# CLASSIFICAÇÃO E ANÁLISE DE DADOS

## *Métodos e Aplicações III - CLADMAp III*


**Editores**

Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Marcelo

Ana Sousa Ferreira

Paulo Infante

Adelaide Figueiredo

# CLASSIFICAÇÃO E ANÁLISE DE DADOS

## MÉTODOS E APLICAÇÕES III - CLADMAp III

**Editores**
Helena Bacelar-Nicolau
Fernanda Sousa
Carlos Marcelo
Ana Sousa Ferreira
Paulo Infante
Adelaide Figueiredo

Associação Portuguesa de
Classificação e Análise de Dados

**Título**
Classificação e Análise de Dados – Métodos e Aplicações III

**Editores**
Helena Bacelar-Nicolau (Universidade de Lisboa)
Fernanda Sousa (Universidade do Porto)
Carlos Marcelo (Instituto Nacional de Estatística)
Ana Sousa Ferreira (Universidade de Lisboa)
Paulo Infante (Universidade de Évora)
Adelaide Figueiredo (Universidade do Porto)

**Impressão**
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA

**1.ª Edição**
**Lisboa, Abril de 2019**
**ISSN 2183-8801**
**Depósito legal 454535/19**
Tiragem: 200 exemplares

# Avaliação de métodos de estimação da variância em amostras complexas

**Eládio Muianga[1] · Anabela Afonso[2]**

**Resumo** Em amostras complexas nem sempre é possível obter uma expressão analítica para o estimador da variância dos estimadores, existindo na literatura alguns métodos para obter aproximações para esse estimador. Neste trabalho estudou-se o desempenho de alguns desses métodos. São simulados dados, a partir da informação real da atividade económica Moçambicana, e usam-se diferentes esquemas de amostragem com o objetivo de contribuir para as investigações por amostragem em Moçambique. O estimador *Taylor* apresentou o melhor desempenho e o estimador *bootstrap* foi o menos preciso.

**Palavras-chave:** Amostras Complexas, Enviesamento, Erro Quadrático Médio, Inferência, Estimadores da Variância.

## 1 Introdução

A necessidade de conhecer uma população impulsiona o processo de recolha e análise de informação. Usualmente, é muito difícil, ou impossível, estudar a totalidade da população, daí a importância do seu estudo com recurso a amostras. Conceber um estudo por amostragem é um processo complexo, desde antes da recolha dos dados até a fase de análise dos mesmos.

As amostras complexas combinam um conjunto de métodos probabilísticos de amostragem para a seleção de uma amostra representativa da população (Szwarcwald e Damacena, 2008). Estas amostras têm pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção

[1] Mestrado em Modelação Estatística e Análise de Dados, Instituto Nacional de Estatística de Moçambique, *eladio.muianga@ine.gov.mz*
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, *aafonso@uevora.pt*

desiguais, ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008). Comparando com a amostragem aleatória simples, sabe-se que a amostragem estratificada, quando usada de forma conveniente, em geral, produz estimativas mais precisas e que a amostragem por conglomerados acarreta uma perda de precisão das estimativas. Para medir o efeito do delineamento de amostragem, Kish (1965) propôs a utilização da medida *deff* (*design effect*). Esta medida consiste na razão entre a variância do estimador do delineamento de amostragem complexo e a variância do estimador considerando uma amostragem aleatória simples. Esta medida para além de quantificar a perda, ou o ganho, de precisão da estimativa também pode ser usada para determinar a dimensão da amostra complexa com base na dimensão da amostra aleatória simples.

Contudo, para as amostras complexas nem sempre é possível obter uma expressão analítica para o estimador da variância dos estimadores dos parâmetros. Ao longo dos anos foram propostas aproximações para este estimador, ajustadas à natureza complexa do plano da amostra, sendo as mais utilizadas para estimar a variância dos totais e médias estimados: o método de linearização *Taylor* e as técnicas de reamostragem e replicação (*Jackknife* e *bootstrap*).

Neste trabalho avalia-se o desempenho, e respetivas propriedades, dos estimadores mais usuais da variância do totais e médias amostrais em amostras complexas. Este estudo tem como objetivo contribuir para as investigações de amostragem em Moçambique. Para tal, usam-se dados simulados a partir da realidade da atividade económica Moçambicana e consideram-se diferentes esquemas de amostragem.

## 2 Estimação da variância

Um requisito básico em todas formas de análise, senão a principal exigência nas investigações práticas, é que uma medida de precisão deve ser fornecida para cada estimativa derivada dos dados de investigação. A medida de precisão mais usada é a variância do estimador (Wolter, 2007).

Os totais e as médias da população podem ser facilmente estimados a partir dos pesos de amostragem. Estimar variâncias é um processo mais complexo pois, em amostras complexas com várias etapas de estratificação e conglomerados, a variância do estimador dos totais e médias deve ser calculada para cada nível e, em seguida, deve ser combinada segundo o delineamento de amostragem (Lohr, 2010).

Estimar a variância dos estimadores em amostras complexas é um tema atual e bastante complexo em termos gerais, uma vez que, de um modo geral, não existe uma expressão analítica para um estimador centrado e eficiente da variância do

estimador. Para estimar a variância dos totais e médias estimados existem, basicamente, duas abordagens:

- Analítica, usando o método de linearização de *Taylor*;
- Métodos de reamostragem ou replicação (*Jackknife,* réplicas equilibradas repetidas (*BRR*) e *bootstrap*).

Nos métodos de reamostragem duas ou mais subamostras são selecionadas de uma dada população, ou eventualmente, de uma amostra. Com base em cada uma das amostras estima-se o parâmetro de interesse. É a partir da combinação das estimativas obtidas que se obtém uma estimativa da variância. Estes métodos diferem entre si na forma de gerar réplicas de amostras.

A escolha de um método para estimar a variância envolve um equilíbrio de fatores tais como a precisão e o custo. Nenhum dos métodos para estimar a variância do estimador é o melhor no geral (Wolter, 2007, pág. 366). Por isso, num bom julgamento, no qual está envolvida a escolha de um método para estimar a variância, não será surpresa se o estatístico recomendar métodos diferentes para diferentes aplicações da investigação.

Seja $\hat{\theta}$ um estimador para o parâmetro populacional $\theta$ e $Var(\hat{\theta})$ a sua variância. De seguida, apresentam-se os três métodos de estimação que são usualmente mais utilizados para estimar $Var(\hat{\theta})$, quando o parâmetro é uma média ($\mu$) ou um total ($\tau$): o método de *Linearização em série de Taylor* e os métodos de replicação (reamostragem) *Jackknife* e *bootstrap.*

## 2.1 Método de linearização de *Taylor*

Este método de linearização baseia-se na expansão em série de *Taylor* a qual permite obter uma aproximação linear para a estatística de interesse.

Sejam $\hat{\theta}_j$, $j = 1, \dots, k$, estimadores não enviesados para os $\theta_j$, $j = 1, \dots, k$ parâmetros populacionais. Seja $\hat{\theta} = \sum_{j=1,\dots,k}^{k} \hat{\theta}_j$ e $a_j$ e $a_l$ constantes de linearização; então pelo método de linearização de *Taylor* a variância do estimador pode ser estimada por (Lohr, 2010):

$$\widehat{Var}_T(\hat{\theta}) = \sum_{j=1}^{k} a_j \widehat{Var}(\hat{\theta}_j) + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} a_j a_l \widehat{Cov}(\hat{\theta}_j, \hat{\theta}_l).$$

A precisão da aproximação da linearização depende do tamanho da amostra, o que origina a que a variância do estimador por vezes seja subestimada se a amostra não for grande o suficiente (Lohr, 2010).

Quando o estimador $\hat{\theta}$ é uma função não linear das observações, o estimador *Taylor* da $Var(\hat{\theta})$ é enviesado, mas tipicamente consistente (Wolter, 2007).

## 2.2 Método de *Jackknife*

Este método foi proposto por Quenouille (1956) como um método para reduzir o enviesamento dos estimadores, num contexto da Estatística Clássica. Posteriormente, Tukey (1958) propôs usá-lo para estimar variâncias e calcular intervalos de confiança.

A ideia geral deste método consiste em tomar como ponto de partida uma amostra de dimensão $n$ a partir da qual serão consideradas todas as subamostras possíveis em que se eliminam $k$ elementos de cada vez. Com base em cada uma destas subamostras é obtida uma estimativa para o parâmetro de interesse e a variância é estimada usando a informação de todas estas estimativas.

No caso do método *delete-1*, considera-se $\hat{\theta}_{(j)}$ um estimador com a mesma forma que $\hat{\theta}$ sem a observação $j$ e $n$ a dimensão da amostra, sendo o estimador de *Jackknife* dado por (Lohr, 2010):

$$\widehat{Var}_{JK}(\hat{\theta}) = \frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{(j)} - \bar{\hat{\theta}}\right)^2.$$

Portanto, a variância do estimador é estimada com base na variabilidade entre as estimativas obtidas a partir das subamostras constituídas.

No caso de uma amostra por grupos, de forma a não perder a estrutura dos grupos, em vez de se eliminar uma observação de cada vez na formação das subamostras, eliminam-se todas as observações pertencentes ao mesmo grupo.

## 2.3 Método de *bootstrap*

O método de *bootstrap* foi proposto por Efron (1979). É um método de reamostragem de computação intensiva que tem sido bastante aplicado em muito devido aos avanços computacionais.

Tal como no método de *Jackknife*, neste método a variância pretendida será estimada a partir de subamostras, designadas por réplicas, que serão extraídas de uma amostra inicial.

Seja $R$ o número de réplicas *bootstrap* e $\hat{\theta}_r^*$ um estimador de $\theta$ calculado usando o vetor dos pesos replicados. A variância estimada pelo método de *bootstrap* é dada por (Lohr, 2010):

$$\widehat{Var}_B(\hat{\theta}) = \frac{1}{B-1}\sum_{l=1}^{R}\left(\hat{\theta}_r^* - \hat{\theta}\right)^2.$$

Para algumas configurações, o método de *bootstrap* pode exigir mais cálculos do que o método de *Jackknife*, uma vez que tipicamente *R* representa um número muito grande de réplicas. No entanto, em outras investigações de larga escala se, por exemplo, for retirada uma amostra aleatória estratificada, o método de *bootstrap* pode exigir menos cálculos do que o *Jackknife*. De notar que, a estimativa de variância *bootstrap* difere quando é retirado um conjunto diferente de amostras *bootstrap*.

## 3 Métodos

Devido à confidencialidade no acesso aos dados reais, foi adotada uma metodologia de geração de números pseudoaleatórios para a criação do universo de estudo, "Empresas do sector do comércio", com base na informação de uma amostra representativa da estrutura deste sector fornecida pelo Instituto Nacional de Estatística de Moçambique (INE-M).

Foram simulados três conjuntos de dados populacionais de empresas por região, classificador de atividade económica CAE Rev-2, número de trabalhadores (*NT* e *NTCat*) e volume de negócio (*VN*). Assumiu-se que a variável de interesse *VN* dependia linearmente de *NT*, i.e., $VN = \alpha + \beta NT + \varepsilon$, com os valores de $\alpha$ e $\beta$ a variarem segundo a população e $\varepsilon \sim N(0; \sigma NT)$, com $\sigma > 0$.

As principais características das populações geradas são (Tabela 1):

I. Com caraterísticas semelhantes às da população real;
II. Considerou-se uma igual distribuição de empresas por Região e CAE;
III. Introduziu-se uma maior dispersão na distribuição da variável de interesse.

**Tabela 1 –** Características das populações geradas.

| Variável | Categorias | População | | |
|---|---|---|---|---|
| | | **I** | **II** | **III** |
| **Região** | Norte | 14% | 33% | 14% |
| | Centro | 47% | 34% | 47% |
| | Sul | 39% | 33% | 39% |
| **CAE** | 45 | 6% | 32% | 6% |
| | 46 | 5% | 34% | 5% |
| | 47 | 89% | 34% | 89% |
| **NTCat*** | Pequena | 69% | 69% | 40% |
| | Média | 20% | 20% | 42% |
| | Grande | 10% | 10% | 18% |

* Dimensão das empresas: pequena (NT < 50), média (50 ≤ NT < 250), e grande (NT ≥ 250).

Com vista a estimar o parâmetro de interesse, i.e., a média do *VN*, de cada um dos universos populacionais foram sorteadas 10 000 réplicas de amostras de acordo com quatro delineamentos de amostragem:

1. Aleatório (sem reposição) estratificado pela variável região;
2. Aleatório (sem reposição) estratificado por duas variáveis (região e CAE);
3. Por grupos a duas etapas (região e CAE);
4. Multietápico (estratificado por região e por grupos de CAE a duas etapas).

Para analisar a qualidade dos estimadores recorreu-se a duas propriedades fundamentais: o não enviesamento e a precisão. A precisão foi medida pelo erro quadrático médio ($EQM$) e pela raiz do erro quadrático médio escalado ($REQME$):

$$REQME(\hat{\theta}) = \frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})} \times 100.$$

O $EQM$ é uma medida de precisão que depende da escala da variável em estudo enquanto o $REQME$ é uma medida cujo resultado é expresso em percentagem. Deste modo, esta medida de desempenho escalado, que combina o enviesamento e a precisão, permite comparar o desempenho do estimador quando as populações têm características diferentes (Walther & Moore, 2005).

Todo o estudo foi realizado com auxílio do *software R Project*.

## 4 Resultados

### 4.1 Distribuição dos estimadores da variância da média estimada

As distribuições de amostragem obtidas para os estimadores *Taylor* $(\widehat{Var}_T)$, *Jackknife* $(\widehat{Var}_{JK})$ e *bootstrap* $(\widehat{Var}_B)$ para a variância de $\hat{\mu}$, para cada uma das três populações, são apresentadas na Figura 1. Além disso, quando possível, também é apresentada a distribuição de amostragem do estimador usualmente indicado na literatura $(\widehat{Var}_L)$ para os delineamentos de amostragem considerados neste trabalho e cuja expressão é apresentada, por ex., em Lohr (2010).

Os resultados empíricos do presente estudo mostram que não existe um padrão consistente no comportamento da distribuição dos estimadores da variância por população.

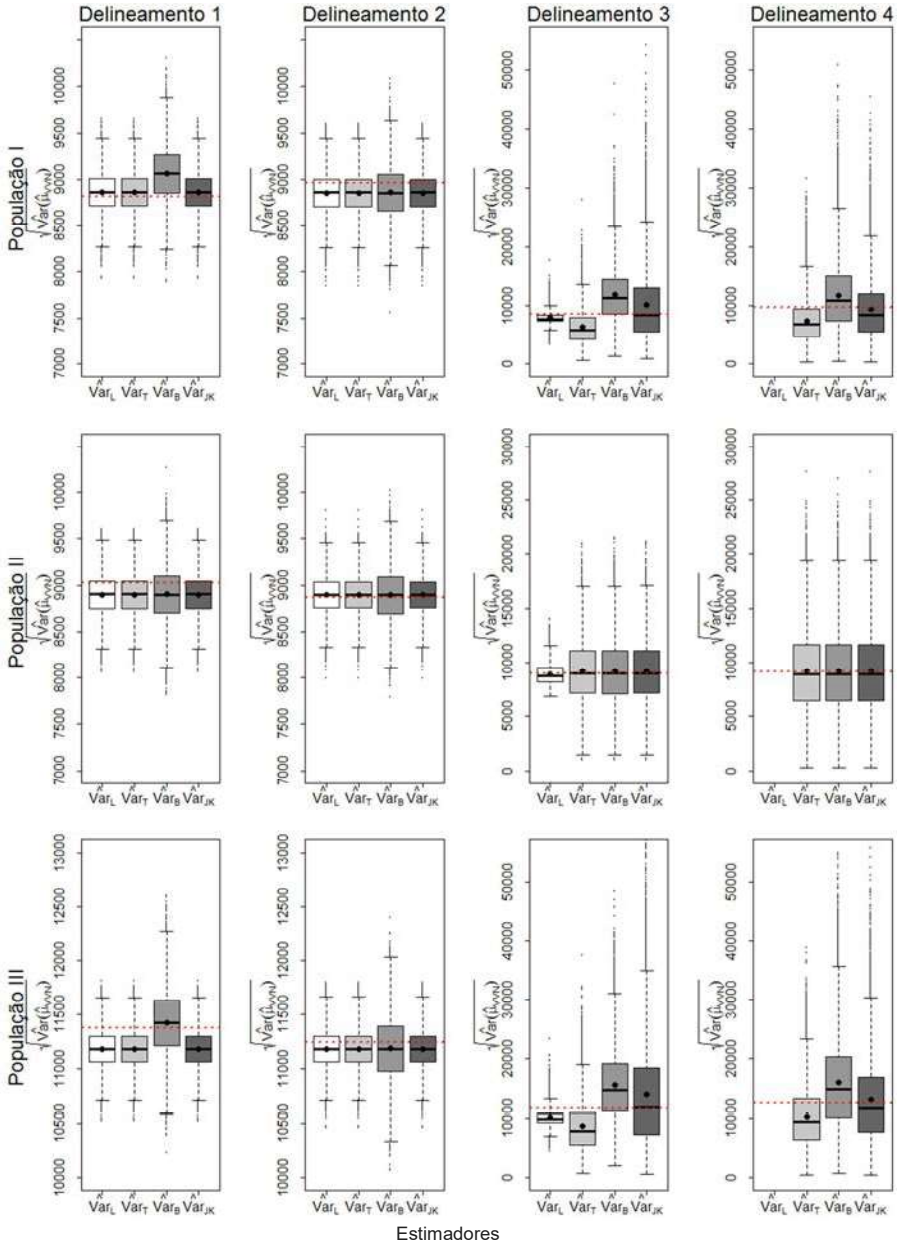


**Figura 1 –** Distribuição empírica de $\sqrt{\widehat{Var}(\hat{\mu}_{VVN})}$.

Nos delineamentos de amostragem estratificados (1 e 2) as distribuições são aproximadamente simétricas e as distribuições dos estimadores *Taylor*, *Jackknife* e o indicado na literatura são similares. Resultados diferentes podem ser visualizados nos planos multietápicos (delineamentos 3 e 4), que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais. A distribuição dos estimadores varia nas três populações em estudo. A assimetria da distribuição depende da população e, em todas as populações, predominam os valores atípicos superiores.

Nas populações I e III os estimadores *bootstrap* e *Jackknife* são enviesados no delineamentos multietápicos, sendo o estimador *bootstrap* o mais enviesado. Na população II, a distribuição dos estimadores é simétrica apresentando o estimadores *Taylor* e *Jackknife* distribuições similares.

## 4.2 Análise comparativa dos estimadores da variância

Nos delineamentos de amostragem estratificados (1 e 2) os estimadores apresentam enviesamentos idênticos nas três populações (Tabela 2), mas o estimador *bootstrap* é o menos preciso, apresentando os maiores valores de $REQME$ (Tabela 3). Os restantes três estimadores apresentam valores iguais de $REQME$.

Nos delineamentos multietápicos (3 e 4), o enviesamento do estimador indicado na literatura depende das características da população, mas é o estimador mais preciso (menor $REQME$). O estimador *bootstrap* é, de uma forma geral, o mais enviesado. No delineamento a duas etapas (3) o estimador *Jackknife* é o menos preciso enquanto no delineamento 4 o estimador *bootstrap* foi o mais impreciso.

**Tabela 2 –** Enviesamento dos estimadores.

| População | Delineamento | $Env\left(\widehat{Var}(\hat{\mu}_{VVN})\right)$ % | | | |
|---|---|---|---|---|---|
| | | $\widehat{Var}_L$ | $\widehat{Var}_T$ | $\widehat{Var}_B$ | $\widehat{Var}_{JK}$ |
| I | 1 | 0,96 | 0,96 | 5,71 | 0,96 |
| | 2 | -2,41 | -2,41 | -2,28 | -2,41 |
| | 3 | -10,96 | -31,10 | 125,96 | 96,68 |
| | 4 | - | -25,73 | 87,51 | 25,89 |
| II | 1 | -2,65 | -2,65 | -2,56 | -2,65 |
| | 2 | 0,68 | 0,68 | 0,77 | 0,68 |
| | 3 | 0,96 | 16,03 | 16,06 | 16,36 |
| | 4 | - | 18,82 | 18,86 | 18,84 |
| III | 1 | -3,43 | -3,43 | 0,93 | -3,43 |
| | 2 | -1,00 | -1,00 | -0,88 | -1,00 |
| | 3 | -20,04 | -29,70 | 109,32 | 102,27 |
| | 4 | - | -14,98 | 100,54 | 43,12 |

**Tabela 3 –** Precisão dos estimadores.

| População | Delineamento | $REQME\left(\widehat{Var}(\hat{\mu}_{VVN})\right)$ % | | | |
|---|---|---|---|---|---|
| | | $\widehat{Var}_L$ | $\widehat{Var}_T$ | $\widehat{Var}_B$ | $\widehat{Var}_{JK}$ |
| I | 1 | 5,10 | 5,10 | 9,11 | 5,10 |
| | 2 | 5,45 | 5,45 | 6,91 | 5,45 |
| | 3 | 27,87 | 75,84 | 225,13 | 296,43 |
| | 4 | - | 86,72 | 223,89 | 162,75 |
| II | 1 | 5,44 | 5,44 | 6,95 | 5,44 |
| | 2 | 4,85 | 4,85 | 6,61 | 4,85 |
| | 3 | 24,11 | 73,45 | 73,71 | 73,88 |
| | 4 | - | 96,38 | 96,62 | 96,39 |
| III | 1 | 4,59 | 4,59 | 5,60 | 4,59 |
| | 2 | 3,30 | 3,30 | 5,55 | 3,30 |
| | 3 | 33,81 | 80,88 | 202,19 | 296,82 |
| | 4 | - | 91,26 | 227,57 | 177,61 |

## 5 Conclusão

Nos delineamentos de amostragem estratificados os resultados dos estimadores *Taylor* e *Jackknife* coincidiram com o indicado na literatura. Apresentaram melhores resultados quanto ao enviesamento e à precisão das estimativas do que o estimador *bootstrap*. Verificou-se, à semelhança do referido por Pessoa & Silva (1998), que neste tipo de delineamentos as estimativas *Jackknife* são iguais às estimativas *Taylor*.

Nos planos multietápicos, que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais, o estimador *Taylor* mostrou ser menos enviesado e mais preciso do que os estimadores *Jackknife* e *bootstrap*. Estes dois últimos estimadores são enviesados, dependendo o enviesamento do tipo de população, e, além disso, muito imprecisos.

Os estimadores indicados na literatura para os delineamentos estratificados e por grupos em duas etapas foram os que garantiram maior confiabilidade nas estimativas. Seguiu-se o estimador *Taylor* com resultados mais fiáveis para as aproximações.

Em Moçambique grande parte dos estudos realizados pelo Instituto Nacional de Estatística são baseados em amostras, havendo a necessidade de garantir um elevado grau de precisão para as estimativas geradas. Com este trabalho pretendeu-se dar um contributo no sentido de minorar essa necessidade, auxiliando na escolha do estimador a usar na produção de estatísticas oficiais no país.

## Agradecimentos

Anabela Afonso é membro do CIMA, centro de investigação financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal, no âmbito do projeto «UID/MAT/04674/2019 (CIMA)».

## Referências

EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7(1), 1-26.

LAVRAKAS, P. J. (2008). *Encyclopedia of Survey Research Methods*, SAGE Publications, Thousand Oaks.

LORH, S. L. (2010). *Sampling: Design and Analysis*, Second Edition, Michelle Julet, Boston.

PESSOA, D. G. C. & SILVA P. L. N. (1998). *Análise de Dados Amostrais Complexos*, Associação Brasileira de Estatística, São Paulo.

QUENOUILLE, M. H. (1956). Notes on Bias in Estimation, *Biometrika*, 43, 353–360.

SZWARCWALD, C. L. & DAMACENA, G. N. (2008). Amostras Complexas em Inquéritos Populacionais: Planeamento e Implicações na Análise Estatística dos Dados. *Revista Brasileira de Epidemiologia*, 11(11), 38–45.

TUKEY, J. W. (1958). Bias and Confidence in Not-quite large Samples, *Annals of Mathematical Statistics*, 29, 614.

WALTHER, B. A. & MOORE, J. L. (2005). The Concepts of bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance, *Ecography*, 28, 815–829.

WOLTER K. M. (2007). *Introduction to Variance Estimation*, Springer-Verlag, New York.