Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-08-20

Contents

Prefácio

Chapter 1

Introdução

Chapter 2

Referencial para Inferência

2.1 Modelagem - Primeiras Idéias

Com o objetivo de dar uma primeira ideia sobre o assunto a ser tratado neste livro vamos considerar, numa situação simples, algumas abordagens alternativas de 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 a seguir, e uma descrição esquemática resumida é apresentada na Tabela 2.1.

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.).

Table 2.1: Representação esquemática da abordagem 1

Abordagem 1 - Modelagem Clássica	
Dados Amostrais	$\begin{array}{ccc} Y_1 & & Y_n \\ \downarrow & \ldots, & \downarrow \\ y_1 & & y_n \end{array}$
Modelo Paramétrico/ Hipóteses	$Y_1, \ldots, Y_n$ vetores aleatórios IID com distribuição $f(y;\theta)$, onde $\theta \in \Theta$
Objetivo	Inferir sobre θ usando observações $y_1, \ldots, y_n$

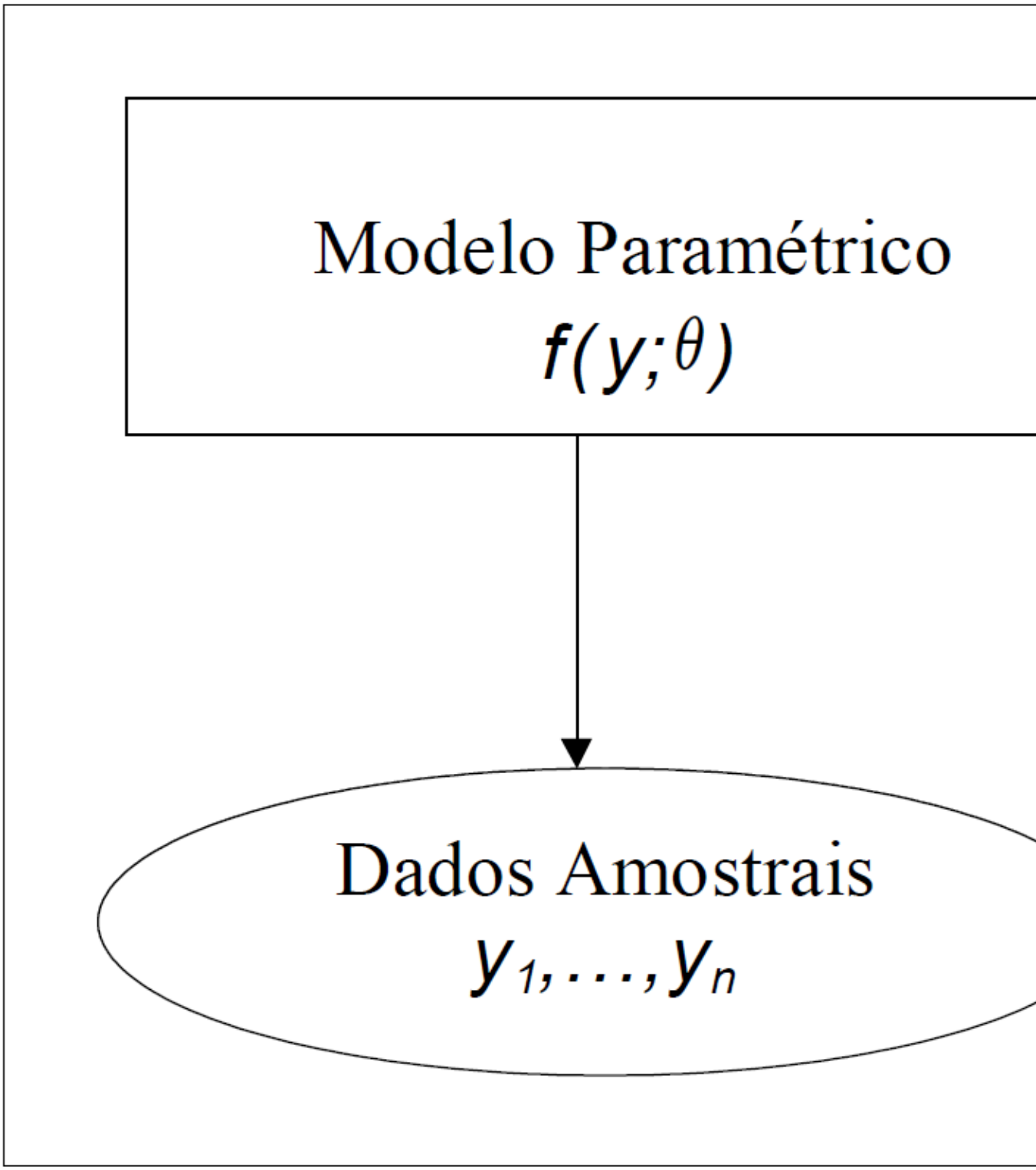

Figure 2.1: Modelagem Clássica

Table 2.2: Representação esquemática da abordagem 2

Abordagem 2 - Amostragem Probabilística	
Dados Amostrais	$\begin{array}{ccc} Y_{i_1} & & Y_{i_n} \\ \downarrow & , \ldots , & \downarrow \\ y_{i_1} & & y_{i_n} \end{array}$
Hipóteses/Modelo	extraídos de $y_1, \ldots, y_N$ segundo $p(a)$
Objetivo	Inferir sobre sobre funções $g(y_1, \ldots, y_N)$ usando $y_{i_1}, \ldots, y_{i_n}$

Vários exemplos discutidos ao longo dos Capítulos 6, 7 e 8 ilustram situações deste tipo. No Capítulo 9 o foco vai ser a estimação não paramétrica da forma da função $f(y;\theta)$.

Investigando a existência de diferenciais de salários por sexo e raça

Uma questão de grande interesse para o debate sobre a existência de desigualdades numa sociedade diz respeito à possível existência de diferenciais de salários entre pessoas de sexo e raça distintos, após controlar por características do trabalhador tais como escolaridade, ocupação e experiência, e da firma, tais como tamanho, setor de atividade e outras. (Rodrigues, 2003) examinou este problema empregando modelos de regressão para explicar o logaritmo do salário hora dos trabalhadores empregados, ajustados a dados obtidos através da Pesquisa sobre Padrões de Vida do IBGE, e tomando como variáveis explicativas características do trabalhador, do posto de trabalho e da empresa. A autora concluiu que não se pode rejeitar a hipótese de existência de discriminação racial e de sexo no mercado de trabalho, pois trabalhadores igualmente produtivos inseridos em trabalhos de características similares apresentavam diferenciais de salários com base em atributos não produtivos, como o sexo, a raça, e o estado civil, por exemplo. Tais conclusões foram obtidas mediante testes de hipóteses sobre valores dos parâmetros do modelo ajustado.

2.1.2 Abordagem 2 - Amostragem Probabilística

A abordagem adotada pelos praticantes de amostragem (amostristas) considera uma população finita $U = \{1, \ldots, N\}$, da qual é selecionada uma amostra $a = \{i_1, \ldots, i_n\}$, segundo um plano amostral caracterizado por $p(a)$, probabilidade de ser selecionada a amostra a, suposta calculável para todas as possíveis amostras. Os valores $y_1, \ldots, y_N$ das variáveis de interesse Y na população finita são considerados fixos, porém desconhecidos. Sem perda de generalidade, podemos reindexar a população de tal forma que a amostra observada seja formada pelos índices $s = \{1, \ldots, n\}$ |

A partir dos valores observados na amostra, 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 no Tabela 2.2, e uma representação gráfica resumida na Figura 2.2.

Investigando a existência de diferenciais de salários por sexo e raça

Ainda no contexto do exemplo da investigação da existência de diferenciais de salários por sexo e raça, os dados da Pesquisa sobre Padrões de Vida do IBGE podem ser utilizados para obter uma tabela cruzada tendo como entradas de linha o sexo dos trabalhadores, e como entrada das colunas um indicador de trabalhadores de raça / cor = branco, sendo as celas da tabela utilizadas para apresentar estimativas dos valores médios dos salários dos trabalhadores nas várias classes definidas pelos valores dos indicadores de linha e coluna.

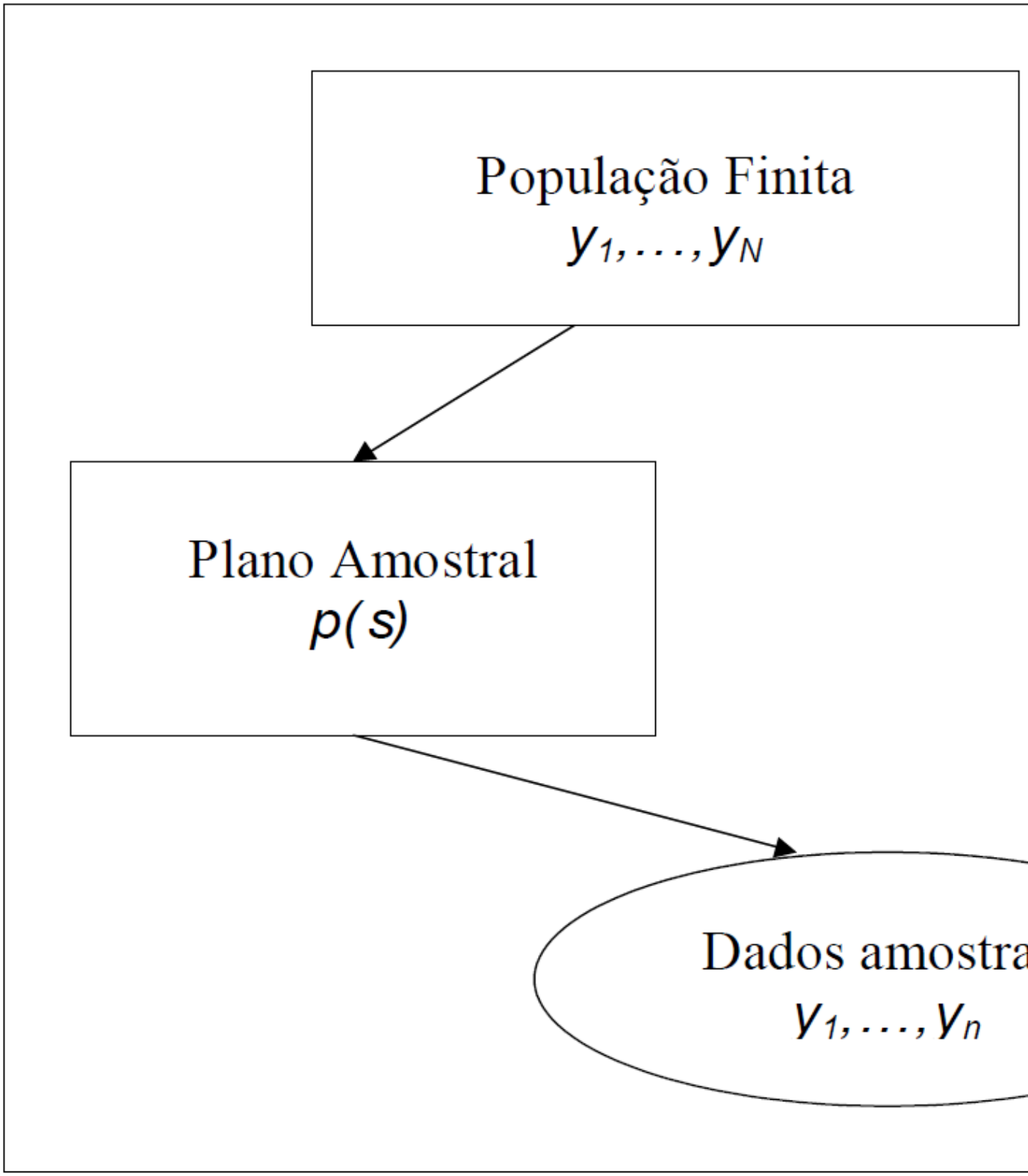

Figure 2.2: Amostragem Probabilística

Tabelas como esta são habitualmente produzidas como resultado da realização de pesquisas amostrais pelas agências oficiais de estatísticas no Brasil e em muitos outros países. Embora capazes de revelar diferenças nos salários médios de trabalhadores de sexo e raça distintos, tais tabelas são insuficientes para investigar a questão da discriminação de gênero ou raça no mercado de trabalho, pois tais diferenças de salários médios poderiam ser explicadas por diferenças de características dos trabalhadores tais como escolaridade, que poderiam ter origem fora do mercado de trabalho. Por outro lado, ilustram bem os alvos de inferência freqüentemente definidos para pesquisas amostrais de populações finitas. Aqui o que se busca estimar é o valor médio de uma característica de interesse (no caso, o salário) para a população finita de onde foi extraída a amostra disponível para análise. Apesar de úteis, tais estimativas representam apenas medidas descritivas da população alvo.

2.1.3 Discussão das Abordagens 1 e 2

A primeira abordagem (Modelagem Clássica), nos termos descritos, foi proposta como modelo para 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, o conceito de aleatoriedade é geralmente introduzido para modelar os erros (não controláveis) no processo de medição.

A segunda abordagem (Amostragem Probabilística) é utilizada principalmente no contexto de estudos sócio-econômicos, para levantamento de dados por agências governamentais produtoras de informações estatísticas. Nesta abordagem, a aleatoriedade é introduzida no processo pelo pesquisador para obtenção dos dados, através do planejamento amostral $p(a)$ utilizado (Neyman, 1934) e as distribuições das estatísticas de interesse são derivadas a partir dessa `distribuição de aleatorização`. Tais planos amostrais podem ser complexos, gerando observações com as características i) a iv) do Capítulo 1. Os dados obtidos são utilizados principalmente para descrição da população finita, sendo calculadas estimativas de totais, médias, razões, etc. Sob essa abordagem, os pontos i) a iv) do Capítulo 1 são devidamente considerados tanto na estimação de parâmetros descritivos desses tipos, como também na estimação de variâncias dos estimadores.

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 dessa abordagem ter sido inicialmente concebida e aplicada para problemas de inferência descritiva sobre populações finitas, é cada vez mais comum, porém, a utilização de dados obtidos através de pesquisas amostrais complexas para fins analíticos, com a aplicação de métodos de análise desenvolvidos e apropriados para a `abordagem 1`.

Diante do exposto, podemos considerar algumas questões de interesse.

- É adequado aplicar métodos de análise da `abordagem 1`, concebidos para observações 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 2`? 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, também é de interesse a questão da robustez da modelagem, traduzida nas seguintes perguntas.

- O que acontece quando o modelo adotado na `abordagem 1` não é verdadeiro?
- Neste caso, qual a interpretação do parâmetro na `abordagem 1`?
- Ainda neste caso, as quantidades descritivas populacionais da `abordagem 2` 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 `abordagem 1`, e além disso

Table 2.3: Representação esquemática da abordagem 3

Abordagem 3 - Modelagem de Superpopulação	
Dados amostrais	$\begin{array}{ccc} Y_{i_1} & & Y_{i_n} \\ \downarrow & , \ldots , & \downarrow \\ y_{i_1} & & y_{i_n} \end{array}$
População e esquema de seleção	Extraídos de $y_1, \ldots, y_N$ segundo $p(s)$
Modelo para população	$Y_1, \ldots, Y_N$ vetores aleatórios IID com distribuição $f(y;\theta)$, onde $\theta \in \Theta$
Parâmetro-alvo	associar $\theta \longleftrightarrow g(Y_1, \ldots Y_N)$
Objetivo	Inferir sobre $g(Y_1, \ldots Y_N)$ a partir de $y_{i_1}, \ldots, y_{i_n}$ usando $p(s)$

incorpora na análise os pontos i) a iii) do Capítulo 1 mediante aproveitamento da estrutura do planejamento amostral como na `abordagem 2`.

2.1.4 Abordagem 3 - Modelagem de Superpopulação

Nesta abordagem, os valores $y_1, \ldots, y_N$ das variáveis de interesse Y na população finita são considerados observações ou realizações dos vetores aleatórios $Y_1, \ldots, Y_N$, supostos 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 `abordagem 1`, o modelo probabilístico aqui é especificado para descrever o mecanismo aleatório que gera `a população`, não `a amostra`, muito embora na maioria das aplicações práticas a população não será jamais observada por inteiro. Não obstante, ao formular modelos para a 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 populações similares, quer existam ao mesmo tempo, quer se refiram a estados futuros (ou passados) da mesma população.

Utilizando um plano amostral definido por $p(a)$, 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 inferências sobre o parâmetro θ, considerando os pontos i) a iii) do Capítulo 1. Veja uma representação gráfica resumida desta abordagem na Figura 2.3.

Adotando o modelo de superpopulação e considerando métodos usuais disponíveis na `abordagem 1`, podemos utilizar funções de $y_1, \ldots, y_N$, digamos $g(y_1, \ldots, y_N)$, para fazer inferências sobre θ. Desta forma, definimos estatísticas $(y_1, \ldots, y_N)$ (no sentido da `abordagem 1`) que são quantidades descritivas populacionais (parâmetros populacionais no contexto da `abordagem 2`), que passam a ser os novos parâmetros-alvo. O passo seguinte é utilizar métodos disponíveis na `abordagem 2` para fazer inferência sobre $g(y_1, \ldots, y_N)$ baseada em $y_{i_1}, \ldots, y_{i_n}$. 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 no Tabela 2.3.

A descrição da abordagem adotada neste livro foi apresentada de maneira propositadamente simplificada e vaga nesta seção, mas será aprofundada ao longo do texto. Admitiremos que o leitor esteja familiarizado com a `abordagem 1` e com as noções básicas da `abordagem 2`. A título de recordação, serão apresentados no Capítulo 2.4 alguns resultados básicos da Teoria de Amostragem. A ênfase do texto, porém, será na apresentação da `abordagem 3`, sendo para isto apresentados os elementos indispensáveis das `abordagens 1 e 2`.

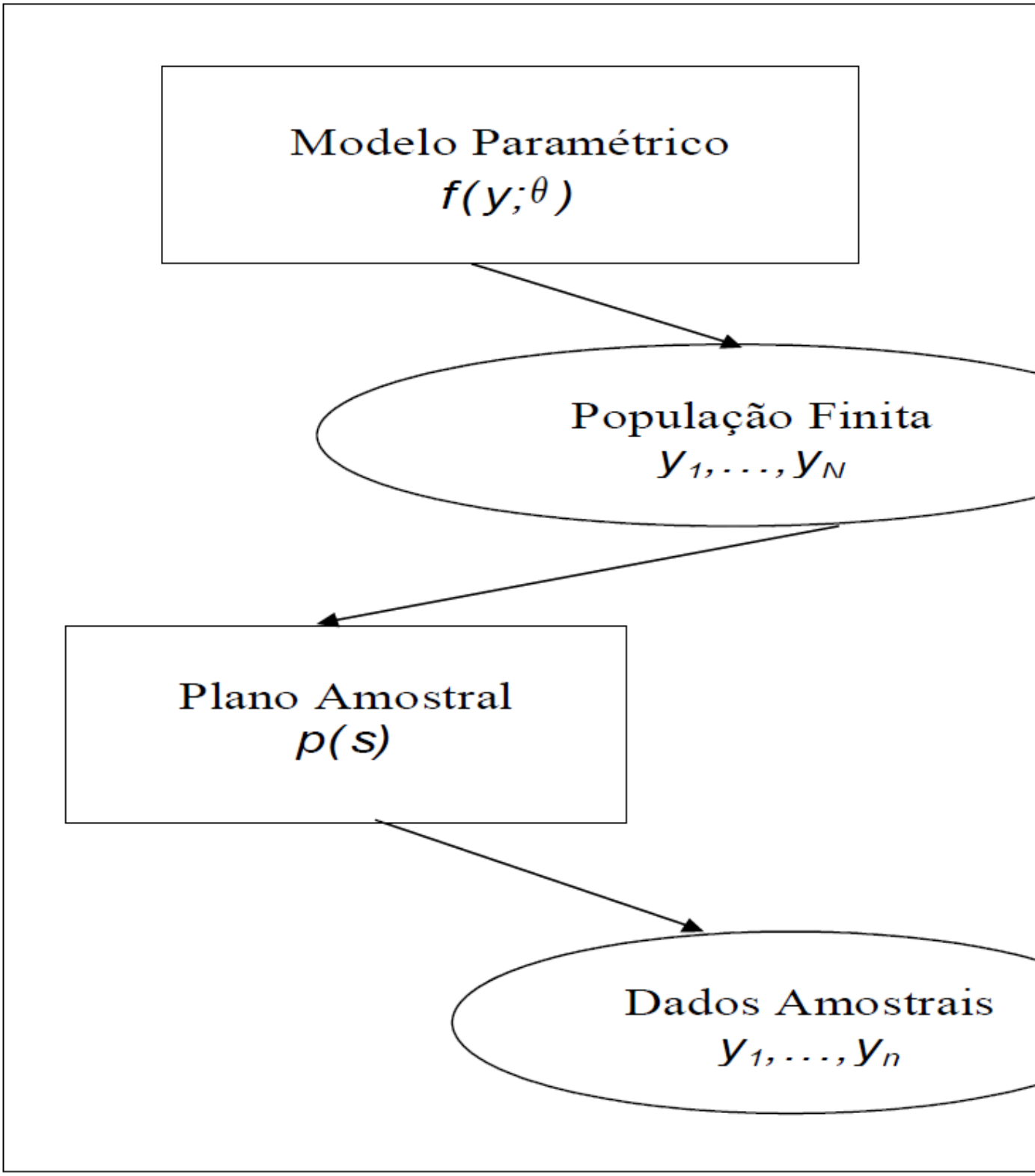

Figure 2.3: Modelagem de Superpopulação

Ao construir e ajustar modelos a partir de dados de pesquisas amostrais `complexas`, tais como as executadas pelo IBGE, o usuário precisa incorporar as informações sobre pesos e planos amostrais utilizados. Em geral, ao publicar os resultados das pesquisas, os pesos são considerados, sendo possível produzir estimativas pontuais `corretas` utilizando os pacotes tradicionais. Por outro lado, para construir intervalos de confiança e testar hipóteses sobre parâmetros de modelos, seria preciso o conhecimento das estimativas de variâncias e covariâncias das estimativas, obtidas a partir 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 na maioria métodos aproximados para estimar matrizes de covariância, tais como os de Máxima Pseudo-Verossimilhança e de Linearização, que serão 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 contra-indicações. 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 faremos uso intensivo da library `survey` disponível no R, mas o leitor encontrará funcionalidade semelhante em vários outros pacotes. Nossa escolha se deveu a dois fatores principais: primeiro ao fato do pacote R ser aberto, livre e gratuito, dispensando o usuário de custos de licenciamento, bem como possibilitando aos interessados o acesso ao código fonte e a capacidade de modificar as rotinas de análise, caso necessário. O segundo fator é de natureza mais técnica, porém transitória. No presente momento, a library `survey` é a coleção de rotinas mais completa e genérica para análise de dados amostrais complexos existente, dispondo de rotinas capazes de ajustar os modelos usuais mas também de ajustar modelos não convencionais mediante maximização 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 pacote R, mas acreditamos que os benefícios do aprendizado desta nova ferramenta compensam largamente os custos adicionais do aprendizado.

2.2 Fontes de Variação

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

1. `modelo de superpopulação`, que descreve o processo subjacente que por hipótese gera as medidas verdadeiras de qualquer unidade 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. `planejamento amostral`, que estabelece o mecanismo pelo qual unidades da população são selecionadas para participar da pesquisa por amostra.

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 são disponibilizados 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) será considerada no capítulo onze. 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 serão considerados valores corretos ou verdadeiros. Admitiremos 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 (Montanari, 1987).

Portanto, o referencial aqui adotado considera apenas duas fontes alternativas de variação: o modelo de superpopulação (1) e o plano amostral (3). Estas fontes alternativas de variação, descritas nesta seção apenas de forma esquemática, são discutidas com maiores detalhes a seguir.

A fonte de variação (1) será 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) será considerada porque a atenção será focalizada na análise de dados obtidos através de pesquisas amostrais 'complexas. Aqui a discussão se restringirá a planos amostrais aleatorizados ou de amostragem probabilística, não sendo considerados métodos intencionais ou outros métodos não-aleatórios de seleção de amostras.

2.3 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 tomaremos $U = \{1, ..., N\}$. Uma pesquisa cobrindo n elementos distintos numa amostra a, $a = \{i_1, ..., i_n\} \subset U$, é realizada para medir os valores de P variáveis de interesse da pesquisa, doravante denominadas simplesmente variáveis da pesquisa.

Denote por $\mathbf{y}_i = (y_{i1}, ..., y_{iP})'$ o vetor $P \times 1$ de valores das variáveis da pesquisa e por $\mathbf{x}_i = (x_{i1}, ..., x_{iQ})'$ o 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 planejamento amostral e a estimação a partir da amostra, como se discutirá com mais detalhes adiante. Denote por $\mathbf{y}_U$ a matriz $N \times P$ formada empilhando os vetores transpostos das observações 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 (marginal) de superpopulação, que doravante denotaremos simplesmente por $f(\mathbf{y}_U; \theta)$, ou de forma abreviada, por M para indicar que se trata do modelo `marginal` de superpopulação. Esperanças e variâncias definidas com respeito à distribuição do modelo marginal 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. Representando por $f(\mathbf{y}_U, \ldots, \mathbf{x}_U; \eta)$ a função de densidade de probabilidade 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 elementos 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:

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

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

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 de amostragem encontrada nos livros clássicos tais como (Cochran, 1977) e outros. Aqui esta abordagem é chamada de abordagem baseada no planejamento amostral ou *abordagem de aleatorização*, pois neste caso a única fonte de variação (aleatoriedade) é proveniente do planejamento amostral. Em geral, a distribuição conjunta de $\left(\mathbf{Y}_U,\mathbf{X}_U\right)$ não precisa ser degenerada como em (2.4), embora o referencial aqui adotado seja suficientemente geral para permitir considerar esta possibilidade.

Se todos os elementos fossem pesquisados (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. 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*.

2.4 Planejamento 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 $a = \{i_1,\ldots,i_n\}$ dos distintos elementos na amostra, onde n $(1 \leq n \leq N)$ é o número de elementos na amostra a, chamado 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 a da população finita U é chamado plano amostral. Uma forma de caracterizá-lo é através da função $p(.)$, onde $p(a)$ dá a probabilidade de selecionar a amostra a no conjunto A de todas as amostras possíveis. Só mecanismos amostrais envolvendo alguma forma de seleção probabilística bem definida serão aqui considerados, e portanto supõe-se que $0 \leq p(a) \leq 1 \; \forall a \in A$ e $\sum_{a \in A} p(a) = 1$.

Esta caracterização do plano amostral $p(a)$ 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`, veja Seção 2.5. Uma notação mais explícita para indicar esta possibilidade envolveria escrever $p(a)$ como $p\left[a|(\mathbf{y}_U, \mathbf{x}_U)\right]$. Tal notação será 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 $\mathbf{\Delta}_a = [I(1 \in a), \ldots, I(N \in a)]'$ um vetor aleatório de indicadores dos elementos incluídos na amostra a. Então o plano amostral pode ser alternativamente caracterizado pela distribuição de probabilidade de $\mathbf{\Delta}_a$ denotada por $f\left[\delta_a|\left(\mathbf{y}_U, \mathbf{x}_U\right)\right]$, onde δ_a é qualquer realização particular de $\mathbf{\Delta}_a$ tal que ${\delta_a}'\mathbf{1}_N = n$, e $\mathbf{1}_N$ é o vetor unitário de dimensão N.

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

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

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

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

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

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 das expressões das variâncias dos estimadores. Contudo, esta não é uma hipótese crucial, pois há planos amostrais que não a sati fazem e para os quais estão disponíveis aproximações e estimadores satisfatórios das variâncias dos estimadores de totais e de médias.

2.5 Planos Amostrais Informativos e Ignoráveis

Ao fazer inferência usando dados de pesquisas amostrais precisamos distinguir duas situações que requerem tratamento diferenciado. 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, e sendo esse indicador uma das variáveis de pesquisa, que é considerada no mecanismo de seleção da amostra. Os métodos que discutiremos 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

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

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 clássica descrita em 2.1 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. Veja o Exemplo 1 adiante.

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

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 se verá 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 5.

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 das estatísticas amostrais p ou com respeito à distribuição mista Mp.

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

Example 1 *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:

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

	y		
x	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:

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

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:

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

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

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

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 será 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 amostra. 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:

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

	y		
x	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

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

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:

$$f_a(x) = Pr(X_i = x|i \in a) = 1/2, \ \forall x \in \Omega \text{ e } \forall i \in U \tag{2.12}$$

onde a designação f_a 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:

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

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 e na amostra ser a mesma.

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

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

	y		
x	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 (mas também na amostra) é dada por

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

	y		
x	0	1	Total
0	0.875	0.125	1
1	0.5	0.5	1

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

Table 2.8: Distribuição de probabilidades marginal de y na população - $f_U(y)$

y	0	1
$f_U(y)$	0.8	0.2
$f_a(y)$	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 será equivocada, pois a alocação igual da amostra nos estratos leva 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 conseqüência, a proporção de valores de y iguais a 1 na amostra (0.3125) é 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 da amostra em relação à correspondente distribuição na população.

Note que caso a inferência requerida fosse sobre parâmetros da distribuição condicional de y dado x, a amostragem seria ignorável, isto é, $f_a(y|x) = f_U(y|x)$. Assim fica evidenciado também que a noção de que o plano amostral pode ser ignorável depende da inferência desejada. No nosso exemplo, o plano amostral é ignorável para inferência sobre a distribuição condicional de y dado x, mas não é ignorável para inferência sobre a distribuição marginal de y.

Chapter 3

Estimação Baseada no Plano Amostral

Chapter 4

Efeitos do Plano Amostral

Chapter 5

Ajuste de Modelos Paramétricos

Chapter 6

Modelos de Regressão

Chapter 7

Testes de Qualidade de Ajuste

Chapter 8

Testes em Tabelas de Duas Entradas

Chapter 9

Estimação de densidades

Chapter 10

Modelos Hierárquicos

Chapter 11

Não-Resposta

Chapter 12

Diagnóstico de ajuste de modelo

Chapter 13

Agregação vs. Desagregação

Chapter 14

Pacotes para Analisar Dados Amostrais

Chapter 15

Placeholder

Chapter 16

Placeholder

Bibliography

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. John Wiley, Nova Iorque.

Chambers, R. and Skinner, C., editors (2003). *Analysis of Survey Data*. John Wiley, Chichester.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley, Nova Iorque.

Montanari, G. E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55:191–202.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society A*, 97:558–606.

Rodrigues, S. C. (2003). Análise da estrutura salarial revelada pela PPV incorporando peso e plano amostral. Master's thesis, Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71:495–506.