

# Análise de Dados Amostrais Complexos

Djalma Galvão Carneiro Pessoa (Consultor)- IBGE  
Pedro Luis do Nascimento Silva (Pesquisador)- IBGE

17 de Março de 1998



# Índice

0.1	Prefácio . . . . .	6
0.2	Agradecimentos . . . . .	7
1	Introdução . . . . .	9
1.1	Motivação . . . . .	9
1.2	Estrutura do Livro . . . . .	14
2	Referencial para Inferência . . . . .	17
2.1	Modelagem - Primeiras Idéias . . . . .	17
2.1.1	Abordagem 1 - Modelagem Clássica . . . . .	17
2.1.2	Abordagem 2 - Amostragem Probabilística . . . . .	19
2.1.3	Discussão das Abordagens 1 e 2 . . . . .	19
2.1.4	Abordagem 3 - Modelagem de Superpopulação . . . . .	22
2.2	Fontes de Variação . . . . .	24
2.3	Modelos de Superpopulação . . . . .	25
2.4	Planejamento Amostral . . . . .	28
2.5	Planos Amostrais Informativos e Ignoráveis . . . . .	29
3	Estimação Baseada no Plano Amostral . . . . .	33
3.1	Estimação de Totais . . . . .	33
3.2	Por que Estimar Variâncias . . . . .	37
3.3	Linearização de Taylor para Estimar Variâncias . . . . .	39
3.4	Método do Conglomerado Primário . . . . .	42
3.5	Métodos de Replicação . . . . .	43
4	Efeitos do Plano Amostral . . . . .	47
4.1	Introdução . . . . .	47
4.2	Efeito do Plano Amostral (EPA) de Kish . . . . .	48
4.3	Efeito do Plano Amostral Ampliado . . . . .	51
4.4	Intervalos de Confiança e Testes de Hipóteses . . . . .	61

4.5	Efeitos Multivariados de Plano Amostral . . . . .	64
5	Ajuste de Modelos Paramétricos . . . . .	71
5.1	Introdução . . . . .	71
5.2	Método de Máxima Verossimilhança (MV) . . . . .	73
5.3	Ponderação de Dados Amostrais . . . . .	74
5.4	Método de Máxima Pseudo-Verossimilhança . . . . .	78
5.5	Robustez do Procedimento MPV . . . . .	83
5.6	Desvantagens da Inferência Baseada em Aleatorização . . . . .	84
6	Modelos de Regressão . . . . .	87
6.1	Modelo de Regressão Linear Normal . . . . .	87
6.1.1	Especificação do Modelo . . . . .	87
6.1.2	Pseudo-parâmetros do Modelo . . . . .	88
6.1.3	Estimadores de MPV dos Parâmetros do Modelo . . . . .	90
6.1.4	Estimação da Variância de Estimadores de MPV . . . . .	91
6.2	Modelo de Regressão Logística . . . . .	92
6.3	Teste de Hipóteses . . . . .	100
7	Testes de Qualidade de Ajuste . . . . .	103
7.1	Introdução . . . . .	103
7.2	Teste para uma Proporção . . . . .	104
7.2.1	Correção de Estatísticas Clássicas . . . . .	104
7.2.2	Estatística de Wald . . . . .	108
7.3	Teste para Várias Proporções . . . . .	109
7.3.1	Estatística de Wald Baseada no Plano Amostral . . . . .	110
7.3.2	Situações Instáveis . . . . .	110
7.3.3	Estatística de Pearson com Ajuste de Rao-Scott . . . . .	111
8	Testes em Tabelas de Duas entradas . . . . .	119
8.1	Introdução . . . . .	119
8.2	Tabelas 2x2 . . . . .	120
8.2.1	Teste de Independência . . . . .	120
8.2.2	Teste de Homogeneidade . . . . .	121
8.2.3	Efeitos de Plano Amostral nas Celas . . . . .	121
8.3	Tabelas de Duas Entradas (Caso Geral) . . . . .	123
8.3.1	Teste de Homogeneidade . . . . .	123
8.3.2	Teste de Independência . . . . .	127
8.3.3	Estatística de Wald Baseada no Plano Amostral . . . . .	128
8.3.4	Estatística de Pearson com Ajuste de Rao-Scott . . . . .	129

9	Agregação vs. Desagregação	135
9.1	Introdução . . . . .	135
9.2	Modelagem da Estrutura Populacional . . . . .	136
9.3	Modelos Hierárquicos . . . . .	139
9.4	Análise Desagregada: Prós e Contras . . . . .	151
10	Pacotes para Análise de Dados Amostrais	155
10.1	Introdução . . . . .	155
10.2	Pacotes Computacionais . . . . .	155

## 0.1 Prefácio

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, 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.

## 0.2 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 usuários de dados obtidos através de pesquisas amostrais por agências produtoras de informações estatísticas. Tais dados são comumente utilizados em análises descritivas envolvendo o cálculo de estimativas para totais, proporções, médias e razões, nas quais, em geral, são devidamente considerados os pesos distintos das observações e o planejamento da amostra que lhes deu origem.

Outro uso destes dados, denominado secundário, é a construção de modelos, feita geralmente por analistas que trabalham fora das agências produtoras dos dados. Neste caso, o foco é, essencialmente, estabelecer a natureza de relações ou associações entre variáveis. Para isto, a estatística clássica conta com um arsenal de ferramentas de análise, já incorporado aos principais pacotes estatísticos disponíveis. O uso destes pacotes se faz, entretanto, sob condições que não refletem a complexidade usualmente envolvida nas pesquisas amostrais de populações finitas. Em geral, partem de hipóteses básicas que só são válidas quando os dados são obtidos através de amostras aleatórias simples com reposição (AASC). Tais pacotes estatísticos não consideram os seguintes aspectos relevantes no caso de amostras complexas:

- i) probabilidades distintas de seleção das unidades;
- ii) conglomeração das unidades;
- iii) estratificação;
- iv) não-resposta e outros ajustes.

As estimativas pontuais de parâmetros da população são influenciadas por pesos distintos das observações. Além disso, as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos. Ao ignorar estes aspectos, os pacotes tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

A seguir vamos apresentar um exemplo de uso de dados de uma pesquisa amostral real para ilustrar como os pontos i) a iv) mencionados afetam a inferência sobre quantidades descritivas populacionais tais como médias, proporções, razões e totais.

### Exemplo 1.1 *Distribuição dos pesos da amostra da PPV*

Os dados deste exemplo são relativos à distribuição dos pesos na amostra da Pesquisa Domiciliar sobre Padrões de Vida (PPV), realizada pelo IBGE nas Regiões Nordeste e Sudeste do Brasil nos anos 96-97. Segundo Albieri e Bianchini(1997), "A Pesquisa Domiciliar sobre Padrões de Vida (PPV) foi realizada nas Regiões Nordeste e Sudeste do País, considerando 10 estratos geográficos, a saber: Região Metropolitana de Fortaleza, Região Metropolitana de Recife, Região Metropolitana de Salvador, restante da área urbana do Nordeste, restante da área rural do Nordeste, Região Metropolitana de Belo Horizonte, Região Metropolitana do Rio de Janeiro, Região Metropolitana de São Paulo, restante da área urbana do Sudeste e restante da área rural do Sudeste.

O plano amostral foi de dois estágios, com estratificação das unidades primárias e probabilidade de seleção proporcional ao tamanho e seleção aleatória das unidades de segundo estágio. A unidade primária é o setor da base geográfica do Censo Demográfico de 91 e a unidade do segundo estágio é o domicílio. O tamanho da amostra para cada estrato geográfico foi fixado em 480 domicílios. Em cada estrato geográfico foi fixado em 60 o número de setores a serem selecionados e 8 domicílios em cada setor, com exceção para os estratos que correspondem ao restante da área rural de cada Região onde fixou-se em 30 o número de setores e em 16 o número de domicílios a serem selecionados por setor, em função da dificuldade de acesso a esses setores, o que implicaria em aumento de custo".

Os setores de cada um dos 10 estratos geográficos foram subdivididos em 3 estratos de acordo com a renda média mensal do chefe da família por setor, perfazendo um total de 30 estratos geográficos versus renda. Em seguida foi feita uma alocação proporcional, com base no número de domicílios particulares permanentes ocupados do estrato de renda no universo de cada estrato geográfico, obtidos pelo Censo de 91. No final foram obtidos 554 setores na amostra, distribuídos tal como revela a Tabela 1.1.

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

Estrato Geográfico	Número de setores	
	População	Amostra
1-RM Fortaleza	2.268	62
2-RM Recife	2.310	61
3-RM Salvador	2.188	61
4-Restante Nordeste Urbano	15.068	61
5-Restante Nordeste Rural	23.723	33
6-RM Belo Horizonte	3.287	62
7-RM Rio de Janeiro	10.427	61
8-RM São Paulo	14.964	61
9-Restante Sudeste Urbano	25.902	61
10-Restante Sudeste Rural	12.038	21
Total	112.175	554

Tabela 1.2: Distribuição dos pesos da amostra da PPV

Região	Mínimo	Q1	Mediana	Q3	Máximo
Nordeste	724	1.159	1.407	6.752	15.348
Sudeste	991	2.940	5.892	10.496	29.234
Nordeste + Sudeste	724	1.364	4.034	8.481	29.234

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.

No cálculo dos pesos foram consideradas as probabilidades de inclusão dos elementos na amostra bem como correções devido a não-resposta. Contudo, a grande variabilidade dos pesos amostrais da PPV é devida 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 bem distintos. Por exemplo, a razão entre o maior e o menor peso é cerca de 40 vezes. 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, etc.

Por outro lado, 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). Contudo, a variabilidade dos pesos produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da conglomeração e 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 tipos usando a amostra da PPV juntamente com estimativas das respectivas variâncias. Essas estimativas de variâncias foram calculadas sob duas estratégias: considerando amostragem aleatória simples (portanto ignorando o plano amostral efetivamente adotado), e 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 usando o pacote SUDAAN (Shah et al., 1992) para cada uma das estimativas consideradas. Essa razão fornece uma medida do efeito de ignorar o plano amostral (EPA). Os resultados são apresentados na Tabela 1.3.

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.

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

"Parâmetro" Populacional	Estimativa	Desvio padrão	EPA
1) Número médio de pessoas por domicílio	3,62	0,05	2,64
2) % de domicílios alugados	16,70	1,15	2,97
3) Número total de pessoas que avaliaram seu estado de saúde como ruim	1.208.123	146.681	3,37
4) Total de analfabetos de 7 a 14 anos	1.174.220	127.982	2,64
5) Total de analfabetos de mais de 14 anos	4.792.344	318.877	4,17
6) % de analfabetos de 7 a 14 anos	11,87	1,18	2,46
7) % de analfabetos de mais de 14 anos	10,87	0,67	3,86
8) Total de mulheres de 12 a 49 anos que tiveram filhos	10.817.590	322.947	2,02
9) Total de mulheres de 12 a 49 anos que tiveram filhos vivos	10.804.511	323.182	2,02
10) Total de mulheres de 12 a 49 anos que tiveram filhos mortos	709.145	87.363	2,03
11) Número médio de filhos tidos por mulheres de 12 a 49 anos	1,39	0,03	1,26
12) Razão de dependência	0,53	0,01	1,99

O objetivo deste livro é analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise de dados, e apresentar os ajustes necessários, de modo a considerar de forma apropriada os pontos i) a iii) mencionados nesta seção. Para isto serão apresentados exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também pacotes estatísticos especializados, no caso o SUDAAN (Shah et al., 1992) e o STATA (Stata, 1997). A comparação dos resultados das análises feitas das duas formas permitirá avaliar o impacto de não se considerar os pontos i) a iii) anteriormente citados. O ponto iv) não será diretamente tratado neste texto. O leitor interessado na análise de dados sujeitos a não-resposta pode consultar Kalton(1983a), Little e Rubin(1987), e Rubin(1987), por exemplo.

## 1.2 Estrutura do Livro

O livro está organizado em dez capítulos. Este primeiro capítulo discute a motivação para estudar o assunto e apresenta uma idéia geral 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 Nascimento Silva(1996, cap. 2) e Skinner, Holt e Smith(1989, cap. 1). Esse referencial tem evoluído ao longo dos anos como uma forma de permitir a incorporação de idéias e procedimentos de análise e inferência usualmente associados à Estatística Clássica à prática da 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 na seção 2.1.

No terceiro capítulo apresentamos uma revisão sucinta, a título de 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 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 Särndal, Swensson e Wretman(1992, cap. 2 e 3), Wolter(1985)

e Cochran(1977).

No Capítulo 4 introduzimos o conceito de *Efeito do Plano Amostral (EPA)*, que permite avaliar o impacto de ignorar a estruturação 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 dos dados estruturados ou sob a distribuição de aleatorização do plano amostral efetivamente utilizado (plano amostral verdadeiro). Aqui a referência principal foi o livro Skinner, Holt e Smith(1989), complementado com o texto de Lehtonen e Pahkinen(1995).

No Capítulo 5 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 regulares de interesse, como também das variâncias dessas estimativas. As referências básicas utilizadas nesse capítulo foram Skinner(1989b), Pfeffermann(1993), Binder(1983) e Nascimento Silva(1996, cap. 6).

O Capítulo 6 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 90, onde ajustamos um modelo de regressão logística. Neste exemplo, são feitas comparações entre resultados de ajustes obtidos através de um programa especializado, o SUDAAN (Shah et al., 1992), e através de um programa de uso geral, o S-Plus. As referências centrais são Nascimento Silva(1996, cap. 6) e Binder(1983), além de Pessoa, Nascimento Silva e Duarte(1997).

O sétimo e oitavo capítulos 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 a estatística de Wald baseada no plano amostral. As referências básicas usadas nesses capítulos foram os livros Skinner, Holt e Smith (1989, cap. 4) e Lehtonen e Pahkinen(1995, cap. 7).

No Capítulo 9 discutimos uma forma alternativa de analisar dados de pesquisas complexas, denominada *análise desagregada*. Nesta abordagem, incorporamos explicitamente na análise vários aspectos do plano amostral

utilizado através do emprego de modelos hierárquicos (Bryk e Raudenbush, 1992). 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 e teste de hipóteses.

Por último, no Capítulo 10, 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. Vários 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 idéias 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 médias, totais, 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 idéias 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 idéias 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, o livro de Skinner, Holt e Smith(1989) é a referência central a pesquisar. Para aqueles querendo um tratamento ainda mais prático que o nosso, o livro de Lehtonen e Pahkinen(1995) pode ser uma opção interessante.

## Capítulo 2

# Referencial para Inferência

### 2.1 Modelagem - Primeiras Idéias

Com o objetivo de dar uma primeira idéia sobre o assunto a ser tratado neste livro vamos considerar, numa situação simples, algumas abordagens de análise estatística.

#### 2.1.1 Abordagem 1 - Modelagem Clássica

Seja  $Y$  uma variável de interesse, e sejam  $y_1, \dots, y_n$ ,  $n$  observações desta variável. Em Inferência Estatística, o modelo usual considera  $y_1, \dots, y_n$  como valores (realizações) de variáveis aleatórias  $Y_1, \dots, Y_n$ . Aqui  $Y_1, \dots, Y_n$  são variáveis aleatórias independentes e identicamente distribuídas (IID), com a mesma distribuição de  $Y$ , digamos com função de densidade ou de frequência  $f(y; \theta)$ , onde  $\theta \in \Theta$  é o parâmetro indexador da distribuição  $f$ , e  $\Theta$  é o espaço paramétrico. A partir das observações  $y_1, \dots, y_n$ , são feitas inferências a respeito do parâmetro  $\theta$ . Uma representação gráfica 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  $\theta$  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 do parâmetro  $\theta$ , e a inferência sobre  $\theta$  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.).

Figura 2.1: Modelagem Clássica

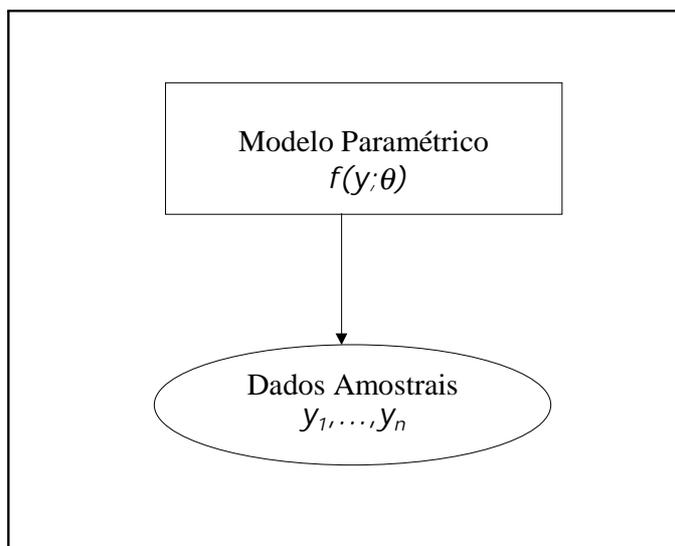


Tabela 2.1: Representação esquemática da abordagem 1

Abordagem 1 - Modelagem Clássica	
Dados Amostrais	$Y_1 \quad Y_n$ $\downarrow \quad \dots, \quad \downarrow$ $y_1 \quad y_n$
Modelo Paramétrico/ Hipóteses	$Y_1, \dots, Y_n$ variáveis aleatórias IID com distribuição $f(y; \theta)$ , onde $\theta \in \Theta$
Objetivo	Inferir sobre $\theta$ usando observações $y_1, \dots, y_n$

Tabela 2.2: Representação esquemática da abordagem 2

Abordagem 2 - Amostragem Probabilística	
Dados Amostrais	$Y_1$ $Y_n$ $\downarrow$ , ... , $\downarrow$ $y_1$ $y_n$
Hipóteses/Modelo	extraídos de $y_1, \dots, y_N$ segundo $p(s)$
Objetivo	Inferir sobre sobre funções $g(y_1, \dots, y_N)$ usando $y_1, \dots, y_n$

### 2.1.2 Abordagem 2 - Amostragem Probabilística

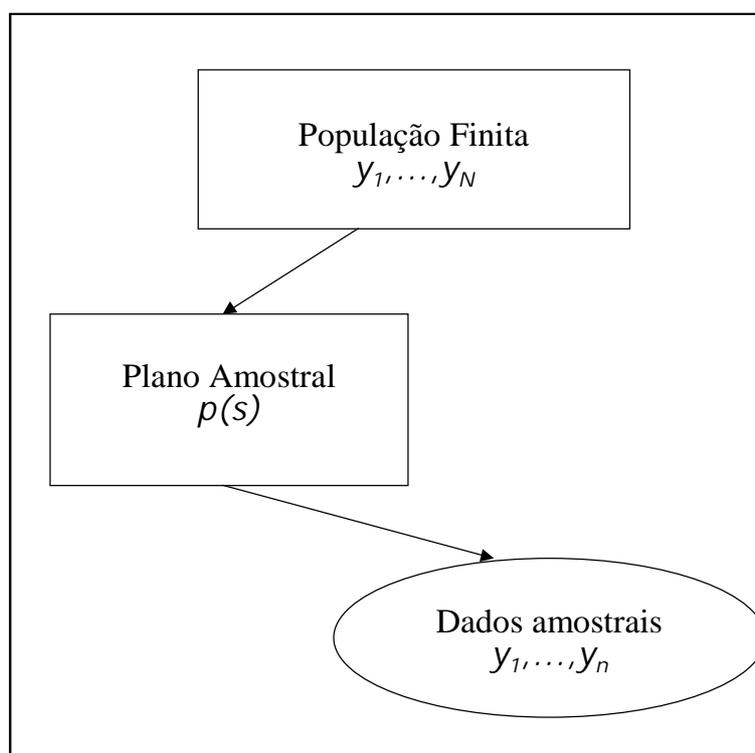
A abordagem adotada pelos praticantes de amostragem (amostristas) considera uma população finita  $U = \{1, \dots, N\}$ , da qual é selecionada uma amostra  $s = \{i_1, \dots, i_n\}$ , segundo um plano amostral caracterizado por  $p(s)$ , probabilidade de ser selecionada a amostra  $s$ , assumida calculável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

A partir dos valores observados na amostra, denotados por  $y_1, \dots, y_n$ , são feitas inferências a respeito de funções dos valores populacionais, digamos  $g(y_1, \dots, y_N)$ . 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  $g(y_1, \dots, y_N) = \sum_{i=1}^N y_i$ , médias  $g(y_1, \dots, y_N) = N^{-1} \sum_{i=1}^N y_i$ , proporções, etc. Uma descrição esquemática resumida dessa abordagem é apresentada no Tabela 2.2, e uma representação gráfica resumida na Figura 2.2.

### 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 con-

Figura 2.2: Amostragem Probabilística



ceito 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  $p(s)$  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. Nessa abordagem, os pontos i) a iv) do Capítulo 1 são devidamente considerados na estimação de parâmetros descritivos desse tipo, e também na estimação de variâncias dos estimadores.

Esta abordagem 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 da população finita, 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*?

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 interpretação ou utilidade?

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 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  $y_1, \dots, y_N$  da variável de interesse  $Y$  na população finita são considerados observações ou realizações das variáveis aleatórias  $Y_1, \dots, Y_N$ , supostas IID com distribuição  $f(y; \theta)$ , onde  $\theta \in \Theta$ . Este modelo é denominado *modelo de superpopulação*. Utilizando um plano amostral definido por  $p(s)$ , obtemos os valores na amostra  $y_1, \dots, y_n$ . A partir de  $y_1, \dots, y_n$  (não considerados IID, em geral) queremos fazer inferências sobre o parâmetro  $\theta$ , 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  $y_1, \dots, y_N$ , digamos  $g(y_1, \dots, y_N)$ , para fazer inferências sobre  $\theta$ . Desta forma, definimos estatísticas  $g(y_1, \dots, y_N)$  (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  $g(y_1, \dots, y_N)$  baseada em  $y_1, \dots, y_n$ . Note que não é possível basear a inferência nos valores populacionais  $y_1, \dots, y_N$ , já que estes não são conhecidos. Este último passo adiciona a informação sobre o plano amostral utilizado, contida em  $p(s)$ , à informação estrutural contida em  $\{f(y; \theta); \theta \in \Theta\}$ . Uma representação esquemática dessa abordagem é apresentada no Tabela 2.3.

A descrição da abordagem adotada neste livro foi apresentada de maneira propositalmente 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 3 alguns resultados básicos da Teoria de Amostragem. A ênfase

Figura 2.3: Modelagem de Superpopulação

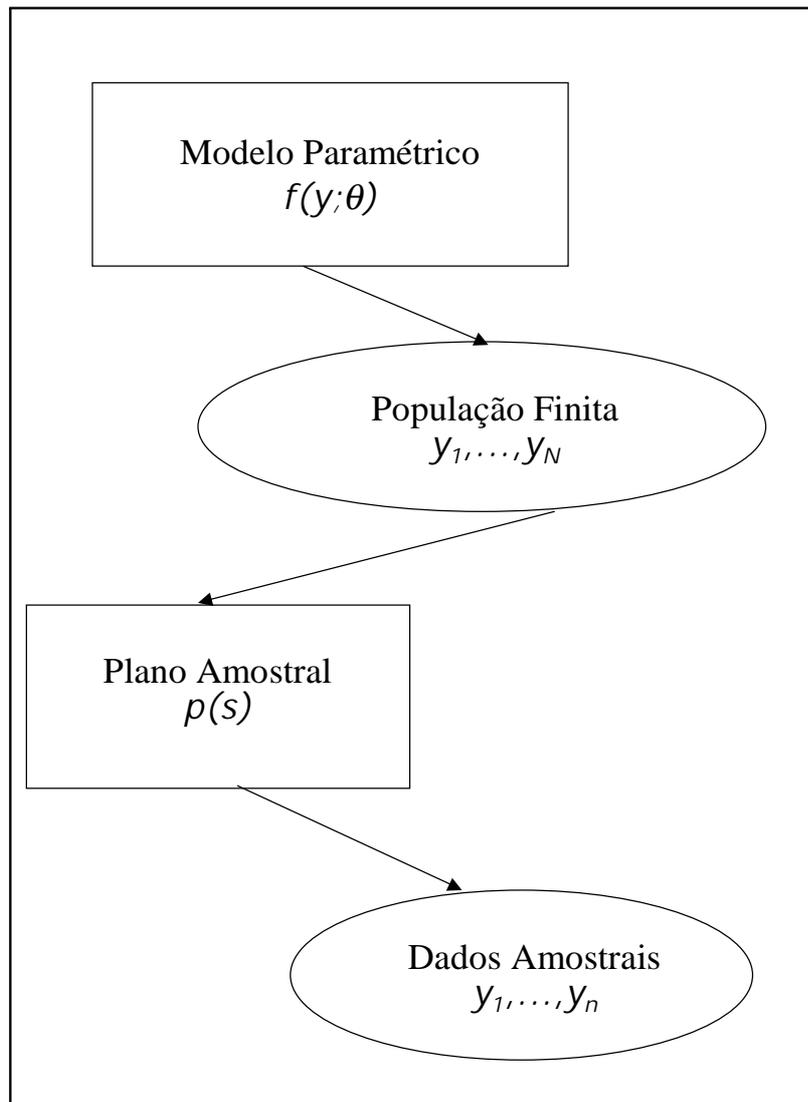


Tabela 2.3: Representação esquemática da abordagem 3

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

do texto, porém, será na apresentação da *abordagem 3*, sendo para isto apresentados os elementos indispensáveis das *abordagens 1 e 2*.

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. Tais pacotes utilizam métodos gerais 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.

## 2.2 Fontes de Variação

Este capítulo estabelece um referencial para inferência em pesquisas amostrais que será usado no restante deste texto. Cassel, Särndal e Wretman (1977, p.1) 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 gerou 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 poderia 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 interesse deste texto, as fontes (2) e (4) não serão consideradas no referencial adotado aqui. 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. 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  $\{1, \dots, N\}$  um conjunto de rótulos que identificam univocamente os  $N$  elementos distintos de uma população-alvo finita  $U$ . Sem perda de gene-

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

Denote por  $\mathbf{y}_i = (y_{i1}, \dots, y_{iR})^0$  o vetor  $R \times 1$  de valores das variáveis da pesquisa e por  $\mathbf{x}_i = (x_{i1}, \dots, x_{iQ})^0$  o vetor  $Q \times 1$  de variáveis auxiliares da  $i$ -ésima unidade da população, respectivamente, para  $i = 1, \dots, N$ . 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.

Quando se supõe que  $\mathbf{y}_1, \dots, \mathbf{y}_N$  são a realização conjunta de vetores aleatórios  $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ , a distribuição conjunta de probabilidade de  $\mathbf{Y}_1, \dots, \mathbf{Y}_N$  é um modelo (marginal) de superpopulação, que doravante denotaremos simplesmente por  $M$ . Analogamente,  $\mathbf{x}_1, \dots, \mathbf{x}_N$  pode ser considerada uma realização conjunta de vetores aleatórios  $\mathbf{X}_1, \dots, \mathbf{X}_N$ . Neste caso, as distribuições das variáveis da pesquisa podem ser alternativamente especificadas condicionando-se nos valores das variáveis auxiliares, ou seja condicionando-se em  $\mathbf{X}_i = \mathbf{x}_i$  para todo  $i \in U$ .

O referencial aqui adotado permite a especificação da distribuição conjunta combinada das variáveis da pesquisa e das variáveis auxiliares. Denote por  $f[(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N); \boldsymbol{\eta}]$  a função de densidade de probabilidade de  $(\mathbf{Y}_1, \mathbf{X}_1), \dots, (\mathbf{Y}_N, \mathbf{X}_N)$ , onde  $\boldsymbol{\eta}$  é um vetor de parâmetros. Como os vetores  $\mathbf{Y}_i$  e  $\mathbf{X}_i$  não têm a mesma dimensão, uma notação mais precisa para esta função de densidade seria  $f[(\mathbf{y}'_1, \mathbf{x}'_1)', \dots, (\mathbf{y}'_N, \mathbf{x}'_N)'; \boldsymbol{\eta}]$ . Por simplicidade, a forma anterior será preferida daqui por diante.

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:

$$f[(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N); \boldsymbol{\eta}] = \prod_{i \in U} f(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\eta}) \quad (2.1)$$

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

onde  $\boldsymbol{\lambda}$  e  $\boldsymbol{\phi}$  são vetores de parâmetros.

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

$$f(\mathbf{y}_1, \dots, \mathbf{y}_N; \boldsymbol{\theta}) = \prod_{i \in U} \int f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\lambda}) f(\mathbf{x}_i; \boldsymbol{\phi}) d\mathbf{x}_i = \prod_{i \in U} f(\mathbf{y}_i; \boldsymbol{\theta}) \quad (2.3)$$

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

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

$$P[(\mathbf{Y}_i, \mathbf{X}_i) = (\mathbf{y}_i, \mathbf{x}_i)] = 1 \quad \forall i \in U \quad (2.4)$$

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

Este modelo foi considerado em Cassel, Särndal e Wretman(1977, p. 2-3), 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. Chamaremos aqui esta abordagem 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  $(\mathbf{Y}_i, \mathbf{X}_i)$  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  $(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N)$ . Sob a hipótese de resposta completa, a única fonte de incerteza seria devida ao fato de que  $(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N)$  é uma realização de  $(\mathbf{Y}_1, \mathbf{X}_1), \dots, (\mathbf{Y}_N, \mathbf{X}_N)$ . Os dados observados poderiam então ser usados para fazer inferências sobre  $\boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\lambda}$  ou  $\boldsymbol{\theta}$  usando procedimentos padrões.

Inferência sobre quaisquer dos parâmetros  $\boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\lambda}$  ou  $\boldsymbol{\theta}$  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  $g(\mathbf{y}_1, \dots, \mathbf{y}_N)$  dos valores das variáveis da pesquisa, o modelo de superpopulação não é estritamente necessário, embora possa ser útil. Inferência para tais quantidades, chamadas parâmetros da população finita ou quantidades descritivas populacionais (QDPs), é chamada *inferência descritiva*.

## 2.4 Planejamento Amostral

Embora censos sejam algumas vezes realizados para coletar dados sobre certas populações, a vasta maioria das pesquisas são 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  $s = \{i_1, \dots, i_n\}$  dos distintos elementos na amostra, onde  $n$  ( $1 \leq n \leq N$ ) é o número de elementos na amostra  $s$ , chamado tamanho da amostra;
2. os valores na amostra das variáveis da pesquisa  $\mathbf{y}_{i_1}, \dots, \mathbf{y}_{i_n}$ ;
3. com informação auxiliar completa, os valores das variáveis auxiliares na amostra  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_n}$  e na população  $\mathbf{x}_1, \dots, \mathbf{x}_N$ ; alternativamente, com informação auxiliar parcial, os valores na amostra  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_n}$ , mais os totais/médias destas variáveis na população.

O mecanismo usado para selecionar a amostra  $s$  da população finita  $U$  é chamado planejamento amostral. Uma forma de caracterizá-lo é através da função  $p(\cdot)$ , 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, e 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  $\mathbf{x}_1, \dots, \mathbf{x}_N$  bem como dos valores das variáveis da pesquisa na população  $\mathbf{y}_1, \dots, \mathbf{y}_N$  (amostragem *informativa*, veja Seção 2.5). Uma notação mais explícita para indicar esta possibilidade possivelmente envolveria escrever  $p(s)$  como  $p[s | (\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N)]$ . Tal notação será evitada por razões de simplicidade.

Denotamos por  $I(A)$  a função indicadora que assume o valor 1 quando o evento  $A$  ocorre e 0 caso contrário. Seja  $\mathbf{\Delta}_s = [I(1 \in s), \dots, I(N \in s)]'$  um vetor aleatório de indicadores dos elementos incluídos na amostra  $s$ . Então o plano amostral pode ser alternativamente caracterizado pela distribuição de probabilidade de  $\mathbf{\Delta}_s$  denotada por  $f[\mathbf{\delta}_s | (\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N)]$ , onde  $\mathbf{\delta}_s$  é qualquer realização particular de  $\mathbf{\Delta}_s$  tal que  $\mathbf{\delta}_s' \mathbf{1}_N = n$ , e  $\mathbf{1}_N$  é o vetor unitário de dimensão  $N$ .

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

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

e denotamos por  $\pi_{ij}$  a probabilidade de inclusão conjunta das unidades  $i$  e  $j$ , dada por

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

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  $\pi_{ij}$  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 satisfazem 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 esse indicador uma das variáveis de pesquisa, que é considerada no mecanismo de seleção da amostra. Os métodos que descreveremos 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

$$f[\boldsymbol{\delta}_s | (\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_N, \mathbf{x}_N)] = f(\boldsymbol{\delta}_s | \mathbf{x}_1, \dots, \mathbf{x}_N). \quad (2.7)$$

Entre os planos amostrais *não-informativos*, ainda precisamos ainda distinguir duas outras situações de interesse. Quando o plano amostral é amo-

stragem 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 na Seção 2.1.1 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 2.1 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 Sugden e Smith, 1984). 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 idéia 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.

**Exemplo 2.1** *Amostragem com probabilidades proporcionais ao tamanho de população bivariada normal (Pfeffermann, 1993)*

Vamos considerar as  $N$  observações da população  $(y_i; x_i)$  como determinações de vetores aleatórios IID com distribuição  $\mathbf{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$ . Seja  $s = \{(y_i, x_i) : i = 1, \dots, n\}$  uma amostra de  $n$  unidades selecionada por esquema probabilístico. Deseja-se estimar  $\mu_Y = E_M(Y)$ . No caso de amostragem aleatória simples com reposição,  $\bar{y} = \sum_{i=1}^n y_i/n$  é um estimador não viciado de  $\mu_Y$  que tem propriedades ótimas sob o modelo. Neste caso, podemos ignorar o esquema de seleção da amostra no processo de inferência sem qualquer problema.

Vamos supor agora que a amostra seja selecionada com probabilidade proporcional a  $x_i$  com reposição (note que a regra de seleção depende apenas da variável auxiliar  $x$  e não da variável de pesquisa  $y$ ). Isto pode ser conseguido mediante um esquema de seleção em que  $n$  sorteios independentes são realizados, e em cada sorteio a probabilidade de escolher a unidade  $i$  da população para a amostra é dada por  $p_i = x_i / \sum_{i=1}^N x_i$ .

Neste caso, se  $CORR_M(Y; X) = \sigma_{yx}/\sigma_y\sigma_x > 0$  então  $P(Y_i > \mu_Y \mid i \in s) > 1/2$  e portanto  $E_M(\bar{y}) > \mu_Y$ , mostrando que ignorar o esquema amostral torna viciado o procedimento de inferência baseado no estimador usual de média amostral.

Supondo que os  $x_i$  são conhecidos para todas as unidades da população, podemos usar o estimador

$$\bar{y}_{reg} = \bar{y} + b(\bar{X} - \bar{x})$$

onde  $b$  é o estimador usual de mínimos quadrados do coeficiente de regressão  $\beta = \sigma_{yx}/\sigma_y^2$  e  $\bar{x}$  e  $\bar{X}$  são as médias amostral e populacional da variável auxiliar  $x$ , respectivamente. O estimador  $\bar{y}_{reg}$  é não viciado para  $\mu_Y$  sob o modelo, e tem propriedades razoáveis em termos da distribuição de aleatorização para grandes amostras.

Este exemplo ilustrou que um determinado procedimento de inferência clássica pode não funcionar adequadamente na presença de um plano amostral que, de alguma forma, interfira com a estrutura estocástica da amostra, para a qual o modelo assumido na população passa a não se adequar.



## 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 Särndal, Swensson e Wretman(1992, seção 2.8).

Consideremos o problema de estimar o vetor  $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$  de totais das  $R$  variáveis da pesquisa na população, a partir de uma amostra observada  $s$ . Naturalmente, qualquer estimador viável do total  $\mathbf{Y}$  só pode depender dos valores das variáveis de pesquisa observados na amostra, contidos em  $\mathbf{y}_{i_1}, \dots, \mathbf{y}_{i_n}$ , 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  $\pi$ -ponderado* (veja Särndal, Swensson e Wretman, 1992, p.42), dado por

$$\hat{\mathbf{Y}}_{\pi} = \sum_{i \in s} \mathbf{y}_i / \pi_i . \quad (3.1)$$

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

*aleatorização e variância de aleatorização.*

O estimador  $\pi$ -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

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

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

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

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

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

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

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

desde que  $\pi_{ij} > 0 \quad \forall i, j \in U$ , como suposto neste livro (Seção 2.4).

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

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

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 planejamento 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

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

e

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

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

$$\hat{\mathbf{Y}}_\pi = \frac{N}{n} \sum_{i \in s} \mathbf{y}_i = N\bar{\mathbf{y}} , \quad (3.7)$$

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

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

onde  $f = n/N$  é a fração amostral e

$$\bar{\mathbf{y}} = n^{-1} \sum_{i \in s} \mathbf{y}_i , \quad (3.10)$$

$$\mathbf{S}_y = N^{-1} \sum_{i \in U} (\mathbf{y}_i - \bar{\mathbf{Y}}) (\mathbf{y}_i - \bar{\mathbf{Y}})^0 , \quad (3.11)$$

$$\bar{\mathbf{Y}} = N^{-1} \sum_{i \in U} \mathbf{y}_i = N^{-1} \mathbf{Y} , \quad (3.12)$$

$$\hat{\mathbf{S}}_y = n^{-1} \sum_{i \in s} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})^0 . \quad (3.13)$$

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

$$\hat{\mathbf{Y}}_w = \sum_{i \in s} w_i \mathbf{y}_i \quad (3.14)$$

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

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

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

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

e

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

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  $\hat{\mathbf{Y}}_w = \sum_{i \in s} w_i \mathbf{y}_i$  com pesos  $w_i$  dados respectivamente por

$$w_i^R = \frac{\pi_i^{-1} \sum_{k \in U} x_k}{\sum_{k \in s} \pi_k^{-1} x_k} = \frac{\pi_i^{-1} X}{\hat{X}_\pi} \quad (3.17)$$

e

$$w_i^{REG} = \pi_i^{-1} g_i, \quad (3.18)$$

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

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 Nascimento 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, Swensson e Wretman (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

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

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 (3.4) 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 idéia 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 10).

### 3.3 Linearização de Taylor para Estimar Variâncias

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

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

onde  $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i = (Y_1, \dots, Y_R)^0$  é um vetor de totais de  $R$  variáveis de pesquisa.

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

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

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

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

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

Assumindo então que  $\boldsymbol{\theta}$  é da forma

$$\boldsymbol{\theta} = \mathbf{A}\mathbf{Y} ,$$

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

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

Este estimador é não-viciado e tem variância de aleatorização

$$V_p \left( \hat{\boldsymbol{\theta}} \right) = \mathbf{A} V_p \left( \hat{\mathbf{Y}}_\pi \right) \mathbf{A}^0 ,$$

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

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  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\mathbf{Y}}_\pi)$ . Para maiores detalhes sobre esse método, veja por exemplo Särndal, Swensson e Wretman(1992, p. 172), Wolter(1985, p. 221) ou Bishop, Fienberg e Holland (1977, p. 486).

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

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

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  $r = 1, \dots, R$ .

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  $\mathbf{\Delta g}(\mathbf{Y}) (\hat{\mathbf{Y}}_\pi - \mathbf{Y})$  é uma função linear de  $\hat{\mathbf{Y}}_\pi$ , segue imediatamente que

$$V_p(\hat{\boldsymbol{\theta}}) \simeq \mathbf{\Delta g}(\mathbf{Y}) V_p(\hat{\mathbf{Y}}_\pi) \mathbf{\Delta g}(\mathbf{Y})^0 \quad (3.20)$$

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

$$\hat{V}_p(\hat{\boldsymbol{\theta}}) = \mathbf{\Delta g}(\hat{\mathbf{Y}}_\pi) \hat{V}_p(\hat{\mathbf{Y}}_\pi) \mathbf{\Delta g}(\hat{\mathbf{Y}}_\pi)^0, \quad (3.21)$$

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

### 3.3. LINEARIZAÇÃO DE TAYLOR PARA ESTIMAR VARIÂNCIAS 41

#### 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  $\mathbf{Y} = (Y_1, \dots, Y_u)^0$  e  $\mathbf{X} = (X_1, \dots, X_u)^0$  vetores de totais e consideremos o vetor de razões  $\mathbf{R} = \left( \frac{Y_1}{X_1}, \dots, \frac{Y_u}{X_u} \right)$ . Conhecendo estimativas das matrizes  $V_p(\hat{\mathbf{Y}}_\pi)$ ,  $V_p(\hat{\mathbf{X}}_\pi)$  e  $COV_p(\hat{\mathbf{Y}}_\pi; \hat{\mathbf{X}}_\pi)$ , queremos calcular a matriz de variância de

$$\hat{\mathbf{R}} = \left( \frac{\hat{Y}_{1\pi}}{\hat{X}_{1\pi}}, \dots, \frac{\hat{Y}_{u\pi}}{\hat{X}_{u\pi}} \right)^0.$$

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

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

onde  $\mathbf{y} = (y_1, \dots, y_u)^0$  e  $\mathbf{x} = (x_1, \dots, x_u)^0$ . A matriz jacobiana de  $\mathbf{g}(\mathbf{y}, \mathbf{x})$  é a matriz  $u \times 2u$  dada por

$$\Delta \mathbf{g}(\mathbf{y}, \mathbf{x}) = \left[ \text{diag} \left( \frac{1}{x_1}, \dots, \frac{1}{x_u} \right) \quad \text{diag} \left( -\frac{y_1}{x_1^2}, \dots, -\frac{y_u}{x_u^2} \right) \right].$$

Seja  $\mathbf{D}_x = \text{diag}(x_1, \dots, x_u)$  a matriz diagonal de dimensão  $u \times u$  formada a partir do vetor  $\mathbf{x} = (x_1, \dots, x_u)^0$ . Usando essa notação, podemos escrever o vetor  $\hat{\mathbf{R}}$  de estimadores das razões como

$$\hat{\mathbf{R}} = \left( \frac{\hat{Y}_{1\pi}}{\hat{X}_{1\pi}}, \dots, \frac{\hat{Y}_{u\pi}}{\hat{X}_{u\pi}} \right)^0 = \mathbf{g}(\hat{\mathbf{Y}}_\pi, \hat{\mathbf{X}}_\pi)$$

e a correspondente matriz jacobiana como

$$\Delta \mathbf{g}(\hat{\mathbf{Y}}_\pi, \hat{\mathbf{X}}_\pi) = \left[ \mathbf{D}_{\hat{\mathbf{R}}} \mathbf{D}_{\hat{\mathbf{Y}}_\pi}^{-1} \quad -\mathbf{D}_{\hat{\mathbf{R}}} \mathbf{D}_{\hat{\mathbf{X}}_\pi}^{-1} \right].$$

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

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

Efetutando os produtos das matrizes em blocos obtemos

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

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 idéia 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, Hurwitz e Madow(1953), é considerar apenas a variação entre informações disponíveis a nível das unidades primárias de amostragem (UPAs), isto é, a nível dos conglomerados primários, e admitir que estes teriam sido selecionadas com reposição da população. Esta idéia 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 é 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  $h$ ,  $h = 1, \dots, H$ . Denotando por  $\pi_{hi}$  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  $h$ ,  $h = 1, \dots, H$ . 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

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

e um estimador não viciado da variância de aleatorização correspondente por

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

onde  $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / \pi_{hi}$  para  $h = 1, \dots, H$ . (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, Swensson e Wretman(1992, p. 153).

### 3.5 Métodos de Replicação

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

A idéia 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  $\theta$  é o parâmetro-alvo, e  $\hat{\theta}_g$  é um estimador não viciado de  $\theta$  baseado

na  $g$ -ésima replicação ( $g = 1, \dots, G$ ), segue-se que

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

é um estimador não viciado de  $\theta$  e

$$\hat{V}_R(\hat{\theta}_R) = \frac{1}{G(G-1)} \sum_{g=1}^G (\hat{\theta}_g - \hat{\theta}_R)^2 \quad (3.24)$$

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

Note que desde que as replicações sejam construídas de forma independente conforme indicado, os estimadores  $\hat{\theta}_R$  e  $\hat{V}_R(\hat{\theta}_R)$  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 idéia é aplicada exatamente é a pesquisa de preços para formar o Índice de Preços ao Consumidor (do inglês *Consumer Price Index - CPI*) do *US Bureau of 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 idéia 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  $\theta$  baseado na amostra completa (denotado  $\hat{\theta}$ ), e um estimador de variância mais conservador que o estimador  $\hat{V}_R(\hat{\theta}_R)$  anteriormente apresentado, dado por

$$\hat{V}_{RG}(\hat{\theta}) = \frac{1}{G(G-1)} \sum_{g=1}^G (\hat{\theta}_g - \hat{\theta})^2. \quad (3.25)$$

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, 1956) como uma técnica para redução de vício de estimadores, num contexto da Estatística Clássica. A idéia 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

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

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

$$\hat{V}_{J1}(\hat{\theta}) = \frac{1}{G(G-1)} \sum_{g=1}^G (\hat{\theta}_{(g)} - \hat{\theta}_J)^2 \quad (3.26)$$

ou

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

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

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

**Observação 3.2** *O estimador  $\widehat{V}_{J2}(\widehat{\theta})$  é mais conservador que o estimador  $\widehat{V}_{J1}(\widehat{\theta})$ .*

**Observação 3.3** *É comum aplicar a técnica fazendo o número de grupos igual ao tamanho da amostra, isto é, tomando  $G = n$  e portanto eliminando uma observação da amostra de cada vez ao calcular os pseudo-valores. Essa regra deve ser aplicada considerando o número de 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*).

## Capítulo 4

# Efeitos do Plano Amostral

### 4.1 Introdução

O cálculo de desvio padrão e o uso de testes de hipóteses desempenham papel fundamental em estudos analíticos. Além de estimativas pontuais, na inferência analítica é necessário transmitir a idéia de precisão associada a essas estimativas e construir intervalos de confiança associados. Valores de desvios padrões, ou alternativamente comprimentos de intervalos de confiança, permitem avaliar a precisão da estimação. O cálculo do desvio padrão também possibilita a construção de estatísticas para testar hipóteses relativas a parâmetros do modelo (tradição de modelagem) ou de parâmetros da população finita (tradição de amostragem). Testes de hipóteses são também usados na fase de seleção de modelos.

Os pacotes mais comuns de análise estatística incluem em suas saídas valores de estimativas pontuais e seus desvios padrões, além de  $p$ valores relativos a hipóteses de interesse. Contudo, as fórmulas usadas nestes pacotes para o cálculo dos desvios padrões e obtenção de testes são, em geral, baseadas nas hipóteses de independência e de igualdade de distribuição (IID) das observações, ou equivalentemente, de amostragem aleatória simples com reposição (AASC). Tais hipóteses quase nunca valem para dados obtidos através de pesquisas por amostragem, como as que realizam o IBGE e outras agências produtoras de estatísticas.

Este capítulo trata de avaliar o impacto sobre desvios padrões, intervalos de confiança e níveis de significância de testes usuais quando há afastamentos das hipóteses IID mencionadas, devidos ao uso de planos amostrais complexos para obter os dados. Como veremos, o impacto pode ser muito grande em algumas situações, justificando os cuidados que devem ser tomados na

análise de dados deste tipo. Neste capítulo, usaremos como referência básica Skinner(1989a).

## 4.2 Efeito do Plano Amostral (EPA) de Kish

Para medir o efeito do plano amostral sobre a variância de um estimador, Kish(1965) propôs uma medida que denominou *Efeito do Plano Amostral (EPA)* (em inglês, *design effect* ou, abreviadamente, *deff*). O objetivo desta medida é comparar planos amostrais no estágio de planejamento da pesquisa. O EPA de Kish é uma razão entre variâncias (de aleatorização) de um estimador, calculadas para dois planos amostrais alternativos. Vamos considerar um estimador  $\hat{\theta}$  e calcular a variância de sua distribuição induzida pelo plano amostral complexo (verdadeiro)  $V_{VERD}(\hat{\theta})$  e a variância da distribuição do estimador induzida pelo plano de amostragem aleatória simples  $V_{AAS}(\hat{\theta})$ .

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

$$\mathbf{EPA}_{Kish}(\hat{\theta}) = \frac{V_{VERD}(\hat{\theta})}{V_{AAS}(\hat{\theta})}. \quad (4.1)$$

Para ilustrar o conceito do  $\mathbf{EPA}_{Kish}(\hat{\theta})$ , vamos considerar um exemplo.

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

Nascimento Silva e Moura(1990) estimaram o  $\mathbf{EPA}_{Kish}$  para estimadores de totais de várias variáveis sócio-econômicas a nível das Regiões Metropolitanas (RMs) utilizando dados do questionário de amostra do Censo Demográfico de 1980. Essas medidas estimadas do efeito do plano amostral foram calculadas para três esquemas amostrais alternativos, todos considerando amostragem conglomerada de domicílios em dois estágios, tendo o setor censitário como unidade primária e o domicílio como unidade secundária de amostragem. Duas das alternativas consideraram seleção de setores com equiprobabilidade via amostragem aleatória simples sem reposição (AC2AAS) e fração amostral constante de domicílios no segundo estágio (uma usando o estimador simples ou  $\pi$ -ponderado do total, e outra usando

o estimador de razão para o total calibrando no número total de domicílios da população), e uma terceira alternativa considerou a seleção de setores com probabilidades proporcionais ao tamanho (número de domicílios por setor), denominada AC2PPT, e a seleção de 15 domicílios em cada setor da amostra, e empregando o correspondente estimador  $\pi$ -ponderado. Os resultados para algumas variáveis a nível da Região Metropolitana do Rio de Janeiro são apresentados na Tabela 4.1 a título de ilustração. Note que a população alvo considera apenas moradores em domicílios particulares permanentes na Região Metropolitana do Rio de Janeiro.

Os valores apresentados na Tabela 4.1 para a RM do Rio de Janeiro são similares aos observados para as demais RMs, se consideradas as mesmas variáveis. Nota-se grande variação dos valores do EPA, cujos valores mínimo e máximo são de 1,28 e 111,27 respectivamente. Para algumas variáveis (1,2,4,5 e 9), o EPA varia consideravelmente entre as diferentes alternativas de plano amostral, enquanto para outras variáveis (3,6,7 e 8) as variações entre os planos amostrais é mínima.

Os valores elevados do EPA observados para algumas variáveis realçam a importância de considerar o plano amostral verdadeiro ao estimar variâncias e desvios padrões associados às estimativas pontuais. Isso ocorre porque estimativas ingênuas de variância baseadas na hipótese de AAS subestimam substancialmente as variâncias corretas.

Outra regularidade encontrada nesse valores é que o EPA para o plano amostral AC2AAS com estimador simples apresenta sempre os valores mais elevados, revelando que este esquema é menos eficiente que os competidores considerados. Em geral, o EPA é menor para o esquema AC2PPT, com valores próximos aos do esquema AC2AAS com estimador de razão.

Os valores dos EPAs calculados por Nascimento Silva e Moura(1990) podem ser usados para planejar pesquisas amostrais (ao menos nas regiões metropolitanas), pois permitem comparar e antecipar o impacto do uso de alguns esquemas amostrais alternativos sobre a precisão de estimadores de totais de várias variáveis relevantes. Permitem também calcular tamanhos amostrais para garantir determinado nível de precisão, sem emprego de fórmulas complicadas. Portanto, tais valores seriam úteis como informação de apoio ao planejamento de novas pesquisas por amostragem, antes que as respectivas amostras sejam efetivamente selecionadas.

Entretanto, esses valores têm pouca utilidade em termos de usos analíticos dos dados da amostra do Censo Demográfico 80. É que tais valores, embora tendo sido estimados com essa amostra, foram calculados para planos

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

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

amostrais distintos do que foi efetivamente adotado para seleção da amostra do censo. A amostra de domicílios usada no censo é estratificada por setor censitário com seleção sistemática de uma fração fixa (25% no Censo 80) dos domicílios de cada setor. Já os planos amostrais considerados na tabulação dos EPAs eram planos amostrais em dois estágios, com seleção de setores no primeiro estágio, os quais foram considerados por sua similaridade com os esquemas adotados nas principais pesquisas domiciliares do IBGE tais como a PNAD e a PME (Pesquisa Mensal de Emprego). Portanto, a utilidade maior dos valores tabulados dos EPAs seria a comparação de planos amostrais alternativos para planejamento de pesquisas futuras, e não a análise dos resultados da amostra do censo 80.

### 4.3 Efeito do Plano Amostral Ampliado

O que se observou no Exemplo 4.1 com respeito à dificuldade de uso dos EPAs de Kish calculados para fins analíticos também se aplica para outras situações e é uma deficiência estrutural do conceito de EPA proposto por Kish. Para tentar contornar essa dificuldade, é necessário considerar um conceito ampliado de EPA, correspondente ao conceito de *misspecification effect (meff)* proposto por Skinner, Holt e Smith (1989, p. 24), que apresentamos e discutimos nesta seção.

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

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

compará-la a  $V_{VERD}(\hat{\theta})$ .

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

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

$$\mathbf{EPA}(\hat{\theta}, v_0) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)}. \quad (4.2)$$

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

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

Em geral, são esperadas as seguintes conseqüências sobre o  $\mathbf{EPA}$  ao ignorar o plano amostral efetivamente adotado e admitir que a seleção da amostra foi AAS:

1. Ignorar os pesos em  $v_0$  pode inflacionar o  $\mathbf{EPA}$ ;
2. Ignorar conglomeração em  $v_0$  pode inflacionar o  $\mathbf{EPA}$ ;
3. Ignorar estratificação em  $v_0$  pode reduzir o  $\mathbf{EPA}$ .

Combinações destes aspectos num mesmo plano amostral, resultando na especificação incorreta do plano amostral subjacente a  $v_0$ , podem inflacionar ou reduzir o  $\mathbf{EPA}$ . Nesses casos é difícil prever o impacto de ignorar o plano amostral (ou modelo) verdadeiro sobre a análise baseada em hipóteses IID. Por essa razão, é recomendável ao menos estimar os EPAs antes de concluir a análise padrão, para poder então avaliar se há impactos importantes a considerar.

Tabela 4.2: Definição da estratificação da população de empresas

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

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

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

- 1) pessoal ocupado em 31/12/94 (PO);
- 2) total de salários pagos no ano de 94 (SAL);
- 3) receita total no ano de 94 (REC).

A idéia é considerar o problema de estimar as médias populacionais das variáveis SAL e REC (variáveis de pesquisa, nesse exemplo), usando amostras estratificadas simples com alocação desproporcional, implicando em unidades amostrais com pesos desiguais numa situação bastante simples. A variável PO é a variável de estratificação. As médias populacionais das variáveis de pesquisa (SAL e REC) são conhecidas, porém supostas desconhecidas para efeitos do presente exercício, em que se supõe que amostragem seria usada para sua estimação.

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

Foram então selecionadas de cada um dos estratos amostras aleatórias simples sem reposição de 30 empresas, implicando em uso de alocação igual e em frações amostrais desiguais, em vista dos diferentes tamanhos populacionais dos estratos. Como o estrato 1 contém cerca de 21% das observações da população, a proporção de 50% das observações da amostra no estrato 1 (das maiores empresas) na amostra é bem maior do que seria esperado sob amostragem aleatória simples da população em geral. Desta forma, a média amostral de uma variável de pesquisa  $y$  qualquer (SAL ou REC) dada por

$$\bar{y} = \frac{1}{n} \sum_{h=1}^2 \sum_{i \in s_h} y_{hi}$$

tenderia a superestimar a média  $\bar{Y}$  da população dada por  $\bar{Y} = \frac{1}{N} \sum_{h=1}^2 \sum_{i \in U_h} y_{hi}$ , onde  $y_{hi}$  é o valor da variável de pesquisa  $y$  para a  $i$ -ésima observação do estrato  $h$  ( $h = 1, 2$ ). Neste caso, um estimador não-viciado da média populacional  $\bar{Y}$  seria dado por

$$\bar{y}_w = \sum_{h=1}^2 W_h \bar{y}_h$$

onde  $W_h = \frac{N_h}{N}$  é a proporção de observações da população no estrato  $h$  e  $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$  é a média amostral dos  $y$ 's no estrato  $h$  ( $h = 1, 2$ ).

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

No presente exercício, a variância do estimador  $\bar{y}_w$  pode ser obtida de duas formas: calculando a expressão da variância utilizando os dados de todas as unidades da população (que são conhecidos, mas admitidos desconhecidos para fins do exercício de estimação de médias via amostragem) e por simulação.

A variância de  $\bar{y}_w$  sob a distribuição de aleatorização verdadeira é dada por

$$V_p(\bar{y}_w) = \sum_{h=1}^2 W_h^2 (1 - f_h) \frac{S_h^2}{n_h} \quad (4.3)$$

onde  $f_h = n_h/N_h$ ,  $n_h$  é o número de observações na amostra no estrato  $h$ , e  $S_h^2 = \frac{1}{N_h-1} \sum_{i \in U_h} (y_{hi} - \bar{Y}_h)^2$  é a variância populacional da variável de pesquisa  $y$  dentro do estrato  $h$ , com  $\bar{Y}_h = \frac{1}{N_h-1} \sum_{i \in U_h} y_{hi}$  representando a média populacional da variável  $y$  dentro do estrato  $h$ .

Um estimador usual da variância de  $\bar{y}_w$  sob amostragem aleatória simples é  $v_0 = (1 - f) \frac{s^2}{n}$  onde  $s^2 = \frac{1}{n-1} \sum_{h=1}^2 \sum_{i \in s_h} (y_{hi} - \bar{y})^2$  e  $f = \sum_{h=1}^2 n_h / \sum_{h=1}^2 N_h = n/N$ .

O cálculo do **EPA** foi feito também por meio de simulação. Geramos 500 amostras de tamanho 60, segundo o plano amostral estratificado consi-

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

Quantidade de interesse	Salários	Receitas
1) Média populacional $\bar{Y}$	78,328	2,107
2) Média de $\bar{y}$ sobre 500 amostras	160,750	4,191
3) Média de $\bar{y}_w$ sobre 500 amostras	76,700	2,054

derado. Para cada uma das 500 amostras e cada uma das duas variáveis de pesquisa (SAL e REC) foram calculados:

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

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

Em contraste com os valores dos parâmetros populacionais, calculamos a média das médias amostrais não ponderadas ( $\bar{y}$ ) dos salários e das receitas obtidas nas 500 amostras simuladas, obtendo os valores na segunda linha da Tabela 4.3. Como previsto, observamos um vício para cima na estimativa destas médias, da ordem de 105% para os salários e de 98,9% para as receitas.

Usamos também o estimador  $\bar{y}_w$  para estimar a média dos salários e das receitas na população, obtendo para esse estimador as médias apresentadas na terceira linha da Tabela 4.3. Observamos ainda um pequeno vício da ordem de -1,95% e -2,51% para os salários e receitas, respectivamente. Note que o estimador  $\bar{y}_w$  é não-viciado sob o plano amostral adotado, entretanto o pequeno vício observado na simulação não pode ser ignorado pois é

Tabela 4.4: Propriedades dos estimadores de variância do estimador  $\bar{y}_w$ 

Quantidade de interesse	Salários	Receitas
1) Variância populacional $V_{AES}(\bar{y}_w)$	244,18	0,43500
2) Média de $\hat{V}_{AES}(\bar{y}_w)$ usando 500 amostras	231,84	0,32569
3) Valor esperado de $v_0$ usando população	1.613,3	1,1880
4) Média de $v_0$ usando 500 amostras	1.636,1	1,2121

significativamente diferente de 0 ao nível de significância de 5%, apesar do tamanho razoável da simulação (500 replicações).

Além dos estimadores pontuais, o interesse maior da simulação foi comparar valores de estimadores de variância, e conseqüentemente de medidas do efeito do plano amostral. Como o estimador pontual dado pela média amostral não ponderada ( $\bar{y}$ ) é grosseiramente viciado, não consideramos estimativas de variância para esse estimador, mas tão somente para o estimador não-viciado dado pela média ponderada  $\bar{y}_w$ . Para esse último, consideramos dois estimadores de variância, a saber o estimador ingênuo sob a hipótese de AAS (dado por  $v_0$ ) e um estimador não viciado da variância sob o plano amostral  $\hat{V}_{AES}(\bar{y}_w)$ , que foi obtido substituindo as variâncias dentro dos estratos  $S_h^2$  por estimativas amostrais não viciadas dadas por  $s_h^2 = \frac{1}{n_h} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ ,  $h = 1, 2$ , na fórmula de  $V_{AES}(\bar{y}_w)$  conforme definida em (4.3).

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

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

nas estimativas de simulação como nos valores populacionais das variâncias, cujos cálculos estão ilustrados a seguir:

$$\begin{aligned} \mathbf{EPA}(\overline{SAL}_w, v_0(\overline{SAL}_w)) &= \frac{231,84}{1.636,1} = 0,142 \\ \mathbf{EPA}(\overline{REC}_w, v_0(\overline{REC}_w)) &= \frac{0,32569}{1,2121} = 0,269 \\ \mathbf{EPA}(\overline{SAL}_w, v_0(\overline{SAL}_w)) &= \frac{244,18}{1.613,3} = 0,151 \quad \text{e} \\ \mathbf{EPA}(\overline{REC}_w, v_0(\overline{REC}_w)) &= \frac{0,43500}{1,1880} = 0,366. \end{aligned}$$

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

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

Essas conclusões são largamente aceitas pelos amostristas e produtores de dados baseados em pesquisas amostrais para o caso da estimação de médias e totais, e respectivas variâncias. Entretanto ainda há exemplos de usos indevidos de dados amostrais nos quais os pesos são ignorados, em particular para a estimação de variâncias associadas a estimativas pontuais de médias e totais. Tal situação se deve ao uso ingênuo de pacotes estatísticos padrões desenvolvidos para analisar amostras IID, sem a devida consideração dos pesos e plano amostral.

*Observação 4.1 Neste exemplo não foi feito uso analítico dos dados e sim descritivo, onde é usual incorporar os pesos no cálculo de estimativas e variâncias.*

Tabela 4.5: Valores dos Efeitos de Plano Amostral (EPA) para as médias de Salário e de Receita

Variável	Estimativa	Simulação	População
Salário	Variância	231,84	244,18
	<i>EPA</i>	0,142	0,151
Receita	Variância	0,32569	0,43500
	<i>EPA</i>	0,269	0,366

Observação 4.2 *Não seria esperado usar um estimador ponderado para a média e não considerar os pesos no cálculo de variâncias, como fizemos neste exemplo.*

Observação 4.3 *O exemplo mostra que ignorar a estratificação ao calcular  $v_0$  diminui o EPA.*

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

Exemplo 4.3 *População conglomerada com conglomerados de tamanho 2 (Skinner, Holt e Smith, 1989, p. 25)*

Considere uma população de conglomerados de tamanho 2, isto é, onde as unidades (elementares ou de referência) estão grupadas em pares (exemplos de tais populações incluem pares de irmãos gêmeos, casais, jogadores numa dupla de vôlei de praia ou tênis, etc.). Suponha que os valores de uma variável de pesquisa medida nessas unidades têm média  $\theta$  e variância  $\sigma^2$ , além de uma correlação  $\rho$  entre os valores dentro de cada par (correlação intraclasse, veja Nascimento Silva e Moura, 1990, cap. 2 e Haggard, 1958). Suponha que um único par é sorteado ao acaso da população e que os valores  $y_1$  e  $y_2$  são observados para as duas unidades do par selecionado. O modelo assumido pode então ser representado como

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

Um estimador não viciado para  $\theta$  é dado por  $\hat{\theta} = (y_1 + y_2)/2$ , a média amostral. Assumindo a (falsa) hipótese de que o esquema amostral é AASC

de unidades individuais e não de pares, ou equivalentemente, que  $y_1$  e  $y_2$  são observações de variáveis aleatórias IID, a variância de  $\hat{\theta}$  é dada por

$$V_{AAS}(\hat{\theta}) = \sigma^2/2$$

com um estimador não viciado dado por

$$v_0(\hat{\theta}) = (y_1 - y_2)^2/4.$$

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

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

e o valor esperado do estimador de variância  $v_0(\hat{\theta})$  é dado por

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

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

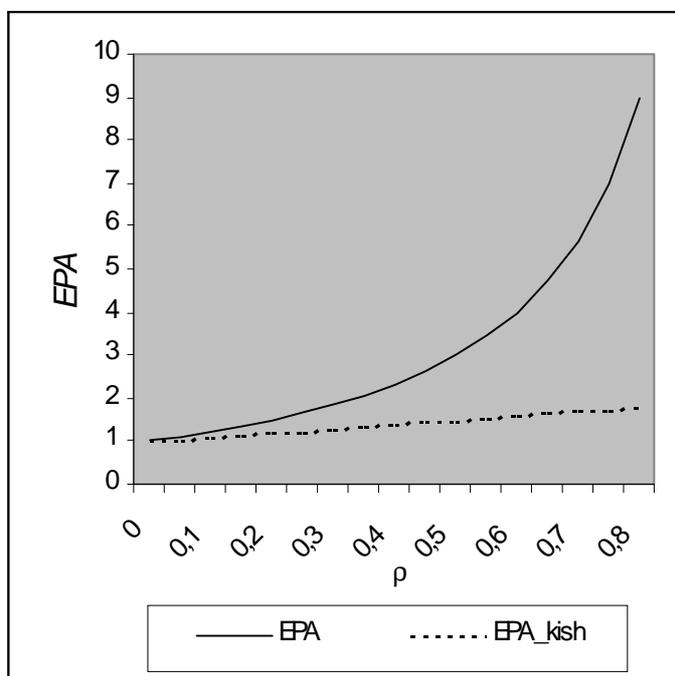
$$\mathbf{EPA}_{Kish}(\hat{\theta}) = 1 + \rho$$

e

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

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

Este exemplo ilustra bem dois aspectos distintos do uso de medidas como o efeito de plano amostral. O primeiro é que as duas medidas são distintas, *embora os respectivos estimadores baseados numa particular amostra coincidam*. No caso particular deste exemplo, o  $\mathbf{EPA}_{Kish}(\hat{\theta})$  cresce pouco com o valor do coeficiente de correlação intraclasse  $\rho$ , o que implica que um plano amostral conglomerado como o adotado (seleção ao acaso de um par da população) seria menos eficiente que um plano amostral aleatório simples (seleção de duas unidades ao acaso da população), mas a perda de eficiência

Figura 4.1: Valores de  $\mathbf{EPA}(\hat{\theta}, v_0)$  e  $\mathbf{EPA}_{Kish}(\hat{\theta})$  para conglomeração

seria modesta. Já se o interesse é medir, a posteriori, o efeito da má especificação do plano amostral no estimador de variância, o impacto, medido pelo **EPA**  $(\hat{\theta}, v_0)$ , seria muito maior.

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

## 4.4 Intervalos de Confiança e Testes de Hipóteses

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

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

que, sob a hipótese de que as observações são IID, muitas vezes é  $N(0; 1)$ .

Neste caso, um intervalo de confiança de nível de confiança aproximado  $(1 - \alpha)$  é dado por  $[\hat{\theta} - z_{\alpha/2} v_0^{1/2}, \hat{\theta} + z_{\alpha/2} v_0^{1/2}]$ , onde  $z_\alpha$  é definido por  $\int_{z_\alpha}^{+\infty} \varphi(t) dt = \alpha$ , onde  $\varphi$  é a função de densidade da distribuição normal padrão.

Vamos analisar o efeito de um plano amostral complexo sobre o intervalo de confiança. No caso de um plano amostral complexo, a distribuição que é aproximadamente normal é a de

$$\frac{\hat{\theta} - \theta}{[\hat{V}_{VERD}(\hat{\theta})]^{1/2}} .$$

Por outro lado, para obter a variância da distribuição assintótica de  $t_0$  note que

$$\frac{\hat{\theta} - \theta}{v_0^{1/2}} = \frac{\hat{\theta} - \theta}{[\hat{V}_{VERD}(\hat{\theta})]^{1/2}} \times \frac{[\hat{V}_{VERD}(\hat{\theta})]^{1/2}}{v_0^{1/2}} .$$

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

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

Como o primeiro fator tende para uma  $N(0; 1)$ , a variância assintótica de  $t_0$  é aproximadamente igual ao quadrado do segundo fator, isto é, a  $\frac{\widehat{V}_{VERD}(\hat{\theta})}{v_0}$  que é um estimador para  $\mathbf{EPA}(\hat{\theta}, v_0)$ . Porém quando a amostra é grande esse valor aproxima o  $\mathbf{EPA}(\hat{\theta}, v_0) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)}$ , pois  $v_0$  é aproximadamente igual a  $E_{VERD}(v_0)$  e  $\widehat{V}_{VERD}(\hat{\theta})$  é aproximadamente igual a  $V_{VERD}(\hat{\theta})$ . Logo temos que a distribuição assintótica verdadeira de  $t_0$  é dada por

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

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

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

onde  $\Phi$  é a função de distribuição acumulada de uma  $N(0; 1)$ . Calculamos esta probabilidade para alguns valores do  $\mathbf{EPA}$ , que apresentamos na Tabela 4.6.

À medida que o valor do **EPA**  $(\hat{\theta}, v_0)$  aumenta, a probabilidade real de cobertura diminui, sendo menor que o valor nominal para valores de **EPA**  $(\hat{\theta}, v_0)$  maiores que 1.

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

#### Exemplo 4.4 *Teste de hipótese sobre proporção*

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

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

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

$$Z = |\hat{p} - p_0| / \sqrt{p_0(1 - p_0)/n}, \quad (4.4)$$

onde o denominador é o desvio padrão da estimativa  $\hat{p}$  sob a hipótese nula.

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

Para o teste baseado na amostragem aleatória simples, ignoramos a conglomeração e usamos na fórmula do desvio padrão o tamanho total da amostra de unidades elementares (empregados), isto é,  $n = 1.000$ . O valor da estatística de teste  $Z$  definida em (4.4) é, portanto,

$$Z_{bin} = |0,84 - 0,8| / \sqrt{0,8(1 - 0,8) / 1.000} = 3,162 > Z_{0,025} = 1,96 \quad (4.5)$$

onde  $\sqrt{0,8(1 - 0,8) / 1.000} = 0,0126$  é o desvio padrão de  $\hat{p}$  sob a hipótese nula. Este resultado sugere a rejeição da hipótese  $H_0$ .

Por outro lado, é razoável admitir que se uma empresa for coberta por plano de saúde, cada empregado dessa empresa terá acesso ao plano. Essa é uma informação importante que foi ignorada no teste anterior. De fato, selecionar mais de uma pessoa numa empresa não aumenta nosso conhecimento sobre a cobertura por plano de saúde no local. Portanto, o *tamanho efetivo* da amostra é  $\bar{n} = 50$ , em contraste com o valor 1.000 usado no teste anterior. O termo *tamanho efetivo* foi introduzido em Kish(1965) para designar o tamanho de uma amostra aleatória simples necessário para estimar  $p$  com a mesma precisão obtida por uma amostra conglomerada de tamanho  $n$  (neste caso, igual a 1.000) unidades elementares.

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

$$Z_p = |\hat{p} - p_0| / \sqrt{p_0(1 - p_0) / 50} = 0,707,$$

onde o valor  $\sqrt{0,8(1 - 0,8) / 50} = 0,0566$  é muito maior que o valor do desvio padrão obtido no teste anterior. Portanto, o valor observado de  $Z_p$  é menor que o de  $Z_{bin}$ , e o novo teste sugere a não rejeição da mesma hipótese nula.

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

## 4.5 Efeitos Multivariados de Plano Amostral

O conceito de efeito de plano amostral introduzido em (4.2) é relativo a inferências sobre um parâmetro univariado  $\theta$ . Consideremos agora o problema de

estimação de um vetor  $\boldsymbol{\theta}$  de  $K$  parâmetros. Seja  $\hat{\boldsymbol{\theta}}$  um estimador de  $\boldsymbol{\theta}$  e seja  $\mathbf{V}_0$  um estimador da matriz  $K \times K$  de covariância de  $\hat{\boldsymbol{\theta}}$ , baseado nas hipóteses de independência e igualdade de distribuição das observações (IID), ou equivalentemente, de amostragem aleatória simples com reposição (AASC). É possível generalizar a equação (4.2), definindo o *efeito multivariado do plano amostral de  $\hat{\boldsymbol{\theta}}$  e  $\mathbf{V}_0$*  como

$$\mathbf{EMPA}(\hat{\boldsymbol{\theta}}, \mathbf{V}_0) = \boldsymbol{\Delta} = \mathbf{E}_{VERD}(\mathbf{V}_0)^{-1} \mathbf{V}_{VERD}(\hat{\boldsymbol{\theta}}), \quad (4.6)$$

onde  $\mathbf{E}_{VERD}(\mathbf{V}_0)$  é o valor esperado de  $\mathbf{V}_0$  e,  $\mathbf{V}_{VERD}(\hat{\boldsymbol{\theta}})$  é a matriz de covariância de  $\hat{\boldsymbol{\theta}}$ , ambas calculadas com respeito à distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado, ou alternativamente sob o modelo "correto".

Os autovalores  $\delta_1 \geq \dots \geq \delta_K$  da matriz  $\boldsymbol{\Delta}$  são denominados *efeitos generalizados do plano amostral*. A partir deles, e utilizando resultados padrões de teoria das matrizes (Johnson e Wichern, 1988, p. 64) é possível definir limitantes para os efeitos (univariados) do plano amostral para combinações lineares  $\mathbf{c}'\hat{\boldsymbol{\theta}}$  das componentes de  $\hat{\boldsymbol{\theta}}$ . Temos os seguintes resultados:

$$\begin{aligned} \delta_1 &= \max \mathbf{EPA}(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c}), \\ \delta_K &= \min \mathbf{EPA}(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c}). \end{aligned}$$

No caso particular onde  $\boldsymbol{\Delta} = \mathbf{I}_{K \times K}$ , temos  $\delta_1 = \dots = \delta_K = 1$  e os efeitos (univariados) do plano amostral das combinações lineares para componentes de  $\hat{\boldsymbol{\theta}}$  são todos iguais a 1. Para ilustrar esse conceito, vamos reconsiderar o Exemplo 4.2 de estimação de médias com amostragem estratificada desproporcional anteriormente apresentado, mas agora considerando a natureza multivariada do problema (há duas variáveis de pesquisa).

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

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

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

onde

$$S_{SAL,REC}^{(h)} = \frac{1}{N_h - 1} \sum_{i \in U_h} (SAL_{hi} - \overline{SAL}_h) (REC_{hi} - \overline{REC}_h) .$$

Substituindo os valores conhecidos na população das variáveis  $SAL_{hi}$  e  $REC_{hi}$ , obtemos para esta covariância o valor

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

e portanto a matriz de variância dos estimadores ponderados da média fica igual a

$$\mathbf{V}_{AES}(\overline{SAL}_w, \overline{REC}_w) = \begin{bmatrix} 244,18 & 3,2358 \\ 3,2358 & 0,4350 \end{bmatrix} \quad (4.7)$$

onde os valores das variâncias em (4.7) foram os calculados no Exemplo 4.2 e coincidem, respectivamente, com os valores usados nos numeradores de  $\mathbf{EPA}(\overline{SAL}_w)$  e de  $\mathbf{EPA}(\overline{REC}_w)$  lá apresentados. Para calcular o  $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0)$  é preciso agora obter  $\mathbf{E}_{VERD}(\mathbf{V}_0)$ .

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

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

A partir da simulação foram obtidos os seguintes resultados:

$$\mathbf{E}_{AES}(\mathbf{V}_0) = \begin{bmatrix} 1785,3 & 27,734 \\ 27,734 & 1,2852 \end{bmatrix}, \quad (4.8)$$

$$\mathbf{V}_{AES}(\hat{\theta}) = \begin{bmatrix} 250,41 & 3,2683 \\ 3,2683 & 0,42267 \end{bmatrix} \text{ e} \quad (4.9)$$

$$\mathbf{\Delta} = [\mathbf{E}_{AES}(\mathbf{V}_0)]^{-1} \mathbf{V}_{AES}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} 0,1516 & -4,931 \\ -0,0007277 & 0,4353 \end{bmatrix}. \quad (4.10)$$

Os autovalores  $\delta_1 = 0,447$  e  $\delta_2 = 0,139$  de  $\mathbf{\Delta}$  fornecem os efeitos generalizados do plano amostral.

Da mesma forma que o **EPA**  $(\hat{\boldsymbol{\theta}}, v_0)$  definido em (4.2) para o caso uniparamétrico foi utilizado para corrigir níveis de confiança de intervalos e níveis de significância de testes, o **EMPA**  $(\hat{\boldsymbol{\theta}}, \mathbf{V}_0)$  definido em (4.6) pode ser utilizado para corrigir níveis de confiança de regiões de confiança e níveis de significância de testes de hipóteses no caso multiparamétrico. Para ilustrar, vamos considerar o problema de testar a hipótese  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ , onde  $\boldsymbol{\mu}$  é o vetor de médias de um vetor de variáveis de pesquisa  $\mathbf{y}$ . A estatística de teste usualmente adotada para este caso é a  $T^2$  de Hottelling dada por

$$T^2 = n (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^0 \mathbf{S}_y^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0), \quad (4.11)$$

onde

$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{n} \sum_{i \in s} \mathbf{y}_i, & \mathbf{S}_y &= \frac{1}{n-1} \sum_{i \in s} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})^0, & \text{e} \\ \boldsymbol{\mu}_0 &= (\mu_{10}, \mu_{20}, \dots, \mu_{K0})^0. \end{aligned}$$

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

Contudo, se for utilizado um plano amostral complexo,  $T^2$  tem aproximadamente a distribuição da variável  $\sum_{i=1}^K \delta_i Z_i^2$ , onde  $Z_1, \dots, Z_K$  são variáveis aleatórias independentes com distribuição normal padrão e os  $\delta_i$  são os autovalores da matriz  $\mathbf{\Delta} = \Sigma_{AAS}^{-1} \Sigma$ , onde  $\Sigma_{AAS} = E_p(\mathbf{S}_y/n)$  e  $\Sigma = V_p(\bar{\mathbf{y}})$ .

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

$$P(\chi^2(K) > \chi_\alpha^2(K) / \delta) \quad (4.12)$$

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

A Tabela 4.7 apresenta os níveis de significância reais para  $\alpha = 5\%$  para vários valores de  $K$  e  $\delta$ . Mesmo quando os valores dos  $\delta_i$  são distintos,

os valores da Tabela 4.7 podem ser devidamente interpretados. Para isso, consideremos o  $p$ valor do teste da hipótese  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ , sob a hipótese de amostragem aleatória simples com reposição e sob o plano amostral efetivamente utilizado. Por definição este valor é dado por

$$p\text{valor}_{AAS}(\bar{\mathbf{y}}) = P \left[ \chi^2(K) > (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^0 \boldsymbol{\Sigma}_{AAS}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \right]$$

e  $H_0$  é rejeitada com nível de significância  $\alpha$  se  $p\text{valor}_{AAS} < \alpha$ .

O verdadeiro  $p$ valor pode ser definido analogamente como

$$p\text{valor}_{VERD}(\bar{\mathbf{y}}) = P \left[ \chi^2(K) > (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^0 \boldsymbol{\Sigma}_{VERD}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \right]. \quad (4.13)$$

Os valores na Tabela 4.7 podem ser usados para quantificar a diferença entre estes  $p$ valores. Consideremos a região crítica do teste de nível  $\alpha$  baseado na hipótese de AAS:

$$\begin{aligned} RC_{AAS}(\bar{\mathbf{y}}) &= \left\{ \bar{\mathbf{y}} : (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^0 \boldsymbol{\Sigma}_{AAS}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) > \chi_\alpha^2(K) \right\} \\ &= \left\{ \bar{\mathbf{y}} : p\text{valor}_{AAS}(\bar{\mathbf{y}}) < \alpha \right\}. \end{aligned} \quad (4.14)$$

Pode-se mostrar que o máximo de  $p\text{valor}_{VERD}(\bar{\mathbf{y}})$  quando  $\bar{\mathbf{y}}$  pertence à  $RC_{AAS}(\bar{\mathbf{y}})$  é dado por:

$$\max_{\bar{\mathbf{y}} \in RC_{AAS}(\bar{\mathbf{y}})} p\text{valor}_{VERD}(\bar{\mathbf{y}}) = P \left( \chi^2(K) > \chi_\alpha^2(K) / \delta_1 \right). \quad (4.15)$$

Observe que o segundo membro de (4.15) é da mesma forma que o segundo membro de (4.12). Logo, os valores da Tabela 4.7 podem ser interpretados como valores máximos de  $p\text{valor}_{VERD}(\bar{\mathbf{y}})$  para  $\bar{\mathbf{y}}$  na região  $RC_{AAS}(\bar{\mathbf{y}})$ , considerando-se  $\delta_1$  no lugar de  $\delta$ .

Tabela 4.7: Níveis de significância (%) verdadeiros do teste  $T^2$  para o nível nominal de 5% assumindo autovalores iguais para  $\Delta$

$\delta$	$K$			
	1	2	3	4
0.9	4	4	3	3
1.0	5	5	5	5
1.5	11	14	16	19
2.0	17	22	27	32
2.5	22	30	37	44
3.0	26	37	46	53



## Capítulo 5

# Ajuste de Modelos Paramétricos

### 5.1 Introdução

Nos primórdios do uso "moderno" de pesquisas por amostragem, os dados obtidos eram usados principalmente para estimar funções simples dos valores das variáveis de interesse nas populações finitas, tais como totais, médias, razões, etc. Isto caracterizava o uso dos dados dessas pesquisas para *inferência descritiva*. Recentemente, os dados de pesquisas amostrais têm sido cada vez mais utilizados também para propósitos analíticos. *Inferências analíticas* baseadas numa pesquisa amostral são aquelas que envolvem a estimação de parâmetros num modelo (de superpopulação) (Kalton, 1983b; Binder et al., 1987).

Quando os valores *amostrais* das variáveis da pesquisa podem ser considerados como realizações de vetores aleatórios independentes e identicamente distribuídos (IID), modelos podem ser especificados, ajustados, testados e reformulados usando procedimentos estatísticos padrões como os apresentados, por exemplo, em Bickel e Doksum(1977) e Garthwaite, Jolliffe e Jones(1995). Neste caso, métodos e pacotes estatísticos padrões podem ser usados para executar os cálculos de estimativas de parâmetros e medidas de precisão correspondentes, bem como diagnóstico e verificação da adequação das hipóteses dos modelos.

Na prática das pesquisas amostrais, contudo, as hipóteses de modelo IID para as observações amostrais são raramente adequadas. Com maior frequência, modelos alternativos com hipóteses mais complexas e/ou estimadores especiais devem ser considerados a fim de acomodar aspectos da

estrutura da população e/ou do plano amostral. Além disso, usualmente estão disponíveis informações sobre variáveis auxiliares, utilizadas ou não na especificação do plano amostral, que podem ser incorporadas com proveito na estimação dos parâmetros ou na própria formulação modelo.

Os exemplos apresentados no Capítulo 4 demonstram claramente a inadequação de ignorar o plano amostral ao efetuar análises de dados de pesquisas amostrais. Os valores dos EPAs calculados, tanto para estimadores de medidas descritivas tais como médias e totais, como para estatísticas analíticas usadas em testes de hipóteses e os correspondentes efeitos nos níveis de significância reais, revelam que ignorar o plano amostral pode levar a decisões erradas e a avaliações inadequadas da precisão das estimativas amostrais.

Embora as medidas propostas no Capítulo 4 para os efeitos de plano amostral sirvam para avaliar o impacto de ignorar o plano amostral nas inferências descritivas ou mesmo analíticas baseadas em dados amostrais, elas não resolvem o problema de como incorporar o plano amostral nessas análises. No caso das inferências descritivas usuais para médias, totais e proporções, o assunto é amplamente tratado na literatura de amostragem e o interessado em maiores detalhes pode consultar livros clássicos como Cochran(1977), ou mais recentes como Särndal, Swensson e Wretman(1992). Já os métodos requeridos para inferências analíticas só recentemente foram consolidados em livro (Skinner, Holt e Smith, 1989). Este capítulo apresenta um dos métodos centrais disponíveis para ajuste de modelos paramétricos regulares considerando dados amostrais complexos, baseado no trabalho de Binder(1983). Antes de descrever esse método, entretanto, fazemos breve discussão sobre o papel dos pesos na análise de dados amostrais, considerando o trabalho de Pfeffermann(1993).

Primeiramente, porém, fazemos uma revisão sucinta do método de Máxima Verossimilhança (MV) para ajustar modelos dentro da abordagem de modelagem clássica, necessária para compreensão adequada do material subsequente. Essa revisão não pretende ser exaustiva ou detalhada, mas tão somente recordar os principais resultados aqui requeridos. Para uma discussão mais detalhada do método de Máxima Verossimilhança para estimação em modelos paramétricos regulares veja, por exemplo, Garthwaite, Jolliffe e Jones(1995).

## 5.2 Método de Máxima Verossimilhança (MV)

Seja  $\mathbf{y}_i = (y_{i1}, \dots, y_{iR})^0$  um vetor  $R \times 1$  dos valores observados das variáveis de interesse observadas para a unidade  $i$  da amostra, gerado por um vetor aleatório  $\mathbf{Y}_i$ , para  $i = 1, \dots, n$ , onde  $n$  é o tamanho da amostra. Suponha que os vetores aleatórios  $\mathbf{Y}_i$ , para  $i = 1, \dots, n$ , são independentes e identicamente distribuídos (IID) com distribuição comum  $f(\mathbf{y}; \boldsymbol{\theta})$ , onde  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^0$  é um vetor  $K \times 1$  de parâmetros desconhecidos de interesse. Sob essas hipóteses, a verossimilhança amostral é dada por

$$l(\boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i; \boldsymbol{\theta})$$

e a correspondente log-verossimilhança por

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \log [f(\mathbf{y}_i; \boldsymbol{\theta})] .$$

Calculando as derivadas parciais de  $L(\boldsymbol{\theta})$  com relação a cada componente de  $\boldsymbol{\theta}$  e igualando a 0, obtemos um sistema de equações

$$\partial L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \sum_{i=1}^n \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0}$$

onde  $\mathbf{u}_i(\boldsymbol{\theta}) = \partial \log [f(\mathbf{y}_i; \boldsymbol{\theta})] / \partial \boldsymbol{\theta}$  é o vetor dos escores da unidade  $i$ , de dimensão  $K \times 1$ .

Sob condições de regularidade (Cox e Hinkley, 1974, p. 281), a solução  $\hat{\boldsymbol{\theta}}$  deste sistema de equações é o Estimador de Máxima Verossimilhança (EMV) de  $\boldsymbol{\theta}$ . A variância assintótica do estimador  $\hat{\boldsymbol{\theta}}$  sob o modelo adotado, denominado aqui abreviadamente modelo  $M$ , é dada por

$$V_M(\hat{\boldsymbol{\theta}}) \simeq [J(\boldsymbol{\theta})]^{-1}$$

e um estimador consistente dessa variância é dado por

$$\hat{V}_M(\hat{\boldsymbol{\theta}}) = [J(\hat{\boldsymbol{\theta}})]^{-1} ,$$

onde

$$J(\boldsymbol{\theta}) = \sum_{i=1}^n \partial \mathbf{u}_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$$

e

$$J(\hat{\boldsymbol{\theta}}) = J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} .$$

### 5.3 Ponderação de Dados Amostrais

O papel da ponderação *na análise de dados amostrais* é alvo de controvérsia entre os estatísticos. Apesar de incorporada comumente na inferência descritiva, não há concordância com respeito a seu uso na inferência analítica, havendo um espectro de opiniões entre dois extremos. Num extremo estão os *modelistas*, que consideram o uso de pesos irrelevante, e no outro os *amostristas*, que incorporam pesos em qualquer análise.

#### Exemplo 5.1 *Uso analítico dos dados da Pesquisa Nacional por Amostragem de Domicílios (PNAD)*

A título de ilustração, consideremos uma pesquisa com uma amostra complexa como a da PNAD do IBGE, que emprega uma amostra estratificada de domicílios em três estágios, tendo como unidades primárias de amostragem (UPAs) os municípios, que são estratificados segundo as unidades da federação (UFs), e regiões menores dentro das UFs (veja IBGE, 1981, p. 67).

A seleção de municípios dentro de cada estrato é feita com probabilidades desiguais, proporcionais ao tamanho, havendo inclusive municípios incluídos na amostra com certeza (chamados de municípios auto-representativos). Da mesma forma, a seleção de setores (unidades secundárias de amostragem ou USAs) dentro de cada município é feita com probabilidades proporcionais ao número de domicílios em cada setor segundo o último censo disponível. Dentro de cada setor, a seleção de domicílios é feita por amostragem sistemática simples (portanto, com equiprobabilidade). Todas as pessoas moradoras em cada domicílio da amostra são pesquisadas.

A amostra de domicílios e de pessoas dentro de cada estrato é *autoponderada*, isto é, tal que todos os domicílios e pessoas dentro de um mesmo estrato têm igual probabilidade de seleção. Entretanto, as probabilidades de inclusão (e conseqüentemente os pesos) variam bastante entre as várias regiões de pesquisa. A Tabela 5.1 revela como variam essas probabilidades de seleção entre as regiões cobertas pela amostra da PNAD de 93. Como se pode observar, tais probabilidades de inclusão chegam a ser 5 vezes maiores em Belém do que em São Paulo, e portanto variação semelhante será observada nos pesos.

Se  $\pi_i$  representa a probabilidade de inclusão na amostra do  $i$ -ésimo domicílio da população,  $i = 1, \dots, N$ , então

$$\pi_i = \pi_{\text{município}|\text{estrato}} \times \pi_{\text{setor}|\text{município}} \times \pi_{\text{domicílio}|\text{setor}}$$

Tabela 5.1: Probabilidades de seleção da amostra da PNAD de 1993 segundo regiões

Região da pesquisa	Probabilidade de seleção
RM de Belém	1/150
RMs de Fortaleza, Recife, Salvador e Porto Alegre Distrito Federal	1/200
RMs de Belo Horizonte e Curitiba	1/250
Rondônia, Acre, Amazonas, Roraima, Amapá, Tocantins, Sergipe, Mato Grosso do Sul, Mato Grosso e Goiás	1/300
Pará	1/350
RM do Rio de Janeiro, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Bahia, Minas Gerais, Espírito Santo e Rio de Janeiro	1/500
Paraná, Santa Catarina, Rio Grande do Sul	1/550
RM de São Paulo, Maranhão, São Paulo	1/750

isto é, a probabilidade global de inclusão de um domicílio (e conseqüentemente de todas as pessoas nele moradoras) é dada pelo produto das probabilidades condicionais de inclusão nos vários estágios de amostragem.

A estimação do total populacional  $Y$  de uma variável de pesquisa  $y$  num dado estrato usando os dados da PNAD é feita rotineiramente com estimadores ponderados de tipo razão  $\hat{Y}_R = \hat{Y}_\pi X / \hat{X}_\pi = \sum_{i \in s} w_i^R y_i$  (tal como definidos por 3.15), com pesos dados por  $w_i^R = \pi_i^{-1} X / \hat{X}_\pi$  (veja 3.17), onde  $X$  é o total da população no estrato obtido por métodos demográficos de projeção, utilizado como variável auxiliar, e  $\hat{X}_\pi$  e  $\hat{Y}_\pi$  são os estimadores  $\pi$ -ponderados de  $X$  e  $Y$  respectivamente. Para estimar para conjuntos de estratos basta somar as estimativas para cada estrato incluído no conjunto. Para estimar médias e proporções, os pesos são também incorporados da forma apropriada. No caso, a estimação de médias é feita usando estimadores ponderados da forma

$$\bar{y}^R = \frac{\sum_{i \in s} w_i^R y_i}{\sum_{i \in s} w_i^R}$$

e a estimação de proporções é caso particular da estimação de médias quando a variável de pesquisa  $y$  é do tipo indicador (isto é, só toma valores 0 e 1).

Estimadores ponderados (como por exemplo os usados na PNAD) são preferidos pelos praticantes de amostragem por sua simplicidade e por serem não viciados (ao menos aproximadamente) com respeito à distribuição de aleatorização induzida pela seleção da amostra, independentemente dos valores assumidos pelas variáveis de pesquisa na população. Já para a modelagem de relações entre variáveis de pesquisa, o uso dos pesos induzidos pelo planejamento amostral ainda não é freqüente ou aceito sem controvérsia.

Um exemplo de modelagem desse tipo com dados da PNAD em que os pesos e o desenho amostral não foram considerados na análise é encontrado em Leote(1996). Essa autora empregou modelos de regressão logística para traçar um perfil sócio-econômico da mão-de-obra empregada no mercado informal de trabalho urbano no Rio de Janeiro, usando dados do suplemento sobre trabalho da PNAD-90. Todos os ajustes efetuados ignoraram os pesos e o plano amostral da pesquisa. O problema foi revisitado por Pessoa, Nascimento Silva e Duarte(1997), quando então esses aspectos foram devidamente incorporados na análise. Um resumo desse trabalho é discutido no Capítulo 6.

Vamos supor que haja interesse em regredir uma determinada variável de pesquisa  $y$  contra algumas outras variáveis de pesquisa num vetor de regressores  $\mathbf{z}$ . Seria natural indagar se, como no caso do total e da média, os pesos amostrais poderiam desempenhar algum papel na estimação dos

parâmetros do modelo (linear) de regressão? Uma possibilidade de incluir os pesos seria estimar os coeficientes da regressão por:

$$\hat{\beta}_w = \left( \sum_{i \in s} w_i \mathbf{z}'_i \mathbf{z}_i \right)^{-1} \sum_{i \in s} w_i \mathbf{z}'_i y_i = (\mathbf{Z}'_s \mathbf{W}_s \mathbf{Z}_s)^{-1} \mathbf{Z}'_s \mathbf{W}_s \mathbf{Y}_s \quad (5.1)$$

em lugar do estimador de mínimos quadrados ordinários (MQO) dado por

$$\hat{\beta} = \left( \sum_{i \in s} \mathbf{z}'_i \mathbf{z}_i \right)^{-1} \sum_{i \in s} \mathbf{z}'_i y_i = (\mathbf{Z}'_s \mathbf{Z}_s)^{-1} \mathbf{Z}'_s \mathbf{Y}_s \quad (5.2)$$

onde  $w_i = \pi_i^{-1}$ ,  $y_i$  é o valor da variável resposta e  $\mathbf{z}_i$  é o vetor de regressores para a observação  $i$ ,  $\mathbf{Z}_s$  e  $\mathbf{Y}_s$  são respectivamente a matriz e vetor com os valores amostrais dos  $\mathbf{z}_i$  e  $y_i$ , e  $\mathbf{W}_s = \text{diag}\{w_i; i \in s\}$  é a matriz diagonal com os pesos amostrais.

Não é possível justificar o estimador  $\hat{\beta}_w$  em (5.1) com base em critério de otimalidade, tal como ocorre com os estimadores usuais de Máxima Verossimilhança ou de Mínimos Quadrados Ordinários (MQO), se uma modelagem clássica IID fosse adotada para a amostra.

De um ponto de vista formal (matemático), o estimador  $\hat{\beta}_w$  em (5.1) é equivalente ao estimador de Mínimos Quadrados Ponderados (MQP) com pesos  $w_i$ . Entretanto, esses estimadores diferem de maneira acentuada. Os estimadores de MQP são usualmente considerados quando o modelo de regressão é heteroscedástico, isto é, quando os resíduos têm variâncias desiguais. Neste caso, os pesos adequados seriam dados pelos inversos das variâncias dos resíduos correspondentes a cada uma das observações, e portanto em geral diferentes dos pesos iguais aos inversos das correspondentes probabilidades de seleção. Além desta diferença de interpretação do papel dos pesos no estimador, outro aspecto em que os dois estimadores diferem de forma acentuada é na estimação da precisão, com o estimador MQP acoplado a um estimador de variância baseado no modelo e o estimador  $\hat{\beta}_w$  acoplado a estimadores de variância que incorporam o planejamento amostral e os pesos, tal como se verá mais adiante.

O estimador  $\hat{\beta}_w$  foi proposto formalmente por Fuller(1975), que o concebeu como uma função de estimadores de totais populacionais. A mesma idéia subsidiou vários outros autores que estudaram a estimação de coeficientes de regressão partindo de dados amostrais complexos, tais como Nathan e Holt(1980), Pfeiffermann e Nathan(1981). Uma revisão abrangente da literatura existente sobre estimação de parâmetros em modelos de regressão

linear com dados amostrais complexos pode ser encontrada em Nascimento Silva(1996, cap. 6).

Apesar dessas dificuldades, será que é possível justificar o uso de pesos na inferência baseada em modelos? Se for o caso, sob que condições? Seria possível desenvolver diretrizes para o uso de pesos em inferência analítica partindo de dados amostrais complexos? A resposta para essas perguntas é afirmativa, ao menos quando a questão da robustez da inferência é relevante. Em inferências analíticas partindo de dados amostrais complexos, os pesos podem ser usados para proteger:

1. contra *planos amostrais não-ignoráveis*, que poderiam introduzir ou causar *vícios*;
2. contra a *má especificação do modelo*.

A *robustez dos procedimentos* que incorporam pesos é obtida pela mudança de foco da inferência para *quantidades da população finita*, que definem parâmetros-alvo alternativos aos parâmetros do modelo de superpopulação, conforme já discutido na Seção 2.1.4.

A questão da construção dos pesos não será tratada neste texto, usando-se sempre como peso o inverso da probabilidade de inclusão na amostra. É possível utilizar pesos de outro tipo como, por exemplo, aqueles de razão empregados na estimação da PNAD, ou mesmo pesos de regressão. Para esses casos, há que fazer alguns ajustes da teoria aqui exposta (veja Nascimento Silva, 1996, cap. 6).

Há várias formas alternativas de incorporar os pesos amostrais no processo de inferência. A principal que será adotada ao longo deste texto será o método de *Máxima Pseudo-Verossimilhança*, que descrevemos na próxima seção.

## 5.4 Método de Máxima Pseudo-Verossimilhança

Suponha que os vetores observados  $\mathbf{y}_i$  das variáveis de pesquisa do elemento  $i$  são gerados por vetores aleatórios  $\mathbf{Y}_i$ , para  $i \in U$ . Suponha também que  $\mathbf{Y}_1, \dots, \mathbf{Y}_N$  são IID com densidade  $f(\mathbf{y}, \boldsymbol{\theta})$ . Se todos os elementos da população finita  $U$  fossem conhecidos, as funções de verossimilhança e de log-verossimilhança *populacionais* seriam dadas respectivamente por

$$l_U(\boldsymbol{\theta}) = \prod_{i \in U} f(\mathbf{y}_i; \boldsymbol{\theta}) \quad (5.3)$$

e

$$L_U(\boldsymbol{\theta}) = \sum_{i \in U} \log [f(\mathbf{y}_i; \boldsymbol{\theta})] \quad . \quad (5.4)$$

As equações de verossimilhança *populacionais* correspondentes são dadas por

$$\sum_{i \in U} \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0} \quad (5.5)$$

onde

$$\tilde{\mathbf{u}}_i(\boldsymbol{\theta}) = \partial \log [f(\mathbf{y}_i; \boldsymbol{\theta})] / \partial \boldsymbol{\theta} \quad (5.6)$$

é o vetor  $K \times 1$  dos escores do elemento  $i, i \in U$ .

Sob condições de regularidade (Cox e Hinkley, 1974, p. 281), a solução  $\boldsymbol{\theta}_U$  deste sistema é o *Estimador de Máxima Verossimilhança* de  $\boldsymbol{\theta}$  no caso de um *censo*. Podemos considerar  $\boldsymbol{\theta}_U$  como uma *Quantidade Descritiva Populacional Correspondente (QDPC)* a  $\boldsymbol{\theta}$ , no sentido definido por Pfeffermann(1993), sobre a qual se deseja fazer inferências com base em informações da amostra. Essa definição da QDPC  $\boldsymbol{\theta}_U$  pode ser generalizada para contemplar outras abordagens de inferência além da abordagem clássica baseada em maximização da verossimilhança. Basta para isso especificar outra regra ou critério a otimizar e então definir a QDPC como a solução ótima segundo essa nova regra. Tal generalização, discutida em Pfeffermann(1993), não será aqui considerada para manter a simplicidade.

A QDPC  $\boldsymbol{\theta}_U$  definida com base em (5.5) não é calculável a menos que um censo seja realizado. Entretanto, desempenha papel fundamental nessa abordagem inferencial, por constituir-se num *pseudo-parâmetro*, eleito como alvo da inferência num esquema que incorpora o planejamento amostral. Isto se justifica porque, sob certas condições de regularidade,  $\boldsymbol{\theta}_U - \boldsymbol{\theta} = o_p(1)$ . Como em pesquisas por amostragem o tamanho da população é geralmente grande, um estimador adequado para  $\boldsymbol{\theta}_U$  será geralmente adequado também para  $\boldsymbol{\theta}$ .

Seja  $\mathbf{T} = \sum_{i \in U} \mathbf{u}_i(\boldsymbol{\theta})$  a soma dos vetores de escores na população, o qual é um vetor de totais populacionais. Para estimar este vetor de totais, podemos então usar um estimador linear ponderado da forma  $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta})$  (veja capítulo 3) onde  $w_i$  são pesos propriamente definidos. Com essa notação, podemos agora obter um estimador para  $\boldsymbol{\theta}_U$  resolvendo o sistema de equações obtido igualando o estimador  $\hat{\mathbf{T}}$  do total  $\mathbf{T}$  a zero.

Definição 5.1 *O estimador de Máxima Pseudo-Verossimilhança  $\hat{\boldsymbol{\theta}}_{MPV}$  de  $\boldsymbol{\theta}_U$  (e conseqüentemente de  $\boldsymbol{\theta}$ ) será a solução das equações de Pseudo-Verossimilhança dadas por*

$$\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0} . \quad (5.7)$$

Através da linearização de Taylor (veja Seção 3.3) e considerando os resultados de Binder(1983), podemos obter a variância de aleatorização assintótica do estimador  $\hat{\boldsymbol{\theta}}_{MPV}$  e seu estimador correspondente, dados respectivamente por:

$$V_p(\hat{\boldsymbol{\theta}}_{MPV}) \simeq [J(\boldsymbol{\theta}_U)]^{-1} V_p \left[ \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta}_U) \right] [J(\boldsymbol{\theta}_U)]^{-1} \quad (5.8)$$

e

$$\hat{V}_p(\hat{\boldsymbol{\theta}}_{MPV}) = [\hat{J}(\hat{\boldsymbol{\theta}}_{MPV})]^{-1} \hat{V}_p \left[ \sum_{i \in s} w_i \mathbf{u}_i(\hat{\boldsymbol{\theta}}_{MPV}) \right] [\hat{J}(\hat{\boldsymbol{\theta}}_{MPV})]^{-1} , \quad (5.9)$$

onde

$$J(\boldsymbol{\theta}_U) = \left. \frac{\partial T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U} = \sum_{i \in U} \left. \frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U} , \quad (5.10)$$

$$\hat{J}(\hat{\boldsymbol{\theta}}_{MPV}) = \left. \frac{\partial \hat{T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}} = \sum_{i \in s} w_i \left. \frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}} , \quad (5.11)$$

$V_p \left[ \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta}_U) \right]$  é a matriz de variância (de aleatorização) do estimador do total populacional dos escores e  $\hat{V}_p \left[ \sum_{i \in s} w_i \mathbf{u}_i(\hat{\boldsymbol{\theta}}_{MPV}) \right]$  é um estimador consistente para esta variância. Binder(1983) mostrou também que a distribuição assintótica de  $\hat{\boldsymbol{\theta}}_{MPV}$  é Normal Multivariada, isto é, que

$$\left[ \hat{V}_p(\hat{\boldsymbol{\theta}}_{MPV}) \right]^{-1/2} (\hat{\boldsymbol{\theta}}_{MPV} - \boldsymbol{\theta}_U) \sim \mathbf{NM}(\mathbf{0}; \mathbf{I}) , \quad (5.12)$$

o que fornece uma base para a inferência sobre  $\boldsymbol{\theta}_U$  (ou  $\boldsymbol{\theta}$ ) usando amostras grandes.

Muitos modelos paramétricos, com vários planos amostrais e estimadores de totais diferentes, podem ser ajustados resolvendo-se as equações de Pseudo-Verossimilhança (5.7), satisfeitas algumas condições de regularidade enunciadas em Binder(1983, apêndice) e revistas em Nascimento Silva(1996,

p. 126). Entretanto, os estimadores de MPV não serão únicos, já que existem diversas maneiras de se definir os pesos  $w_i$ .

Os pesos  $w_i$  devem ser tais que os estimadores de total em (5.7) sejam assintoticamente normais e não-viciados, e possuam estimadores de variância consistentes, conforme requerido para a obtenção da distribuição assintótica dos estimadores MPV. Os pesos mais usados são os do estimador  $\pi$ -ponderado ou de Horvitz-Thompson para totais, dados pelo inverso das probabilidades de inclusão dos indivíduos, ou seja  $w_i = \pi_i^{-1}$ . Tais pesos satisfazem essas condições sempre que  $\pi_i > 0$  e  $\pi_{ij} > 0 \quad \forall i, j \in U$  e algumas condições adicionais de regularidade são satisfeitas (veja Fuller, 1984).

Assim, um procedimento padrão para ajustar um modelo paramétrico regular  $f(\mathbf{y}; \boldsymbol{\theta})$  pelo método da Máxima Pseudo-Verossimilhança seria dado pelos passos indicados a seguir.

1. Resolver  $\sum_{i \in s} \pi_i^{-1} \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0}$  e calcular o estimador pontual  $\hat{\boldsymbol{\theta}}_\pi$  do parâmetro  $\boldsymbol{\theta}$  no modelo  $f(\mathbf{y}; \boldsymbol{\theta})$  (ou do pseudo-parâmetro  $\boldsymbol{\theta}_U$  correspondente).
2. Calcular a matriz de variância estimada

$$\hat{V}_p(\hat{\boldsymbol{\theta}}_\pi) = [\hat{J}(\hat{\boldsymbol{\theta}}_\pi)]^{-1} \hat{V}_p \left[ \sum_{i \in s} \pi_i^{-1} \mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi) \right] [\hat{J}(\hat{\boldsymbol{\theta}}_\pi)]^{-1}, \quad (5.13)$$

onde

$$\hat{V}_p \left[ \sum_{i \in s} \pi_i^{-1} \mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi) \right] = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} [\mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi)] [\mathbf{u}_j(\hat{\boldsymbol{\theta}}_\pi)]^0 \quad (5.14)$$

e

$$\hat{J}(\hat{\boldsymbol{\theta}}_\pi) = \left. \frac{\partial \hat{T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\pi} = \sum_{i \in s} \pi_i^{-1} \left. \frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\pi}. \quad (5.15)$$

3. Usar  $\hat{\boldsymbol{\theta}}_\pi$  e  $\hat{V}_p(\hat{\boldsymbol{\theta}}_\pi)$  para calcular regiões ou intervalos de confiança e/ou estatísticas de teste baseadas na distribuição normal e utilizá-las para fazer inferência sobre os componentes de  $\boldsymbol{\theta}$ .

**Observação 5.1** *No Método de Máxima Pseudo-Verossimilhança, os pesos amostrais são incorporados na análise através das equações de estimação dos parâmetros (5.7) e através das equações de estimação da matriz de covariância dos estimadores(5.13)-(5.15).*

**Observação 5.2** *O plano amostral é também incorporado no método de estimação MPV através da expressão para a variância do total dos escores sob o plano amostral (5.14), onde as propriedades do plano amostral estão resumidas nas probabilidades de inclusão de primeira e segunda ordem, isto é, os  $\pi_i$  e os  $\pi_{ij}$  respectivamente.*

**Observação 5.3** *Sob probabilidades de seleção iguais, os pesos  $\pi_i^{-1}$  serão constantes e o estimador pontual  $\hat{\theta}_\pi$  será idêntico ao estimador de Máxima Verossimilhança (MV) ordinário para uma amostra de observações IID com distribuição  $f(\mathbf{y}; \boldsymbol{\theta})$ . Entretanto, o mesmo não ocorre em se tratando da variância do estimador  $\hat{\theta}_\pi$ , que difere da variância sob o modelo do estimador usual de MV.*

#### Vantagens do procedimento de MPV

O procedimento MPV proporciona estimativas *baseadas no plano amostral* para a variância assintótica dos estimadores dos parâmetros, as quais são razoavelmente simples de calcular e são consistentes sob *condições fracas* no plano amostral e na especificação do modelo. Mesmo quando o estimador pontual de MPV coincide com o estimador usual de Máxima Verossimilhança, a estimativa da variância obtida pelo procedimento de MPV pode ser preferível aos estimadores usuais da variância baseados no modelo, que ignoram o plano amostral.

O procedimento MPV fornece estimativas *robustas*, no sentido de que em muitos casos a quantidade  $\boldsymbol{\theta}_U$  da população finita permanece um alvo válido para inferência, mesmo quando o modelo especificado por  $f(\mathbf{y}; \boldsymbol{\theta})$  não proporciona uma descrição adequada para a distribuição das variáveis de pesquisa na população.

#### Desvantagens do método de MPV

Este procedimento requer conhecimento de informações detalhadas sobre os elementos da amostra, tais como pertinência a estratos e conglomerados ou unidades primárias de amostragem, e suas probabilidades de inclusão ou pesos. Tais informações nem sempre estão disponíveis para usuários de dados de pesquisas amostrais, seja por razões operacionais ou devido às regras de proteção do sigilo de informações individuais.

As propriedades dos estimadores MPV não são conhecidas para pequenas amostras. Este problema pode não ser importante em análises que usam os dados de pesquisas feitas pelas agências oficiais de estatística, desde que em tais análises seja utilizada a amostra inteira, ou no caso de subdomínios estudados separadamente, que as amostras usadas sejam suficientemente grandes nestes domínios.

Outra dificuldade é que métodos usuais de diagnóstico de ajuste de modelos (tais como gráficos de resíduos) e outros procedimentos da inferência clássica (tais como testes estatísticos de Razões de Verossimilhança) não podem ser utilizados.

## 5.5 Robustez do Procedimento MPV

Nesta seção vamos examinar a questão da robustez dos estimadores obtidos pelo procedimento MPV. É essa robustez que justifica o emprego desses estimadores frente aos estimadores usuais de MV, pois nas situações práticas da análise de dados amostrais complexos as hipóteses usuais de modelo IID para as observações amostrais raramente são verificadas.

Vamos agora analisar com mais detalhes a terceira abordagem para a inferência analítica. Nela, postulamos um modelo como na primeira abordagem e a inferência é direcionada aos parâmetros do modelo. Porém, em vez de acharmos um estimador ótimo sob o modelo, achamos um estimador na classe dos estimadores consistentes para a QDPC, onde a consistência é referida à distribuição de aleatorização do estimador. Por que usar a QDPC? A resposta é exatamente para obter maior robustez. Para entender porque essa abordagem oferece maior robustez, vamos considerar dois casos.

- Caso 1: o modelo para a população é adequado.

Então quando  $N \rightarrow \infty$  a QDPC  $\theta_U$  converge para o parâmetro  $\theta$ , isto é,  $\theta_U - \theta \rightarrow \mathbf{0}$  em probabilidade, segundo a distribuição de probabilidades do modelo  $M$ . Se  $\hat{\theta}_{MPV}$  for consistente, então quando  $n \rightarrow \infty$  temos que  $\hat{\theta}_{MPV} - \theta_U \rightarrow \mathbf{0}$  em probabilidade, segundo a distribuição de aleatorização  $p$ . Juntando essas condições obtemos que

$$\hat{\theta}_{MPV} \xrightarrow{P} \theta$$

em probabilidade segundo a mistura  $Mp$ . Esse resultado segue porque

$$\hat{\theta}_{MPV} - \theta = (\hat{\theta}_{MPV} - \theta_U) + (\theta_U - \theta) = O_p(n^{-1/2}) + O_p(N^{-1/2}) = O_p(n^{-1/2}).$$

- Caso 2: o modelo para a população não é válido.

Nesse caso, o parâmetro  $\theta$  do modelo não tem interpretação substantiva significativa, porém a QDPC  $\theta_U$  é uma entidade definida na população finita (real) com interpretação clara, independente da validade do modelo.

Como  $\hat{\theta}_{MPV}$  é consistente para a QDPC  $\theta_U$ , a inferência baseada no procedimento MPV segue válida para este pseudo-parâmetro, independente da inadequação do modelo para a população. Skinner(1989b, p. 81) discute essa situação, mostrando que  $\theta_U$  pode ainda ser um alvo válido para inferência mesmo quando o modelo  $f(\mathbf{y}; \theta)$  especificado para a população é inadequado, ao menos no sentido de que  $f(\mathbf{y}; \theta_U)$  forneceria a *melhor aproximação possível* (em certo sentido) para o verdadeiro modelo que gera as observações populacionais ( $f^*(\mathbf{y}; \eta)$ , digamos). Skinner(1989b) reconhece que a *melhor aproximação possível* entre um conjunto de aproximações ruins ainda seria uma aproximação ruim, e portanto que a escolha do elenco de modelos especificados pela distribuição  $f(\mathbf{y}; \theta)$  deve seguir os cuidados necessários para garantir que esta escolha forneça uma aproximação razoável da realidade.

Observação 5.4 *Consistência referente à distribuição de aleatorização.*

Consistência na teoria clássica tem a ver com comportamento limite de um estimador quando o tamanho da amostra cresce, isto é, quando  $n \rightarrow \infty$ . No caso de populações finitas, temos que considerar o que ocorre quando crescem o tamanho da amostra e também o tamanho da população, isto é, quando  $n \rightarrow \infty$  e  $N \rightarrow \infty$ . Neste caso, é preciso definir a maneira pela qual  $N \uparrow$  e  $n \uparrow$  preservando a estrutura do plano amostral. Para evitar um desvio indesejado que a discussão deste problema traria, vamos supor que  $N \uparrow$  e  $n \uparrow$  de uma forma bem definida. Os leitores interessados poderão consultar: Särndal, Swensson e Wretman(1992, p. 166), Brewer(1979), Isaki e Fuller(1982), Robinson e Särndal(1983), Hájek(1960) e Skinner, Holt e Smith(1989, p. 18-19).

## 5.6 Desvantagens da Inferência Baseada em Aleatorização

Se o modelo postulado para os dados amostrais for correto, o uso de estimadores ponderados pode resultar em perda substancial de eficiência comparado com o estimador ótimo, sob o modelo. Em geral, a perda de eficiência aumenta quando diminui o tamanho da amostra e aumenta a variação dos pesos. Há casos onde a ponderação é a única alternativa. Por exemplo, se os dados disponíveis já estão na forma de estimativas amostrais ponderadas, então o uso de pesos é inevitável. Um exemplo clássico é discutido a seguir.

## 5.6. DESVANTAGENS DA INFERÊNCIA BASEADA EM ALEATORIZAÇÃO 85

Exemplo 5.2 *Análise secundária de tabelas de contingência.*

A pesquisa *Canada Health Survey* usa um plano amostral estratificado com vários estágios de seleção. Nessa pesquisa, a estimativa de contagem na cela  $k$  de uma tabela de contingência qualquer é dada por

$$\hat{N}_k = \sum_a \left( N_a / \hat{N}_a \right) \left[ \sum_h \sum_i \sum_j w_{hij} Y_{ka(hij)} \right] = \sum_a \left( N_a / \hat{N}_a \right) \hat{N}_{ka}$$

onde  $Y_{ka(hij)} = 1$  se a  $j$ -ésima unidade da UPA  $i$  do estrato  $h$  pertence à  $k$ -ésima cela e ao  $a$ -ésimo grupo de idade-sexo, e 0 (zero) caso contrário;

$N_a / \hat{N}_a$  – são fatores de ajustamento de pós-estratificação que usam contagens censitárias  $N_a$  de idade-sexo para diminuir as variâncias dos estimadores.

Quando as contagens *expandidas*  $\hat{N}_k$  são usadas, os testes de homogeneidade e de qualidade de ajuste de modelos loglineares baseados em amostragem Multinomial e Poisson independentes não são mais válidos. A estatística clássica  $X^2$  não tem mais distribuição  $\chi^2$  e sim uma soma ponderada  $\sum_k \delta_k X_k$  de variáveis  $X_k$  IID com distribuição  $\chi^2(1)$ . Esse exemplo será rediscutido com mais detalhes na Seção 7.3.3.

A importância desse exemplo é ilustrar que mesmo quando o usuário pensa estar livre das complicações causadas pelo plano amostral e pesos, ele precisa estar atento à forma como foram gerados os dados que pretende modelar ou analisar, sob pena de realizar inferências incorretas. Este exemplo tem também grande importância prática, pois um grande número de pesquisas domiciliares por amostragem produz como principal resultado conjunto de tabelas com contagens e proporções, as quais foram obtidas mediante ponderação pelas agências produtoras. Este é o caso, por exemplo, da PNAD, da amostra do Censo Demográfico e de inúmeras outras pesquisas do IBGE e de agências estatísticas congêneres.



## Capítulo 6

# Modelos de Regressão

### 6.1 Modelo de Regressão Linear Normal

O problema considerado nesta seção é o de estimar os parâmetros num modelo de regressão linear normal especificado para um subconjunto das variáveis da pesquisa. O procedimento de máxima pseudo-verossimilhança, descrito na Seção 5.4, é aplicado. Os resultados são derivados considerando pesos ordinários dados pelo inverso das probabilidades de inclusão das unidades na amostra. Resultados mais gerais considerando outros tipos de pesos (tais como os derivados de estimadores de razão ou regressão, por exemplo) estão discutidos em Nascimento Silva(1996, cap. 6).

#### 6.1.1 Especificação do Modelo

Vamos supor que os dados da  $i$ -ésima unidade da população pesquisada incluam um vetor  $\mathbf{z}_i = (z_{i1}, \dots, z_{iP})^0$  de dimensão  $P \times 1$  com os valores de variáveis  $\mathbf{z}$ , que são **preditoras** ou **explanatórias** num modelo de regressão  $M$ . Este modelo tem o objetivo de prever ou explicar os valores de uma variável da pesquisa  $y$ , que é considerada como **variável resposta**. Denotemos por  $Y_i$  e  $\mathbf{Z}_i$  a variável e o vetor aleatórios que geram  $y_i$  e  $\mathbf{z}_i$ , para  $i \in U$ . Sem perda de generalidade, suponhamos também que a primeira componente do vetor  $\mathbf{z}_i$  de variáveis preditoras é sempre igual a 1, de modo a incluir sempre um termo de intercepto nos modelos de regressão linear considerados (tal hipótese não é essencial, mas será adotada no restante deste capítulo). Suponhamos agora que  $(Y_i, \mathbf{Z}_i^0)^0$ ,  $i \in U$ , são vetores aleatórios independentes

e identicamente distribuídos tais que

$$f(y_i | \mathbf{z}_i; \boldsymbol{\beta}, \sigma_e) = (2\pi\sigma_e)^{-1/2} \exp \left[ - \left( y_i - \mathbf{z}_i' \boldsymbol{\beta} \right)^2 / 2\sigma_e \right] \quad (6.1)$$

onde  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  e  $\sigma_e > 0$  são parâmetros desconhecidos do modelo.

Observe que (6.1) constitui-se numa especificação (parcial) de um modelo marginal para um conjunto de variáveis da pesquisa, e não faz nenhuma referência direta à forma como elas se relacionam com variáveis auxiliares  $\mathbf{x}$  que eventualmente possam estar disponíveis. A atenção é focalizada na estimação de  $\boldsymbol{\beta}$  e  $\sigma_e$  e sua interpretação com respeito ao modelo agregado (6.1).

Modelos como (6.1) já foram considerados por vários autores, por exemplo Holt, Smith e Winter(1980), Nathan e Holt (1980), Skinner(1989b, p.81), Chambers(1986, 1995). Eles são simples, mesmo assim frequentemente usados pelos analistas de dados, pelo menos como uma primeira aproximação. Além disto, eles satisfazem todas as condições padrões de regularidade. Assim eles são adequados a uma aplicação de procedimentos de máxima pseudo-verossimilhança descritos na Seção 5.4.

As funções escores para  $\boldsymbol{\beta}$  e  $\sigma_e$  correspondentes ao modelo (6.1) podem ser facilmente obtidas como

$$\begin{aligned} \partial \log [f(y_i | \mathbf{z}_i; \boldsymbol{\beta}, \sigma_e)] / \partial \boldsymbol{\beta} &= \mathbf{z}_i (y_i - \mathbf{z}_i' \boldsymbol{\beta}) / \sigma_e \\ &\propto \mathbf{z}_i (y_i - \mathbf{z}_i' \boldsymbol{\beta}) = \mathbf{u}_i(\boldsymbol{\beta}) \end{aligned} \quad (6.2)$$

e

$$\begin{aligned} \partial \log [f(y_i | \mathbf{z}_i; \boldsymbol{\beta}, \sigma_e)] / \partial \sigma_e &= \left[ (y_i - \mathbf{z}_i' \boldsymbol{\beta})^2 - \sigma_e \right] / 2\sigma_e^2 \\ &\propto (y_i - \mathbf{z}_i' \boldsymbol{\beta})^2 - \sigma_e = u_i(\sigma_e) . \end{aligned} \quad (6.3)$$

### 6.1.2 Pseudo-parâmetros do Modelo

Se todos os elementos da população tivessem sido pesquisados, os EMVs de  $\boldsymbol{\beta}$  e  $\sigma_e$  do censo, denotados por  $\mathbf{B}$  e  $S_e$  respectivamente, poderiam ser facilmente obtidos como soluções das equações de verossimilhança do censo dadas por

$$\sum_{i \in U} \mathbf{u}_i(\mathbf{B}) = \sum_{i \in U} \mathbf{z}_i (y_i - \mathbf{z}_i' \boldsymbol{\beta}) = \mathbf{z}_U' \mathbf{y}_U - \left( \mathbf{z}_U' \mathbf{z}_U \right) \mathbf{B} = \mathbf{0} \quad (6.4)$$

e

$$\sum_{i \in U} u_i (S_e) = \sum_{i \in U} \left[ (y_i - \mathbf{z}'_i \mathbf{B})^2 - S_e \right] = (\mathbf{y}_U - \mathbf{z}'_U \mathbf{B})^0 (\mathbf{y}_U - \mathbf{z} \mathbf{z}'_U \mathbf{B}) - N S_e = 0 \quad (6.5)$$

onde  $\mathbf{z}_U = (\mathbf{z}_1, \dots, \mathbf{z}_N)^0$  e  $\mathbf{y}_U = (y_1, \dots, y_N)^0$ .

Se  $\mathbf{z}'_U \mathbf{z}_U$  for não-singular, as soluções para estas equações são facilmente obtidas como

$$\mathbf{B} = \left( \mathbf{z}'_U \mathbf{z}_U \right)^{-1} \mathbf{z}'_U \mathbf{y}_U \quad (6.6)$$

e

$$S_e = N^{-1} \sum_{i \in U} (y_i - \mathbf{z}'_i \mathbf{B})^2 = N^{-1} (\mathbf{y}_U - \mathbf{z}'_U \mathbf{B})^0 (\mathbf{y}_U - \mathbf{z}'_U \mathbf{B}) . \quad (6.7)$$

Com uma parametrização que isole o termo correspondente ao intercepto (primeira coluna do vetor  $\mathbf{z}_i$ ) do modelo de regressão (6.1), pode ser facilmente mostrado (Nascimento Silva, 1996, p. 142) que os EMV de  $\beta_2$  (igual a  $\beta$  excluído o primeiro componente),  $\beta_1$  e  $\sigma_e$  são dados respectivamente por

$$\mathbf{B}_2 = \mathbf{S}_z^{-1} \mathbf{S}_{zy} , \quad (6.8)$$

$$B_1 = \bar{Y} - \bar{\mathbf{Z}}^0 \mathbf{B}_2 , \quad (6.9)$$

e

$$S_e = N^{-1} \sum_{i \in U} \left( y_i - B_1 - \mathbf{z}'_i \mathbf{B}_2 \right)^2 = N^{-1} \sum_{i \in U} e_i^2 , \quad (6.10)$$

onde  $\bar{Y} = N^{-1} \sum_{i \in U} y_i$ ,  $\bar{\mathbf{Z}} = N^{-1} \sum_{i \in U} \mathbf{z}_i$ ,  $\mathbf{S}_z = N^{-1} \sum_{i \in U} (\mathbf{z}_i - \bar{\mathbf{Z}}) (\mathbf{z}_i - \bar{\mathbf{Z}})^0$ ,  $\mathbf{S}_{zy} = N^{-1} \sum_{i \in U} (\mathbf{z}_i - \bar{\mathbf{Z}}) (y_i - \bar{Y})$  e  $e_i = y_i - B_1 - \mathbf{z}'_i \mathbf{B}_2 = (y_i - \bar{Y}) - (\mathbf{z}_i - \bar{\mathbf{Z}})^0 \mathbf{B}_2$ , sendo neste trecho os vetores de variáveis preditoras tomados sem o termo constante referente ao intercepto.

Os EMVs do censo dados em (6.8) a (6.10) coincidem com os estimadores de mínimos quadrados ordinários, sob as hipóteses mais fracas do modelo dadas por (6.11) a seguir (ver Nathan e Holt, 1980), onde se dispensou a hipótese de normalidade dos erros, isto é

$$E_M(Y_i | \mathbf{z}_i = \mathbf{z}_i) = \beta_1 + \mathbf{z}'_i \beta_2 \quad (6.11)$$

$$V_M(Y_i | \mathbf{z}_i = \mathbf{z}_i) = \sigma_e$$

$$COV_M(Y_i, Y_j | \mathbf{z}_i = \mathbf{z}_i, \mathbf{z}_j = \mathbf{z}_j) = 0 \quad \forall i \neq j \in U.$$

## 6.1.3 Estimadores de MPV dos Parâmetros do Modelo

Quando apenas uma amostra de unidades da população é observada, são usados pesos  $w_i$  para obter estimadores de máxima pseudo-verossimilhança de  $\beta$  e  $\sigma_e$ , ou alternativamente de  $\mathbf{B}$  e  $S_e$ , se as quantidades descritivas populacionais correspondentes forem escolhidas para alvo da inferência. Se os pesos  $w_i$  satisfizerem às condições de regularidade discutidas na Seção 5.4, será imediato obter as equações de pseudo-verossimilhança correspondentes ao modelo (6.1) como

$$\begin{aligned} \sum_{i \in s} w_i \mathbf{u}_i(\hat{\mathbf{B}}_w) &= \sum_{i \in s} w_i \mathbf{z}_i (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_w) \\ &= \mathbf{z}_s^0 \mathbf{W}_s \mathbf{y}_s - (\mathbf{z}_s^0 \mathbf{W}_s \mathbf{y}_s) \hat{\mathbf{B}}_w = 0 \end{aligned} \quad (6.12)$$

e

$$\begin{aligned} \sum_{i \in s} w_i u_i(s_e^w) &= \sum_{i \in s} w_i \left[ (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_w)^2 - s_e^w \right] \\ &= (\mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_w)^0 \mathbf{W}_s (\mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_w) - (\mathbf{1}_s^0 \mathbf{W}_s \mathbf{1}_s) s_e^w = 0 \end{aligned} \quad (6.13)$$

onde  $\mathbf{z}_s$  e  $\mathbf{y}_s$  são os análogos amostrais de  $\mathbf{z}_U$  e  $\mathbf{y}_U$ , respectivamente,  $\mathbf{W}_s = \text{diag}[(w_{i_1}, \dots, w_{i_n})]$  é uma matriz diagonal  $n \times n$  com os pesos dos elementos da amostra na diagonal principal, e  $\hat{\mathbf{B}}_w$  e  $s_e^w$  são estimadores MPV de  $\beta$  e  $\sigma_e$  respectivamente.

Supondo que  $\mathbf{z}_s^0 \mathbf{W}_s \mathbf{z}_s$  é não-singular e resolvendo (6.12) e (6.13) em  $\hat{\mathbf{B}}_w$  e  $s_e^w$  obtemos as seguintes expressões para os estimadores MPV dos parâmetros do modelo:

$$\hat{\mathbf{B}}_w = (\mathbf{z}_s^0 \mathbf{W}_s \mathbf{z}_s)^{-1} \mathbf{z}_s^0 \mathbf{W}_s \mathbf{y}_s \quad (6.14)$$

e

$$\begin{aligned} s_e^w &= (\mathbf{1}_s^0 \mathbf{W}_s \mathbf{1}_s)^{-1} (\mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_w)^0 \mathbf{W}_s (\mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_w) \\ &= (\mathbf{1}_s^0 \mathbf{W}_s \mathbf{1}_s)^{-1} \mathbf{y}_s^0 \left[ \mathbf{W}_s - \mathbf{W}_s \mathbf{z}_s (\mathbf{z}_s^0 \mathbf{W}_s \mathbf{z}_s)^{-1} \mathbf{z}_s^0 \mathbf{W}_s \right] \mathbf{y}_s \end{aligned} \quad (6.15)$$

sendo a segunda expressão para  $s_e^w$  obtida mediante substituição do valor de  $\hat{\mathbf{B}}_w$  em (6.14) na primeira linha de (6.15).

Observe que a hipótese de não-singularidade de  $\mathbf{z}_s^0 \mathbf{W}_s \mathbf{z}_s$  não seria satisfeita se  $w_i = 0$  para algum  $i \in s$ . Para evitar que se percam de vista

as questões principais com relação à estimação dos parâmetros do modelo, admitiremos de agora em diante que  $\mathbf{z}_s^0 \mathbf{W}_s \mathbf{z}_s$  é não-singular.

Estimadores pontuais dos parâmetros do modelo podem ser derivados a partir de (6.14) e (6.15) para vários esquemas de ponderação de interesse pela simples substituição da matriz apropriada de ponderação  $\mathbf{W}_s$ . Se todos os elementos da pesquisa têm o mesmo peso (como no caso de planos amostrais autoponderados), ou seja,  $w_i = \bar{w}$  e  $\mathbf{W}_s = \bar{w} \mathbf{I}_n$ , os estimadores pontuais não dependem do valor  $\bar{w}$  dos pesos. Neste caso, eles ficam reduzidos às expressões correspondentes dos estimadores de mínimos quadrados ordinários (que são também estimadores de máxima verossimilhança sob normalidade) dos parâmetros do modelo, dados por:

$$\hat{\mathbf{B}} = \left( \mathbf{z}_s^0 \mathbf{z}_s \right)^{-1} \mathbf{z}_s^0 \mathbf{y}_s \quad (6.16)$$

e

$$s_e = n^{-1} \left( \mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}} \right)^0 \left( \mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}} \right). \quad (6.17)$$

Substituindo  $\mathbf{W}_s$  em (6.14) e (6.15) por  $\text{diag}(\pi_i : i \in s) = \mathbf{\Pi}_s^{-1}$ , onde os  $\pi_i$  em geral não são todos iguais, obtemos estimadores, chamados de mínimos quadrados  $\pi$ -ponderados, dados por:

$$\hat{\mathbf{B}}_\pi = \left( \mathbf{z}_s^0 \mathbf{\Pi}_s^{-1} \mathbf{z}_s \right)^{-1} \mathbf{z}_s^0 \mathbf{\Pi}_s^{-1} \mathbf{y}_s \quad (6.18)$$

e

$$s_e^\pi = \left( \mathbf{1}_s^0 \mathbf{\Pi}_s^{-1} \mathbf{1}_s \right)^{-1} \left( \mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_\pi \right)^0 \mathbf{\Pi}_s^{-1} \left( \mathbf{y}_s - \mathbf{z}_s \hat{\mathbf{B}}_\pi \right). \quad (6.19)$$

#### 6.1.4 Estimação da Variância de Estimadores de MPV

O exercício de ajustar um modelo não estará completo sem a avaliação da precisão e significância das estimativas dos parâmetros. Para isto é necessária a estimação das variâncias correspondentes. Nesta seção concentramos nossa atenção na estimação das variâncias dos estimadores de MPV dos coeficientes de regressão  $\boldsymbol{\beta}$ . As expressões a seguir são obtidas por aplicação direta dos resultados gerais fornecidos na Seção 5.4, observando-se que os escores correspondentes a  $\boldsymbol{\beta}$  no *ajuste do censo* do modelo (6.1) são dados por  $\mathbf{u}_i(\mathbf{B}) = \mathbf{z}_i (y_i - \mathbf{z}_i' \mathbf{B}) = \mathbf{z}_i e_i$ , onde  $e_i = (y_i - \bar{Y}) - (\mathbf{z}_i - \bar{\mathbf{Z}})' \mathbf{B}$  para  $i \in U$ , com o Jacobiano correspondente dado por

$$\begin{aligned} J(\mathbf{B}) &= \sum_{i \in U} \partial \mathbf{z}_i (y_i - \mathbf{z}_i' \boldsymbol{\beta}) / \partial \boldsymbol{\beta} \Big|_{\boldsymbol{\beta}=\mathbf{B}} \\ &= \partial (\mathbf{z}'_U \mathbf{y}_U - \mathbf{z}'_U \mathbf{z}_U \boldsymbol{\beta}) / \partial \boldsymbol{\beta} \Big|_{\boldsymbol{\beta}=\mathbf{B}} = -\mathbf{z}'_U \mathbf{z}_U. \end{aligned} \quad (6.20)$$

Substituindo em (5.8) e (5.9) os valores dos escores, do jacobiano e dos estimadores  $\pi$ -ponderados correspondentes, obtemos as seguintes expressões para a variância assintótica de aleatorização do estimador de MPV padrão  $\widehat{\mathbf{B}}_\pi$  e seu estimador consistente, dadas por

$$V_p(\widehat{\mathbf{B}}_\pi) = (\mathbf{z}'_U \mathbf{z}_U)^{-1} V_p \left( \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i e_i \right) (\mathbf{z}'_U \mathbf{z}_U)^{-1} \quad (6.21)$$

e

$$\widehat{V}_p(\widehat{\mathbf{B}}_\pi) = (\mathbf{z}'_s \mathbf{\Pi}_s^{-1} \mathbf{z}_s)^{-1} \widehat{V}_p \left( \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i e_i \right) (\mathbf{z}'_s \mathbf{\Pi}_s^{-1} \mathbf{z}_s)^{-1}, \quad (6.22)$$

onde

$$V_p \left( \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i e_i \right) = \sum_{i \in U} \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} e_i \mathbf{z}_i \mathbf{z}'_j e_j, \quad (6.23)$$

$$\widehat{V}_p \left( \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i e_i \right) = \sum_{i \in s} \sum_{j \in s} \left( \pi_i^{-1} \pi_j^{-1} - \pi_{ij}^{-1} \right) \widehat{e}_i \mathbf{z}_i \mathbf{z}'_j \widehat{e}_j, \quad (6.24)$$

e  $\widehat{e}_i = y_i - \mathbf{z}'_i \widehat{\mathbf{B}}_\pi$  para  $i \in s$ .

Isto completa a especificação de um procedimento de máxima pseudo-verossimilhança para ajustar modelos normais de regressão como (6.1). Este procedimento é bastante flexível e aplicável numa ampla gama de planos amostrais.

## 6.2 Modelo de Regressão Logística

No modelo de regressão logística, a variável resposta  $y$  é binária, isto é, assume os valores 0 e 1. Considerando um vetor  $\mathbf{z}$  de variáveis explanatórias tal como o empregado no modelo de regressão linear discutido na Seção 6.1, o modelo de superpopulação é dado por

$$f(y_i | \mathbf{z}_i, \boldsymbol{\beta}) = [p(\mathbf{z}'_i \boldsymbol{\beta})]^{y_i} [1 - p(\mathbf{z}'_i \boldsymbol{\beta})]^{1-y_i}, \quad (6.25)$$

onde,

$$p(\mathbf{z}'_i \boldsymbol{\beta}) = P(Y_i = 1 | \mathbf{Z}_i = \mathbf{z}_i) = \exp(\mathbf{z}'_i \boldsymbol{\beta}) / [1 + \exp(\mathbf{z}'_i \boldsymbol{\beta})].$$

A função escore de  $\boldsymbol{\beta}$  é

$$\mathbf{u}_i(\boldsymbol{\beta}) = \partial \log(y_i | \mathbf{z}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = [y_i - p(\mathbf{z}'_i \boldsymbol{\beta})] \mathbf{z}_i \quad (6.26)$$

e portanto a equação de verossimilhança do censo correspondente é dada por

$$\sum_{i \in U} \mathbf{u}_i(\boldsymbol{\beta}) = \sum_{i \in U} [y_i - p(\mathbf{z}'_i \boldsymbol{\beta})] \mathbf{z}_i = \mathbf{0}. \quad (6.27)$$

O estimador de MPV do vetor de coeficientes  $\boldsymbol{\beta}$  no modelo (6.25) é a solução da equação

$$\sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\beta}) = \sum_{i \in s} w_i [y_i - p(\mathbf{z}'_i \boldsymbol{\beta})] \mathbf{z}_i = \mathbf{0}, \quad (6.28)$$

onde  $w_i$  é o peso da  $i$ -ésima observação amostral.

A matriz de covariância do estimador de MPV de  $\boldsymbol{\beta}$  pode ser obtida conforme indicado na Seção 5.4, bastando substituir os valores dos escores  $\mathbf{u}_i(\boldsymbol{\beta}) = [y_i - p(\mathbf{z}'_i \boldsymbol{\beta})] \mathbf{z}_i$  e do jacobiano correspondentes. Para maiores detalhes, o leitor interessado pode consultar Binder(1983), que aborda o problema da estimação da matriz de covariância dos estimadores de MPV na família de modelos lineares generalizados, da qual o modelo de regressão logística é caso particular.

Vale observar que, tal como no caso da modelagem clássica, a obtenção dos estimadores de MPV dos parâmetros no modelo de regressão logística depende da solução por métodos numéricos de um sistema de equações. Portanto é importante dispor de um pacote computacional adequado para efetuar os cálculos. Hoje em dia já estão disponíveis vários pacotes com essa funcionalidade, conforme se discute no Capítulo 10.

**Exemplo 6.1** *Análise do perfil sócio-econômico das pessoas ocupadas no setor informal da economia na área urbana do Rio de Janeiro*

Utilizando dados do Suplemento Trabalho da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 90, Leote(1996) analisou o perfil sócio-econômico das pessoas ocupadas no setor informal da economia na área urbana do Rio de Janeiro.

Os dados utilizados são relativos a pessoas que:

- moravam em domicílios urbanos do estado do Rio de Janeiro;
- trabalhavam em atividades mercantis (não foram incluídos trabalhadores domésticos);
- na semana da pesquisa estavam trabalhando ou não estavam trabalhando por estarem de férias, licença, etc., mas tinham trabalho;
- desenvolviam atividades não agrícolas.

Tabela 6.1: Descrição das variáveis explicativas

Fatores	Níveis	Descrição dos níveis
Sexo (sx)	sx(1)	Homens
	sx(2)	Mulheres
Anos de estudo (ae)	ae(1)	Até 4
	ae(2)	De 5 a 8
	ae(3)	9 ou mais
Horas trabalhadas (ht)	ht(1)	Menos de 40
	ht(2)	De 40 a 48
	ht(3)	Mais de 48
Idade em anos completos (id)	id(1)	Até 17
	id(2)	De 18 a 25
	id(3)	De 26 a 49
	id(4)	50 ou mais
Rendimento médio mensal (re)	re(1)	Menos de 1
	re(2)	De 1 a 5
	re(3)	Mais de 5

As pessoas que trabalhavam em locais com até cinco pessoas ocupadas foram classificadas no setor informal, independente da posição de ocupação delas, enquanto as que trabalhavam em locais com mais de cinco pessoas ocupadas foram classificadas no setor formal. O trabalho refere-se ao trabalho principal. Para a variável renda considerou-se a soma dos rendimentos de todos os trabalhos.

Foi considerada uma amostra de 6.507 pessoas (após a exclusão de 9 registros considerados atípicos), classificadas de acordo com as variáveis descritas na Tabela 6.1, todas tratadas como fatores na análise. A variável ht foi considerada como a soma de horas trabalhadas em todos os trabalhos, por semana. A variável re compreende a renda média mensal de todos os trabalhos, em salários mínimos.

Os fatores considerados foram tomados como explicativos e a variável resposta foi o indicador de pertinência ao setor informal da economia. Foi ajustado um modelo logístico (Agresti, 1990) para explicar a probabilidade de uma pessoa pertencer ao setor informal da economia.

Para a seleção do modelo foi usada a função `glm` do `S-Plus`, aplicada aos dados tabelados. O modelo final selecionado foi escolhido passo a passo, incluindo em cada passo as interações que produziam maior decréscimo do desvio residual, considerando a perda de graus de liberdade. O modelo selecionado foi

$$\log\left(\frac{p_{ijklm}}{1-p_{ijklm}}\right) = \mu + \beta_i^{sx} + \beta_j^{ae} + \beta_k^{ht} + \beta_l^{id} + \beta_m^{re} \quad (6.29)$$

$$+ \beta_{ij}^{sx.id} + \beta_{ik}^{sx.ht} + \beta_{jk}^{ae.ht} + \beta_{kl}^{ht.id} + \beta_{km}^{ht.re},$$

onde  $p_{ijklm}$  é a probabilidade de pertencer ao setor informal correspondente à combinação de níveis das variáveis explicativas, sendo  $i=1, 2$  o nível de  $sx$ ;  $j=1, 2, 3$  o nível de  $ae$ ;  $k=1, 2, 3$  o nível de  $ht$ ;  $l=1, 2, 3, 4$  o nível de  $id$  e  $m=1, 2, 3$  o nível de  $re$ .

Os efeitos foram adicionados seqüencialmente na ordem da Tabela 6.1. Depois de introduzidos os efeitos principais, as interações de dois fatores foram introduzidas na ordem definida pela *função step do S-Plus*.

O  $p$ valor do teste de nulidade das interações não incluídas no modelo é 0,0515, aceitando-se a hipótese de nulidade destes efeitos ao nível  $\alpha = 0,05$ . O modelo obtido difere do selecionado em Leote(1996) só pela inclusão de mais um efeito, referente à interação  $ae:ht$ .

Uma descrição detalhada do plano amostral da PNAD 90 foi apresentada no Exemplo 5.1. Como se pode observar dessa descrição, o plano amostral da PNAD apresenta todos os aspectos de um plano amostral complexo, incluindo estratificação (geográfica), seleção de unidades primárias (municípios, ou setores nos municípios auto-representativos) ou secundárias (setores nos municípios não auto-representativos) com probabilidades desiguais, conglomeração (de domicílios em setores, e de pessoas nos domicílios) e seleção sistemática sem reposição de unidades. Nesse caso, fica difícil admitir a priori com confiança as hipóteses usuais de modelagem das observações amostrais como IID. Por esse motivo foram considerados métodos alternativos de modelagem e ajuste.

Apresentamos a seguir as estimativas dos efeitos principais e interações do modelo selecionado e seus respectivos desvios padrões, calculadas pela `PROC LOGISTIC` do pacote `SUDAAN`. Para facilitar a comparação incluímos na Tabela 6.2 os valores correspondentes estimados pelo `S-Plus`.

As estimativas calculadas pelo pacote `SUDAAN` são feitas pelo *Método de Máxima Pseudo-Verossimilhança*, resolvendo a equação (6.28). As estimativas dos desvios padrões são obtidas das variâncias calculadas pelo

método de linearização descrito na Seção 5.4, equação (5.9), considerando os escores tal como apresentados na equação (6.26). Para esses cálculos, os estimadores de variância considerados levaram em conta os pesos das observações, mas utilizaram uma aproximação que consiste em considerar que as unidades primárias de amostragem foram selecionadas com reposição, especificando a opção WR do pacote SUDAAN. Veja Shah et al.(1993, p. 4) e Wolter(1985, eq. 7.7.2).

Na Tabela 6.3 são apresentadas as probabilidades de significância dos testes de nulidade dos efeitos do modelo. Todos os efeitos incluídos no modelo são significativos, nos níveis usuais de significância. A PROC LOGISTIC do pacote SUDAAN não inclui testes para os efeitos principais, por não ser possível separar tais efeitos das interações. A coluna de *p*valores da Tabela 6.3, obtida pela PROC LOGISTIC do pacote SUDAAN, utiliza a estatística de Wald baseada no plano amostral com correção. Mais detalhes são encontrados em Shah et al.(1993).

Os testes da Tabela 6.3 indicam a significância de todas as interações de 2 fatores que entraram no modelo selecionado. O teste de qualidade global de ajuste, na primeira linha da Tabela 6.3, indica a necessidade de serem introduzidas novas interações.

Para comparação, apresentamos na Tabela 6.4 algumas estimativas de razões de vantagens, relevantes na análise, calculadas tanto pela função glm do S-Plus como pela PROC LOGISTIC do pacote SUDAAN e, na Tabela 6.5, os correspondentes intervalos de confiança de 95%. Na construção destes intervalos foi necessário utilizar estimativas pontuais dos efeitos bem como a matriz de covariância estimada dos estimadores dos efeitos do modelo. Deste modo, estes intervalos sumarizam, ao mesmo tempo, discrepâncias existentes tanto nas estimativas pontuais dos efeitos como nas variâncias e covariâncias das estimativas.

Além dos ajustes aqui comparados, foram feitos (embora não apresentados) os seguintes ajustes com a utilização do S-Plus: 1) dados individuais (resposta 0-1) considerando os pesos; 2) dados da tabela estimada considerando os pesos e 3) dados individuais com pesos normalizados. Em todas

Tabela 6.2: Estimativas dos efeitos e dos respectivos desvios padrões obtidas pelo SUDAAN e pelo S-Plus

Variáveis independentes e efeitos	Ajuste no SUDAAN		Ajuste no S-Plus	
	Estimativa do efeito	Desvio Padrão	Estimativa do efeito	Desvio Padrão
Intercepto	-0,515	0,260	-0,514	0,269
sx	0,148	0,222	0,156	0,228
ae1	0,745	0,165	0,740	0,165
ae2	0,496	0,156	0,497	0,159
ht1	-0,377	0,317	-0,386	0,312
ht2	-0,697	0,275	-0,698	0,268
id1	-0,239	0,540	-0,243	0,492
id2	-0,729	0,302	-0,724	0,314
id3	0,227	0,231	0,227	0,234
re1	0,286	0,277	0,293	0,245
re2	0,065	0,144	0,062	0,145
ht1.re1	1,529	0,356	1,531	0,332
ht2.re1	0,338	0,320	0,336	0,284
ht1.re2	0,490	0,233	0,498	0,221
ht2.re2	-0,115	0,183	-0,112	0,178
ht1.id1	-1,420	0,605	-1,408	0,515
ht2.id1	-0,413	0,506	-0,397	0,465
ht1.id2	-0,124	0,354	-0,129	0,351
ht2.id2	-0,109	0,279	-0,106	0,286
ht1.id3	-0,220	0,248	-0,216	0,253
ht2.id3	-0,537	0,205	-0,533	0,201
sx.id1	0,878	0,348	0,870	0,335
sx.id2	0,300	0,231	0,294	0,226
sx.id3	-0,259	0,190	-0,263	0,186
sx.ht1	-0,736	0,206	-0,737	0,211
sx.ht2	-0,089	0,185	-0,093	0,182
ae1.ht1	0,792	0,240	0,792	0,239
ae2.ht1	0,739	0,227	0,735	0,226
ae1.ht2	0,026	0,197	0,029	0,196
ae2.ht2	0,089	0,183	0,087	0,189

Tabela 6.3: Testes de hipóteses de nulidade dos efeitos do modelo

Contraste	Graus de liberdade	Graus de liberdade ajustados	Estatística F ajustada	<i>p</i> valor da estatística F ajustada
Modelo Global	30	26,132	37,510	0,000
Bondade do ajuste	29	25,692	28,179	0,000
ht:re	4	3,946	6,040	0,000
ht:id	6	5,764	4,110	0,001
sx:id	3	2,969	7,168	0,000
sx:ht	2	1,993	9,166	0,000
ae:ht	4	3,959	4,814	0,001

Tabela 6.4: Estimativas das razões de vantagens, variando-se os níveis de ae para níveis fixos de ht

ht	Mudança de nível de ae	S-Plus	SUDAAN
1	1 para 2	0,741	0,739
1	2 para 3	0,291	0,291
2	1 para 2	0,831	0,830
2	2 para 3	0,558	0,557
3	1 para 2	0,785	0,780
3	2 para 3	0,608	0,608

Tabela 6.5: Intervalos de confiança de 95% para razões de vantagens, variando-se os níveis de ae para níveis fixos de ht

ht	Mudança de nível de ae	S-Plus	SUDAAN
1	1 para 2	(0,530; 1,036).	(0,516; 1,059)
1	2 para 3	(0,213; 0,399)	(0,212; 0,398)
2	1 para 2	(0,697; 0,991)	(0,693; 0,994)
2	2 para 3	(0,457; 0,680)	(0,452; 0,687)
3	1 para 2	(0,586; 1,050)	(0,577; 1,053)
3	2 para 3	(0,445; 0,831)	(0,448; 0,827)

Tabela 6.6: Distribuição de frequências dos pesos da amostra da PNAD-90 - Parte Urbana do Rio de Janeiro

Valor do peso	Frequência
674	127
675	784
711	3288
712	712

estas análises, como esperado, as estimativas pontuais dos efeitos coincidiram com as obtidas pela PROC LOGISTIC do pacote SUDAAN. Pode-se notar que, neste exemplo, há estreita concordância entre as estimativas pontuais obtidas pelos dois pacotes.

A concordância das estimativas dos coeficientes pode ser explicada, em parte, pela pequena variabilidade dos pesos das unidades, tal como se pode verificar na Tabela 6.6, que apresenta a distribuição de frequências dos pesos.

Como foi visto na Tabela 6.2, o impacto do plano amostral nas estimativas de precisão é um pouco maior. As maiores diferenças entre os dois métodos ocorrem na estimação dos desvios das estimativas dos parâmetros do primeiro nível de idade (até 17 anos) e da interação deste com horas trabalhadas (tanto no nível de menos de 40 horas semanais como no nível de 40 a 48 horas semanais trabalhadas). Esta diferenciação maior no caso dos desvios padrões já era esperada. Quando não levamos em conta os pesos nem o plano amostral na estimação dos parâmetros, podemos até chegar em uma estimativa pontual dos coeficientes bem próxima de quando levamos ambos em conta, mas as estimativas dos desvios padrões são mais sensíveis a esta diferença entre as análises. A tendência revelada é de subestimação dos desvios padrões pelo S-Plus ao ignorar o plano amostral e a variação dos pesos.

Neste exemplo, foi utilizada a função glm do S-Plus na seleção do modelo. Feita a seleção, o mesmo modelo foi ajustado através da PROC LOGISTIC do SUDAAN. O propósito foi imitar uma situação onde o modelo já tivesse sido selecionado e ajustado por usuário secundário dos dados, sem considerar os pesos e o plano amostral, tal como é usual. Outra possibilidade seria repetir o processo de seleção do modelo usando-se a PROC LOGISTIC do SUDAAN. Isto poderia ser feito passo a passo, incluindo

efeitos e interações que melhorassem mais a qualidade de ajuste, tal como foi feito automaticamente pela função `step` do `S-Plus`. Este procedimento possibilitaria comparar a seleção de modelos quando são considerados os pesos e o plano amostral na análise.

Diferentemente dos pacotes mais usados de análise estatística, tais como `SAS`, `S-Plus`, `BMDP`, etc., o `SUDAAN` não possui, atualmente, ferramentas usuais de diagnóstico de ajuste de modelos, como gráficos de resíduos padronizados, etc., tornando mais difícil seu uso na etapa de seleção de modelos. Considerando-se a maior dificuldade de seleção de modelos através do `SUDAAN`, preferiu-se usá-lo aqui apenas para ajustar um modelo já selecionado.

### 6.3 Teste de Hipóteses

Nas seções 6.1 e 6.2 discutimos formas de introduzir pesos e plano amostral em procedimentos de estimação pontual e de variâncias ao ajustar modelos com dados de pesquisas amostrais complexas. Neste contexto, procedimentos estatísticos de teste de hipóteses devem, também, sofrer adaptações. Nesta seção, esse problema será abordado de forma sucinta, para modelos de regressão.

De modo geral, testes de hipóteses em regressão surgem inicialmente na seleção de modelos e também para fornecer evidência favorável ou contrária a indagações levantadas pelo pesquisador.

Denotemos por  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$  o vetor de parâmetros num modelo de regressão. Como é sabido, para testar a hipótese  $H_0 : \beta_j = 0$ , para algum  $j \in \{1, \dots, P\}$ , usamos um teste  $t$ , e para testar a hipótese  $H_0 : (\beta_{j_1}, \dots, \beta_{j_R})' = \mathbf{0}_R$ , onde  $(j_1, \dots, j_R) \subset (1, \dots, P)$  e  $\mathbf{0}_R$  é o vetor zero  $R$ -dimensional, usamos um teste  $\mathbf{F}$ . Tais testes  $t$  e  $\mathbf{F}$ , sob as hipóteses do modelo clássico de regressão com erros normais, são testes da Razão de Máxima Verossimilhança.

É pois natural tentar adaptar testes de Razão de Máxima Verossimilhança para pesquisas amostrais complexas, tal como foi feito na derivação de estimadores de MPV a partir de estimadores de Máxima Verossimilhança. A principal dificuldade é que no contexto de pesquisas complexas, devido aos pesos distintos das observações e ao plano amostral utilizado, a função de verossimilhança usual não representa a distribuição conjunta das observações. Apesar desta dificuldade ter sido contornada na derivação de estimadores de MPV, a adaptação fica bem mais difícil no caso de testes da Razão de Máxima Verossimilhança.

Por essa causa, é mais fácil basear os testes na estatística Wald, que mede a distância entre uma estimativa pontual e o valor hipotetizado do parâmetro numa métrica definida pela matriz de covariância do estimador. Pesos e plano amostral podem ser incorporados facilmente nessa estatística, bastando para isto utilizar estimativas apropriadas (consistentes sob aleatorização) dos parâmetros e da matriz de covariância, tais como as que são geradas pelo método de MPV. É essa abordagem que vamos adotar aqui.

Considere o problema de testar a hipótese linear geral

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}, \quad (6.30)$$

onde  $\mathbf{C}$  é uma matriz de dimensão  $R \times P$  de posto pleno  $R = P - Q$  e  $\mathbf{c}$  é um vetor  $R \times 1$ .

Um caso particular de interesse é testar a hipótese aninhada  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}_R$ , onde  $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)$ , com  $\boldsymbol{\beta}_1$  de dimensão  $Q \times 1$  e  $\boldsymbol{\beta}_2$  de dimensão  $R \times 1$ ,  $\mathbf{C} = \begin{bmatrix} \mathbf{0}_{R \times Q} & \mathbf{I}_R \end{bmatrix}$  e  $\mathbf{c} = \mathbf{0}_R$ , sendo  $\mathbf{0}_{R \times Q}$  matriz de zeros de dimensão  $R \times Q$  e  $\mathbf{I}_R$  a matriz identidade de ordem  $R$ .

A estatística de Wald clássica para testar a hipótese nula (6.30) é definida por

$$X_W^2 = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{c})' (\mathbf{C}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{C}')^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{c}), \quad (6.31)$$

onde os estimadores  $\hat{\boldsymbol{\beta}}$  e  $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$  são obtidos pela teoria de mínimos quadrados ordinários. Sob  $H_0$ , a distribuição assintótica da estatística  $X_W^2$  é  $\chi^2(R)$ .

Quando os dados são obtidos através de pesquisas amostrais complexas, a estatística  $X_W^2$  deixa de ter distribuição assintótica  $\chi^2(R)$ , e usar esta última como distribuição de referência implica na obtenção de testes com níveis de significância incorretos. Esse problema é solucionado substituindo-se na expressão de  $X_W^2$ ,  $\hat{\boldsymbol{\beta}}$  pela estimativa MPV  $\hat{\mathbf{B}}_\pi$  de  $\boldsymbol{\beta}$  dada em (6.18), e  $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$  pela estimativa da matriz de covariância do estimador de MPV  $\hat{\mathbf{V}}_p(\hat{\mathbf{B}}_\pi)$  dada em (6.22). Tais estimativas consideram os pesos diferentes das observações e o plano amostral efetivamente utilizado. A normalidade assintótica do estimador de MPV de  $\boldsymbol{\beta}$  e a consistência do estimador da matriz de covariância correspondente (Binder, 1983) implicam que

$$X_W^2 \sim \chi^2(R), \text{ sob } H_0.$$

Esta aproximação não leva em conta o erro amostral na estimação de  $\mathbf{V}(\hat{\boldsymbol{\beta}})$ . Uma alternativa é usar a aproximação

$$X_W^2/R \sim \mathbf{F}(R, v),$$

onde  $v = m - H$  é o número de UPAs da amostra menos o número de estratos considerados no plano amostral para seleção das UPAs, que fornece uma medida de graus de liberdade apropriada para amostras complexas quando o método do conglomerado primário é empregado para estimar variâncias.

Com a finalidade de melhorar a aproximação da distribuição da estatística de teste, podem ser utilizados ajustes e correções da estatística  $X_W^2$ , que são apresentados com mais detalhes nos Capítulos 7 e 8 para o caso da análise de dados categóricos.

A especificação de um procedimento para testar hipóteses sobre os parâmetros de um modelo de regressão completa a abordagem para ajuste de modelos desse tipo partindo de dados amostrais complexos. Entretanto, uma das partes importantes da teoria clássica para modelagem é a que trata do diagnóstico dos modelos ajustados, muitas vezes empregando recursos gráficos. Nessa parte a abordagem baseada em MPV e em estatísticas de Wald deixa a desejar, pois não é possível adaptar de maneira simples as técnicas clássicas de diagnóstico. Por exemplo, é difícil considerar pesos ao plotar os resíduos do ajuste dum modelo via MPV. Essa é questão que ainda merece maior investigação e por enquanto é uma desvantagem da abordagem aqui preconizada.

## Capítulo 7

# Testes de Qualidade de Ajuste

### 7.1 Introdução

Tabelas de distribuições de frequências ocorrem comumente na análise de dados de pesquisas complexas. Tais tabelas são formadas pela classificação e cálculo de frequências dos dados da amostra disponível segundo níveis de uma variável categórica - tabelas de uma entrada - ou segundo celas de uma classificação cruzada de duas (ou mais) variáveis categóricas - tabelas de duas (ou mais) entradas. Neste capítulo concentraremos a atenção em tabelas de uma entrada, ou equivalentemente nas frequências absolutas e relativas (ou proporções) correspondentes.

Em muitos casos, o objetivo da análise é testar hipóteses de bondade de ajuste de modelos para descrever essas distribuições de frequências. Sob a hipótese de observações IID (distribuição Multinomial) ou equivalentemente, de amostragem aleatória simples, inferências válidas para testar tais hipóteses podem ser baseadas na estatística padrão de teste qui-quadrado de Pearson. Tais testes podem ser facilmente executados usando procedimentos prontos em pacotes estatísticos padrões tais como o SAS, S-Plus, SPSS, GLIM e outros.

No caso de planos amostrais complexos, entretanto, os procedimentos de teste precisam ser ajustados devido aos efeitos de conglomeração, estratificação e/ou pesos desiguais. Neste capítulo examinaremos o impacto do plano amostral sobre as estatísticas de teste usuais notando que, em alguns casos, os valores observados dessas estatísticas de teste podem ser muito grandes, acarretando inferências incorretas, conforme já ilustrado no Exem-

plo 4.4 do Capítulo 4. Isto ocorre porque a probabilidade de erros do tipo I (rejeitar a hipótese nula quando esta é verdadeira) é muito maior que o nível nominal de significância  $\alpha$  especificado.

Para obter inferências válidas usando amostras complexas podemos introduzir correções na estatística de teste de Pearson, tais como os ajustes de Rao-Scott, ou alternativamente usar outras estatísticas de teste que já incorporem o plano amostral, tais como a estatística de Wald. Os dois enfoques serão ilustrados através de um exemplo introdutório simples de teste de bondade de ajuste. Os resultados discutidos neste capítulo são adequados tanto para uma abordagem de aleatorização, em que os parâmetros se referem à população finita em questão, quanto para uma abordagem baseada em modelos, em que os parâmetros especificam algum modelo de superpopulação.

## 7.2 Teste para uma Proporção

### 7.2.1 Correção de Estatísticas Clássicas

No Exemplo 4.4 do Capítulo 4 a estatística de teste  $Z_{bin}$ , que foi utilizada para comparar com um valor hipotético pré-fixado a proporção de empregados cobertos por plano de saúde, resultou num teste mais liberal do que o teste que empregou a estatística  $Z_p$ , baseada no plano amostral efetivamente adotado. A causa disto foi o fato de  $Z_{bin}$  não considerar o efeito de conglomeração existente. Vamos examinar com mais detalhes o comportamento assintótico da estatística de teste  $Z_{bin}$ , construindo a estatística de teste  $X_p^2$  de Pearson para o exemplo correspondente. Para isto, consideremos a Tabela 7.1 contendo a distribuição de frequências, onde  $n_j$  e  $p_{oj}$  são as frequências (absolutas) observadas na amostra e as proporções hipotéticas nas categorias de interesse, respectivamente.

As proporções populacionais desconhecidas nas categorias são  $p_j = N_j/N$ , onde  $N$  é o tamanho total da população de empregados e  $N_j$  é o número de elementos da população na categoria  $j$ ,  $j = 1, 2$ . Os parâmetros populacionais  $p_j$  poderiam também ser considerados como pseudo-parâmetros, se vistos como estimativas de censo para as probabilidades desconhecidas ( $\pi_j$ , digamos) no contexto de um modelo de superpopulação.

Tabela 7.1: Freqüências observadas e proporções hipotéticas

Categoria	$j$	$n_j$	$p_{0j}$
Cobertos por plano de saúde	1	840	0,8
Não cobertos	2	160	0,2
Todos os empregados	$\sum$	1.000	1,0

A estatística de teste de Pearson para a hipótese simples de bondade de ajuste  $H_0 : p_j = p_{0j}$ ,  $j = 1, 2$ , é dada por

$$X_P^2 = \sum_{j=1}^2 (n_j - n p_{0j})^2 / (n p_{0j}) = n \sum_{j=1}^2 (\hat{p}_j - p_{0j})^2 / p_{0j}, \quad (7.1)$$

onde as proporções  $\hat{p}_j = n_j/n$  são estimativas amostrais usuais das proporções populacionais  $p_j$ , para  $j = 1, 2$ .

Como há apenas duas categorias e as proporções devem somar 1, observa-se que  $p_2 = 1 - p_1$ ,  $\hat{p}_2 = 1 - \hat{p}_1$  e  $p_{02} = 1 - p_{01}$ . Isto acarreta na equivalência entre as estatísticas  $Z_{bin}$  e  $X_P^2$  demonstrada pela relação

$$X_P^2 = n \sum_{j=1}^2 (\hat{p}_j - p_{0j})^2 / p_{0j} = \frac{(\hat{p} - p_0)^2}{p_0(1 - p_0)/n} = Z_{bin}^2 \quad (7.2)$$

onde  $\hat{p} = \hat{p}_1$  e  $p_0 = p_{01}$  para simplicidade e coerência com a notação do Exemplo 4.4.

Sob a hipótese de observações IID, a distribuição assintótica da estatística  $X_P^2$  é qui-quadrado ( $\chi^2$ ). Neste caso, em que há apenas duas categorias e uma restrição (soma das proporções igual a 1), a distribuição da estatística  $X_P^2$  em (7.2) tem apenas um grau de liberdade.

Rao e Scott(1981) obtiveram resultados gerais para a distribuição assintótica da estatística de teste  $X_P^2$  de Pearson sob planos amostrais complexos. Com apenas duas celas, a distribuição assintótica da estatística de teste  $X_P^2$  é a distribuição da variável aleatória  $dW$ , onde  $W$  tem distribuição  $\chi^2(1)$  (qui-quadrado com um grau de liberdade) e  $d$  é o efeito de plano amostral (EPA) da estimativa  $\hat{p}$  da proporção  $p$ . O efeito de plano amostral nesse caso é dado por  $d = V_p(\hat{p})/V_{bin}(\hat{p})$ .

Para uma amostra de empregados selecionada por amostragem aleatória simples, teríamos  $d = 1$  pois  $V_p(\hat{p})$  e  $V_{bin}(\hat{p})$  seriam iguais. Neste caso,

a estatística  $X_P^2$  de teste seria assintoticamente  $\chi^2(1)$ . Como a amostra foi efetivamente selecionada por amostragem de conglomerados, devido à correlação intraclasse positiva o efeito de plano amostral  $d$  é maior que um, e portanto a distribuição assintótica da estatística de teste  $X_P^2$  não é mais  $\chi^2(1)$ .

Considerando que o impacto da correlação intraclasse positiva na distribuição assintótica da estatística  $X_P^2$  de Pearson pode levar a inferências incorretas caso se utilize a distribuição assintótica usual, o próximo passo é derivar um procedimento de teste válido. Isto é feito introduzindo uma correção em  $X_P^2$ . Para isto, observe que a esperança assintótica de  $X_P^2$  é  $E_p(X_P^2) = d$ . Como  $E_p(X_P^2/d) = E(\chi^2(1)) = 1$ , obtemos então a correção simples de Rao-Scott para  $X_P^2$  dividindo o valor observado da estatística de teste pelo efeito do plano amostral  $d$ , isto é,

$$X_P^2(d) = X_P^2/d, \quad (7.3)$$

que tem, no caso de duas celas, distribuição assintótica  $\chi^2(1)$ .

Outra estatística comumente usada para testar a mesma hipótese de bondade de ajuste no caso de proporções é a estatística do teste da Razão de Verossimilhança (RV), dada por

$$X_{RV}^2 = 2n \sum_{j=1}^2 \hat{p}_j \log(\hat{p}_j/p_{0j}) = 2n \log\left(\frac{\hat{p}(1-\hat{p})}{p_0(1-p_0)}\right). \quad (7.4)$$

No caso de amostragem aleatória simples, a estatística  $X_{RV}^2$  é também distribuída assintoticamente como  $\chi^2(1)$ , quando a hipótese nula é verdadeira. Para planos amostrais complexos, a estatística corrigida correspondente é

$$X_{RV}^2(d) = X_{RV}^2/d. \quad (7.5)$$

Vamos calcular os valores das estatísticas de Pearson e de RV, com suas correções de Rao-Scott, para os dados do Exemplo 4.4. Para as correções, primeiro é preciso calcular o efeito do plano amostral

$$\begin{aligned} d &= V_p(\hat{p})/V_{bin}(\hat{p}) = \frac{p(1-p)/m}{p(1-p)/n} \\ &= \frac{0,0032}{0,00016} = 20 \end{aligned}$$

onde  $m = 50$  é o número de empregados por empresa (tamanho do conglomerado) e  $n = 1.000$  é o número de empregados na amostra.

O valor da estatística de teste de Pearson é

$$X_P^2 = \frac{(0,84 - 0,80)^2}{(0,80 \times 0,20)/1.000} = 10$$

com *p*valor 0,0016. O valor da estatística de teste de Pearson com a correção de Rao-Scott  $X_P^2(d)$  é então dado por

$$X_P^2(d) = X_P^2/d = 10/20 = 0,5$$

com *p*valor 0,4795. Observe que  $Z_p^2 = 0,707^2 = 0,5$ , e também que  $X_P^2(d) = Z_{bin}^2/d = 3,162^2/20 = 0,5$  ou seja,  $Z_p^2 = X_P^2(d)$  conforme esperado. Os valores da estatística do teste da Razão de Verossimilhança e sua correção de Rao-Scott são dados respectivamente por

$$X_{RV}^2 = 2 \times 1.000 \times \log \left( \frac{0,84 \times 0,16}{0,80 \times 0,20} \right) = 10,56,$$

com *p*valor 0,0012, e

$$X_{RV}^2(d) = X_{LR}^2/d = 10,56/20 = 0,528,$$

com *p*valor de 0,4675.

Como se pode notar, as estatísticas baseadas na Razão de Verossimilhança oferecem resultados semelhantes às versões correspondentes baseadas na estatística de Pearson. Em ambos os casos, as decisões baseadas nas estatísticas sem correção seriam incorretas no sentido de rejeitar a hipótese nula. Também em ambos os casos a correção de Rao-Scott produziu efeito semelhante.

O efeito de plano amostral  $d = 20$  observado neste exemplo é muito grande e pouco comum na prática. Isto ocorreu neste caso porque o coeficiente de correlação intraclasse assume o valor máximo  $\rho = 1$  (todos os valores dentro de um conglomerado são iguais, e portanto a homogeneidade é máxima). Na prática, as correlações intraclasse observadas são usualmente positivas mas menores que um, e portanto as estimativas de efeito de plano amostral  $\hat{d}$  correspondentes são maiores que um. Para conglomerados de tamanho médio igual a 20 ( $m = 20$ ) como neste exemplo, os valores típicos de  $\hat{d}$  são menores que 3, tendo em correspondência correlações intraclasse estimadas positivas  $\hat{\rho} < 0,1$ .

Os resultados do exemplo discutido nesta seção ilustram bem a importância de considerar o plano amostral na construção de estatísticas de teste para proporções simples, embora num caso um tanto extremo. Ilustram também um dos enfoques existentes para tratar do problema, a saber a correção de estatísticas de teste usuais (de Pearson e da Razão de Verossimilhança).

### 7.2.2 Estatística de Wald

Como alternativa à estatística de teste de Pearson, podemos usar a estatística de bondade de ajuste  $X_N^2$  de Neyman. No caso de duas celas, ela se reduz a

$$X_N^2 = n \sum_{j=1}^2 (\hat{p}_j - p_{0j})^2 / \hat{p}_j = \frac{(\hat{p} - p_0)^2}{\hat{p}(1 - \hat{p})/n} . \quad (7.6)$$

Note que a expressão de  $X_N^2$  em (7.6) pode ser obtida substituindo-se no denominador de  $X_P^2$  em (7.2) a proporção hipotética  $p_0$  pela proporção estimada  $\hat{p}$ .

A estatística de Neyman é um caso particular da estatística de bondade de ajuste de Wald. Esta última estatística difere das estatísticas de Pearson, da Razão de Verossimilhança e de Neyman por incorporar automaticamente o plano amostral. Para o caso de duas celas, ela se reduz a

$$X_W^2 = (\hat{p} - p_0)^2 / \hat{V}_p(\hat{p}), \quad (7.7)$$

onde  $\hat{V}_p(\hat{p})$  é uma estimativa da variância de aleatorização de  $\hat{p}$ , correspondente ao plano amostral efetivamente utilizado.

O efeito do termo  $\hat{V}_p(\hat{p})$ , que aparece no denominador de  $X_W^2$ , é incorporar na estatística de bondade de ajuste o efeito do plano amostral utilizado. No caso particular de amostragem aleatória simples, usamos no lugar de  $\hat{V}_p(\hat{p})$  a variância  $\hat{V}_{bin}(\hat{p}) = \hat{p}(1 - \hat{p})/n$ . Neste caso, estatística resultante  $X_{bin}^2$  coincide com a estatística  $X_N^2$  de Neyman.

Para o plano amostral de conglomerados considerado no Exemplo 4.4, a estatística  $X_W^2$ , sem qualquer ajuste auxiliar, já é distribuída assintoticamente como qui-quadrado com um grau de liberdade. O valor da estatística de Wald para esse exemplo é

$$X_W^2 = (0,84 - 0,80)^2 / 0,002743 = 0,583 .$$

Observe que o valor desta estatística é bem próximo dos valores das estatísticas de Pearson e da Razão de Verossimilhança com a correção de Rao-Scott.

A estatística de Wald, pelo uso de uma estimativa apropriada da variância, reflete a complexidade do plano amostral e fornece uma estatística de teste assintoticamente válida, não necessitando que seja feito qualquer ajuste auxiliar. Esta pode ser considerada uma vantagem em relação às estatísticas com correção de Rao-Scott. Entretanto, no caso de mais de duas celas, pode haver desvantagens no uso da estatística de Wald baseada no plano amostral, devido à instabilidade nas estimativas de variância em pequenas amostras.

Tabela 7.2: Valores e  $p$ valores de estatísticas de teste alternativas para os dados do Exemplo 4.4

Estatística de teste	gl	Valor observado	$p$ valor
Pearson $X_P^2$	1	10,00	0,0016
Pearson ajustada $X_P^2(d)$	1	0,500	0,4795
RV $X_{RV}^2$	1	10,56	0,0012
RV ajustada $X_{RV}^2(d)$	1	0,528	0,4675
Wald $X_W^2$	1	0,583	0,4451

Reproduzimos na Tabela 7.2 os resultados para todas as estatísticas de teste consideradas até agora, para facilidade de comparação.

Nesta seção foram apresentadas as duas principais abordagens para incorporar o efeito do plano amostral na estatística de teste:

1. a metodologia de ajuste de Rao-Scott para as estatísticas de teste de Pearson e da Razão de Verossimilhança;
2. e a estatística de Wald baseada no plano amostral.

Ambas as abordagens são facilmente generalizáveis para tabelas de uma ou duas entradas com número de linhas e colunas maior que dois. Vamos considerar na próxima seção o caso geral de testes de bondade de ajuste e apresentar mais detalhes sobre as estatísticas de teste alternativas. Depois, introduziremos os testes de independência e de homogeneidade para tabelas de duas entradas. A ênfase será dada nos procedimentos baseados na estatísticas de teste de Wald baseadas no plano amostral e nas estatísticas de Pearson e da RV com os vários ajustes de Rao-Scott.

### 7.3 Teste para Várias Proporções

Neste seção vamos considerar extensões do problema de testes de bondade de ajuste, aumentando o número de proporções envolvidas. O caso de tabelas de duas entradas será considerado no capítulo seguinte.

A hipótese de bondade de ajuste para  $J \geq 2$  celas pode ser escrita como  $H_0 : p_j = p_{0j}$ ,  $j = 1, \dots, J$ , onde  $p_j = N_j/N$  são as proporções

populacionais desconhecidas nas celas e  $p_{0j}$  são as proporções hipotéticas das celas. Essa hipótese pode também ser escrita, usando notação vetorial, como  $H_0 : \mathbf{p} = \mathbf{p}_0$ , onde  $\mathbf{p} = (p_1, \dots, p_{J-1})^0$  é o vetor de proporções populacionais desconhecidas e  $\mathbf{p}_0 = (p_{01}, \dots, p_{0, J-1})^0$  é o vetor de proporções hipotéticas.

O vetor de estimativas consistentes das proporções das celas, baseado em  $n$  observações, é denotado por  $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{J-1})^0$ , onde  $\hat{p}_j = \hat{n}_j/n$ . Os  $\hat{n}_j$  são as frequências ponderadas nas celas, considerando as diferentes probabilidades de inclusão dos elementos e ajustes por não-resposta, onde os pesos amostrais são normalizados de modo que  $\sum_{j=1}^J \hat{n}_j = n$ . Se  $n$  não for fixado de antemão, os  $\hat{p}$  serão estimadores de razões, o que é comum quando trabalhamos com subgrupos da população. Observe que apenas  $J - 1$  componentes são incluídos em cada um dos vetores  $\mathbf{p}$ ,  $\mathbf{p}_0$  e  $\hat{\mathbf{p}}$ , pois a soma das proporções nas  $J$  categorias é igual a 1, e portanto a proporção na  $J$ -ésima categoria é obtida por diferença.

### 7.3.1 Estatística de Wald Baseada no Plano Amostral

A estatística de Wald baseada no plano amostral  $X_W^2$ , para o teste da hipótese simples de bondade de ajuste, foi anteriormente introduzida no caso de duas celas como uma alternativa à estatística de Pearson ajustada. No caso de mais de duas celas, a estatística de bondade de ajuste de Wald é dada por

$$X_W^2 = (\hat{\mathbf{p}} - \mathbf{p}_0)^0 \hat{\mathbf{V}}_p^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0), \quad (7.8)$$

onde  $\hat{\mathbf{V}}_p$  denota um estimador consistente da matriz de covariância de aleatorização verdadeira  $\mathbf{V}_p$  do estimador  $\hat{\mathbf{p}}$  do vetor de proporções  $\mathbf{p}$ . Uma estimativa  $\hat{\mathbf{V}}_p$  pode ser obtida pelo método de linearização, usando-se por exemplo o pacote SUDAAN.

Sob a hipótese nula  $H_0$ , a estatística  $X_W^2$  tem distribuição assintótica qui-quadrado com  $J - 1$  graus de liberdade, fornecendo assim um procedimento de teste válido no caso de amostras complexas. Na prática, espera-se que  $X_W^2$  funcione adequadamente se o número de unidades primárias de amostragem selecionadas for grande e o número de celas componentes do vetor  $\mathbf{p}$  for relativamente pequeno. Neste caso, podemos obter um estimador estável de  $\mathbf{V}_p$ . Observe que (7.7) é um caso particular de (7.8).

### 7.3.2 Situações Instáveis

Se o número  $m$  de unidades primárias de amostragem disponíveis for pequeno, pode ocorrer um problema de instabilidade na estimativa  $\hat{\mathbf{V}}_p$ , devido

ao pequeno número de graus de liberdade  $f = m - H$  disponível para a estimação da variância. A instabilidade da estimativa  $\widehat{\mathbf{V}}_p$  pode tornar a estatística de Wald muito liberal.

É comum contornar esta instabilidade corrigindo a estatística de Wald, mediante emprego da chamada *estatística de Wald F-corrigida*. Há duas propostas alternativas de estatísticas F-corrigidas de Wald. A primeira é dada por

$$F_{1,p} = \frac{f - J + 2}{f(J - 1)} X_W^2, \quad (7.9)$$

que tem distribuição assintótica de referência  $F$  com  $J - 1$  e  $f - J + 2$  graus de liberdade. A segunda é dada por

$$F_{2,p} = \frac{X_W^2}{(J - 1)}, \quad (7.10)$$

que tem distribuição assintótica de referência  $F$  com  $J - 1$  e  $f$  graus de liberdade. No caso  $J = 2$ , as duas correções reproduzem a estatística original.

O efeito de uma correção  $\mathbf{F}$  à estatística  $X_W^2$  pode ser visualizado facilmente no caso de duas celas. Se  $f$  for pequeno, então o  $p$ valor de  $X_W^2$ , obtido a partir de uma distribuição  $F$  com 1 e  $f$  graus, é maior que o  $p$ valor obtido numa distribuição qui-quadrado com um grau de liberdade. Quando  $f$  aumenta a diferença diminui, tornando a correção desprezível, quando  $f$  for grande.

Thomas e Rao(1987) analisaram o desempenho das diferentes estatísticas de teste de bondade de ajuste, no caso de instabilidade. Eles verificaram que a estatística de Wald F-corrigida  $F_{1,p}$  não apresentou, em geral, o melhor desempenho nesta comparação, contudo, comportou-se relativamente bem nos casos padrões, onde a instabilidade não era muito grave. As estatísticas F-corrigidas de Wald são bastante utilizadas na prática, e estão implementadas em pacotes para análise de dados de pesquisas amostrais complexas.

### 7.3.3 Estatística de Pearson com Ajuste de Rao-Scott

O exemplo introdutório serviu para mostrar que, na presença de efeitos de plano amostral importantes, as estatísticas clássicas de teste precisam ser ajustadas para terem a mesma distribuição assintótica de referência que a obtida para o caso de amostragem aleatória simples. Inicialmente, vamos considerar a estatística de teste  $X_P^2$  de Pearson. Essa estatística pode ser escrita em forma matricial como

$$X_P^2 = n \sum_{j=1}^J (\widehat{p}_j - p_{0j})^2 / p_{0j} = n (\widehat{\mathbf{p}} - \mathbf{p}_0)^0 \mathbf{P}_0^{-1} (\widehat{\mathbf{p}} - \mathbf{p}_0) \quad (7.11)$$

onde  $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^0$  e  $\mathbf{P}_0/n$  é a matriz  $(J-1) \times (J-1)$  de covariância multinomial de  $\hat{\mathbf{p}}$  sob a hipótese nula, e  $\text{diag}(\mathbf{p}_0)$  representa uma matriz diagonal com elementos  $p_{0j}$  na diagonal.

A matriz de covariância  $\mathbf{P}_0/n$  é uma generalização do caso  $J = 2$  celas para o caso de mais de duas celas ( $J > 2$ ). Observe que a expressão de  $X_P^2$  tem a mesma forma da estatística de Wald, com  $\mathbf{P}_0/n$  no lugar de  $\hat{\mathbf{V}}_p$ . No caso de apenas duas celas,  $X_P^2$  reduz-se à fórmula simples antes considerada  $X_P^2 = (\hat{p}_1 - p_{01})^2 / [p_{01}(1 - p_{01})/n]$ , onde o denominador corresponde à variância da binomial sob a hipótese nula.

Para examinar a distribuição assintótica da estatística  $X_P^2$  de Pearson, vamos generalizar os resultados anteriores, do caso de duas celas para o caso  $J > 2$ . Neste caso,  $X_P^2$  é assintoticamente distribuído como uma soma ponderada  $\delta_1 W_1 + \delta_2 W_2 + \dots + \delta_{J-1} W_{J-1}$  de  $J-1$  variáveis aleatórias independentes  $W_j$ , cada uma tendo distribuição qui-quadrado com um grau de liberdade. Os pesos  $\delta_j$  são os autovalores da matriz de efeito multivariado de plano amostral  $\mathbf{\Delta} = \mathbf{P}_0^{-1}\mathbf{V}_p$ , onde  $\mathbf{V}_p/n$  é a matriz de covariância do estimador  $\hat{\mathbf{p}}$  do vetor de proporção  $\mathbf{p}$  baseada no plano amostral verdadeiro. Tais autovalores são também chamados efeitos generalizados de plano amostral. Observe que, em geral, eles não coincidem com os efeitos univariados de plano amostral  $d_j$ .

No caso de amostragem aleatória simples, os efeitos generalizados de plano amostral  $\delta_j$  são todos iguais a um, pois neste caso  $\mathbf{\Delta} = \mathbf{I}$ , matriz identidade. Neste caso, a soma  $\sum_{j=1}^{J-1} \delta_j W_j$  se reduz a  $\sum_{j=1}^{J-1} W_j$ , cuja distribuição é  $\chi^2$  com  $J-1$  graus de liberdade. Assim, sob amostragem aleatória simples, a estatística  $X_P^2$  é distribuída assintoticamente como qui-quadrado com  $J-1$  graus de liberdade.

No caso de plano amostral mais complexo, envolvendo estratificação e/ou conglomeração, os efeitos generalizados de plano amostral não são iguais a um. Devido aos efeitos de conglomeração, os  $\delta_j$  tendem a ser maiores que um, e assim a distribuição assintótica da variável aleatória  $\sum_{j=1}^{J-1} \delta_j W_j$  diferente de uma qui-quadrado com  $J-1$  graus de liberdade. Desta forma, a estatística  $X_P^2$  requer correções semelhantes às introduzidas no caso de duas celas. No caso geral, há mais de uma possibilidade de correção e consideraremos as *correções de primeira ordem e de segunda ordem de Rao-Scott*, desenvolvidas por Rao e Scott(1981). A correção de primeira ordem tem por objetivo corrigir a esperança assintótica da estatística  $X_P^2$  de Pearson, e a de segunda ordem também envolve correção da variância. Tecnicamente, os dois ajustes são baseados nos autovalores da matriz de efeito multivariado de plano amostral estimada  $\hat{\mathbf{\Delta}}$ .

Inicialmente, consideramos um *ajuste simples de EPA médio* à estatística  $X_P^2$ , devido a Fellegi (1980) e Holt, Scott e Ewings(1980), e o ajuste de primeira ordem de Rao-Scott. Estes ajustes são úteis nos casos em que não é possível obter uma estimativa adequada  $\hat{V}_p$  para a matriz de covariância de aleatorização. Quando esta estimativa está disponível, deve-se usar o ajuste mais preciso de segunda ordem.

O ajuste de EPA médio é baseado nos efeitos univariados de plano amostral estimados  $\hat{d}_j$  das estimativas  $\hat{p}_j$ . O ajuste da estatística (7.11) é feito dividindo o valor observado da estatística  $X_P^2$  de Pearson pela média  $\hat{d}$  dos efeitos univariados de plano amostral:

$$X_P^2(\hat{d}) = X_P^2/\hat{d}. \quad (7.12)$$

onde  $\hat{d} = \sum_{j=1}^J \hat{d}_j/J$  é um estimador da média  $\bar{d}$  dos efeitos de plano amostral desconhecidos.

Estimamos os efeitos do plano amostral por  $\hat{d}_j = \hat{V}_p(\hat{p}_j) / (\hat{p}_j(1 - \hat{p}_j)/n)$ , onde  $\hat{V}_p(\hat{p}_j)$  é a estimativa da variância de aleatorização do estimador de proporção  $\hat{p}_j$ . Este ajustamento requer que estejam disponíveis as estimativas dos efeitos de plano amostral dos estimadores das proporções das  $J$  celas. A correlação intraclasse positiva fornece uma média  $\hat{d}$  maior que 1 e, portanto, o ajuste do EPA médio tende a remover a liberalidade de  $X_P^2$ .

O ajuste do EPA médio não corrige exatamente a esperança assintótica de  $X_P^2$ , pois a média dos efeitos univariados de plano amostral não é igual à média dos efeitos generalizados de plano amostral. Sob a hipótese nula, a esperança assintótica de  $X_P^2$  é  $E(X_P^2) = \sum_{j=1}^{J-1} \delta_j$ , logo  $E(X_P^2/\bar{\delta}) = E(\chi^2(J-1)) = J-1$ , onde a média dos autovalores é  $\bar{\delta} = \sum_{j=1}^{J-1} \delta_j / (J-1)$ . Este raciocínio conduz ao ajuste de primeira ordem de Rao-Scott para  $X_P^2$ , dado por

$$X_P^2(\hat{\delta}) = X_P^2/\hat{\delta}, \quad (7.13)$$

onde  $\hat{\delta}$  é um estimador da média  $\bar{\delta}$  dos autovalores desconhecidos da matriz de efeitos multivariados de plano amostral  $\Delta$ .

Podemos estimar a média dos efeitos generalizados usando os efeitos univariados de plano amostral estimados, pela equação

$$(J-1)\hat{\delta} = \sum_{j=1}^J \frac{\hat{p}_j}{p_{0j}} (1 - \hat{p}_{0j}) \hat{d}_j,$$

sem estimar os próprios autovalores. Alternativamente,  $\hat{\delta}$  pode ser obtido a partir da estimativa da matriz de efeitos multivariados  $\hat{\Delta} = n\mathbf{P}_0^{-1}\hat{\mathbf{V}}_p$ , pela equação  $\hat{\delta} = \text{tr}(\hat{\Delta}) / (J - 1)$ , isto é, dividindo o traço de  $\hat{\Delta}$  pelo número de graus de liberdade.

A estatística ajustada  $X_P^2(\hat{\delta})$  só tem distribuição assintoticamente qui-quadrado com  $(J - 1)$  graus de liberdade se os autovalores forem iguais. Na prática, esta estatística funciona bem se a variação dos autovalores estimados for pequena. No cálculo de  $X_P^2(\hat{\delta})$  só são necessários os efeitos multivariados de plano amostral dos  $\hat{p}_j$  que aparecem na diagonal da matriz  $\hat{\Delta}$ . Assim, esta estatística é adequada em análises secundárias de tabelas de contingência, se forem divulgadas as estimativas de efeito de plano amostral correspondentes. O ajuste de primeira ordem de Rao-Scott  $X_P^2(\hat{\delta})$  é mais exato do que o ajuste do EPA médio da estatística  $X_P^2(\hat{d})$ , que é considerada uma alternativa conservadora de  $X_P^2(\hat{\delta})$ .

A correção de primeira ordem de Rao-Scott (7.13) é introduzida na estatística de Pearson com o objetivo de tornar a média assintótica da estatística ajustada igual ao número de graus de liberdade da distribuição de referência. Se a variação dos autovalores estimados  $\hat{\delta}_j$  for grande, então será também necessária uma correção da variância de  $X_P^2$ . Isto é obtido através de uma *correção de segunda ordem de Rao-Scott*, baseada no método de Satterthwaite(1946). A estatística de Pearson com ajuste de Rao-Scott de segunda ordem é dada por

$$X_P^2(\hat{\delta}, \hat{a}^2) = X_P^2(\hat{\delta}) / (1 + \hat{a}^2), \quad (7.14)$$

onde  $\hat{a}^2$  é um estimador do quadrado do coeficiente de variação  $a^2$  dos autovalores desconhecidos dado por

$$\hat{a}^2 = \sum_{j=1}^{J-1} \hat{\delta}_j^2 / \left( (J-1) \hat{\delta}^2 \right) - 1 .$$

Um estimador da soma dos quadrados dos autovalores é dado por

$$\sum_{j=1}^{J-1} \hat{\delta}_j^2 = \text{tr}(\hat{\Delta}^2) = n^2 \sum_{j=1}^J \sum_{k=1}^J \hat{V}_p^2(\hat{p}_j, \hat{p}_k) / p_{0j} p_{0k} ,$$

onde  $\hat{V}_p(\hat{p}_j, \hat{p}_k)$  são os estimadores das covariâncias de aleatorização de  $\hat{p}_j$  e  $\hat{p}_k$ . Os graus de liberdade também devem ser corrigidos. A estatística

$X_P^2(\hat{\delta}, \hat{a}^2)$  é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite dados por  $gl_S = (J - 1) / (1 + \hat{a}^2)$ .

Observe que, para o ajuste de segunda ordem, é necessária estimativa completa da matriz de variância  $\hat{\mathbf{V}}_p$ , enquanto que para o ajuste de primeira ordem só precisamos conhecer estimativas das variâncias  $\hat{V}_p$ .

Em situações instáveis, pode ser necessário fazer uma correção F ao ajuste de primeira ordem de Rao-Scott (7.13). A estatística F-corrigida é definida por

$$FX_P^2(\hat{\delta}) = X_W^2 / ((J - 1) \hat{\delta}) \quad . \quad (7.15)$$

A estatística  $FX_P^2(\hat{\delta})$  tem distribuição de referência  $F$  com  $J - 1$  e  $f$  graus de liberdade. Thomas e Rao(1987) observaram que esta estatística, em situações instáveis, é melhor que a estatística sem correção de primeira ordem.

**Exemplo 7.1** *Teste de bondade de ajuste para a distribuição etária da PPV 96-97 na Região Sudeste.*

Vamos considerar um teste da bondade de ajuste da distribuição das idades para a Pesquisa sobre Padrões de Vida (PPV) 96/97, para os subgrupos de 0 a 14; de 15 a 29; de 30 a 44; de 45 a 59 e de 60 e mais anos de idade. As proporções correspondentes para a população foram obtidas da Contagem Populacional de 96. Na Região Sudeste, o número de estratos é  $H = 15$  e o número total de conglomerados (setores) na amostra da PPV é  $m = 276$  e portanto  $f = m - H = 261$ . As informações utilizadas neste exemplo são apresentadas na Tabela 7.3.

Os valores dos EPAs observados na PPV (coluna 5 da Tabela 7.3) mostram que o plano amostral não pode ser ignorado na análise. Queremos testar a hipótese  $H_0 : \mathbf{p} = \mathbf{p}_0$  usando as estimativas de proporções obtidas pela amostra da PPV. O vetor de proporções populacionais  $\mathbf{p}_0$  foi obtido dos resultados da Contagem Populacional de 96, que é uma pesquisa censitária. Neste exemplo, vamos calcular a estatística de Pearson e suas correções, e também a estatística de Wald baseada no plano amostral. Calculamos a matriz  $\hat{\mathbf{V}}_p$  pela aplicação do método de linearização de Taylor descrito na

Tabela 7.3: Vetores de proporções por classes de idade da PPV 96/97 e Contagem 96 e EPAs calculados para a PPV - Região Sudeste

Idade	Contagem	PPV		
	Proporção ( $\mathbf{p}_0$ )	Frequência	Proporção ( $\hat{\mathbf{p}}$ )	EPA
0 a 14 anos	0,2842	2.516	0,2845	2,2859
15 a 29 anos	0,2774	2.360	0,2678	2,1864
30 a 44 anos	0,2263	2.018	0,2225	2,2539
45 a 59 anos	0,1261	1.177	0,1316	1,9903
60 anos e mais	0,0860	832	0,0935	3,1627
Total	1,0000	8903	0,9999	—

Seção 3.3 através da fórmula (3.22) obtendo

$$\hat{\mathbf{V}}_p(\hat{\mathbf{p}}) = 10^{-6} \times \left[ \begin{array}{cccc|c} 52,27 & -3,899 & -5,672 & -19,29 & -23,41 \\ -3,899 & 48,16 & -29,35 & -3,399 & -11,52 \\ -5,672 & -29,35 & 43,80 & -8,226 & -0,5558 \\ -19,29 & -3,399 & -8,226 & 25,55 & 5,366 \\ \hline -23,41 & -11,52 & -0,5558 & 5,366 & 30,12 \end{array} \right].$$

Para obter a estatística de Pearson (7.11), vamos calcular a matriz de covariância populacional e uma estimativa dessa matriz de covariância sob suposição de distribuição multinomial, dada por  $\mathbf{P}_0/n = \frac{\text{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^0}{8.903}$ , resultando em

$$\mathbf{P}_0/n = 10^{-6} \times \left[ \begin{array}{cccc|c} 22,85 & -8,855 & -7,223 & -4,025 & -2,745 \\ -8,855 & 22,51 & -7,051 & -3,929 & -2,680 \\ -7,223 & -7,051 & 19,67 & -3,205 & -2,186 \\ -4,025 & -3,929 & -3,205 & 12,38 & -1,218 \\ \hline -2,745 & -2,680 & -2,186 & -1,218 & 8,829 \end{array} \right].$$

Para obter os diversos ajustes desta estatística precisamos usar os valores dos EPAs, listados na coluna 5 da Tabela 7.3. Estes valores foram obtidos através do pacote SUDAAN. Para obter as diferentes correções da estatística de Pearson, precisamos calcular as seguintes quantidades:

$$\hat{d} = \sum_{j=1}^5 \hat{d}_j/5 = 2,376 ,$$

$$\hat{\delta} = \sum_{j=1}^5 \frac{\hat{p}_j}{4p_{0j}} (1 - \hat{p}_{0j}) \hat{d}_j = 2,457 \text{ ,}$$

$$1 + \hat{a}^2 = 8903^2 \sum_{j=1}^5 \sum_{k=1}^5 \left( \hat{V}_p^2(\hat{p}_j, \hat{p}_k) / p_{0j} p_{0k} \right) / (4 \times 2,457^2) = 1,253 \text{ .}$$

Podemos então calcular a estatística  $X_P^2$  de Pearson usando (7.11), resultando em

$$X_P^2 = 11,64$$

com 4 g.l. e um *p*valor 0,020 .

A estatística de Pearson com ajustamento de EPA médio é calculada usando (7.12), resultando em

$$X_P^2(\hat{d}) = 11,64/2,376 = 4,901$$

com 4 g.l. e um *p*valor 0,298 .

A estatística de Pearson com ajustamento de Rao-Scott de primeira ordem, dada por (7.13), resulta em

$$X_P^2(\hat{\delta}) = 11,64/2,457 = 4,74$$

com 4 g.l. e um *p*valor 0,315 .

O ajustamento de Rao-Scott de primeira ordem F-corrigido para a estatística de Pearson, dado por (7.15), resulta em

$$FX_P^2(\hat{\delta}) = 4,74/4 = 1,185$$

com 4 e 261 g.l e um *p*valor 0,318 .

O ajustamento de Rao-Scott de segunda ordem para a estatística de Pearson, dado por (7.14), resulta em

$$X_P^2(\hat{\delta}, \hat{a}^2) = 4,74/1,253 = 3,784$$

com  $4/1,253 = 3,19$  g.l. e *p*valor 0,314 .

A estatística de Wald baseada no plano amostral (veja equação 7.8) resulta em

$$X_W^2 = 5,691$$

com 4 g.l. e um *p*valor 0,223 .

Tabela 7.4: Valores e  $p$ valores de estatísticas alternativas de teste

Estatística	Tipo	Valor	Distribuição	$p$ valor
$X_P^2$	Adequada para IID	11,640	$\chi_4^2$	0,020
$X_P^2(\hat{d})$	Ajustes e correções da	4,901	$\chi_4^2$	0,298
$X_P^2(\hat{\delta})$		4,740	$\chi_4^2$	0,315
$FX_P^2(\hat{\delta})$	Estatística	1,850	$F_{4,261}$	0,318
$X_P^2(\hat{\delta}, \hat{a}^2)$	$X_P^2$	3,784	$\chi_{3,19}^2$	0,314
$X_W^2$	Baseadas no plano amostral	5,691	$\chi_4^2$	0,223
$F_{1,p}$		1,406	$F_{4,259}$	0,232
$F_{2,p}$		1,423	$F_{4,261}$	0,228

As estatísticas F-corrigidas de Wald, definidas em (7.9) e (7.10), resultam em

$$F_{1,p} = \frac{261 - 5 + 2}{261 \times 4} \times 5,690661 = 1,406$$

com 4 e 259 g.l. e um  $p$ valor 0,232, e

$$F_{2,p} = 5,691/4 = 1,423$$

com 4 e 261 gl e um  $p$ valor 0,228.

A Tabela 7.4 resume os valores das diversas estatísticas de teste calculadas, bem como das informações comparativas com as respectivas distribuições de referência.

Examinando os resultados da Tabela 7.4, verificamos que o teste clássico de Pearson rejeita a hipótese nula  $H_0$  no nível  $\alpha = 5\%$ , diferentemente de todos os outros testes. Os valores das estatísticas com ajustes de Rao-Scott (com ou sem correção F) são semelhantes e parecem corrigir exageradamente o  $p$ valor dos testes. A estatística de Wald baseada no plano amostral e suas correções F, que têm valores quase iguais, produzem uma correção menor no  $p$ valor do teste. Nesse exemplo, como o número de graus de liberdade (dado pelo número de unidades primárias na amostra menos o número de estratos)  $f = m - H = 261$  é grande, a correção F tem pouco efeito, tanto nas estatísticas com ajustes de primeira e segunda ordem de Rao-Scott, como na estatística Wald.

## Capítulo 8

# Testes em Tabelas de Duas entradas

### 8.1 Introdução

Os principais testes em tabelas de duas entradas são os de homogeneidade e de independência. O *teste de homogeneidade* é apropriado para estudar a igualdade das distribuições condicionais de uma variável resposta categórica correspondentes a diferentes níveis de uma variável preditora também categórica. O *teste de independência* é adequado para estudar a associação entre duas variáveis categóricas. Enquanto o primeiro teste se refere às distribuições condicionais da variável resposta para níveis fixados da variável preditora, o segundo se refere à distribuição conjunta das duas variáveis categóricas que definem as celas da tabela. Apesar de conceitualmente distintas, as duas hipóteses podem ser testadas, no caso de amostragem aleatória simples, utilizando a mesma estatística de teste multinomial de Pearson.

Nos testes de homogeneidade e de independência para tabelas de frequências  $L \times C$  obtidas por amostragem aleatória simples, a estatística de teste de Pearson tem distribuição assintótica qui-quadrado com  $(L - 1)(C - 1)$  graus de liberdade, isto é  $\chi^2((L - 1)(C - 1))$ . Para pesquisas com planos amostrais complexos, esta propriedade assintótica padrão não é válida. Por exemplo, testes definidos em tabelas de frequências obtidas mediante amostragem por conglomerados são mais liberais (rejeitam mais) relativamente aos níveis nominais de significância, devido à correlação intraclasses positiva das variáveis usadas para definir a tabela. Além disso, para planos amostrais complexos, as estatísticas de teste das duas hipóteses devem ser corrigidas de formas diferentes.

Tabela 8.1: Tabela 2x2 de proporções

Var 1	Var 2		
	1	2	Total
1	$p_{11}$	$p_{12}$	$p_{1+}$
2	$p_{21}$	$p_{22}$	$p_{2+}$
Total	$p_{+1}$	$p_{+2}$	1

Neste capítulo, apresentamos versões modificadas de procedimentos clássicos de testes para dados categóricos, de maneira a incorporar os efeitos de plano amostral na análise. Procedimentos mais recentes, baseados em ajustes de modelos regressivos, estão disponíveis em pacotes especializados como o SUDAAN (procedimento CATAN, para dados tabelados, e procedimento LOGISTIC, para regressão com respostas individuais binárias, por exemplo), porém não serão aqui considerados.

## 8.2 Tabelas 2x2

Para fixar idéias, vamos considerar inicialmente uma tabela de contingência  $2 \times 2$ , isto é, com  $L = 2$  e  $C = 2$ , representada pela Tabela 8.1. A entrada  $p_{lc}$  na Tabela 8.1 representa a proporção populacional de unidades no nível  $l$  da variável 1 e  $c$  da variável 2, ou seja  $p_{lc} = \frac{N_{lc}}{N}$ , onde  $N_{lc}$  é o número de observações na cela  $(l, c)$  na população,  $N$  é o tamanho da população e  $\sum_l \sum_c p_{lc} = 1$ . Vamos denotar, ainda, as proporções marginais na tabela por  $p_{l+} = \sum_c p_{lc}$  e  $p_{+c} = \sum_l p_{lc}$ .

### 8.2.1 Teste de Independência

A hipótese de independência corresponde a

$$H_0 : p_{lc} = p_{l+}p_{+c} \quad \forall l, c = 1, 2.$$

A estatística de teste de Pearson para testar esta hipótese, no caso de amostragem aleatória simples, é dada por

$$X_P^2(I) = n \sum_{l=1}^2 \sum_{c=1}^2 \frac{(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c})^2}{\hat{p}_{l+}\hat{p}_{+c}}$$

onde  $\hat{p}_{lc} = n_{lc}/n$ ,  $n_{lc}$  é o número de observações da amostra na cela  $(l, c)$  da tabela,  $n$  é o tamanho total da amostra,  $\hat{p}_{l+} = \sum_c \hat{p}_{lc}$  e  $\hat{p}_{+c} = \sum_l \hat{p}_{lc}$ .

Sob a hipótese nula, a estatística  $X_P^2(I)$  tem distribuição de referência qui-quadrado com um grau de liberdade. Observe que esta estatística mede uma distância (em certa escala) entre os valores observados na amostra e os valores esperados (estimados) sob a hipótese nula de independência.

### 8.2.2 Teste de Homogeneidade

No caso do teste de independência, as duas variáveis envolvidas são consideradas como respostas. No teste de homogeneidade, uma das variáveis, a variável 2, por exemplo, é considerada a resposta enquanto a variável 1 é considerada explicativa. Vamos agora analisar a distribuição da variável 2 (coluna) para cada nível da variável 1 (linha). Considerando ainda uma tabela  $2 \times 2$ , queremos testar a hipótese

$$H_0 : p_{1c} = p_{2c} \quad c = 1, 2 \quad .$$

onde agora  $p_{lc}$  representa a proporção na linha  $l$  de unidades na coluna  $c$ . Com as restrições usuais de que as proporções nas linhas somam 1, isto é,  $p_{11} + p_{12} = p_{21} + p_{22} = 1$ , a hipótese nula considerada se reduz a  $p_{11} = p_{21}$  e novamente temos apenas um grau de liberdade.

Para o teste de homogeneidade, usamos a seguinte estatística de teste de Pearson:

$$X_P^2(H) = \sum_{l=1}^2 \sum_{c=1}^2 \frac{n_{l+} (\hat{p}_{lc} - \hat{p}_{+c})^2}{\hat{p}_{+c}} \quad ,$$

onde  $n_{l+} = \sum_c n_{lc}$  para  $l = 1, 2$  e  $\hat{p}_{lc} = n_{lc}/n_{l+}$  para  $l = 1, 2$  e  $c = 1, 2$ .

Esta estatística mede a distância entre valores observados e esperados sob a hipótese nula de homogeneidade e tem, também, distribuição de referência qui-quadrado com um grau de liberdade. Embora as expressões de  $X_P^2(I)$  e  $X_P^2(H)$  sejam distintas, seus valores numéricos são iguais.

### 8.2.3 Efeitos de Plano Amostral nas Celas

Para relacionar os testes tratados neste capítulo com o teste de qualidade de ajuste apresentado no capítulo anterior, observe que os testes de independência e de homogeneidade são definidos sobre o vetor de proporções de distribuições multinomiais. No caso de independência, temos uma distribuição multinomial com vetor de probabilidades  $(p_{11}, p_{12}, p_{21}, p_{22})$ , e no caso do teste de homogeneidade, temos duas multinomiais (no caso binomiais) com

vetores de probabilidades  $(p_{11}, p_{12})$  e  $(p_{21}, p_{22})$ . O processo de contagem que gera estas multinomias pressupõe que as observações individuais (indicadores de classe) são independentes e com mesma distribuição. Estas hipóteses só são válidas no caso de amostragem aleatória simples com reposição.

Quando os dados são gerados através de um plano amostral complexo, surgem efeitos de conglomeração e estratificação que devem ser considerados no cálculo das estatísticas de teste. Neste caso, as frequências nas celas da tabela são estimadas, levando em conta os pesos dos elementos da amostra bem como o plano amostral efetivamente utilizado.

Denotemos por  $\hat{N}_{lc}$  o estimador do número de observações na cela  $(l, c)$  na população, e designemos por  $\hat{n}_{lc} = \left(\hat{N}_{lc}/\hat{N}\right) \times n$  o valor padronizado de  $\hat{N}_{lc}$ ,

de modo que  $\sum_{l=1}^L \sum_{c=1}^C \hat{n}_{lc} = n$ . Sejam, agora, os estimadores das proporções nas celas dados por  $\hat{p}_{lc} = \hat{n}_{lc}/n$  no caso do teste de independência e por  $\hat{p}_{lc} = \hat{n}_{lc}/n_{l+}$  no caso do teste de homogeneidade. As estatísticas  $X_P^2(I)$  e  $X_P^2(H)$  calculadas com as estimativas  $\hat{n}_{lc}$  no lugar dos valores  $n_{lc}$  não têm, como antes, distribuição assintótica qui-quadrado com um grau de liberdade.

Por outro lado, é importante observar que as agências produtoras de dados estatísticos geralmente apresentam os resultados de suas pesquisas em tabelas contendo as estimativas  $\hat{N}_{lc}$ , como ilustrado no Exemplo 5.2 do Capítulo 5. Se calcularmos as estatísticas  $X_P^2(I)$  e  $X_P^2(H)$  a partir dos valores dos  $\hat{N}_{lc}$  fornecidos, com a estimativa do tamanho da população  $\hat{N}$  no lugar de  $n$ , os resultados assintóticos obtidos para amostragem aleatória simples com reposição (IID) deixarão de ser válidos. Devemos calcular as estatísticas de teste  $X_P^2(I)$  e  $X_P^2(H)$  a partir dos  $\hat{n}_{lc}$  anteriormente definidos, que correspondem aos  $\hat{N}_{lc}$  padronizados para totalizar  $n$ .

As estatísticas baseadas nos valores estimados  $\hat{n}_{lc}$  podem ser corrigidas para ter distribuição de referência qui-quadrado com um grau de liberdade, no caso de tabela  $2 \times 2$ . Mas, é importante observar que os efeitos de plano amostral e as correções a serem considerados são distintos para as duas estatísticas  $X_P^2(I)$  e  $X_P^2(H)$ .

Para ilustrar esse ponto vamos considerar o *ajuste de EPA médio*, que será apresentado na próxima seção para o caso de tabelas  $L \times C$ . Este ajuste, no caso da estatística  $X_P^2(I)$ , se baseia no EPA médio das estimativas das proporções nas celas  $\hat{p}_{lc} = \hat{n}_{lc}/n$ , enquanto que para a estatística  $X_P^2(H)$  ele se baseia no EPA médio das estimativas das proporções nas linhas  $\hat{p}_{lc} = \hat{n}_{lc}/n_{l+}$ .

Os valores das estatísticas  $X_P^2(I)$  e  $X_P^2(H)$  são iguais no caso IID, mas para planos amostrais complexos, as estatísticas corrigidas pelo EPA médio

Tabela 8.2: Proporções de linhas em tabela  $L \times C$ 

Região	1	2	...	$c$	...	$C$	Total
1	$p_{11}$	$p_{12}$	...	$p_{1c}$	...	$p_{1C}$	1
2	$p_{21}$	$p_{22}$	...	$p_{2c}$	...	$p_{2C}$	1
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
$l$	$p_{l1}$	$p_{l2}$	...	$p_{lc}$	...	$p_{lC}$	1
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
$L$	$p_{L1}$	$p_{L2}$	...	$p_{Lc}$	...	$p_{LC}$	1

são distintas, apesar de terem, para tabelas  $2 \times 2$ , a mesma distribuição de referência qui-quadrado com um grau de liberdade. Adiante apresentaremos um exemplo numérico para ilustrar este ponto.

### 8.3 Tabelas de Duas Entradas (Caso Geral)

#### 8.3.1 Teste de Homogeneidade

O teste de homogeneidade pode ser usado para comparar distribuições de uma variável categórica ( $C$  categorias) para um conjunto de  $L$  regiões não superpostas, a partir de amostras independentes obtidas através de um plano amostral com vários estágios. Vamos considerar uma tabela  $L \times C$  e supor que as colunas da tabela correspondem às classes da variável resposta e as linhas correspondem às regiões, de modo que as somas das proporções nas linhas na tabela de proporções são iguais a 1. A tabela para a população é da forma da Tabela 8.2

Note que aqui as proporções que aparecem nas linhas da tabela são proporções calculadas em relação à frequência total da linha, e não proporções calculadas em relação ao total da tabela como na seção anterior. Portanto,  $p_{lc} = N_{lc}/N_{l+}$  para todo  $l = 1, \dots, L$  e  $c = 1, \dots, C$ .

Vamos considerar o caso em que  $L = 2$  regiões devem ser comparadas. Seja  $\mathbf{p}_l = (p_{l1}, \dots, p_{l,C-1})'$  o vetor de proporções da  $l$ -ésima região, sem incluir a proporção referente à última categoria ( $p_{lC}$ ),  $l = 1, 2$ . A hipótese de igualdade das distribuições da resposta nas duas regiões pode ser expressa

como  $H_0 : \mathbf{p}_1 = \mathbf{p}_2$ , com  $C - 1$  componentes em cada vetor, pois em cada região a soma das proporções é 1.

Seja  $\mathbf{p}_0 = (p_{+1}, \dots, p_{+C-1})'$  o vetor comum de proporções sob  $H_0$ , desconhecido. Denotemos por  $\hat{\mathbf{p}}_l = (\hat{p}_{l1}, \dots, \hat{p}_{lC-1})'$  os vetores de proporções estimadas ( $l = 1, 2$ ), baseados em amostras independentes para as diferentes regiões, onde  $\hat{p}_{lc} = \hat{N}_{lc} / \hat{N}_{l+}$  é um estimador consistente da proporção  $p_{lc}$  na população correspondente, e  $\hat{N}_{lc}$  e  $\hat{N}_{l+}$  são estimadores ponderados das frequências nas celas e nas marginais de linha da tabela, respectivamente, de modo que  $\sum_{c=1}^C \hat{N}_{lc} = \hat{N}_{l+}$ . Estes estimadores levam em consideração as probabilidades desiguais de inclusão na amostra e os ajustes por não-resposta. Observe que, se os tamanhos das amostras dos subgrupos regionais não forem fixados, os  $\hat{p}_{lc}$  são estimadores de razão.

Sejam  $\hat{\mathbf{V}}_p(\hat{\mathbf{p}}_1)$  e  $\hat{\mathbf{V}}_p(\hat{\mathbf{p}}_2)$  estimadores consistentes das matrizes de variância de aleatorização dos vetores  $\hat{\mathbf{p}}_1$  e  $\hat{\mathbf{p}}_2$ , respectivamente. A estatística de Wald baseada no plano amostral  $X_W^2(H)$  para efetuar o teste de homogeneidade no caso de duas regiões ( $L = 2$ ) é dada por

$$X_W^2(H) = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)' \left[ \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_2) \right]^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \quad (8.1)$$

pois as amostras são disjuntas e supostas independentes.

No caso, a estatística de Wald  $X_W^2(H)$  tem distribuição assintótica qui-quadrado com  $(2 - 1) \times (C - 1)$  graus de liberdade. Quando o número de unidades primárias de amostragem na amostra de cada região é grande, a estatística de Wald funciona adequadamente. Caso contrário, ocorre problema de instabilidade e usamos, alternativamente, uma estatística F-corrigida de Wald. Freitas et al.(1997) descrevem uma aplicação da estatística  $X_W^2(H)$  para testar a hipótese de igualdade das pirâmides etárias estimadas pela Pesquisa de Padrões de Vida 96/97 (PPV) e da Pesquisa Nacional por Amostra de Domicílios 95 para as regiões Sudeste e Nordeste. Tal comparação fez parte do processo de avaliação da qualidade dos resultados da PPV.

Designemos por  $f = m - H$  o número total de graus de liberdade disponível para estimar  $\left[ \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_2) \right]$ , onde  $m$  e  $H$  são os números totais de conglomerados e de estratos nas amostras das duas regiões, respectivamente. As correções F da estatística  $X_W^2(H)$  são dadas por

$$F_{1,p} = \frac{f - (C - 1) + 1}{f(C - 1)} X_W^2(H), \quad (8.2)$$

que tem distribuição de referência  $F$  com  $(C - 1)$  e  $(f - (C - 1) + 1)$  graus de liberdade e, ainda,

$$F_{2,p} = X_W^2(H) / (C - 1) \quad (8.3)$$

que tem distribuição de referência  $F$  com  $(C - 1)$  e  $f$  graus de liberdade.

As estatísticas  $F_{1,p}$  e  $F_{2,p}$  podem amenizar o efeito de instabilidade, quando  $f$  não é grande relativamente ao número de classes ( $C$ ) da variável resposta.

No caso de  $L = 2$  regiões, a estatística de teste de homogeneidade de Pearson é dada por

$$X_P^2(H) = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)^0 \left( \hat{\mathbf{P}}/\hat{n}_{1+} + \hat{\mathbf{P}}/\hat{n}_{2+} \right)^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2) \quad , \quad (8.4)$$

onde  $\hat{\mathbf{P}} = \mathbf{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0 \hat{\mathbf{p}}_0^0$  e  $\hat{\mathbf{p}}_0$  é o estimador do vetor comum de proporções sob a hipótese de homogeneidade.

Neste caso,  $\hat{\mathbf{P}}/\hat{n}_{1+}$  é o estimador da matriz de covariância de  $\hat{\mathbf{p}}_0$  na primeira região e  $\hat{\mathbf{P}}/\hat{n}_{2+}$  na segunda. Observe que (8.4) e (8.1) têm a mesma forma, diferindo só no estimador da matriz de covariância usado para definir a métrica de distância. No caso da estatística  $X_P^2(H)$ , o estimador da matriz de covariância baseia-se nas hipóteses relativas à distribuição multinomial, apropriadas para a amostragem aleatória simples. A distribuição de referência da estatística  $X_P^2(H)$  é qui-quadrado com  $(C - 1)$  graus de liberdade.

Para introduzir em  $X_P^2(H)$  o ajuste de EPA médio e o ajuste de Rao-Scott de primeira ordem, é preciso calcular estimativas de efeitos de plano amostral das estimativas das proporções nas linhas em ambas as regiões. O ajuste de segunda ordem de Rao-Scott, por sua vez, depende da matriz de efeito multivariado do plano amostral. As estimativas de efeitos de plano amostral na região  $l$  são da forma

$$\hat{d}_{lc} = \hat{n}_{l+} \hat{V}_{lc} / (\hat{p}_{+c} (1 - \hat{p}_{+c})), \quad l = 1, 2 \text{ e } c = 1, \dots, C, \quad (8.5)$$

onde  $\hat{V}_{lc}$  é o  $c$ -ésimo elemento da diagonal de  $\hat{\mathbf{V}}_p(\hat{\mathbf{p}}_l)$ .

A matriz estimada de efeito multivariado de plano amostral é

$$\hat{\Delta} = \frac{\hat{n}_{1+} \times \hat{n}_{2+}}{\hat{n}_{1+} + \hat{n}_{2+}} \hat{\mathbf{P}}^{-1} \left( \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\hat{\mathbf{p}}_2) \right) \quad . \quad (8.6)$$

A estatística de Pearson com ajuste de EPA médio é dada por

$$X_P^2(H; \hat{d}) = X_P^2(H) / \hat{d} \quad , \quad (8.7)$$

onde  $\hat{d} = \sum_{l=1}^2 \sum_{c=1}^C \hat{d}_{lc} / 2C$  é a média das estimativas dos efeitos univariados de plano amostral.

Usando os autovalores  $\hat{\delta}_c$  de  $\hat{\Delta}$ , o ajuste de primeira ordem de Rao-Scott é dado por

$$X_P^2(H; \hat{\delta}) = X_P^2(H) / \hat{\delta} \quad , \quad (8.8)$$

onde

$$\hat{\delta} = \frac{tr(\hat{\Delta})}{(C-1)} = \frac{1}{C-1} \sum_{l=1}^2 \left( 1 - \frac{\hat{n}_{l+}}{\hat{n}_{1+} + \hat{n}_{2+}} \right) \sum_{c=1}^C \frac{\hat{p}_{lc}}{\hat{p}_{+c}} (1 - \hat{p}_{lc}) \hat{d}_{lc}$$

é um estimador da média  $\bar{\delta}$  dos autovalores  $\delta_c$  da matriz  $\Delta$ , desconhecida, de efeito multivariado do plano amostral. Como a soma dos autovalores de  $\hat{\Delta}$  é igual ao traço de  $\hat{\Delta}$ , esta correção pode ser obtida sem ser necessário calcular os autovalores.

As distribuições de referência, tanto de  $X_P^2(H; \hat{d})$  como de  $X_P^2(H; \hat{\delta})$ , são qui-quadrado com  $(C-1)$  graus de liberdade. Estes ajustes corrigem a estatística  $X_P^2(H)$  de modo a obter estatísticas com valor esperado igual ao da distribuição qui-quadrado de referência. Tal correção é apropriada quando houver pouca variação das estimativas dos autovalores  $\hat{\delta}_c$ . Quando isto não ocorrer, pode ser introduzido o ajuste de segunda ordem de Rao-Scott, que para a estatística de Pearson é dado por

$$X_P^2(H; \hat{\delta}, \hat{a}^2) = X_P^2(H; \hat{\delta}) / (1 + \hat{a}^2) \quad (8.9)$$

onde  $\hat{a}^2$  é o quadrado do coeficiente de variação dos quadrados das estimativas dos autovalores  $\hat{\delta}_c$ , dado por

$$\hat{a}^2 = \sum_{c=1}^C \hat{\delta}_c^2 / \left( (C-1) \hat{\delta}^2 \right) - 1 \quad ,$$

onde a soma dos quadrados dos autovalores pode ser obtida a partir do traço de  $\hat{\Delta}^2$

$$\sum_{c=1}^C \hat{\delta}_c^2 = tr(\hat{\Delta}^2) \quad .$$

A estatística de Pearson com a correção de segunda ordem de Rao-Scott  $X_P^2(H; \hat{\delta}, \hat{a}^2)$  tem distribuição de referência qui-quadrado com graus de liberdade com ajuste de Satterhwaite  $gl_S = (C-1) / (1 + \hat{a}^2)$ .

Tabela 8.3: Proporções por cela na população

Variável 1	Variável 2						Total
	1	2	...	$c$	...	$C$	
1	$p_{11}$	$p_{12}$	...	$p_{1c}$	...	$p_{1C}$	$p_{1+}$
2	$p_{21}$	$p_{22}$	...	$p_{2c}$	...	$p_{2C}$	$p_{2+}$
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
$l$	$p_{l1}$	$p_{l2}$	...	$p_{lc}$	...	$p_{lC}$	$p_{l+}$
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
$L$	$p_{L1}$	$p_{L2}$	...	$p_{Lc}$	...	$p_{LC}$	$p_{L+}$
Total	$p_{+1}$	$p_{+2}$	...	$p_{+c}$	...	$p_{+C}$	1

Quando as estimativas  $\hat{\mathbf{V}}_p(\hat{\mathbf{p}}_1)$  e  $\hat{\mathbf{V}}_p(\hat{\mathbf{p}}_2)$  das matrizes de covariâncias regionais são baseadas em números relativamente pequenos de unidades primárias de amostragem selecionadas, pode-se usar a estatística F-corrigida de Pearson. Ela é dada, no caso de duas regiões, por

$$FX_P^2(H; \hat{\delta}) = X_P^2(H; \hat{\delta}) / (C - 1) ,$$

e tem distribuição de referência  $F$  com  $(C - 1)$  e  $f$  graus de liberdade.

### 8.3.2 Teste de Independência

Vamos considerar o teste de independência no caso geral de tabela  $L \times C$ , onde os dados são extraídos de uma única população, sem fixar marginais. Consideremos a Tabela 8.3 com as proporções nas celas a nível da população, onde agora novamente se tem  $p_{lc} = N_{lc}/N$ .

Estamos interessados em testar a hipótese de independência

$$H_0 : p_{lc} = p_{l+}p_{+c}, \quad l = 1, \dots, L - 1, \quad c = 1, \dots, C - 1 ,$$

onde  $p_{l+} = \sum_{c=1}^C p_{lc}$ ,  $p_{+c} = \sum_{l=1}^L p_{lc}$  e  $\sum_{c=1}^C \sum_{l=1}^L p_{lc} = 1$ .

Vamos escrever a hipótese de independência numa forma alternativa mas equivalente, usando contrastes de proporções:

$$H_0 : f_{lc} = p_{lc} - p_{l+}p_{+c} = 0, \quad l = 1, \dots, L - 1, \quad c = 1, \dots, C - 1 .$$

Consideremos o vetor  $\mathbf{f}$  com  $(L - 1)(C - 1)$  componentes formado pelos contrastes  $f_{lc}$  arranjados em ordem de linhas:

$$\mathbf{f} = (f_{11}, \dots, f_{1C-1}, \dots, f_{L-11}, \dots, f_{L-1C-1})' .$$

Um teste da hipótese de independência pode ser definido em termos da distância entre uma estimativa consistente do vetor de contrastes  $\mathbf{f}$  e o vetor nulo com mesmo número de componentes. O vetor de estimativa consistente de  $\mathbf{f}$  é denotado por  $\hat{\mathbf{f}} = \left( \hat{f}_{11}, \dots, \hat{f}_{1\ C-1}, \dots, \hat{f}_{L-1\ 1}, \dots, \hat{f}_{L-1\ C-1} \right)^0$ , onde  $\hat{f}_{lc} = \hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c}$ , onde  $\hat{p}_{lc} = \hat{n}_{lc}/n$ . Os  $\hat{n}_{lc}$  são as frequências ponderadas nas celas, considerando as diferentes probabilidades de inclusão e ajustes por não-resposta, onde os pesos amostrais são normalizados de modo que  $\sum_{c=1}^C \sum_{l=1}^L \hat{n}_{lc} = n$ . Se  $n$  não for fixado de antemão, os  $\hat{p}_{lc}$  serão estimadores de razões. Apenas  $(L-1)(C-1)$  componentes são incluídos no vetores  $\mathbf{f}$  e  $\hat{\mathbf{f}}$ , pois a soma das proporções nas celas da tabela é igual a 1.

### 8.3.3 Estatística de Wald Baseada no Plano Amostral

A estatística de Wald baseada no plano amostral  $X_W^2(I)$ , para o teste de independência, tem a forma da expressão (7.8), com  $\hat{\mathbf{f}}$  no lugar de  $\hat{\mathbf{p}}$ , o vetor  $\mathbf{0}_{(L-1)(C-1)}$  no lugar de  $\mathbf{p}_0$  e a estimativa baseada no plano amostral  $\hat{\mathbf{V}}_f$  da matriz de covariância de  $\hat{\mathbf{f}}$  no lugar de  $\hat{\mathbf{V}}_p$ . Assim, a estatística de teste de independência de Wald é dada por

$$X_W^2(I) = \hat{\mathbf{f}}' \hat{\mathbf{V}}_f^{-1} \hat{\mathbf{f}} \quad , \quad (8.10)$$

que é assintoticamente  $\chi^2((L-1)(C-1))$ .

A estimativa  $\hat{\mathbf{V}}_f$  da matriz de covariância de  $\hat{\mathbf{f}}$  pode ser obtida pelo método de linearização de Taylor apresentado na Seção 3.3, considerando o vetor de contrastes  $\mathbf{f}$  como uma função (não-linear) do vetor  $\mathbf{p}$ , isto é,  $\mathbf{f} = \mathbf{g}(\mathbf{p}) = \mathbf{g}(p_{11}, \dots, p_{1\ C-1}, \dots, p_{L-1\ 1}, \dots, p_{L-1\ C-1})$ . Assim, a matriz de covariância de  $\hat{\mathbf{f}}$  pode ser estimada por

$$\hat{\mathbf{V}}_f = \Delta \mathbf{g}(\hat{\mathbf{p}}) \hat{\mathbf{V}}_p^{-1} \Delta \mathbf{g}(\hat{\mathbf{p}})^0 \quad , \quad (8.11)$$

onde  $\Delta \mathbf{g}(\mathbf{p})$  é a matriz jacobiana de dimensão  $(L-1)(C-1) \times (L-1)(C-1)$  dada por

$$\Delta \mathbf{g}(\mathbf{p}) = [\partial \mathbf{g} / \partial p_{11}, \dots, \partial \mathbf{g} / \partial p_{1\ C-1}, \dots, \partial \mathbf{g} / \partial p_{L-1\ 1}, \dots, \partial \mathbf{g} / \partial p_{L-1\ C-1}]$$

e  $\hat{\mathbf{V}}_p$  é uma estimativa consistente da matriz de covariância de  $\hat{\mathbf{p}}$ .

É possível ainda introduzir, no caso de se ter o número  $m$  de unidades primárias pequeno, correção na estatística de Wald, utilizando as propostas alternativas de estatísticas F-corrigidas, como em 7.9 e 7.10, com  $(L-1)(C-1)$  no lugar de  $J-1$ , obtendo-se

$$F_{1,p} = \frac{f - (L-1)(C-1) - 1}{f(L-1)(C-1)} X_W^2(I) \quad ,$$

que tem distribuição assintótica  $\mathbf{F}$  com  $(L - 1)(C - 1)$  e  $f - (L - 1)(C - 1) - 1$  graus de liberdade e

$$F_{2,p} = \frac{X_W^2(I)}{(L - 1)(C - 1)},$$

que tem distribuição assintótica  $\mathbf{F}$  com  $(L - 1)(C - 1)$  e  $f$  graus de liberdade.

### 8.3.4 Estatística de Pearson com Ajuste de Rao-Scott

Na presença de efeitos de plano amostral importantes, as estatísticas clássicas de teste precisam ser ajustadas para terem a mesma distribuição assintótica de referência que a obtida para o caso de amostragem aleatória simples.

A estatística de teste de independência  $X_P^2(I)$  de Pearson para a tabela  $L \times C$  é dada por

$$X_P^2(I) = n \sum_{l=1}^L \sum_{c=1}^C \frac{(\hat{p}_{lc} - \hat{p}_{l+} \hat{p}_{+c})^2}{\hat{p}_{l+} \hat{p}_{+c}}.$$

Esta estatística pode ser escrita em forma matricial como

$$X_P^2(I) = n \hat{\mathbf{f}}' \hat{\mathbf{P}}_{\text{of}} \hat{\mathbf{f}}, \quad (8.12)$$

onde

$$\begin{aligned} \hat{\mathbf{P}}_{\text{of}} &= \mathbf{\Delta g}(\hat{\mathbf{p}}) \hat{\mathbf{P}}_0 \mathbf{\Delta g}(\hat{\mathbf{p}})' , \\ \hat{\mathbf{P}}_0 &= \text{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0 \hat{\mathbf{p}}_0' , \end{aligned} \quad (8.13)$$

$\hat{\mathbf{P}}_0/n$  estima a matriz  $(L - 1)(C - 1) \times (L - 1)(C - 1)$  de covariância multinomial de  $\hat{\mathbf{p}}$  sob a hipótese nula,  $\hat{\mathbf{p}}_0$  é o vetor com componentes  $\hat{p}_{l+}$  e  $\hat{p}_{+c}$ , e  $\text{diag}(\hat{\mathbf{p}}_0)$  representa a matriz diagonal com elementos  $\hat{p}_{l+}$  e  $\hat{p}_{+c}$  na diagonal.

Observemos que a forma de  $X_P^2(I)$  como expressa em (8.12) é semelhante à da estatística de Wald dada em (8.10), a diferença sendo a estimativa da matriz de covariância de  $\hat{\mathbf{f}}$  usada em cada uma dessas estatísticas.

Como nos testes de qualidade de ajuste e de homogeneidade no caso de plano amostral complexo, podemos introduzir correções simples na estatística de Pearson em (8.12) para obter estatísticas de teste com distribuições assintóticas conhecidas.

Inicialmente, vamos considerar ajustes baseados nos efeitos univariados de plano amostral estimados,  $\hat{d}_{lc}$ , das estimativas das proporções nas celas

$\hat{p}_{lc}$ . O ajuste mais simples é feito dividindo-se o valor da estatística  $X_P^2$  de Pearson pela média  $\hat{d}$ , dos efeitos univariados de plano amostral:

$$X_P^2(I; \hat{d}) = X_P^2(I) / \hat{d} ,$$

onde  $\hat{d} = \sum_{c=1}^C \sum_{l=1}^L \hat{d}_{lc} / (LC)$  é um estimador da média dos efeitos univariados de plano amostral desconhecidos.

Estimamos os efeitos do plano amostral por  $\hat{d}_{lc} = \hat{V}_p(\hat{p}_{lc}) / (\hat{p}_{lc}(1 - \hat{p}_{lc}) / n)$ , onde  $\hat{V}_p(\hat{p}_{lc})$  é a estimativa da variância de aleatorização do estimador de proporção  $\hat{p}_{lc}$ . Este ajustamento requer que estejam disponíveis as estimativas dos efeitos de plano amostral dos estimadores das proporções nas  $L \times C$  celas da tabela.

A seguir vamos apresentar as correções de primeira e de segunda ordem de Rao-Scott para a estatística  $X_P^2(I)$  de Pearson para o teste de independência. Estas correções baseiam-se nos autovalores da matriz estimada de efeito multivariado de plano amostral, dada por

$$\hat{\Delta} = n \hat{\mathbf{P}}_{\text{of}}^{-1} \hat{\mathbf{V}}_f, \quad (8.14)$$

onde  $\hat{\mathbf{V}}_f$  foi definido em (8.11) e  $\hat{\mathbf{P}}_{\text{of}}$  definido em (8.13).

O ajuste de Rao-Scott de primeira ordem para  $X_P^2(I)$  é dado por

$$X_P^2(I; \hat{\delta}) = X_P^2(I) / \hat{\delta} , \quad (8.15)$$

onde  $\hat{\delta}$  é um estimador da média  $\bar{\delta}$  dos autovalores desconhecidos da matriz  $\hat{\Delta}$  de efeitos multivariados de plano amostral.

Podemos estimar a média dos efeitos generalizados, usando os efeitos univariados nas celas e nas marginais da tabela, por

$$\hat{\delta} = \frac{1}{(L-1)(C-1)} \sum_{l=1}^L \sum_{c=1}^C \frac{\hat{p}_{lc}(1 - \hat{p}_{lc})}{\hat{p}_{l+}\hat{p}_{+c}} \hat{d}_{lc} - \sum_{l=1}^L (1 - \hat{p}_{l+}) \hat{d}_{l+} - \sum_{c=1}^C (1 - \hat{p}_{+c}) \hat{d}_{+c} ,$$

sem precisar calcular a matriz de efeitos multivariados de plano amostral. A distribuição assintótica de  $X_P^2(I; \hat{\delta})$ , sob  $H_0$ , é qui-quadrado com  $(L-1)(C-1)$  graus de liberdade.

O ajuste de Rao-Scott de segunda ordem é definido por

$$X_P^2(I; \hat{\delta}; \hat{a}^2) = X_P^2(I) / \left( \hat{\delta} (1 + \hat{a}^2) \right) ,$$

onde  $\hat{\delta}_.$  é um estimador da média dos autovalores de  $\hat{\Delta}$ , dado por

$$\hat{\delta}_. = \frac{\text{tr}(\hat{\Delta})}{(L-1)(C-1)}$$

e  $\hat{a}^2$  é um estimador do quadrado do coeficiente de variação dos autovalores desconhecidos de  $\Delta$ ,  $\delta_k$ ,  $k = 1, \dots, (L-1)(C-1)$ , dado por

$$\hat{a}^2 = \frac{\sum_{k=1}^{(L-1)(C-1)} \delta_k^2}{\left((L-1)(C-1)\hat{\delta}_.\right)^2} - 1 .$$

Um estimador da soma dos quadrados dos autovalores é

$$\sum_{k=1}^{(L-1)(C-1)} \delta_k^2 = \text{tr}(\hat{\Delta}^2) .$$

A estatística  $X_P^2(I; \hat{\delta}_.; \hat{a}^2)$  é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite  $gl_S = (L-1)(C-1) / (1 + \hat{a}^2)$ .

Em situações instáveis, pode ser necessário fazer uma correção F ao ajuste de primeira ordem de Rao-Scott (8.15). A estatística F-corrigida é definida por

$$FX_P^2(\hat{\delta}_.) = X_P^2(\hat{\delta}_.) / (L-1)(C-1) . \quad (8.16)$$

A estatística (8.16) tem distribuição de referência  $F$  com  $(L-1)(C-1)$  e  $f$  graus de liberdade.

**Exemplo 8.1** *Correções de EPA médio das estatísticas  $X_P^2(I)$  e  $X_P^2(H)$ .*

Considerando os dados do Exemplo 6.1 do Capítulo 6, vamos testar a hipótese de independência entre as variáveis Sexo (sx) e Rendimento médio mensal (re). Vamos fazer também um teste de homogeneidade, para comparar as distribuições de renda para os dois sexos.

A variável sx tem dois níveis: sx(1)-Homens, sx(2)-Mulheres e a variável re tem três níveis: re(1)- Menos de salário mínimo, re(2) - de 1 a 5 salário mínimos e re(3)- mais de 5 salários mínimos. A Tabela 8.4 apresenta as frequências nas celas para a amostra pesquisada.

Tabela 8.4: Freqüências amostrais por celas na PNAD 90

Sexo	Renda Mensal			Total
	1	2	3	
1	476	2.527	1.273	4.276
2	539	1.270	422	2.231
Total	1.015	3.797	1.695	6.507

Tabela 8.5: Proporções nas linhas, desvios padrões e EPAs

Sexo	Renda Mensal			Total
	1	2	3	
1	0,111	0,591	0,298	1,00
	57,269	102,576	111,213	
	1,420	1,861	2,527	
2	0,240	0,570	0,190	1,00
	125,026	119,375	111,410	
	1,909	1,297	1,800	
Amostra completa	0,155	0,584	0,261	1,00
	68,977	82,001	96,1300	
	2,358	1,800	3,117	

No teste de homogeneidade das distribuições de renda, consideramos fixadas as marginais 4.276 e 2.231 da variável Sexo na tabela de freqüências amostrais. Usando o programa Stata, calculamos as estimativas das proporções nas linhas da tabela. Nestas estimativas são considerados os pesos das unidades da amostra e o plano amostral utilizado na pesquisa (PNAD 90), conforme descrito no Exemplo 6.1 do Capítulo 6.

Vamos considerar o teste de homogeneidade entre as variáveis Sexo e Renda e calcular o efeito de plano amostral médio das estimativas das proporções nas celas da tabela. A Tabela 8.5 contém, em cada cela, as estimativas: da proporção na cela, do desvio-padrão da estimativa da proporção na cela ( $\times 10.000$ ), e do efeito de plano amostral da estimativa de proporção na cela.

Vamos calcular, a título de ilustração, uma das celas de tabela de efei-

Tabela 8.6: Proporções nas celas, desvios padrões e EPAs

Sexo	Renda Mensal			Total
	1	2	3	
1	0,073	0,388	0,196	0,657
	38,343	80,435	71,772	55,814
	1,414	1,772	2,128	0,899
2	0,082	0,195	0,065	0,343
	44,401	51,582	40,219	55,814
	1,695	1,101	1,729	0,899
Total	0,155	0,584	0,261	
	68,977	82,001	96,130	1,000
	2,358	1,800	3,117	

tos de plano amostral, digamos a cela (1,1). A estimativa da variância do estimador da proporção de linha nesta cela é  $(0,0057269)^2$ . Sob amostragem aleatória simples com reposição, a estimativa da variância do estimador de proporção de linha na cela é:  $0,111(1 - 0,111)/4.276$ . A estimativa do efeito de plano amostral do estimador de proporção na cela é portanto igual a

$$\frac{(0,0057269)^2}{0,111(1 - 0,111)/4.276} \cong 1,420 .$$

A estimativa do efeito médio de plano amostral para corrigir a estatística  $X_P^2(H)$  é  $\hat{d} = 1,802$ , calculada tomando a média dos EPAs das celas correspondentes aos níveis 1 e 2 da variável sx.

Vamos agora considerar o teste de independência entre as variáveis Sexo e Renda e calcular o efeito de plano amostral médio das estimativas das proporções nas celas da tabela. A Tabela 8.6 contém, em cada cela, as estimativas: da proporção na cela, do desvio-padrão da estimativa da proporção na cela ( $\times 10.000$ ), e do efeito de plano amostral da estimativa de proporção na cela.

Vamos calcular, a título de ilustração, o efeito de plano amostral na cela (1,1) da Tabela 8.6. A estimativa da variância do estimador de proporção nesta cela é  $(0,0038343)^2$ . Sob amostragem aleatória simples com reposição, a estimativa da variância do estimador de proporção na cela é:

0,073(1 - 0,073)/6.507. A estimativa do efeito de plano amostral do estimador de proporção na cela é

$$\frac{(0,0038343)^2}{0,073(1 - 0,073)/6.507} \cong 1,414 .$$

Portanto, a estimativa do efeito médio de plano amostral requerida para corrigir a estatística  $X_P^2(I)$  é  $\hat{d} = 1,640$ , calculada tomando a média dos EPAs das celas correspondentes aos níveis 1 e 2 da variável *sex*.

Calculando as estatísticas  $X_P^2(I)$  e  $X_P^2(H)$  para os testes clássicos de independência e homogeneidade a partir da Tabela 8.6, obtemos os valores  $X_P^2(I) = X_P^2(H) = 227,025$ , com distribuição de referência  $\chi^2(2)$ , resultado que indica rejeição da hipótese de independência entre *sex* e *re*, bem como da hipótese de igualdade de distribuição de renda para os dois sexos a partir do teste de homogeneidade. O valor comum das estatísticas  $X_P^2(I)$  e  $X_P^2(H)$  foi calculado sem considerar os pesos e o plano amostral. Considerando estes últimos, mediante a correção de EPA médio das estatísticas clássicas, obtemos os valores  $X_P^2(I; \hat{d}) = \frac{224,849}{1,640} = 137,117$  e  $X_P^2(H; \hat{d}) = \frac{224,830}{1,802} = 124,742$ , que também indicam a rejeição das hipóteses de independência e de homogeneidade.

Vale ressaltar que apesar de todos os testes mencionados indicarem forte rejeição das hipóteses de independência e de homogeneidade, os valores das estatísticas de teste 137,117 e 124,742, calculados considerando os pesos e plano amostral, são bem menores que o valor 227,025 obtido para o caso de amostra IID. Sob a hipótese nula, a distribuição de referência de todas essas estatísticas de teste é  $\chi^2(2)$ , mostrando novamente que a estatística de teste calculada sob a hipótese de amostra IID tem maior tendência a rejeitar a hipótese nula.

A partir da Tabela 8.6, examinando as estimativas das proporções nas celas da tabela para cada sexo, observamos uma ordenação estocástica das distribuições de renda para os dois sexos, com proporções maiores em valores mais altos para o nível 1 da variável *sexo*, que é o sexo masculino.

## Capítulo 9

# Agregação vs. Desagregação

### 9.1 Introdução

Há duas abordagens principais para tratar a estrutura dos dados de pesquisas amostrais complexas. Numa delas, encaramos a estrutura dos dados como fator complicador ou aspecto indesejado, que invalida o uso de procedimentos padrões de análise, e mantemos inalterados os objetivos básicos da análise. Os métodos descritos nos capítulos anteriores se baseiam nesta abordagem, denominada de *análise agregada* ou *marginal*, pois os parâmetros de interesse são obtidos tomando-se a média ao longo de alguns aspectos da estrutura da população.

Na outra abordagem, denominada *análise desagregada*, mudamos os objetivos, incorporando mais explicitamente a estrutura da população no procedimento de análise, construindo modelos para descrever a relação entre as variáveis de interesse. A complexidade da estrutura da população é então usada como evidência de que modelos simples e procedimentos padrões são também, em geral, inadequados.

Para considerar a estrutura da população, os modelos requeridos são geralmente mais elaborados e às vezes requerem alteração dos alvos da inferência. Nos modelos modificados, os antigos parâmetros são abandonados e novos parâmetros são introduzidos, num processo iterativo que se baseia nos dados da pesquisa. Efeitos de conglomeração não mais são vistos como complicadores, que se interpõem entre dados e procedimentos bem aceitos, e sim como parte integral da estrutura da população, que deve ser adequadamente modelada e que pode contribuir para melhorar nossa compreensão das relações entre as variáveis.

Este capítulo se dedica a apresentar uma introdução à abordagem de aná-

lise desagregada, em contraposição aos procedimentos indicados nos capítulos anteriores. Para um exame mais detalhado do tema, o leitor deve consultar Skinner, Holt e Smith(1989, cap. 10 a 13) e Bryk e Raudenbush(1992).

## 9.2 Modelagem da Estrutura Populacional

Para introduzir a abordagem de análise desagregada, vamos considerar um modelo simples de regressão linear, definido por

$$E_M(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i \quad (9.1)$$

onde  $\beta_0$  e  $\beta_1$  são parâmetros desconhecidos e  $Y_i$  e  $X_i$  são as variáveis resposta e preditora para a  $i$ -ésima unidade da população, respectivamente. Modelos dessa forma são frequentemente considerados na prática para representar relações entre variáveis, e a inferência é dirigida aos parâmetros  $\beta_0$  e  $\beta_1$ .

Vamos agora considerar o caso bem simples de uma população com unidades divididas em dois grupos disjuntos (ou estratos), seja para fins de amostragem estratificada (emprego de planos amostrais com estratificação das unidades elementares) ou mesmo apenas para fins de análise. Um exemplo simples é o caso de populações humanas, em que pessoas são separadas em grupos de acordo com o sexo.

Neste caso simples, para incorporar ao modelo efeitos de estratificação basta introduzir uma variável preditora de tipo indicador  $Z$ , que indica se uma unidade pertence ao estrato 1, digamos. O modelo modificado fica então definido como

$$E_M(Y_i | X_i = x_i, Z_i = z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \quad (9.2)$$

onde  $z_i = 1$  se a unidade pertence ao estrato 1 e  $z_i = 0$  caso contrário. Observe que neste novo modelo aparecem dois novos parâmetros, a saber  $\beta_2$  e  $\beta_3$ .

Se  $\beta_3 = 0$ , o efeito do estrato é modificar o intercepto de  $\beta_0$  para  $\beta_0 + \beta_2$ , quando  $z_i$  passa de 0 a 1. Se  $\beta_3 \neq 0$ , além da variação do intercepto, há também modificação na declividade, que passa de  $\beta_1$  para  $\beta_1 + \beta_3$  quando  $z_i$  passa de 0 a 1.

Modelos com o efeito de estratificação aqui ilustrado podem ser facilmente generalizados para o caso de mais de dois estratos, bastando para isso adicionar de forma similar variáveis indicadoras de pertinência aos diversos estratos, exceto o último. Tais modelos podem ser úteis em uma variedade de situações de interesse prático. Um caso importante é o do emprego

de planos amostrais estratificados. Nesse caso, o analista pode optar por modificar seu modelo agregado (9.1) em favor de um modelo desagregado da forma (9.2), pois acredita que este último representa melhor a realidade subjacente. Se o plano amostral for do tipo amostragem estratificada simples e os estratos (de seleção) coincidirem com os do modelo (de análise), a inferência para os parâmetros do modelo pode ser feita usando procedimentos e pacotes padrões, sem maiores problemas. O mesmo já não ocorre se os estratos de análise diferem dos de seleção ou se o plano amostral empregado envolver outros aspectos de complexidade, tais como conglomeração e/ou probabilidades desiguais de seleção dentro dos estratos.

Outro caso de interesse prático é aquele em que os estratos de análise são definidos por razões substantivas ligadas à modelagem pretendida, independentemente de como foi selecionada a amostra da pesquisa que gerou os dados (este caso englobaria inclusive dados coletados mediante censos). Nesse caso, os efeitos de estratificação são intrínsecos ao modelo e a estimação dos parâmetros correspondentes é o alvo da inferência desejada. Um exemplo típico é a análise de efeitos de sexo sobre relações entre educação (medida em termos de anos de estudo, por exemplo) e renda, que sustenta discussões sobre preconceito contra mulheres no mercado de trabalho (estamos simplificando aqui a situação, pois em geral se precisa remover efeitos de profissão, posição na ocupação, número de horas trabalhadas e outros que afetam a renda de assalariados). Em casos como este, em que dados de pesquisas amostrais domiciliares são frequentemente usados para ajustar modelos com efeitos de estratificação, os estratos de análise (pessoas classificadas por sexo) são formados a posteriori, porque as pessoas da amostra não são selecionadas em grupos devido à inexistência de cadastros que suportassem esse tipo de plano amostral. Na prática, as amostras selecionadas são de domicílios e nestes investigadas todas as pessoas moradoras.

Uma outra situação de interesse prático que pode requerer modificação dos modelos de interesse é a ocorrência de efeitos de conglomeração. Estes podem tanto se originar de necessidades administrativas que motivam a adoção de planos amostrais conglomerados (vide o caso das pesquisas por amostragem domiciliar, em que municípios, setores e domicílios formam conglomerados de pessoas, estas últimas as unidades de análise de interesse da modelagem), quanto de necessidades substantivas, em que os grupos de unidades elementares fazem parte de uma estrutura populacional cujas propriedades se deseja modelar de forma mais explícita. Um exemplo é o caso de estudos demográficos sobre mortalidade infantil, em que os filhos tidos por uma determinada mulher são considerados um conglomerado e se pretende identificar algum efeito potencial do tamanho dos conglomerados sobre os

eventos de interesse, no caso a mortalidade infantil.

Efeitos de conglomeração podem ser introduzidos no modelo (9.1) de maneira simples, bastando para isso considerar um modelo da forma

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \quad , \quad (9.3)$$

no qual  $j$  denota conglomerado e  $i$  denota indivíduo no conglomerado.

Em dados de pesquisas amostrais, os erros  $\varepsilon_{ij}$  não satisfazem, em geral, a hipótese de IID. Além disso, no modelo (9.3),  $\beta_0$  e  $\beta_1$  não variam para os diferentes conglomerados. Pode ser adequado supor que  $\beta_0$  e  $\beta_1$  variam entre conglomerados. Isto pode ser obtido substituindo  $\beta_0$  e  $\beta_1$  em (9.3) por coeficientes aleatórios, que dependem dos conglomerados, isto é, adotando-se o modelo

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \\ \beta_{0j} = \beta_0 + \eta_{0j} \\ \beta_{1j} = \beta_1 + \eta_{1j} \end{cases} \quad (9.4)$$

com  $\beta_0$  e  $\beta_1$  fixos e desconhecidos e  $\varepsilon_{ij}$ ,  $\eta_{0j}$  e  $\eta_{1j}$  variáveis aleatórias, satisfazendo

$$\begin{aligned} E_M(\varepsilon_{ij}) &= E_M(\eta_{0j}) = E_M(\eta_{1j}) = 0 & (9.5) \\ V_M(\varepsilon_{ij}) &= \sigma^2, V_M(\eta_{0j}) = \sigma_0^2, V_M(\eta_{1j}) = \sigma_1^2, \\ COV_M(\varepsilon_{ij}, \eta_{0j}^0) &= COV_M(\varepsilon_{ij}, \eta_{1j}^0) = 0, \\ COV_M(\varepsilon_{ij}, \varepsilon_{i^0 j^0}) &= 0, \quad j \neq j^0 \text{ ou } i \neq i^0, \end{aligned}$$

e

$$COV_M(\eta_{0j}, \eta_{1j^0}) = \begin{cases} \sigma_{01} & j = j^0 \\ 0 & j \neq j^0 \end{cases} . \quad (9.6)$$

Podemos juntar as expressões em (9.4) e reescrever o modelo como

$$\begin{aligned} Y_{ij} &= (\beta_0 + \eta_{0j}) + (\beta_1 + \eta_{1j}) x_{ij} + \varepsilon_{ij} & (9.7) \\ &= \beta_0 + \beta_1 x_{ij} + \eta_{0j} + \eta_{1j} x_{ij} + \varepsilon_{ij} . \end{aligned}$$

Em (9.7), os coeficientes  $\beta_0$  e  $\beta_1$  são fixos e os coeficientes  $\eta_{0j}$  e  $\eta_{1j}$  são aleatórios, sendo o modelo denominado de efeitos mistos: fixos e aleatórios (veja por exemplo Longford, 1993; Diggle, Liang e Zeger, 1994; Bryk e Raudenbush, 1992).

Em (9.5) e (9.6) os valores de  $\sigma_0^2$ ,  $\sigma_1^2$ ,  $\sigma_{01}$  e  $\sigma^2$  servem para medir a variação intra-conglomerados não explicada pelo modelo. O modelo pode ser mais elaborado, na tentativa de reduzir as variações não explicadas  $\sigma_0^2$ ,

$\sigma_1^2$  e talvez reduzir a covariância  $\sigma_{01}$ . Para isto, podemos introduzir no modelo uma outra variável preditora  $a_j$ , definida no nível de conglomerados, e considerar o novo modelo dado por

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}a_j + \eta_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}a_j + \eta_{1j} \end{cases} \quad (9.8)$$

Mais uma vez o objetivo básico da inferência se altera, pois agora está centralizado nos parâmetros  $(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}, \sigma_0^2, \sigma_1^2, \sigma^2, \sigma_{01})$ , com intervalos de confiança e testes de hipóteses relativos a estes parâmetros. O modelo (9.8) é de efeitos mistos, com efeitos fixos  $(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$  e efeitos aleatórios  $(\eta_{0j}, \eta_{1j})$ .

Modelos de efeitos mistos da forma (9.8) podem ser generalizados de diversas maneiras: mais variáveis preditoras  $x$  podem ser introduzidas na equação que descreve os valores individuais da variável resposta  $y$ ; efeitos de estratificação podem ser adicionados mediante introdução de variáveis indicadoras de pertinência a estratos  $z$ , como no modelo (9.2); mais variáveis preditoras  $a$  podem ser introduzidas nas equações que descrevem a variação dos parâmetros aleatórios a nível dos conglomerados; maior número de níveis de conglomeração podem ser considerados; etc. Aqui, o modelo "simples" da forma (9.8) já basta para ilustrar a maior complexidade envolvida na modelagem ao se tentar incorporar efeitos de conglomeração nessa abordagem desagregada.

Entre os modelos disponíveis para incorporar generalizações dos tipos aqui discutidos, uma classe de modelos bastante ampla e que tem sido objeto de grande interesse na literatura recente é a classe dos modelos hierárquicos, cujas idéias básicas introduziremos na próxima seção.

### 9.3 Modelos Hierárquicos

Modelos hierárquicos são indicados quando a estrutura populacional é hierárquica, isto é, quando as unidades elementares de análise estão grupadas em unidades maiores, que por sua vez também podem ou não pertencer a uma estrutura de grupos, numa hierarquia bem definida. Algumas vezes, tal hierarquia é uma propriedade intrínseca da população estudada. Um exemplo interessante de estrutura populacional hierárquica é um sistema educacional. Nele, os estudantes são naturalmente agrupados em turmas, as turmas agrupadas em escolas, as escolas agrupadas por distritos escolares ou municípios, e assim por diante. O uso de modelos hierárquicos para descrever

tais estruturas tem motivação nas próprias estruturas, independentemente do procedimento amostral usado para a obtenção dos dados eventualmente observados.

Adotando como referência básica Skinner, Holt e Smith (1989, cap. 11), vamos apresentar um resumo de alguns modelos hierárquicos básicos, iniciando com o caso de variáveis contínuas. Ainda no contexto de estudantes e turmas do exemplo discutido nesta seção, vamos considerar um modelo hierárquico de dois níveis com as seguintes variáveis:

- *ESC* - escore do aluno num teste de Matemática, considerada como variável resposta;
- *SEX* - sexo do aluno;
- *CSA* - classe social do aluno;
- *CST* - classe social média dos alunos da turma;
- *EXP* - anos de experiência do professor de Matemática.

Observe que as variáveis *SEX* e *CSA* se referem ao aluno (nível 1 do modelo), enquanto as variáveis *CST* e *EXP* se referem à turma (nível 2 do modelo) à qual o aluno pertence. A variável *EXP* é uma característica do professor, ao passo que *CST* é uma variável "contextual", baseada numa característica dos alunos agregada para o nível da turma.

Para fixar idéias, vamos considerar um modelo (nível aluno, ou nível 1) diferente para cada turma, explicando *ESC* pelas variáveis *SEX* e *CSA*:

$$ESC_{ij} = \beta_{0j} + \beta_{1j}SEX_{ij} + \beta_{2j}CSA_{ij} + \varepsilon_{ij}, \quad (9.9)$$

onde  $i = 1, \dots, n_j$  denota o aluno dentro da turma e  $j = 1, \dots, J$  denota a turma.

É possível que os coeficientes  $\beta_{0j}$ ,  $\beta_{1j}$  e  $\beta_{2j}$  variem entre as turmas. Além disso, parte desta variação tem uma componente não-sistemática, mas os coeficientes podem também depender de características das turmas. Vamos considerar as variáveis *CST* e *EXP*, medidas no nível da turma (nível 2), para explicar parte da variação dos coeficientes, através das seguintes equações (nível 2):

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}CST_j + \gamma_{02}EXP_j + \eta_{0j} , \\ \beta_{1j} = \gamma_{10} + \gamma_{11}CST_j + \gamma_{12}EXP_j + \eta_{1j} , \\ \beta_{2j} = \gamma_{20} + \gamma_{21}CST_j + \gamma_{22}EXP_j + \eta_{2j} , \end{cases} \quad (9.10)$$

onde  $\eta_{0j}$ ,  $\eta_{1j}$  e  $\eta_{2j}$  são erros no nível 2 satisfazendo as condições em (9.5). As equações (9.9) e (9.10) definem um modelo hierárquico, que pode ser escrito de forma equivalente como

$$\begin{aligned} ESC_{ij} = & \gamma_{00} + \gamma_{01}CST_j + \gamma_{02}EXP_j & (9.11) \\ & + (\gamma_{10} + \gamma_{11}CST_j + \gamma_{12}EXP_j)SEX_{ij} \\ & + (\gamma_{20} + \gamma_{21}CST_j + \gamma_{22}EXP_j)CSA_{ij} \\ & + \eta_{0j} + \eta_{1j}SEX_{ij} + \eta_{2j}CSA_{ij} + \varepsilon_{ij} . \end{aligned}$$

A presença dos erros aleatórios  $\eta_{0j}$ ,  $\eta_{1j}$  e  $\eta_{2j}$  (de nível 2), torna (9.11) um modelo misto. Se os erros fossem suprimidos em (9.10), o modelo especificado só teria efeitos fixos e a estimação dos parâmetros não traria qualquer problema. Entretanto, a exclusão dos erros de nível 2 em (9.10) não seria razoável, pois as variáveis definidas no nível 2 não determinam completamente os coeficientes dentro das turmas. Este aspecto fundamental do modelo deve ser incorporado no procedimento de estimação dos respectivos parâmetros de interesse.

Supondo que só os interceptos dos modelos dentro das turmas variam com as turmas, obtemos o seguinte modelo simplificado:

$$\begin{aligned} ESC_{ij} = & \gamma_{00} + \gamma_{01}CST_j + \gamma_{02}EXP_j & (9.12) \\ & + \beta_{1j}SEX_{ij} + \beta_{2j}CSA_{ij} + (\eta_{0j} + \varepsilon_{ij}) . \end{aligned}$$

Além da parte fixa, o modelo contém uma *parte residual*, com os erros aleatórios com média zero:  $\eta_{0j}$ , que representa o desvio da média dos indivíduos da turma  $j$  com relação à média total, e  $\varepsilon_{ij}$ , que é o desvio do  $i$ -ésimo aluno com relação à média da turma  $j$ . Vamos supor, ainda, que os  $\eta_{0j}$  e os  $\varepsilon_{ij}$  são independentes entre si e os  $\eta_{0j}$  são independentes dos  $\varepsilon_{ij}$ , com

$$E_M(\eta_{0j}) = E_M(\varepsilon_{ij}) = 0, V_M(\eta_{0j}) = \sigma_0^2, V_M(\varepsilon_{ij}) = \sigma^2, \forall i, j. \quad (9.13)$$

A parte aleatória do modelo (9.12) é o termo  $\eta_{0j} + \varepsilon_{ij}$ , com distribuição tendo parâmetros  $\sigma_0^2$  e  $\sigma^2$  a serem estimados.

O modelo (9.12) não permite estudar interações entre variáveis nos dois níveis hierárquicos. Para isto teríamos de supor, por exemplo, que a diferença de desempenho entre sexos varia com as turmas, requerendo que o modelo fosse alterado, fazendo  $\beta_{1j}$  depender das variáveis  $CST_j$  e  $EXP_j$ . Isto introduziria mais erros aleatórios no modelo e mais parâmetros a serem estimados.

Voltando ao modelo básico de dois níveis hierárquicos (9.11), verificamos que há uma correlação positiva entre respostas de alunos na mesma turma,

mas uma correlação nula entre respostas de alunos em turmas diferentes. Assim

$$COV_M(ESC_{ij}, ESC_{ij^0}) = COV_M[(\eta_{0j} + \varepsilon_{ij}), (\eta_{0j} + \varepsilon_{ij^0})] = \sigma_0^2, \quad (9.14)$$

pois supusemos que  $\varepsilon_{ij}$  e  $\varepsilon_{ij^0}$  são independentes. Por outro lado, condicionando na parte fixa do modelo, tem-se que

$$V_M(ESC_{ij}) = V_M(ESC_{ij^0}) = \sigma_0^2 + \sigma^2,$$

e portanto

$$CORR_M(ESC_{ij}, ESC_{ij^0}) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}. \quad (9.15)$$

A expressão (9.15) define a *correlação intraclassa* usual que, em nosso exemplo, mede o grau de similaridade entre alunos dentro das turmas ou o grau de conglomeração da variável resposta  $ESC_{ij}$  por turmas. Ela é a fração da variância residual atribuída à variância intra-conglomerado.

No caso de se supor adicionalmente a normalidade dos dados, foram propostos vários métodos para estimação dos parâmetros do modelo (9.11), entre os quais os métodos de Máxima Verossimilhança, de Máxima Verossimilhança Restrita e o Método Iterativo de Mínimos Quadrados Generalizados. Detalhes destes métodos de estimação não serão abordados neste texto, e o leitor interessado deve consultar, por exemplo, Bryk e Raudenbush(1992).

Nas considerações anteriores, não mencionamos explicitamente o plano amostral utilizado. Os modelos hierárquicos, ao incorporarem covariáveis características da estrutura populacional e também do plano amostral, tais como efeitos de estratificação e de conglomeração, tornam o plano amostral ignorável, condicionalmente nestas características, no sentido definido por Rubin(1976). Este raciocínio, porém, não é aplicável quando unidades em qualquer nível da hierarquia são selecionadas com probabilidades desiguais, de formas não consideradas pelas covariáveis. Por exemplo, quando as unidades são selecionadas com probabilidades proporcionais a uma medida de tamanho que é relacionada à variável resposta. Pfeffermann et al.(1998) apresentam uma forma de incorporar pesos no ajuste de modelos hierárquicos para compensar diferentes probabilidades de inclusão das unidades na amostra.

**Exemplo 9.1** *Plano amostral de pesquisa educacional (Lethonen e Pahkinen, 1995, p. 297).*

Os dados deste exemplo se referem a uma pesquisa de avaliação de escolas (SNACS), na qual foi analisado o desempenho em Matemática de alunos da sexta série. A população de conglomerados consistiu em 4.126 escolas, da qual foi selecionada uma amostra de 53 escolas, que produziu 1.071 alunos, numa população de 60.934 alunos. O tamanho total da amostra de alunos não foi fixado, de início. O plano amostral utilizado foi uma amostra estratificada de escolas (conglomerados) com um estágio, selecionada de um cadastro de escolas. Foi usada estratificação regional e as amostras nos estratos foram proporcionais ao tamanhos dos estratos.

A variável resposta binária *DESEMP*, indica se o aluno atingiu ou não um nível de conhecimento desejado em matemática. As variáveis explicativas quantitativas são:

- *EXP* - tempo de experiência do professor;
- *TEMP* - tempo em minutos gasto pelo aluno em trabalhos de casa, no tempo livre.

Cada preditor foi categorizado em três categorias, da forma a seguir:

- *EXP* - 1-10; 11-20 e 21 ou mais anos de experiência;
- *TEMP* - 0-14; 15-30; 31 ou mais minutos.

Observe que a variável *EXP* se refere ao professor, enquanto a variável *TEMP* se refere ao aluno.

Vários modelos foram ajustados, com complexidades e abordagens diferentes. Considerando inicialmente a abordagem agregada, e tomando as variáveis *EXP* e *TEMP* como contínuas, assumindo valores 1, 2 e 3, foi ajustado o modelo logístico

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1(TEMP)_j + \beta_2(EXP)_j, \quad (9.16)$$

no qual  $\beta_0, \beta_1$  e  $\beta_2$  são coeficientes a serem estimados, e o índice  $j$  se refere a um domínio de estudo,  $j = 1, \dots, 9$  (tais domínios não foram identificados na referência citada).

Podemos analisar os dados considerando as observações como IID, ignorando a existência de conglomerados e de pesos distintos, que chamaremos

Tabela 9.1: Análise do modelo (9.16) sob hipótese de observações IID, ignorando complexidades do plano amostral

Variável	Coefficiente	Desvio Padrão	Teste-t	pvalor	EPA
Intercepto	2,912	0,427	6,82	0,000	1
<i>TEMP</i>	-0,894	0,174	-5,14	0,000	1
<i>EXP</i>	0,254	0,127	2,00	0,045	1

Tabela 9.2: Análise agregada do modelo (9.16), usando MPV e estatísticas de Wald (PROC LOGISTIC do SUDAAN )

Variável	Coefficiente	Desvio Padrão	Teste-t	pvalor	EPA
Intercepto	2,899	0,578	5,02	0,00	1,83
<i>TEMP</i>	-0,906	0,211	-4,29	0,00	1,47
<i>EXP</i>	0,271	0,181	1,50	0,14	2,03

aqui de *análise ingênua*. Neste caso, os coeficientes podem ser estimados pelo método padrão de Máxima Verossimilhança, utilizando um dos pacotes padrões. Os resultados dessa análise são apresentados na Tabela 9.1.

Os *p*valores da Tabela 9.1 indicam que os coeficientes são significativamente diferentes de 0 ao nível de significância  $\alpha = 5\%$ , sugerindo que todas as variáveis preditoras têm poder de explicação, e portanto devem permanecer no modelo.

Outra opção é a *análise agregada*, que incorpora o plano amostral e os pesos através do método de MPV para estimar parâmetros, e do uso de estatísticas baseadas no plano amostral para testar hipóteses. Esta abordagem pode ser usada também na etapa de seleção de modelos, com testes de significância baseados, por exemplo, na estatística de Wald ou ajustes desta, no caso de instabilidade. Os resultados dessa análise são apresentados na Tabela 9.2.

Tabela 9.3: Análise agregada do modelo (9.16), usando estatísticas de Wald corrigidas (programa PC CARP)

Variável	Coefficiente	Desvio Padrão	Teste-t	pvalor	EPA
Intercepto	2,899	0,597	4,86	0,00	1,95
<i>TEMP</i>	-0,906	0,219	-4,14	0,00	1,58
<i>EXP</i>	0,271	0,186	1,46	0,15	2,14

A coluna de *p*valores da Tabela 9.2 indica que o coeficiente de *TEMP* é significativamente diferente de zero, e também que a hipótese de nulidade do coeficiente de *EXP* não é rejeitada no nível de significância  $\alpha = 5\%$ .

Neste exemplo,  $f = 38$  graus de liberdade para a estimação da matriz de covariância  $9 \times 9$  baseada no plano amostral. Como o estimador poderia ser instável, foi calculada a estatística corrigida de Wald. A correção usada foi a implementada no pacote PC-CARP, que difere da correção F antes mencionada. Os resultados da análise permanecem os mesmos que os da análise reportada na Tabela 9.2, conforme se pode verificar consultando os valores da Tabela 9.3. Em ambos os casos, observamos EPAs moderados com máximo em torno de 2.

Finalmente, a *análise desagregada* que, neste exemplo, usaria um *modelo hierárquico* com dois níveis, a saber:

- nível 1 - alunos;
- nível 2 - turmas.

No modelo hierárquico, para cada nível se considera uma variação aleatória. Denotando por  $p_{jk}$  a probabilidade de um aluno da turma  $k$  no domínio  $j$  atingir o nível desejado em Matemática, podemos modificar o modelo (9.16) para incluir o efeito da turma empregando

$$\log \left( \frac{p_{jk}}{1 - p_{jk}} \right) = \beta_0 + \beta_1 TEMP_{jk} + \beta_2 EXP_j + u_k . \quad (9.17)$$

O erro aleatório  $u_k \sim N(0, \sigma_u^2)$  representa a variação aleatória no nível 2. A variação no nível 1, entre alunos, é introduzida da seguinte forma. Sob

Tabela 9.4: Análise desagregada do modelo (9.18) via programa ML3

Variável	Coeficiente	Desvio Padrão	Teste-t	pvalor	EPA
Intercepto	2,941	0,538	5,47	0,00	1,58
<i>TEMP</i>	-0,927	0,179	-5,18	0,00	1,06
<i>EXP</i>	0,254	0,188	1,35	0,19	2,19

a hipótese binomial, em cada domínio, a proporção de alunos atingindo o nível adequado tem variância  $p_{jk}(1-p_{jk})/n_{jk}$ . Vamos supor que a variação residual no nível 1 é denotada por  $\sigma_e^2$ , e também que a variabilidade entre alunos acarreta uma variação extra-binomial

$$\frac{p_{jk}(1-p_{jk})\sigma_e^2}{n_{jk}}. \quad (9.18)$$

Os resultados desta análise são apresentados na Tabela 9.4.

A coluna de *p*valores da Tabela 9.4 indica novamente que o coeficiente de *EXP* não é significativamente diferente de zero ao nível  $\alpha = 5\%$ . A variação no nível 2 foi estimada por  $\sigma_u^2 = 0,42$  com desvio-padrão 0,189, e portanto é significativa ao nível  $\alpha = 5\%$ . Isto sugere a existência de diferença de avaliação dos professores sobre o aprendizado de Matemática dos alunos.

No ajuste da Tabela 9.4, a variação entre alunos foi tomada como  $\sigma_e^2 = 1$ . Uma alternativa seria estimar também este valor a partir dos dados.

Este exemplo ilustra bem o efeito de ignorar efeitos de plano amostral, ao fazer a análise ingênua, cujas conclusões levariam a incluir a variável *EXP* no modelo quando esta parece não ser importante, como revelaram as análises alternativas que levaram em conta o plano amostral ou a estrutura da população (análises agregada e desagregada, respectivamente).

### Exemplo 9.2 (*Bryk e Raudenbush, 1992, Cap. 5*)

Os dados consistiram em respostas de 8.000 professores aninhados em 357 escolas. A média de professores por escola foi  $8.000/357 = 22$  professores por escola. Os níveis da estrutura hierárquica considerados e os índices usados para representá-los foram:

- Unidade Primária de Amostragem (UPA) = Escola;
- Unidade Elementar = Professor;
- $i$  = Professor e  $j$  = Escola.

Foram observadas as seguintes variáveis:

- Variável resposta

$y_{ij}$  = Eficiência do Professor;

- Variáveis preditoras, relativas à Escola (unidade de nível 2)

$x_{1j}$  = Experiência Acadêmica Média antes da Escola Secundária;

$x_{2j}$  = Status Sócio-Econômico Médio;

$x_{3j}$  = Proporção Alta de Minorias;

$x_{4j}$  = Tamanho;

$x_{5j}$  = Mistura Étnica;

$x_{6j}$  = Mistura de Status Sócio-Econômico;

$x_{7j}$  = Grau de Organização Comunitária (*Comunit*).

A idéia deste exemplo é ilustrar como diversos modelos alternativos podem ser usados para analisar os dados sobre eficiência do professor, buscando explicação em variáveis que refletem a estrutura da escola onde atua. Um primeiro modelo que se poderia ajustar, considerando a estrutura hierárquica da população, é o modelo de análise de variância com um fator e com efeitos aleatórios

Modelo I: Análise de Variância com um Fator e com Efeitos Aleatórios.

Modelo de nível 1

$$Y_{ij} = \beta_{0j} + r_{ij} ,$$

Modelo de nível 2

$$\beta_{0j} = \gamma_{00} + u_{0j} .$$

Ou juntando as duas equações

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} ,$$

com  $E_M(r_{ij}) = E_M(u_{0j}) = 0$  e  $V_M(r_{ij}) = \sigma^2$  ,  $V_M(u_{0j}) = \tau_{00}$ .

Foram obtidas as seguintes estimativas para os parâmetros deste modelo:

$$\hat{\sigma}^2 = 0,915; \hat{\tau}_{00} = 0,084 \text{ e}$$

Tabela 9.5: Efeitos da escola na eficácia do professor (Modelo II)

Coefficiente	Estimativa	Desvio Padrão
$\hat{\gamma}_{01}$	0,044	0,020
$\hat{\gamma}_{02}$	0,133	0,023
$\hat{\gamma}_{03}$	0,031	0,046
$\hat{\gamma}_{04}$	-0,066	0,027
$\hat{\gamma}_{05}$	-0,014	0,019
$\hat{\gamma}_{06}$	-0,028	0,023
$\hat{\sigma}^2$	0,915	-
$\hat{\tau}_{00}$	0,055	-

$$\text{Correlação Intra-Escola} = \frac{\hat{\tau}_{00}}{\hat{\sigma}^2 + \hat{\tau}_{00}} = 0,092 .$$

Logo cerca de apenas 9% da variação na eficiência do professor é explicada pelas diferenças entre as escolas. Como essa proporção da variação explicada é pequena, na tentativa de aumentar o poder explicativo do modelo, vamos introduzir no modelo algumas variáveis explicativas referentes à escola, a saber as variáveis  $(x_1, \dots, x_6)$ .

**Modelo II:** Modelo Hierárquico com dois níveis, usando as variáveis  $(x_1, \dots, x_6)$ , definidas no nível 2, para explicar a variação da eficiência média do professor por escola.

Modelo de nível 1:

$$Y_{ij} = \beta_{0j} + r_{ij} .$$

Modelo explanatório de nível 2:

$$\beta_{0j} = \gamma_{00} + \sum_{k=1}^6 \gamma_{0j} x_{kj} + u_{0j} .$$

A Tabela 9.5 apresenta as estimativas dos parâmetros para este modelo.

Com essas estimativas, a proporção da variação total entre escolas do parâmetro  $\beta_{0j}$  (nível médio da eficácia dos professores por escola) explicada pelas variáveis  $(x_1, \dots, x_6)$  aumentou para

$$\frac{0,084 - 0,055}{0,084} = 35\% .$$

Tabela 9.6: Efeitos da escola na eficácia do professor (Modelo III)

Coeficiente	Estimativa	Desvio Padrão
$\hat{\gamma}_{01}$	0,038	0,017
$\hat{\gamma}_{02}$	0,015	0,022
$\hat{\gamma}_{03}$	-0,055	0,040
$\hat{\gamma}_{04}$	0,061	0,026
$\hat{\gamma}_{05}$	-0,014	0,016
$\hat{\gamma}_{06}$	-0,001	0,020
$\hat{\gamma}_{07}$	0,504	0,045
$\hat{\sigma}^2$	0,915	-
$\hat{\tau}_{00}$	0,031	-

Embora esse aumento do poder explicativo do modelo já tenha sido substancial, ainda é relativamente baixa a proporção de variação explicada, e portanto consideramos um terceiro modelo, em que foi adicionada ao Modelo II a variável preditora  $x_7 = Comunit$  ao nível da escola.

**Modelo III:** Modelo Hierárquico com dois níveis, usando as variáveis  $(x_1, \dots, x_7)$ , definidas no nível 2, para explicar a variação da eficiência média do professor por escola.

Modelo de nível 1

$$Y_{ij} = \beta_{0j} + r_{ij} ,$$

Modelo Explanatório de nível 2

$$\beta_{0j} = \gamma_{00} + \sum_{k=1}^7 \gamma_{0j} x_{kj} + u_{0j} .$$

A Tabela 9.6 apresenta as estimativas dos coeficientes do Modelo III e seus respectivos desvios padrões.

A proporção da variação total entre escolas do parâmetro  $\beta_{0j}$  (nível médio da eficácia dos professores por escola) explicada pelas variáveis  $(x_1, \dots, x_7)$  aumentou para

$$\frac{0,084 - 0,031}{0,084} = 63\% .$$

O incremento na variação explicada devido à introdução da variável *Comunit* no modelo de nível 2, foi de  $63\% - 35\% = 28\%$ , sugerindo que

Tabela 9.7: Efeitos da escola na eficácia do professor (Modelo IV)

Coefficiente	Estimativa	Desvio Padrão
$\hat{\gamma}_1$	0,040	0,013
$\hat{\gamma}_2$	0,015	0,017
$\hat{\gamma}_3$	-0,056	0,031
$\hat{\gamma}_4$	0,062	0,021
$\hat{\gamma}_5$	-0,014	0,013
$\hat{\gamma}_6$	0,002	0,016
$\hat{\gamma}_7$	0,507	0,035

essa variável é importante para explicar a variação na eficácia do professor. O Modelo III já atinge um nível razoável de poder explicativo e poderia ser considerado satisfatório para algumas finalidades.

Agora vamos ver o que teria ocorrido caso um analista procurasse ajustar um modelo aos dados de forma bastante ingênua, ignorando a estrutura hierárquica da população.

**Modelo IV:** Análise de Regressão Simples (nível 1).

Vamos considerar um modelo de regressão linear simples, com o resultado do professor  $Y_{ij}$  dependendo das características  $(x_{1j}, \dots, x_{7j})$  da escola, que teriam seus valores "repetidos" para os professores de uma mesma escola. Esse modelo pode ser escrito como

$$Y_{ij} = \gamma_0 + \sum_{k=1}^7 \gamma_k x_{kj} + e_{ij} .$$

A Tabela 9.7 apresenta as estimativas de Mínimos Quadrados Ponderados de  $(\gamma_0, \gamma_1, \dots, \gamma_7)$  com pesos dados por  $n_j =$  número de professores da escola  $j$ .

A proporção de variação explicada pelo Modelo IV é de apenas 5,4%. Os coeficientes da análise de nível 2 (Modelo III, Tabela 9.6) foram bem semelhantes neste exemplo, o que pode ser explicado em parte pela pequena variação do número de professores por escola.

A introdução da variável *Comunit*, neste modelo, só aumentou a quantidade de variação explicada em 2,5%. A julgar por este resultado, a importância da variável *Comunit* é pequena. Este resultado é enganador devido

ao fato de usar, no cálculo da razão de variação explicada, a variação total  $(\tau_{00} + \sigma^2)$  no denominador. No modelo hierárquico correspondente (Modelo III) este denominador é  $\tau_{00}$ , que é a parte explicável da variação. A estatística de variação explicada da análise hierárquica fornece uma evidência mais clara para se julgar a importância de preditores do nível 2.

Com este exemplo, procuramos ilustrar uma situação em que a estrutura populacional hierárquica não pode ser ignorada na modelagem, sob pena de se chegar a conclusões incorretas sobre a importância de determinadas variáveis preditoras num modelo de regressão, quando este é ajustado de forma ingênua a dados provenientes de uma estrutura hierárquica. Os modelos hierárquicos II e III aqui considerados são modelos de interceptos aleatórios do tipo  $y_{ij} = \beta_{0j} + \mathbf{x}_j^T \boldsymbol{\beta} + v_{ij}$ , onde os coeficientes de regressão  $\boldsymbol{\beta}$  são considerados fixos e apenas os interceptos  $\beta_{0j} = \beta_0 + u_j$  são efeitos aleatórios. Tais modelos poderiam ser generalizados mediante suposição de que os coeficientes de regressão nas variáveis preditoras  $\boldsymbol{\beta}$  também são aleatórios. O Modelo V a seguir é desse tipo.

Modelo V: Modelo de Coeficientes Aleatórios

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + v_{ij},$$

onde

$$\boldsymbol{\beta}_j = \mathbf{x}_j^T \boldsymbol{\gamma} + \boldsymbol{\delta}_j,$$

com  $\boldsymbol{\delta}_j (Q \times 1)$  e  $E_M(\boldsymbol{\delta}_j) = \mathbf{0}$ ,  $E_M(v_{ij}) = 0$ ,  $V_M(v_{ij}) = \sigma_1^2$ ,  $V_M(\boldsymbol{\delta}_j) = \boldsymbol{\Delta}$ ,  $\boldsymbol{\Delta}$  podendo ser não-diagonal. Não foram fornecidas estimativas dos parâmetros para este último modelo na referência citada. Sua formulação foi aqui incluída apenas para indicar que o estudo de modelos para a situação prática de interesse não se esgotaria nas alternativas de modelagem aqui consideradas.

## 9.4 Análise Desagregada: Prós e Contras

Vamos inicialmente listar algumas dificuldades na análise de dados de pesquisas complexas, indicando em cada caso como a análise desagregada poderia ajudar a solucionar o problema. Esta parte é um resumo da seção introdutória do Capítulo 5 do livro de Bryk e Raudenbush (1992), denominada "Pontos básicos sobre efeitos organizacionais da pesquisa".

### Vício de agregação

Pode ocorrer quando a variável tem significados diferentes e, portanto, pode ter efeitos distintos em níveis organizacionais diversos. Por exemplo,

numa pesquisa educacional, a classe social média de uma escola pode ter um efeito sobre o desempenho do aluno diferente do efeito da classe social individual do aluno. Esta última fornece uma medida dos recursos intelectuais e materiais do ambiente familiar de cada aluno individualmente. Já a classe social média dos alunos da escola é uma proxy da medida dos recursos da escola e de seu ambiente normativo. Modelos Hierárquicos ajudam a solucionar este confundimento, fornecendo uma decomposição de qualquer relação entre variáveis, tais como desempenho e classe social, em componentes separadas no nível 1 (*indivíduo*) e no nível 2 (*organização*).

#### Desvios padrões mal estimados

Podem ocorrer com dados estruturados em vários níveis, quando não consideramos a dependência entre respostas individuais dentro da mesma organização. Tal dependência pode aparecer pelas experiências compartilhadas dentro da organização ou pela forma como os indivíduos são arrematados pela organização. Modelos Hierárquicos solucionam este problema incorporando no modelo estatístico um efeito aleatório único para cada unidade organizacional. As estimativas dos desvios padrões dependem da variabilidade destes efeitos aleatórios ou, na terminologia de pesquisas amostrais, as estimativas dos desvios padrões são ajustadas pela correlação intraclasse (ou pelo efeito do plano amostral), que decorre da amostragem por conglomerado.

#### Heterogeneidade de regressão

Pode ocorrer quando as relações entre características individuais e resultados variam ao longo das organizações. Embora este fenômeno seja, frequentemente, considerado como de distúrbio do ponto de vista metodológico, as causas da heterogeneidade da regressão são muitas vezes de interesse substantivo. Modelos hierárquicos possibilitam ao pesquisador estimar um conjunto de coeficientes de regressão para cada unidade organizacional e, então, modelar a variação de conjuntos de coeficientes entre organizações como resultados multivariados a serem explicados por fatores organizacionais.

Se nos casos citados a abordagem de análise desagregada pode ser vista como uma solução que apresenta vantagens quando comparada com as abordagens tradicionais, em outras situações essa abordagem apresenta desvantagens claras quando comparada, por exemplo, com a abordagem agregada. A seguir listamos algumas dessas situações e discutimos suas implicações para a modelagem desagregada.

#### Complexidade do Modelo Desagregado

Os exemplos de modelagem desagregada discutidos anteriormente e na

maioria dos livros sobre modelos hierárquicos são relativamente simples, ao menos em termos do número de variáveis consideradas. Apesar disso, representam situações mais complexas que as cobertas pelos pacotes padrões até recentemente, e frequentemente requerem o emprego de pacotes ou procedimentos especializados para seu ajuste e análise.

Quando a modelagem for feita com a finalidade de incorporar aspectos do planejamento amostral, tais como estratificação, conglomeração e probabilidades desiguais de inclusão, a situação desejável é incorporar na formulação do modelo as informações necessárias para que o plano amostral seja ignorável na etapa de estimação dos parâmetros. Mesmo quando se pode incorporar no modelo as informações sobre a estrutura populacional, há casos nos quais o plano amostral é não ignorável e pesos precisam ser incorporados para ajustar o modelo (veja Pfeffermann et al., 1998).

Para poder incorporar no modelo as informações sobre a estrutura populacional e/ou sobre o plano amostral, é geralmente necessário considerar variáveis indicadoras de pertinência a estratos, medidas de tamanho usadas para definir as probabilidades de inclusão e também informações sobre a estrutura de conglomeração da população e da amostra. Incluir todas essas variáveis num modelo pode apresentar desafios não triviais ao analista: a especificação detalhada da forma do modelo, a estimação de seus inúmeros parâmetros dada apenas uma amostra das unidades da população, a interpretação das estimativas dos parâmetros e o diagnóstico do ajuste efetuado podem todas se tornar tarefas bastante complexas. Essa dificuldade é também mencionada por Skinner, Holt e Smith (1989, p. 9).

#### Disponibilidade da Informação Desagregada

Outra dificuldade da abordagem desagregada é que esta abordagem requer conhecimento detalhado das variáveis consideradas no planejamento amostral, tais como as identidades dos estratos, conglomerados (em vários níveis) e probabilidades de seleção (possivelmente nos vários estágios de amostragem) para cada unidade amostral. Tais informações muitas vezes não estão disponíveis por razões de proteção da confidencialidade das informações ou outras razões práticas. Modelagem de dados de pesquisas amostrais por analistas secundários é geralmente realizada em condições em que as informações sobre o plano amostral são parcial ou completamente ignoradas. Este é o caso das várias pesquisas para as quais já existe a prática de disseminar arquivos de microdados nos quais, entretanto, as informações de identificação do plano amostral são omitidas (total ou parcialmente) para evitar a revelação indesejada de informações individuais "sensitivas".

Estas dificuldades não podem ser ignoradas quando se optar por uma

abordagem desagregada para analisar dados de pesquisas amostrais complexas. Sua consideração foi uma das razões que nos levou a discutir neste livro com maior detalhe a abordagem agregada, que também depende do acesso a informações como as citadas aqui, mas que pode ser aplicada nalguns casos em que uma abordagem desagregada seria impossível. Para citar um exemplo, numa pesquisa amostral conglomerada em dois ou mais estágios, quando se pretende estimar variâncias pelo método do conglomerado primário numa abordagem agregada de análise, basta conhecer estratos e pertinência a Unidades Primárias de Amostragem, bem como os pesos das unidades individuais. Esse conhecimento pode ser insuficiente para permitir a modelagem de todos os níveis da hierarquia na população, numa abordagem desagregada.

Apesar desta dificuldade, entretanto, há muitas situações em que uma abordagem desagregada pode oferecer alternativa adequada de análise, não podendo ser desprezada e devendo figurar no arsenal de que dispõe o analista para interpretar os dados da melhor maneira possível. Os progressos recentes nas técnicas e pacotes de modelagem hierárquica têm levado essas técnicas cada vez mais para o domínio da aplicação prática, e a maior disponibilidade de resultados de pesquisas amostrais na forma de arquivos de microdados deve contribuir com essa tendência. Para isso é imprescindível que as agências produtoras de dados estatísticos baseados em pesquisas (amostrais ou mesmo censitárias) passem a fornecer nesses arquivos de microdados as informações sobre a estrutura populacional necessárias à modelagem. Isto é um desafio pois precisa ser feito sem permitir que ocorra a revelação de informações sensíveis individuais, e requer o uso de técnicas apropriadas.

## Capítulo 10

# Pacotes para Análise de Dados Amostrais

### 10.1 Introdução

Os métodos usados na coleta dos dados de pesquisas por amostragem introduzem uma complexidade na análise, que deve ser considerada na obtenção de estimativas dos parâmetros de interesse e de seus níveis de precisão associados. Ao longo deste texto foi discutido o impacto causado pela complexidade do plano amostral sobre as análises estatísticas. Foi dada ênfase em mostrar como a utilização das técnicas de análise estatística disponíveis nos pacotes estatísticos padrões de uso generalizado podem conduzir a conclusões incorretas. Foram também sugeridos ajustes dos procedimentos para o caso de dados amostrais complexos, que muitas vezes requerem pacotes especializados para serem adotados. Neste capítulo fazemos breve revisão dos pacotes computacionais especializados para a análise de dados de pesquisas amostrais complexas.

### 10.2 Pacotes Computacionais

Hoje em dia estão disponíveis diversos pacotes especializados para analisar dados obtidos através de pesquisas amostrais. Vários aspectos importantes podem diferenciá-los, tais como: ambiente computacional; método de estimação de variância; abrangência de planos amostrais que podem ser tratados; elenco de técnicas estatísticas disponíveis, etc.

Carlson(1998) cita alguns aspectos importantes que influenciam na escolha de pacote computacional especializado, tanto de um ponto de vista prá-

tico quanto da facilidade de uso. De um ponto de vista prático, um pacote deve idealmente:

- operar num ambiente computacional familiar ao usuário;
- conter as técnicas de análise estatística requeridas;
- ser capaz de tratar conjuntos de dados criados por pacote estatístico padrão, base de dados ou planilha, bem como arquivos em formato de texto (ASCII).

Carlson(1998) ressalta ainda que quanto mais fácil o uso do pacote, mais fácil o seu uso inadequado. Menciona também outras características importantes, sugerindo que do ponto de vista da facilidade de uso um pacote deve ainda:

- ter documentação bem redigida;
- ter capacidade de lidar com planos amostrais não-padrões;
- ter documentação técnica detalhada e completa, incluindo as fórmulas usadas para as estimativas pontuais e respectivas estimativas de variância.

Para analisar dados de pesquisas amostrais, é comum criar arquivos de dados através de pacotes de uso geral tais como SAS, SPSS ou outro, e depois importá-los para uso em pacotes especializados. Pode haver, ainda, necessidade de utilizar o pacote padrão tendo como entrada deste os resultados gerados pelo pacote especializado. Sem dúvida, tais tarefas seriam facilitadas caso os pacotes de uso geral contivessem ferramentas de análise apropriadas para dados de pesquisas amostrais complexas, o que não ocorre na maioria dos casos. Uma exceção à regra parece ser o caso do pacote STATA (descrito mais adiante), que já vem com um conjunto de funções ou procedimentos para análise de dados amostrais complexos integrados à parte básica do pacote (veja Stata, 1997, cap. 36).

Por outro lado, a utilização de qualquer um desses programas especializados só se torna possível se forem incluídas no arquivo de dados variáveis que informem a estrutura do plano amostral, identificando ao menos o estrato, a UPA e o peso de cada unidade da amostra. Além disso, para maior facilidade do usuário, o arquivo deve ser ordenado por estrato e também por UPA dentro de estrato.

Essas informações devem ser fornecidas pela agência produtora dos dados. Se isto não for possível por razões de sigilo ou outras razões práticas, a agência deve prover, quando solicitada, informações tais como desvios padrões e/ou coeficientes de variação e/ou efeitos de plano amostral das estimativas de interesse. Alternativamente, pode fornecer mecanismos abreviados ou aproximados de avaliação da precisão das estimativas, tais como funções de variância generalizadas (do inglês *generalised variance functions*, veja Wolter, 1985, cap. 5) ou então tabelas com estimativas dos desvios padrões, CVs ou EPAs para uma grande quantidade de variáveis, ou divulgar ao menos o efeito de plano amostral médio para certos tipos de variáveis e para certos domínios de estudo.

Por último, a documentação dos arquivos de microdados de uso público deve sempre conter avisos sobre a necessidade de considerar o plano amostral no cálculo de estimativas. Para que tais avisos sejam efetivos e possam ser acatados, os usuários devem ter acesso ao conhecimento detalhado das características do plano amostral, incluindo:

- a estratificação utilizada;
- os estágios de amostragem;
- os mecanismos de seleção em cada estágio, inclusive se as unidades foram selecionadas com ou sem reposição;
- as probabilidades de seleção em cada estágio, sejam iguais ou distintas;
- as escalas de mensuração das variáveis, se contínuas, categóricas ou ordinais;
- as categorias e escalas de resposta, no caso de variáveis categóricas ou ordinais.

Os pacotes especializados disponíveis diferem, ainda, quanto à abrangência de métodos de análise estatística. Alguns estimam as variâncias amostrais e estatísticas relacionadas como efeitos de plano amostral, efeitos de especificação incorreta, homogeneidade intraconglomerado, só para estimadores de médias, totais e proporções para a totalidade da amostra, para domínios e diferenças entre domínios. Outros estimam também variâncias de estatísticas na regressão e na regressão logística. Quase todos fornecem testes estatísticos baseados nessas variâncias amostrais. Poucos calculam estimativas de variâncias e estatísticas de teste associadas em análise de sobrevivência, tabelas de contigência, modelos de equações generalizadas de estimação e razões padronizadas.

## 158CAPÍTULO 10. PACOTES PARA ANÁLISE DE DADOS AMOSTRAIS

A seguir transcrevemos do artigo Lepkowski e Bowles(1996) uma lista dos pacotes especializados mais utilizados para análise de dados de pesquisas amostrais.

### SUDAAN

Statistical Software Center  
Research Triangle Institute  
3040 Cornwallis Road  
Research Triangle Park  
NC 27709-2194  
USA  
e-mail: SUDAAN@rti.org  
internet: [www.rti.org/patents/sudaan.html](http://www.rti.org/patents/sudaan.html)

SUDAAN (sigla de *SURvey DATA ANalysis*) é um pacote computacional para análise de dados correlacionados, incluindo dados de pesquisas amostrais complexas. Possibilita a estimação de várias características populacionais e de seus erros amostrais, incluindo médias, proporções, razões, quantis, tabelas cruzadas, razões de vantagens (do inglês *odds ratios*), além de modelos de regressão linear e logística, modelos de riscos proporcionais e análise de tabelas de contingência.

SUDAAN usa aproximações de linearização de Taylor para estimação de variâncias, e permite também empregar o método do conglomerado primário. Permite tratar o caso de seleção de unidades de primeiro estágio com ou sem reposição, incluindo componentes de variância, bem como planos de amostragem aleatória simples e amostragem estratificada de unidades elementares. SUDAAN está disponível para PCs sob DOS e também sob Windows. Também estão disponíveis versões para computadores de grande porte. Os preços variam em função do tipo de instituição, tipo e número de licenças. Por exemplo, o preço de uma só licença nova da versão 6.53 de PC do SUDAAN para empresas comerciais e agências governamentais é US\$995 e a versão 7.0 de Windows custa US\$1495.

### Stata

Stata Corporation  
702 University Drive East  
College Station  
TX 77840  
USA  
e-mail: [stata@stata.com](mailto:stata@stata.com)  
internet: [www.stata.com](http://www.stata.com)

Stata é um sistema computacional programável de análise estatística, que recentemente introduziu comandos para o cálculo de estimativas de desvios padrões de várias estatísticas para dados amostrais complexos. O programa está disponível em ambientes DOS e Windows com comandos por teclado. Telas e menus de ajuda estão disponíveis na versão em Windows. Stata usa aproximação de linearização de Taylor para estimação de variâncias. Seu preço de lista é US\$945 para usuários comerciais e US\$395 para usuários acadêmicos.

Os comandos atuais de análise incluem `svymean`, `svytotal`, `svyratio`, and `svyprop` para estimação de médias, totais, razões e proporções, além dos comandos `svyreg`, `svylogit`, e `svyprobt` para análise de regressão linear, logística e probit respectivamente. Os comandos `svylc` and `svytest` permitem a estimação de combinações lineares de parâmetros e testes de hipóteses. O comando `svydes` possibilita ao usuário descrever o plano amostral específico adotado e deve ser usado antes de qualquer dos comandos de estimação e análise citados anteriormente.

Há intenção de acrescentar comandos para estimar funções de distribuição e quantis, análise de tabelas de contingência, recursos para compensação de dados ausentes e outras análises.

#### WesVarPC

Westat, Inc.

1650 Research Blvd.

Rockville, MD 20850-3129

USA

e-mail: [WESVAR@westat.com](mailto:WESVAR@westat.com)

internet: [www.westat.com/wesvarpc/index.html](http://www.westat.com/wesvarpc/index.html)

WesVarPC é um sistema computacional estatístico projetado pela Westat, Inc. para análise de dados de pesquisas amostrais complexas. O programa opera em ambiente Windows (3.1, 3.11, e 95) e é completamente comandado por menus. Seu plano amostral básico é estratificado com vários estágios de conglomeração. WesVarPC usa o método do conglomerado primário combinado com técnicas de replicação para estimação de variâncias, incluindo os métodos de *jackknife*, meias amostras balanceadas (do inglês *balanced half samples*), e a modificação de Fay do método de meias amostras balanceadas. Os dados podem ser lidos em arquivos formato ASCII, DBF, SPSS para Windows, SAS Transport, ou formato PC SAS para DOS.

WesVarPC requer que uma nova versão do conjunto de dados seja criada num formato especial WesVarPC. Para isto é necessário especificar réplicas

e, se a pós-estratificação for incorporada na estimação de variâncias, pesos de réplicas devem também ser criados. WesVarPC permite a análise de tabelas de contingência, regressão linear e regressão logística. Há um sistema completo de comandos por menu para criar novas variáveis, o que amplia o conjunto de estatísticas possíveis de usar no WesVarPC. A saída tem formato de lista com uma linha para cada estatística. Este formato é adequado para publicação, e pode ser arquivado para processamento em planilha ou em outro programa.

### CENVAR

International Programs Center

U.S. Bureau of the Census

Washington, DC 20233-8860, USA

e-mail :IMPS@census.gov

internet : [www.census.gov/ftp/pub/ipc/www/imps.html](http://www.census.gov/ftp/pub/ipc/www/imps.html)

CENVAR é um componente do sistema computacional estatístico IMPS (*Integrated Microcomputer Processing System*) para apuração, gerenciamento e análise de dados de pesquisas complexas. Pode ser utilizado com os seguintes planos amostrais: amostragem aleatória simples; amostragem estratificada; e amostragem de conglomerados em vários estágios com probabilidades iguais ou distintas de seleção. Estes planos amostrais são todos tratados através do método do conglomerado primário combinado com a aproximação de linearização de Taylor para estimação de variâncias. CENVAR é uma versão parcial do programa PC CARP, desenvolvido pela Iowa State University, que descrevemos mais adiante.

CENVAR pode ser obtido gratuitamente através do endereço internet fornecido. Os dados devem ser lidos de arquivos em formato ASCII, com uso de um dicionário IMPS.

CENVAR pode produzir desvios padrões para estimativas de médias, proporções e totais para toda a amostra bem como para domínios especificados num formato tabular. Além disso, fornece desvios padrões, limites de confiança de 95%, coeficientes de variação, efeitos de plano amostral e também tamanhos de amostras considerados nos cálculos (frequências não expandidas).

**PC CARP**

Sandie Smith

Statistical Laboratory

219 Snedecor Hall

Iowa State University

Ames, IA 50011

USA

e-mail : sandie@iastate.edu

internet: [www.statlib.iastate.edu/survey/software/pccarp.html](http://www.statlib.iastate.edu/survey/software/pccarp.html)

PC CARP é um programa para computadores tipo PC desenvolvido pela *Iowa State University* para implementar métodos de análise de dados amostrais complexos (seu nome vem da sigla em inglês *CARP* - *Complex Analysis Regression Program*).

PC CARP pode ser usado para estimar desvios padrões de estimativas de totais, médias, proporções, quantis, razões e diferenças de razões, além de frequências e estatísticas de teste para tabelas de duas entradas. PC CARP é completado por um conjunto de três outros programas que ampliam o escopo de análises disponíveis: PC CARPL para regressão logística; POSTCARP para estimativas de totais, razões e diferenças de razões via pós-estratificação; e EV CARP para análise de regressão considerando erros de medição nas variáveis preditoras. O programa opera em um ambiente DOS com comandos por teclado. Os programas são projetados para lidar com amostras estratificadas de conglomerados em vários estágios, e com correção de população finita para até dois estágios de seleção. PC CARP usa a aproximação de linearização de Taylor para estimação de variâncias. O conjunto de programas pode ser adquirido do *Statistical Laboratory* da *Iowa State University* por US\$300. Os dados devem ser lidos em arquivo formato ASCII, mediante a criação de um dicionário próprio.

**VPLX**

Robert E. Fay

Room 3067, Bldg. 3

U.S. Bureau of the Census

Washington, DC 20233-9001

USA

e-mail: rfay@census.gov

internet: [www.census.gov/sdms/www/vwelcome.html](http://www.census.gov/sdms/www/vwelcome.html)

VPLX é um programa isolado para estimação de variâncias, projetado e usado pelo *US Bureau of the Census* para dados de pesquisa amostrais

complexas. Opera em ambiente DOS com comandos pelo teclado. O VPLX é fundamentalmente projetado para amostras estratificadas em vários estágios, e adota o método do conglomerado primário, combinado com técnicas de replicação para estimação de variâncias, incluindo procedimentos baseados nos métodos de grupos aleatórios, de *jackknife*, e de replicação balanceada. O VPLX pode ser obtido gratuitamente no endereço internet. Os dados devem ser lidos em arquivos formato ASCII mediante a criação de um dicionário próprio.

VPLX pode produzir desvios padrões para estimativas de médias, proporções e totais, tanto para a totalidade da amostra como para domínios especificados.

### CLUSTERS

Vijay Verma

World Fertility Survey

105 Park Road, Teddington (Middlesex), TW11 OAW, United Kingdom

e-mail: vjverma@essex.ac.uk

CLUSTERS é um programa isolado desenvolvido originalmente pela equipe da *World Fertility Survey* e depois aperfeiçoado por Vijay Verma e Mick Verma. O principal plano amostral é amostragem estratificada de conglomerados em vários estágios. CLUSTERS usa o método do conglomerado primário combinado com a aproximação de linearização de Taylor para estimação de variâncias. Os dados devem ser lidos de arquivos em formato ASCII, mediante a criação de um dicionário de formato próprio. CLUSTERS pode produzir estimativas de desvios padrões para médias e proporções, para toda a amostra bem como para domínios, e também para diferenças entre domínios especificados num formato tabular. Além dos desvios padrões, CLUSTER fornece estimativas dos coeficientes de variação, dos efeitos de plano amostral e tamanhos de amostras considerados nos cálculos (frequências não expandidas), bem como estatísticas de correlações intraclasse.

### Epi Info

Andrew G. Dean, MD

Epidemiology Program Office, Mailstop C08

Centers for Disease Control and Prevention

Atlanta, GA 30333

U.S.A

e-mail: AGD1@epo.em.cdc.gov ou EpiInfo@cdc1.cdc.gov

internet: [www.cdc.gov/epo/epi/epi.html](http://www.cdc.gov/epo/epi/epi.html)

Epi Info é um pacote estatístico para epidemiologia, desenvolvido pelo *US Centers for Disease Control and Prevention*, para apuração, gerenciamento e análise de dados epidemiológicos, incluindo análise de dados de pesquisas amostrais complexas (componente CSAMPLE). Seu plano amostral básico é amostragem de conglomerados em vários estágios, através do método do conglomerado primário combinado com a aproximação de linearização de Taylor para estimação de variâncias.

Epi Info pode ser obtido gratuitamente do endereço internet fornecido. A leitura de dados pode ser feita de arquivos em formatos DBF, Lotus, ou ASCII. O pacote pode produzir estimativas de desvios padrões para estimativas de médias e proporções, tanto para a totalidade da amostra como para domínios especificados através de tabelas de duas entradas. A saída inclui apenas frequências não expandidas, proporções e médias expandidas, desvios padrões, limites de confiança de 95% e efeitos de plano amostral.



# Bibliografia

- [1] Agresti A. (1990). *Categorical Data Analysis*. Nova Iorque: Wiley.
- [2] Albieri S. e Bianchini Z.M. (1997). *Aspectos de amostragem relativos à pesquisa domiciliar sobre padrões de vida*. Rio de Janeiro: IBGE, Departamento de Metodologia, mimeo.
- [3] Bickel P.J. e Doksum K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. São Francisco: Holden-Day.
- [4] Binder D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, p. 279-292.
- [5] Binder D.A., Kovar J.G., Kumar S., Paton D. e Baaren A.V. (1987). Analytic uses of survey data: a review. Em MacNeill I.B. e Umphrey G.J. (eds), *Applied Probability, Stochastic Processes and Sampling Theory*, p. 243-264.
- [6] Bishop Y.M.M., Fienberg S.E. e Holland P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: The MIT Press.
- [7] Brewer K.W.R. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, p. 911-915.
- [8] Bryk A.S. e Raudenbush S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage.
- [9] Carlson B.L. (1998). Software for statistical analysis of sample survey data. A aparecer em: Armitage P. e Colton T. (eds), *Encyclopaedia of Biostatistics*, Wiley.

- [10] Cassel C.M., Särndal C.E. e Wretman J.H. (1977). *Foundations of Inference in Survey Sampling*. Nova Iorque: Wiley.
- [11] Chambers R.L. (1986). Design-adjusted parameter estimation. *Journal of the Royal Statistical Society A*, 149, p. 161-173.
- [12] Chambers R.L. (1995). *Regression analysis with sample survey data*. Manuscrito inédito cedido por cortesia do autor.
- [13] Cochran W.G. (1977). *Sampling Techniques*, Third Edition. Nova Iorque: Wiley.
- [14] Cox D.R. e Hinkley D.V. (1974). *Theoretical Statistics*. Londres: Chapman & Hall.
- [15] Deming W.E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, 51, p. 24-53.
- [16] Diggle P.J., Liang K.Y. e Zeger S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- [17] Draper N.R. e Smith H. (1981). *Applied Regression Analysis*, Second Edition. Nova Iorque: Wiley.
- [18] Fellegi I.P. (1980). Approximate tests of independence and goodness-of-fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, p. 261-268.
- [19] Freitas M.P.S., Duarte R.P.N., Pessoa D.G.C., Albieri S. e Nascimento Silva P.L.D. (1997). *Comparando distribuições etárias em pesquisas por amostragem: PNAD 95 e PPV 96/97*. Rio de Janeiro: IBGE, Departamento de Metodologia, mimeo.
- [20] Fuller W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, p. 117-132.
- [21] Fuller W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, p. 97-118.
- [22] Garthwaite P.H., Jolliffe I.T. e Jones B. (1995). *Statistical Inference*. Nova Iorque: Prentice Hall.
- [23] Haggard E.A. (1958). *Intraclass Correlation and the Analysis of Variance*. Nova Iorque: Dryden Press.

- [24] Hájek J. (1960). Limiting distributions in simple random sampling from finite populations. *Pub.Math. Inst. Hung. Acad. Sci.*, 5, p. 361-374.
- [25] Hansen M.H., Hurwitz W.N. e Madow W.G. (1953). *Sample Survey Methods and Theory*. Nova Iorque: Wiley.
- [26] Holt D. (1989). Introduction to part C. Em: Skinner C.J. Holt D. e Smith T.M.F. (eds), *Analysis of Complex Surveys*, Chichester, Wiley, p. 209-215.
- [27] Holt D. Scott A. e Ewings P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society A*, 143, p. 303-320.
- [28] Holt D., Smith T.M.F. e Winter P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A*, 143, p. 474-487.
- [29] IBGE (1981). *Metodologia da Pesquisa Nacional por Amostra de Domicílios na Década de 70*. Rio de Janeiro: IBGE, Série Relatórios Metodológicos, 1.
- [30] IBGE (1985). *Amostra de Uso Público do Censo Demográfico de 1980 - Metodologia e Manual do Usuário*. Rio de Janeiro: IBGE.
- [31] Isaki C.T. e Fuller W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, p. 89-96.
- [32] Johnson R.A. e Wichern D.W. (1988). *Applied Multivariate Statistical Analysis*, Second Edition. Englewood Cliffs, New Jersey: Prentice Hall.
- [33] Kalton G. (1983a). *Compensating for missing survey data*. Ann Arbor, Michigan: The University of Michigan, Institute for Social Research, Survey Research Center.
- [34] Kalton G. (1983b). Models in the practice of survey sampling. *International Statistical Review*, 51, p. 175-188.
- [35] Kish L. (1965). *Survey Sampling*. Nova Iorque: Wiley.
- [36] Kish L. e Frankel M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society B*, 36, p. 1-37.
- [37] Lehtonen R. e Pahkinen E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

- [38] Leote R.M.D. (1996). *Um perfil sócio-econômico das pessoas ocupadas no setor informal na área urbana do Rio de Janeiro*. Rio de Janeiro: IBGE, Escola Nacional de Ciências Estatísticas, Relatório Técnico 02/96.
- [39] Lepkowski J. e Bowles J. (1996). Sampling error software for personal computers. *The Survey Statistician*, 35, p. 10-17.
- [40] Little R.J.A. e Rubin D.B. (1987). *Statistical Analysis with Missing Data*. Nova Iorque: Wiley.
- [41] Longford N.T. (1993). *Random Coefficient Models*. Oxford: Clarendon Press.
- [42] Mahalanobis P.C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhyā*, 4, p. 511-531.
- [43] Mahalanobis P.C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London B*, 231, p. 329-451.
- [44] Montanari G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55, p. 191-202.
- [45] Nascimento Silva P.L.D. (1996). *Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys*. Southampton: University of Southampton, Department of Social Statistics, Tese de Doutorado.
- [46] Nascimento Silva P.L.D. e Moura F.A.S. (1990). *Efeitos de conglomeração da malha setorial do censo demográfico 80*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Série Textos para Discussão, 32.
- [47] Nathan G. e Holt D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B*, 42, p. 377-386.
- [48] 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, p. 558-606.
- [49] Pessoa D.G.C., Nascimento Silva P.L.D. e Duarte R.P.N. (1997). Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. *Revista Brasileira de Estatística*, a aparecer.
- [50] Pfeiffermann D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61, p. 317-337.

- [51] Pfeffermann D. e Nathan G. (1981). Regression analysis of data from complex samples. *Journal of the American Statistical Association*, 76, p. 681-689.
- [52] Pfeffermann D., Skinner C. J., Holmes D.J., Goldstein H.e Rasbash J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, 60, p. 23-40.
- [53] Quenoille M.H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, 20, p. 355-375.
- [54] Quenoille M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, p. 353-360.
- [55] Rao J.N.K. e Scott A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two way tables. *Journal of the American Statistical Association*, 76, p. 221-230.
- [56] Robinson P.M. e Särndal C.E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā B*, 45, p. 240-248.
- [57] Rubin D.B. (1976). Inference and missing data. *Biometrika*, 63, p. 581-592.
- [58] Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Nova Iorque: Wiley.
- [59] Särndal C.E., Swensson B. e Wretman J.H. (1992). *Model Assisted Survey Sampling*. Nova Iorque: Springer-Verlag.
- [60] Satterthwaite F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, p. 110-114.
- [61] Shah B.V., Barnwell B.G., Hunt P.N. e LaVange L.M. (1992). *SUDAAN User's Manual - Professional Software for SURvey DATA ANalysis for multi-stage sample designs - release 6.0*. Research Triangle Park, NC: Research Triangle Institute.
- [62] Shah B.V., Folsom R.E., LaVange L.M., Wheelless S.C., Boyle K.E. e Williams R.L. (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*. Research Triangle Park, NC: Research Triangle Institute.

- [63] Skinner C. J. (1989a). Introduction to part A. Em: Skinner C.J., Holt D. e Smith T.M.F. (eds) *Analysis of Complex Surveys*, Chichester, Wiley, p. 23-57.
- [64] Skinner C. J. (1989b). Domain Means, Regression and Multivariate Analysis. Em: Skinner C.J., Holt D. e Smith T.M.F. (eds), *Analysis of Complex Surveys*, Chichester, Wiley, p. 59-87.
- [65] Skinner C.J., Holt D. e Smith T.M.F. (eds) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.
- [66] Stata (1997). *Stata User's Guide, Release 5*. College Station, Texas: Stata Press.
- [67] Sudman S. (1976). *Applied Sampling*. Nova Iorque: Academic Press.
- [68] Sugden R.A. e Smith T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, p. 495-506.
- [69] Thomas D.R e Rao J.N.K. (1987). Small-sample comparison of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, p. 630-636.
- [70] US Bureau of Labor Statistics (1984). *BLS Handbook of Methods - Volume II - The Consumer Price Index*. Washington DC: BLS Bulletin 2134-2.
- [71] Westat (1996). *A User's Guide to WesVarPC, version 2.0*. Rockville, Mariland: Westat, Inc.
- [72] Wolter K.M. (1985). *Introduction to Variance Estimation*. Nova Iorque: Springer-Verlag.