Análise de Dados Amostrais com R

Djalma Galvão Carneiro Pessoa e Pedro Luis do Nascimento Silva

Editores
Pedro Luis do N. Silva, Antonio José R. Dias, Zélia M. Bianchini e Sonia Albieri

11 de fevereiro de 2022, 11:44:59

Sumário

Bem-vindo

Uma preocupação básica de toda instituição produtora de informações estatísticas é com a utilização "correta" de seus dados. Isso pode ser interpretado de várias formas, algumas delas com reflexos até na confiança do público e na própria sobrevivência do órgão. Do nosso ponto de vista, enfatizamos um aspecto técnico particular, mas nem por isso menos importante para os usuários dos dados.

A revolução da informática, com a resultante facilidade de acesso ao computador, criou condições extremamente favoráveis à utilização de dados estatísticos produzidos por órgãos como o IBGE. Algumas vezes esses dados são utilizados para fins puramente descritivos. Outras vezes, porém, sua utilização é feita para fins analíticos, envolvendo a construção de modelos, quando o objetivo é extrair conclusões aplicáveis também a populações distintas daquela da qual se extraiu a amostra. Neste caso, é comum empregar, sem grandes preocupações, pacotes computacionais padrões disponíveis para a seleção e ajuste de modelos. É neste ponto que entra a nossa preocupação com o uso adequado dos dados produzidos pelo IBGE.

O que torna tais dados especiais para quem pretende usá-los para fins analíticos? Esta é a questão básica que é amplamente discutida ao longo deste texto. A mensagem principal que pretendemos transmitir é que certos cuidados precisam ser tomados para utilização correta dos dados de pesquisas amostrais como as que o IBGE realiza.

O que torna especiais dados como os produzidos pelo IBGE é que estes são obtidos através de pesquisas amostrais complexas de populações finitas que envolvem: **probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não resposta e outros ajustes**. Os sistemas tradicionais de análise ignoram estes aspectos, podendo produzir estimativas incorretas tanto dos parâmetros como para as variâncias destas estimativas. Quando utilizamos a amostra para estudos analíticos, as opções disponíveis nos pacotes estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações independentes e identicamente distribuídas - IID. Além disso, a variabilidade dos pesos produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da estratificação e conglomeração.

O objetivo deste livro é analisar o impacto das simplificações feitas ao utilizar procedimentos e pacotes usuais de análise de dados e apresentar os ajustes necessários desses procedimentos de modo a incorporar na análise, de forma apropriada, os aspectos aqui ressaltados. Para isto são apresentados exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também pacotes estatísticos especializados. A comparação dos resultados das análises feitas das duas formas permite avaliar o impacto de ignorar o plano amostral na análise dos dados resultantes de pesquisas amostrais complexas.

Agradecimentos

A elaboração de um texto como esse não se faz sem a colaboração de muitas pessoas. Em primeiro lugar, agradecemos à Comissão Organizadora do SINAPE por ter propiciado a oportunidade ao selecionar nossa proposta de minicurso. Agradecemos também ao IBGE por ter proporcionado as condições e os meios usados

para a produção da monografia, bem como o acesso aos dados detalhados e identificados que utilizamos em vários exemplos.

No plano pessoal, agradecemos a Zélia Bianchini pela revisão do manuscrito e sugestões que o aprimoraram. Agradecemos a Marcos Paulo de Freitas e Renata Duarte pela ajuda com a computação de vários exemplos. Agradecemos a Waldecir Bianchini, Luiz Pessoa e Marinho Persiano pela colaboração na utilização do processador de textos. Aos demais colegas do Departamento de Metodologia do IBGE, agradecemos o companheirismo e solidariedade nesses meses de trabalho na preparação do manuscrito.

Finalmente, agradecemos a nossas famílias pela aceitação resignada de nossas ausências e pelo incentivo à conclusão da empreitada.

Lista de Figuras

Lista de Siglas

Sigla	Descrição
AAS	Amostragem Aleatória Simples Sem Reposição
AASC	Amostragem Aleatória Simples Com Reposição
AS-PPT	Amostragem Sistemática com Probabilidades Proporcionais ao Tamanho
EMPV	Estimadores de Máxima Pseudo Verossimilhança
EPA	Efeito do Plano Amostral
IBGE	Instituto Brasileiro de Geografia e Estatística
IID	Independentes e Identicamente Distribuídas
MCP	Método do Conglomerado Primário
MINITAB	Minitab Statistical Software
MPV	Método de Máxima Pseudo Verossimilhança
NIC.br	Núcleo de Informação e Coordenação do Ponto BR
PNAD Contínua	Pesquisa Nacional por Amostra de Domicílios Contínua
QDPs	Quantidades Descritivas Populacionais
R	Software Estatístico R
SAS	Statistical Analysis System
SINAPE	Simpósio Nacional de Probabilidade e Estatística
SPSS	Statistical Package for the Social Science
STATA	Software for Statistics and Data Science
TCL	Teorema Central do Limite
TIC	Tecnologias de Informação e Comunicação
TICDOM	Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros
UPAs	Unidades Primárias de Amostragem

Lista de Tabelas

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 de 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. conglomeração 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 conglomeração, 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 1.1 considera o uso de dados de uma pesquisa amostral real, realizada pelo Núcleo de Informação e Coordenação do Ponto BR - NIC.br, 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. NIC.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 de 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

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.

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

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, dividindo-se o maior peso pelo menor encontra-se uma razão da ordem de 66 mil. Os pesos também variam bastante entre as regiões, sendo a razão entre as medianas dos pesos das regiões Norte e Sul igual a 1,65 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 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 diferentes das unidades.**

Na Tabela 1.3 apresentamos as estimativas dos seguintes parâmetros populacionais: porcentagem de domicílios com computador de mesa; porcentagem de domicílios com notebook; porcentagem de domicílios com tablete; porcentagem de domicílios com algum computador (de mesa, notebook ou tablete); total de domicílios com algum computador (de mesa, notebook ou tablete); número médio de computadores por domicílio que tem computador.

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 (2021). 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 1.3.

Tabela 1.3: Estimativas de parâmetros populacionais e EPAs

Parâmetro	Est_por_AAS	Est_Verd	DP_Est_Verd	EPA
Porcentagem de domicílios com computador de mesa	14,21	16,17	0,46	3,64
Porcentagem de domicílios com notebook	22,84	26,05	0,66	5,30
Porcentagem de domicílios com tablete	11,24	12,95	0,36	2,64
Porcentagem de domicílios com computador	35,34	39,36	0,67	4,38
Total de domicílios com computador (milhões)	25,10	27,95	1,37	36,90
Número médio de computadores por domicílio que tem computador	1,55	1,63	0,02	3,73

Os resultados mostram que há diferenças entre as estimativas pontuais dos parâmetros considerados, com uma tendência de subestimar quando se ignoram 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 padrão.

Note que as variáveis e parâmetros cujas estimativas foram apresentadas na Tabela 1.3 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 e 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, Pessoa e Silva (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 2, 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 Silva et al. (2020), o Capítulo 1 em Skinner et al. (1989) e os Capítulos 1 e 2 em Chambers e 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 2.

No Capítulo 3 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 4 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 e 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 Silva et al. (2020).

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 (2021), para ajustar modelos de regressão linear. As referências centrais são o Capítulo 6 em Silva et al. (2020) 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 (2021), para ajustar modelos de regressão logística. As referências centrais são o Capítulo 6 em Silva et al. (2020) 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 em Lehtonen e 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 doze 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. A seguir, a lista dos dez capítulos da parte 2.

Capítulo 10 - Gráficos

Capítulo 11 - Estimação de funções de densidade

Capítulo 12 - Estimação de funções de distribuição e quantis

Capítulo 13 - Estimação de medidas de desigualdade e pobreza

Capítulo 14 - Estimação de fluxos

Capítulo 15 - Modelos multiníveis

Capítulo 16 - Modelos para dados longitudinais

Capítulo 17 - Modelos de teoria da resposta ao item

Capítulo 18 - Modelos de séries temporais

Capítulo 19 - Modelos de redes neurais

Capítulo 20 - Modelos log-lineares para tabelas

Capítulo 21 - Aplicações

O Capítulo **??** aborda a elaboração de alguns tipos de gráficos de uso frequente quando os dados elementares provêm 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 de fluxos em pesquisas repetidas sujeitas a não resposta. Apresenta os recursos do pacote `surf` (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.

O Capítulo **??** trata do ajuste de modelos para dados longitudinais.

O Capítulo **??** trata do ajuste de modelos da Teoria da Resposta ao Item (TRI).

O Capítulo **??** trata do ajuste de modelos séries temporais a dados de pesquisas amostrais repetidas.

O Capítulo **??** trata do ajuste de modelos de redes neurais.

O Capítulo **??** trata do ajuste de modelos log-lineares a dados de tabelas de contingência.

O Capítulo **??** apresenta algumas aplicações de modelos e métodos descritos em capítulos anteriores no contexto de pesquisas sobre TICs no Brasil.

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 e Skinner (2003) são as referências centrais a consultar. Para aqueles querendo um tratamento ainda mais prático que o nosso, os livros de Lehtonen e 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.

Capítulo 2

Referencial para Inferência

2.1 Modelagem - Primeiras ideias

Com o objetivo de dar uma primeira ideia sobre o assunto a ser tratado neste livro vamos considerar, em situações simples, algumas abordagens alternativas para modelagem e análise estatística. A ideia é apresentar a principal abordagem que vamos considerar, a de *Modelagem de Superpopulação*, em contraste com as alternativas que poderiam ser consideradas, mas que tornariam difícil incorporar adequadamente as características que diferenciam dados obtidos com amostras complexas de outros.

2.2 Abordagem 1 - Modelagem Clássica

Seja y uma variável de pesquisa (ou de interesse), e sejam n observações desta variável para uma amostra de unidades de pesquisa denotadas por $y_1, \ldots, y_n$. Em Inferência Estatística, a abordagem que aqui chamamos de *Modelagem Clássica*
considera $y_1, \ldots, y_n$ como valores (realizações) de variáveis aleatórias $Y_1, \ldots, Y_n$.

Podemos formular modelos bastante sofisticados para a distribuição conjunta destas variáveis aleatórias, mas para simplificar a discussão, vamos inicialmente supor que $Y_1, \ldots, Y_n$ são variáveis aleatórias independentes e identicamente distribuídas - IID, com a mesma distribuição caracterizada pela função de densidade ou de frequência $f(y;\theta)$, onde $\theta \in \Theta$ é o parâmetro (um vetor de dimensão $K \times 1$) indexador da distribuição f, e Θ é o espaço paramétrico. A partir das observações $y_1, \ldots, y_n$, são feitas inferências a respeito do parâmetro θ.

Uma representação gráfica esquemática dessa abordagem é apresentada na Figura 2.1, e uma descrição esquemática resumida é apresentada na Tabela 2.1.

Tabela 2.1: Representação esquemática da abordagem *Modelagem Clássica*

Dados Amostrais (observações)	$y_1, \ldots, y_n$
Modelo Paramétrico/ Hipóteses	$Y_1, \ldots, Y_n$ variáveis aleatórias IID com distribuição $f(y, \theta)$ onde $\theta \in \Theta$
Objetivo	Inferir sobre θ usando as observações $y_1, \ldots, y_n$

Do ponto de vista matemático, o parâmetro θ serve para indexar os elementos da família de distribuições

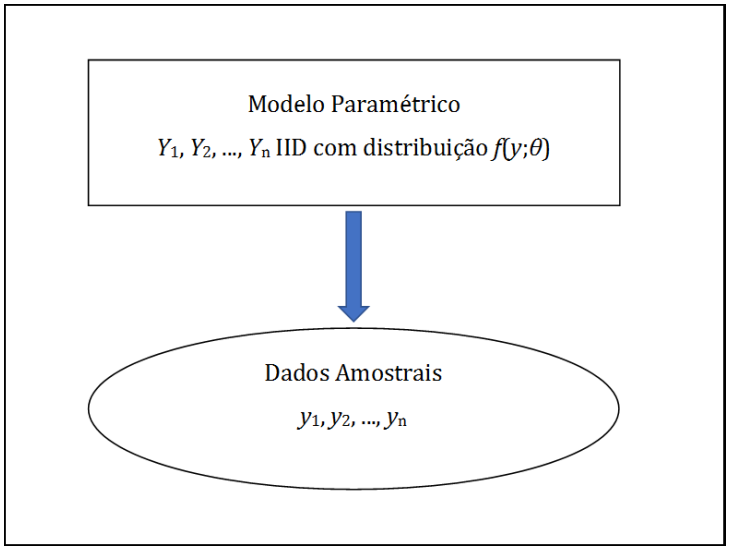

Figura 2.1: Representação esquemática da *Modelagem Clássica*

$\{f(y;\theta); \theta \in \Theta\}$. Na prática, as questões relevantes da pesquisa são traduzidas em termos de perguntas sobre o valor ou região a que pertence o parâmetro θ, e a inferência sobre θ a partir dos dados ajuda a responder tais questões.

Esta abordagem é útil em estudos analíticos tais como, por exemplo, na investigação da natureza da associação entre variáveis (modelos de regressão linear ou logística, modelos log-lineares, etc.). Vários exemplos discutidos ao longo dos Capítulos **??**, **??** e **??** ilustram situações deste tipo. No Capítulo **??** o foco é a estimação não paramétrica da forma da função $f(y;\theta)$.

Inferência sob modelos do tipo descrito nesta seção forma o conteúdo de um curso introdutório de inferência estatística. Mais detalhes podem ser consultados, por exemplo, em Casella e Berger (2010) e Magalhães e Lima (2015).

Exemplo 2.1. Estimação da proporção de sucessos em ensaios de Bernoulli

Para dar um exemplo concreto de modelagem do tipo descrito aqui, considere uma sequência de n ensaios de Bernoulli, em que a cada ensaio a resposta é o indicador de ocorrência de um evento de interesse - por exemplo, o indivíduo amostrado já foi vacinado contra uma doença especificada.

Se considerarmos que os resultados desses ensaios podem ser modelados como uma sequência de variáveis aleatórias $Y_1, \ldots, Y_n$ IID, com distribuição de Bernoulli dada por $f(y;\theta) = \theta^y \times (1-\theta)^{(1-y)}$, com $\theta \in (0;1)$, podemos usar a amostra observada $y_1, \ldots, y_n$ para fazer inferência sobre θ.

Sob o modelo especificado, é fácil deduzir que $T = \sum_{i=1}^{n} Y_i$ tem distribuição Binomial de parâmetros (n, θ). Logo, $\overline{T} = T/n$ tem média θ. Portanto, considerando o método dos momentos, $\overline{T}$ pode ser usado para estimar o parâmetro de interesse θ. Ademais, como n é conhecido, sabemos que a variância da sua distribuição de probabilidades é dada por $\theta \times (1-\theta)/n$, podendo ser estimada sem viés usando $\overline{T} \times (1-\overline{T})/(n-1)$.

Para amostras de tamanho grande ($n \to \infty$), podemos usar o Teorema Central do Limite - TCL para obter intervalos de confiança de nível especificado para θ e também testar hipóteses sobre regiões de interesse.

2.3 Abordagem 2 - Amostragem Probabilística

A abordagem adotada pelos praticantes de *Amostragem Probabilística* (amostristas) considera uma população finita $U = \{1, \ldots, N\}$, da qual é selecionada uma amostra $s = \{i_1, \ldots, i_n\}$, segundo um plano amostral caracterizado por $p(s)$, probabilidade de ser selecionada a amostra s, suposta calculável para todas as possíveis amostras. Os valores $y_1, \ldots, y_N$ da variável de interesse y na *população finita* são considerados fixos, porém desconhecidos.

A partir dos valores observados na amostra s, denotados por $y_{i_1}, \ldots, y_{i_n}$, são feitas inferências a respeito de funções dos valores populacionais, digamos $g(y_1, \ldots, y_N)$. Os valores de tais funções são quantidades descritivas populacionais - QDPs, também denominadas *parâmetros da população finita* pelos amostristas.

Em geral, o objetivo desta abordagem é fazer estudos descritivos utilizando funções g particulares, tais como totais $g(y_1, \ldots, y_N) = \sum_{i=1}^{N} y_i$, médias $g(y_1, \ldots, y_N) = N^{-1} \sum_{i=1}^{N} y_i$, proporções, razões, etc. Uma descrição esquemática resumida dessa abordagem é apresentada na Tabela 2.2, e uma representação gráfica resumida na Figura 2.2.

Tabela 2.2: Representação esquemática da abordagem *Amostragem Probabilística*

Dados Amostrais	$y_{i_1}, \ldots, y_{i_n}$
Modelo / Hipóteses	Dados extraídos de $y_1, \ldots, y_N$ segundo $p(s)$
Objetivo	Inferir sobre funções $g(y_1, \ldots, y_N)$ usando $y_{i_1}, \ldots, y_{i_n}$

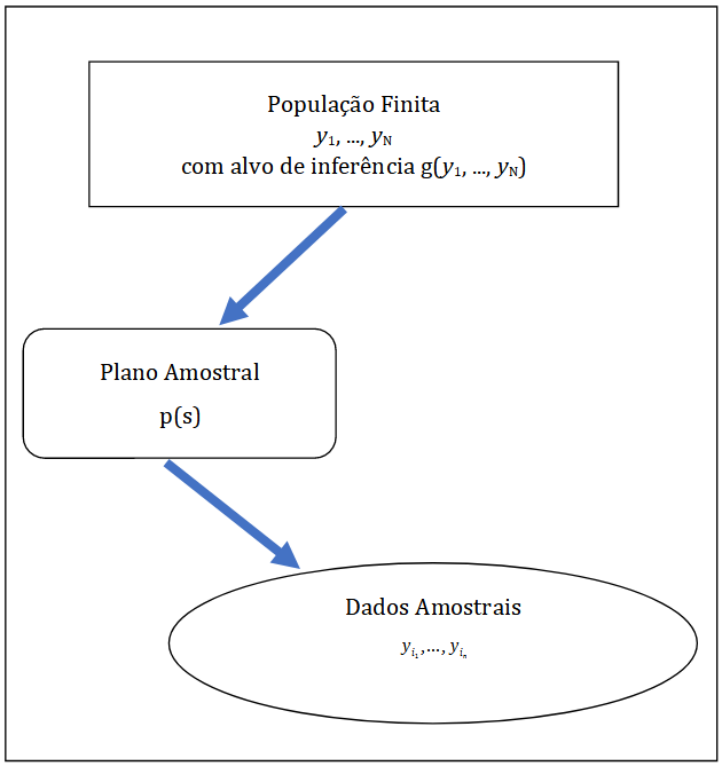

Figura 2.2: Representação esquemática da *Amostragem Probabilística*

Esta abordagem é largamente empregada na produção de estatísticas públicas e oficiais, por agências e instituições de muitos países. Uma das alegadas vantagens dessa abordagem é o fato de que as distribuições de referência usadas para inferência são controladas pelos amostristas que planejam as pesquisas por

amostragem e, portanto, a inferência pode ser considerada não paramétrica e não dependente de modelos que precisariam ser especificados pelo analista.

Uma revisão detalhada da amostragem probabilística pode ser encontrada em Silva et al. (2020). Nessa abordagem, a inferência é geralmente guiada também por distribuições dos estimadores aproximadas usando o TCL.

Exemplo 2.2. Estimação do total com amostragem estratificada simples

Considere o cenário de uma população U que foi estratificada em H grupos com base numa variável de estratificação x. Dos estratos formados, foram selecionadas de forma independente amostras aleatórias simples de tamanhos $n_1, \ldots, n_h, \ldots, n_H$. Nessa população, denotando por U_h o h-ésimo estrato, de tamanho N_h, o total populacional da variável de pesquisa y pode ser escrito como:

$$T_y = \sum_{h=1}^{H} \sum_{i \in U_h} y_i \tag{2.1}$$

O estimador padrão (tipo Horvitz-Thompson) para este parâmetro na amostragem estratificada simples é dado por:

$$\widehat{T}_y = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} y_i \tag{2.2}$$

onde s_h é a amostra das unidades do estrato $h, h = 1, 2, \ldots, H$.

Este estimador pode ser usado para fazer inferência sobre o total populacional, como descrito, por exemplo, na seção 11.2 de Silva et al. (2020). A distribuição do estimador $\widehat{T}_y$ obtida considerando o plano amostral $p(s)$ é denominada de *distribuição de aleatorização* e é geralmente aproximada usando o TCL para viabilizar a inferência.

2.4 Discussão das abordagens 1 e 2

A primeira abordagem (*Modelagem Clássica*), nos termos descritos, foi inicialmente proposta para dados de medidas na Física e Astronomia, onde em geral o pesquisador tem relativo controle sobre os experimentos, e onde faz sentido falar em replicação ou repetição do experimento. Neste contexto, a ideia de aleatoriedade é geralmente introduzida para modelar os erros (não controláveis) do processo de medição, e as distribuições de estatísticas de interesse são derivadas a partir da *distribuição do modelo* especificado.

A segunda abordagem (*Amostragem Probabilística*) é utilizada principalmente no contexto de estudos socioeconômicos observacionais, para levantamento de dados por agências produtoras de informações estatísticas públicas ou oficiais. Nesta abordagem, a aleatoriedade é introduzida pelo pesquisador no processo conduzido para obtenção dos dados, através do *plano amostral* $p(s)$ utilizado para selecionar as unidades de uma população finita U para observação ou medição, e as distribuições das estatísticas de interesse são derivadas a partir dessa *distribuição de aleatorização*.

Os planos amostrais podem ser complexos, gerando observações afetadas pelas características i) a iv) mencionadas no Capítulo 1. Os dados obtidos são utilizados principalmente para descrição da população finita, mediante o cálculo de estimativas de *parâmetros descritivos* usuais tais como totais, médias, proporções, razões, etc.

Sob a abordagem de *Amostragem Probabilística*, os pontos i) a iv) do Capítulo 1 são devidamente considerados tanto na estimação dos parâmetros descritivos como, também, na estimação de variâncias dos estimadores,

permitindo a inferência pontual e por intervalos de confiança baseada na distribuição assintótica normal dos estimadores habitualmente considerados.

A abordagem de *Amostragem Probabilística* é essencialmente não paramétrica, pois não supõe uma distribuição paramétrica particular para as observações da amostra. Por outro lado, essa abordagem tem a desvantagem de fazer inferências restritas à particular população finita considerada.

Apesar da abordagem de *Amostragem Probabilística* ter sido inicialmente concebida e aplicada para problemas de inferência descritiva sobre populações finitas, é cada vez mais comum, porém, a utilização dos dados obtidos através de pesquisas amostrais complexas para fins analíticos, com a aplicação de métodos de análise desenvolvidos e apropriados para a abordagem de *Modelagem Clássica*. Nesse contexto, é relevante considerar algumas questões de interesse:

- É adequado aplicar métodos de análise da *Modelagem Clássica*, concebidos para observações de variáveis aleatórias IID, aos dados obtidos através de pesquisas amostrais complexas?
- Em caso negativo, seria possível corrigir estes métodos, tornando-os aplicáveis para tratar dados amostrais complexos?
- Ou seria mais adequado fazer uso analítico dos dados dentro da abordagem de *Amostragem Probabilística*? E neste caso, como fazer isto, visto que nesta abordagem não é especificado um modelo para a distribuição das variáveis de pesquisa *na população*?

Além destas questões, também é de interesse a questão da *robustez da inferência*, traduzida nas seguintes perguntas:

- O que acontece quando o modelo adotado na *Modelagem Clássica* não é verdadeiro?
- Neste caso, qual a interpretação dos parâmetros na *Modelagem Clássica*?
- Ainda neste caso, as quantidades descritivas populacionais da *Amostragem Probabilística* poderiam ter alguma utilidade ou interpretação?

O objeto deste livro é exatamente discutir respostas para as questões aqui enumeradas. Para isso, vamos considerar uma abordagem que propõe um modelo parametrizado como na *Modelagem Clássica*, mas formulado para descrever os dados da *população*, e não os da amostra. Essa abordagem incorpora na análise os pontos i) a iii) do Capítulo 1 mediante aproveitamento da estrutura do plano amostral, como feito habitualmente na *Amostragem Probabilística*. Essa abordagem, denominada de *Modelagem de Superpopulação*, foi primeiro proposta em Brewer (1963) e Royall (1970), e é bem descrita, por exemplo, em Binder (1983) e Valliant et al. (2000).

2.5 Abordagem 3 - Modelagem de Superpopulação

Nesta abordagem, os valores $y_1, \ldots, y_N$ da variável de interesse y na população finita são considerados observações ou realizações das variáveis aleatórias $Y_1, \ldots, Y_N$, supostas IID com distribuição $f(y;\theta)$, onde $\theta \in \Theta$. Este modelo é denominado *Modelo de Superpopulação*. Note que, em contraste com o que se faz na *Modelagem Clássica*, o modelo probabilístico é aqui especificado para descrever o mecanismo aleatório que gera a *população*, não a amostra.

Na maioria das aplicações práticas, a população de interesse, embora considerada finita, jamais é observada por inteiro. Não obstante, ao formular o modelo para descrever propriedades da população, nossas perguntas e respostas descritas em termos de valores ou regiões para o parâmetro θ passam a se referir à população de interesse ou a populações similares, quer existam ao mesmo tempo, quer se refiram a estados futuros (ou passados) da mesma população. Vale realçar também que pesquisas por amostragem "consistem em

selecionar parte de uma população para observar, de modo que seja possível estimar alguma coisa sobre toda a população", conforme Thompson (1992).

Utilizando um plano amostral definido por $p(s)$, obtemos os valores das variáveis de pesquisa na amostra $y_{i_1}, \ldots, y_{i_n}$. A partir de $y_{i_1}, \ldots, y_{i_n}$, em geral não considerados como observações de vetores aleatórios IID, queremos fazer inferência sobre o parâmetro θ, considerando os pontos i) a iii) do Capítulo 1. Ver uma representação gráfica resumida desta abordagem na Figura 2.3.

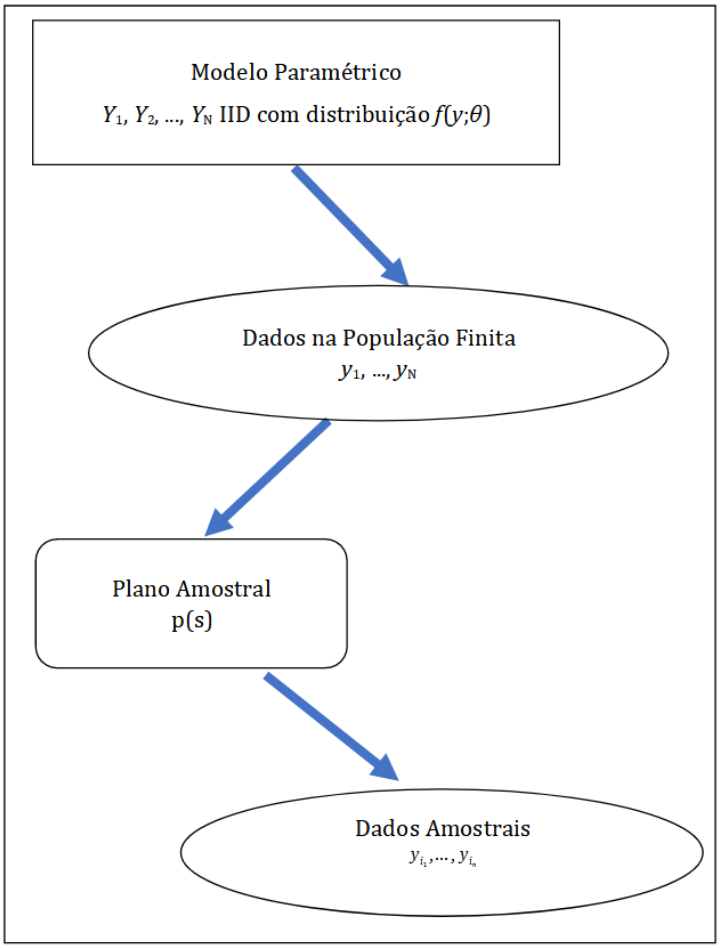

Figura 2.3: Representação esquemática da *Modelagem de Superpopulação*

Adotando o *Modelo de Superpopulação* e considerando métodos usuais disponíveis na *Modelagem Clássica*, podemos utilizar funções de $y_1, \ldots, y_N$, digamos $g(y_1, \ldots, y_N)$, para fazer inferência sobre θ. Desta forma, definimos estatísticas $g(y_1, \ldots, y_N)$ (no sentido da *Modelagem Clássica*) que são quantidades descritivas populacionais (parâmetros populacionais no contexto da *Amostragem Probabilística*), que passam a ser os novos parâmetros-alvo.

O passo seguinte é utilizar métodos disponíveis na *Amostragem Probabilística* para fazer inferência sobre $g(y_1, \ldots, y_N)$ com base nas observações (dados amostrais) $y_{i_1}, \ldots, y_{i_n}$. Note que não é possível basear a inferência nos valores populacionais $y_1, \ldots, y_N$, já que estes não são conhecidos ou observados. Este último passo adiciona a informação sobre o plano amostral utilizado, contida em $p(s)$, à informação estrutural contida no modelo $\{f(y;\theta); \theta \in \Theta\}$.

Uma representação esquemática dessa abordagem é apresentada na Tabela 2.3.

Tabela 2.3: Representação esquemática da *Modelagem de Superpopulação*

Dados Amostrais	$y_{i_1}, \ldots, y_{i_n}$
População e esquema de seleção	Selecionados de $y_1, \ldots, y_N$ segundo $p(s)$
Modelo para população	$Y_1, \ldots, Y_N$ variáveis aleatórias IID com distribuição $f(y, \theta)$, onde $\theta \in \Theta$
Parâmetro-alvo	Associar $\theta \Leftrightarrow g(Y_1, \ldots, Y_N)$
Objetivo	Inferir sobre $g(y_1, \ldots, y_N)$ partir de $y_{i_1}, \ldots, y_{i_n}$ usando $p(s)$

A descrição da abordagem adotada neste livro foi apresentada de maneira propositalmente simplificada e vaga nesta seção, mas é aprofundada ao longo do texto. Admitimos que o leitor esteja familiarizado com a *Modelagem Clássica* e com as noções básicas da *Amostragem Probabilística*.

A título de recordação, são apresentados na Seção 2.8 alguns resultados básicos da *Amostragem Probabilística*. A ênfase do texto, porém, é a apresentação da *Modelagem de Superpopulação*, sendo para isto apresentados os elementos indispensáveis das abordagens de *Modelagem Clássica* e da *Amostragem Probabilística*.

Ao construir e ajustar modelos a partir de dados de pesquisas amostrais *complexas*, tais como as executadas pelo IBGE e outras instituições similares, o usuário precisará incorporar as informações sobre pesos e sobre a estrutura dos planos amostrais utilizados para obtenção dos dados. Em geral, ao publicar os resultados das pesquisas, os pesos são considerados, sendo possível produzir estimativas pontuais *corretas* utilizando os pacotes computacionais tradicionais. Por outro lado, para construir intervalos de confiança e testar hipóteses sobre parâmetros de modelos, é necessário conhecer estimativas de variâncias e covariâncias das estimativas, obtidas levando em conta a estrutura do plano amostral utilizado.

Mesmo conhecendo o plano amostral, geralmente não é simples incorporar pesos e plano amostral na análise sem o uso de pacotes especializados, ou de rotinas específicas já agora disponíveis em alguns dos pacotes mais comumente utilizados (por exemplo, SAS, STATA, SPSS, ou R entre outros). Tais pacotes especializados ou rotinas específicas utilizam, em geral, métodos aproximados para estimar matrizes de covariância. Entre esses métodos, destacam-se o de Máxima Pseudo-Verossimilhança, a Linearização de Taylor, o método do Conglomerado Primário e métodos de reamostragem, que são descritos mais adiante.

Em outras palavras, o uso dos pacotes usuais para analisar dados produzidos por pesquisas com planos amostrais complexos, tal como o uso de muitos remédios, pode ter contraindicações. Cabe ao usuário *ler a bula* e identificar situações em que o uso de tais pacotes pode ser inadequado e buscar opções de rotinas específicas ou de pacotes especializados capazes de incorporar adequadamente a estrutura do plano amostral nas análises.

Ao longo deste livro fazemos uso intensivo do pacote *survey* e outros disponíveis no R, mas o leitor pode encontrar funcionalidade semelhante em alguns outros sistemas. Nossa escolha se deveu a dois fatores principais: primeiro ao fato do sistema R ser aberto, livre e gratuito, dispensando o usuário de custos de licenciamento, bem como possibilitando aos interessados o acesso ao código fonte e à capacidade de modificar as rotinas de análise, caso necessário. O segundo fator é de natureza mais técnica, porém transitória. No presente momento, o pacote *survey* do R é a coleção de rotinas mais completa e genérica existente para análise de dados amostrais complexos, dispondo de funções capazes de ajustar os modelos usuais, mas também de ajustar modelos não convencionais, mediante a maximização numérica de verossimilhanças especificadas pelo usuário.

Sabemos, entretanto, que muitos usuários habituados à facilidade de uso de pacotes com interfaces gráficas do tipo *aponte e clique* terão dificuldade adicional de adaptar-se à linguagem de comandos utilizada pelo

sistema R, mas acreditamos que os benefícios do aprendizado desta nova ferramenta compensarão largamente os custos adicionais do aprendizado.

O emprego de ferramentas de análise como o pacote *survey* permite aos usuários focar sua atenção mais na seleção, análise e interpretação dos modelos ajustados do que nas dificuldades técnicas envolvidas nos cálculos correspondentes. É com este espírito que escrevemos este texto, que busca apresentar os métodos, ilustrando seu uso com exemplos reais e orientando sobre o uso adequado das ferramentas de modelagem e análise disponíveis no sistema R.

2.6 Fontes de variação

Esta seção estabelece o referencial para inferência em pesquisas amostrais que é usado no restante deste texto. Cassel et al. (1977) sugerem que um referencial para inferência poderia considerar três fontes de aleatoriedade (incerteza, variação), incluindo:

1. *Modelo de Superpopulação*, que descreve o processo subjacente que, por hipótese, gera as medidas verdadeiras para todas as unidades da população considerada;

2. *Processo de Medição*, que diz respeito aos instrumentos e métodos usados para obter as medidas de qualquer unidade da população;

3. *Plano Amostral*, que estabelece o mecanismo pelo qual unidades da população são selecionadas para participar da amostra da pesquisa ou estudo.

Uma quarta fonte de incerteza que precisa ser acrescentada às anteriores é o

4. *Mecanismo de resposta*, ou seja, o mecanismo que controla se valores de medições de unidades selecionadas para a amostra são obtidos / observados ou não.

Para concentrar o foco nas questões de maior interesse deste texto, as fontes (2) e (4) não são consideradas no referencial adotado para a maior parte dos capítulos. Para o tratamento das dificuldades causadas por não resposta, a fonte (4) é considerada no Capítulo xx.

Assim sendo, exceto onde explicitamente indicado, de agora em diante admitiremos que não há *erros de medição*, implicando que os valores observados de quaisquer variáveis de interesse são considerados valores corretos ou verdadeiros. Admitimos ainda que há *resposta completa*, implicando que os valores de quaisquer variáveis de interesse estão disponíveis para todos os elementos da amostra selecionada depois que a pesquisa foi realizada. Hipóteses semelhantes são adotadas, por exemplo, em Binder (1983) e Montanari (1987).

Portanto, o referencial aqui adotado considera apenas duas fontes de variação: o *Modelo de Superpopulação* (1) e o *Plano Amostral* (3). Estas fontes de variação, descritas nesta seção apenas de forma esquemática, são discutidas com maiores detalhes a seguir.

A fonte de variação (1) é considerada porque usos analíticos das pesquisas são amplamente discutidos neste texto, os quais só têm sentido quando é especificado um modelo estocástico para o processo subjacente que gera as medidas na população. A fonte de variação (3) é considerada porque a atenção é focalizada na análise de dados obtidos através de pesquisas amostrais complexas. Aqui a discussão se restringe a planos amostrais aleatorizados ou de *Amostragem Probabilística*, não sendo considerados métodos intencionais ou outros métodos não aleatórios algumas vezes usados para seleção de amostras.

2.7 Modelos de Superpopulação

Seja $\{1, ..., N\}$ um conjunto de rótulos que identificam univocamente os N elementos distintos de uma população-alvo finita U. Sem perda de generalidade tomemos $U = \{1, ..., N\}$. Uma pesquisa cobrindo n elementos distintos numa amostra s, $s = \{i_1, ..., i_n\} \subset U$, é realizada para medir os valores de Q variáveis de interesse da pesquisa, doravante denominadas simplesmente *variáveis da pesquisa*.

Denotemos por $\mathbf{y}_i = (y_{i1}, ..., y_{iP})'$ um vetor $Q \times 1$ de valores das variáveis da pesquisa e por $\mathbf{x}_i = (x_{i1}, \ldots, x_{iQ})'$ um vetor $Q \times 1$ de variáveis auxiliares da i-ésima unidade da população, respectivamente, para $i = 1, ..., N$. Aqui as variáveis auxiliares são consideradas como variáveis contendo a informação requerida para o plano amostral e a estimação a partir da amostra, como se discute com mais detalhes adiante.

Denotemos por $\mathbf{y}_U$ a matriz $N \times Q$ formada empilhando os vetores transpostos das observações das variáveis de pesquisa correspondentes a todas as unidades da população, e por $\mathbf{Y}_U$ a correspondente matriz de vetores aleatórios geradores das observações na população.

Quando se supõe que $\mathbf{y}_1, \ldots, \mathbf{y}_N$ são a realização conjunta de vetores aleatórios $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, a distribuição conjunta de probabilidade de $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ é um *Modelo de Superpopulação* (marginal), que doravante denotaremos simplesmente por $f(\mathbf{y}_U; \theta)$, ou de forma abreviada, por M. Esperanças e variâncias definidas com respeito à distribuição do modelo M são denotadas E_M e V_M respectivamente.

Analogamente, $\mathbf{x}_1, \ldots, \mathbf{x}_N$ pode ser considerada uma realização conjunta de vetores aleatórios $\mathbf{X}_1, \ldots, \mathbf{X}_N$. As matrizes $N \times Q$ formadas empilhando os vetores transpostos das observações das variáveis auxiliares correspondentes a todas as unidades da população, $\mathbf{x}_U$, e a correspondente matriz $\mathbf{X}_U$ de vetores aleatórios geradores das variáveis auxiliares na população são definidas de forma análoga às matrizes $\mathbf{y}_U$ e $\mathbf{Y}_U$.

O referencial aqui adotado permite a especificação da distribuição conjunta combinada das variáveis da pesquisa e das variáveis auxiliares. Representamos por $f(\mathbf{y}_U, \mathbf{x}_U; \eta)$ a função de densidade de probabilidade conjunta de $(\mathbf{Y}_U, \mathbf{X}_U)$, onde η é um vetor de parâmetros.

Um tipo importante de modelo de superpopulação é obtido quando os vetores aleatórios correspondentes às observações de unidades diferentes da população são supostos independentes e identicamente distribuídos - IID. Neste caso, o modelo de superpopulação pode ser escrito como:

$$\begin{aligned} f(\mathbf{y}_U, \mathbf{x}_U; \eta) &= \prod_{i \in U} f(\mathbf{y}_i, \mathbf{x}_i; \eta) \\ &= \prod_{i \in U} f(\mathbf{y}_i | \mathbf{x}_i; \lambda) f(\mathbf{x}_i; \phi) \end{aligned} \tag{2.3}$$

onde λ e ϕ são vetores de parâmetros.

Sob (2.3), o modelo marginal correspondente das variáveis da pesquisa seria obtido integrando nas variáveis auxiliares:

$$f(\mathbf{y}_U; \theta) = f(\mathbf{y}_1, \ldots, \mathbf{y}_N; \theta) = \prod_{i \in U} \int f(\mathbf{y}_i | \mathbf{x}_i; \lambda) f(\mathbf{x}_i; \phi) \, \mathbf{dx}_i = \prod_{i \in U} f(\mathbf{y}_i; \theta) \tag{2.4}$$

onde $f(\mathbf{y}_i; \theta) = \int f(\mathbf{y}_i | \mathbf{x}_i; \lambda) f(\mathbf{x}_i; \phi) \, \mathbf{dx}_i$ e $\theta = h(\lambda, \phi)$ é função de λ e ϕ.

Outro tipo especial de modelo de superpopulação é o modelo de população fixa, que supõe que os valores numa população finita são fixos mas desconhecidos. Este modelo pode ser descrito por:

$$P\left[(\mathbf{Y}_U, \mathbf{X}_U) = (\mathbf{y}_U, \mathbf{x}_U)\right] = 1 \tag{2.5}$$

ou seja, uma distribuição degenerada é especificada para $(\mathbf{Y}_U, \mathbf{X}_U)$.

Este modelo foi considerado em Cassel et al. (1977), que o chamaram de *abordagem de população fixa*, e afirmaram ser esta a abordagem subjacente ao desenvolvimento da teoria da *Amostragem Probabilística* encontrada nos livros clássicos tais como Cochran (1977) e outros.

Aqui esta abordagem é chamada de *abordagem baseada no plano amostral* ou *abordagem de aleatorização*, pois neste caso a única fonte de variação (aleatoriedade) é proveniente do plano amostral. Em geral, a distribuição conjunta de $(\mathbf{Y}_U, \mathbf{X}_U)$ não precisa ser degenerada como especificada em (2.5), embora o referencial aqui adotado seja suficientemente geral para permitir considerar esta possibilidade.

Se todas as unidades da população U fossem pesquisadas (ou seja, se fosse executado um *censo*), os dados observados seriam $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N)$. Sob a hipótese de resposta completa, a única fonte de incerteza seria devida ao fato de que $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N)$ é uma realização de $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_N, \mathbf{X}_N)$. Os dados observados poderiam então ser usados para fazer inferências sobre η, ϕ, λ ou θ usando procedimentos padrões.

Inferência sobre quaisquer dos parâmetros η, ϕ, λ ou θ do modelo de superpopulação é chamada *inferência analítica*. Este tipo de inferência só faz sentido quando o modelo de superpopulação não é degenerado como em (2.5). Usualmente seu objetivo é explicar a relação entre variáveis não apenas para a população finita sob análise, mas também para outras populações que poderiam ter sido geradas pelo modelo de superpopulação adotado. Vários exemplos de inferência analítica são discutidos ao longo deste livro.

Se o objetivo da inferência é estimar quantidades que fazem sentido somente para a população finita sob análise, tais como funções $g(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ dos valores das variáveis da pesquisa, o modelo de superpopulação não é estritamente necessário, embora possa ser útil. Inferência para tais quantidades, chamadas parâmetros da população finita ou quantidades descritivas populacionais - QDPs, é chamada *inferência descritiva*.

Vale notar que a especificação do modelo de superpopulação aqui proposta serve tanto para o caso da abordagem clássica para inferência, como também para o caso da abordagem Bayesiana. Neste caso, a especificação do modelo M precisaria ser completada mediante a especificação de distribuições *a priori* para os parâmetros do modelo.

2.8 Plano amostral

Embora *censos* sejam algumas vezes realizados para coletar dados sobre certas populações, a vasta maioria das pesquisas realizadas é de pesquisas amostrais, nas quais apenas uma amostra de elementos da população (usualmente uma pequena parte) é investigada. Neste caso, os dados disponíveis incluem:

1. O conjunto de rótulos $s = \{i_1, \ldots, i_n\}$ dos distintos elementos na amostra, onde n, $1 \leq n \leq N$, é o número de elementos na amostra s, também chamado de *tamanho da amostra*.
2. Os valores na amostra das variáveis da pesquisa $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_n}$.
3. Os valores das variáveis auxiliares na população $\mathbf{x}_1, \ldots, \mathbf{x}_N$, quando a informação auxiliar é dita *completa*; alternativamente, os valores das variáveis auxiliares na amostra $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}$, mais os totais ou médias destas variáveis na população, quando a informação auxiliar é dita *parcial*.

O mecanismo usado para selecionar a amostra s da população finita U é chamado *plano amostral*. Uma forma de caracterizá-lo é através da função $p(.)$, onde $p(s)$ dá a probabilidade de selecionar a amostra s no conjunto S de todas as amostras possíveis. Só mecanismos amostrais envolvendo alguma forma de

seleção probabilística bem definida são aqui considerados. Portanto, supõe-se que $0 \leq p(s) \leq 1 \; \forall \, s \in S$ e $\sum_{s \in S} p(s) = 1$.

Esta caracterização do plano amostral $p(s)$ é bem geral, permitindo que o mecanismo de seleção amostral dependa dos valores das variáveis auxiliares $\mathbf{x}_1, \ldots, \mathbf{x}_N$ bem como dos valores das variáveis da pesquisa na população $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (*amostragem informativa*, ver Seção 2.9). Uma notação mais explícita para indicar esta possibilidade envolveria escrever $p(s)$ como $p\left[s|(\mathbf{y}_U, \mathbf{x}_U)\right]$. Tal notação é evitada por razões de simplicidade.

Denotamos por $I(B)$ a função indicadora que assume o valor 1 quando o evento B ocorre e 0 caso contrário. Seja $\boldsymbol{\Delta}_s = [I(1 \in s), \ldots, I(N \in s)]'$ um vetor aleatório de indicadores dos elementos incluídos na amostra s. Então o plano amostral pode ser alternativamente caracterizado pela distribuição de probabilidade de $\boldsymbol{\Delta}_s$ denotada por $f\left[\delta_s|\left(\mathbf{y}_U, \mathbf{x}_U\right)\right]$, onde δ_s é qualquer realização particular de $\boldsymbol{\Delta}_s$ tal que ${\delta_s}'\mathbf{1}_N = n$, e $\mathbf{1}_N$ é o vetor unitário de dimensão N.

Notação adicional necessária nas seções posteriores é agora introduzida. Denotamos por π_i a probabilidade de inclusão da unidade i na amostra s, isto é,

$$\pi_i = P\left(i \in s\right) = \sum_{s \ni i} p(s) \tag{2.6}$$

e denotamos por π_{ij} a probabilidade de inclusão conjunta na amostra s das unidades i e j, dada por

$$\pi_{ij} = P\left(i \in s, j \in s\right) = \sum_{s \ni i,j} p(s) \tag{2.7}$$

para todo $i \neq j \in U$. Note que $\pi_{ii} = \pi_i \; \forall \, i \in U$.

Uma hipótese básica assumida com relação aos planos amostrais aqui considerados é que $\pi_i > 0$ e $\pi_{ij} > 0$ $\forall \, i, j \in U$. A hipótese de π_{ij} ser positiva é adotada para simplificar a apresentação de expressões para estimadores de variância dos estimadores dos parâmetros de interesse. Contudo, esta não é uma hipótese crucial, pois há planos amostrais que não a satisfazem e para os quais estão disponíveis aproximações e estimadores satisfatórios das variâncias dos estimadores de totais e de médias.

2.9 Planos amostrais informativos e ignoráveis

Ao fazer inferência usando dados de pesquisas amostrais precisamos distinguir duas situações que requerem tratamentos diferentes. Uma dessas situações ocorre quando o plano amostral empregado para coletar os dados é *informativo*, isto é, quando o mecanismo de seleção das unidades amostrais pode depender dos valores das variáveis de pesquisa.

Um exemplo típico desta situação é o dos *estudos de caso-controle*, em que a amostra é selecionada de tal forma que há *casos* (unidades com determinada condição) e *controles* (unidades sem essa condição), sendo de interesse a modelagem do indicador de presença ou ausência da condição em função de variáveis preditoras, sendo esse indicador uma das variáveis de pesquisa, que é considerada no mecanismo de seleção da amostra. Os métodos que discutimos ao longo deste livro não são adequados, em geral, para esse tipo de situação e, portanto, uma hipótese fundamental adotada ao longo deste texto é que os planos amostrais considerados são *não informativos*, isto é, não podem depender diretamente dos valores das variáveis da pesquisa. Logo eles satisfazem:

$$f\left[\delta_s|\left(\mathbf{y}_U, \mathbf{x}_U\right)\right] = f\left(\delta_s|\mathbf{x}_U\right) \tag{2.8}$$

Entre os planos amostrais *não informativos*, precisamos ainda distinguir duas outras situações de interesse. Quando o plano amostral é Amostragem Aleatória Simples Com Reposição - AASC, o modelo adotado para a amostra é o mesmo que o modelo adotado para a população antes da amostragem. Quando isto ocorre, o plano amostral é dito *ignorável*, porque a inferência baseada na amostra utilizando a abordagem de *Modelagem Clássica* descrita na Seção 2.2 pode prosseguir sem problemas. Entretanto, esquemas amostrais desse tipo são raramente empregados na prática, por razões de eficiência e custo. Em vez disso, são geralmente empregados planos amostrais envolvendo estratificação, conglomeração e probabilidades desiguais de seleção (*amostragem complexa*).

Com amostragem complexa, porém, os modelos para a população e a amostra podem ser muito diferentes (plano amostral *não ignorável*), mesmo que o mecanismo de seleção não dependa das variáveis de pesquisa, mas somente das variáveis auxiliares. Neste caso, ignorar o plano amostral pode viciar a inferência. Ver o Exemplo 2.3 adiante.

A definição precisa de ignorabilidade e as condições sob as quais um plano amostral é *ignorável* para inferência são bastante discutidas na literatura - ver por exemplo Sugden e Smith (1984) ou os Capítulos 1 e 2 de Chambers e Skinner (2003). Porém, testar a ignorabilidade do plano amostral é muitas vezes complicado. Em caso de dificuldade, o uso dos *pesos amostrais* tem papel fundamental, como se vê mais adiante.

Uma forma simples de lidar com os efeitos do plano amostral na estimação pontual de quantidades descritivas populacionais de interesse é incorporar pesos adequados na análise, como pode ser visto em Silva et al. (2020) e no Capítulo 3. Essa forma porém, não resolve por si só o problema de estimação da precisão das estimativas pontuais, nem mesmo o caso da estimação pontual de parâmetros em *modelos de superpopulação*, o que vai requerer métodos específicos discutidos no Capítulo **??**.

Como incluir os pesos para proteger contra planos amostrais *não ignoráveis* e a possibilidade de má especificação do modelo? Uma ideia é modificar os estimadores dos parâmetros de modo que sejam consistentes (em termos da *distribuição de aleatorização*) para quantidades descritivas da população finita da qual a amostra foi extraída, que por sua vez seriam boas aproximações para os parâmetros dos modelos de interesse. Afirmações probabilísticas são então feitas com respeito à *distribuição de aleatorização p* das estatísticas amostrais ou com respeito à distribuição mista ou combinada Mp.

A seguir apresentamos um exemplo com a finalidade de ilustrar uma situação de plano amostral *não ignorável*.

Exemplo 2.3. Efeito da amostragem estratificada simples com alocação desproporcional

Considere N observações de uma população finita U onde são consideradas de interesse duas variáveis binárias $(x_i; y_i)$. Suponha que na população os vetores aleatórios $(X_i; Y_i)$ são independentes e identicamente distribuídos com distribuição de probabilidades conjunta dada por:

Tabela 2.4: Distribuição de probabilidades conjunta na população $P(Y_i = y; X_i = x)$

$x \downarrow \mid y \rightarrow$	0	1	Total
0	η_{00}	η_{01}	η_{0+}
1	η_{10}	η_{11}	η_{1+}
Total	η_{+0}	η_{+1}	1

que também pode ser representada por:

$$\begin{aligned} f_U(x;y) &= P(X=x;Y=y) \\ &= \eta_{00}^{(1-x)(1-y)} \times \eta_{01}^{(1-x)y} \times \eta_{10}^{x(1-y)} \times (1-\eta_{00}-\eta_{01}-\eta_{10})^{xy} \end{aligned}$$

onde a designação f_U é utilizada para denotar a distribuição *na população*.

Note agora que a distribuição marginal da variável Y *na população* é Bernoulli com parâmetro $1-\eta_{00}-\eta_{10}$, ou alternativamente:

$$f_U(y) = P(Y=y) = (\eta_{00}+\eta_{10})^{(1-y)} \times (1-\eta_{00}-\eta_{10})^{y}$$

De forma análoga, a distribuição marginal da variável X *na população* também é Bernoulli, mas com parâmetro $1-\eta_{00}-\eta_{01}$, ou alternativamente:

$$f_U(x) = P(X=x) = (\eta_{00}+\eta_{01})^{(1-x)} \times (1-\eta_{00}-\eta_{01})^{x}$$

Seja N_{xy} o número de unidades na população com a combinação de valores observados $(x;y)$, onde x e y tomam valores em $\Omega = \{0;1\}$. É fácil notar então que o vetor de contagens populacionais $\mathbf{N} = (N_{00}, N_{01}, N_{10}, N_{11})'$ tem distribuição Multinomial com parâmetros N e $\eta = (\eta_{00}, \eta_{01}, \eta_{10}, 1-\eta_{00}-\eta_{01}-\eta_{10})'$.

Após observada uma realização do modelo que dê origem a uma população, como seria o caso da realização de um *censo* na população, a proporção de valores de y iguais a 1 observada no censo seria dada por $N_{+1}/N = 1-(N_{00}-N_{10})/N$. E a proporção de valores de x iguais a 1 na população seria igual a $N_{1+}/N = 1-(N_{00}-N_{01})/N$.

Agora suponha que uma amostra estratificada simples *com reposição* de tamanho n inteiro e par seja selecionada da população, onde os estratos são definidos com base nos valores da variável x, e onde a alocação da amostra nos estratos é dada por $n_0 = n_1 = n/2$, sendo n_x o tamanho da amostra no estrato correspondente ao valor x usado como índice. Esta alocação é dita *alocação igual*, pois o tamanho total da amostra é repartido em partes iguais entre os estratos definidos para seleção e, no caso, há apenas dois estratos. A alocação desta amostra é desproporcional exceto no caso em que $N_{0+} = N_{1+}$.

Nosso interesse aqui é ilustrar o efeito que uma *alocação desproporcional* pode causar na análise dos dados amostrais, caso não sejam levadas em conta na análise informações relevantes sobre a estrutura do plano amostral. Para isto, vamos precisar obter a *distribuição amostral* da variável de interesse Y. Isto pode ser feito em dois passos. Primeiro, note que a distribuição condicional de Y dado X *na população* é dada por:

Tabela 2.5: Distribuição de probabilidades condicional de y dado x na população - $P(Y_i = y|X_i = x)$

$x\downarrow \mid y\rightarrow$	0	1	Total
0	η_{00}/η_{0+}	η_{01}/η_{0+}	1
1	η_{10}/η_{1+}	η_{11}/η_{1+}	1

ou, alternativamente

$$\begin{aligned} f_U(y|x) &= P(Y=y|X=x) \\ &= (1-x) \times \frac{\eta_{00}^{(1-y)}\eta_{01}^{y}}{\eta_{00}+\eta_{01}} + x \times \frac{\eta_{10}^{(1-y)}(1-\eta_{00}-\eta_{01}-\eta_{10})^{y}}{1-\eta_{00}-\eta_{01}} \end{aligned}$$

Dado o plano amostral acima descrito, a distribuição marginal de X *na amostra* é Bernoulli com parâmetro 1/2. Isto segue devido ao fato de que a amostra foi alocada igualmente com base nos valores de x na população e, portanto, sempre teremos metade da amostra com valores de x iguais a 0 e metade com valores iguais a 1. Isto pode ser representado como:

$$f_s(x) = P(X_i = x|i \in s) = 1/2, \ \forall x \in \Omega \text{ e } \forall i \in U$$

onde a designação f_s é utilizada para denotar a distribuição *na amostra*.

Podemos usar a informação sobre a distribuição condicional de Y dado X *na população* e a informação sobre a distribuição marginal de X *na amostra* para obter a distribuição marginal de Y *na amostra*, que é dada por:

$$\begin{aligned} f_s(y) &= P(Y_i = y|i \in s) \\ &= \sum_{x=0}^{1} P(X_i = x; Y_i = y|i \in s) \\ &= \sum_{x=0}^{1} P[Y_i = y|(X_i = x)e(i \in s)] \times P(X_i = x|i \in s) \\ &= \sum_{x=0}^{1} P(Y_i = y|X_i = x) \times f_s(x) \\ &= \sum_{x=0}^{1} f_U(y|x) f_s(x) \\ &= \frac{1}{2} \times \left[\frac{\eta_{00}^{(1-y)}\eta_{01}^{y}}{\eta_{00}+\eta_{01}} + \frac{\eta_{10}^{(1-y)}(1-\eta_{00}-\eta_{01}-\eta_{10})^{y}}{1-\eta_{00}-\eta_{01}} \right] \end{aligned}$$

Isto mostra que a distribuição marginal de Y *na amostra* é diferente da distribuição marginal de Y *na população*, mesmo quando o plano amostral é especialmente simples e utiliza amostragem aleatória simples com reposição dentro de cada estrato definido pela variável X. Isto ocorre devido à *alocação desproporcional* da amostra, apesar de a distribuição condicional de Y dado X *na população* ser a mesma que a distribuição condicional de Y dado X *na amostra*.

Um exemplo numérico facilita a compreensão. Se a distribuição conjunta de X e Y na população é dada por:

Tabela 2.6: Distribuição de probabilidades conjunta na população $f_U(x; y)$

$x \downarrow \mid y \rightarrow$	0	1	Total
0	0,7	0,1	0,8
1	0,1	0,1	0,2
Total	0,8	0,2	1

segue-se que a distribuição condicional de Y dado X *na população* (e também *na amostra*) é dada por

Tabela 2.7: Distribuição de probabilidades condicional de Y dado X na população - $f_U(y|x)$

$x \downarrow \mid y \rightarrow$	0	1	Total
0	0,875	0,125	1
1	0,500	0,500	1

e que a distribuição marginal de Y *na população* e *na amostra* são dadas por

Tabela 2.8: Distribuição de probabilidades marginal de Y na população e na amostra - $f_U(y)$ e $f_s(y)$

y	0	1
$f_U(y)$	0,8000	0,2000
$f_s(y)$	0,6875	0,3125

Assim, inferência sobre a distribuição de Y *na população* levada a cabo a partir dos dados da *amostra observada* sem considerar a estrutura do plano amostral seria equivocada, pois a alocação igual da amostra nos estratos levaria à observação de uma proporção maior de valores de X iguais a 1 na amostra (1/2) do que a correspondente proporção existente na população (1/5). Em consequência, a proporção de valores de Y iguais a 1 na amostra (0,3125) seria 56% maior que a correspondente proporção *na população* (0,2).

Este exemplo é propositalmente simples, envolve apenas duas variáveis com distribuição Bernoulli, mas ilustra bem como a amostragem pode modificar distribuições de variáveis *na amostra* em relação à correspondente distribuição *na população*. Isto ocorre mesmo em situações em que a amostragem não é *informativa* (pois a seleção da amostra não depende dos valores da variável y), mas onde o plano amostral não é *ignorável* para inferência sobre a distribuição marginal de Y.

Caso a inferência requerida fosse sobre parâmetros da distribuição condicional de Y dado X, a amostragem seria *ignorável*, isto é, $f_s(y|x) = f_U(y|x)$. Assim, fica evidenciado também que a questão de saber se o plano amostral pode ou não ser ignorado depende da inferência desejada. No nosso exemplo, o plano amostral seria ignorável para inferência sobre a distribuição condicional de Y dado X, mas não seria ignorável para inferência sobre a distribuição marginal de Y.

Feita esta discussão sobre o referencial para inferência que adotamos neste livro, segue-se uma revisão rápida dos métodos de estimação da *amostragem probabilística*. Como já indicado, o leitor interessado numa discussão mais detalhada pode consultar Silva et al. (2020).

Capítulo 3

Estimação Baseada no Plano Amostral

3.1 Estimação de totais

Devido a sua importância para os desenvolvimentos teóricos em vários dos capítulos subsequentes, alguns resultados básicos relativos à estimação de totais da população finita numa abordagem baseada no plano amostral são relembrados nesta seção. A referência básica usada foi a Seção 2.8 de Särndal et al. (1992). O leitor pode também consultar o Capítulo 3 de Silva et al. (2020).

Consideremos o problema de estimar o vetor $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$ de totais das Q variáveis da pesquisa na população, a partir de uma amostra observada s. Naturalmente, qualquer estimador viável do total $\mathbf{Y}$ só pode depender dos valores das variáveis de pesquisa observados na amostra, contidos em $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_n}$, mas não dos valores dessas variáveis para os elementos não pesquisados ($i \in U - s$).

Um estimador usual baseado no plano amostral para o total $\mathbf{Y}$ é o estimador de Horvitz-Thompson (ver Capítulo 2 deste livro e Seção 3.7 de Silva et al. (2020)), dado por:

$$\widehat{\mathbf{Y}}_{HT} = \sum_{i \in s} \mathbf{y}_i / \pi_i = \sum_{i \in s} d_i \mathbf{y}_i \tag{3.1}$$

onde $d_i = 1/\pi_i$ é o *peso básico* da unidade i.

Na abordagem baseada no planejamento amostral, as propriedades de uma estatística ou estimador são avaliadas com respeito a sua *distribuição de aleatorização*. Denotemos por $E_p(.)$ e $V_p(.)$ os operadores de esperança e variância referentes à distribuição de probabilidades induzida pelo planejamento amostral $p(s)$, que chamaremos daqui por diante de *esperança de aleatorização* e *variância de aleatorização*.

O estimador $\widehat{\mathbf{Y}}_{HT}$ é não viciado para o total $\mathbf{Y}$ com respeito à distribuição de aleatorização, isto é:

$$E_p\left(\widehat{\mathbf{Y}}_{HT}\right) = \mathbf{Y}$$

Além disto, sua variância de aleatorização é dada por

$$V_p\left(\widehat{\mathbf{Y}}_{HT}\right) = \sum_{i \in U} \sum_{j \in U} \left(\frac{d_i d_j}{d_{ij}} - 1\right) \mathbf{y}_i \mathbf{y}_j' \tag{3.2}$$

Um estimador não viciado para a variância de aleatorização de $\widehat{\mathbf{Y}}_{HT}$ é dado por:

$$\widehat{V}_p\left(\widehat{\mathbf{Y}}_{HT}\right) = \sum_{i\in s}\sum_{j\in s}\left(d_i d_j - d_{ij}\right)\mathbf{y}_i\mathbf{y}_j' \tag{3.3}$$

O estimador de variância em (3.3) é um estimador não viciado da variância de aleatorização de $\widehat{\mathbf{Y}}_{HT}$, isto é

$$E_p\left[\widehat{V}_p\left(\widehat{\mathbf{Y}}_{HT}\right)\right] = V_p\left(\widehat{\mathbf{Y}}_{HT}\right) \tag{3.4}$$

desde que $\pi_{ij} > 0 \quad \forall\, i \neq j \in U$, como vamos supor neste livro.

Exemplo 3.1. Amostragem Aleatória Simples Sem Reposição - AAS

Quando o plano amostral empregado num levantamento é amostragem aleatória simples sem reposição - AAS, as expressões apresentadas para o estimador de total, sua variância e estimadores desta variância simplificam bastante, porque as probabilidades de inclusão e os pesos básicos das unidades ficam iguais a

$$\pi_i = \frac{n}{N} \text{ e } d_i = \frac{N}{n} \;\; \forall\, i \in U \tag{3.5}$$

e

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \text{ e } d_{ij} = \frac{N(N-1)}{n(n-1)} \;\; \forall\, i \neq j \in U \tag{3.6}$$

Essas probabilidades de inclusão e pesos básicos levam às seguintes expressões simplificadas para o caso AAS:

$$\widehat{\mathbf{Y}}_{AAS} = \frac{N}{n}\sum_{i\in s}\mathbf{y}_i = N\,\overline{\mathbf{y}} \tag{3.7}$$

onde

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i\in s}\mathbf{y}_i \tag{3.8}$$

$$V_{AAS}\left(\widehat{\mathbf{Y}}_{AAS}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\mathbf{S}_y \tag{3.9}$$

onde

$$\mathbf{S}_y = \frac{1}{N-1}\sum_{i\in U}\left(\mathbf{y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{y}_i - \overline{\mathbf{Y}}\right)' \tag{3.10}$$

$$\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i\in U}\mathbf{y}_i = \frac{1}{N}\mathbf{Y} \tag{3.11}$$

Sob AAS, o estimador da variância do estimador de total simplifica para:

$$\widehat{V}_{AAS}\left(\widehat{\mathbf{Y}}_{AAS}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\widehat{\mathbf{S}}_y \tag{3.12}$$

onde

$$\widehat{\mathbf{S}}_y = \frac{1}{n-1} \sum_{i \in s} (\mathbf{y}_i - \overline{\mathbf{y}}) (\mathbf{y}_i - \overline{\mathbf{y}})' \tag{3.13}$$

Vários outros estimadores de totais estão disponíveis na literatura de amostragem, porém os que são comumente usados na prática são estimadores ponderados (lineares) da forma

$$\widehat{\mathbf{Y}}_w = \sum_{i \in s} w_i \mathbf{y}_i \tag{3.14}$$

onde w_i é um peso associado à unidade i da amostra ($i \in s$).

O estimador de Horvitz-Thompson é um caso particular de $\widehat{\mathbf{Y}}_w$ em (3.14) quando os pesos w_i são da forma

$$w_i^{HT} = d_i = 1/\pi_i \quad \forall\ i \in s.$$

Outros dois estimadores de totais comumente usados pelos praticantes de amostragem são o *estimador de razão simples* $\widehat{\mathbf{Y}}_R$ e o *estimador de regressão simples* $\widehat{\mathbf{Y}}_{REG}$, dados respectivamente por

$$\widehat{\mathbf{Y}}_R = \sum_{i \in s} w_i^R\ \mathbf{y}_i \tag{3.15}$$

com

$$w_i^R = d_i \times \frac{\sum_{i \in U} x_i}{\sum_{i \in s} d_i\ x_i} = d_i \times \frac{X}{\widehat{X}_{HT}} \tag{3.16}$$

e

$$\widehat{\mathbf{Y}}_{REG} = \sum_{i \in s} w_i^{REG}\ \mathbf{y}_i \tag{3.17}$$

onde

$$w_i^{REG} = d_i \times g_i \tag{3.18}$$

sendo

$$g_i = 1 + x_i\ (X - \widehat{X}_{HT}) / \sum_{i \in s} d_i x_i^2$$

O fator multiplicativo de ajuste de regressão g_i depende de conhecermos o total populacional $\sum_{i \in U} x_i = X$ de uma variável auxiliar x, e do estimador tipo Horvitz-Thompson para esse total dado por $\widehat{X}_{HT} = \sum_{i \in s} d_i\, x_i$.

O estimador de regressão descrito em (3.17) é um caso particular do *estimador de regressão generalizado*, obtido quando se consideram vetores de variáveis auxiliares em vez de uma única variável auxiliar x como aqui. Para uma discussão detalhada do *estimador de regressão generalizado* ver o Capítulo 3 de Silva (1996), ou o excelente livro de Särndal et al. (1992). Por sua vez, o *estimador de regressão generalizado* é caso

particular da família mais ampla dos *estimadores de calibração*, definidos por Deville e Särndal (1992). Mais informações sobre esta família de estimadores no Capítulo 13 de Silva et al. (2020).

Para completar a descrição dos procedimentos de inferência para médias e totais baseados em estimadores ponderados do tipo razão ou regressão, é necessário identificar estimadores para as variâncias de aleatorização correspondentes. Entretanto, os estimadores de razão e regressão são viciados sob a distribuição de aleatorização para pequenas amostras. Em ambos os casos, o vício é desprezível para amostras grandes, e estão disponíveis expressões assintóticas para as respectivas variâncias de aleatorização.

Partindo destas expressões foram então construídos estimadores amostrais das variâncias dos estimadores de razão e regressão, que podem ser encontrados na excelente revisão sobre o tema contida em Särndal et al. (1992), Seção 6.6 e Capítulo 7. Apesar de sua importância para os praticantes de amostragem, a discussão detalhada desse problema não está incluída neste livro.

O problema da estimação das variâncias de aleatorização para estimadores como os de razão e regressão nos remete a uma questão central da teoria da amostragem. Trata-se dos métodos disponíveis para estimar variâncias de estimadores "complexos". O caso dos estimadores de razão e regressão para totais e médias foi resolvido faz tempo, e não há muito o que discutir aqui. Entretanto, a variedade de métodos empregados para estimação de variâncias merece uma discussão em separado, pois as técnicas de ajuste consideradas neste livro para incorporar pesos e plano amostral na inferência partindo de dados de pesquisas amostrais complexas depende em grande medida da aplicação de tais técnicas.

3.2 Estimação de variâncias - motivação

Em Amostragem, como de resto na Estatística Clássica, a estimação de variâncias é um componente *essencial* da abordagem inferencial adotada: sem estimativas de variância, nenhuma indicação da precisão (e, portanto, da qualidade) das estimativas de interesse está disponível. Nesse caso, uma tentação que assola muitos usuários incautos é esquecer que os resultados são baseados apenas em dados de uma amostra da população e, portanto, sujeitos a incerteza, que não pode ser quantificada sem medidas de precisão amostral.

Em geral, a obtenção de estimativas de variâncias (alternativamente, de desvios padrões ou mesmo de coeficientes de variação) é requerida para que intervalos de confiança possam ser calculados, e outras formas de inferência realizadas. Intervalos de confiança elaborados com estimativas amostrais são geralmente baseados em aproximações assintóticas da distribuição amostral do estimador pela distribuição normal, usando resultados análogos ao TCL para populações finitas - ver Fuller (2009), tais que intervalos da forma

$$IC\left[\widehat{\theta}; 1-\alpha\right] = \left[\ \widehat{\theta} \mp z_{\alpha/2}\sqrt{\widehat{V}_p\left(\widehat{\theta}\right)}\ \right]$$

têm probabilidade de cobertura aproximada $1-\alpha$, com $z_{\alpha/2}$ sendo o quantil que deixa área de $1-\alpha/2$ à sua esquerda na distribuição Normal padrão.

Estimativas de variância podem ser úteis também para outras finalidades, tais como a detecção de problemas não antecipados, tais como observações suspeitas, celas raras em tabelas de contingência, etc.

A estimação de variâncias para os casos padrões de amostragem, isto é, quando os estimadores são lineares nas observações amostrais, não viciados e todas as probabilidades de inclusão conjuntas são não nulas, é tratada em todos os livros de amostragem convencionais. Apesar disso, os pacotes estatísticos usuais, tais como SAS, SPSS, MINITAB e outros, por muito tempo não ofereciam rotinas prontas para estimar variâncias considerando o plano amostral, nem mesmo para estatísticas simples como estimadores de totais e médias.

Felizmente tal situação mudou, e agora já é possível contar com ferramentas no SAS (procedimentos *survey*), no SPSS (módulo *Complex Samples*) e no STATA (funções *svy*). Mas, a nosso ver, é no pacote *survey* do sistema R que estão disponíveis as melhores ferramentas para estimação de parâmetros a partir de dados de amostras complexas.

Para alguns planos amostrais utilizados na prática, as probabilidades de inclusão conjuntas podem ser nulas (caso de amostragem sistemática) ou difíceis de calcular (caso de alguns esquemas de seleção com probabilidades desiguais). Nesses casos, as expressões fornecidas na Seção 3.1 para os estimadores das variâncias dos estimadores de totais não são mais adequadas.

Em muitos outros casos, como se vê no restante deste livro, os parâmetros de interesse são "não lineares" (diferentes de totais, médias e proporções, por exemplo). Casos comuns que consideremos mais adiante são a estimação de razões, coeficientes de modelos de regressão etc. Nesses casos é comum que as estatísticas empregadas para estimar tais parâmetros também sejam "não lineares".

Finalmente, alguns estimadores de variância podem, em alguns casos, produzir estimativas negativas da variância, que são inaceitáveis de um ponto de vista prático (tais como o estimador da expressão (3.3) para alguns esquemas de seleção com probabilidades desiguais e determinadas configurações peculiares da amostra).

Em todos esses casos, é requerido o emprego de técnicas especiais de estimação de variância. É de algumas dessas técnicas que tratam as seções seguintes deste capítulo. A seleção das técnicas discutidas aqui não é exaustiva, e um tratamento mais completo e aprofundado da questão pode ser encontrado no livro de Wolter (2007). Discutimos inicialmente a técnica de *Linearização de Taylor*, em seguida uma abordagem comumente adotada para estimar variâncias para planos amostrais estratificados e conglomerados em vários estágios, com seleção de unidades primárias com probabilidades desiguais, denominada *Método do Conglomerado Primário* (do inglês *Ultimate Cluster*). Por último, tratamos brevemente de uma técnica baseada na ideia de pseudo replicações da amostra, denominada *Bootstrap*. A combinação dessas três idéias suporta os desenvolvimentos teóricos dos algoritmos empregados pelo pacote *survey* do sistema R para estimação de variâncias - ver Lumley (2006) e Lumley (2010).

3.3 Linearização de Taylor (ou Delta) para estimar variâncias

Um problema que ocorre frequentemente é o de estimar um vetor de parâmetros $\theta = (\theta_1, \ldots, \theta_K)$ de uma população finita U, que pode ser escrito na forma:

$$\theta = \mathbf{g}(\mathbf{Y})$$

onde $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$ é o vetor de totais de Q variáveis de pesquisa.

Poderíamos usar como estimador para o vetor de parâmetros θ o estimador $\widehat{\theta}$ dado por:

$$\widehat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right) = \mathbf{g}\left(\sum_{i \in s} d_i \, \mathbf{y}_i\right)$$

No caso particular em que $\mathbf{g}(\bullet)$ é uma função linear dos totais das variáveis de pesquisa, isto é:

$$\theta = \mathbf{AY}$$

onde $\mathbf{A}$ é uma matriz de constantes de dimensão $K \times Q$, o estimador $\widehat{\theta}$ de θ neste caso seria

$$\widehat{\theta} = \mathbf{A}\widehat{\mathbf{Y}}_{HT}$$

Nesse caso particular, é fácil estudar as propriedades do estimador $\widehat{\theta}$. Este estimador é não viciado e tem variância de aleatorização dada por:

$$V_p\left(\widehat{\theta}\right) = \mathbf{A}\left[V_p\left(\widehat{\mathbf{Y}}_{HT}\right)\right]\mathbf{A}'$$

onde $V_p\left(\widehat{\mathbf{Y}}_{HT}\right)$ é dado em (3.2).

Quando $\mathbf{g}(\bullet)$ é uma função não linear, podemos usar a técnica de *Linearização de Taylor* (ou Método Delta) para obter aproximações assintóticas para a variância de $\widehat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right)$. Para maiores detalhes sobre esse método, ver por exemplo p. 172 de Särndal et al. (1992) ou p. 486 de Bishop et al. (1975).

Vamos considerar a expansão de $\mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right)$ em torno de $\mathbf{Y}$, até o termo de primeira ordem, desprezando o resto, dada por:

$$\widehat{\theta} \doteq \widehat{\theta}_L = \mathbf{g}(\mathbf{Y}) + \mathbf{\Delta g}(\mathbf{Y})\left(\widehat{\mathbf{Y}}_{HT} - \mathbf{Y}\right) \tag{3.19}$$

onde $\mathbf{\Delta g}(\mathbf{Y})$ é a matriz Jacobiana $K \times Q$ cuja q-ésima coluna é $\partial \mathbf{g}(\mathbf{Y})/\partial Y_q$, para $q = 1, \ldots, Q$.

A ideia básica do método de linearização é aproximar a variância do estimador $\widehat{\theta}$ pela variância do *estimador linearizado* $\widehat{\theta}_L$ dado pelo lado direito da expressão (3.19). Para obter a variância do estimador linearizado, note que $\mathbf{g}(\mathbf{Y})$ é uma constante, e que

$$\begin{aligned}
\mathbf{\Delta g}(\mathbf{Y})\left(\widehat{\mathbf{Y}}_{HT} - \mathbf{Y}\right) &= \mathbf{\Delta g}(\mathbf{Y})\widehat{\mathbf{Y}}_{HT} - \mathbf{\Delta g}(\mathbf{Y})\mathbf{Y} \\
&= \textstyle\sum_{i\in s} d_i\, \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i - \sum_{i\in U} \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i \\
&= \textstyle\sum_{i\in s} d_i\, \mathbf{z}_i - \sum_{i\in U} \mathbf{z}_i = \widehat{\mathbf{Z}}_{HT} - \mathbf{Z}
\end{aligned}$$

onde $\mathbf{z}_i = \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i$.

Logo, a variância aproximada por linearização do estimador $\widehat{\theta}$ pode ser obtida usando a expressão (3.2)

$$V_p\left(\widehat{\theta}\right) \doteq V_p\left(\widehat{\mathbf{Z}}_{HT}\right)$$

Este resultado segue porque na expressão do lado direito o único termo que tem variância de aleatorização é $\widehat{\mathbf{Z}}_{HT}$.

Um estimador consistente de $V_p\left(\widehat{\theta}\right)$ é dado por:

$$\widehat{V}_p\left(\widehat{\theta}\right) = \widehat{V}_p\left(\widehat{\mathbf{Z}}_{HT}\right) \tag{3.20}$$

onde $\widehat{V}_p\left(\widehat{\mathbf{Z}}_{HT}\right)$ é dado em (3.3), onde substituímos o vetor de variáveis resposta original $\mathbf{y}_i$ pelo vetor de variáveis linearizadas $\mathbf{z}_i = \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i$.

Linearização de Taylor pode ser trabalhosa, porque para cada parâmetro/estimador de interesse são requeridas derivações e cálculos específicos. Felizmente, grande parte das situações de interesse prático estão hoje cobertas por pacotes estatísticos especializados na estimação de medidas descritivas e parâmetros de modelos, e suas respectivas variâncias de aleatorização empregando o método de linearização, de modo que essa desvantagem potencial tende a se diluir.

Linearização de Taylor pode não ser imediatamente possível, pois pode ocorrer que as quantidades de interesse não podem ser expressas como funções de totais ou médias populacionais (este é o caso de quantis de distribuições, por exemplo). Para estes casos é necessário recorrer a outras técnicas de estimação de variâncias, como discutido, por exemplo, em Wolter (2007).

3.4 Equações de estimação

Até aqui, falamos da estimação de totais e de parâmetros que podem ser escritos como funções de totais. O caminho para obter resultados gerais referentes a muitos outros parâmetros de interesse é o que discutimos nesta seção.

Se um parâmetro populacional de interesse θ_U é uma solução única de um sistema de equações de estimação definidas como

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \mathbf{0} \tag{3.21}$$

para uma função $\mathbf{u}(\bullet)$ conhecida, então é possível estimar o parâmetro θ_U usando o estimador $\widehat{\theta}$ obtido resolvendo as equações de estimação amostrais:

$$\sum_{i \in s} d_i \, \mathbf{u_i}(\theta) = \mathbf{0} \tag{3.22}$$

O estimador $\widehat{\theta}$ é consistente para θ_U, e adiante mostraremos como o método de Linearização de Taylor pode ser usado para estimar a sua variância. Antes, porém, vamos usar alguns exemplos para ilustrar casos particulares relevantes de como aplicar essa ideia.

Exemplo 3.2. Estimação de médias populacionais

Para ilustrar a aplicação da abordagem de equações de estimação, considere o caso em que a função $\mathbf{u_i}(\theta) = y_i - \theta$. Nesse caso, as equações de estimação populacionais (3.21) simplificam para:

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \sum_{i \in U} (y_i - \theta) = \mathbf{0}$$

Resolvendo esta equação, obtemos:

$$\theta_U = \frac{1}{N} \sum_{i \in U} y_i = \overline{Y}$$

A solução das equações de estimação amostrais fornece:

$$\widehat{\theta} = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} = \overline{y}_{H\grave{a}jek}$$

que é o conhecido estimador de Hàjek da média populacional.

Exemplo 3.3. Estimação de razões populacionais

Considere agora o caso em que a função $\mathbf{u_i}(\theta) = y_i - \theta z_i$. Nesse caso, as equações de estimação populacionais (3.21) simplificam para:

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \sum_{i \in U} (y_i - \theta z_i) = \mathbf{0}$$

Resolvendo esta equação, obtemos:

$$\theta_U = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} z_i} = \frac{Y}{Z} = R$$

A solução das equações de estimação amostrais correspondentes fornece:

$$\widehat{\theta} = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i z_i} = \frac{\widehat{Y}_{HT}}{\widehat{Z}_{HT}} = \widehat{R}$$

Os exemplos apresentados ilustram que a estimação de médias e razões populacionais são casos particulares simples da abordagem mais geral de *equações de estimação*. Essa abordagem também se mostrará útil quando lidamos com a estimação de parâmetros sob vários tipos de modelos paramétricos, que está apresentada nos capítulos seguintes deste livro. É também graças a ela que foi possível desenvolver software genérico para estimação a partir de amostras complexas, como é o caso do pacote *survey* do sistema R.

A estimação de variâncias nesse caso pode ser feita usando o método de Linearização de Taylor, empregando a estratégia de calcular variáveis linearizadas z definidas como na Seção 3.3. Esta é a estratégia adotada no pacote *survey* do sistema R.

3.5 Método do Conglomerado Primário

A ideia central do Método do Conglomerado Primário (do inglês *Ultimate Cluster*) para estimação de variâncias para estimadores de totais e médias em planos amostrais de múltiplos estágios, proposto por Hansen et al. (1953), é considerar apenas a variação entre informações disponíveis no nível das unidades primárias de amostragem - UPAs, isto é, dos *conglomerados primários*, e admitir que estes teriam sido selecionados com reposição da população de UPAs. Esta ideia é simples, porém bastante poderosa, porque permite acomodar uma enorme variedade de planos amostrais envolvendo estratificação, amostragem conglomerada e seleção com probabilidades desiguais (com ou sem reposição) tanto das UPAs como das demais unidades de amostragem.

Os requisitos fundamentais para permitir a aplicação deste método são que estejam disponíveis estimadores não viciados dos totais da variável de interesse para cada um dos conglomerados primários selecionados, e que pelo menos dois destes sejam selecionados em cada estrato (se a amostra for estratificada no primeiro estágio).

Embora o método tenha sido originalmente proposto para estimação de totais, pode ser aplicado também para estimar (por linearização) quantidades populacionais que possam ser representadas como funções de totais, conforme discutido na Seção 3.3. De fato, esse método fornece a base para ferramentas dos sistemas estatísticos para cálculo de variâncias considerando o plano amostral, tais como o pacote *survey* do R, as funções *svy* do STATA, o módulo *Complex Samples* do SPSS, as procs *Survey* do SAS, entre outros.

Para descrever o método, considere um plano amostral em vários estágios, no qual n_h unidades primárias de amostragem - UPAs foram selecionadas no estrato h, com $h = 1, \ldots, H$. Denotemos por π_{hi} a probabilidade de inclusão na amostra da UPA (conglomerado primário) i do estrato h, e por $\widehat{Y}_{hi}$ um estimador não viciado

do total Y_{hi} da variável de pesquisa y na i-ésima UPA do estrato h. Então um estimador não viciado do total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$ da variável de pesquisa y na população é dado por

$$\widehat{Y}_{CP} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \widehat{Y}_{hi}/\pi_{hi} \tag{3.23}$$

e um estimador não viciado da variância de aleatorização correspondente por

$$\widehat{V}_p\left(\widehat{Y}_{CP}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\widehat{Y}_{hi}}{\pi_{hi}} - \frac{\widehat{Y}_h}{n_h}\right)^2 \tag{3.24}$$

onde $\widehat{Y}_h = \sum_{i=1}^{n_h} \widehat{Y}_{hi}/\pi_{hi}$ para $h = 1, \ldots, H$. (Ver por exemplo, Shah et al. (1993), p. 4).

Embora na prática a seleção das UPAs seja geralmente feita sem reposição, o estimador do Método do Conglomerado Primário - MCP aqui apresentado pode fornecer uma aproximação razoável da correspondente variância de aleatorização, especialmente nos casos em que as frações amostrais de UPAs são pequenas nos estratos. Isso ocorre porque planos amostrais sem reposição são em geral mais eficientes que planos com reposição de igual tamanho.

Tal aproximação é largamente utilizada pelos praticantes de amostragem para estimar variâncias de quantidades descritivas usuais tais como totais e médias (com a devida adaptação) devido à sua simplicidade, comparada com a complexidade muito maior envolvida com o emprego de estimadores de variância que tentam incorporar todas as etapas de planos amostrais conglomerados em vários estágios. Uma discussão sobre a qualidade dessa aproximação e alternativas pode ser encontrada em Särndal et al. (1992), p. 153.

3.6 Métodos de replicação

A ideia de usar métodos indiretos ou de replicação para estimar variâncias em amostragem não é nova. Mahalanobis (1939), Mahalanobis (1944) e Deming (1956) foram os precursores e muitos desenvolvimentos importantes se seguiram. Hoje em dia várias técnicas baseadas nessa ideia são rotineiramente empregadas por praticantes de amostragem, e inclusive formam a base para pacotes especializados de estimação tais como WesVarPC (ver Westat (1996)).

A ideia básica original foi construir a amostra de tamanho n como a união de G amostras de tamanho n/G cada uma, selecionadas de forma independente e usando o mesmo plano amostral, onde G é o número de *réplicas*. Nesse caso, se θ é o parâmetro-alvo, e $\widehat{\theta}_g$ é um estimador não viciado de θ baseado na g-ésima réplica ($g = 1, \ldots, G$), segue-se que

$$\widehat{\theta}_G = \frac{1}{G} \sum_{g=1}^{G} \widehat{\theta}_g$$

é também um estimador não viciado de θ e

$$\widehat{V}_G\left(\widehat{\theta}_G\right) = \frac{1}{G\,(G-1)} \sum_{g=1}^{G} \left(\widehat{\theta}_g - \widehat{\theta}_G\right)^2 \tag{3.25}$$

é um estimador não viciado (de replicação) da variância do estimador $\widehat{\theta}_G$.

Note que desde que as réplicas sejam construídas de forma independente conforme indicado, os estimadores $\widehat{\theta}_G$ e $\widehat{V}_G\left(\widehat{\theta}_G\right)$ são não viciados qualquer que seja o plano amostral empregado para selecionar a amostra de cada réplica, o que faz desta uma técnica flexível e genérica. Além disso, a abordagem de replicação é bastante geral, pois os estimadores aos quais se aplica não precisam ser necessariamente expressos como funções de totais, como ocorre com a técnica de Linearização de Taylor discutida na Seção 3.3.

Apesar destas vantagens, a aplicação prática desta técnica de forma exata é restrita porque, em geral, é menos eficiente, inconveniente e mais caro selecionar G amostras (réplicas) independentes com o mesmo esquema, se comparado à seleção de uma única amostra de tamanho n diretamente. Além disto, se o número de réplicas G for pequeno, o estimador de variância pode ser instável.

Mesmo quando a amostra não foi selecionada exatamente dessa forma, a construção de réplicas a posteriori para fins de estimação de variâncias em situações complexas é também uma ideia simples de aplicar, poderosa e flexível, por acomodar uma ampla gama de planos amostrais e situações de estimação de interesse. Quando as réplicas são construídas após a pesquisa (a posteriori), mediante repartição (por sorteio) da amostra pesquisada em G grupos mutuamente exclusivos de igual tamanho, estas são chamadas de *réplicas dependentes* ou *grupos aleatórios* (do inglês *random groups*). As expressões fornecidas para o estimador de replicação e sua variância são também empregadas nesse caso como uma aproximação, mas não possuem as mesmas propriedades do caso de réplicas independentes.

Uma pesquisa importante e de grande porte em que esta ideia é aplicada é a pesquisa de preços para formar o índice de Preços ao Consumidor (do inglês *Consumer Price Index - CPI*) do *US Bureau of Labour Statistics (2020)*, p. 46, que utiliza duas ou mais réplicas para formar a amostra de itens cujos preços são pesquisados.

É importante observar que a repartição da amostra em grupos aleatórios a posteriori precisa considerar o plano amostral empregado e pode não ser possível em algumas situações. Idealmente, tal repartição deveria ser feita respeitando estratos e alocando UPAs inteiras (isto é, com todas as respectivas unidades subordinadas). Wolter (1985), p. 31, discute algumas regras sobre como fazer para respeitar o plano amostral ao fazer a repartição da amostra a posteriori, porém recomendamos que o interessado no uso dessa técnica exerça cautela.

Além da modificação da interpretação das réplicas no caso de serem formadas a posteriori, é comum também nesse caso empregar um estimador para o parâmetro θ baseado na amostra completa (denotado $\widehat{\theta}$), e um estimador de variância mais conservador que o estimador $\widehat{V}_G\left(\widehat{\theta}_G\right)$ anteriormente apresentado, dado por

$$\widehat{V}_G\left(\widehat{\theta}\right) = \frac{1}{G\left(G-1\right)} \sum_{g=1}^{G} \left(\widehat{\theta}_g - \widehat{\theta}\right)^2 \tag{3.26}$$

Um exemplo de aplicação desta técnica pode ser encontrado na forma recomendada para estimação de variâncias a partir das Amostras de Uso Público do Censo Demográfico Brasileiro de 80 (ver IBGE (1985)).

Nesta seção descreveremos duas outras dessas técnicas baseadas em replicações. A primeira é o método de *jackknife*. Este método foi originalmente proposto por Quenoille (1949) e Quenoille (1956) como uma técnica para redução de vício de estimadores, num contexto da Estatística Clássica. A ideia central consiste em repartir a amostra (a posteriori, como no caso do método dos grupos aleatórios) em G grupos mutuamente exclusivos de igual tamanho n/G. Em seguida, para cada grupo formado calcular os chamados pseudo estimadores dados por

$$\widehat{\theta}_{(g)} = G\widehat{\theta} - \left(G-1\right)\widehat{\theta}_g$$

onde $\widehat{\theta}_g$ é um estimador de θ obtido da amostra após eliminar os elementos do grupo g, empregando a mesma forma funcional adotada no cálculo do estimador $\widehat{\theta}$ que considera a amostra inteira.

A estimação da variância por esse método pode então ser feita de duas maneiras alternativas, usando um dos estimadores dados por

$$\widehat{V}_{J1}\left(\widehat{\theta}\right) = \frac{1}{G\left(G-1\right)} \sum_{g=1}^{G} \left(\widehat{\theta}_{(g)} - \widehat{\theta}_J\right)^2 \tag{3.27}$$

ou

$$\widehat{V}_{J2}\left(\widehat{\theta}\right) = \frac{1}{G\left(G-1\right)} \sum_{g=1}^{G} \left(\widehat{\theta}_{(g)} - \widehat{\theta}\right)^2 \tag{3.28}$$

onde $\widehat{\theta}_J = \frac{1}{G}\sum_{g=1}^{G} \widehat{\theta}_{(g)}$ é um estimador pontual *jackknife* para θ, alternativo ao estimador da amostra inteira $\hat{\theta}$.

Observaçao 3.1: A descrição do método *jackknife* aqui apresentada não cobre o caso de planos amostrais estratificados, que é mais complexo. Para detalhes sobre este caso, consulte Wolter (1985), pág. 174.

Observaçao 3.2: O estimador $\widehat{V}_{J2}\left(\widehat{\theta}\right)$ é mais conservador que o estimador $\widehat{V}_{J1}\left(\widehat{\theta}\right)$.

Observaçao 3.3: É comum aplicar a técnica fazendo o número de grupos igual ao tamanho da amostra, isto é, tomando $G = n$ e portanto eliminando uma observação da amostra de cada vez ao calcular os pseudo valores. Essa regra deve ser aplicada considerando o número de UPAs quando o plano amostral é em múltiplos estágios, pois as UPAs devem sempre ser eliminadas com todas as unidades subordinadas.

Os estimadores de variância do método *jackknife* fornecem resultadoS idênticos aos dos estimadores usuais de variância quando aplicados para o caso de estimadores lineares nas observações amostrais. Além disso, suas propriedades são razoáveis para vários outros casos de estimadores não lineares de interesse (ver, por exemplo, Cochran (1977), p. 321 e Wolter (1985), p. 306). A situação merece maiores cuidados para o caso de quantis ou estatísticas de ordem, tais como a mediana e o máximo, pois neste caso essa técnica não funciona bem Wolter (1985), p. 163.

O pacote WesVarPC - Westat (1996) - baseia suas estimativas de variância principalmente no método *jackknife*, embora também possua uma opção para usar outro método conhecido como de replicações de meias amostras balanceadas (do inglês *balanced half-sample replication*).

O outro método de replicação que vamos considerar é uma variante do método *bootstrap* proposta por Rao et al. (1992). O método consiste dos seguintes passos:

1) Selecione amostras aleatórias simples com reposição de m_h das n_h UPAs de cada estrato $h = 1, \ldots, H$.
2) Calcule as contagens m^*_{hi} de vezes que cada UPA i aparece na amostra selecionada no estrato h; note que $\sum_i m^*_{hi} = m_h$ para todo estrato h;
3) Defina pesos *bootstrap* para as unidades da amostra selecionada em (1) usando:

$$w^*_{hik} = \left[1 - \left(\frac{m_h}{n_h - 1}\right)^{1/2} + \left(\frac{m_h}{n_h - 1}\right)^{1/2} \times \frac{n_h}{m_h} \times m^*_{hi}\right] \times w_{hik} \tag{3.29}$$

onde w_{hik} é o peso da unidade k da UPA i do estrato h. Note que quando uma UPA i não é selecionada, sua contagem m^*_{hi} é igual a zero, e o terceiro termo dentro do colchete é nulo.

4) Calcule uma estimativa $\widehat{\theta}_b$ para o parâmetro de interesse usando os pesos *bootstrap* w^*_{hik} em lugar dos pesos originais w_{hik}.

5) Repita os passos 1) a 4) um número B grande de vezes.

6) Estime a variância do estimador $\widehat{\theta}$ com:

$$\widehat{V}_B\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b - \widehat{\theta}\right)^2 \tag{3.30}$$

A Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua do IBGE passou a usar este método *bootstrap* para estimação da precisão dos indicadores que divulga a partir do terceiro trimestre de 2021, ver IBGE (2021).

Embora computacionalmente mais custoso que o método da Linearização de Taylor, o método *bootstrap* aqui descrito tem como vantagem a aplicação em casos onde o estimador não é função suave de totais populacionais, tais como separatrizes (quantis), algumas medidas de desigualdade e pobreza etc. Além disso, o método pode ser aplicado com qualquer software que permita implementar o algoritmo descrito, e não requer pacotes especializados. Vale mencionar, entretanto, que este método está disponível no pacote *survey* do sistema R. Sua utilização é ilustrada em capítulos posteriores.

XXX Parei aqui

3.7 Laboratório de R

Vamos utilizar dados da Pesquisa de Padrão de Vida (PPV) do IBGE para ilustrar alguns métodos de estimação de variâncias. Vamos considerar a estimação da proporção de analfabetos na faixa etária acima de 14 anos. Os dados da pesquisa encontram-se no data frame `ppv1`. A variável *analf2* é indicadora da condição de analfabetismo na faixa etária acima de 14 anos e a variável *faixa2* é indicadora da faixa etária acima de 14 anos. Queremos estimar a proporção de analfabetos na faixa etária acima de 14 anos na região Sudeste. Antes apresentamos o método de estimação de variância por linearização de Taylor

Vamos criar duas variáveis:

- analf - variável indicadora da condição de analfabetismo: *v04a01* ou `v04a02` igual a 2;
- faixa - variável indicadora de faixa etária entre 7 e 14 anos.

```
library(survey)
ppv_dat <- readRDS("./data/ppv.rds") # carrega dados
# cria objeto de desenho
ppv_plan<-svydesign(ids = ~nsetor, strata = ~estratof,
data = ppv_dat, nest = TRUE, weights = ~pesof)
# atualiza objeto de desenho com novas variáveis
ppv_plan<-update(ppv_plan,
  analf=(v04a01 == 2 | v04a02 == 2)*1,
  faixa=(v02a08 >= 7 & v02a08 <= 14) *1,
  analf.faixa= (analf==1 & faixa==1)*1
)
```

Como estamos interessados em estimativas relativas à Região Sudeste, vamos restringir o desenho a esse domínio:

```
ppv_se_plan <- subset(ppv_plan, regiao == 2)
```

Vamos estimar os totais das variáveis *analf.faixa* e *faixa*:

```
analf_faixa_tot_est<-svytotal(~analf.faixa+faixa ,ppv_se_plan )
Vcov.Y1.Y2<-vcov(analf_faixa_tot_est)
```

Substituindo os valores na expressão (3.21), obtemos a estimativa da variância da razão de totais das variáveis *analf.faixa* e *faixa*.

```
y1hat<-coef(analf_faixa_tot_est)[1]
y2hat<-coef(analf_faixa_tot_est)[2]
Var.raz<-(1/y2hat)*(1/y2hat)*Vcov.Y1.Y2[1,1]+2*(1/y2hat)*(-y1hat/y2hat^2)*Vcov.Y1.Y2[1,2]+
(-y1hat/y2hat^2)*(-y1hat/y2hat^2)*Vcov.Y1.Y2[2,2]
# estimativa do desvio-padrão
sqrt(Var.raz)
```

```
##  faixa
## 0,0118
```

Podemos calcular diretamente o desvio-padrão:

```
svyratio(~analf.faixa, ~faixa, ppv_se_plan)
```

```
## Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv_se_plan)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##              faixa
## analf.faixa 0,0118
```

A estimativa do desvio-padrão obtida por meio da função *svyratio* coincide com a obtida diretamente pelo método de linearização, e é igual a *r round(sqrt(Var.raz),digits=5)*. O método default para estimar variâncias usado pela library *survey* (Lumley, 2021) do R é o de linearização de Taylor.

A library *survey* dispõe de métodos alternativos para a estimação de variância. Vamos utilizar os métodos de replicação de *Jackknife* e de *Bootstrap* para estimar esta variância de razão. Inicialmente, vamos converter o objeto de desenho *ppv1_se_plan* em um objeto de desenho de replicação de tipo *Jackknife*, contendo as réplicas de pesos que fornecem correspondentes réplicas de estimativas.

```
ppv_se_plan_jkn<-as.svrepdesign(ppv_se_plan,type="JKn")
svyratio(~analf.faixa, ~faixa, ppv_se_plan_jkn)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_jkn)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##             [,1]
## [1,] 0,0118
```

Para o tipo *Bootstrap*, temos:

```
ppv_se_plan_boot<-as.svrepdesign(ppv_se_plan,type="bootstrap")
svyratio(~analf.faixa, ~faixa, ppv_se_plan_boot)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_boot)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##        [,1]
## [1,] 0,0103
```

Vamos apresentar mais detalhes sobre a obtenção dos estimadores de *Jackknife* e *Bootstrap* na library *survey* (Lumley, 2021). A classe do objeto *ppv_se_plan_jkn* é *svyrep.design* e ele contém as seguintes componentes:

```
class(ppv_se_plan_jkn)
```

```
## [1] "svyrep.design"
```

```
names(ppv_se_plan_jkn)
```

```
##  [1] "repweights"       "pweights"         "type"             "rho"
##  [5] "scale"            "rscales"          "call"             "combined.weights"
##  [9] "selfrep"          "mse"              "variables"        "degf"
```

A componente *repweights* é uma lista com duas componentes: *weights* e *index*. A componente *weights* é uma matriz de dimensão 276×276, onde 276 é o número de conglomerados primários do plano amostral da PPV na região Sudeste. A partir desta matriz, podemos obter 276 réplicas de pesos de desenho de Jackknife.

```
ppv_se_dat<-ppv_se_plan_jkn$variables
nrow(ppv_se_dat)
```

```
## [1] 8903
```

```
ncong<-sum(with(ppv_se_dat,tapply( nsetor,estratof, function(t) length(unique(t)))))
ncong
```

```
## [1] 276
```

O argumento *compress* da função *as.svrepdesign* permite especificar se, na saída da função, a matriz `weights` é na forma comprimida ou não. Na aplicação feita foi usado o valor default que é a forma comprimida. A forma não comprimida da matriz *weights* tem *r nrow(ppv_se_dat)* linhas e *r ncong* colunas. A forma comprimida permite economizar memória, e pode ser facilmente convertida para a forma não comprimida, utilizando-se a componente`index`.

No método *jackknife*, cada um dos conglomerados primários é removido, e a réplica correspondente dos pesos é o produto do peso amostral original por um fator apropriado, definido da forma a seguir. Suponhamos que foi removido um conglomerado no estrato h, então os pesos do plano amostral serão multiplicados por:

- 0 para as unidades no conglomerado removido;
- $m_h/(m_h - 1)$ para unidades pertencentes a outros conglomerados do estrato h;
- 1 para unidades em estratos $h' \neq h$.

Podemos obter a matriz de fatores de correção do peso amostral na forma não comprimida da seguinte maneira:

```
fact_peso_comp_mat<-ppv_se_plan_jkn$repweights[[1]]
ind_cong <-ppv_se_plan_jkn$repweights[[2]]
fat_pesos_mat<- fact_peso_comp_mat[ind_cong,]
str(fat_pesos_mat)
```

```
##  num [1:8903, 1:276] 0 0 1,06 1,06 1,06 ...
```

Podemos obter matriz de réplicas de pesos multiplicando cada coluna dessa matriz pelos pesos do plano amostra:

```
rep_pesos_mat<-weights(ppv_se_plan)*fat_pesos_mat
```

Utilizando esta matriz de réplicas de pesos, podemos obter réplicas correspondentes de estimativas da razão.

```
rep_est_raz<-numeric(ncol(rep_pesos_mat))
for (i in 1:ncol(rep_pesos_mat)){
rep_est_raz[i]<-sum(rep_pesos_mat[,i]*ppv_se_dat$analf.faixa)/sum(rep_pesos_mat[,i]*ppv_se_da
}
```

A partir destas réplicas de estimativas da razão, finalmente estimamos a variância:

```
mean_raz<-mean( rep_est_raz[ppv_se_plan_jkn$rscales>0])
var_jack_raz<- sum((rep_est_raz-mean_raz)^2*ppv_se_plan_jkn$rscales)*ppv_se_plan_jkn$scale
round(sqrt(var_jack_raz),5)
```

```
## [1] 0,0118
```

A library *survey* (Lumley, 2021) fornece uma função para estimar a variância de uma função de totais a partir das réplicas de pesos:

```
var_raz_rep<-withReplicates(ppv_se_plan_jkn, function(w,ppv_se_dat) sum(w*ppv_se_dat$analf.fa
var_raz_rep
```

```
##      theta   SE
## [1,] 0,119 0,01
```

Resultado que coincide com a estimativa obtida pela aplicação da função *svyratio*.

A vantagem de utilizar métodos de replicação é a facilidade com que estimamos a variância de qualquer característica da população, cujo estimador pontual é conhecido. Por exemplo, se quisermos estimar a variância da razão das taxas de analfabetos nas faixas etárias de 0 a 14 anos e acima de 14 anos podemos usar as mesmas réplicas de pesos:

```
withReplicates (ppv_se_plan_jkn,function(w,ppv_se_dat) with(ppv_se_dat,
(sum(w*(analf==1&faixa==1))/sum(w*(faixa==1)))/(sum(w*(analf==1&faixa==0))/sum(w*(faixa==0)))
))
```

```
##      theta   SE
## [1,] 0,504 0,05
```

O erro padrão da razão entre razões estimada no exemplo anterior pode ser estimado por linearização de Taylor, usando-se a função *svycontrast()* da library survey:

```
# cria variáveis dummies:
ppv_se_plan <- update(ppv_se_plan,
```

```
num1 = as.numeric(analf==1 & faixa==1),
num2 = as.numeric(analf==1 & faixa==0),
den1 = as.numeric (faixa == 1),
den2 = as.numeric(faixa == 0)
)
# estima totais e matriz de covariância de estimativas de totais
comp.tot <- svytotal(~num1+num2+den1+den2, ppv_se_plan)

# estima razão de razões:
svycontrast(comp.tot, quote((num1/den1)/(num2/den2)))
```

```
##          nlcon   SE
## contrast 0,504 0,05
```

Capítulo 4

Efeitos do Plano Amostral

4.1 Introdução

A estimação do desvio ou erro padrão de estimativas, de intervalos de confiança e o uso de testes de hipóteses desempenham papel fundamental em estudos analíticos. Além de estimativas pontuais, na inferência analítica é necessário conhecer e comunicar a precisão associada às estimativas e construir intervalos de confiança para os parâmetros de interesse. Valores de desvios padrões, ou alternativamente margens de erro iguais à semiamplitude de intervalos de confiança, permitem avaliar a precisão da estimação. Em muitos casos, a estimação do desvio padrão ou da variância também possibilita a construção de estatísticas para testar hipóteses relativas aos parâmetros do modelo (tradição de modelagem) ou aos parâmetros da população finita (tradição de amostragem). Testes de hipóteses são também usados na fase de seleção de variáveis ou efeitos para inclusão ou manutenção nos modelos ajustados.

Os sistemas mais comuns de análise estatística incluem em suas saídas valores de estimativas pontuais e seus desvios padrões, intervalos de confiança e valores-p relativos a hipóteses de interesse. Contudo, as fórmulas usadas nestes sistemas para a estimação dos desvios padrões e obtenção de estatísticas de testes são, em geral, baseadas nas hipóteses de independência e de igualdade de distribuição (IID) das observações ou, equivalentemente, do emprego de Amostragem Aleatória Simples Com Reposição - AASC. Tais hipóteses quase nunca valem para dados obtidos através de pesquisas por amostragem, como as que realizam o IBGE e outras agências produtoras de estatísticas públicas e/ou oficiais. Os principais motivos para isso são o emprego de estratificação, conglomeração, planos amostrais com probabilidades desiguais de seleção e correções para não resposta, que dão origem a observações amostrais com pesos distintos, e/ou que não são independentes quando provenientes dos mesmos grupos usados como conglomerados nalguma das etapas de amostragem.

Este capítulo trata de como avaliar o impacto sobre estimativas de desvios padrões, intervalos de confiança e níveis de significância de testes usuais quando há afastamentos das hipóteses IID mencionadas, devidos ao uso de planos amostrais complexos para obter os dados. Como veremos, o impacto pode ser muito grande em algumas situações, justificando os cuidados que devem ser tomados na análise de dados provenientes de amostras complexas. Neste capítulo, usamos como referência básica Skinner (1989a).

4.2 Efeito do Plano Amostral - EPA de Kish

Para medir o efeito do plano amostral sobre a variância de um estimador, Kish (1965) propôs uma medida que denominou **Efeito do Plano Amostral** - EPA (em inglês, *design effect* ou, abreviadamente, *deff*). O objetivo desta medida é comparar planos amostrais no estágio de planejamento da pesquisa. O EPA de

Kish é uma razão entre variâncias (de aleatorização) de um estimador, calculadas para dois planos amostrais alternativos. Vamos considerar um estimador $\hat{\theta}$ para um parâmetro único θ e calcular a variância de sua distribuição induzida pelo plano amostral complexo (verdadeiro) $V_{VERD}\left(\hat{\theta}\right)$ e a variância da distribuição do estimador induzida pelo plano de amostragem aleatória simples $V_{AAS}\left(\hat{\theta}\right)$. Nessa comparação, está implícita a ideia de que o estimador $\hat{\theta}$ é não enviesado ou ao menos consistente para o parâmetro θ tanto sob o plano amostral complexo como sob AAS.

Definição 4.1. O **Efeito do Plano Amostral -EPA de Kish** para um estimador $\hat{\theta}$ é

$$EPA_{Kish}\left(\hat{\theta}\right) = \frac{V_{VERD}\left(\hat{\theta}\right)}{V_{AAS}\left(\hat{\theta}\right)} \tag{4.1}$$

Para ilustrar o conceito do $EPA_{Kish}\left(\hat{\theta}\right)$, vamos considerar um exemplo.

Exemplo 4.1. Efeitos de plano amostral de Kish para estimadores de totais com amostragem conglomerada em dois estágios.

Silva e Moura (1990) estimaram o EPA_{Kish} para estimadores de totais de várias variáveis socioeconômicas no nível das Regiões Metropolitanas - RMs - utilizando dados do questionário de amostra do Censo Demográfico de 1980. Essas medidas estimadas do efeito do plano amostral foram calculadas para três estratégias amostrais alternativas, todas considerando amostragem conglomerada de domicílios em dois estágios, tendo o setor censitário como unidade primária de amostragem - UPA - e o domicílio como unidade secundária de amostragem - USA. Duas das alternativas consideraram seleção de setores com equiprobabilidade via amostragem aleatória simples sem reposição - AC2AAS - e fração amostral constante de domicílios no segundo estágio (uma usando o estimador HT do total, e outra usando o estimador de razão para o total calibrando no número total de domicílios da população). Uma terceira estratégia, denominada AC2PPT, considerou a seleção de setores com probabilidades proporcionais ao tamanho - PPT - usando o número de domicílios por setor como medida de tamanho, a seleção de 15 domicílios em cada setor da amostra usando AAS e empregando o correspondente estimador HT.

A título de ilustração, resultados referentes à Região Metropolitana do Rio de Janeiro são apresentados na Tabela 4.1 para algumas variáveis. Note que a população-alvo para a qual se faz inferência considera apenas moradores em domicílios particulares permanentes na Região Metropolitana do Rio de Janeiro.

Tabela 4.1: Efeitos de plano amostral de Kish para variáveis selecionadas
Região Metropolitana do Rio de Janeiro

Variável	AC2AAS + Estimador HT	AC2AAS + Estimador de Razão	AC2PPT + Estimador HT
1. Número total de moradores	10,74	2,00	1,90
2. Número de moradores ocupados	5,78	1,33	1,28
3. Rendimento monetário mensal	5,22	4,92	4,49
4. Número total de filhos nascidos vivos de mulheres com 15 anos ou mais	4,59	2,02	1,89
5. Número de domicílios que têm fogão	111,27	1,58	1,55
6. Número de domicílios que têm telefone	7,11	7,13	6,41
7. Valor do aluguel ou prestação mensal	7,22	7,02	6,45
8. Número de domicílios que têm automóvel e renda $<$ 5SM	1,80	1,67	1,55
9. Número de domicílios que têm geladeira e renda $\geq$ 5SM	46,58	2,26	2,08

Os valores apresentados na Tabela 4.1 para a RM do Rio de Janeiro são similares aos observados para as demais RMs, se consideradas as mesmas variáveis, conforme os resultados obtidos por Silva e Moura (1990). Nota-se grande variação dos valores do EPA_{Kish} cujos valores mínimo e máximo são de 1,28 e 111,27 respectivamente. Porém, em todos os casos, os valores dos EPA_{Kish} são maiores que 1, indicando que as três estratégias consideradas levariam a estimativas menos precisas que as que poderiam ser obtidas usando AAS com os mesmos tamanhos de amostra. Quanto maior o valor do EPA_{Kish}, menor a precisão das estimativas obtidas com cada estratégia em comparação com a de uma AAS de igual tamanho.

Os valores elevados do EPA_{Kish} observados para algumas variáveis realçam a importância de considerar o plano amostral verdadeiro ao estimar variâncias e desvios padrões associados às estimativas pontuais. Isso ocorre porque estimativas ingênuas de variância baseadas na hipótese de AASC podem subestimar as variâncias corretas com viés substancial.

Uma outra análise revela que, para algumas variáveis (1, 2, 4, 5 e 9), o EPA_{Kish} varia consideravelmente entre as diferentes estratégias, enquanto para outras variáveis (3, 6, 7 e 8) as variações entre as estratégias é pequena. Esta análise ilustra o uso pretendido por Kish (1965), qual seja o de permitir comparar estratégias de amostragem em etapas de planejamento de pesquisas por amostragem, usando a precisão da AAS como marco de referência para estratégias alternativas sendo consideradas.

Outra regularidade encontrada nesse valores é que o EPA_{Kish} para a estratégia que usa o plano amostral AC2AAS com o estimador HT apresenta sempre os valores mais elevados, revelando que esta estratégia é menos eficiente que as duas competidoras consideradas. Em geral, o EPA_{Kish} é menor para a estratégia AC2PPT com o correspondente estimador HT, com valores próximos aos da estratégia AC2AAS com estimador de razão.

Os valores dos EPA_{Kish} calculados por Silva e Moura (1990) podiam ser usados para planejar pesquisas amostrais (ao menos nas regiões metropolitanas), pois permitiam comparar e antecipar o impacto do uso de algumas estratégias amostrais alternativas sobre a precisão de estimadores de totais de várias variáveis relevantes. Permitiam também calcular tamanhos amostrais para garantir determinado nível de precisão, sem emprego de fórmulas complicadas, como discutido, por exemplo, em Silva et al. (2020), seção 12.10.4. Portanto, tais valores seriam úteis como informação de apoio ao planejamento de novas pesquisas por amostragem, antes que as respectivas amostras fossem efetivamente selecionadas.

Entretanto, esses valores teriam pouca utilidade em termos de usos analíticos dos dados da amostra do Censo Demográfico 1980 - CD80. É que tais valores, embora tendo sido estimados com essa amostra, foram calculados para planos amostrais distintos do que foi efetivamente adotado para seleção da amostra daquele censo.

A amostra de domicílios usada no CD80 foi estratificada por setor censitário com seleção sistemática de uma fração fixa (1/4 ou 25%) dos domicílios de cada setor. Já os planos amostrais considerados por Silva e Moura (1990) na estimação dos EPA_{Kish} que apresentaram eram planos amostrais em dois estágios, com seleção de setores no primeiro estágio, e domicílios no segundo estágio. Tais planos foram considerados por sua similaridade com os planos amostrais adotados nas principais pesquisas domiciliares do IBGE, tais como a PNAD e a Pesquisa Mensal de Emprego - PME. Portanto, a utilidade maior dos valores tabulados dos EPA_{Kish} seria a comparação de planos amostrais alternativos para planejamento de pesquisas futuras, e não a análise dos resultados da amostra do CD80.

4.3 Efeito do Plano Amostral Ampliado

O que se observou no Exemplo 4.1 com respeito à dificuldade ou mesmo impossibilidade de uso dos EPA_{Kish} calculados para fins analíticos também se aplica para outras situações e é uma deficiência estrutural do conceito de EPA proposto por Kish (1965). Para tentar contornar essa dificuldade, é necessário considerar um conceito ampliado de EPA, correspondente ao conceito de *misspecification effect* - `meff` - proposto por Skinner et al. (1989) na p. 24, que apresentamos e discutimos nesta seção.

Para introduzir este conceito ampliado de EPA, que tem utilidade também para fins de inferência analítica, vamos agora considerar um modelo subjacente às observações usadas para o cálculo do estimador pontual $\hat{\theta}$ para um parâmetro único θ. Designemos por $v_0 = \widehat{V}_{IID}\left(\hat{\theta}\right)$ um estimador usual (consistente) da variância de $\hat{\theta}$ calculado sob a hipótese (ingênua) de que as observações da amostra disponível são IID. A inadequação da hipótese de IID poderia ser consequência ou da estrutura da população ou do efeito de um plano amostral complexo, ou ambos. Em qualquer dos casos, a estimativa v_0 da variância de $\hat{\theta}$ calculada sob a hipótese de observações IID se afastaria da variância de $\hat{\theta}$ sob o plano amostral (ou modelo) verdadeiro, denotada $V_{VERD}\left(\hat{\theta}\right)$. Note que $V_{VERD}\left(\hat{\theta}\right) = V_M\left(\hat{\theta}\right)$ na abordagem baseada em modelos e $V_{VERD}\left(\hat{\theta}\right) = V_p\left(\hat{\theta}\right)$ na abordagem de aleatorização, onde o plano amostral p considerado é diferente de AASC.

Para avaliar se este afastamento tende a ser grande ou pequeno, vamos considerar a distribuição de v_0 com relação à distribuição de aleatorização verdadeira (ou do modelo verdadeiro) e localizar v_0 com relação a esta distribuição de referência. Como em geral seria complicado obter esta distribuição, vamos tomar uma medida de centro ou locação da distribuição de v_0 e compará-la ao valor de $V_{VERD}\left(\hat{\theta}\right)$.

Podemos desta forma introduzir uma medida de efeito da especificação incorreta do plano amostral (ou do modelo) sobre a estimativa v_0 da variância do estimador $\hat{\theta}$.

Definição 4.2. O efeito da especificação incorreta do plano amostral (ou do modelo) sobre a estimativa v_0 da variância do estimador $\hat{\theta}$ é

$$EPA\left(\hat{\theta}, v_0\right) = \frac{V_{VERD}\left(\hat{\theta}\right)}{E_{VERD}\left(v_0\right)} \tag{4.2}$$

Desta forma, o $EPA\left(\hat{\theta}, v_0\right)$ mede a tendência de v_0 a subestimar ou superestimar $V_{VERD}\left(\hat{\theta}\right)$, variância *verdadeira* de $\hat{\theta}$. Valores de $EPA\left(\hat{\theta}, v_0\right)$ maiores que 1 sinalizam que v_0 tende a subestimar a variância $V_{VERD}\left(\hat{\theta}\right)$. Quanto mais afastado de 1 for o valor de $EPA\left(\hat{\theta}, v_0\right)$, mais incorreta será considerada a especificação do plano amostral ou do modelo considerado.

Enquanto a medida proposta por Kish (1965) baseia-se nas distribuições induzidas pela aleatorização dos planos amostrais comparados, o $EPA\left(\hat{\theta}, v_0\right)$ pode ser calculado com respeito a distribuições de aleatorização ou do modelo envolvido, bastando calcular V_{VERD} e E_{VERD} da definição (4.2) com relação à distribuição de referência correspondente.

Em geral, são esperadas as seguintes consequências sobre o EPA ao ignorar o plano amostral efetivamente adotado e admitir que a seleção da amostra foi AASC:

1. Ignorar os pesos em v_0 pode inflacionar o EPA;
2. Ignorar conglomeração em v_0 pode inflacionar o EPA;
3. Ignorar estratificação em v_0 pode reduzir o EPA.

Combinações destes aspectos num mesmo plano amostral complexo, resultando na especificação incorreta do plano amostral como se fosse AASC ou da hipótese de observações IID que levam ao estimador ingênuo de variância v_0, podem inflacionar ou reduzir o EPA. Nesses casos é difícil prever o impacto de ignorar o plano amostral (ou modelo) verdadeiro sobre a análise baseada em hipóteses IID. Por essa razão, é recomendável ao menos estimar os EPA antes de concluir a análise padrão, para então poder avaliar se há impactos importantes a considerar.

Embora pensadas com objetivos distintos, as duas definições de EPA consideradas têm em comum a ideia de que são medidas adimensionais, estabelecidas mediante comparação de um valor de interesse - $V_{VERD}\left(\hat{\theta}\right)$ - com valores de referência - $V_{AAS}\left(\hat{\theta}\right)$ ou $E_{VERD}\left(v_0\right)$ - que nos servem de baliza para avaliar o resultado, por serem quantidades habitualmente usadas ou fáceis de calcular. Essa ideia é recorrente na Estatística, de buscar referir quantidades novas a valores ou distribuições de referência conhecidos, como veremos na sequência deste livro.

Apesar dessa diferença de conceitos do EPA de Kish e do EPA ampliado, na prática vai quase sempre ser necessário estimar valores de EPA usando dados de uma amostra. O estimador habitualmente usado para os dois tipos de EPA será então o mesmo, dado por:

$$\widehat{EPA}\left(\hat{\theta}, v_0\right) = \frac{\widehat{V}_{VERD}\left(\hat{\theta}\right)}{v_0} \tag{4.3}$$

onde o numerador é um estimador não viciado (ou ao menos consistente) para a verdadeira variância do estimador sob o plano amostral de fato empregado para obter a amostra usada na inferência, e no denominador usamos a ideia de que a estimativa v_0 obtida na amostra é não viciada para estimar o valor esperado dessa quantidade aleatória sob o plano amostral de fato empregado na obtenção da amostra. Apesar de resultar numa estimativa de EPA idêntica para os dois conceitos, a interpretação do resultado será bastante distinta dependendo de qual dos dois parâmetros se buscava estimar. No caso do EPA de Kish ser o alvo, o interesse seria pela comparação de eficiência relativa entre o plano amostral efetivamente

empregado e AAS. No caso do EPA ampliado ser o alvo, o interesse é em avaliar quão viciada pode ser a aplicação de métodos padrões de forma ingênua para estimar a variância de estimadores dos parâmetros de interesse da análise.

Exemplo 4.2. Efeitos de plano amostral para estimação de médias na amostragem estratificada simples com alocação desproporcional

Neste exemplo consideramos uma população de $N = 749$ empresas, para as quais foram observadas as seguintes variáveis:

1. pessoal ocupado em 31/12/1994 - PO;
2. total de salários pagos no ano de 1994 - SAL;
3. receita total no ano de 1994 - REC.

A ideia é considerar o problema de estimar as médias populacionais das variáveis SAL e REC (variáveis de pesquisa consideradas neste exemplo), usando amostras estratificadas simples com alocação desproporcional, implicando em unidades amostrais com pesos desiguais numa situação bastante simples. A variável PO é a variável de estratificação. Como temos dados de todas as empresas da população, as médias populacionais das variáveis de pesquisa são conhecidas. Entretanto, para efeitos do presente exemplo, vamos supor que seriam desconhecidas e que amostragem seria usada para sua estimação.

Para estimar estas médias, as empresas da população foram divididas em dois estratos, definidos a partir da variável PO, conforme indicado na Tabela 4.2.

Tabela 4.2: Definição da estratificação da população de empresas

Estrato	Condição	Tamanho
1	empresas com PO > 21	161 empresas
2	empresas com PO ≤ 21	588 empresas

A ideia seria então selecionar, de cada um dos estratos, amostras aleatórias simples sem reposição de tamanhos $n_h = 30$ empresas, implicando em uso de *alocação igual* e em frações amostrais desiguais, em vista dos diferentes tamanhos populacionais dos estratos. Como o estrato 1 contém cerca de 21% das observações da população, a proporção de 50% das observações *da amostra total* sendo alocada no estrato 1 (das maiores empresas) da amostra é bem maior do que seria esperado sob amostragem aleatória simples da população de empresas.

Em consequência, a média amostral simples $\overline{y}$ de qualquer das duas variáveis de pesquisa (SAL ou REC) dada por

$$\overline{y} = \frac{1}{n}\sum_{h=1}^{2}\sum_{i\in s_h} y_i = \sum_{h=1}^{2}\frac{n_h}{n}\overline{y}_h$$

tenderia a superestimar a média populacional $\overline{Y}$ correspondente dada por

$$\overline{Y} = \frac{1}{N}\sum_{h=1}^{2}\sum_{i\in U_h} y_i = \sum_{h=1}^{2}\frac{N_h}{N}\overline{Y}_h$$

onde:

- y_i é o valor da variável de pesquisa y, $y \in \{SAL; REC\}$, para a $i-$ésima observação;
- n é o tamanho total da amostra;
- N o tamanho total da população;
- s_h e U_h representam os conjuntos de unidades na amostra e na população do estrato h respectivamente;
- n_h é o número de unidades na amostra do estrato h;
- N_h é o número de unidades na população do estrato h;
- $\overline{y}_h = \frac{1}{n_h}\sum_{i\in s_h} y_i$ é a média dos $y's$ na amostra no estrato $h \in \{1;2\}$ e
- $\overline{Y}_h = \frac{1}{N_h}\sum_{i\in U_h} y_i$ é a média populacional dos $y's$ no estrato $h \in \{1;2\}$.

É fácil verificar que o vício do estimador $\overline{y}$ para a média populacional $\overline{Y}$ é dado por:

$$B_{AES}(\overline{y}) = \sum_{h=1}^{2}\left(\frac{n_h}{n} - \frac{N_h}{N}\right)\overline{Y}_h$$

É fácil notar que sob alocação proporcional da amostra estratificada o vício seria nulo, o que não ocorre no presente exemplo. Como aqui a amostra estratificada carrega proporcionalmente mais unidades do estrato de maiores empresas, o valor deste viés será positivo. De fato, o valor exato deste vício pode ser calculado no presente exemplo onde conhecemos todos os dados das unidades da população.

Para amostras estratificadas simples como a considerada neste exemplo, o estimador não viciado tipo Horvitz-Thompson da média populacional $\overline{Y}$ seria dado por

$$\overline{y}_w = \sum_{h=1}^{2} W_h\overline{y}_h = \frac{1}{N}\sum_{h=1}^{2}\sum_{i\in s_h}\frac{N_h}{n_h}y_i$$

onde $W_h = N_h/N$ é a proporção de unidades da população no estrato $h \in \{1;2\}$.

Com a finalidade de ilustrar o cálculo do EPA, vamos considerar o estimador não viciado $\overline{y}_w$ e calcular sua variância sob o plano amostral realmente utilizado (amostragem estratificada simples - AES - com alocação igual, mas desproporcional). Essa variância poderá então ser comparada com o valor esperado (sob a distribuição induzida pelo plano amostral estratificado) do estimador da variância obtido sob a hipótese de amostragem aleatória simples.

Neste exemplo, a variância do estimador $\overline{y}_w$ pode ser obtida de duas formas: calculando a expressão da variância utilizando os dados de todas as unidades da população (que são conhecidos, mas admitidos desconhecidos para fins do exercício de estimação de médias via amostragem) e por simulação.

A variância de $\overline{y}_w$ sob a distribuição de aleatorização do plano amostral 'verdadeiro' - AES - é dada por - ver a seção 11.2.3 de Silva et al. (2020):

$$V_{AES}\left(\overline{y}_w\right) = \sum_{h=1}^{2} W_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)S_h^2 \tag{4.4}$$

onde $S_h^2 = \frac{1}{N_h-1}\sum_{i\in U_h}\left(y_i - \overline{Y}_h\right)^2$ é a variância populacional da variável de pesquisa y dentro do estrato h.

Conforme a seção 11.2.3 de Silva et al. (2020), esta variância pode ser estimada sem viés sob AES usando o estimador

$$\widehat{V}_{AES}\left(\overline{y}_w\right) = \sum_{h=1}^{2} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2 \tag{4.5}$$

onde $\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_i - \overline{y}_h)^2$.

Um estimador ingênuo da variância de $\overline{y}$ que seria usado sob amostragem aleatória simples é dado por

$$v_0 = \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{S}^2$$

onde $\widehat{S}^2 = \frac{1}{n-1} \sum_{h=1}^{2} \sum_{i \in s_h} (y_i - \overline{y})^2$.

Note que tanto o estimador média amostral simples $\overline{y}$ como o correspondente estimador de variância v_0 devem ser viciados no contexto da amostragem estratificada proposta neste exemplo. Mas a ideia do exemplo é justamente ilustrar os problemas que surgem quando se adota uma inferência padrão, dependente das hipóteses de observações IID (ou de AASC), como a que está disponível nos sistemas padrões de análise de dados, sem levar em conta os aspectos do plano amostral de fato usado para obtenção dos dados considerados na análise.

O valor do EPA foi também estimado por meio de simulação. O motivo para usar simulação é que a obtenção do valor esperado sob AES de v_0 de forma analítica é trabalhosa. Geramos então 500 amostras de tamanho $n = 60$, segundo o plano amostral estratificado considerado. Para cada uma das 500 amostras e cada uma das duas variáveis de pesquisa (SAL e REC) foram calculadas as estatísticas:

1. média amostral ($\overline{y}$);
2. estimativa ponderada da média populacional ($\overline{y}_w$);
3. estimativa da variância da estimativa ponderada da média ($\overline{y}$) considerando observações IID v_0;
4. estimativa da variância da estimativa ponderada da média ($\overline{y}_w$) considerando o plano amostral verdadeiro $\hat{V}_{AES}\left(\overline{y}_w\right)$.

Note que na apresentação dos resultados os valores dos salários foram expressos em milhares de Reais (R$ $1.000,00$) e os valores das receitas em milhões de Reais (R$ $1.000.000,00$). Como a população é conhecida, os parâmetros populacionais de interesse podem ser calculados, obtendo-se os valores na primeira linha da Tabela 4.3.

Tabela 4.3: Propriedades dos estimadores da média das variáveis de pesquisa

Quantidade de interesse	Salários	Receitas
Média Populacional	78,3	2,11
Média de estimativas de média AAS	163,3	4,18
Média de estimativas de média AES	77,8	2,10

Em contraste com os valores dos parâmetros populacionais, calculamos a média das médias amostrais não ponderadas ($\overline{y}$) dos salários e das receitas obtidas nas 500 amostras simuladas, obtendo os valores na segunda linha da Tabela 4.3. Como previsto, observamos um viés para cima na estimação das médias populacionais, da ordem de $108,5\%$ para os salários e de $98,1\%$ para as receitas, demonstrando o que era

esperado e o quanto seria inadequado neste exemplo ignorar os pesos amostrais na estimação da média populacional.

XXX Rever estes parágrafos após calcular erros padrões e estatísticas t de teste XXX para o vício do estimador média ponderada

Usamos também o estimador $\overline{y}_w$ para estimar a média dos salários e das receitas na população, obtendo para esse estimador as médias apresentadas na terceira linha da Tabela 4.3. Observamos ainda um pequeno vício relativo da ordem de $-0,6\%$ e $-0,5\%$ para os salários e receitas, respectivamente.

Note que o estimador $\overline{y}_w$ é não viciado sob o plano amostral adotado, entretanto o pequeno vício observado na simulação não pode ser ignorado pois é significantemente diferente de 0 ao nível de significância de 5%, apesar do tamanho razoável da simulação (500 replicações).

XXX

Além das estimativas pontuais, o interesse maior da simulação foi comparar valores de estimativas de variância e, consequentemente, de medidas do efeito do plano amostral. Como o estimador pontual dado pela média amostral não ponderada ($\overline{y}$) é grosseiramente viciado, não consideramos estimativas de variância para esse estimador, mas tão somente para o estimador não viciado dado pela média ponderada $\overline{y}_w$. Para esse último, consideramos dois estimadores de variância, a saber o estimador ingênuo sob a hipótese de AASC (dado por v_0, mas com o valor da estimativa de média usado no seu cálculo substituído por $\overline{y}_w$) e um estimador não viciado da variância sob o plano amostral AES efetivamente empregado $\hat{V}_{AES}\left(\overline{y}_w\right)$ dado pela expressão (4.5).

Como neste exercício a população é conhecida, podemos calcular $V_{AES}\left(\overline{y}_w\right)$ através das variâncias de y dentro dos estratos $h = 1, 2$ ou estimar essa variância através da simulação. Esses valores são apresentados respectivamente na primeira e segunda linhas da Tabela 4.4, para as duas variáveis de pesquisa consideradas.

Tabela 4.4: Propriedades dos estimadores de variância do estimador ponderado da média

Quantidade de interesse	Salários	Receitas
Variância do estimador AES	244	0,435
Média de estimativas de variância AES	243	0,505
Valor esperado AES do estimador AAS de variância	1.613	1,188
Média de estimativas de variância AAS	1.714	1,242

Os valores de $E_{VERD}\left[v_0\left(\overline{SAL}_w\right)\right]$ e de $E_{VERD}\left[v_0\left(\overline{REC}_w\right)\right]$ foram também calculados a partir das variâncias dentro e entre estratos na população, resultando nos valores na linha 3 da Tabela 4.4, e estimativas desses valores baseadas nas 500 amostras da simulação são apresentadas na linha 4 da Tabela 4.4. Os valores para o EPA foram calculados tanto com base nas estimativas de simulação como nos valores populacionais das variâncias, cujos cálculos estão ilustrados a seguir:

$EPA\left[\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right] =$

```
## 242,792358381168/1713,8055430219=0,142
```

$EPA\left[\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right] =$

```
## 0,505416004004538/1,24157646252246=0,407
```

$EPA\left[\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right] =$

```
## 244,17580090709/1613,3=0,151
```

$EPA\left[\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right] =$

```
## 0,434994493392157/1,188=0,366
```

A Tabela 4.5 resume os principais resultados deste exercício, para o estimador ponderado da média $\overline{y}_w$. Apesar das diferenças entre os resultados da simulação e suas contrapartidas calculadas considerando conhecidos os valores da população, as conclusões da análise são similares:

1. ignorar os pesos na estimação da média provoca vícios substanciais, que não podem ser ignorados; portanto, o uso do estimador simples de média ($\overline{y}$) é desaconselhado;
2. ignorar os pesos na estimação da variância do estimador ponderado $\overline{y}_w$ também provoca vícios substanciais, neste caso, superestimando a variância por ignorar o efeito de estratificação; os efeitos de plano amostral são substancialmente menores que 1 para as duas variáveis de pesquisa consideradas (salários e receita); portanto o uso do *estimador ingênuo* de variância v_0 é desaconselhado.

Essas conclusões são largamente aceitas pelos amostristas e produtores de dados baseados em pesquisas amostrais para o caso da estimação de médias e totais, e respectivas variâncias. Entretanto, ainda há exemplos de usos indevidos de dados amostrais nos quais os pesos e outros aspectos importantes dos planos amostrais complexos de fato empregados para obter os dados são ignorados nas análises, em particular para a estimação de variâncias associadas a estimativas pontuais de médias e de parâmetros de modelos. Tal situação se deve ao uso ingênuo de sistemas estatísticos padrões desenvolvidos para analisar amostras IID, sem a devida consideração dos pesos e outros aspectos da estrutura do plano amostral empregado para obtenção dos dados.

Tabela 4.5: Valores dos Efeitos de Plano amostral - EPA para as médias de Salário e Receita

Variável	Estimativa	Simulação	População
Salário	Variância	242,792	244,176
Salário	EPA	0,142	0,151
Receita	Variância	0,505	0,435
Receita	EPA	0,407	0,366

Observação. Neste exemplo não foi feito uso analítico dos dados e sim descritivo, onde é usual incorporar os pesos no cálculo de estimativas e variâncias. Não seria esperado usar um estimador ponderado para a média e não considerar os pesos no cálculo de variâncias, como fizemos neste exemplo.

Observação. O exemplo mostra que ignorar a estratificação ao calcular v_0 diminui o EPA.

Um outro exemplo relevante é utilizado a seguir para ilustrar o fato de que o conceito do EPA adotado aqui é mais abrangente do que o definido por Kish (1965), em particular porque a origem do efeito pode estar na estrutura da população e não no plano amostral usado para obter os dados.

Exemplo 4.3. População conglomerada com conglomerados de tamanho 2 - Skinner et al. (1989), p. 25

Considere uma população de unidades agrupadas em conglomerados de tamanho 2, isto é, onde as unidades (elementares ou de referência) estão grupadas em pares (exemplos de tais populações incluem pares de irmãos gêmeos, casais, jogadores numa dupla de vôlei de praia ou de tênis, etc.). Suponha que os valores de uma variável de pesquisa medida nessas unidades têm média θ e variância σ^2, além de uma correlação ρ entre os valores dentro de cada par (correlação intraclasse, veja Silva e Moura (1990), cap. 2 e Haggard

(1958)). Suponha que um único par é sorteado ao acaso da população e que os valores y_1 e y_2 são observados para as duas unidades do par selecionado. O modelo assumido M pode então ser representado como

$$M = \begin{cases} E_M(Y_i) = \theta \\ V_M(Y_i) = \sigma^2 \\ CORR_M(Y_1; Y_2) = \rho \end{cases} \qquad i = 1, 2.$$

Um estimador não viciado para θ é a média amostral dada por $\widehat{\theta} = (y_1 + y_2)/2$. Assumindo a (falsa) hipótese de que o esquema amostral é AASC de unidades individuais e não de pares ou, equivalentemente, que y_1 e y_2 são observações de variáveis aleatórias IID, a variância de $\widehat{\theta}$ é dada por

$$V_{AASC}\left(\widehat{\theta}\right) = \sigma^2/2$$

com um estimador não viciado dado por

$$v_0\left(\widehat{\theta}\right) = (y_1 - y_2)^2/4$$

Entretanto, na realidade a variância de $\widehat{\theta}$ é dada por

$$V_{VERD}\left(\widehat{\theta}\right) = V_M\left(\widehat{\theta}\right) = \sigma^2(1+\rho)/2$$

O valor esperado do estimador de variância $v_0\left(\widehat{\theta}\right)$ é dado por

$$E_{VERD}\left[v_0\left(\widehat{\theta}\right)\right] = \sigma^2(1-\rho)/2$$

Consequentemente, considerando as equações (4.1) e (4.2), tem-se que

$$EPA_{Kish}\left(\hat{\theta}\right) = 1 + \rho$$

e o efeito do plano amostral ampliado é

$$EPA\left(\hat{\theta}, v_0\right) = (1+\rho)/(1-\rho)$$

A Figura 4.1 mostra os valores de $EPA_{Kish}\left(\hat{\theta}\right)$ e $EPA\left(\hat{\theta}, v_0\right)$ para valores de ρ entre 0 e $0,8$. Como se pode notar, o efeito da especificação inadequada do plano amostral ou da estrutura populacional pode ser severo, com valores de $EPA\left(\hat{\theta}, v_0\right)$ chegando a 9. Um aspecto importante a notar é que o $EPA_{Kish}\left(\hat{\theta}\right)$ tem variação muito mais modesta que o $EPA\left(\hat{\theta}, v_0\right)$.

Este exemplo ilustra bem dois aspectos distintos do uso de medidas como as duas propostas para avaliar o efeito de plano amostral. O primeiro é que as duas medidas são distintas, embora as respectivas estimativas baseadas numa particular amostra seriam coincidentes, conforme discutido ao apresentar o estimador na expressão (4.3). No caso particular deste exemplo, o $EPA_{Kish}\left(\hat{\theta}\right)$ cresce pouco com o valor do coeficiente de correlação intraclasse ρ, o que implica que um plano amostral conglomerado como o adotado (seleção ao acaso de um par da população) seria menos eficiente que um plano amostral aleatório simples (seleção de duas unidades ao acaso da população), mas a perda de eficiência seria modesta. Isto se dá porque os tamanhos dos conglomerados são bem pequenos - duas unidades. Já se o interesse é medir, a posteriori, o efeito da má especificação do plano amostral no estimador de variância, o impacto, medido pelo $EPA\left(\hat{\theta}, v_0\right)$,

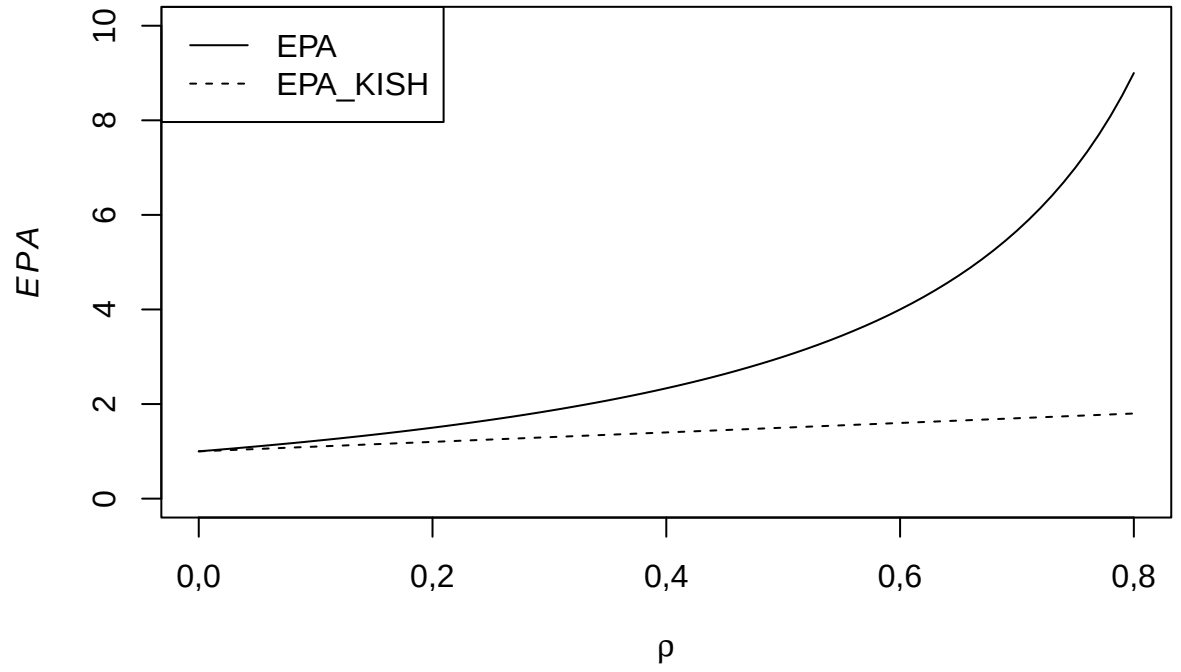

Figura 4.1: Valores de EPA e EPA de Kish para conglomeração

seria muito maior. Assim, a imprudência na estimação de variâncias (e erros padrões, intervalos de confiança, testes de significância, etc.) teria consequências bem mais sérias para as análises.

Vale ainda notar que o $EPA\left(\hat{\theta}, v_0\right)$ mede o impacto da má especificação do plano amostral ou do modelo para a estrutura populacional. Neste exemplo, ignorar a estrutura da população (o fato de que as observações são pareadas) poderia provocar subestimação da variância do estimador de média, que seria tanto maior quanto maior fosse o coeficiente de correlação intraclasse ρ. Efeitos como esse são comuns devido ao planejamento amostral, mesmo em populações onde a conglomeração é imposta artificialmente pelo amostrista.

4.4 Efeitos sobre Intervalos de Confiança e Testes de Hipóteses Uniparamétricos

A partir da estimativa pontual $\hat{\theta}$ de um parâmetro θ (da população finita ou do modelo de superpopulação) é possível construir um intervalo de confiança de nível de confiança aproximado $(1-\alpha)$ a partir da distribuição assintótica de

$$t_0 = \frac{\hat{\theta} - \theta}{v_0^{1/2}}$$

que, sob a hipótese de que as observações são IID, tem distribuição bem aproximada pela $N(0;1)$ para amostras grandes.

Neste caso, um intervalo de nível de confiança aproximado $(1-\alpha)$ é dado por $\left[\hat{\theta} \mp z_{\alpha/2} v_0^{1/2}\right]$, onde $z_{\alpha/2}$ é o

quantil que deixa área igual a $\alpha/2$ à sua direita sob a curva da função de densidade da distribuição normal padrão ou $N(0;1)$.

Vamos agora analisar o efeito de um plano amostral complexo sobre o intervalo de confiança. Nesse caso, a distribuição que é aproximadamente normal é a da estatística pivô dada por:

$$t = \frac{\hat{\theta} - \theta}{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}}$$

Por outro lado, para obter a variância da distribuição assintótica de t note que

$$t_0 = \frac{\hat{\theta} - \theta}{v_0^{1/2}} = \frac{\hat{\theta} - \theta}{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}} \times \frac{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}}{v_0^{1/2}} = t \times \left[\widehat{EPA}\left(\hat{\theta}, v_0\right)\right]^{1/2}$$

Como a distribuição de t tende para uma $N\left(0;1\right)$, a variância assintótica de t_0 é aproximadamente igual ao quadrado do limite do segundo fator no lado direito da expressão, isto é, a $EPA\left(\hat{\theta}, v_0\right)$. Logo temos que a distribuição assintótica de t_0 é dada por

$$t_0 \sim N\left[0; EPA\left(\hat{\theta}, v_0\right)\right]$$

Dependendo do valor de $EPA\left(\hat{\theta}, v_0\right)$, o intervalo de confiança baseado na distribuição assintótica verdadeira de t_0 pode ser bem distinto daquele baseado na distribuição assintótica obtida sob a hipótese de observações IID. Em geral, a probabilidade de cobertura assintótica do intervalo $\left[\hat{\theta} \mp z_{\alpha/2} v_0^{1/2}\right]$ será aproximadamente igual a

$$2\Phi\left(z_{\alpha/2} / \left[EPA\left(\hat{\theta}, v_0\right)\right]^{1/2}\right) - 1$$

onde Φ é a função de distribuição acumulada da distribuição normal padrão. Calculamos esta probabilidade para alguns valores do EPA, que apresentamos na Tabela 4.6.

Tabela 4.6: Probabilidades de cobertura para níveis nominais de 0,95 e 0,99

$EPA\left(\hat{\theta}, v_0\right)$	$1-\alpha = 0,95$	$1-\alpha = 0,99$
0,90	0,96	0,99
0,95	0,96	0,99
1,0	0,95	0,99
1,5	0,89	0,96
2,0	0,83	0,93
2,5	0,78	0,90
3,0	0,74	0,86
3,5	0,71	0,83
4,0	0,67	0,80

À medida que o valor do $EPA\left(\hat{\theta}, v_0\right)$ aumenta, a probabilidade real de cobertura diminui, sendo menor que o valor nominal para valores de $EPA\left(\hat{\theta}, v_0\right)$ maiores que 1, como esperado, já que nestes casos o estimador ingênuo v_0 subestima a variância verdadeira do estimador.

Utilizando a correspondência existente entre intervalos de confiança e testes de hipóteses, podemos derivar os níveis de significância nominais e reais subtraindo de 1 os valores da Tabela 4.6. Por exemplo, para $\alpha = 0,05$ e $EPA\left(\hat{\theta}, v_0\right) = 2$, o nível de significância real seria aproximadamente $1 - 0,83 = 0,17$.

Exemplo 4.4. Teste de hipótese sobre proporção

Vamos considerar um exemplo hipotético de teste de hipótese sobre uma proporção, semelhante ao de Sudman (1976), apresentado na p. 196 de Lehtonen e Pahkinen (1995). Uma amostra de $m = 50$ conglomerados é extraída de uma grande população de empresas industriais (conglomerados de empregados). Suponhamos que cada empresa $i = 1, \ldots, 50$ da amostra tenha $n_i = 20$ empregados. O tamanho total da amostra de empregados (unidades elementares) é $n = \sum_{i \in s} n_i = 1.000$. Queremos estudar o acesso dos trabalhadores das empresas a planos de saúde.

Usando-se conhecimento do ano anterior, foi estabelecida a hipótese de que a proporção de trabalhadores cobertos por planos de saúde é 80%, ou seja $H_0 : p = p_0 = 0,8$. Vamos adotar o nível de significância $\alpha = 5\%$.

A estimativa obtida na pesquisa foi $\widehat{p} = n_A/n = 0,84$, onde $n_A = 840$ é o número de trabalhadores na amostra com acesso a planos de saúde. Ignorando o plano amostral e a conglomeração das unidades elementares na população, podemos considerar um teste binomial e usar a aproximação normal $N(0;1)$ para a estatística de teste

$$Z_{AASC} = |\widehat{p} - p_0| / \sqrt{p_0 \left(1 - p_0\right) / n} \tag{4.6}$$

onde o denominador é o desvio padrão da estimativa $\widehat{p}$ sob a hipótese nula, e consideramos como hipótese alternativa $H_A : \ p \neq 0,8$.

Vamos calcular o valor da estatística Z_{AASC}, supondo que tenha sido usada amostragem aleatória simples com reposição (AASC) de empregados. Vamos também considerar uma abordagem baseada no plano amostral de conglomerados. O desvio padrão de $\widehat{p}$, no denominador de Z_{AASC}, será baseado na hipótese de distribuição binomial, com tamanhos amostrais diferentes para as duas abordagens.

Para o teste baseado na AASC, ignoramos a conglomeração e usamos na fórmula do desvio padrão do estimador $\widehat{p}$ da proporção o tamanho total da amostra de unidades elementares (empregados), isto é, $n = 1.000$. O valor da estatística de teste Z_{AASC} definida em (4.6) fica, portanto, igual a

$$Z_{AASC} = |0,84 - 0,8| / \sqrt{0,8 \left(1 - 0,8\right) / 1.000} = 3,162 > z_{0,975} = 1,96 \tag{4.7}$$

onde $\sqrt{0,8 \left(1 - 0,8\right) / 1.000} = 0,0126$ é o desvio padrão de $\widehat{p}$ sob a hipótese nula e $z_{0,975}$ é o quantil que deixa área de $0,975$ à sua esquerda sob a densidade da distribuição normal padrão. Este resultado indica a rejeição da hipótese H_0 ao nível de significância estabelecido.

Por outro lado, é razoável admitir que se uma empresa oferece cobertura por plano de saúde, cada empregado dessa empresa terá acesso ao benefício. Essa é uma informação importante que foi ignorada no teste anterior. De fato, selecionar mais de uma pessoa numa empresa não aumenta nosso conhecimento sobre a cobertura por plano de saúde no local. Portanto, o *tamanho efetivo* da amostra é $\overline{n} = 50$, em contraste com o valor 1.000 usado no teste anterior. O termo *tamanho efetivo* foi introduzido em Kish (1965) para designar o

tamanho de uma amostra aleatória simples necessário para estimar p com a mesma precisão obtida por uma amostra conglomerada de tamanho n (neste caso, igual a 1.000) unidades elementares.

Usando o tamanho efetivo de amostra, temos a estatística de teste baseada no plano amostral verdadeiro

$$Z_{VERD} = |\widehat{p} - p_0| / \sqrt{p_0 \left(1 - p_0\right) / \overline{n}} = |0,84 - 0,8| / \sqrt{0,8 \left(1 - 0,8\right) / 50} = 0,707$$

onde o valor $\sqrt{0,8 \left(1 - 0,8\right) / 50} = 0,0566$ é muito maior que o valor do desvio padrão empregado no cálculo da estatística de teste Z_{AASC}. Em consequência, o valor observado de Z_{VERD} é menor que o limite de aceitação $z_{0,975} = 1,96$, e a nova estatística de teste indica a não rejeição da mesma hipótese nula.

Neste exemplo, portanto, se verifica que ignorar a conglomeração pode induzir a uma decisão incorreta de rejeitar a hipótese nula, quando a mesma não seria rejeitada se o plano amostral de fato empregado fosse corretamente incorporado na análise. Efeitos desse tipo são mais difíceis de antecipar para inferência analítica, particularmente quando os planos amostrais empregados envolvem combinações de estratificação, conglomeração, probabilidades desiguais de seleção e calibração de pesos. Por esse motivo, a recomendação é procurar sempre considerar o plano amostral na análise, ao menos como forma de verificar se as conclusões obtidas por formas ingênuas de análise ignorando os pesos e plano amostral seriam mantidas.

4.5 Efeitos Multivariados de Plano Amostral

O conceito de efeito de plano amostral introduzido em (4.2) é relativo a inferências sobre um único parâmetro θ. Consideremos agora o problema de estimação de um vetor $\mathbf{\Theta}$ contendo K parâmetros. Seja $\hat{\mathbf{\Theta}}$ um estimador de $\mathbf{\Theta}$ e seja $\mathbf{V}_0$ um estimador da matriz $K \times K$ de covariância de $\hat{\mathbf{\Theta}}$, baseado nas hipóteses de independência e igualdade de distribuição das observações - IID, ou equivalentemente, de amostragem aleatória simples com reposição - AASC. É possível generalizar a equação (4.2), definindo o *Efeito Multivariado do Plano Amostral* de $\hat{\mathbf{\Theta}}$ e seu estimador ingênuo de variância $\mathbf{V}_0$ como:

$$\mathbf{EMPA}(\hat{\mathbf{\Theta}}, \mathbf{V}_0) = \mathbf{\Delta} = \left[E_{VERD}\left(\mathbf{V}_0\right)\right]^{-1} \times \mathbf{V}_{VERD}(\hat{\mathbf{\Theta}}) \tag{4.8}$$

onde $E_{VERD}\left(\mathbf{V}_0\right)$ é o valor esperado de $\mathbf{V}_0$ e $\mathbf{V}_{VERD}(\hat{\mathbf{\Theta}})$ é a matriz de covariância de $\hat{\mathbf{\Theta}}$, ambos calculados com respeito à distribuição induzida pelo plano amostral efetivamente utilizado (verdadeiro), ou alternativamente sob o *modelo correto*.

Os autovalores $\delta_1 \geq \ldots \geq \delta_K$ da matriz $\mathbf{\Delta}$ são denominados *efeitos generalizados do plano amostral*. A partir deles, e utilizando resultados padrões de teoria das matrizes (Johnson e Wichern (1988), p.64) é possível definir limites para os efeitos (univariados) do plano amostral para qualquer combinação linear $\mathbf{c}'\widehat{\mathbf{\Theta}}$ das componentes de $\widehat{\mathbf{\Theta}}$. Segundo Skinner (1989a), pag. 43, temos os seguintes resultados:

$$\begin{aligned} \delta_1 &= \max\left\{EPA(\mathbf{c}'\widehat{\mathbf{\Theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c})\right\} \\ \delta_K &= \min\left\{EPA(\mathbf{c}'\widehat{\mathbf{\Theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c})\right\} \end{aligned}$$

No caso particular onde $\mathbf{\Delta} = \mathbf{I}_K$, temos $\delta_1 = \ldots = \delta_K = 1$ e os *efeitos (univariados) do plano amostral* das combinações lineares para componentes de $\hat{\mathbf{\Theta}}$ são todos iguais a 1.

Para ilustrar esse conceito, vamos reconsiderar o Exemplo 4.2 de estimação de médias com amostragem estratificada desproporcional anteriormente apresentado, mas agora considerando a natureza multivariada do problema (há duas variáveis de pesquisa - SAL e REC).

Exemplo 4.5. Efeitos Multivariados do Plano Amostral para as médias de Salários e de Receitas

Vamos considerar as variáveis Salário (em R$ 1.000) e Receita (em R$ 1.000.000) definidas na população de empresas do Exemplo 4.2 e calcular a matriz $\mathbf{EMPA}\left(\hat{\boldsymbol{\Theta}}, \mathbf{V}_0\right)$, onde $\hat{\boldsymbol{\Theta}} = \left(\overline{SAL}_w, \overline{REC}_w\right)'$. Neste exemplo, os dados populacionais são conhecidos e, portanto, podemos calcular a covariância dos estimadores $\left(\overline{SAL}_w, \overline{REC}_w\right)$. Usando a mesma notação do Exemplo 4.2, temos que:

$$COV_{AES}(\overline{SAL}_w, \overline{REC}_w) = \sum_{h=1}^{2} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{SAL,REC}^{(h)}$$

onde

$$S_{SAL,REC}^{(h)} = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(SAL_i - \overline{SAL}_h\right)\left(REC_i - \overline{REC}_h\right)$$

Substituindo os valores conhecidos na população das variáveis SAL_i e REC_i, obtemos para esta covariância o valor

$$COV_{AES}\left(\overline{SAL}_w, \overline{REC}_w\right) = 3,2358$$

e portanto a matriz de variância $\mathbf{V}_{AES}\left(\overline{SAL}_w, \overline{REC}_w\right)$ dos estimadores ponderados da média fica igual a

```
       SAL  REC
SAL 244,18 3,24
REC   3,24 0,43
```

onde os valores das variâncias usadas nos cálculos com a expressão (4.8) foram os obtidos no Exemplo 4.2 e coincidem, respectivamente, com os valores usados nos numeradores de $EPA\left(\overline{SAL}_w\right)$ e de $EPA\left(\overline{REC}_w\right)$ lá apresentados. Para calcular o $\mathbf{EMPA}(\hat{\boldsymbol{\Theta}}, \mathbf{V}_0)$ é preciso agora obter $E_{VERD}\left(\mathbf{V}_0\right)$.

Neste exemplo, a matriz de efeito do plano amostral $\mathbf{EMPA}\left(\hat{\boldsymbol{\Theta}}, \mathbf{V}_0\right) = \boldsymbol{\Delta}$ pode também ser calculada através de simulação, de modo análogo ao que foi feito no Exemplo 4.2. Para isto, foram utilizadas as 500 amostras de tamanho 60 segundo o plano amostral descrito no Exemplo 4.2. Para cada uma das 500 amostras foram calculadas estimativas:

1. da variância da média amostral ponderada do salário e da receita assumindo observações IID;
2. da covariância entre médias ponderadas do salário e da receita assumindo observações IID;
3. da variância da média amostral ponderada do salário e da receita considerando o plano amostral verdadeiro;
4. da covariância entre médias ponderadas do salário e da receita considerando o plano amostral verdadeiro.

A partir da simulação foram obtidos os seguintes resultados:

- A matriz de covariância das médias amostrais ponderadas de salário e da receita, supondo observações IID $E_{AES}\left(\mathbf{V}_0\right)$:

```
      SAL  REC
SAL 1713,8 26,3
REC   26,3  1,2
```

- A matriz de covariância das médias ponderadas de salário e da receita considerando o plano amostral verdadeiro $\mathbf{V}_{AES}\left(\hat{\boldsymbol{\Theta}}\right)$:

```
      SAL REC
SAL 242,8 3,1
REC   3,1 0,5
```

- A matriz Δ definida em (4.8)

$$\mathbf{\Delta} = \left[E_{AES}\left(\mathbf{V}_0\right)\right]^{-1} \times \mathbf{V}_{AES}(\hat{\theta})$$

```
        sal      rec
[1,]  0,153 -0,00656
[2,] -0,740  0,54597
```

Os autovalores 1 e 1,02 de $\mathbf{\Delta}$ fornecem os efeitos generalizados do plano amostral.

Da mesma forma que o $EPA\left(\hat{\theta}, v_0\right)$ definido em (4.2) para o caso uniparamétrico foi utilizado para corrigir níveis de confiança de intervalos e níveis de significância de testes, o $\mathbf{EMPA}(\hat{\mathbf{\Theta}}, \mathbf{V}_0)$ definido em (4.8) pode ser utilizado para corrigir níveis de regiões de confiança e também de significância de testes de hipóteses no caso multiparamétrico. Para ilustrar, vamos considerar o problema de testar a hipótese $H_0 : \mathbf{\Theta} = \mathbf{\Theta}_0$, onde $\mathbf{\Theta}$ é o vetor de médias de um vetor de variáveis de pesquisa $\mathbf{y}$. A estatística de teste usualmente adotada para este caso - ver Johnson e Wichern (1988), p.171 - é a T^2 de Hottelling dada por

$$T^2 = n\left(\overline{\mathbf{y}} - \mathbf{\Theta}_0\right)' \mathbf{S}_y^{-1}\left(\overline{\mathbf{y}} - \mathbf{\Theta}_0\right) \tag{4.9}$$

onde:

$$\begin{array}{rcl} \overline{\mathbf{y}} & = & \frac{1}{n}\sum_{i\in s}\mathbf{y}_i \\ \mathbf{S}_y & = & \frac{1}{n-1}\sum_{i\in s}\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)' \\ \mathbf{\Theta}_0 & = & \left(\theta_{0;1}, \theta_{0;2}, \ldots, \theta_{0;K}\right)' \end{array}$$

Sob H_0, se as observações $\mathbf{y}_i$ forem IID normais, a estatística T^2 tem a distribuição $\frac{(n-1)K}{n-K}F(K; n-K)$, onde $F(K; n-K)$ denota uma variável aleatória com distribuição F com K e $(n-K)$ graus de liberdade. Mesmo se as observações $\mathbf{y}_i$ não tiverem distribuição normal, T^2 tem distribuição assintótica $\chi^2(K)$ quando $n \to \infty$, Johnson e Wichern (1988), p.191.

Contudo, se for utilizado um plano amostral complexo, T^2 tem aproximadamente a distribuição da variável $\sum_{k=1}^{K}\delta_k Z_k^2$, onde $Z_1, \ldots, Z_K$ são variáveis aleatórias independentes com distribuição normal padrão e os δ_k são os autovalores da matriz $\mathbf{\Delta} = \mathbf{\Sigma}_{AASC}^{-1}\mathbf{\Sigma}_p$, onde $\mathbf{\Sigma}_{AASC} = E_p(\mathbf{S}_y/n)$ e $\mathbf{\Sigma}_p = V_p(\overline{\mathbf{y}})$.

Vamos analisar o efeito do plano amostral sobre o nível de significância deste teste. Para simplificar, consideremos o caso em que $\delta_1 = \ldots = \delta_K = \delta$. Neste caso, o nível de significância real é dado aproximadamente por

$$P\left(T^2 > \chi_\alpha^2(K)/\delta\right) \tag{4.10}$$

onde $\chi_\alpha^2(K)$ é o quantil superior α de uma distribuição χ^2 com K graus de liberdade, isto é, o valor tal que $P\left[\chi^2(K) > \chi_\alpha^2(K)\right] = \alpha$.

A Tabela 4.7 apresenta os níveis de significância reais para $\alpha = 5\%$ para vários valores de K e δ. Mesmo quando os valores dos δ_k são distintos, os valores da Tabela 4.7 podem ser devidamente interpretados. Para

isso, consideremos o valor-p do teste da hipótese $H_0 : \boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, sob a hipótese de AASC e sob o plano amostral efetivamente utilizado. Por definição este valor é dado por

$$\text{valor-}p_{AASC}\left(\overline{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right)' \boldsymbol{\Sigma}_{AASC}^{-1}\left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right)\right]$$

e H_0 é rejeitada com nível de significância α se valor-$p_{AASC} < \alpha$.

O verdadeiro valor-p_{AASC} pode ser definido analogamente como

$$\text{valor-}p_{VERD}\left(\overline{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right)' \boldsymbol{\Sigma}_{VERD}^{-1}\left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right)\right] \tag{4.11}$$

Os valores na Tabela 4.7 podem ser usados para quantificar a diferença entre estes valores-p. Consideremos a região crítica do teste de nível α baseado na hipótese de AASC:

$$\begin{array}{rcl} RC_{AASC}\left(\overline{\mathbf{y}}\right) & = & \left\{\overline{\mathbf{y}}: \quad \left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right)' \boldsymbol{\Sigma}_{AASC}^{-1}\left(\overline{\mathbf{y}} - \boldsymbol{\Theta}_0\right) > \chi_\alpha^2\left(K\right)\right\} \\ & = & \left\{\overline{\mathbf{y}}: \quad \text{valor-}p_{AASC}\left(\overline{\mathbf{y}}\right) < \alpha\right\} \end{array} \tag{4.12}$$

Pode-se mostrar - ver Skinner (1989a), pag. 43 - que o máximo de valor-$p_{VERD}\left(\overline{\mathbf{y}}\right)$ quando $\overline{\mathbf{y}}$ pertence à $RC_{AASC}\left(\overline{\mathbf{y}}\right)$ é dado por:

$$\max_{\overline{\mathbf{y}} \in RC_{AASC}(\overline{\mathbf{y}})} \text{valor-}p_{VERD}\left(\overline{\mathbf{y}}\right) = P\left[\chi^2(K) > \chi_\alpha^2(K)/\delta_1\right] \tag{4.13}$$

Observe que o segundo membro de (4.13) é da mesma forma que o segundo membro de (4.10). Logo, os valores da Tabela 4.7 podem ser interpretados como valores máximos valor-$p_{VERD}\left(\overline{\mathbf{y}}\right)$ para $\overline{\mathbf{y}}$ na região $RC_{AASC}\left(\overline{\mathbf{y}}\right)$, considerando-se δ_1 no lugar de δ.

Tabela 4.7: Níveis de significância verdadeiros, em porcentagem, do teste T^2 para o nível nominal de 5 por cento assumindo autovalores iguais a δ

δ	K=1	K=2	K=3	K=4
0,9	4	4	3	3
1,0	5	5	5	5
1,5	11	14	16	19
2,0	17	22	27	32
2,5	22	30	37	44
3,0	26	37	46	53

Os valores apresentados na Tabela 4.7 mostram que, mesmo quando os efeitos do plano amostral são modestos (valores iguais a 1,5, por exemplo) os níveis de significância reais são bem maiores que o nível nominal especificado. Além disso, esses níveis crescem rapidamente com o número de parâmetros K considerados na estatística de teste. Vale então reforçar o alerta aos analistas de dados para que não deixem de levar em conta o plano amostral empregado para obter os dados usados em suas análises, sob pena de tomarem decisões incorretas com frequências muito acima das que seriam aceitáveis em aplicações dos testes de hipóteses considerados.

4.6 Laboratório de R

Utilizando o R, obtemos a seguir alguns resultados descritos nos Exemplos 4.2 e 4.5. Na simulação, usamos o pacote *sampling*, Tillé e Matei (2016), para gerar amostras estratificadas de tamanho $n = 60$, com estratos definidos na Tabela 4.2, para obter os valores nas Tabelas 4.3 e 4.4.

```
# carrega library
library(survey)
# carrega dados
#library(anamco)
popul_dat <- readRDS(file = 'data/popul.rds') # carrega dados
N <- nrow(popul_dat)
n1 <- 30
n2 <- 30
nh = c(n1, n2)
n <- sum(nh)
Nh <- table(popul_dat$estrat)
fh <- nh/Nh
Wh <- Nh/N
f <- n/N
popul_dat$sal <- popul_dat$sal/1000
popul_dat$rec <- popul_dat$rec/1e+06
library(sampling)
# define espaços para salvar resultados
est_aas <- c(0, 0)
est_aes <- c(0, 0)
cov_mat_aas_est <- matrix(0, 2, 2)
cov_mat_aes_est <- matrix(0, 2, 2)
set.seed(123)
# gera amostras com dois estratos de tamanho 30
for (i in 1:500) {
  s <- strata(popul_dat, "estrat", c(30, 30), method = "srswor")
  dados <- getdata(popul_dat, s)
  # média amostral de salário e de receita
  est_aas <- est_aas + c(mean(dados$sal), mean(dados$rec))
  # estimador v0
  cov_mat_aas_est <- cov_mat_aas_est + (1 - f) * cov(cbind(dados$sal,
    dados$rec))/n

  # yhat_aes estimador não viciado
  popul_plan <- svydesign(~1, strata = ~estrat, data = dados,
    fpc = ~Prob)
  # estimador não viciado da média de salario e receita
  sal_rec_aes_est <- svymean(~sal + rec, popul_plan)
  est_aes <- est_aes + coef(sal_rec_aes_est)
  cov_mat_aes_est <- cov_mat_aes_est + attr(sal_rec_aes_est,
    "var")
}
```

```r
# média populacional
med_pop <- round(c(mean(popul_dat$sal), mean(popul_dat$rec)), 3)


# Calcula médias das estimativas na simulação

## Média das estimativas pontuais para as 500 amostras aas
mean_est_aas <- round(est_aas/500,3)
mean_est_aas
```

```
## [1] 163,30   4,18
```

```r
## Média das estimativas pontuais para as 500 amostras aes
mean_est_aes <- round(est_aes/500,3)
mean_est_aes
```

```
##  sal  rec
## 77,8  2,1
```

```r
# Média das estimativas de matriz de covariância para as 500
# amostras aas
mean_cov_mat_aas_est <- round(cov_mat_aas_est/500, 3)
mean_cov_mat_aas_est
```

```
##         [,1]  [,2]
## [1,] 1713,8 26,28
## [2,]   26,3  1,24
```

```r
# Média das estimativas de matriz de covariância para as 500
# amostras aes
mean_cov_mat_aes_est <- round(cov_mat_aes_est/500, 3)
mean_cov_mat_aes_est
```

```
##       sal   rec
## sal 242,8 3,103
## rec   3,1 0,505
```

```r
## Matriz de covariância populacional
mat_cov_pop <- by(popul_dat, popul_dat$estrat, function(t) var(cbind(t$sal,
  t$rec)))



## Matriz de covariância considerando o plano amostral
## verdadeiro
mat_cov_aleat_verd <- (Wh[1]^2 * (1 - fh[1])/nh[1]) * mat_cov_pop[[1]] +
  (Wh[2]^2 * (1 - fh[2])/nh[2]) * mat_cov_pop[[2]]
mat_cov_aleat_verd <- round(mat_cov_aleat_verd,3)


## estimativa de efeitos generalizados do plano amostral
```

```
DELTA = solve(mean_cov_mat_aas_est) %*% mean_cov_mat_aes_est
epa <-round(eigen(DELTA)$values,3)
```

Exemplo 4.6. Teste da igualdade de médias para duas populações

Para exemplificar o material descrito na Seção 4.4, vamos utilizar o data frame `amolim`, contendo dados da `Amostra do Censo Experimental de Limeira`.

```
# Carrega dados
amolim <- readRDS("./data/amolim.rds")
dim(amolim)
```

```
## [1] 706  14
```

```
names(amolim)
```

```
##  [1] "setor"  "np"     "domic"  "sexo"   "renda"  "lrenda" "raca"   "estudo"
##  [9] "idade"  "na"     "peso"   "domtot" "peso1"  "pesof"
```

- Objeto de desenho para os dados da Amostra do Censo Experimental de Limeira:

```
library(survey)
amolim.des <- svydesign(data=amolim,
                        id=~setor+domic,
                        weights=~pesof)
```

- Vamos estimar, a renda média por raça:

```
svyby(~renda, ~raca, amolim.des, svymean)
```

```
##   raca  renda    se
## 1    1 110406 11262
## 2    2  73560  8207
```

- Vamos estimar, a renda média por sexo:

```
svyby(~renda, ~sexo, amolim.des, svymean)
```

```
##   sexo  renda    se
## 1    1 108746 11696
## 2    2  40039  4042
```

- Vamos testar a igualdade de rendas por sexo:

```
svyttest(renda ~ sexo, amolim.des)
```

```
##
##  Design-based t-test
##
## data:  renda ~ sexo
## t = -6, df = 23, p-value = 0,000005
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -92694 -44719
## sample estimates:
```

```
## difference in mean
##             -68707
```

- Vamos testar a igualdade de rendas por raça:

```
svyttest(renda ~ raca, amolim.des)
```

```
##
##  Design-based t-test
##
## data:  renda ~ raca
## t = -4, df = 23, p-value = 0,0006
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -56039 -17653
## sample estimates:
## difference in mean
##             -36846
```

Referências

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292.

Bishop, Y. M. M.; Fienberg, S. E. e Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of statistics*, *5*(3).

Casella, G. e Berger, R. L. (2010). *Inferência Estatística*. Cengage Learning.

Cassel, C. M.; Särndal, C. E. e Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. John Wiley & Sons.

Chambers, R. L. e Skinner, C. J. (Orgs.). (2003). *Analysis of Survey Data*. John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed, p. 428). John Wiley & Sons.

Deming, W. E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, *51*, 24–53.

Deville, J. C. e Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley.

Haggard, E. A. (1958). *Intraclass Correlation and the Analysis of Variance*. Dryden Press.

Hansen, M. H.; Hurwitz, W. N. e Madow, W. G. (1953). *Sample Survey Methods and Theory*. John Wiley & Sons.

Heeringa, S. G.; West, B. T. e Berglund, P. A. (2010). *Applied Survey Data Analysis*. Taylor & Francis. Disponível em: https://books.google.com.br/books?id=QNmIvnTLlxcC (Acesso em 1/12/2021)

IBGE. (1985). *Amostra de Uso Público do Censo Demográfico de 1980 - Metodologia e Manual do Usuário*. IBGE.

IBGE. (2021). *Sobre a alteração do método de calibração dos fatores de expansão da PNAD Contínua* . Instituto Brasileiro de Geografia e Estatística - IBGE; IBGE, Diretoria de Pesquisas. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101882 (Acesso em 28/11/2021)

Johnson, R. A. e Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Kalton, G. (1983). *Compensating for missing survey data*. The University of Michigan, Institute for Social Research, Survey Research Center.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.

Lehtonen, R. e Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley & Sons.

Little, R. J. A. e Rubin, D. B. (2002). *Statistical Analysis with missing data*. John Wiley & Sons.

Lumley, T. (2006). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R* (p. 276). John Wiley & Sons.

Lumley, T. (2021). *survey: Analysis of Complex Survey Samples*. Disponível em: https://CRAN.R-project.org/package=survey R package version >3.5.0, (Acesso em 1/12/2021)

Magalhães, M. N. e Lima, A. C. P. (2015). *Noções de Probabilidade e Estatística* (7ª edição, 3ª reimpressão

revista). Edusp - Editora da Universidade de São Paulo.
Mahalanobis, P. C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhya*, *4*, 511–531.
Mahalanobis, P. C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London B*, *231*, 329–451.
Montanari, G. E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, *55*, 191–202.
NIC.br. (2020). *Pesquisa Sobre o Uso das Tecnologias da Informação e da Comunicação no Brasil* (p. 344). Disponível em: https://cetic.br/media/docs/publicacoes/2/20201123121817/tic_dom_2019_livro_eletronico.pdf (Acesso em 1/12/2021)
Pessoa, D. G. C. e Silva, P. L. N. (1998). *Análise de dados amostrais complexos* (p. 170). Associação Brasileira de Estatística.
Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, *61*, 317–337.
Quenoille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, *20*, p. 355–375.
Quenoille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.
Rao, J. N. K.; Wu, C. F. J. e Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(2), 209–217.
Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387.
Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.
Särndal, C. E.; Swensson, B. e Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.
Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* (p. 430). Chapman & Hall / CRC.
Shah, B. V.; Folsom, R. E.; LaVange, L. M.; Wheeless, S. C.; Boyle, K. E. e Williams, R. L. (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*.
Silva, P. L. N. (1996). *Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys* [Tese de doutorado]. University of Southampton, Department of Social Statistics.
Silva, P. L. N.; Bianchini, Z. M. e Dias, A. J. R. (2020). *Amostragem: teoria e prática usando R*. Disponível em: https://amostragemcomr.github.io/livro/ (Acesso em 1/12/2021)
Silva, P. L. N. e Moura, F. A. S. (1990). *Efeitos de conglomeração da malha setorial do censo demográfico 80* (Série Textos para Discussão nº 32). IBGE, Diretoria de Pesquisas.
Skinner, C. J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys* (p. 23–57). John Wiley & Sons.
Skinner, C. J.; Holt, D. e Smith, T. M. F. (Orgs.). (1989). *Analysis of Complex Surveys*. John Wiley & Sons.
Sudman, S. (1976). *Applied Sampling*. Academic Press.
Sugden, R. A. e Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, *71*, 495–506.
Thompson, S. K. (1992). *Sampling*. John Wiley & Sons.
Tillé, Y. e Matei, A. (2016). *sampling: Survey Sampling*. Disponível em: https://CRAN.R-project.org/package=sampling R package version 2.8
US Bureau of Labour Statistics. (2020). *Consumer price index - Handbook of methods* (p. 67). US Bureau of Labour Statistics.
Valliant, R.; Dorfman, A. H. e Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*.
Westat. (1996). *A User's Guide to WesVarPc, version 2.0*. Westat, Inc.
Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag.
Wolter, K. M. (2007). *Introduction to Variance Estimation* (Second, p. 428). Springer-Verlag.