

# Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2018-05-01



# Sumário

<b>Prefácio</b>	<b>5</b>
Agradecimentos . . . . .	5
<b>1 Introdução</b>	<b>7</b>
1.1 Motivação . . . . .	7
1.2 Objetivos do Livro . . . . .	11
1.3 Estrutura do Livro . . . . .	12



# Prefácio

Placeholder

## Agradecimentos



# Capítulo 1

## Introdução

### 1.1 Motivação

Este livro trata de questões e ideias de grande importância para os analistas de dados obtidos através de pesquisas amostrais, tais como as conduzidas por agências produtoras de informações estatísticas oficiais ou públicas. Tais dados são comumente utilizados em análises descritivas envolvendo a obtenção de estimativas para totais, médias, proporções e razões. Nessas análises, em geral, são devidamente incorporados os pesos distintos das observações e a estrutura do plano amostral empregado para obter os dados considerados.

Nas últimas décadas tornou-se muito mais frequente um outro tipo de uso de dados de pesquisas amostrais. Tal uso, denominado secundário e/ou analítico, envolve a construção e ajuste de modelos, geralmente feito por analistas que trabalham fora das agências produtoras dos dados. Neste caso, o foco da análise busca estabelecer a natureza de relações ou associações entre variáveis ou testar hipóteses. Para tais fins, a estatística clássica conta com um vasto arsenal de ferramentas de análise, já incorporadas aos principais sistemas estatísticos disponíveis (tais como MINITAB, R, SAS, SPSS, etc).

Muitas ferramentas de análise convencionais disponíveis nesses sistemas estatísticos geralmente partem de hipóteses básicas sobre as amostras disponíveis que só são válidas quando os dados foram obtidos através de Amostras Aleatórias Simples Com Reposição (AASC). Por exemplo, a hipótese de observações Independentes e Identicamente Distribuídas (IID) corresponde justamente ao caso de observações selecionadas por AASC de uma população especificada. Tais hipóteses são geralmente inadequadas para modelar observações provenientes pesquisas amostrais de populações finitas, pois desconsideram os seguintes aspectos relevantes dos planos amostrais usualmente empregados nessas pesquisas:

- i.) **probabilidades desiguais de seleção das unidades;**

- ii.) **conglomeramento das unidades;**
- iii.) **estratificação;**
- iv.) **calibração ou imputação para não-resposta e outros ajustes.**

Em amostragem de populações finitas, a abordagem probabilística emprega pesos para as observações amostrais que dependem das probabilidades de seleção das unidades, que podem ser desiguais. Em consequência, as estimativas pontuais de parâmetros descritivos da população ou mesmo de parâmetros de modelos são influenciadas por pesos distintos das observações.

Além disso, as estimativas de variância (ou da precisão dos estimadores) são influenciadas pela conglomeramento, estratificação e pesos, ou no caso de não resposta, também por eventual imputação de dados faltantes ou reponderação das observações disponíveis para compensar a não resposta. Ao ignorar estes aspectos, as ferramentas convencionais dos sistemas estatísticos tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

O exemplo a seguir considera o uso de dados de uma pesquisa amostral real conduzida pelo IBGE para ilustrar como os pontos i) a iv) acima mencionados afetam a inferência sobre quantidades descritivas populacionais tais como totais, médias, proporções e razões.

Pesquisa TIC Domicílios 2019 do NIC.br

Os dados deste exemplo são relativos à distribuição dos pesos de domicílios na amostra da Pesquisa TIC Domicílios 2019 do NIC.br (TICDOM 2019), realizada pelo NIC.br. (de Informação e Coordenação do Ponto BR, 2020) apresenta os resultados da pesquisa, e seu capítulo intitulado ‘Relatório Metodológico’ descreve os métodos e o plano amostral empregado na pesquisa, que foi estratificado e conglomerado em múltiplos estágios, com alocação desproporcional da amostra nos estratos.

As Unidades Primárias de Amostragem (UPAs) foram municípios ou setores censitários da Base Operacional Geográfica do IBGE conforme usada para o Censo Demográfico de 2010. A seleção de municípios quando estes eram UPAs foi feita usando Amostragem Sistemática com Probabilidades Proporcionais ao Tamanho (AS-PPT) - ver a seção 10.6 de (Silva et al., 2020). A seleção dos setores dentro de cada município também foi feita com AS-PPT. Dentro cada setor censitário selecionado, quinze domicílios foram selecionados por amostragem aleatória simples sem reposição, após a atualização do cadastro de domicílios do setor.

A amostra da pesquisa foi planejada e dimensionada visando ao fornecimento de estimativas com precisão adequada para as cinco macrorregiões do Brasil. Os tamanhos da amostra planejada de setores e domicílios para as macrorregiões são apresentados na tabela 1.1.



Tabela 1.1: Tamanhos da amostra de setores e domicílios por macrorregião

Macrorregião	Setores	Domicílios
Norte	201	3.015
Nordeste	617	9.255
Sudeste	863	12.945
Sul	337	5.055
Centro-Oeste	196	2.940
Total	2.214	33.210

Tabela 1.2: Resumos da distribuição dos pesos de domicílios por macrorregião

Macrorregião	Mínimo	Quartil1	Mediana	Quartil3	Máximo
Norte	1,8	1.957	2.898	4.359	82.627
Nordeste	103,8	1.283	2.057	3.314	40.118
Sudeste	36,0	1.814	2.583	3.583	27.993
Sul	20,0	1.028	1.756	2.706	118.715
Centro-Oeste	140,8	1.153	2.401	3.640	29.029
Total	1,8	1.546	2.470	3.636	118.715

A Tabela 1.2 apresenta um resumo das distribuições dos pesos amostrais dos domicílios pesquisados na TICDOM 2019 para as macrorregiões separadamente, e também para o conjunto da amostra da pesquisa.

No cálculo dos pesos amostrais foram consideradas as probabilidades de inclusão dos domicílios na amostra, bem como as correções de calibração para compensar a não-resposta. Contudo, a grande variabilidade dos pesos amostrais da TICDOM 2019 é devida, principalmente, à variabilidade das probabilidades de inclusão na amostra, ilustrando desta forma o ponto i) citado anteriormente nesta seção. Tal variabilidade é devida à alocação desproporcional da amostra entre os estratos geográficos, e ao emprego de contagens defasadas de domicílios nos setores para definir probabilidades de seleção dos mesmos.

Nas análises de dados desta pesquisa, deve-se considerar que há domicílios com pesos muito diferentes. Por exemplo, a razão entre o maior e o menor peso é cerca de 66 mil vezes. Os pesos também variam bastante entre as regiões, com mediana 1,65 vezes maior na região Sudeste quando comparada com a região Norte, em função da alocação desproporcional da amostra nas regiões. Os maiores pesos são também muito maiores que os pesos medianos, com essa razão sendo 48 vezes para o conjunto da amostra.

Tais pesos são utilizados para *expandir* os dados, multiplicando-se cada observação pelo seu respectivo peso. Assim, por exemplo, para *estimar* quantos

domicílios *da população* pertencem a determinado conjunto ( *domínio* ), basta somar os pesos dos domicílios da amostra que pertencem a este conjunto. É possível ainda incorporar os pesos, de maneira simples e natural, quando se quer estimar medidas descritivas simples da população, tais como totais, médias, proporções, razões, etc. Os métodos para estimação de parâmetros descritivos da população como os aqui citados são cobertos com maior detalhe em (Silva et al., 2020).

Por outro lado, quando se quer utilizar a amostra para estudos analíticos, as opções padrão disponíveis nos sistemas estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações IID. Por exemplo, os procedimentos padrão disponíveis para estimar a média populacional permitem utilizar pesos distintos das observações amostrais, mas tratariam tais pesos como se fossem frequências de observações repetidas na amostra, e portanto interpretariam a soma dos pesos como tamanho amostral, situação que, na maioria das vezes, geraria inferências incorretas sobre a precisão das estimativas resultantes. Isto ocorre porque o tamanho da amostra é muito menor que a soma dos pesos amostrais usualmente encontrados nos arquivos de microdados de pesquisas disseminados por agências de estatísticas oficiais ou públicas, como é o caso da pesquisa TICDOM 2019 aqui considerada. Em tais pesquisas, a opção mais frequente é disseminar pesos que, quando somados, estimam o total de unidades *da população*.

Além disso, a variabilidade dos pesos para distintas observações amostrais produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da conglomeração e da estratificação - pontos ii) e iii) mencionados anteriormente.

Para exemplificar o impacto de ignorar os pesos e o plano amostral ao estimar quantidades descritivas populacionais, tais como totais e proporções, calculamos estimativas de quantidades desses diferentes tipos usando a amostra da TICDOM 2019 juntamente com estimativas das respectivas variâncias. Tais estimativas de variância foram calculadas sob duas estratégias:

- a) **considerando Amostragem Aleatória Simples (AAS)** , e portanto ignorando o plano amostral efetivamente adotado na pesquisa; e
- b) **considerando o plano amostral da pesquisa e os pesos diferenciados das unidades.**

Na Tabela ?? apresentamos as estimativas dos seguintes parâmetros populacionais: 1. Porcentagem de domicílios com computador de mesa; 2. Porcentagem de domicílios com notebook; 3. Porcentagem de domicílios com tablete; 4. Porcentagem de domicílios com algum computador (de mesa, notebook ou tablete); 5. Total de domicílios com algum computador (de mesa, notebook ou tablete); 6. Número médio de celulares por domicílio;

Tabela 1.3: Estimativas de parâmetros populacionais e EPAs

Parâmetro	Est_por_AAS	Est_Verd	DP_Est_Verd
Porcentagem de domicílios com computador de mesa	14,21	16,17	0,4
Porcentagem de domicílios com notebook	22,84	26,05	0,6
Porcentagem de domicílios com tablete	11,24	12,95	0,3
Porcentagem de domicílios com computador	35,34	39,36	0,6
Total de domicílios com computador (milhões)	25,10	27,95	1,3
Número médio de computadores por domicílio que tem computador	1,55	1,63	0,0

A razão entre as estimativas de variância obtidas sob o plano amostral verdadeiro (de fato usado na pesquisa) e sob AAS foi estimada para cada uma das estimativas consideradas usando o pacote `survey` do R (Lumley, 2017). Essa razão fornece uma medida do efeito de ignorar o plano amostral. Os resultados das estimativas pontuais (`Est_por_AAS` e `Est_Verd` para as estimativas considerando AAS e o plano amostral verdadeiro, respectivamente), do desvio padrão da estimativa considerando o plano amostral verdadeiro (`DP_Est_Verd`) e do efeito do plano amostral (EPA) são apresentados na Tabela ??.

Os resultados mostram que há diferenças entre as estimativas pontuais dos parâmetros considerados, com uma tendência de subestimar quando se ignora os pesos e o plano amostral efetivamente usado na pesquisa. As estimativas dos EPAs variam entre 2,64 e 5,30, se deixarmos de fora o EPA maior que 30 observado para a estimativa da contagem de domicílios com computador. Estes valores indicam que ignorar o plano amostral na estimação da precisão levaria também à subestimação dos erros padrões.

Note que as variáveis e parâmetros cujas estimativas foram apresentadas na Tabela ?? não foram escolhidas de forma a acentuar os efeitos ilustrados, mas tão somente para representar distintos parâmetros (totais, médias, proporções) e variáveis de interesse. Os resultados apresentados para as estimativas de EPA ilustram bem o cenário típico em pesquisas amostrais complexas: o impacto do plano amostral sobre a inferência varia conforme a variável e o tipo de parâmetro de interesse. Note ainda que todas as estimativas de EPA apresentaram valores superiores a 2.

## 1.2 Objetivos do Livro

Este livro tem três objetivos principais:

- 1) **Apresentar uma coleção de métodos e recursos computacionais disponíveis no R para análise de dados de pesquisas amostrais,**

equipando o analista para trabalhar com tais dados, reduzindo assim o risco de inferências incorretas;

- 2) Ilustrar e analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise de dados quando estes são provenientes de pesquisas amostrais complexas;
- 3) Ilustrar o potencial analítico de muitas das pesquisas produzidas por agências de estatísticas públicas para responder questões de interesse, mediante uso de ferramentas de análise estatística agora já bastante difundidas, aumentando assim o valor adicionado destas pesquisas.

Para alcançar tais objetivos, adotamos uma abordagem fortemente ancorada na apresentação de exemplos de análises de dados obtidos em pesquisas amostrais, usando os recursos do sistema estatístico R (<http://www.r-project.org/>).

A comparação dos resultados de análises feitas das duas formas (considerando ou ignorando o plano amostral) permite avaliar o impacto de não se considerar os pontos i) a iv) anteriormente citados. O ponto iv) não é tratado de forma completa neste texto. O leitor interessado na análise de dados sujeitos a não-resposta pode consultar (Kalton, 1983), (Little and Rubin, 2002), (Rubin, 1987), (Särndal et al., 1992), ou (Schafer, 1997), por exemplo.

### 1.3 Estrutura do Livro

O livro está organizado em duas partes. A primeira parte representa uma segunda edição atualizada e revisada do conteúdo do livro publicado em 1998 (?). A segunda parte é uma coletânea de textos reunidos para cobrir temas não tratados no livro anterior, que foram produzidos por autores convidados, como forma de prestar homenagem ao Prof. Djalma Pessoa.

A parte 1 é composta por nove capítulos. Este primeiro capítulo discute a motivação para estudar o assunto e apresenta uma ideia geral dos objetivos e da estrutura do livro.

No Capítulo ??, procuramos dar uma visão das diferentes abordagens utilizadas na análise estatística de dados de pesquisas amostrais. Apresentamos um referencial para inferência com ênfase no *Modelo de Superpopulação* que incorpora, de forma natural, tanto uma estrutura estocástica para descrever a geração dos dados populacionais (modelo) como o plano amostral efetivamente utilizado para obter os dados amostrais (plano amostral). As referências básicas para seguir este capítulo são o capítulo 2 em (Nascimento Silva, 1996), o capítulo 1 em (Skinner et al., 1989) e os capítulos 1 e 2 em (Chambers and Skinner, 2003).

Esse referencial tem evoluído ao longo dos anos como uma forma de permitir a incorporação de ideias e procedimentos de análise e inferência usualmente

associados à Estatística Clássica à prática da análise e interpretação de dados provenientes de pesquisas amostrais. Apesar dessa evolução, sua adoção não é livre de controvérsia e uma breve revisão dessa discussão é apresentada no Capítulo ??.

No Capítulo ?? apresentamos uma revisão sucinta, para recordação, de alguns resultados básicos da Teoria de Amostragem, requeridos nas partes subsequentes do livro. São discutidos os procedimentos básicos para estimação de totais considerando o plano amostral, e em seguida revistas algumas técnicas para estimação de variâncias que são necessárias e úteis para o caso de estatísticas complexas, tais como razões e outras estatísticas requeridas na inferência analítica com dados amostrais. As referências centrais para este capítulo são os capítulos 2 e 3 em (Särndal et al., 1992), (Silva et al., 2020), (Wolter, 1985) e (Cochran, 1977).

No Capítulo ?? introduzimos o conceito de *Efeito do Plano Amostral (EPA)*, que permite avaliar o impacto de ignorar a estrutura dos dados populacionais ou do plano amostral sobre a estimativa da variância de um estimador. Para isso, comparamos o estimador da variância apropriado para dados obtidos por Amostragem Aleatória Simples (hipótese de AAS) com o valor esperado deste mesmo estimador sob a distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado (plano amostral verdadeiro). Aqui a referência principal foi o livro (Skinner et al., 1989), complementado com o texto de (Lehtonen and Pahkinen, 1995).

No Capítulo ?? estudamos a questão do uso de pesos ao analisar dados provenientes de pesquisas amostrais complexas, e introduzimos um método geral, denominado *Método de Máxima Pseudo Verossimilhança (MPV)*, para incorporar os pesos e o plano amostral na obtenção não só de estimativas de parâmetros dos modelos de interesse mais comuns, como também das variâncias dessas estimativas. As referências básicas utilizadas nesse capítulo foram (Skinner et al., 1989), (Pfeffermann, 1993), (Binder, 1983) e o capítulo 6 em (Nascimento Silva, 1996).

O Capítulo ?? trata da obtenção de *Estimadores de Máxima Pseudo Verossimilhança (EMPV)* e da respectiva matriz de covariância para os parâmetros em modelos de regressão linear quando os dados vêm de pesquisas amostrais complexas. Apresentamos alguns exemplos de aplicação desse método ilustrando o uso do pacote `survey` (Lumley, 2017) para ajustar modelos de regressão linear. As referências centrais são o capítulo 6 em (Nascimento Silva, 1996) e (Binder, 1983).

O Capítulo ?? trata da obtenção de *Estimadores de Máxima Pseudo Verossimilhança (EMPV)* e da respectiva matriz de covariância para os parâmetros em modelos de regressão logística quando os dados vêm de pesquisas amostrais complexas. Apresentamos alguns exemplos de aplicação desse método ilustrando o uso do pacote `survey` (Lumley, 2017) para ajustar modelos de regressão logística. As referências centrais são o capítulo 6 em (Nascimento Silva, 1996) e

(Binder, 1983).

Os Capítulos ?? e ?? tratam da análise de dados categóricos, dando ênfase à adaptação dos testes clássicos para proporções, de independência e de homogeneidade em tabelas de contingência, para lidar com dados provenientes de pesquisas amostrais complexas. Apresentamos correções das estatísticas clássicas e também a estatística de Wald baseada no plano amostral. As referências básicas usadas nesses capítulos foram o capítulo 4 em (Skinner et al., 1989) e o capítulo 7 (Lehtonen and Pahkinen, 1995). Também são apresentadas as ideias básicas de como efetuar ajuste de modelos log-lineares a dados de frequências em tabelas de múltiplas entradas.

A parte 2 é composta por mais dez capítulos, todos escritos por autores convidados. Todos estes temas foram objeto de avanços importantes tanto no desenvolvimento de métodos como no de ferramentas computacionais para sua implementação no ambiente do sistema R, desde que foi publicado o livro inicial.

10 Gráficos 11 Estimação de funções de densidade 12 Estimação de funções de distribuição e quantis 13 Estimação de medidas de desigualdade e pobreza 14 Modelos multiníveis 15 Modelos de teoria da resposta ao item 16 Estimação de fluxos 17 Modelos de séries temporais 18 Modelos de redes neurais 19 Modelos log-lineares para tabelas

O Capítulo ?? aborda a elaboração de alguns tipos de gráficos de uso frequente quando os dados elementares provém de pesquisas amostrais. Entre os gráficos cobertos estão histogramas, boxplots, diagramas de dispersão e gráficos tipo quantil-quantil (qq-plots).

O Capítulo ?? trata da estimação de densidades, ferramenta que tem assumido importância cada dia maior com a maior disponibilidade de microdados de pesquisas amostrais para analistas fora das agências produtoras. Também é apresentada ferramenta para elaboração de gráficos das densidades estimadas.

O Capítulo ?? trata da estimação de funções de distribuição empíricas e também de quantis. Também é apresentada ferramenta para elaboração de gráficos das funções de distribuição estimadas.

O Capítulo ?? trata da estimação de medidas de desigualdade e pobreza, enfatizando o uso destas em análises baseadas na renda de domicílios ou pessoas. Apresenta os recursos do pacote `convey` (inserir referência).

O Capítulo ?? trata da estimação e ajuste de modelos hierárquicos ou multiníveis considerando o plano amostral. Modelos hierárquicos têm sido bastante utilizados para explorar situações em que as relações entre variáveis de interesse em uma certa população de unidades elementares (por exemplo, crianças em escolas, pacientes em hospitais, empregados em empresas, moradores em regiões, etc.) são afetadas por efeitos de grupos determinados ao nível de unidades conglomeradas (os grupos). Ajustar e interpretar tais modelos é tarefa mais difícil

que o mero ajuste de modelos lineares, mesmo em casos onde os dados são obtidos de forma exaustiva ou por AAS, e ainda mais complicada quando se trata de dados obtidos através de pesquisas com planos amostrais complexos. Diferentes abordagens estão disponíveis para ajuste de modelos hierárquicos nesse caso, e este capítulo apresenta uma revisão de tais abordagens, ilustrando com aplicações a dados de pesquisas amostrais de escolares.

Uma das características que procuramos dar ao livro foi o emprego de exemplos com dados reais, retirados principalmente da experiência do IBGE com pesquisas amostrais complexas. Sem prejuízo na concentração de exemplos que se utilizam de dados de pesquisas do IBGE, incluímos também exemplos que consideram aplicações a dados de pesquisas realizadas por outras instituições. Nas duas décadas desde a primeira edição deste livro foram muitas as iniciativas de realizar pesquisas por amostragem em várias áreas, tendo a educação e a saúde como as mais proeminentes.

Para facilitar a localização e replicação dos exemplos pelos leitores, estes foram em sua maioria introduzidos em seções denominadas *Laboratório* ao final de cada um dos capítulos. Os códigos em R dos exemplos são todos fornecidos, o que torna simples a replicação dos mesmos pelos leitores. Optamos pelo emprego do sistema R que, por ser de acesso livre e gratuito, favorece o amplo acesso aos interessados em replicar nossas análises e também em usar as ferramentas disponíveis para implementar suas próprias análises de interesse com outros conjuntos de dados.

Embora a experiência de fazer inferência analítica com dados de pesquisas amostrais complexas já tenha alguma difusão no Brasil, acreditamos ser fundamental difundir ainda mais essas ideias para alimentar um processo de melhoria do aproveitamento dos dados das inúmeras pesquisas realizadas pelo IBGE e instituições congêneres, que permita ir além da tradicional estimação de totais, médias, proporções e razões. Esperamos com esse livro fazer uma contribuição a esse processo.

Uma dificuldade em escrever um livro como este vem do fato de que não é possível começar do zero: é preciso assumir algum conhecimento prévio de ideias e conceitos necessários à compreensão do material tratado. Procuramos tornar o livro acessível para um estudante de fim de curso de graduação em Estatística. Por essa razão, optamos por não apresentar provas de resultados e, sempre que possível, apresentar os conceitos e ideias de maneira intuitiva, juntamente com uma discussão mais formal para dar solidez aos resultados apresentados.

As provas de vários dos resultados aqui discutidos se restringem a material disponível apenas em artigos em periódicos especializados estrangeiros e portanto, são de acesso mais difícil. Ao leitor em busca de maior detalhamento e rigor, sugerimos consultar diretamente as inúmeras referências incluídas ao longo do texto. Para um tratamento mais profundo do assunto, os livros de (Skinner et al., 1989) e (Chambers and Skinner, 2003) são as referências centrais a consultar. Para aqueles querendo um tratamento ainda mais prático que

o nosso, os livros de (Lehtonen and Pahkinen, 1995) e (Heeringa et al., 2010) podem ser opções interessantes, sendo que este último apresenta os recursos do sistema STATA para análise de dados amostrais.



# Referências Bibliográficas

- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.
- Chambers, R. and Skinner, C., editors (2003). *Analysis of Survey Data*. John Wiley, Chichester.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley, Nova Iorque.
- de Informação e Coordenação do Ponto BR, N. N. (2020). *Pesquisa Sobre O Uso Das Tecnologias Da Informação E Da Comunicação No Brasil*.
- Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis.
- Kalton, G. (1983). Compensating for missing survey data. Technical report, The University of Michigan, Institute for Social Research, Survey Research Center, Ann Arbor, Michigan.
- Lehtonen, R. and Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley and Sons, Chichester.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with missing data*. John Wiley and Sons, Nova Iorque.
- Lumley, T. (2017). *survey: Analysis of Complex Survey Samples*. R package version 3.32-1.
- Nascimento Silva, P. L. D. (1996). *Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys*. PhD thesis, University of Southampton, Department of Social Statistics.
- Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61:317–337.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Nova Iorque.

- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall / CRC.
- Silva, P. L. d. N., Bianchini, Z. M., and Dias, A. J. R. (2020). *Amostragem: teoria e prática usando R*.
- Skinner, C. J., Holt, D., and Smith, T. M. F., editors (1989). *Analysis of Complex Surveys*. John Wiley and Sons, Chichester.
- Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, Nova Iorque.
- Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, Nova Iorque.