Análise de Dados Amostrais

Djalma G. C. Pessoa, Pedro Luis do Nascimento Silva,
Antonio José Ribeiro Dias, Zélia Magalhães Bianchini e Sonia Albieri

19 de outubro de 2021, 14:46:11

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.

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 (2017). Essa razão fornece uma medida do efeito de ignorar o plano amostral. Os resultados das estimativas pontuais (Est_por_AAS e Est_Verd para as estimativas considerando AAS e o plano amostral verdadeiro, respectivamente), do desvio padrão da estimativa considerando o plano amostral verdadeiro (DP_Est_Verd) e do Efeito do Plano Amostral - EPA são apresentados na Tabela 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 (1996), 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 **??** 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 (1996).

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

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

Os Capítulos **??** e **??** tratam da análise de dados categóricos, dando ênfase à adaptação dos testes clássicos para proporções, de independência e de homogeneidade em tabelas de contingência, para lidar com dados provenientes de pesquisas amostrais complexas. Apresentamos correções das estatísticas clássicas e também a estatística de Wald baseada no plano amostral. As referências básicas usadas nesses capítulos foram o Capítulo 4 em Skinner et al. (1989) e o Capítulo 7 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 dez capítulos, todos escritos por autores convidados. Todos estes temas foram objeto de avanços importantes tanto no desenvolvimento de métodos como no de ferramentas computacionais para sua implementação no ambiente do sistema R, desde que foi publicado o livro inicial. 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 - Modelos multiníveis

Capítulo 15 - Modelos de teoria da resposta ao item

Capítulo 16 - Estimação de fluxos

Capítulo 17 - Modelos de séries temporais

Capítulo 18 - Modelos de redes neurais

Capítulo 19 - Modelos log-lineares para tabelas

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 e ajuste de modelos hierárquicos ou multiníveis considerando o plano amostral. Modelos hierárquicos têm sido bastante utilizados para explorar situações em que as relações entre variáveis de interesse em uma certa população de unidades elementares (por exemplo, crianças em escolas, pacientes em hospitais, empregados em empresas, moradores em regiões, etc.) são afetadas por efeitos de grupos determinados ao nível de unidades conglomeradas (os grupos). Ajustar e interpretar tais modelos é tarefa mais difícil que o mero ajuste de modelos lineares, mesmo em casos onde os dados são obtidos de forma exaustiva ou por AAS, e ainda mais complicada quando se trata de dados obtidos através de pesquisas com planos amostrais complexos. Diferentes abordagens estão disponíveis para ajuste de modelos hierárquicos nesse caso, e este capítulo apresenta uma revisão de tais abordagens, ilustrando com aplicações a dados de pesquisas amostrais de escolares.

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

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

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

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

As provas de vários dos resultados aqui discutidos se restringem a material disponível apenas em artigos em periódicos especializados estrangeiros e, portanto, são de acesso mais difícil. Ao leitor em busca de maior detalhamento e rigor, sugerimos consultar diretamente as inúmeras referências incluídas ao longo do texto. Para um tratamento mais profundo do assunto, os livros de Skinner et al. (1989) e Chambers 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 $\{f(y;\theta); \theta \in \Theta\}$. Na prática, as questões relevantes da pesquisa são traduzidas em termos de perguntas

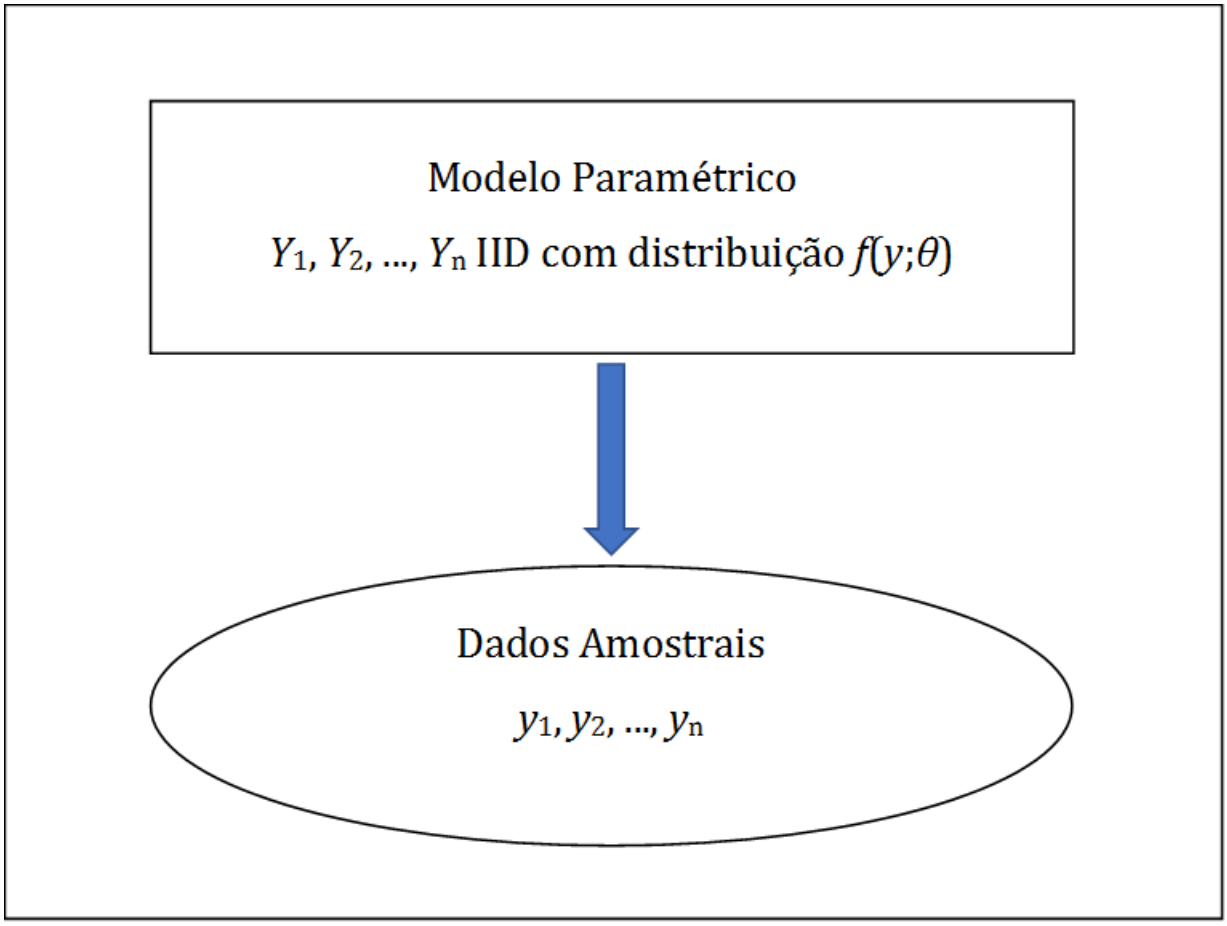

Figura 2.1: Representação esquemática da *Modelagem Clássica*

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 $y_{i_1}, \ldots, y_{i_n}$

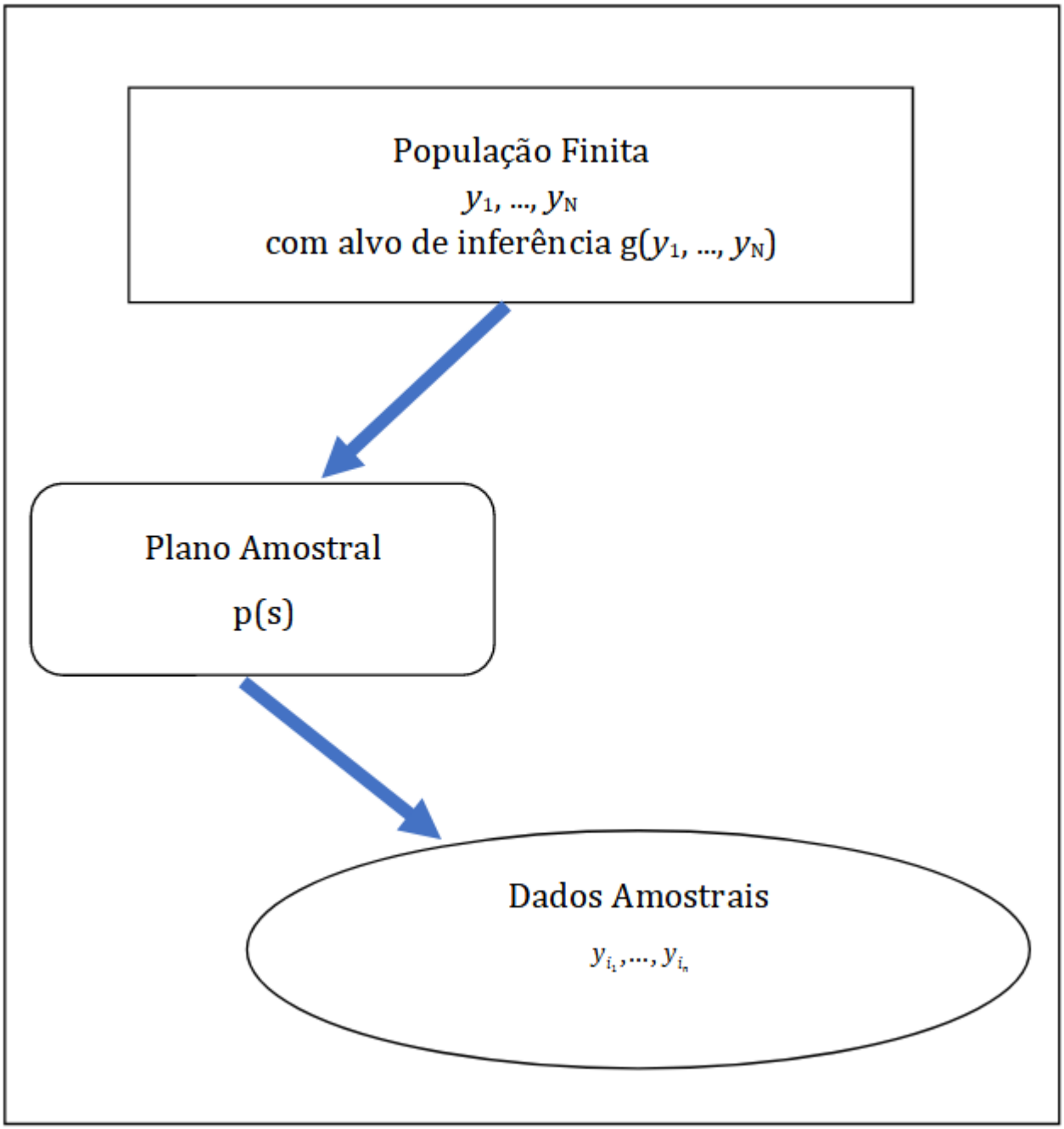

Figura 2.2: Representação esquemática da *Amostragem Probabilística*

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:

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

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

$$\widehat{T_y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} y_i,$$

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.

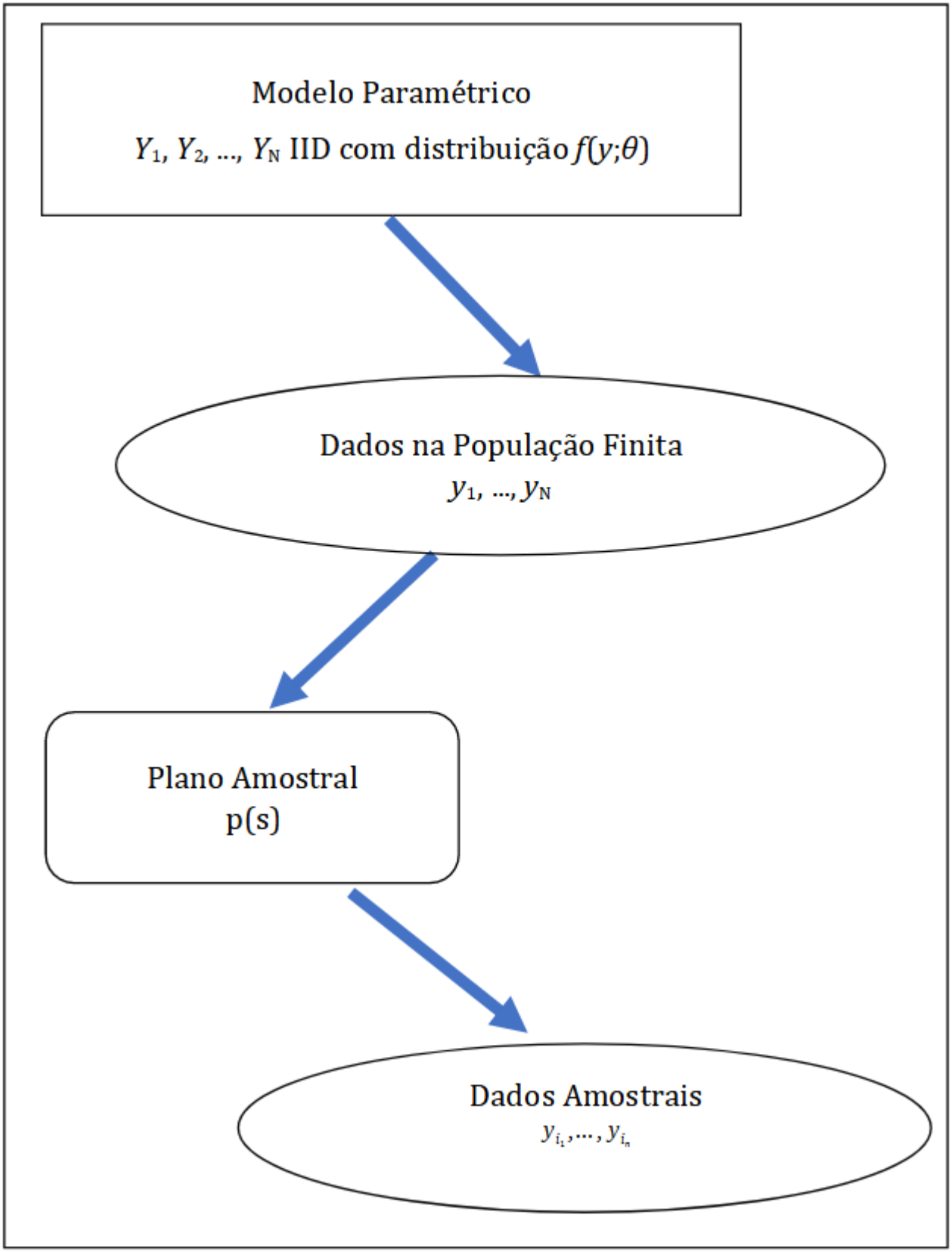

Figura 2.3: Representação esquemática da *Modelagem de Superpopulação*

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 serão 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 P variáveis de interesse da pesquisa, doravante denominadas simplesmente *variáveis da pesquisa*.

Denotemos por $\mathbf{y}_i = (y_{i1}, ..., y_{iP})'$ um vetor $P \times 1$ 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 $N \times P$ 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 serão 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:

$$f(\mathbf{y}_U, \mathbf{x}_U; \eta) = \prod_{i \in U} f(\mathbf{y}_i, \mathbf{x}_i; \eta) \tag{2.1}$$

$$= \prod_{i \in U} f(\mathbf{y}_i | \mathbf{x}_i; \lambda) f(\mathbf{x}_i; \phi) \tag{2.2}$$

onde λ e ϕ são vetores de parâmetros.

Sob (2.2), o modelo marginal correspondente das variáveis da pesquisa seria obtido integrando nas variáveis auxiliares:

$$f(\mathbf{y}_U;\theta) = f(\mathbf{y}_1,\ldots,\mathbf{y}_N;\theta) = \prod_{i\in U}\int f\left(\mathbf{y}_i|\mathbf{x}_i;\lambda\right) f\left(\mathbf{x}_i;\phi\right)\mathbf{dx}_i = \prod_{i\in U} f\left(\mathbf{y}_i;\theta\right) \tag{2.3}$$

onde $f\left(\mathbf{y}_i;\theta\right) = \int f\left(\mathbf{y}_i|\mathbf{x}_i;\lambda\right) f\left(\mathbf{x}_i;\phi\right)\mathbf{dx}_i$ e $\theta = h\left(\lambda,\phi\right)$ 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:

$$P\left[\left(\mathbf{Y}_U,\mathbf{X}_U\right) = \left(\mathbf{y}_U,\mathbf{x}_U\right)\right] = 1 \tag{2.4}$$

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

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 $\left(\mathbf{Y}_U,\mathbf{X}_U\right)$ não precisa ser degenerada como especificada em (2.4), 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 $\left(\mathbf{y}_1,\mathbf{x}_1\right),\ldots,\left(\mathbf{y}_N,\mathbf{x}_N\right)$. Sob a hipótese de resposta completa, a única fonte de incerteza seria devida ao fato de que $\left(\mathbf{y}_1,\mathbf{x}_1\right),\ldots,\left(\mathbf{y}_N,\mathbf{x}_N\right)$ é uma realização de $\left(\mathbf{Y}_1,\mathbf{X}_1\right),\ldots,\left(\mathbf{Y}_N,\mathbf{X}_N\right)$. 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 (2.4). 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 serão 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 $g\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)$ 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 serão 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,

$$\pi_i = P\left(i \in s\right) = \sum_{s \ni i} p(s) \tag{2.5}$$

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

$$\pi_{ij} = P\left(i \in s, j \in s\right) = \sum_{s \ni i,j} p(s) \tag{2.6}$$

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:

$$f\left[\delta_s|\left(\mathbf{y}_U, \mathbf{x}_U\right)\right] = f\left(\delta_s|\mathbf{x}_U\right) \tag{2.7}$$

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 (1996) 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:

$$\begin{aligned} f_U(x;y) &= P(X=x;Y=y) \\ &= \eta_{00}^{(1-x)(1-y)} \times \eta_{01}^{(1-x)y} \times \eta_{10}^{x(1-y)} \times (1-\eta_{00}-\eta_{01}-\eta_{10})^{xy} \end{aligned} \tag{2.8}$$

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:

$$f_U(y) = P(Y=y) = (\eta_{00}+\eta_{10})^{(1-y)} \times (1-\eta_{00}-\eta_{10})^{y} \tag{2.9}$$

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

$$f_U(x) = P(X=x) = (\eta_{00}+\eta_{01})^{(1-x)} \times (1-\eta_{00}-\eta_{01})^{x} \tag{2.10}$$

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

$$\begin{aligned} f_U(y|x) &= P(Y = y|X = x) \\ &= (1-x) \times \frac{\eta_{00}^{(1-y)}\eta_{01}^{y}}{\eta_{00}+\eta_{01}} + x \times \frac{\eta_{10}^{(1-y)}(1-\eta_{00}-\eta_{01}-\eta_{10})^{y}}{1-\eta_{00}-\eta_{01}} \end{aligned} \tag{2.11}$$

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:

$$f_s(x) = P(X_i = x|i \in s) = 1/2, \ \forall x \in \Omega \text{ e } \forall i \in U \tag{2.12}$$

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:

$$\begin{aligned} f_s(y) &= P(Y_i = y|i \in s) \\ &= \sum_{x=0}^{1} P(X_i = x; Y_i = y|i \in s) \\ &= \sum_{x=0}^{1} P[Y_i = y|(X_i = x)e(i \in s)] \times P(X_i = x|i \in s) \\ &= \sum_{x=0}^{1} P(Y_i = y|X_i = x) \times f_s(x) \\ &= \sum_{x=0}^{1} f_U(y|x) f_s(x) \\ &= \frac{1}{2} \times \left[\frac{\eta_{00}^{(1-y)}\eta_{01}^{y}}{\eta_{00}+\eta_{01}} + \frac{\eta_{10}^{(1-y)}(1-\eta_{00}-\eta_{01}-\eta_{10})^{y}}{1-\eta_{00}-\eta_{01}} \right] \end{aligned} \tag{2.13}$$

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 serão reproduzidos nesta seção. A referência básica usada foi a Seção 2.8 de Särndal et al. (1992).

Consideremos o problema de estimar o vetor $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$ de totais das P variáveis da pesquisa na população, a partir de uma amostra observada a. 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.

Um estimador usual baseado no plano amostral para o total $\mathbf{Y}$ é o estimador de Horvitz-Thompson, também chamado *estimador π -ponderado* (veja p.42 de Särndal et al. (1992)), dado por:

$$\hat{\mathbf{Y}}_\pi = \sum_{i \in a} \mathbf{y}_i / \pi_i. \tag{3.1}$$

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

O estimador π-ponderado $\hat{\mathbf{Y}}_\pi$ é não-viciado para o total $\mathbf{Y}$ com respeito à distribuição de aleatorização, isto é

$$E_p\left(\hat{\mathbf{Y}}_\pi\right) = \mathbf{Y}.$$

Além disto, sua variância de aleatorização é dada por

$$V_p\left(\hat{\mathbf{Y}}_\pi\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{\mathbf{y}_i}{\pi_i} \frac{\mathbf{y}_j'}{\pi_j} \ . \tag{3.2}$$

Uma expressão alternativa da variância de aleatorização de $\hat{\mathbf{Y}}_\pi$, válida quando o plano amostral é de tamanho fixo, é dada por

$$V_p\left(\hat{\mathbf{Y}}_\pi\right) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\pi_{ij}-\pi_i\pi_j\right)\left(\frac{\mathbf{y}_i}{\pi_i}-\frac{\mathbf{y}_j}{\pi_j}\right)\left(\frac{\mathbf{y}_i}{\pi_i}-\frac{\mathbf{y}_j}{\pi_j}\right)^{'}. \tag{3.3}$$

Note que na expressão (3.3) os termos onde $i = j$ não contribuem para a soma. Dois estimadores são usualmente recomendados para estimar a variância de aleatorização de $\hat{\mathbf{Y}}_\pi$. O primeiro é motivado pela expressão (3.2) e é dado por

$$\hat{V}_p\left(\hat{\mathbf{Y}}_\pi\right) = \sum_{i\in a}\sum_{j\in a}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\mathbf{y}_i}{\pi_i}\frac{\mathbf{y}_j^{'}}{\pi_j}. \tag{3.4}$$

O estimador de variância em (3.4) é um estimador não-viciado da variância de aleatorização de $\hat{\mathbf{Y}}_\pi$, isto é

$$E_p\left[\hat{V}_p\left(\hat{\mathbf{Y}}_\pi\right)\right] = V_p\left(\hat{\mathbf{Y}}_\pi\right) \tag{3.5}$$

desde que $\pi_{ij} > 0 \quad \forall i, j \in U$, como suposto neste livro na Seção 2.8.

O segundo estimador da variância, chamado estimador de Sen-Yates-Grundy, é motivado pela expressão (3.3) e é dado por

$$\hat{V}_{SYG}\left(\hat{\mathbf{Y}}_\pi\right) = -\frac{1}{2}\sum_{i\in a}\sum_{j\in a}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{\mathbf{y}_i}{\pi_i}-\frac{\mathbf{y}_j}{\pi_j}\right)\left(\frac{\mathbf{y}_i}{\pi_i}-\frac{\mathbf{y}_j}{\pi_j}\right)^{'}. \tag{3.6}$$

Observe que embora as expressões da variância (3.2) e (3.3) coincidam para planos amostrais de tamanho fixo, o mesmo não vale para os estimadores de variância (3.4) e (3.6), apesar de $\hat{V}_{SYG}\left(\hat{\mathbf{Y}}_\pi\right)$ ser também não-viciado para $V_p\left(\hat{\mathbf{Y}}_\pi\right)$ para planos amostrais de tamanho fixo.

Exemplo 3.1. Amostragem Aleatória Simples Sem Reposição (AAS)

Quando o plano 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 ficam iguais a

$$\pi_i = \frac{n}{N} \quad \forall \quad i \in U,$$

e

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \forall \quad i \neq j \in U\ .$$

Essas probabilidades de inclusão levam às seguintes expressões para o caso AAS:

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

$$V_{AAS}\left(\hat{\mathbf{Y}}_\pi\right) = N^2\frac{1-f}{n}\frac{N}{N-1}\mathbf{S}_y\ , \tag{3.8}$$

$$\hat{V}_p\left(\hat{\mathbf{Y}}_{AAS}\right) = \hat{V}_{SYG}\left(\hat{\mathbf{Y}}_{AAS}\right) = N^2\frac{1-f}{n}\frac{n}{n-1}\hat{\mathbf{S}}_y\ , \tag{3.9}$$

onde $f = n/N$ é a fração amostral e

$$\overline{\mathbf{y}} = n^{-1} \sum_{i \in a} \mathbf{y}_i \ , \tag{3.10}$$

$$\mathbf{S}_y = N^{-1} \sum_{i \in U} \left(\mathbf{y}_i - \overline{\mathbf{Y}}\right) \left(\mathbf{y}_i - \overline{\mathbf{Y}}\right)^{'} \ , \tag{3.11}$$

$$\overline{\mathbf{Y}} = N^{-1} \sum_{i \in U} \mathbf{y}_i = N^{-1} \mathbf{Y} \ , \tag{3.12}$$

$$\widehat{\mathbf{S}}_y = n^{-1} \sum_{i \in a} \left(\mathbf{y}_i - \overline{\mathbf{y}}\right) \left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^{'} \ . \tag{3.13}$$

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

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

onde w_i é um peso associado à unidade i da amostra ($i \in a$). O estimador **π-ponderado** ou de **Horvitz-Thompson** é um caso particular de $\hat{\mathbf{Y}}_w$ em (3.14) quando os pesos w_i são da forma

$$w_i^{HT} = \pi_i^{-1} \quad \forall \ \ i \in a.$$

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

$$\hat{\mathbf{Y}}_R = \left(\sum_{i \in a} \ \pi_i^{-1} \mathbf{y}_i\right) \times \left(\sum_{i \in U} x_i\right) / \left(\sum_{i \in a} \ \pi_i^{-1} x_i\right) \tag{3.15}$$

e

$$\hat{\mathbf{Y}}_{REG} = \sum_{i \in a} \pi_i^{-1} \mathbf{y}_i + \left(\sum_{i \in U} x_i - \sum_{i \in a} \pi_i^{-1} x_i\right) b_{xy} \tag{3.16}$$

onde x é uma variável auxiliar cujo total populacional $\sum_{i \in U} x_i = X$ é conhecido e b_{xy} é um estimador dos coeficientes da regressão linear entre as variáveis de pesquisa $\mathbf{y}$ e a variável auxiliar x.

Ambos os estimadores $\hat{\mathbf{Y}}_R$ e $\hat{\mathbf{Y}}_{REG}$ podem ser escritos na forma $\hat{\mathbf{Y}}_w = \sum\limits_{i \in a} w_i \mathbf{y}_i$ com pesos w_i dados respectivamente por

$$w_i^R = \frac{\pi_i^{-1} \sum_{k \in U} x_k}{\sum\limits_{k \in a} \pi_k^{-1} x_k} = \frac{\pi_i^{-1} X}{\widehat{X}_\pi} \tag{3.17}$$

e

$$w_i^{REG} = \pi_i^{-1} g_i \ , \tag{3.18}$$

onde $\widehat{X}_\pi = \sum_{i \in a} \pi_i^{-1} x_i$ é o estimador π -ponderado de X e $g_i = 1 + x_i \ (X - \widehat{X}_\pi) / \sum_{i \in a} \pi_i^{-1} x_i^2$.

O estimador de regressão descrito em (3.16) 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. Outra forma de generalizar o estimador de regressão é considerar estimadores alternativos dos coeficientes de regressão em lugar do estimador simples b_{xy} empregado aqui. Para uma discussão detalhada do estimador de regressão generalizado veja Silva (1996), Cap.3.

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 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 cap. 7. Apesar de sua importância para os praticantes de amostragem, a discussão detalhada desse problema não será 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 Por que Estimar Variâncias

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 em dados apenas 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 normal, tais que intervalos da forma

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

têm probabilidade de cobertura aproximada $1 - \alpha$.

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, BMDP e outros, não oferecem rotinas prontas para estimar variâncias considerando o plano amostral, nem mesmo para estatísticas simples como estimadores de totais e médias.

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 válidas.

Em muitos outros casos, como se verá no restante deste livro, os parâmetros de interesse são `não lineares` (diferentes de totais, médias e proporções, por exemplo). Casos comuns que consideraremos mais adiante são a estimação de razões, coeficientes 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 (3.5) 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 (1985). Discutimos inicialmente a técnica de `Linearização de Taylor`, em seguida uma abordagem comumente adotada para estimar variâncias para planos amostrais estratificados em vários estágios, com seleção de unidades primárias com probabilidades desiguais, denominada *Método do Conglomerado Primário* (do inglês *Ultimate Cluster*) e finalmente se discute brevemente uma técnica baseada na ideia de pseudo-replicações da amostra, denominada `Jackknife`. A combinação dessas três idéias suporta os desenvolvimentos teóricos dos algoritmos empregados pelos principais pacotes estatísticos especializados em estimação de variâncias de aleatorização (veja discussão no Capítulo **??**.

3.3 Linearização de Taylor para Estimar variâncias

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

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

onde $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i = (Y_1, \ldots, Y_R)'$ é um vetor de totais de R variáveis de pesquisa.

Consideremos estimadores π-ponderados de $\mathbf{Y}$, isto é, estimadores da forma:

$$\widehat{\mathbf{Y}}_\pi = \sum_{i \in s} \mathbf{y}_i / \pi_i \ .$$

Poderíamos usar $\hat{\theta}$ dado por

$$\hat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_\pi\right) = \mathbf{g}(\sum_{i \in s} \mathbf{y}_i / \pi_i) \ .$$

como estimador de θ. No caso particular em que $\mathbf{g}$ é uma função linear, é fácil estudar as propriedades de $\hat{\theta}$.

Assumindo então que θ é da forma

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

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

$$\hat{\theta} = \mathbf{A}\hat{\mathbf{Y}}_\pi \ .$$

Este estimador é não-viciado e tem variância de aleatorização

$$V_p\left(\hat{\theta}\right) = \mathbf{A}V_p\left(\hat{\mathbf{Y}}_\pi\right)\mathbf{A}',$$

onde $V_p\left(\hat{\mathbf{Y}}_\pi\right)$ é dado em (3.2) ou (3.3).

Quando $\mathbf{g}$ é não linear, podemos usar a técnica de Linearização de Taylor (ou Método Delta) para obter aproximações assintóticas para a variância de $\hat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_\pi\right)$. Para maiores detalhes sobre esse método, veja por exemplo p. 172 de Särndal et al. (1992), p. 221 de Wolter (1985) ou p. 486 de Bishop et al. (1975).

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

$$\hat{\theta} \simeq \hat{\theta}_L = \mathbf{g}(\mathbf{Y}) + \mathbf{\Delta g}(\mathbf{Y})\left(\hat{\mathbf{Y}}_\pi - \mathbf{Y}\right) \tag{3.19}$$

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

Tomando as variâncias de aleatorização dos dois lados em (3.19), e notando que no lado direito o único termo que tem variância de aleatorização $\mathbf{\Delta g}(\mathbf{Y})\left(\hat{\mathbf{Y}}_\pi - \mathbf{Y}\right)$ é uma função linear de $\hat{\mathbf{Y}}_\pi$, segue imediatamente que

$$V_p\left(\hat{\theta}\right) \simeq \mathbf{\Delta g}(\mathbf{Y})V_p\left(\hat{\mathbf{Y}}_\pi\right)\mathbf{\Delta g}(\mathbf{Y})' \tag{3.20}$$

onde $V_p\left(\hat{\mathbf{Y}}_\pi\right)$ é dado em (3.2). Um estimador consistente de $V_p\left(\hat{\theta}\right)$ é dado por

$$\hat{V}_p\left(\hat{\theta}\right) = \mathbf{\Delta g}(\hat{\mathbf{Y}}_\pi)\hat{V}_p\left(\hat{\mathbf{Y}}_\pi\right)\mathbf{\Delta g}(\hat{\mathbf{Y}}_\pi)', \tag{3.21}$$

onde $\hat{V}_p\left(\hat{\mathbf{Y}}_\pi\right)$ é dado em (3.4). Um outro estimador consistente seria obtido substituindo $\hat{V}_p\left(\hat{\mathbf{Y}}_\pi\right)$ por $\hat{V}_{SYG}\left(\hat{\mathbf{Y}}_\pi\right)$ dado em (3.6) na expressão (3.21).

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 as quantidades de interesse podem não ser expressas como funções de totais ou médias populacionais (este é o caso de quantis de distribuições, por exemplo).

Exemplo 3.2. Matriz de covariância para um vetor de razões

Para ilustrar a aplicação dos resultados anteriores, consideremos o problema de estimar a matriz de covariância de um vetor de razões. Sejam $\mathbf{Y} = (Y_1, \ldots Y_u)'$ e $\mathbf{X} = (X_1, \ldots, X_u)'$ vetores de totais e consideremos o vetor de razões $\mathbf{R} = \left(\frac{Y_1}{X_1}, \ldots, \frac{Y_u}{X_u}\right)'$. Conhecendo estimativas das matrizes $V_p\left(\hat{\mathbf{Y}}_\pi\right)$, $V_p\left(\hat{\mathbf{X}}_\pi\right)$ e $COV_p\left(\hat{\mathbf{Y}}_\pi; \hat{\mathbf{X}}_\pi\right)$, queremos calcular a matriz de variância de

$$\widehat{\mathbf{R}} = \left(\frac{\hat{Y}_{1\pi}}{\hat{X}_{1\pi}}, \ldots, \frac{\hat{Y}_{u\pi}}{\hat{X}_{u\pi}}\right)' .$$

Consideremos a função $\mathbf{g}:\mathbf{R}^{2u}\to\mathbf{R}^{u}$ dada por

$$\mathbf{g}\left(\mathbf{y},\mathbf{x}\right)=\left(\frac{y_1}{x_1},\ldots,\frac{y_u}{x_u}\right)$$

onde $\mathbf{y}=\left(y_1,\ldots,y_u\right)^{'}$ e $\mathbf{x}=\left(x_1,\ldots,x_u\right)^{'}$. A matriz jacobiana de $\mathbf{g}\left(\mathbf{y},\mathbf{x}\right)$ é a matriz $u\times 2u$ dada por

$$\boldsymbol{\Delta}\mathbf{g}\left(\mathbf{y},\mathbf{x}\right)=\left[\begin{array}{cc} diag\left(\frac{1}{x_1},\ldots,\frac{1}{x_u}\right) & diag\left(-\frac{y_1}{x_1^2},\ldots,-\frac{y_u}{x_u^2}\right)\end{array}\right].$$

Seja $\mathbf{D}_{\mathbf{x}}=diag(x_1,\ldots,x_u)$ a matriz diagonal de dimensão $u\times u$ formada a partir do vetor $\mathbf{x}=\left(x_1,\ldots,x_u\right)^{'}$. Usando essa notação, podemos escrever o vetor $\widehat{\mathbf{R}}$ de estimadores das razões como

$$\widehat{\mathbf{R}}=\left(\frac{\hat{Y}_{1\pi}}{\hat{X}_{1\pi}},\ldots,\frac{\hat{Y}_{u\pi}}{\hat{X}_{u\pi}}\right)^{'}=\mathbf{g}\left(\hat{\mathbf{Y}}_{\pi},\hat{\mathbf{X}}_{\pi}\right)$$

e a correspondente matriz jacobiana como

$$\boldsymbol{\Delta}\mathbf{g}\left(\hat{\mathbf{Y}}_{\pi},\hat{\mathbf{X}}_{\pi}\right)=\left[\begin{array}{cc}\mathbf{D}_{\widehat{\mathbf{R}}}\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1} & -\mathbf{D}_{\widehat{\mathbf{R}}}\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\end{array}\right].$$

A partir deste resultado, aplicando (3.21) podemos escrever:

$$\begin{aligned}\widehat{V}_p\left(\widehat{\mathbf{R}}\right) &\doteq \left[\begin{array}{cc}\mathbf{D}_{\widehat{\mathbf{R}}}\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1} & -\mathbf{D}_{\widehat{\mathbf{R}}}\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\end{array}\right]\\ &\times\left[\begin{array}{cc}\widehat{V}_p\left(\hat{\mathbf{Y}}_{\pi}\right) & \widehat{COV}_p\left(\hat{\mathbf{Y}}_{\pi},\hat{\mathbf{X}}_{\pi}\right)\\ \widehat{COV}_p\left(\hat{\mathbf{X}}_{\pi},\hat{\mathbf{Y}}_{\pi}\right) & \widehat{V}_p\left(\hat{\mathbf{X}}_{\pi}\right)\end{array}\right]\\ &\times\left[\begin{array}{c}\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1}\mathbf{D}_{\widehat{\mathbf{R}}}\\ -\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\mathbf{D}_{\widehat{\mathbf{R}}}\end{array}\right].\end{aligned}$$

Efetuando os produtos das matrizes em blocos obtemos

$$\begin{aligned}\widehat{V}_p\left(\widehat{\mathbf{R}}\right) &= \mathbf{D}_{\widehat{\mathbf{R}}}\left[\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1}\widehat{V}_p\left(\hat{\mathbf{Y}}_{\pi}\right)\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1}+\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\widehat{V}_p\left(\hat{\mathbf{X}}_{\pi}\right)\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\right]\mathbf{D}_{\widehat{\mathbf{R}}}\\ &\quad-\mathbf{D}_{\widehat{\mathbf{R}}}\left[\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1}\widehat{COV}_p\left(\hat{\mathbf{Y}}_{\pi},\hat{\mathbf{X}}_{\pi}\right)\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\right.\\ &\quad\left.+\mathbf{D}_{\hat{\mathbf{X}}_{\pi}}^{-1}\widehat{COV}_p\left(\hat{\mathbf{X}}_{\pi},\hat{\mathbf{Y}}_{\pi}\right)\mathbf{D}_{\hat{\mathbf{Y}}_{\pi}}^{-1}\right]\mathbf{D}_{\widehat{\mathbf{R}}},\end{aligned}\tag{3.22}$$

que fornece o resultado desejado, isto é, uma expressão de estimador para a matriz de variância do estimador $\widehat{\mathbf{R}}$ do vetor de razões de interesse.

3.4 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. Esta ideia é simples, porém bastante poderosa, porque permite acomodar uma enorme variedade de planos amostrais, envolvendo estratificação e seleção com probabilidades desiguais (com ou sem reposição) tanto das unidades primárias 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 vários dos pacotes estatísticos especializados em cálculo de variâncias considerando o plano amostral, tais como SUDAAN, CENVAR, STATA ou PC-CARP (veja discussão no Capítulo 10).

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

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

e um estimador não viciado da variância de aleatorização correspondente por

$$\widehat{V}_p\left(\widehat{Y}_{CP}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\widehat{Y}_{hi}}{\pi_{hi}} - \frac{\widehat{Y}_h}{n_h}\right)^2 \tag{3.23}$$

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

Embora muitas vezes a seleção das unidades primárias possa ter sido feita sem reposição, o estimador de Conglomerados Primários aqui apresentado pode fornecer uma aproximação razoável da correspondente variância de aleatorização. 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 em vários estágios. Uma discussão sobre a qualidade dessa aproximação e alternativas pode ser encontrada em Särndal et al. (1992), p. 153.

3.5 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 (veja Westat (1996)).

A ideia básica é 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 `replicações`. Nesse caso, se θ é o parâmetro-alvo, e $\widehat{\theta}_g$ é um estimador não viciado de θ baseado na g-ésima replicação ($g = 1, \ldots, G$), segue-se que

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

é um estimador não viciado de θ e

$$\widehat{V}_R\left(\widehat{\theta}_R\right) = \frac{1}{G\left(G-1\right)}\sum_{g=1}^{G}\left(\widehat{\theta}_g - \widehat{\theta}_R\right)^2 \tag{3.24}$$

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

Note que desde que as replicações sejam construídas de forma independente conforme indicado, os estimadores $\widehat{\theta}_R$ e $\widehat{V}_R\left(\widehat{\theta}_R\right)$ são não viciados qualquer que seja o plano amostral empregado para selecionar a amostra de cada replicação, 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 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 independentes com o mesmo esquema, se comparado à seleção de uma única amostra de tamanho n diretamente. Além disto, se o número de replicações G for pequeno, o estimador de variância pode ser instável. Uma pesquisa importante e de grande porte em que esta ideia é aplicada exatamente é a pesquisa de preços para formar o índice de Preços ao Consumidor (do inglês *Consumer Price Index - CPI* do Labor Statistics (1984), p. 22, que utiliza duas replicações (meias amostras) para formar a amostra pesquisada.

Mesmo quando a amostra não foi selecionada exatamente dessa forma, a construção de replicações 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 replicações 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 `replicações 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 replicações independentes.

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 deve ser feita respeitando estratos e alocando unidades primárias 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 replicações 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 $\widehat{V}_R\left(\widehat{\theta}_R\right)$ anteriormente apresentado, dado por

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

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 (veja IBGE (1985)).

Nesta seção descreveremos uma outra dessas técnicas baseadas em replicações, talvez a mais conhecida e popular, 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

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

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

$$\widehat{V}_{J1}\left(\widehat{\theta}\right) = \frac{1}{G\,(G-1)} \sum_{g=1}^{G} \left(\widehat{\theta}_{(g)} - \widehat{\theta}_J\right)^2 \tag{3.26}$$

ou

$$\widehat{V}_{J2}\left(\widehat{\theta}\right) = \frac{1}{G\,(G-1)} \sum_{g=1}^{G} \left(\widehat{\theta}_{(g)} - \widehat{\theta}\right)^2 \tag{3.27}$$

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 $\hat{\theta}$.

Observação . 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.

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

Observação . 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 unidades primárias na amostra (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 resultado idêntico 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 (veja, 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*).

3.6 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.

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

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

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

Podemos calcular diretamente o desvio-padrão:

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 (2017), 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.

Para o tipo *Bootstrap*, temos:

Vamos apresentar mais detalhes sobre a obtenção dos estimadores de *Jackknife* e *Bootstrap* na library *survey*, Lumley (2017). A classe do objeto `ppv_se_plan_jkn` é `svyrep.design` e ele contém as seguintes componentes:

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.

O argumento `compress` da função `as.svrepdesign` permite especificar se, na saída da função, a matriz `weights` será 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:

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

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

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

A library `survey` Lumley (2017) fornece uma função para estimar a variância de uma função de totais a partir das réplicas de pesos:

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:

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:

3.7

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. (Eds.). (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.

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: set. 2020.)

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

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

Labor Statistics, U. B. of. (1984). *BLS Handbook of Methods - Volume II - The Consumer Price Index* [BLS Bulletin 2134-2].

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. (2017). *survey: Analysis of Complex Survey Samples*. Disponível em: https://CRAN.R-project.org/package=survey (Acesso: set. 2020.)

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%0Ahttps://cetic.br/pt/publicacoes/indice/pesquisas/%0Ahttps://cetic.br/pt/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nos-domicilios (Acesso: set. 2020.)

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.

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* [Phdthesis]. 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: set. 2020.)

Skinner, C. J.; Holt, D. e Smith, T. M. F. (Eds.). (1989). *Analysis of Complex Surveys*. John Wiley & Sons.

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.

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.