Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-11-30

Sumário

Prefácio

Uma preocupação básica de toda instituição produtora de informações estatísticas é com a utilização ``'correta'' de seus dados. Isso pode ser intrepretado de várias formas, algumas delas com reflexos até na confiança do público e na própria sobrevivência do órgão. Do nosso ponto de vista, como técnicos da área de metodologia do IBGE, 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 será 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 pacotes 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 serão 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 permitirá 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 problema 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 três últimas décadas tem se tornado 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 feitos 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 incorporado aos principais pacotes estatísticos disponíveis (tais como MINITAB, R, SAS, SPSS, etc).

As ferramentas de análise convencionais disponíveis nesses pacotes estatísticos geralmente partem de hipóteses básicas que só são válidas quando os dados foram obtidos através de amostras aleatórias simples com reposição (AASC). Tais hipóteses são geralmente inadequadas para modelar observações provenientes de amostras de populações finitas, pois desconsideram os seguintes aspectos relevantes dos planos amostrais usualmente empregados nas pesquisas amostrais:

i.) **probabilidades distintas de seleção das unidades**;

ii.) **conglomeração das unidades**;

iii.) **estratificação**;

iv.) **calibração ou imputação para não-resposta e outros ajustes**.

As estimativas pontuais de parâmetros descritivos da população ou 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. Ao ignorar estes aspectos, os pacotes tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

O exemplo a seguir considera o uso de dados de uma pesquisa amostral real conduzida pelo IBGE para ilustrar como os pontos i) a iv) acima mencionados afetam a inferência sobre quantidades descritivas populacionais tais como totais, médias, proporções e razões.
Exemplo 1.1. Distribuição dos pesos da amostra da PPV

Tabela 1.1: Número de setores na população e na amostra, por estrato geográfico

Estrato_Geográfico	População	Amostra
Região Metropolitana de Fortaleza	2.263	62
Região Metropolitana de Recife	2.309	61
Região Metropolitana de Salvador	2.186	61
Restante Nordeste Urbano	15.057	61
Restante Nordeste Rural	23.711	33
Região Metropolitana de Belo Horizonte	3.283	62
Região Metropolitana do Rio de Janeiro	10.420	61
Região Metropolitana de São Paulo	14.931	61
Restante Sudeste Urbano	25.855	61
Restante Sudeste Rural	12.001	31
Total	112.016	554

Os dados deste exemplo são relativos à distribuição dos pesos na amostra da Pesquisa sobre Padrões de Vida (PPV), realizada pelo IBGE nos anos 1996-97. (**?**) descrevem resumidamente a PPV, que foi realizada nas Regiões Nordeste e Sudeste do País.

O plano amostral empregado na seleção da amostra da PPV foi estratificado e conglomerado em dois estágios, com alocação desproporcional da amostra nos estratos geográficos. A estratificação considerou inicialmente 10 estratos geográficos conforme listados na Tabela 1.1.

As unidades primárias de amostragem (UPAs) foram os setores censitários da base geográfica do IBGE conforme usada para o Censo Demográfico de 1991. A seleção dos setores dentro de cada estrato foi feita com probabilidade proporcional ao tamanho. Os domicílios foram as unidades de segundo estágio, selecionados por amostragem aleatória simples sem reposição em cada setor selecionado, após a atualização do cadastro de domicílios do setor.

Em cada um dos 10 estratos geográficos, os setores foram subdivididos em três estratos de acordo com a renda média mensal do chefe do domicílio por setor, perfazendo um total de 30 estratos finais para seleção da amostra.

O tamanho da amostra para cada estrato geográfico foi fixado em 480 domicílios, e o número de setores selecionados foi fixado em 60, com 8 domicílios sendo selecionados em cada setor. A exceção ficou por conta dos estratos que correspondiam ao restante da área rural de cada Região, onde foram selecionados 30 setores e com 16 domicílios selecionados por setor, em função da maior dificuldade de acesso a esses setores, o que implicaria em aumento de custo da coleta caso fosse mantido o mesmo tamanho da amostra do segundo estágio em cada setor.

A alocação da amostra dentro de cada estrato geográfico foi proporcional ao número de domicílios particulares permanentes ocupados do estrato de renda no Censo de 1991. No final foram incluídos 554 setores na amostra, distribuídos tal como mostrado na Tabela 1.1.

A Tabela 1.2 apresenta um resumo das distribuições dos pesos amostrais para as Regiões Nordeste (5 estratos geográficos) e Sudeste (5 estratos geográficos) separadamente, e para o conjunto da amostra da PPV.

```
## -- Attaching packages ------------------

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.3.4     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0

## -- Conflicts -- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

Tabela 1.2: Resumos da distribuição dos pesos da amostra da PPV

Região	Mínimo	Quartil 1	Mediana	Quartil 3	Máximo
Nordeste	724	1.194	1.556	6.937	15.348
Sudeste	991	2.789	5.429	9.509	29.234
Nordeste+Sudeste	724	1.403	3.785	8.306	29.234

```
## x dplyr::lag()    masks stats::lag()
```

No cálculo dos pesos amostrais foram consideradas as probabilidades de inclusão dos elementos na amostra, bem como correções devido à não-resposta. Contudo, a grande variabilidade dos pesos amostrais da PPV é devida, principalmente, à variabilidade das probabilidades de inclusão na amostra, ilustrando desta forma o ponto i) citado anteriormente nesta seção.

Na análise de dados desta pesquisa, deve-se considerar que há elementos da amostra com pesos muito distintos. Por exemplo, a razão entre o maior e o menor peso é cerca de 40 vezes. Os pesos também variam bastante entre as regiões, com mediana 3,5 vezes maior na região Sudeste quando comparada com a região Nordeste, em função da alocação desproporcional da amostra nas regiões.

Tais pesos são utilizados para `expandir` os dados, multiplicando-se cada observação pelo seu respectivo peso. Assim, por exemplo, para `estimar` quantos elementos `da população` pertencem a determinado conjunto (domínio), basta somar os pesos dos elementos da amostra que pertencem a este conjunto. É possível ainda incorporar os pesos, de maneira simples e natural, quando estimamos medidas descritivas simples da população tais como totais, médias, proporções, razões, etc.

Por outro lado, quando utilizamos a amostra para estudos analíticos, as opções padrão disponíveis nos pacotes estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações independentes e identicamente distribuídas (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 gera inferências incorretas sobre a precisão das estimativas. Isto ocorre pois 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. Em tais pesquisas, a opção mais freqüente 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, médias, proporções e razões, calculamos estimativas de quantidades desses diferentes tipos usando a amostra da PPV juntamente com estimativas das respectivas variâncias. Tais estimativas de variância foram calculadas sob duas estratégias: a) considerando amostragem aleatória simples (portanto ignorando o plano amostral efetivamente adotado na pesquisa), e b) considerando o plano amostral da pesquisa e os pesos diferenciados das unidades.

A razão entre as estimativas de variância obtidas sob o plano amostral verdadeiro e sob amostragem aleatória simples foi calculada para cada uma das estimativas consideradas usando o pacote `survey` do R (**?**). Essa razão fornece uma medida do efeito de ignorar o plano amostral. Os resultados das estimativas ponderadas e variâncias considerando o plano amostral são apresentados na Tabela 1.3, juntamente com as medidas dos efeitos de plano amostral (EPA).

Exemplos de utilização do pacote `survey` para obtenção de estimativas apresentadas na 1.3 estão na Seção 1.4. As outras estimativas da Tabela 1.3 podem ser obtidas de maneira análoga.

Tabela 1.3: Estimativas de Efeitos de Plano Amostral (EPAs) para variáveis selecionadas da PPV - Região Sudeste

Parâmetro	Estimativa	Erro.Padrão	EPA
1.	3,62	0,05	2,64
2.	10,70	1,15	2,97
3.	1.208.123,00	146.681,00	3,37
4.	1.174.220,00	127.982,00	2,64
5.	4.792.344,00	318.877,00	4,17
6.	11,87	1,18	2,46
7.	10,87	0,67	3,86
8.	10.817.590,00	322.947,00	2,02
9.	10.804.511,00	323.182,00	3,02
10.	709.145,00	87.363,00	2,03
11.	1,39	0,03	1,26
12.	0,53	0,01	1,99

Na Tabela 1.3 apresentamos as estimativas dos seguintes parâmetros populacionais:

1. Número médio de pessoas por domicílio;
2. % de domicílios alugados;
3. Número total de pessoas que avaliaram seu estado de de saúde como ruim;
4. Total de analfabetos de 7 a 14 anos;
5. Total de analfabetos de mais de 14 anos;
6. % de analfabetos de 7 a 14 anos;
7. % de analfabetos de mais de 14 anos;
8. Total de mulheres de 12 a 49 anos que tiveram filhos;
9. Total de mulheres de 12 a 49 anos que tiveram filhos vivos;
10. Total de mulheres de 12 a 49 anos que tiveram filhos mortos;
11. Número médio de filhos tidos por mulheres de 12 a 49 anos;
12. Razão de dependência.

Como se pode observar da quarta coluna da Tabela 1.3, os valores do efeito do plano amostral variam de um modesto 1,26 para o número médio de filhos tidos por mulheres em idade fértil (12 a 49 anos de idade) até um substancial 4,17 para o total de analfabetos entre pessoas de mais de 14 anos. Nesse último caso, usar a estimativa de variância como se o plano amostral fosse amostragem aleatória simples implicaria em subestimar consideravelmente a variância da estimativa pontual, que é mais que 4 vezes maior se consideramos o plano amostral efetivamente utilizado.

Note que as variáveis e parâmetros cujas estimativas são 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, razões) 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, à exceção dos dois menores valores (1,26 e 1,99), todas as demais estimativas de EPA apresentaram valores superiores a 2.

1.2 Objetivos do Livro

Este livro tem três objetivos principais:

1) **Ilustrar e analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise**

de dados quando estes são provenientes de pesquisas amostrais complexas;

2) **Apresentar uma coleção de métodos e recursos computacionais disponíveis para análise de dados amostrais complexos, equipando o analista para trabalhar com tais dados, reduzindo assim o risco de inferências incorretas**;

3) **Ilustrar o potencial analítico de muitas das pesquisas produzidas por agências de estatísticas oficiais 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 complexas, usando os recursos do pacote 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 (**?**), (**?**), (**?**), (**?**), ou Schafer (1997), por exemplo.

1.3 Estrutura do Livro

O livro está organizado em catorze capítulos. Este primeiro capítulo discute a motivação para estudar o assunto e apresenta uma ideia geral dos objetivos e da estrutura do livro.

No segundo capítulo, procuramos dar uma visão das diferentes abordagens utilizadas na análise estatística de dados de pesquisas amostrais complexas. 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 (**?**), o capítulo 1 em (**?**) e os capítulos 1 e 2 em (**?**).

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 **??** 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 (**?**), (**?**) e (**?**).

No Capítulo 1.4 introduzimos o conceito de `Efeito do Plano Amostral (EPA)`, que permite avaliar o impacto de ignorar a estrutura dos dados populacionais ou do plano amostral sobre a estimativa da variância de um estimador. Para isso, comparamos o estimador da variância apropriado para dados obtidos por amostragem aleatória simples (hipótese de AAS) com o valor esperado deste mesmo estimador sob a distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado (plano amostral verdadeiro). Aqui a referência principal foi o livro (**?**), complementado com o texto de (**?**).

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 (**?**), (**?**), (**?**) e o capítulo 6 em (**?**).

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 e de regressão logística, quando os dados vêm de pesquisas amostrais complexas. Apresentamos um exemplo de aplicação com dados do Suplemento Trabalho da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1990, onde ajustamos um modelo de regressão logística. Neste exemplo, são feitas comparações entre resultados de ajustes obtidos através de um programa especializado, o pacote `survey` (?), e através de um programa de uso geral, a função `glm` do R. As referências centrais são o capítulo 6 em (?) e Binder(1983), além de (?).

Os Capítulos ?? e ?? tratam da análise de dados categóricos com ênfase na adaptação dos testes clássicos para proporções, de independência e de homogeneidade em tabelas de contingência, para 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 os o capítulo 4 em (?) e o capítulo 7 (?). Também são apresentadas as ideias básicas de como efetuar ajuste de modelos log-lineares a dados de frequências em tabelas de múltiplas entradas.

O Capítulo ?? trata da estimação de densidades e funções de distribuição, ferramentas que tem assumido importância cada dia maior com a maior disponibilidade de microdados de pesquisas amostrais para analistas fora das agências produtoras.

O Capítulo ?? trata da estimação e ajuste de modelos hierárquicos considerando o plano amostral. Modelos hierárquicos (ou modelos multiníveis) 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, mas ainda mais complicada quando se trata de dados obtidos através de pesquisas amostrais complexas. Várias alternativas de métodos para ajuste de modelos hierárquicos estão disponíveis, e este capítulo apresenta uma revisão de tais abordagens, ilustrando com aplicações a dados de pesquisas amostrais de escolares.

O Capítulo ?? trata da não resposta e suas conseqüências sobre a análise de dados. As abordagens de tratamento usuais, reponderação e imputação, são descritas de maneira resumida, com apresentação de alguns exemplos ilustrativos, e referências à ampla literatura existente sobre o assunto. Em seguida destacamos a importância de considerar os efeitos da não-resposta e dos tratamentos compensatórios aplicados nas análises dos dados resultantes, destacando em particular as ferramentas disponíveis para a estimação de variâncias na presença de dados incompletos tratados mediante reponderação e/ou imputação.

O Capítulo ?? trata de assunto ainda emergente: diagnósticos do ajuste de modelos quando os dados foram obtidos de amostras complexas. A literatura sobre o assunto ainda é incipiente, mas o assunto é importante e procura-se estimular sua investigação com a revisão do estado da arte no assunto.

O Capítulo ?? discute algumas formas alternativas de analisar dados de pesquisas complexas, contrapondo algumas abordagens distintas à que demos preferência nos capítulos anteriores, para dar aos leitores condições de apreciar de forma crítica o material apresentado no restante deste livro. Entre as abordagens discutidas, há duas principais: a denominada `análise desagregada`, e a abordagem denominada `obtenção do modelo amostral` proposta por (?). A chamada `análise desagregada` incorpora explicitamente na análise vários aspectos do plano amostral utilizado, através do emprego de modelos hierárquicos (?). Em contraste, a abordagem adotada nos oito primeiros capítulos é denominada `análise agregada`, e procura `eliminar` da análise efeitos tais como conglomeração induzida pelo plano amostral, considerando tais efeitos como `ruídos` ou fatores de perturbação que `atrapalham` o emprego dos procedimentos clássicos de estimação, ajuste de modelos e teste de hipóteses.

A abordagem de `obtenção do modelo amostral` parte de um modelo de superpopulação e procura derivar o modelo amostral (ou que valeria para as observações da amostra obtida), considerando modelos para as probabilidades de inclusão dadas as variáveis auxiliares e as variáveis resposta de interesse. Uma vez obtidos tais modelos, seu ajuste prossegue por métodos convencionais tais como máxima verossimilhança ou mesmo MCMC (Markov Chain Monte Carlo).

Por último, no Capítulo **??**, listamos alguns pacotes computacionais especializados disponíveis para a análise de dados de pesquisas amostrais complexas. Sem pretender ser exaustiva ou detalhada, essa revisão dos pacotes procura também apresentar suas características mais importantes. Alguns destes programas podem ser adquiridos gratuitamente via `internet`, nos endereços fornecidos de seus produtores. Com isto, pretendemos indicar aos leitores o caminho mais curto para permitir a implementação prática das técnicas e métodos aqui discutidos.

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. Embora a experiência de fazer inferência analítica com dados desse tipo seja ainda incipiente no Brasil, acreditamos ser fundamental difundir 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 (**?**) e (**?**) são as referências centrais a consultar. Para aqueles querendo um tratamento ainda mais prático que o nosso, os livro de (**?**) e (**?**) podem ser opções interessantes.

1.4 Laboratório de R do Capítulo 1.

Exemplo 1.2. Utilização do pacote survey do R para estimar alguns totais e razões na Tabela 1.3

Os exemplos a seguir utilizam dados da Pesquisa de Padrões de Vida (`PPV`) do IBGE, cujo plano amostral encontra-se descrito no Exemplo 1.1. Os dados da `PPV` que usamos aqui estão disponíveis no arquivo (data frame) `ppv` do pacote `anamco`.

```
# Leitura dos dados
library(anamco)
ppv_dat <- ppv
# Características dos dados da PPV
dim(ppv_dat)
```

```
## [1] 19409    13
```

```
names(ppv_dat)
```

```
##  [1] "serie"    "ident"    "codmor"   "v04a01"   "v04a02"   "v04a03"
##  [7] "estratof" "peso1"    "peso2"    "pesof"    "nsetor"   "regiao"
## [13] "v02a08"
```

Inicialmente, adicionamos quatro variáveis de interesse por meio de transformação das variáveis existentes no data frame `ppv_dat`, a saber:

- analf1 - indicador de analfabeto na faixa etária de 7 a 14 anos;
- analf2 - indicador de analfabeto na faixa etária acima de 14 anos;
- faixa1 - indicador de idade entre 7 e 14 anos;
- faixa2 - indicador de idade acima de 14 anos;

```
# Adiciona variáveis ao arquivo da ppv_dat
ppv_dat <- transform(ppv_dat,
analf1 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >= 7 & v02a08 <= 14)) * 1,
analf2 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >14)) * 1,
faixa1 = (v02a08 >= 7 & v02a08 <= 14) *1,
faixa2 = (v02a08 > 14) * 1)
#str(ppv_dat)
```

A seguir, mostramos como utilizar o pacote `survey` (**?**) do R para obter algumas estimativas da Tabela 1.3. Os dados da pesquisa estão contidos no data frame `ppv_dat`, que contém as variáveis que caracterizam o plano amostral:

- estratof - identifica os estratos de seleção;
- nsetor - identifica as unidades primárias de amostragem ou conglomerados;
- pesof - identifica os pesos do plano amostral.

O passo fundamental para utilização do pacote `survey` (**?**) é criar um objeto que guarde as informações relevantes sobre a estrutura do plano amostral junto dos dados. Isso é feito por meio da função `svydesign()`. As variáveis que definem estratos, conglomerados e pesos na `PPV` são `estratof`, `nsetor` e `pesof` respectivamente. O objeto de desenho amostral, `ppv_plan` incorpora as informações da estrutura do plano amostral adotado na `PPV`.

```
# Carrega o pacote survey
library(survey)
# Cria objeto de desenho
ppv_plan <- svydesign(ids = ~nsetor, strata = ~estratof, data = ppv_dat,
                      nest = TRUE, weights = ~pesof)
```

Como todos os exemplos a seguir serão relativos a estimativas para a Região Sudeste, vamos criar um objeto de desenho restrito a essa região.-

```
ppv_se_plan <- subset(ppv_plan, regiao == "Sudeste")
```

Para exemplificar as análises descritivas de interesse, vamos estimar algumas características da população, descritas na Tabela 1.3. Os totais das variáveis `analf1` e `analf2` para a região Sudeste fornecem os resultados mostrados nas linhas 4 e 5 da Tabela 1.3:

- total de analfabetos nas faixas etárias de 7 a 14 anos (`analf1`) e
- total de analfabetos acima de 14 anos (`analf2`).

```
svytotal(~analf1, ppv_se_plan, deff = TRUE)
```

```
##          total      SE   DEff
## analf1 1174220  127982 2,0543
```

```
svytotal(~analf2, ppv_se_plan, deff = TRUE)
```

```
##          total      SE   DEff
## analf2 4792344  318877 3,3237
```

- percentual de analfabetos nas faixas etárias consideradas, que fornece os resultados nas linhas 6 e 7 da Tabela 1.3:

```
svyratio(~analf1, ~faixa1, ppv_se_plan)
```

```
## Ratio estimator: svyratio.survey.design2(~analf1, ~faixa1, ppv_se_plan)
## Ratios=
##          faixa1
## analf1 0,118689
```

```
## SEs=
##            faixa1
## analf1 0,01178896
```

```
svyratio(~analf2, ~faixa2, ppv_se_plan)
```

```
## Ratio estimator: svyratio.survey.design2(~analf2, ~faixa2, ppv_se_plan)
## Ratios=
##           faixa2
## analf2 0,1086871
## SEs=
##             faixa2
## analf2 0,006732254
```

Uma alternativa para obter estimativa por domínios é utilizar a função `svyby()` do pacote `survey` (**?**). Assim, poderíamos estimar os totais da variável `analf1` para as regiões `Nordeste (1)` e `Sudeste (2)` da seguinte forma:

```
svyby(~analf1, ~regiao, ppv_plan, svytotal, deff = TRUE)
```

```
##            regiao  analf1       se DEff.analf1
## Nordeste Nordeste 3512866 352619,5    9,660561
## Sudeste   Sudeste 1174220 127982,2    2,054345
```

Observe que as estimativas de totais e desvios padrão obtidas coincidem com as Tabela 1.3, porém as estimativas de Efeitos de Plano Amostral(EPA) são distintas. Uma explicação detalhada para essa diferença será apresentada no capítulo 4, após a discussão do conceito de efeito do plano amostral e de métodos para sua estimação.

1.5 Laboratório de R do Capítulo 1 - Extra.

Exemplo 1.3. Exemplo anterior usando o pacote srvyr

- Carrega o pacote `srvyr`:

```
library(srvyr)
```

- Cria objeto de desenho:

```
ppv_plan <- ppv_dat %>%
            as_survey_design(strata = estratof, ids = nsetor, nest = TRUE,
                             weights = pesof)
```

Vamos criar novamente as variáveis derivadas necessárias, mas observe que, desta vez, estas variáveis estão sendo adicionadas ao objeto que já contem os dados e as informações da estrutura do plano amostral.

```
ppv_plan <- ppv_plan %>%
            mutate(
analf1 = as.numeric((v04a01 == 2 | v04a02 == 2) & (v02a08 >= 7 & v02a08 <= 14)),
analf2 = as.numeric((v04a01 == 2 | v04a02 == 2) & (v02a08 >14)),
faixa1 = as.numeric(v02a08 >= 7 & v02a08 <= 14),
faixa2 = as.numeric(v02a08 > 14)
)
```

- Estimar a taxa de analfabetos por região para as faixas etárias de 7-14 anos e mais de 14 anos.

```
result1 <- ppv_plan %>%
           group_by(regiao) %>%
```

Tabela 1.4: Proporção de analfabetos para faixas etárias 7-14 anos e mais de 14 anos

regiao	taxa_analf1	taxa_analf1_se	taxa_analf2	taxa_analf2_se
NA	42,3	3,1	33,6	1,6
NA	11,9	1,2	10,9	0,7

```
            summarise(
            taxa_analf1 = 100*survey_ratio(analf1, faixa1),
            taxa_analf2 = 100*survey_ratio(analf2, faixa2)
            )
```

```
## Warning in Ops.factor(left, right): '*' not meaningful for factors

## Warning in Ops.factor(left, right): '*' not meaningful for factors
```

```
knitr::kable(as.data.frame(result1), booktabs = TRUE, row.names = FALSE, digits = 1,
             align = "crrrr", format.args= list(decimal.mark=","),
caption = "Proporção de analfabetos para faixas etárias 7-14 anos e mais de 14 anos")
```

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, numa situação simples, algumas abordagens alternativas para modelagem e análise estatística.

2.1.1 Abordagem 1 - Modelagem Clássica

Seja Y um vetor $P \times 1$ de variáveis de pesquisa (ou de interesse), e sejam n vetores de observações destas variáveis para uma amostra de unidades de interesse denotados 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 vetores de variáveis aleatórias $Y_1, \ldots, Y_n$.

Podemos formular modelos bastante sofisticados para a distribuição conjunta destes vetores aleatórios, mas para simplificar a discussão, vamos inicialmente supor que $Y_1, \ldots, Y_n$ são vetores aleatórios independentes e identicamente distribuídos (IID), com a mesma distribuição de Y, 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 1.

Dados Amostrais	$Y_1 = y_1, \ldots, Y_n = y_n$
Modelo Paramétrico/	$Y_1, \ldots, Y_n$ variáveis aleatórias IID com distribuição $f(y, \theta)$, onde
Hipóteses	$\theta \in \Theta$
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 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 vai ser a estimação não paramétrica da forma da função $f(y;\theta)$.

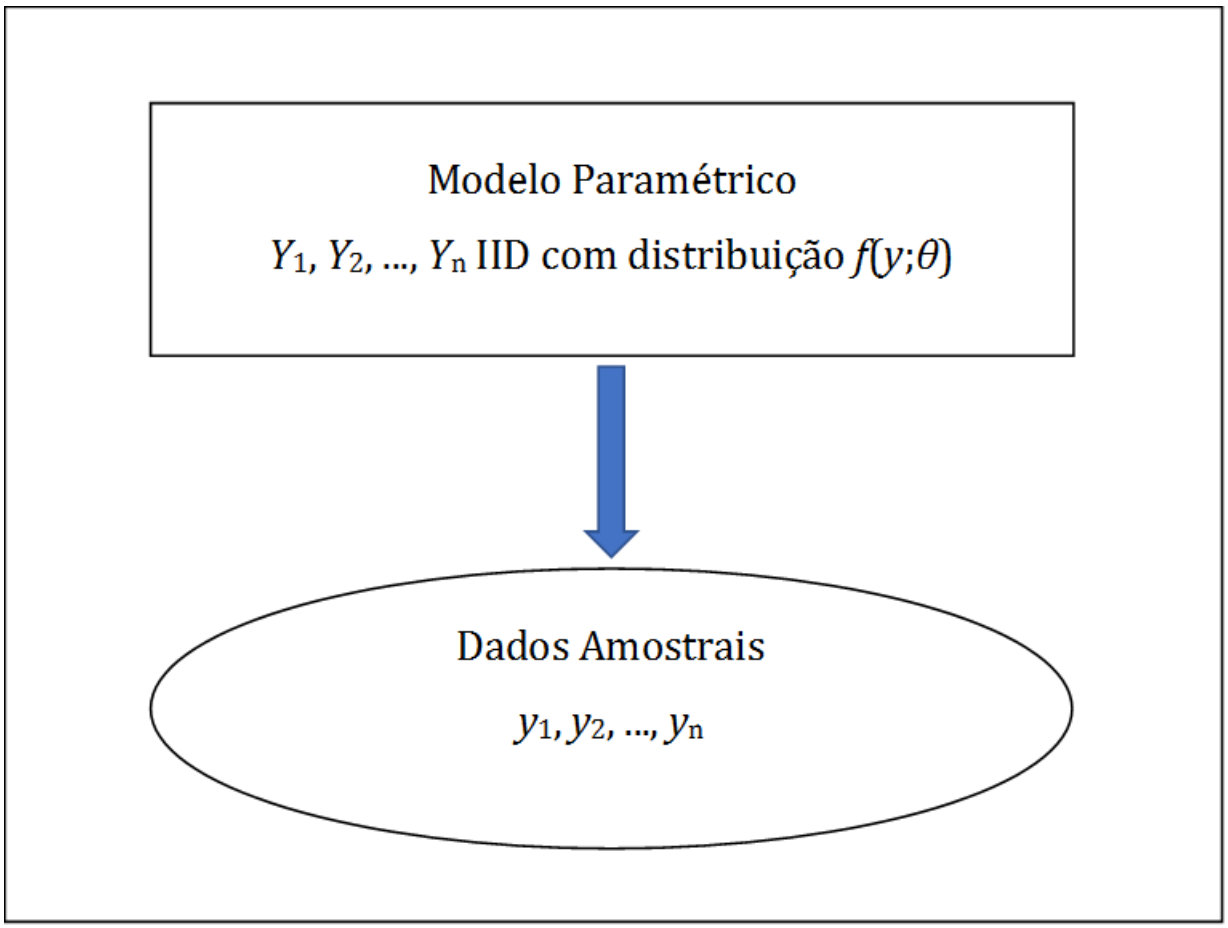

Figura 2.1: Representação esquemática da Modelagem Clássica