^*)$ para cada amostra. Um histograma de $\hat{\theta}^*$ ilustra a incerteza em torno de $\hat{\theta}$. Também podemos resumir a incerteza em torno de $\hat{\theta}$ por meio do cálculo do erro padrão das amostras *bootstrap* definida como

$$\text{ep}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{j=1}^{B}(\hat{\theta}_j^* - \bar{\theta}^*)^2}, \tag{7}$$

em que $\bar{\theta}^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$ é a estimativa *bootstrap*.

Algoritmo:

P1) Escolha B;

P2) Extraia $\mathbf{y}^{*1}, \ldots, \mathbf{y}^{*r}$ amostras *bootstrap* independentes de F_n, em que $\mathbf{y}^{*j} = (y_1^{*j}, \ldots, y_n^{*j})$ é um vetor de n observações que contêm a j-ésima amostra *bootstrap* para $j = 1, \ldots, B$;

P3) Avalie a réplica *bootstrap*, $\hat{\theta}_j^* = T(y^{*j}), j = 1, \ldots, B$, correspondente a cada amostra *bootstrap*;

P4) Determine a incerteza em torno de $\hat{\theta}$ como:

- erro padrão amostral como $\text{ep}(\hat{\theta}^*)$ definido em (7);
- histograma dos $\hat{\theta}_j^*$.

► Paramétrico

Dada a amostra aleatória $Y_1, \ldots, Y_n$ proveniente de uma população com função de probabilidade ou função densidade f que depende de uma parâmetro θ e utilizando um método de estimação de parâmetros, podemos obter uma estimativa paramétrica da função de distribuição acumulada $F(y|\hat{\theta})$. Portanto, podemos realizar o procedimento *bootstrap* utilizando $F(y|\hat{\theta})$ ao invés da empírica F_n como o estimador de F. Denominamos este método de retirar as amostras para determinar a incerteza em torno de θ como *bootstrap* paramétrico.

Algoritmo:

P1) Escolha B;

P2) Extraia $\mathbf{y}^{*1}, \ldots, \mathbf{y}^{*b}$ amostras *bootstrap* independentes de $\hat{F} = F(y|\hat{\xi})$, em que $\mathbf{y}^{*j} = (y_1^{*j}, \ldots, y_n^{*j})$ é um vetor de n observações que contêm a j-ésima amostra *bootstrap* para $j = 1, \ldots, B$;

P3) Avalie a réplica *bootstrap*, $\hat{\theta}_j^*, j = 1, \ldots, B$, correspondente a cada amostra *bootstrap*;

P4) Determine a incerteza em torno de $\hat{\theta}$ como:

- erro padrão amostral como $\text{ep}(\hat{\theta}^*)$ definido em (7);
- histograma dos $\hat{\theta}_j^*$.

★ **Observação**: Para estimação do erro padrão, valores de B variam de 50 200 réplicas *bootstrap*.