Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-10-02

Sumário

Prefácio

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

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

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

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

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

Agradecimentos

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

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

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

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

Capítulo 1

Introdução

1.1 Motivação

Este livro trata de problema de grande importância para os 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 e ajuste 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.) **calibração ou imputação para não-resposta e outros ajustes**.

As estimativas pontuais de parâmetros da população ou de modelos são influenciadas por pesos distintos das observações. Além disso, as estimativas de variância (ou da precisão dos estimadores) são influenciadas pela conglomeração, estratificação e pesos, ou no caso de não resposta, também por eventual imputação de dados faltantes. 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) acima 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 sobre Padrões de Vida (PPV), realizada pelo IBGE nos anos 1996-97. (Albieri and Bianchini, 1997) descrevem resumidamente a Pesquisa sobre Padrões de Vida (PPV), que 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,

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

Estrato	População	Amostra
RM Fortaleza	2263	62
RM Recife	2309	61
RM Salvador	2186	61
Restante Nordeste Urbano	15057	61
Restante Nordeste Rural	23711	33
RM Belo Horizonte	3283	62
RM Rio de Janeiro	10420	61
RM São Paulo	14931	61
Restante Sudeste Urbano	25855	61
Restante Sudeste Rural	12001	31
Total	112016	554

Tabela 1.2: Distribuição dos pesos da amostra da PPV

Região	Mínimo	Q1	Mediana	Q3	Máximo
Nordeste	724	1194	1556	6937	15348
Sudeste	991	2789	5429	9509	29234
Nordeste_Sudeste	724	1403	3785	8306	29234

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 empregado na seleção da amostra da PPV foi de dois estágios, com estratificação das unidades primárias de amostragem (no caso os setores censitários da base geográfica do IBGE conforme usada para o Censo Demográfico de 1991), seleção destes setores com probabilidade proporcional ao tamanho, e seleção aleatória das unidades de segundo estágio (domicílios). O tamanho da amostra para cada estrato geográfico foi fixado em 480 domicílios, e o número de setores selecionados foi fixado em 60, com 8 domicílios selecionados em cada setor. A exceção ficou por conta dos estratos que correspondem ao restante da área rural de cada Região, onde foram selecionados 30 setores e 16 domicílios por setor, em função da dificuldade de acesso a esses setores, o que implicaria em aumento de custo da coleta.

Os setores de cada um dos 10 estratos geográficos foram subdivididos em 3 estratos de acordo com a renda média mensal do chefe do domicílio 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 1991. No final foram obtidos 554 setores na amostra, distribuídos tal como revela a Tabela 1.1.

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

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, razões, etc.

Por outro lado, quando utilizamos a amostra para estudos analíticos, as opções padrão disponíveis nos pacotes estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações independentes e identicamente distribuídas (IID). Por exemplo, os procedimentos padrão disponíveis para estimar a média populacional permitem utilizar pesos distintos das observações amostrais, mas tratariam tais pesos como se fossem frequências de observações repetidas na amostra, e portanto interpretariam a soma dos pesos como tamanho amostral, situação que na maioria das vezes gera inferências incorretas sobre a precisão das estimativas, pois o tamanho da amostra é muito menor que a soma dos pesos amostrais usualmente encontrados nos arquivos de microdados de pesquisas disseminados por agências de estatísticas oficiais. Em tais pesquisas, a opção mais freqüente é disseminar pesos que somados estimam o total de unidades `da população`.

Além disso, a variabilidade dos pesos para distintas observações amostrais produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da conglomeração e estratificação - pontos ii) e iii) mencionados anteriormente.

Para exemplificar o impacto de ignorar os pesos e o plano amostral ao estimar quantidades descritivas populacionais, tais como totais, médias, proporções e razões, calculamos estimativas de quantidades desses diferentes tipos usando a amostra da PPV juntamente com estimativas das respectivas variâncias. 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 para cada uma das estimativas consideradas usando a library `survey` do R (Lumley, 2017) . Essa razão fornece uma medida do efeito de ignorar o plano amostral . Os resultados das estimativas ponderadas e variâncias considerando o plano amostral são apresentados na Tabela 1.3 , juntamente com as medidas dos efeitos de plano amostral (EPA). Exemplos de utilização da library `survey` para obtenção de estimativas apresentadas na 1.3 estão na Seção 4. As outras estimativas da Tabela 1.3 podem ser obtidas de maneira análoga.

Na Tabela 1.3 apresentamos as estimativas dos seguintes parâmetros populacionais:

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

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

Note que as variáveis e parâmetros cujas estimativas são apresentadas na Tabela 1.3 não foram escolhidas de forma a acentuar os efeitos ilustrados, mas tão somente para representar distintos parâmetros (médias, razões, totais, proporções) e variáveis de interesse. Os resultados apresentados para as estimativas de EPA ilustram bem o cenário típico em pesquisas amostrais complexas: o impacto do plano amostral sobre a

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

Parâmetro	Estimativa	DP	EPA
1.	3.62	0.05	2.64
2.	10.70	1.15	2.97
3.	1208123.00	146681.00	3.37
4.	1174220.00	127982.00	2.64
5.	4792344.00	318877.00	4.17
6.	11.87	1.18	2.46
7.	10.87	0.67	3.86
8.	10817590.00	322947.00	2.02
9.	10804511.00	323182.00	3.02
10.	709145.00	87363.00	2.03
11.	1.39	0.03	1.26
12.	0.53	0.01	1.99

inferência varia conforme a variável e o tipo de parâmetro de interesse. Note ainda que à exceção do menor valor, todas as demais estimativas de EPA apresentaram valores superiores a 2.

1.2 Objetivos do Livro

Este livro tem três objetivos principais:

1) **ilustrar e analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise de dados quando estes são provenientes de pesquisas amostrais complexas**;

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

3) **ilustrar o potencial analítico de muitas das pesquisas produzidas por agências de estatísticas oficiais para responder questões de interesse, mediante uso de ferramentas de análise estatística agora já bastante difundidas, aumentando assim o valor adicionado destas pesquisas**.

Para alcançar tais objetivos, adota-se uma abordagem fortemente ancorada na apresentação de exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também recursos do pacote estatístico R (http://www.r-project.org/). A comparação dos resultados das análises feitas das duas formas permite avaliar o impacto de não se considerar os pontos i) a iv) anteriormente citados. O ponto iv) não será tratado de forma completa neste texto. O leitor interessado na análise de dados sujeitos a não-resposta pode consultar (Kalton, 1983a), (Little and Rubin, 2002), (Rubin, 1987), (Särndal et al., 1992), ou Schafer (1997), por exemplo.

Estrutura do Livro

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

No segundo capítulo, procuramos dar uma visão das diferentes abordagens utilizadas na análise estatística de dados de pesquisas amostrais complexas. Apresentamos um referencial para inferência com ênfase no `Modelo de Superpopulação` que incorpora, de forma natural, tanto uma estrutura estocástica para descrever a geração dos dados populacionais (modelo) como o plano amostral efetivamente utilizado para obter os dados

amostrais (plano amostral). As referências básicas para seguir este capítulo são cap2 em (Nascimento Silva, 1996), cap1 em (Skinner et al., 1989) e caps 1 e 2 em (Chambers and Skinner, 2003).

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 no Capítulo 2.

No Capítulo 3 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 caps 2 e 3 em (Särndal et al., 1992), (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 de aleatorização induzida pelo plano amostral efetivamente utilizado (plano amostral verdadeiro). Aqui a referência principal foi o livro (Skinner et al., 1989), complementado com o texto de (Lehtonen and 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 et al., 1989), (Pfeffermann, 1993), (Binder, 1983) e cap.6 em (Nascimento Silva, 1996).

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, a library `survey` (Lumley, 2017), e através de um programa de uso geral, a library `glm` do R. As referências centrais são cap6 em (Nascimento Silva, 1996) e Binder(1983), além de (Pessoa et al., 1997).

Os Capítulos 7 e 8 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 cap. 4, (Skinner et al., 1989) e cap. 7 (Lehtonen and Pahkinen, 1995). Também são apresentadas as idéias básicas de como efetuar ajuste de modelos log-lineares a dados de frequências em tabelas de múltiplas entradas.

O Capítulo 9 trata da estimação de densidades e funções de distribuição, ferramentas que tem assumido importância cada dia maior com a maior disponibilidade de microdados de pesquisas amostrais para analistas fora das agências produtoras.

O Capítulo 10 trata da estimação e ajuste de modelos hierárquicos considerando o plano amostral. Modelos hierárquicos (ou modelos multinível) têm sido bastante utilizados para explorar situações em que as relações entre variáveis de interesse em uma certa população de unidades elementares (por exemplo, crianças em escolas, pacientes em hospitais, empregados em empresas, moradores em regiões, etc.) são afetadas por efeitos de grupos determinados ao nível de unidades conglomeradas (os grupos). Ajustar e interpretar tais modelos é tarefa mais difícil que o mero ajuste de modelos lineares mesmo em casos onde os dados são obtidos de forma exaustiva, mas ainda mais complicada quando se trata de dados obtidos através de pesquisas amostrais complexas. Várias alternativas de métodos para ajuste de modelos hierárquicos estão

disponíveis, e este capítulo apresenta uma revisão de tais abordagens, ilustrando com aplicações a dados de pesquisas amostrais de escolares.

O Capítulo 11 trata da não resposta e suas conseqüências sobre a análise de dados. As abordagens de tratamento usuais, reponderação e imputação, são descritas de maneira resumida, com apresentação de alguns exemplos ilustrativos, e referências à ampla literatura existente sobre o assunto. Em seguida destacamos a importância de considerar os efeitos da não-resposta e dos tratamentos compensatórios aplicados nas análises dos dados resultantes, destacando em particular as ferramentas disponíveis para a estimação de variâncias na presença de dados incompletos tratados mediante reponderação e/ou imputação.

O Capítulo 12 trata de assunto ainda emergente: diagnósticos do ajuste de modelos quando os dados foram obtidos de amostras complexas. A literatura sobre o assunto ainda é incipiente, mas o assunto é importante e procura-se estimular sua investigação com a revisão do estado da arte no assunto.

O Capítulo 13 discute algumas formas alternativas de analisar dados de pesquisas complexas, contrapondo algumas abordagens distintas à que demos preferência nos capítulos anteriores, para dar aos leitores condições de apreciar de forma crítica o material apresentado no restante deste livro. Entre as abordagens discutidas, há duas principais: a denominada `análise desagregada`, e a abordagem denominada `obtenção do modelo amostral` proposta por (Pfeffermann et al., 1998a). A chamada `análise desagregada` incorpora explicitamente na análise vários aspectos do plano amostral utilizado através do emprego de modelos hierárquicos (Bryk and 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, ajuste de modelos e teste de hipóteses.

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

Por último, no Capítulo 14, 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, os livros de (Skinner et al., 1989) e (Chambers and Skinner, 2003) são as referências centrais a pesquisar. Para aqueles querendo um tratamento ainda mais prático que o nosso, o livro de (Lehtonen and Pahkinen, 1995) pode ser uma opção interessante.

1.3 Laboratório de R do Capítulo 1.

Exemplo 1.2. Utilização da library survey do R para estimar totais e razões na PPV

Os exemplos a seguir utilizam dados da Pesquisa de Padrões de Vida (PPV) de 2004 do IBGE, cujo plano amostral encontra-se descrito no Exemplo 1.1 . Os dados da PPV estão no data frame ppv da library anamco.

```
# leitura dos dados
library(anamco)
dim(ppv)
```

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

```
names(ppv)
```

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

Inicialmente, vamos criar variáveis de interesse por meio de transformação das variáveis existentes no data frame ppv. Agumas dessas variáveis são:

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

```
# cria novo data frame ppv1
ppv1 <- transform(ppv,
analf1 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >= 7 & v02a08 <= 14)) * 1,
analf2 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >14)) * 1,
faixa1 = (v02a08 >= 7 & v02a08 <= 14) *1,
faixa2 = (v02a08 > 14) * 1)
```

A seguir, mostramos como utilizar a library survey (Lumley, 2017) do R para obter algumas estimativas da Tabela 1.3. Os dados da pesquisa estão contidos no data frame ppv1, que contém as variáveis que caracterizam o plano amostral:

- nsetor - conglomerados;
- estratof - estratos;
- pesof - pesos do plano amostral;

O passo fundamental para utilização da library survey (Lumley, 2017) é criar um objeto que guarde as informações relevantes do plano amostral. Isso é feito por meio da função svydesign(). As variáveis que definem estratos, conglomerados e pesos na PPV são respectivamente, estratof, nsetor e pesof. O objeto de desenho amostral, ppv.des incorpora as informações do plano amostral adotado na PPV.

```
# carrega library survey
library(survey)
# cria objeto de desenho
ppv.des<-svydesign(ids = ~nsetor, strata = ~estratof,
data = ppv1, nest = TRUE, weights = ~pesof)
```

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

```
ppv.se.des <- subset(ppv.des, regiao == 2)
```

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

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

```
svytotal(~analf1, ppv.se.des, deff = TRUE)
```

```
##          total      SE   DEff
## analf1 1174220  127982 2.0543
```

```
svytotal(~analf2, ppv.se.des, deff = TRUE)
```

```
##          total      SE   DEff
## analf2 4792344  318877 3.3237
```

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

```
svyratio(~analf1, ~faixa1, ppv.se.des)
```

```
## Ratio estimator: svyratio.survey.design2(~analf1, ~faixa1, ppv.se.des)
## Ratios=
##          faixa1
## analf1 0.118689
## SEs=
##            faixa1
## analf1 0.01178896
```

```
svyratio(~analf2, ~faixa2, ppv.se.des)
```

```
## Ratio estimator: svyratio.survey.design2(~analf2, ~faixa2, ppv.se.des)
## Ratios=
##           faixa2
## analf2 0.1086871
## SEs=
##             faixa2
## analf2 0.006732254
```

Uma alternativa para obter estimativa por domínios é utilizar a função `svyby()` da library `survey` (Lumley, 2017). Assim, poderíamos estimar os totais da variável `analf1` para as regiões `Nordeste (1)` e `Sudeste(2)` da seguinte forma:

```
svyby(~analf1, ~regiao, ppv.des, svytotal, deff = TRUE)
```

```
##   regiao  analf1       se DEff.analf1
## 1      1 3512866 352619.5    9.660561
## 2      2 1174220 127982.2    2.054345
```

Observe que as estimativas de totais e desvios padrão obtidas coincidem com as Tabela 1.3, porém as estimativas de Efeitos de Plano Amostral(DEff) são distintas.

Exemplo 1.3. Exemplo anterior usando a library srvyr

- Carrega a library `srvyr`:

```
library(srvyr)
```

- Cria objeto de desenho:

```
ppv.des <- ppv %>% as_survey_design (ids = nsetor, strata = estratof,
nest = TRUE, weights = pesof)
```

- Vamos criar novas variáveis:

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

regiao	taxa_analf1	taxa_analf1_se	taxa_analf2	taxa_analf2_se
1	0.423	0.031	0.336	0.016
2	0.119	0.012	0.109	0.007

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

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

```
res <- ppv.des %>%
group_by(regiao) %>%
summarise(
taxa_analf1 = survey_ratio(analf1, faixa1),
taxa_analf2 = survey_ratio(analf2, faixa2)
)
knitr::kable(as.data.frame(res), booktabs = TRUE, row.names = FALSE, digits = 3,
caption = "Proporção de analfabetos para faixas etárias 7-14 anos e mais de 14 anos")
```

1.3.1 Estimativa do efeito de plano amostral (EPA)

Esse assunto será tratado em detalhes no Capítulo 4 . Por enquanto, apresentaremos uma introdução necessária para compreender os valores na Tabela 1.3.

O efeito de plano amostral (EPA) de Kish é definido na fórmula (4.1). Vamos considerar o caso particular em que $\hat{\theta}$ é um estimador de total de uma variável Y.Ou seja

$$EPA_{Kish}\left(\widehat{Y}\right) = \frac{V_{VERD}\left(\widehat{Y}\right)}{V_{AAS}\left(\widehat{Y}\right)}$$

Na definição do EPA, a estimativa do numerador pode ser obtida usando-se a library `survey` (Lumley, 2017), a partir do objeto de `ppv.se.des` que incorpora as características do plano amostral utilizado para coletar os dados. Não é possível estimar diretamente o denominador, pois o plano amostral AAS (Amostragem Aleatória Simples) não foi adotado na coleta dos dados. Devemos estimar o denominador a partir de dados obtidos através do plano amostral VERD, como se eles tivessem sido obtidos através de AAS.

Supondo conhecido o tamanho da população N e a fração amostral $f = n/N$ pequena, a estimativa da variância de $\widehat{Y}$ é dada na expressão (3.9)

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

onde $\widehat{S}_y = n^{-1}\sum_{i\in s}\left(y_i - \overline{y}\right)^2$ é a estimativa de $S_y = N^{-1}\sum_{i\in U}\left(y_i - \overline{Y}\right)^2$, com $\overline{Y} = N^{-1}Y$.

No lugar dessa estimativa, vamos utilizar os pesos do plano amostral verdadeiro para estimar S_y. Vamos ainda estimar N, em geral é desconhecido, por $\widehat{N} = \sum_{i\in s} w_i$. Dessa forma obtemos a estimativa

$$\begin{aligned}\widehat{V}_{w-AAS}\left(\widehat{Y}\right) &= \widehat{N}^2\left[\sum_{i\in s} w_i\left(y_i-\overline{y}\right)^2/\widehat{N}\right]/(n-1)\\ &= \frac{\widehat{N}}{n-1}\left[\sum_{i\in s} w_i y_i^2-\left(\sum_{i\in s} w_i y_i\right)^2/\widehat{N}\right],\end{aligned}$$

onde $\overline{y}=\sum_{i\in s} w_i y_i/n$.

A expressão acima pode ser calculada facilmente através da seguinte função do R:

```
Vwaas<-function(y,w)
{
#função auxiliar usada em outras funções
#entrada:
#y - valores de variavel na amostra;
#w - pesos amostrais;
#saida:  estimativa de variância de desenho para o total (segundo o SUDAAN)

n1<-length(y)-1
wsum<-sum(y*w)
wsum2<-sum((y^2)*w)
nhat<-sum(w)
vwaas<-(nhat/n1)*(wsum2-wsum^2/nhat)
vwaas
}
```

Vamos utilizar a função `Vwaas` para estimar os valores de `Efeitos do Plano Amostral`das estimativas de totais apresentadas anteriormente. Consideremos o plano amostral `ppv.se.des` anteriormente definido. Vamos usar a função `Vwaas` para obter uma estimativa da variância do total estimado da variável `analf1`. Todos os elementos os elementos necessários estão contidos no objeto `ppv.se.des`:

```
VAAS1<- Vwaas(ppv.se.des$variables[,"analf1"],weights(ppv.se.des))
VAAS2<- Vwaas(ppv.se.des$variables[,"analf2"],weights(ppv.se.des))
```

O efeito de plano amostral da estimativa do total de `analf1` pode agora ser calculada por

```
attr(svytotal(~analf1, ppv.se.des),"var")/VAAS1
```

```
##          analf1
## analf1 2.054049
```

```
attr(svytotal(~analf2, ppv.se.des),"var")/VAAS2
```

```
##         analf2
## analf2 3.32324
```

Esses valores do EPA coincidem com os obtidos acima através da library `survey`(Lumley, 2017) e são distintos daqueles apresentados na Tabela 1.3. Para obter os valores correspondentes aos da Tabela 1.3, através da library `survey`(Lumley, 2017), vamos definir as variáveis:

```
analf1.se<-with(ppv1,((v04a01==2|v04a02==2) & (v02a08>=7&v02a08<=14))&(regiao==2))
analf2.se<-with(ppv1,((v04a01==2|v04a02==2) & (v02a08>14))&(regiao==2))
ppv.des <- update (ppv.des,analf1.se=analf1.se,analf2.se=analf2.se  )
svytotal(analf1.se,ppv.des,deff=T)
```

```
##        total      SE   DEff
```

```
## [1,] 1174220  127982 2.6426
```

```
svytotal(analf2.se,ppv.des,deff=T)
```

```
##        total      SE   DEff
## [1,] 4792344  318877 4.1667
```

Ou, alternativamente,

```
svytotal(~I(ifelse(regiao==2,analf1,0)),ppv.des,deff=T)
```

```
##                                      total      SE   DEff
## I(ifelse(regiao == 2, analf1, 0)) 1174220  127982 2.6426
```

```
svytotal(~I(ifelse(regiao==2,analf2,0)),ppv.des,deff=T)
```

```
##                                      total      SE   DEff
## I(ifelse(regiao == 2, analf2, 0)) 4792344  318877 4.1667
```

Observe que as estimativas de variância para o desenho verdadeiro (numerador do EPA) são iguais quando usamos: a variável `analf1.se` com o objeto de desenho `ppv.des` ou a variável `analf1` com o objeto `ppv.se.des`. Porém na estimativa do denominador do EPA, obtida a partir da função `Vwaas`, obtemos resultados diferentes quando usamos `analf1.se` ou `analf1`, com os pesos correspondentes. No segundo caso, a soma dos pesos não estima N. Deve-se ter o cuidado, quando estimamos em um domínio, de trabalhar com pesos cuja soma seja um estimador do tamanho da população.

Exemplo 1.4. Utilização da library survey do R para estimar taxa de desocupação para um trimestre na PNADC

- Instala library `lodown` (Damico, 2016) do github:

```
library(devtools)
install_github("ajdamico/lodown")
```

- carrega a library para ler os dados da PNADC

```
library(lodown)
```

- Baixa catálogo da PNADC com arquivos disponíveis:

```
pnadc_cat <- get_catalog( "pnadc" , output_dir =tempdir() )
```

Os microdados de interesse são terceiro trimestre de 2016. Vamos ler os microdados e salvá-los em um data frame `x`.

```
lodown( "pnadc" , subset( pnadc_cat , year == 2016 & quarter == '03' ) )
x <- readRDS( paste0( tempdir() , "/pnadc 2016 03.rds" ) )
```

vamos salvar o data frame `x` para uso posterior, :

```
saveRDS(x, file="C:/adac/pnadc/pnadc 2016 03.rds")
```

Partindo do arquivo `pnadc 2016 03.rds`, podemos recuperar o data frame `x`:

```
x <- readRDS("C:/adac/pnadc/pnadc 2016 03.rds")
```

- Carrega a library `survey`

```
library(survey)
```

- Fixa opção para caso de UPA única no estrato

```
options( survey.lonely.psu = "adjust" )
```

- Cria versão inicial de objeto de desenho:

```
pre_w <- svydesign(ids =~upa, strata=~estrato,
  weights=~v1027, data = x, nest=TRUE)
```

- Especifica totais de pós-estratos na população:

```
df_pos <-data.frame(posest=unique(x$posest),
  Freq=unique(x$v1029))
```

- Pós-estratifica objeto de desenho inicial:

```
w <-postStratify(pre_w, ~posest, df_pos)
```

Para calcular a taxa de desocupação, o IBGE considera pessoas de 14 anos ou mais na semana de referência(PIA) e calcula a razão de dois totais:

1. Numerador: total de pessoas desocupadas (vd4002==2)
2. Denominador: total de pessoas na força de trabalho (vd4001==1)

```
# estima taxa de desocupação
taxa_des <- svyratio(~ vd4002=="2" ,
  ~ vd4001 == "1" , w , na.rm = TRUE)
# organiza saída
result <- data.frame(
  100*coef(taxa_des),
  100*SE(taxa_des),
  100*cv(taxa_des)
)
row.names(result)<- NULL
names(result) <-NULL
names(result) <- c("Taxa", "Erro_Padrão", "CV")
# taxa de desocupação
result
```

```
##       Taxa Erro_Padrão        CV
## 1 11.80303   0.1174791 0.9953299
```

Exemplo 1.5. Utilização da library survey do R para análise de microdados da PNS

Leitura dos microdados usando a library `lodown`(Damico, 2016)

```
#library(lodown)
#lodown( "pns" , output_dir = "C:/adac/PNS")
```

Depois de baixar os dados, são criados os seguintes arquivos no diretório `C:/adac/PNS` :

1. `2013 all questionnaire survey design.rds` que contém o objeto de desenho para todos as pessoas na amostra;
2. `2013 all questionnaire survey.rds` que contém os microdados para todos as pessoas da amostra
3. `2013 long questionnaire survey design.rds` que contém o objeto de desenho para uma subamostra de pessoas com 18 anos ou mais que responderam um questionário mais longo;
4. `2013 long questionnaire survey.rds` que contém os microdados para uma subamostra de pessoas com 18 anos ou mais que responderam um questionário mais longo;

Observação. Nos arquivos acima, além das variáveis contidas no dicionário da PNS, foram acrescentadas variáveis derivadas, utilizadas nos exemplos no site asdfree.com

Inicialmente, vamos estimar características de pessoas com 18 anos ou mais mais que responderam o questionário longo e salvar os microdados dessa amostra no data frame `pes_sel`.

```
pes_sel <- readRDS("C:/adac/PNS/2013 long questionnaire survey.rds")
dim(pes_sel)
```

```
## [1] 60202  1019
```

Observação. No data frame pes_sel o nome das variáveis, obtidos por `names(pes_sel)`, estão em minúsculas. No dicionário da PNS os códigos correspondentes estão em maiúsculas.

O data frame `pes_sel` contém as variáveis descritas no dicionário da `PNS` e algumas variáveis obtidas derivadas.

O passo inicial para a análise dos microdados é definir um objeto de desenho que salva as características do plano amostral da pesquisa. Isso é feito por meio da função `svydesign()` da library `survey` (Lumley, 2017).

```
library(survey)
pes_sel_des <-
    svydesign(
        id = ~ upa_pns ,
        strata = ~ v0024 ,
        data = pes_sel ,
        weights = ~ pre_pes_long ,
        nest = TRUE
    )
```

Os pesos do objeto de desenho `pes_sel_des` devem ser modificados de modo que as estimativas dos totais populacionais dos pós-estratos fixados coincidam com os totais populacionais dos pós-estratos conhecidos a partir do Censo Demográfico. O data frame `post_pop` contém na primeira coluna a identidade dos pós-estratos e na segunda seus totais populacionais.

```
post_pop <- unique( pes_sel[ c( 'v00293.y' , 'v00292.y' ) ] )
names( post_pop ) <- c( "v00293.y" , "Freq" )
```

Utilizando a função `postStratify()` da library `survey` (Lumley, 2017) incorpora-se no objeto de desenho `pes_sel_des` as informações contidas no data frame `post_pop`.

```
pes_sel_des_pos <- postStratify( pes_sel_des , ~v00293.y , post_pop )
```

Salvar objeto de desenho pós-estratificado para posterior utilização:

```
saveRDS(pes_sel_des_pos, file = "C:/adac/PNS/pns_des_sel.rds" )
```

```
pes_sel_des_pos <- readRDS("C:/adac/PNS/pns_des_sel.rds")
```

Os comandos acima foram apresentados apenas para ilustrar como, a partir dos microdados da pesquisa, obtemos o objeto de desenho da pesquisa. Não seria necessária a execução desses comandos pois o objeto de desenho `pes_sel_des_pos` já se encontra disponível no arquivo 1 `2013 all questionnaire survey design.rds`.

Para exemplificar, vamos agora reproduzir estimativas na Tabela 6.42.1.1 da publicação em ftp://ftp.ibge.gov.br/PNS/2013/pns2013.pdf.

A variável de interesse tem código Q092 e sua descrição é: ``Algum médico ou profissional de saúde mental (como psiquiatra ou psicólogo) já lhe deu o diagnóstico de depressão?''

Essa variável é de classe `character` e tem dois valores ``1'' e ``2''. Vamos definir uma nova variável `diag_dep` que é igual a 1 se recebe diagnóstico de depressão e 0 caso contrário:

```
diag_dep <- as.numeric (pes_sel$q092=="1")
```

Como mencionado, o objeto de desenho pós-estratificado `pes_sel_des_pos` pode ser lido do arquivo `2013 all questionnaire survey design.rds`. Vamos atualizar o objeto de desenho `pes_sel_des_pos` para incluir a

Tabela 1.5: Proporção de diagnóstico de depressão por sexo

sexo	prop	l.i.c.	l.s.c
masculino	3.9	3.5	4.4
feminino	10.9	10.3	11.6

Tabela 1.6: Proporção de diagnóstico de depressão por situação

situ	prop	l.i.c.	l.s.c
urbano	8.0	7.5	8.4
rural	5.6	4.9	6.3

variável `diad_dep` na componente `variables`:

```
pes_sel_des_pos <- update (pes_sel_des_pos, diag_dep = diag_dep )
```

Para calcular a proporção de pessoas que receberam diagnóstico de depressão para o país inteiro usamos o seguinte comando:

```
diagdepbr <- svymean(~diag_dep, pes_sel_des_pos)
# estimativa em % e erro padrão
round(100* coef(diagdepbr),1)
```

```
## diag_dep
##      7.6
```

```
round(100*SE(diagdepbr),1)
```

```
##          diag_dep
## diag_dep      0.2
```

e para obter um intervalo de confiança aproximado com nível de confiança de 95%, usamos:

```
round(100* c(coef(diagdepbr) - 2*SE(diagdepbr),coef(diagdepbr) + 2*SE(diagdepbr)  ), 1)
```

```
## [1] 7.2 8.1
```

As estimativa de proporção e os limites do intervalo de confiança podem ser obtidos por meio da função:

```
three_stats <- function(z) round(100 * c(coef(z), coef(z) -
  2 * SE(z), coef(z) + 2 * SE(z)), 1)
```

Aplicando a função `three_stats` para estimar a proporção para o país inteiro:

```
three_stats(diagdepbr)
```

```
## diag_dep
##      7.6      7.2      8.1
```

que coincidem com os resultados já obtidos.

Para o país por sexo:

Por situação (rural e urbano)

Por Unidade da Federação:

Usando o objeto de desenho para todas as pessoas que reponderam o questionário curto, salvo no arquivo

Tabela 1.7: Proporção de diagnóstico de depressão por uf

uf	prop	l.i.c.	l.s.c
Rondonia	5.6	4.1	7.2
Acre	5.8	4.5	7.2
Amazonas	2.7	1.9	3.5
Roraima	4.4	3.2	5.7
Para	1.6	1.0	2.1
Amapa	3.4	2.0	4.7
Tocantins	7.1	5.2	8.9
Maranhao	3.8	2.4	5.1
Piaui	3.9	2.8	5.1
Ceara	4.4	3.3	5.4
Rio Grande do Norte	6.9	5.4	8.4
Paraiba	4.8	3.4	6.2
Pernambuco	7.2	5.7	8.6
Alagoas	6.2	4.6	7.9
Sergipe	6.2	4.9	7.5
Bahia	4.0	2.7	5.3
Minas Gerais	11.1	9.0	13.1
Espirito Santo	5.5	3.7	7.2
Rio de Janeiro	6.0	5.0	7.0
Sao Paulo	8.4	7.3	9.5
Parana	11.7	9.4	14.0
Santa Catarina	12.9	9.7	16.0
Rio Grande do Sul	13.2	11.5	15.0
Mato Grosso do Sul	8.8	7.3	10.4
Mato Grosso	6.9	5.1	8.7
Goias	7.1	5.7	8.5
Distrito Federal	6.2	4.9	7.5

Tabela 1.8: Proporção de pessoas com seguro de saúde por nível de instrução

	Prop	l.i.c.	l.s.c.
SinstFundi	0.164	0.157	0.171
FundcMedi	0.228	0.217	0.238
MedcSupi	0.374	0.363	0.384
Supc	0.688	0.672	0.704

em `2013 all questionnaire survey design.rds`, vamos estimar a proporção de pessoas que possuem plano de saúde por grupos de nível de instrução.

```
pes_all_des_pos <- readRDS("C:/adac/PNS/2013 all questionnaire survey design.rds")
```

Proporção de pessoas com seguro de saúde por nível de instrução:

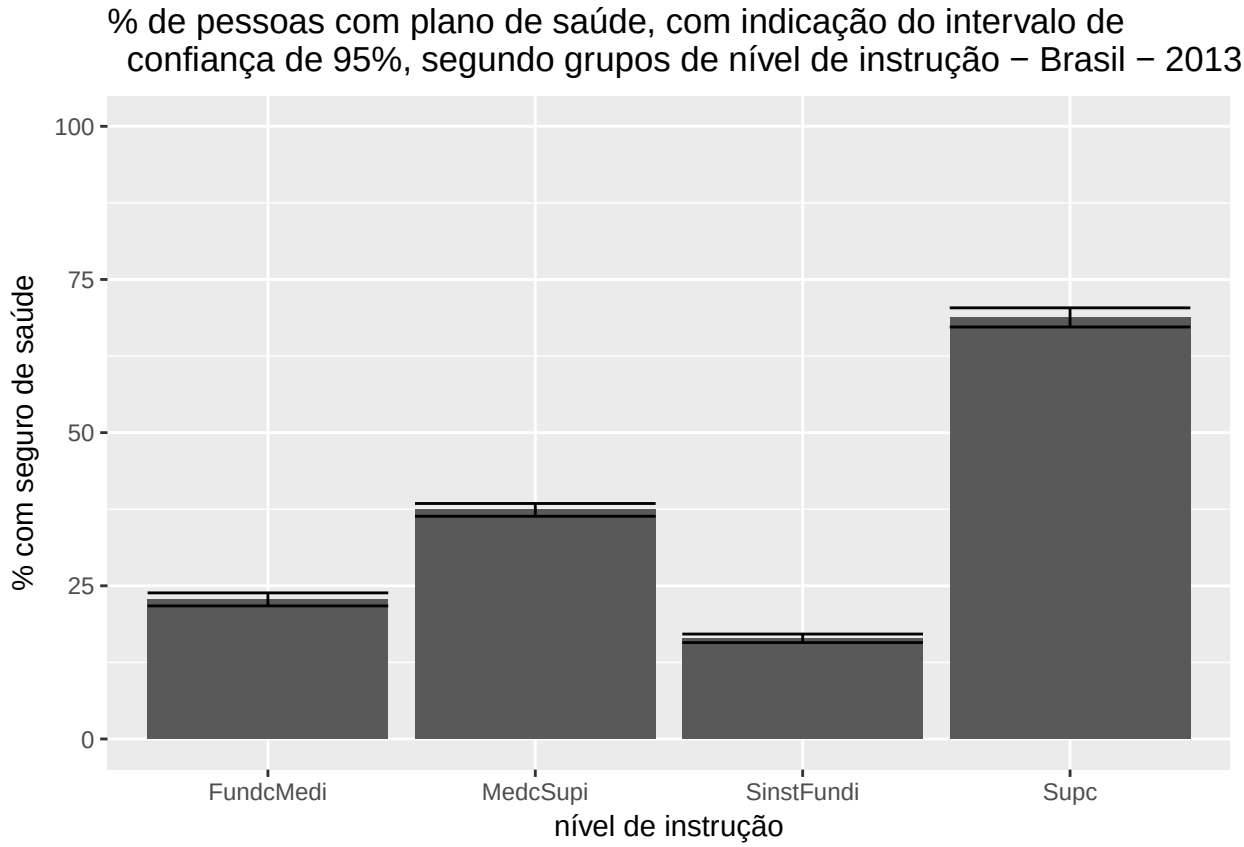
```
byeduc <- data.frame( svyby( ~ as.numeric( i001 == 1 ) , ~ educ ,
design = pes_all_des_pos , vartype = "ci" ,  level = 0.95 ,
svymean , na.rm = TRUE ) )
byeduc <- byeduc[,-1]
names(byeduc) <- c("Prop", "l.i.c.", "l.s.c.")
```

Imprime tabela:

```
knitr::kable(byeduc,booktabs = TRUE,
digits= 3, caption = "Proporção de pessoas com seguro de saúde por nível de instrução")
```

Gráfico de barras usando a library `ggplot2` (Wickham and Chang, 2016):

```
library(ggplot2)
ggplot(
    byeduc,
    aes( x = c("SinstFundi", "FundcMedi", "MedcSupi", "Supc") , y = 100*Prop )
    ) +
    geom_bar( stat = "identity" ) +
    geom_errorbar( aes( ymin = 100*l.i.c. , ymax = 100*l.s.c. ))  +
  ylim(c(0,100))+
    xlab( "nível de instrução" ) +
    ylab( "% com seguro de saúde" )+
  ggtitle("% de pessoas com plano de saúde, com indicação do intervalo de
  confiança de 95%, segundo grupos de nível de instrução - Brasil - 2013")
```


% de pessoas com plano de saúde, com indicação do intervalo de confiança de 95%, segundo grupos de nível de instrução − Brasil − 2013
% com seguro de saúde
100
75
50
25
0
FundcMedi
MedcSupi
SinstFundi
Supc
nível de instrução

Capítulo 2

Referencial para Inferência

2.1 Modelagem - Primeiras Idéias

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

2.1.1 Abordagem 1 - Modelagem Clássica

Seja Y um vetor $P \times 1$ de variáveis de pesquisa (ou de interesse), e sejam n vetores de observações destas variáveis para uma amostra de unidades de interesse denotados por $y_1, \ldots, y_n$. Em Inferência Estatística, a abordagem que aqui chamamos de `Modelagem clássica` considera $y_1, \ldots, y_n$ como valores (realizações) de vetores de variáveis aleatórias $y_1, \ldots, y_n$. Podemos formular modelos bastante sofisticados para a distribuição conjunta destes vetores aleatórios, mas para simplificar a discussão, vamos inicialmente supor que $y_1, \ldots, y_n$ são vetores aleatórios independentes e identicamente distribuídos (IID), com a mesma distribuição de Y, caracterizada pela função de densidade ou de frequência $f(y;\theta)$, onde $\theta \in \Theta$ é o parâmetro (um vetor de dimensão $K \times 1$) indexador da distribuição f, e Θ é o espaço paramétrico. A partir das observações $y_1, \ldots, y_n$, são feitas inferências a respeito do parâmetro θ. Uma representação gráfica esquemática dessa abordagem é apresentada na Figura 2.1 a seguir, e uma descrição esquemática resumida é apresentada na Tabela **??**.

Tabela 2.1: Representação esquemática da abordagem 1.

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

Do ponto de vista matemático, o parâmetro θ serve para indexar os elementos da família de distribuições $\{f(y;\theta); \theta \in \Theta\}$. Na prática, as questões relevantes da pesquisa são traduzidas em termos de perguntas sobre o valor ou região a que pertence o parâmetro θ, e a inferência sobre θ a partir dos dados ajuda a responder tais questões. Esta abordagem é útil em estudos analíticos tais como, por exemplo, na investigação da natureza da associação entre variáveis (modelos de regressão linear ou logística, modelos log-lineares, etc.). Vários exemplos discutidos ao longo dos Capítulos 6, 7 e 8 ilustram situações deste tipo. No Capítulo 9 o foco vai ser a estimação não paramétrica da forma da função $f(y;\theta)$.

Investigando a existência de diferenciais de salários por sexo e raça

Uma questão de grande interesse para o debate sobre a existência de desigualdades numa sociedade diz

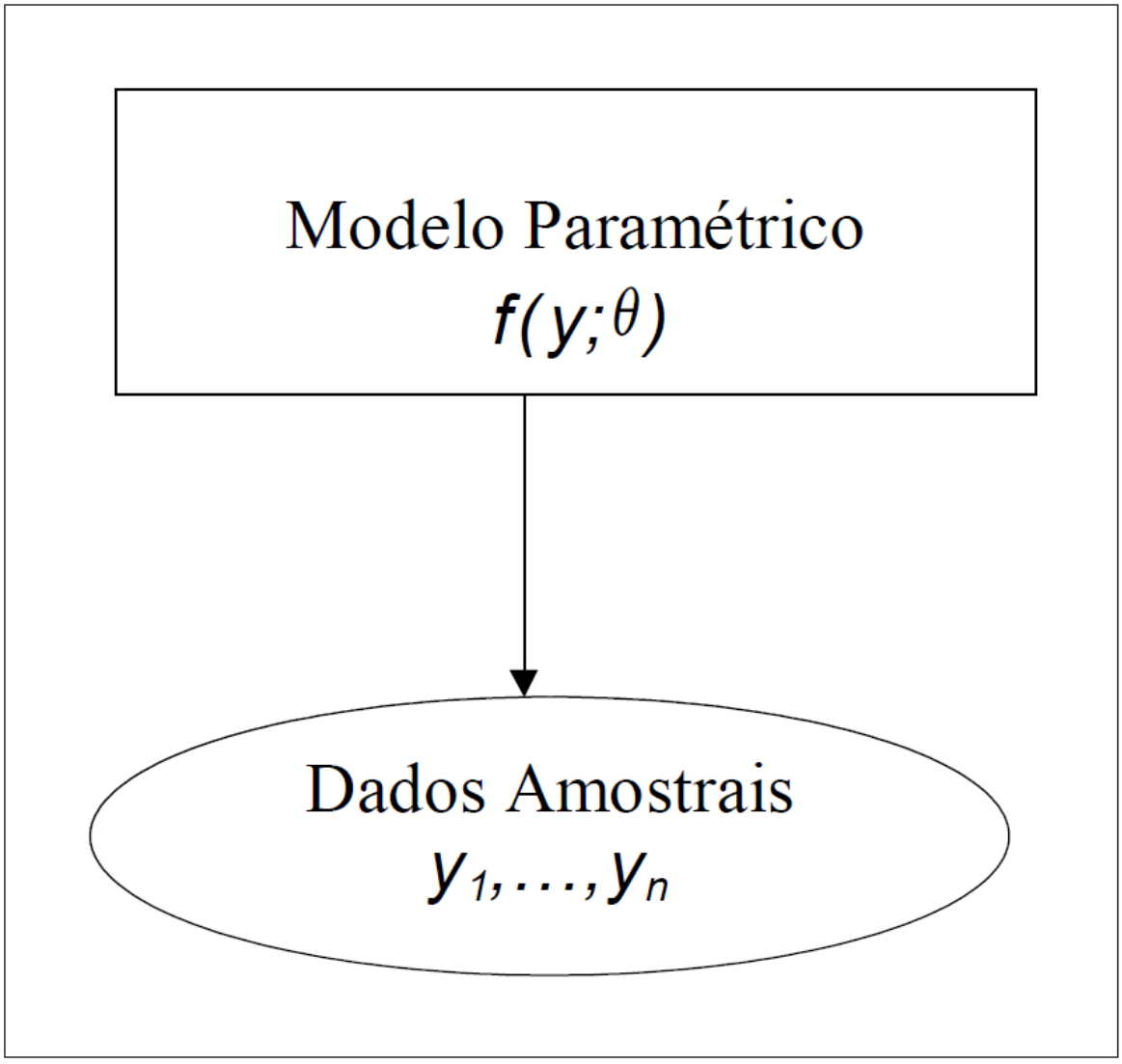


Figura 2.1: Modelagem Clássica

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

2.1.2 Abordagem 2 - Amostragem Probabilística

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

A partir dos valores observados na amostra, denotados por $y_{i_1}, \ldots, y_{i_n}$, são feitas inferências a respeito de funções dos valores populacionais, digamos $g(y_1, \ldots, y_N)$. Os valores de tais funções são quantidades descritivas populacionais (QDPs), também denominadas `parâmetros da população finita` pelos amostristas. Em geral, o objetivo desta abordagem é fazer estudos descritivos utilizando funções g particulares, tais como totais $g(y_1, \ldots, y_N) = \sum_{i=1}^{N} y_i$, médias $g(y_1, \ldots, y_N) = N^{-1} \sum_{i=1}^{N} y_i$, proporções, razões, etc. Uma descrição esquemática resumida dessa abordagem é apresentada no Tabela **??**, e uma representação gráfica resumida na Figura 2.2.

Tabela 2.2: Representação esquemática da abordagem 2.

Dados Amostrais	$Y_1 = y_1, \ldots, Y_n = y_n$
Hipóteses/Modelo Hipóteses	extraídos de $y_1, \ldots, y_N$ segundo $p(s)$
Objetivo	Inferir sobre funções $g(y_1, \ldots, y_N)$ usando $y_1, \ldots, y_n$

Investigando a existência de diferenciais de salários por sexo e raça

Ainda no contexto do exemplo da investigação da existência de diferenciais de salários por sexo e raça, os dados da Pesquisa sobre Padrões de Vida do IBGE podem ser utilizados para obter uma tabela cruzada tendo como entradas de linha o sexo dos trabalhadores, e como entrada das colunas um indicador de trabalhadores de raça / cor = branco, sendo as celas da tabela utilizadas para apresentar estimativas dos valores médios dos salários dos trabalhadores nas várias classes definidas pelos valores dos indicadores de linha e coluna. Tabelas como esta são habitualmente produzidas como resultado da realização de pesquisas amostrais pelas agências oficiais de estatísticas no Brasil e em muitos outros países. Embora capazes de revelar diferenças nos salários médios de trabalhadores de sexo e raça distintos, tais tabelas são insuficientes para investigar a questão da discriminação de gênero ou raça no mercado de trabalho, pois tais diferenças de salários médios poderiam ser explicadas por diferenças de características dos trabalhadores tais como escolaridade, que poderiam ter origem fora do mercado de trabalho. Por outro lado, ilustram bem os alvos de inferência freqüentemente definidos para pesquisas amostrais de populações finitas. Aqui o que se busca estimar é o valor médio de uma característica de interesse (no caso, o salário) para a população finita de onde foi extraída a amostra disponível para análise. Apesar de úteis, tais estimativas representam apenas medidas descritivas da população alvo.

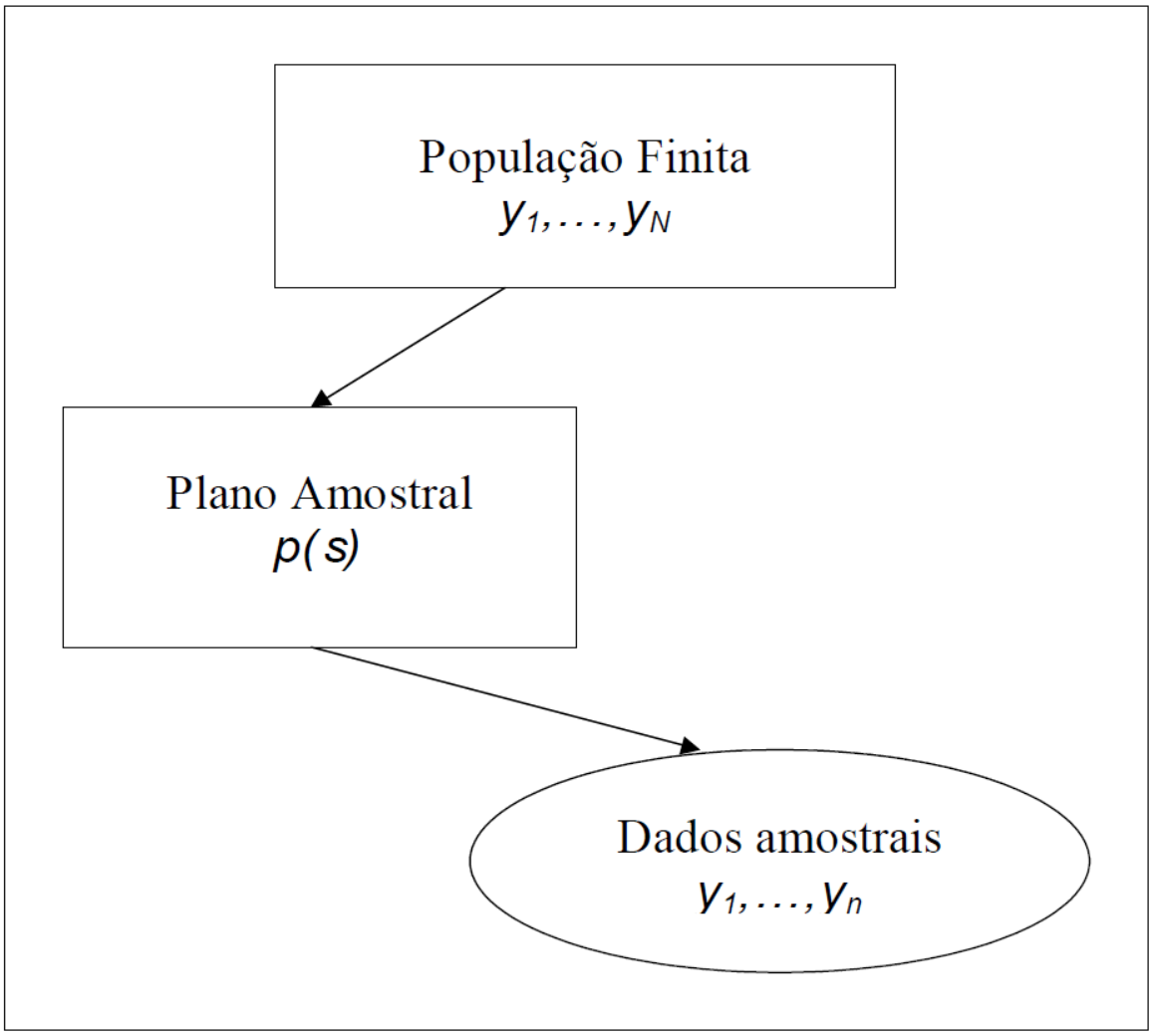


Figura 2.2: Amostragem Probabilística

2.1.3 Discussão das Abordagens 1 e 2

A primeira abordagem (Modelagem Clássica), nos termos descritos, foi proposta como modelo para medidas na Física e Astronomia, onde em geral o pesquisador tem relativo controle sobre os experimentos, e onde faz sentido falar em replicação ou repetição do experimento. Neste contexto, o conceito de aleatoriedade é geralmente introduzido para modelar os erros (não controláveis) no processo de medição.

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

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

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

Diante do exposto, podemos considerar algumas questões de interesse.

- É adequado aplicar métodos de análise da `abordagem 1`, concebidos para observações IID, aos dados obtidos através de pesquisas amostrais complexas?
- Em caso negativo, seria possível corrigir estes métodos, tornando-os aplicáveis para tratar dados amostrais complexos?
- Ou seria mais adequado fazer uso analítico dos dados dentro da `abordagem 2`? E neste caso, como fazer isto, visto que nesta abordagem não é especificado um modelo para a distribuição das variáveis de pesquisa `na população`?

Além destas, também é de interesse a questão da robustez da modelagem, traduzida nas seguintes perguntas.

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

O objeto deste livro é exatamente discutir respostas para as questões aqui enumeradas. Para isso, vamos considerar uma abordagem que propõe um modelo parametrizado como na `abordagem 1`, e além disso incorpora na análise os pontos i) a iii) do Capítulo 1 mediante aproveitamento da estrutura do planejamento amostral como na `abordagem 2`.

2.1.4 Abordagem 3 - Modelagem de Superpopulação

Nesta abordagem, os valores $y_1, \ldots, y_N$ das variáveis de interesse Y na população finita são considerados observações ou realizações dos vetores aleatórios $Y_1, \ldots, Y_N$, supostos IID com distribuição $f(y;\theta)$, onde $\theta \in \Theta$. Este modelo é denominado `Modelo de Superpopulação`. Note que, em contraste com o que se faz na `abordagem 1`, o modelo probabilístico aqui é especificado para descrever o mecanismo aleatório que gera `a população`, não `a amostra`, muito embora na maioria das aplicações práticas a população não será jamais observada por inteiro. Não obstante, ao formular modelos para a população, nossas perguntas e respostas

descritas em termos de valores ou regiões para o parâmetro θ passam a se referir à população de interesse ou populações similares, quer existam ao mesmo tempo, quer se refiram a estados futuros (ou passados) da mesma população.

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

Adotando o modelo de superpopulação e considerando métodos usuais disponíveis na `abordagem 1`, podemos utilizar funções de $y_1, \ldots, y_N$, digamos $g(y_1, \ldots, y_N)$, para fazer inferências sobre θ. Desta forma, definimos estatísticas $(y_1, \ldots, y_N)$ (no sentido da `abordagem 1`) que são quantidades descritivas populacionais (parâmetros populacionais no contexto da `abordagem 2`), que passam a ser os novos parâmetros-alvo. O passo seguinte é utilizar métodos disponíveis na `abordagem 2` para fazer inferência sobre $g\left(y_1, \ldots, y_N\right)$ baseada em $y_{i_1}, \ldots, y_{i_n}$. Note que não é possível basear a inferência nos valores populacionais $y_1, \ldots, y_N$, já que estes não são conhecidos ou observados. Este último passo adiciona a informação sobre o plano amostral utilizado, contida em $p(s)$, à informação estrutural contida no modelo $\{f\left(y;\theta\right);\theta \in \Theta\}$. Uma representação esquemática dessa abordagem é apresentada no Tabela **??**.

Tabela 2.3: Representação esquemática da abordagem 3.

Dados Amostrais	$Y_1 = y_1, \ldots, Y_n = y_n$
População e esquema de seleção	Extraídos de $y_1, \ldots, y_n$ segundo $p(s)$
Modelo para poulação	$Y_1, \ldots, Y_N$ variáveis aleatórias IID com distribuição $f(y, \theta)$, onde $\theta \in \Theta$
Parâmetro-alvo	associar $\theta \longleftrightarrow g\left(Y_1, \ldots Y_N\right)$
Objetivo	Inferir sobre $g\left(Y_1, \ldots Y_N\right)$ partir de $y_{i_1}, \ldots, y_{i_n}$ usando $p\left(s\right)$

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

Ao construir e ajustar modelos a partir de dados de pesquisas amostrais `complexas`, tais como as executadas pelo IBGE, o usuário precisa incorporar as informações sobre pesos e planos amostrais utilizados. Em geral, ao publicar os resultados das pesquisas, os pesos são considerados, sendo possível produzir estimativas pontuais `corretas` utilizando os pacotes tradicionais. Por outro lado, para construir intervalos de confiança e testar hipóteses sobre parâmetros de modelos, seria preciso o conhecimento das estimativas de variâncias e covariâncias das estimativas, obtidas a partir do plano amostral utilizado. Mesmo conhecendo o plano amostral, geralmente não é simples incorporar pesos e plano amostral na análise sem o uso de pacotes especializados, ou de rotinas específicas já agora disponíveis em alguns dos pacotes mais comumente utilizados (por exemplo, SAS, Stata, SPSS, ou R entre outros). Tais pacotes especializados ou rotinas específicas utilizam na maioria métodos aproximados para estimar matrizes de covariância, tais como os de Máxima Pseudo-Verossimilhança e de Linearização, que serão descritos mais adiante.

Em outras palavras, o uso dos pacotes usuais para analisar dados produzidos por pesquisas com planos amostrais complexos, tal como o uso de muitos remédios, pode ter contra-indicações. Cabe ao usuário `ler a bula` e identificar situações em que o uso de tais pacotes pode ser inadequado, e buscar opções de rotinas específicas ou de pacotes especializados capazes de incorporar adequadamente a estrutura do plano amostral nas análises. Ao longo deste livro faremos uso intensivo da library `survey` disponível no R, mas o leitor encontrará funcionalidade semelhante em vários outros pacotes. Nossa escolha se deveu a dois fatores principais: primeiro ao fato do pacote R ser aberto, livre e gratuito, dispensando o usuário de custos de

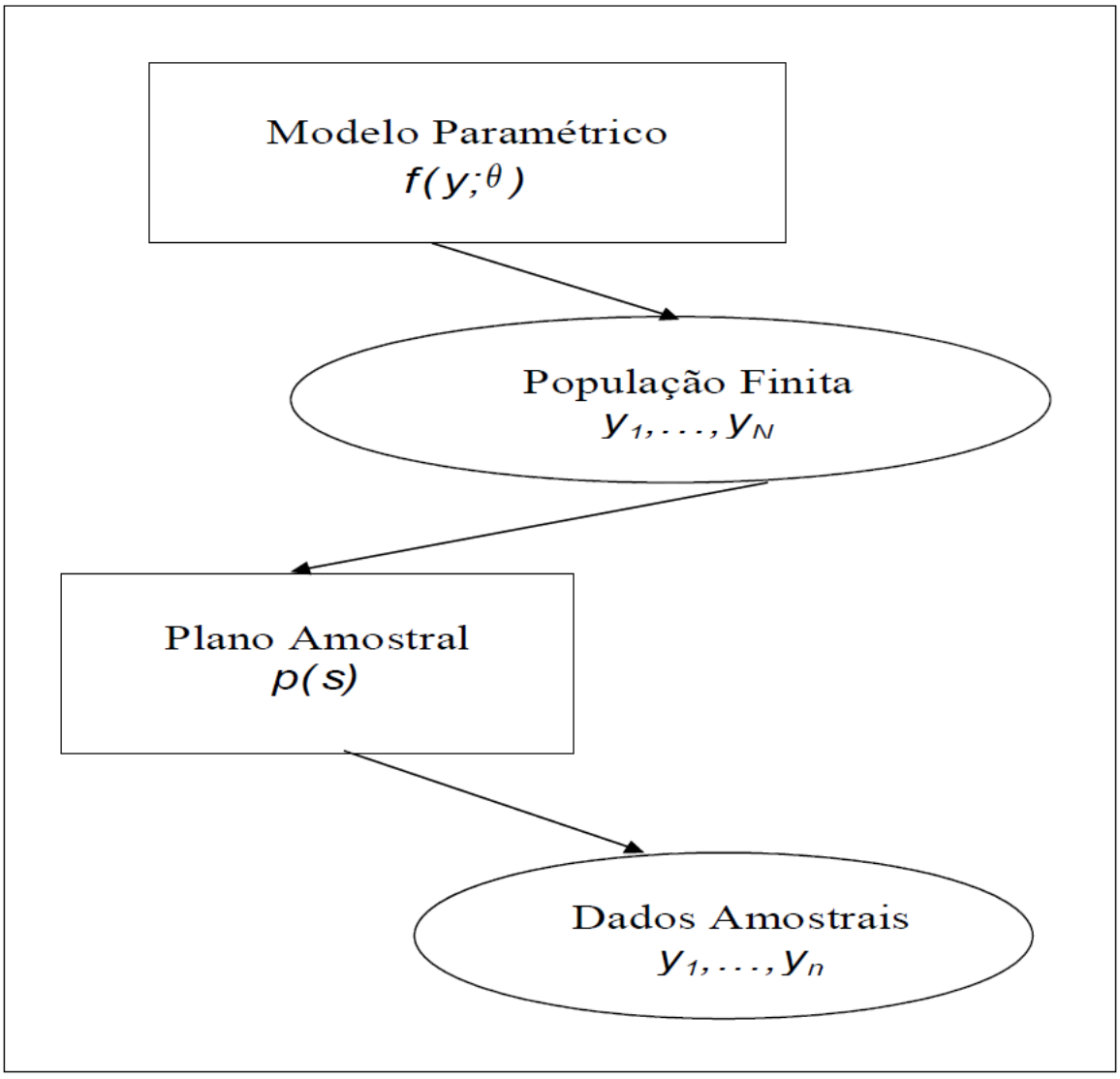


Figura 2.3: Modelagem de Superpopulação

licenciamento, bem como possibilitando aos interessados o acesso ao código fonte e a capacidade de modificar as rotinas de análise, caso necessário. O segundo fator é de natureza mais técnica, porém transitória. No presente momento, a library `survey` é a coleção de rotinas mais completa e genérica para análise de dados amostrais complexos existente, dispondo de rotinas capazes de ajustar os modelos usuais mas também de ajustar modelos não convencionais mediante maximização de verossimilhanças especificadas pelo usuário. Sabemos, entretanto, que muitos usuários habituados à facilidade de uso de pacotes com interfaces gráficas do tipo `aponte e clique` terão dificuldade adicional de adaptar-se à linguagem de comandos utilizada pelo pacote R, mas acreditamos que os benefícios do aprendizado desta nova ferramenta compensam largamente os custos adicionais do aprendizado.

2.2 Fontes de Variação

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

1. `modelo de superpopulação`, que descreve o processo subjacente que por hipótese gera as medidas verdadeiras de qualquer unidade da população considerada;
2. `processo de medição`, que diz respeito aos instrumentos e métodos usados para obter as medidas de qualquer unidade da população;
3. `planejamento amostral`, que estabelece o mecanismo pelo qual unidades da população são selecionadas para participar da pesquisa por amostra.

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

4. `mecanismo de resposta`, ou seja, o mecanismo que controla se valores de medições de unidades selecionadas são disponibilizados ou não.

Para concentrar o foco nas questões de maior interesse deste texto, as fontes (2) e (4) não serão consideradas no referencial adotado para a maior parte dos capítulos. Para o tratamento das dificuldades causadas por não resposta, a fonte (4) será considerada no capítulo onze. Assim sendo, exceto onde explicitamente indicado, de agora em diante admitiremos que não há erros de medição, implicando que os valores observados de quaisquer variáveis de interesse serão considerados valores corretos ou verdadeiros. Admitiremos ainda que há resposta completa, implicando que os valores de quaisquer variáveis de interesse estão disponíveis para todos os elementos da amostra selecionada depois que a pesquisa foi realizada. Hipóteses semelhantes são adotadas, por exemplo, em (Montanari, 1987).

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

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

2.3 Modelos de Superpopulação

Seja $\{1, ..., N\}$ um conjunto de rótulos que identificam univocamente os N elementos distintos de uma população-alvo finita U. Sem perda de generalidade tomaremos $U = \{1, ..., N\}$. Uma pesquisa cobrindo n

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

Denote por $\mathbf{y}_i = (y_{i1}, ..., y_{iP})'$ o vetor $P \times 1$ de valores das variáveis da pesquisa e por $\mathbf{x}_i = (x_{i1}, ..., x_{iQ})'$ o vetor $Q \times 1$ de variáveis auxiliares da i-ésima unidade da população, respectivamente, para $i = 1, ..., N$. Aqui as variáveis auxiliares são consideradas como variáveis contendo a informação requerida para o planejamento amostral e a estimação a partir da amostra, como se discutirá com mais detalhes adiante. Denote por $\mathbf{y}_U$ a matriz $N \times P$ formada empilhando os vetores transpostos das observações correspondentes a todas as unidades da população, e por $\mathbf{Y}_U$ a correspondente matriz de vetores aleatórios geradores das observações na população.

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

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

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

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

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

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

onde λ e ϕ são vetores de parâmetros.

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

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

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

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

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

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

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

baseada no planejamento amostral ou abordagem de aleatorização, pois neste caso a única fonte de variação (aleatoriedade) é proveniente do planejamento amostral. Em geral, a distribuição conjunta de $(\mathbf{Y}_U, \mathbf{X}_U)$ 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 $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N)$. Sob a hipótese de resposta completa, a única fonte de incerteza seria devida ao fato de que $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N)$ é uma realização de $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_N, \mathbf{X}_N)$. Os dados observados poderiam então ser usados para fazer inferências sobre η, ϕ, λ ou θ usando procedimentos padrões.

Inferência sobre quaisquer dos parâmetros η, ϕ, λ ou θ do modelo de superpopulação é chamada `inferência analítica`. Este tipo de inferência só faz sentido quando o modelo de superpopulação não é degenerado como em (2.4). 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 $g(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ dos valores das variáveis da pesquisa, o modelo de superpopulação não é estritamente necessário, embora possa ser útil. Inferência para tais quantidades, chamadas parâmetros da população finita ou quantidades descritivas populacionais (QDPs), é chamada inferência descritiva.

2.4 Planejamento Amostral

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

1. o conjunto de rótulos $a = \{i_1, \ldots, i_n\}$ dos distintos elementos na amostra, onde n $(1 \leq n \leq N)$ é o número de elementos na amostra a, chamado tamanho da amostra;

2. os valores na amostra das variáveis da pesquisa $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_n}$;

3. os valores das variáveis auxiliares na população $\mathbf{x}_1, \ldots, \mathbf{x}_N$, quando a informação auxiliar é dita ``completa''; alternativamente, os valores das variáveis auxiliares na amostra $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}$, mais os totais ou médias destas variáveis na população, quando a informação auxiliar é dita ``parcial''.

O mecanismo usado para selecionar a amostra a da população finita U é chamado plano amostral. Uma forma de caracterizá-lo é através da função $p(.)$, onde $p(a)$ dá a probabilidade de selecionar a amostra a no conjunto A de todas as amostras possíveis. Só mecanismos amostrais envolvendo alguma forma de seleção probabilística bem definida serão aqui considerados, e portanto supõe-se que $0 \leq p(a) \leq 1 \; \forall a \in A$ e $\sum_{a \in A} p(a) = 1$.

Esta caracterização do plano amostral $p(a)$ é bem geral, permitindo que o mecanismo de seleção amostral dependa dos valores das variáveis auxiliares $\mathbf{x}_1, \ldots, \mathbf{x}_N$ bem como dos valores das variáveis da pesquisa na população $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (amostragem `informativa`, veja Seção 2.5. Uma notação mais explícita para indicar esta possibilidade envolveria escrever $p(a)$ como $p[a|(\mathbf{y}_U, \mathbf{x}_U)]$. Tal notação será evitada por razões de simplicidade.

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

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

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

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

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

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

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

2.5 Planos Amostrais Informativos e Ignoráveis

Ao fazer inferência usando dados de pesquisas amostrais precisamos distinguir duas situações que requerem tratamento diferenciado. Uma dessas situações ocorre quando o plano amostral empregado para coletar os dados é `informativo`, isto é, quando o mecanismo de seleção das unidades amostrais pode depender dos valores das variáveis de pesquisa. Um exemplo típico desta situação é o dos `estudos de caso-controle`, em que a amostra é selecionada de tal forma que há `casos` (unidades com determinada condição) e `controles` (unidades sem essa condição), sendo de interesse a modelagem do indicador de presença ou ausência da condição em função de variáveis preditoras, e sendo esse indicador uma das variáveis de pesquisa, que é considerada no mecanismo de seleção da amostra. Os métodos que discutiremos ao longo deste livro não são adequados, em geral, para esse tipo de situação, e portanto uma hipótese fundamental adotada ao longo deste texto é que os planos amostrais considerados são `não-informativos`, isto é, não podem depender diretamente dos valores das variáveis da pesquisa. Logo eles satisfazem

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

Entre os planos amostrais `não-informativos`, precisamos ainda distinguir duas outras situações de interesse. Quando o plano amostral é amostragem aleatória simples com reposição (AASC), o modelo adotado para a amostra é o mesmo que o modelo adotado para a população antes da amostragem. Quando isto ocorre, o plano amostral é dito ignorável, porque a inferência baseada na amostra utilizando a abordagem clássica descrita em 2.1 pode prosseguir sem problemas. Entretanto, esquemas amostrais desse tipo são raramente empregados na prática, por razões de eficiência e custo. Em vez disso, são geralmente empregados planos amostrais envolvendo estratificação, conglomeração e probabilidades desiguais de seleção (amostragem complexa).

Com amostragem complexa, porém, os modelos para a população e a amostra podem ser muito diferentes (plano amostral `não-ignorável`), mesmo que o mecanismo de seleção não dependa das variáveis de pesquisa, mas somente das variáveis auxiliares. Neste caso, ignorar o plano amostral pode viciar a inferência. Veja o Exemplo 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 por exemplo (Sugden and Smith, 1984) ou os Capítulos 1 e 2 de (Chambers and Skinner, 2003). Porém testar a ignorabilidade do plano amostral é muitas vezes complicado. Em caso de dificuldade, o uso de pesos tem papel fundamental.

Uma forma simples de lidar com os efeitos do plano amostral na estimação pontual de quantidades descritivas populacionais de interesse é incorporar pesos adequados na análise, como se verá no Capítulo 3. Essa forma porém, não resolve por si só o problema de estimação da precisão das estimativas pontuais, nem mesmo

o caso da estimação pontual de parâmetros em modelos de superpopulação, o que vai requerer métodos específicos discutidos no Capítulo 5.

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

A seguir apresentamos um exemplo com a finalidade de ilustrar uma situação de plano amostral não-ignorável.
Exemplo 2.1. Efeito da amostragem estratificada simples com alocação desproporcional

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

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

	y		
x	0	1	Total
0	η_{00}	η_{01}	η_{0+}
1	η_{10}	η_{11}	η_{1+}
Total	η_{+0}	η_{+1}	1

que também pode ser representada por:

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

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

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

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

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

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

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

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

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

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

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

	y		
x	0	1	Total
0	η_{00}/η_{0+}	η_{01}/η_{0+}	1
1	η_{10}/η_{1+}	η_{11}/η_{1+}	1

ou, alternativamente

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

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

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

onde a designação f_a é utilizada para denotar a distribuição `na amostra`.

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

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

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

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

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

	y		
x	0	1	Total
0	0.7	0.1	0.8
1	0.1	0.1	0.2
Total	0.8	0.2	1

segue-se que a distribuição condicional de y dado x na população (mas também na amostra) é dada por

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

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

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

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

y	0	1
$f_U(y)$	0.8	0.2

y	0	1
$f_a(y)$	0.6875	0.3125

Assim, inferência sobre a distribuição de y `na população` levada a cabo a partir dos dados da amostra observada sem considerar a estrutura do plano amostral será equivocada, pois a alocação igual da amostra nos estratos leva à observação de uma proporção maior de valores de x iguais a 1 na amostra (1/2) do que a correspondente proporção existente na população (1/5). Em conseqüência, a proporção de valores de y iguais a 1 na amostra (0.3125) é 56% maior que a correspondente proporção na população (0.2).

Este exemplo é propositalmente simples, envolve apenas duas variáveis com distribuição Bernoulli, mas ilustra bem como a amostragem pode modificar distribuições de variáveis da amostra em relação à correspondente distribuição na população.

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

Capítulo 3

Estimação Baseada no Plano Amostral

3.1 Estimação de Totais

Devido a sua importância para os desenvolvimentos teóricos em vários dos capítulos subsequentes, alguns resultados básicos relativos à estimação de totais da população finita numa abordagem baseada no plano amostral serão reproduzidos nesta seção. A referência básica usada foi a Seção 2.8 de (Särndal et al., 1992).

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

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

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

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

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

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

Além disto, sua variância de aleatorização é dada por

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

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

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

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

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

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

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

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

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

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

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

Exemplo 3.1. Amostragem Aleatória Simples Sem Reposição (AAS)

Quando o plano é amostragem aleatória simples sem reposição (AAS), as expressões apresentadas para o estimador de total, sua variância e estimadores desta variância simplificam bastante, porque as probabilidades de inclusão ficam iguais a

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

e

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

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

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

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

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

onde $f = n/N$ é a fração amostral e

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

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

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

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

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

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

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

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

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

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

e

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

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

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

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

e

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

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

O estimador de regressão descrito em (3.16) é um caso particular do estimador de regressão generalizado, obtido quando se consideram vetores de variáveis auxiliares em vez de uma única variável auxiliar x como aqui. Outra forma de generalizar o estimador de regressão é considerar estimadores alternativos dos coeficientes de regressão em lugar do estimador simples b_{xy} empregado aqui. Para uma discussão detalhada do estimador de regressão generalizado veja (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 et al., 1992), Seção 6.6 e cap. 7. Apesar de sua importância para os praticantes de amostragem, a discussão detalhada desse problema não será incluída neste livro.

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

3.2 Por que Estimar Variâncias

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

Em geral, a obtenção de estimativas de variâncias (alternativamente, de desvios padrões ou mesmo de coeficientes de variação) é requerida para que intervalos de confiança possam ser calculados, e outras formas de inferência realizadas. Intervalos de confiança elaborados com estimativas amostrais são geralmente baseados em aproximações assintóticas da distribuição normal, tais que intervalos da forma

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

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

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

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

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

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

Finalmente, alguns estimadores de variância podem, em alguns casos, produzir estimativas negativas da variância, que são inaceitáveis de um ponto de vista prático (tais como o estimador da expressão (3.5) para alguns esquemas de seleção com probabilidades desiguais e determinadas configurações peculiares da amostra).

Em todos esses casos, é requerido o emprego de técnicas especiais de estimação de variância. é de algumas dessas técnicas que tratam as seções seguintes deste capítulo. A seleção das técnicas discutidas aqui não é exaustiva, e um tratamento mais completo e aprofundado da questão pode ser encontrado no livro de (Wolter, 1985). Discutimos inicialmente a técnica de `Linearização de Taylor`, em seguida uma abordagem comumente adotada para estimar variâncias para planos amostrais estratificados em vários estágios, com seleção de unidades primárias com probabilidades desiguais, denominada `Método do Conglomerado Primário` (do inglês Ultimate Cluster, e finalmente se discute brevemente uma técnica baseada na ideia de pseudo-replicações da amostra, denominada `Jackknife`. A combinação dessas três idéias suporta os desenvolvimentos teóricos dos algoritmos empregados pelos principais pacotes estatísticos especializados em estimação de variâncias de aleatorização (veja discussão no Capítulo 13.

3.3 Linearização de Taylor para Estimar variâncias

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

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

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

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

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

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

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

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

Assumindo então que θ é da forma

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

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

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

Este estimador é não-viciado e tem variância de aleatorização

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

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

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

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

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

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

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

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

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

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

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

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

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

Exemplo 3.2. Matriz de covariância para um vetor de razões

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

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

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

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

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

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

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

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

e a correspondente matriz jacobiana como

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

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

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

Efetuando os produtos das matrizes em blocos obtemos

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

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

3.4 Método do Conglomerado Primário

A ideia central do `Método do Conglomerado Primário` (do inglês `Ultimate Cluster`) para estimação de variâncias para estimadores de totais e médias em planos amostrais de múltiplos estágios, proposto por (Hansen et al., 1953), é considerar apenas a variação entre informações disponíveis no nível das unidades primárias de amostragem (UPAs), isto é, dos conglomerados primários, e admitir que estes teriam sido selecionados com reposição da população. Esta ideia é simples, porém bastante poderosa, porque permite acomodar uma enorme variedade de planos amostrais, envolvendo estratificação e seleção com probabilidades desiguais (com ou sem reposição) tanto das unidades primárias como das demais unidades de amostragem. Os requisitos fundamentais para permitir a aplicação deste método são que estejam disponíveis estimadores não viciados dos totais da variável de interesse para cada um dos conglomerados primários selecionados, e que pelo menos dois destes sejam selecionados em cada estrato (se a amostra for estratificada no primeiro estágio).

Embora o método tenha sido originalmente proposto para estimação de totais, pode ser aplicado também para estimar (por linearização) quantidades populacionais que possam ser representadas como funções de totais, conforme discutido na Seção 3.3. De fato, esse método fornece a base para vários dos pacotes estatísticos especializados em cálculo de variâncias considerando o plano amostral, tais como SUDAAN, CENVAR, STATA ou PC-CARP (veja discussão no Capítulo 10).

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

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

e um estimador não viciado da variância de aleatorização correspondente por

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

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

Embora muitas vezes a seleção das unidades primárias possa ter sido feita sem reposição, o estimador de Conglomerados Primários aqui apresentado pode fornecer uma aproximação razoável da correspondente variância de aleatorização. Isso ocorre porque planos amostrais sem reposição são em geral mais eficientes que planos com reposição de igual tamanho. Tal aproximação é largamente utilizada pelos praticantes de amostragem para estimar variâncias de quantidades descritivas usuais tais como totais e médias (com a devida adaptação) devido à sua simplicidade, comparada com a complexidade muito maior envolvida com o emprego de estimadores de variância que tentam incorporar todas as etapas de planos amostrais em vários estágios. Uma discussão sobre a qualidade dessa aproximação e alternativas pode ser encontrada em (Särndal et al., 1992), p. 153.

3.5 Métodos de Replicação

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

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

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

é um estimador não viciado de θ e

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

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

Note que desde que as replicações sejam construídas de forma independente conforme indicado, os estimadores $\widehat{\theta}_R$ e $\widehat{V}_R\left(\widehat{\theta}_R\right)$ são não viciados qualquer que seja o plano amostral empregado para selecionar a amostra de cada replicação, o que faz desta uma técnica flexível e genérica. Além disso, a abordagem de replicação é bastante geral, pois os estimadores aos quais se aplica não precisam ser necessariamente expressos como funções de totais, como ocorre com a técnica de linearização discutida na Seção 3.3. Apesar destas vantagens, a aplicação prática desta técnica de forma exata é restrita porque em geral é menos eficiente, inconveniente e mais caro selecionar G amostras independentes com o mesmo esquema, se comparado à seleção de uma única amostra de tamanho n diretamente. Além disto, se o número de replicações G for pequeno, o estimador de variância pode ser instável. Uma pesquisa importante e de grande porte em que esta ideia é aplicada exatamente é a pesquisa de preços para formar o índice de Preços ao Consumidor (do inglês Consumer Price Index - CPI do (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 ideia simples de aplicar, poderosa e flexível, por acomodar uma ampla gama de planos amostrais e situações de estimação de interesse. Quando as replicações são construídas após a pesquisa (a posteriori), mediante repartição (por sorteio) da amostra pesquisada em G grupos mutuamente exclusivos de igual tamanho, estas são chamadas de `replicações dependentes` ou `grupos aleatórios` (do inglês random groups). As expressões fornecidas para o estimador de replicação e sua variância são também empregadas nesse caso como uma aproximação, mas não possuem as mesmas propriedades do caso de replicações independentes.

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

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

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

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

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

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

onde $\widehat{\theta}_g$ é um estimador de θ obtido da amostra após eliminar os elementos do grupo g, empregando a mesma forma funcional adotada no cálculo do estimador $\widehat{\theta}$ que considera a amostra inteira. A estimação da variância por esse método pode então ser feita de duas maneiras alternativas, usando um dos estimadores dados por

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

ou

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

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

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

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

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

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

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

3.6 Laboratório de R

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

Vamos criar duas variáveis:

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

```
library(survey)
library(anamco) # carrega dados
# cria objeto de desenho
ppv.des<-svydesign(ids = ~nsetor, strata = ~estratof,
data = ppv, nest = TRUE, weights = ~pesof)
# atualiza objeto de desenho com novas variáveis
ppv.des<-update(ppv.des,
  analf=(v04a01 == 2 | v04a02 == 2)*1,
```

```
  faixa=(v02a08 >= 7 & v02a08 <= 14) *1,
  analf.faixa= (analf==1 & faixa==1)*1
)
```

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

```
ppv.se.des <- subset(ppv.des, regiao == 2)
```

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

```
tot.est<-svytotal(~analf.faixa+faixa ,ppv.se.des )
Vcov.Y1.Y2<-vcov(tot.est)
```

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

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

```
##      faixa
## 0.01178896
```

Podemos calcular diretamente o desvio-padrão:

```
svyratio(~analf.faixa, ~faixa, ppv.se.des)
```

```
## Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv.se.des)
## Ratios=
##                faixa
## analf.faixa 0.118689
## SEs=
##                  faixa
## analf.faixa 0.01178896
```

A estimativa do desvio-padrão obtida por meio da função `svyratio` coincide com a obtida diretamente pelo método de linearização, e é igual a 0.01179. O método default para estimar variâncias usado pela library `survey` (Lumley, 2017) do R é o de linearização de Taylor.

A library `survey` dispõe de métodos alternativos para a estimação de variância. Vamos utilizar os métodos de replicação de Jackknife e de Bootstrap para estimar esta variância de razão. Inicialmente, vamos converter o objeto de desenho `ppv1.se.design` em um objeto de desenho de replicação de tipo Jackknife, contendo as réplicas de pesos que fornecem correspondentes réplicas de estimativas.

```
ppv.se.des.jkn<-as.svrepdesign(ppv.se.des,type="JKn")
svyratio(~analf.faixa, ~faixa, ppv.se.des.jkn)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv.se.des.jkn)
## Ratios=
##                faixa
## analf.faixa 0.118689
## SEs=
##                [,1]
## [1,] 0.01181434
```

Para o tipo Bootstrap, temos:

```
ppv.se.des.boot<-as.svrepdesign(ppv.se.des,type="bootstrap")
svyratio(~analf.faixa, ~faixa, ppv.se.des.boot)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv.se.des.boot)
## Ratios=
##                faixa
## analf.faixa 0.118689
## SEs=
##             [,1]
## [1,] 0.01080893
```

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

```
class(ppv.se.des.jkn)
```

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

```
names(ppv.se.des.jkn)
```

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

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

```
ppv.se<-ppv.se.des.jkn$variables
nrow(ppv.se)
```

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

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

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

O argumento `compress` da função `as.svrepdesign` permite especificar se, na saída da função, a matriz `weights` será na forma comprimida ou não. Na aplicação feita foi usado o valor default que é a forma comprimida. A forma não comprimida da matriz `weights` tem 8903 linhas e 276 colunas. A forma comprimida permite economizar memória, e pode ser facilmente convertida para a forma não comprimida, utilizando-se a componente`index`.

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

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

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

```
fact.peso.comp<-ppv.se.des.jkn$repweights[[1]]
ind.cong<-ppv.se.des.jkn$repweights[[2]]
mat.fat.pesos<- fact.peso.comp[ind.cong,]
str(mat.fat.pesos)
```

```
##  num [1:8903, 1:276] 0 0 1.06 1.06 1.06 ...
```

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

```
mat.rep.pesos<-weights(ppv.se.des)*mat.fat.pesos
```

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

```
rep.est.raz<-numeric(ncol(mat.rep.pesos))
for (i in 1:ncol(mat.rep.pesos)){
rep.est.raz[i]<-sum(mat.rep.pesos[,i]*ppv.se$analf.faixa)/sum(mat.rep.pesos[,i]*ppv.se$faixa)
}
```

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

```
mean.raz<-mean( rep.est.raz[ppv.se.des.jkn$rscales>0])
var.jack.raz<- sum((rep.est.raz-mean.raz)^2*ppv.se.des.jkn$rscales)*ppv.se.des.jkn$scale
round(sqrt(var.jack.raz),5)
```

```
## [1] 0.01181
```

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

```
var.raz.rep<-withReplicates(ppv.se.des.jkn, function(w,ppv.se) sum(w*ppv.se$analf.faixa)/sum(w*ppv.se$fa
var.raz.rep
```

```
##        theta     SE
## [1,] 0.11869 0.0118
```

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

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

```
withReplicates(ppv.se.des.jkn,function(w,ppv.se) with(ppv.se,
(sum(w*(analf==1&faixa==1))/sum(faixa==1))/(sum(w*(analf==1&faixa==0))/sum(faixa==0))))
```

```
##        theta     SE
## [1,] 0.50623 0.0494
```

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 ideia 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 ão 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 valores-p 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, 989a).

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) $V_{VERD}\left(\hat{\theta}\right)$ e a variância da distribuição do estimador induzida pelo plano de amostragem aleatória simples $V_{AAS}\left(\hat{\theta}\right)$.

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

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

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

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

(Nascimento Silva and Moura, 1990) estimaram o $\mathbf{EPA}_{Kish}$ para estimadores de totais de várias variáveis sócio-econômicas no nível das Regiões Metropolitanas (RMs) utilizando dados do questionário de amostra do Censo Demográfico de 1980. Essas medidas estimadas do efeito do plano amostral foram calculadas para três 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 π-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 π-ponderado. Os resultados referentes à Região Metropolitana do Rio de Janeiro para algumas variáveis 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.

Plano amostral AC2AAS AC2PPT

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

Variável	Estimador Simples	Estimador de Razão	Estimador π-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	5.22	4.92	4.49
4) Número total de filhos nascidos vivos de mulheres com 15 anos ou mais	4.59	2.02	1.89
5) Número de domicílios que têm fogão	111.27	1.58	1.55
6) Número de domicílios que têm telefone	7.11	7.13	6.41
7) Valor do aluguel ou prestação mensal	7.22	7.02	6.45
8) Número de domicílios que têm automóvel e renda $<$ 5SM	1.80	1.67	1.55
9) Número de domicílios que têm geladeira e renda $\geq$ 5SM	46.58	2.26	2.08

Os valores apresentados na Tabela 4.1 para a RM do Rio de Janeiro são similares aos observados para as demais RMs, se consideradas as mesmas variáveis. 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 and 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 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 p. 24, (Skinner et al., 1989), 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 $v_0 = \widehat{V}_{IID}\left(\hat{\theta}\right)$ um estimador usual (consistente) da variância de $\hat{\theta}$ calculado sob a hipótese (ingênua) de que as observações 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 $V_{VERD}\left(\hat{\theta}\right)$. Note que $V_{VERD}\left(\hat{\theta}\right) = V_M\left(\hat{\theta}\right)$ na abordagem baseada em modelos e $V_{VERD}\left(\hat{\theta}\right) = V_p\left(\hat{\theta}\right)$ na abordagem de aleatorização.

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 $V_{VERD}\left(\hat{\theta}\right)$ 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 $V_{VERD}\left(\hat{\theta}\right)$.

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

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

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

Desta forma, o $\textbf{EPA}\left(\hat{\theta}, v_0\right)$ mede a tendência de v_0 a subestimar ou superestimar $V_{VERD}\left(\hat{\theta}\right)$, variância `verdadeira` de $\hat{\theta}$. Quanto mais afastado de 1 for o valor de $\textbf{EPA}\left(\hat{\theta}, v_0\right)$, 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 $\textbf{EPA}\left(\hat{\theta}, v_0\right)$ pode ser calculado com respeito a distribuições de aleatorização ou do modelo envolvido, bastando calcular V_{VERD} e E_{VERD} da Definição (4.2) com relação à distribuição correspondente.

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

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

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

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 ideia 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.

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

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 populacional $\overline{Y}$ dada por $\overline{Y} = \frac{1}{N} \sum\limits_{h=1}^{2} \sum_{i \in U_h} y_{hi}$, 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 $\bar{y}_h = \frac{1}{n_h} \sum\limits_{i \in s_h} y_{hi}$ 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

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

onde $f_h = n_h / N_h$, n_h é o número de observações na amostra no estrato h, e $S_h^2 = \frac{1}{N_h - 1} \sum\limits_{i \in U_h} \left(y_{hi} - \overline{Y}_h\right)^2$ é a variância populacional da variável de pesquisa y dentro do estrato h, com $\overline{Y}_h = \frac{1}{N_h} \sum\limits_{i \in U_h} y_{hi}$ 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 $s^2 = \frac{1}{n-1} \sum\limits_{h=1}^{2} \sum\limits_{i \in s_h} \left(y_{hi} - \bar{y}\right)^2$ 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 considerado. 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 $\left(\hat{V}_{AES}\ (\bar{y}_w)\right)$.

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

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

Quantidade de interesse	Salários	Receitas
1) Média populacional	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

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 significantemente 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 consequentemente 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 $\hat{V}_{AES}\left(\bar{y}_w\right)$, que foi obtido substituindo as variâncias dentro dos estratos S_h^2 por estimativas amostrais não viciadas dadas por $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)^2$, $h = 1, 2$, na fórmula de $V_{AES}\left(\bar{y}_w\right)$ conforme definida em (4.3).

Como neste exercício a população é conhecida, podemos calcular $V_{AES}\left(\bar{y}_w\right)$ 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.

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

Quantidade de interesse	Salários	Receitas
1) Variância populacional $V_{AES}\left(\bar{y}_w\right)$	244.1	0.43500
2) Média de $\widehat{V}_{AES}\left(\bar{y}_w\right)$ usando 500 amostras	231.84	0.32569
3) Valor esperado de v_0 usando população	1613.3	1.1880
4) Média de v_0 usando 500 amostras	1636.1	1.2121

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

$$\begin{aligned}\textbf{EPA}\left(\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right) &= \frac{231,84}{1.636,1} = 0,142\\ \textbf{EPA}\left(\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right) &= \frac{0,32569}{1,2121} = 0,269\end{aligned}$$

$$\begin{aligned}\textbf{EPA}\left(\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right) &= \frac{244,18}{1.613,3} = 0,151 \text{ e}\\ \textbf{EPA}\left(\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right) &= \frac{0,43500}{1,1880} = 0,366.\end{aligned}$$

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.

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

Variável	Estimativa	Simulação	População
Salário	Variância EPA	231.84 0.142	244.18 0.151
Receita	Variância EPA	0.32569 0.269	0.43500 0.366

Observação. Neste exemplo não foi feito uso analítico dos dados e sim descritivo, onde é usual incorporar os pesos no cálculo de estimativas e variâncias. Não seria esperado usar um estimador ponderado para a média e não considerar os pesos no cálculo de variâncias, como fizemos neste exemplo.
Observação. O exemplo mostra que ignorar a estratificação ao calcular v_0 diminui o **EPA**.

Um outro exemplo relevante é utilizado a seguir para ilustrar o fato de que o conceito do **EPA** adotado aqui é mais abrangente do que o definido por Kish, em particular porque a origem do efeito pode estar na estrutura da população e não no plano amostral usado para obter os dados.
Exemplo 4.3. População conglomerada com conglomerados de tamanho 2 (Skinner et al., 1989), p. 25

Considere uma população de 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 θ e variância σ^2, além de uma correlação ρ entre os valores dentro de cada par (correlação intraclasse, veja (Nascimento Silva and 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

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

Um estimador não viciado para θ é dado por $\widehat{\theta} = (y_1 + y_2)/2$, 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 $\widehat{\theta}$ é dada por

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

com um estimador não viciado dado por

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

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

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

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

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

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

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

e

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

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

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 $\mathbf{EPA}_{Kish}\left(\hat{\theta}\right)$ cresce pouco com o valor do coeficiente de correlação intraclasse ρ, o que implica que um plano amostral conglomerado como o adotado (seleção ao acaso de um par da população) seria menos eficiente que um plano amostral aleatório simples (seleção de duas unidades ao acaso da população), mas a perda de eficiência seria modesta. Já se o interesse é medir, a posteriori, o efeito da má especificação do plano amostral no estimador de variância, o impacto, medido pelo $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$, seria muito maior.

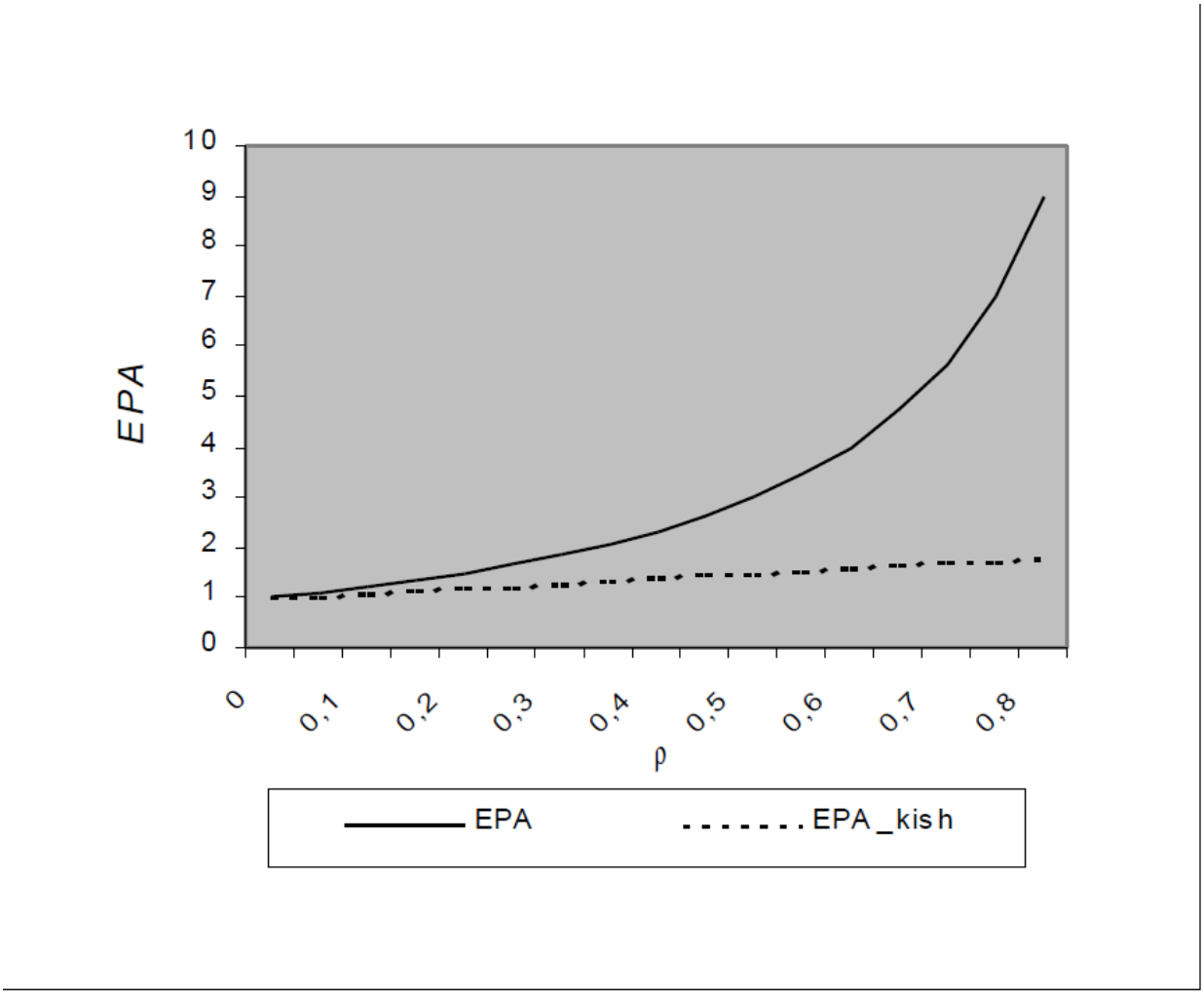


Figura 4.1: Valores de EPA e EPA de Kish para conglomeração

Vale ainda notar que o **EPA** $\left(\hat{\theta}, v_0\right)$ mede o impacto da má especificação do plano amostral ou do modelo para a estrutura populacional. Neste exemplo, ignorar a estrutura da população (o fato de que as observações são pareadas) poderia provocar subestimação da variância do estimador de média, que seria tanto maior quanto maior fosse o coeficiente de correlação intraclasse ρ. Efeitos como esse são comuns 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 θ (da população finita ou do modelo de superpopulação) é possível construir um intervalo de confiança de nível de confiança aproximado $(1-\alpha)$ a partir da distribuição assintótica de

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

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

Neste caso, um intervalo de confiança de nível de confiança aproximado $(1-\alpha)$ é dado por $\left[\hat{\theta} - z_{\alpha/2} v_0^{1/2}, \hat{\theta} + z_{\alpha/2} v_0^{1/2}\right]$, onde z_α é definido por $\int_{z_\alpha}^{+\infty} \varphi(t)\, dt = \alpha$, onde φ 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

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

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

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

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 **EPA** $\left(\hat{\theta}, v_0\right)$. Porém quando a amostra é grande esse valor aproxima o **EPA** $\left(\hat{\theta}, v_0\right) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)}$, pois v_0 é aproximadamente igual a $E_{VERD}(v_0)$ e $\widehat{V}_{VERD}\left(\hat{\theta}\right)$ é aproximadamente igual a $V_{VERD}\left(\hat{\theta}\right)$. Logo temos que a distribuição assintótica verdadeira de t_0 é dada por

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

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

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

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

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

EPA $\left(\hat{\theta}, v_0\right)$	$1-\alpha = 0.95$	$1-\alpha = 0.99$
0.90	0.96	0.99
0.95	0.96	0.99
1.0	0.95	0.99
1.5	0.89	0.96
2.0	0.83	0.93
2.5	0.78	0.90
3.0	0.74	0.86
3.5	0.71	0.83
4.0	0.67	0.80

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

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

Exemplo 4.4. Teste de hipótese sobre proporção

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

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

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

$$Z = |\widehat{p} - p_0| / \sqrt{p_0 \left(1 - p_0\right) / n}, \tag{4.4}$$

onde o denominador é o desvio padrão da estimativa $\widehat{p}$ 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 $\widehat{p}$, 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,

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

onde $\sqrt{0,8\,(1-0,8)\,/1.000} = 0,0126$ é o desvio padrão de $\widehat{p}$ 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 $\overline{n} = 50$, em contraste com o valor 1.000 usado no teste anterior. O termo `tamanho efetivo` foi introduzido em (Kish, 1965) para designar o tamanho de uma amostra aleatória simples necessário para estimar p com a mesma precisão obtida por uma amostra conglomerada de tamanho n (neste caso, igual a 1.000) unidades elementares.

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

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

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 θ. Consideremos agora o problema de estimação de um vetor θ de K parâmetros. Seja $\hat{\theta}$ um estimador de θ e seja $\mathbf{V}_0$ um estimador da matriz $K \times K$ de covariância de $\hat{\theta}$, 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` $\hat{\theta}$ `e` $\mathbf{V}_0$ como

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

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

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

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

No caso particular onde $\mathbf{\Delta} = \mathbf{I}_{K\times K}$, temos $\delta_1 = \ldots = \delta_K = 1$ e os efeitos (univariados) do plano amostral das combinações lineares para componentes de $\hat{\theta}$ 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 $\mathbf{EMPA}\left(\hat{\theta}, \mathbf{V}_0\right)$, onde $\hat{\theta} = \left(\overline{SAL}_w, \overline{REC}_w\right)'$. Neste exemplo, os dados populacionais são conhecidos, e portanto podemos calcular a covariância dos estimadores $\left(\overline{SAL}_w, \overline{REC}_w\right)$. Usando a mesma notação do Exemplo 4.2, temos que

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

onde

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

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

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

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 $\mathbf{EPA}\left(\overline{SAL}_w\right)$ e de $\mathbf{EPA}\left(\overline{REC}_w\right)$ lá apresentados. Para calcular o $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0)$ é preciso agora obter $\mathbf{E}_{VERD}\left(\mathbf{V}_0\right)$.

Neste exemplo, a matriz de efeito do plano amostral $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0) = \mathbf{\Delta}$ pode também ser calculada através de simulação, de modo análogo ao que foi feito no Exemplo 4.2. Para isto, foram utilizadas outras 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:

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

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

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

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** $\left(\hat{\theta}, v_0\right)$ definido em (4.2) para o caso uniparamétrico foi utilizado para corrigir níveis de confiança de intervalos e níveis de significância de testes, o **EMPA**$(\hat{\theta}, \mathbf{V}_0)$ 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 $H_0 : \mu = \mu_0$, onde μ 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

$$T^2 = n\left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{S}_y^{-1}\left(\bar{\mathbf{y}} - \mu_0\right), \tag{4.11}$$

onde

$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{n}\sum_{i\in s}\mathbf{y}_i, \quad \mathbf{S}_y = \frac{1}{n-1}\sum_{i\in s}\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)', \text{ e} \\ \mu_0 &= \left(\mu_{10}, \mu_{20}, \ldots, \mu_{K0}\right)' . \end{aligned}$$

Se as observações $\mathbf{y}_i$ são IID normais, a estatística T^2 tem a distribuição $\frac{(n-1)}{(n-K)}\mathbf{F}\left(K; n-K\right)$ sob H_0, onde $\mathbf{F}\left(K; n-K\right)$ 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 $\chi^2\left(K\right)$ quando $n \to \infty$, (Johnson and 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 $Z_1, \ldots, Z_K$ são variáveis aleatórias independentes com distribuição normal padrão e os δ_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 $\delta_1 = \ldots = \delta_K = \delta$. Neste caso, o nível de significância real é dado aproximadamente por

$$P\left(\chi^2\left(K\right) > \chi_\alpha^2\left(K\right)/\delta\right) \tag{4.12}$$

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

A Tabela 4.5 apresenta os níveis de significância reais para $\alpha = 5\%$ para vários valores de K e δ. Mesmo quando os valores dos δ_i são distintos, os valores da Tabela 4.5 podem ser devidamente interpretados. Para isso, consideremos o $p$valor do teste da hipótese $H_0 : \mu = \mu_0$, sob a hipótese de amostragem aleatória simples com reposição e sob o plano amostral efetivamente utilizado. Por definição este valor é dado por

$$p\text{valor}_{AAS}\left(\bar{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{\Sigma}_{AAS}^{-1}\left(\bar{\mathbf{y}} - \mu_0\right)\right]$$

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

O verdadeiro valor-p pode ser definido analogamente como

$$p\text{valor}_{VERD}\left(\bar{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{\Sigma}_{VERD}^{-1}\left(\bar{\mathbf{y}} - \mu_0\right)\right] . \tag{4.13}$$

Os valores na Tabela 4.5 podem ser usados para quantificar a diferença entre estes valores-p. Consideremos a região crítica do teste de nível α baseado na hipótese de AAS:

$$RC_{AAS}(\bar{\mathbf{y}}) = \left\{\bar{\mathbf{y}} : (\bar{\mathbf{y}} - \mu_0)' \boldsymbol{\Sigma}_{AAS}^{-1} (\bar{\mathbf{y}} - \mu_0) > \chi^2_\alpha(K)\right\} = \{\bar{\mathbf{y}} : p\text{valor}_{AAS}(\bar{\mathbf{y}}) < \alpha\}. \quad (4.14)$$

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:

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

Observe que o segundo membro de (4.15) é da mesma forma que o segundo membro de (4.12). Logo, os valores da Tabela 4.5 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 δ_1 no lugar de δ.

\begin{table}

\caption{Níveis de significância (%) verdadeiros do teste T2 para o nível nominal de 5% assumindo autovalores iguais para delta.}

delta	K=1	K=2	K=3	K=4
0.9	4	4	3	3
1.0	5	5	5	5
1.5	11	14	16	18
2.0	17	22	27	31
2.5	22	30	37	43
3.0	26	37	46	53

\end{table}

4.6 Laboratório de R

Vale a pena incluir as contas que não dependem da simulação?

Exemplo 4.2 Parte de simulação do exemplo para salários.

```
library(survey)
# carrega dados
library(anamco)
data(popul)
```

```
N<-nrow(popul)
n1<-30; n2<-30
nh=c(n1,n2)
n<-sum(nh)
Nh<-table(popul$estrat)
fh<-nh/Nh ; Wh<-Nh/N ; f<- n/N
```

Nessa simulação usamos a library **sampling** (Tillé and Matei, 2016) (incluir?)

```
sal.mat.res<-matrix(NA,500,4)
rec.mat.res<-matrix(NA,500,4)
popul$sal <- popul$sal/1000
popul$rec <- popul$rec/1000000
library(sampling)
```

```
#Geracão dos dados
for(i in 1:500){
s<-strata(popul, "estrat",c(30,30),method= "srswor")
dados<-getdata(popul,s)
# média amostral de salário e de receita
sal_med_y<-mean(dados$sal)
rec_med_y<-mean(dados$rec)
# estimador v0
sal_var_aas<-(1-f)*var(dados$sal)/n
rec_var_aas<-(1-f)*var(dados$rec)/n
# vhat_aes estimador não-viciado
desenho<-svydesign(~1,strata=~estrat,data=dados,fpc=~Prob)
# estimador não-viciado da média de  salario e receita
sal_med_yw<-svymean(~sal,desenho)
rec_med_yw<-svymean(~rec,desenho)
sal.mat.res[i,]<-c(sal_med_y,coef(sal_med_yw),sal_var_aas,SE(sal_med_yw)^2)
rec.mat.res[i,]<-c(rec_med_y,coef(rec_med_yw),rec_var_aas,SE(rec_med_yw)^2)
}
# 1-media de ybar; 2- média de ybar_w; 3- média de v0; 4- média de vhat_aes
sal_res_mean <-colMeans(sal.mat.res)
rec_res_mean <- colMeans(rec.mat.res)
# epa = média de vhat_aes/ média de v0

epa_sal1 <- sal_res_mean[4]/sal_res_mean[3]
epa_sal1
```

```
## [1] 0.1398175
```

```
epa_rec1 <- rec_res_mean[4]/rec_res_mean[3]
epa_rec1
```

```
## [1] 0.3103891
```

Usando a fórmula (4.3)

```
# variância sob distr. de aleatorização verdadeira
vp_sal_h <- aggregate(sal~estrat, popul, var)
vp_rec_h <- aggregate(rec~estrat, popul, var)
vp_sal <- sum(Wh^2*(1-fh)*vp_sal_h[,"sal"]/nh)
vp_rec <- sum(Wh^2*(1-fh)*vp_rec_h[,"rec"]/nh)
```

A seguir, reproduzimos a Tabela 4.3, usando os resultados da simulação:

```
tab_43 <- data.frame(
Quantidade_de_interesse = c("Média_populacional_Ybar", "Média_de_ybar_sobre_ 500_ amostras",
  "Média_de_ybarw_sobre_500_amostras"),
Salários = c(mean(popul$sal), sal_res_mean[1],sal_res_mean[2] ),
Receitas = c(mean(popul$rec),rec_res_mean[1],rec_res_mean[2] )
)
knitr::kable(tab_43, booktabs=TRUE )
```

Quantidade_de_interesse	Salários	Receitas
Média_populacional_Ybar	78.32826	2.106846
Média_de_ybar_sobre_ 500_ amostras	166.71024	4.329960
Média_de_ybarw_sobre_500_amostras	79.35198	2.132731

A seguir reproduzimos a Tabela 4.4, usando os resultados da simulação:

```
tab_44 <- data.frame(
Quantidade_de_interesse = c("Variância_populacional_V_AES(ybarw)", "Média_de_ Vhat_AES(yvarw)",
"Valor_esperado_de_v0_usando_população","Média_de_v0_usando_500_amostras"),
Salários= c(vp_sal, sal_res_mean[4], NA, sal_res_mean[3]),
Receitas = c(vp_rec, rec_res_mean[4], NA, rec_res_mean[3])
 )
knitr::kable(tab_44, booktabs= TRUE)
```

Quantidade_de_interesse	Salários	Receitas
Variância_populacional_V_AES(ybarw)	244.1758	0.4349945
Média_de_ Vhat_AES(yvarw)	248.1744	0.4074866
Valor_esperado_de_v0_usando_população	NA	NA
Média_de_v0_usando_500_amostras	1774.9875	1.3128248

Exemplo 4.6. Teste da igualdade de médias para duas populações

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

```
# carregar dados
library(anamco)
dim(amolim)
```

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

```
names(amolim)
```

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

- Objeto de desenho para os dados da Amostra de Limeira:

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

- Vamos estimar, a renda média por raça:

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

```
##   raca     renda        se
## 1    1 110405.93 11261.845
## 2    2  73559.84  8207.357
```

- Vamos estimar, a renda média por sexo:

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

```
##   sexo     renda        se
## 1    1 108746.01 11695.974
## 2    2  40039.39  4042.393
```

- Vamos testar a igualdade de rendas por sexo:

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

```
##
##  Design-based t-test
##
## data:  renda ~ sexo
```

```
## t = -5.9251, df = 23, p-value = 4.855e-06
## alternative hypothesis: true difference in mean is not equal to 0
## sample estimates:
## difference in mean
##          -68706.63
```

- Vamos testar a igualdade de rendas por raça:

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

```
##
##  Design-based t-test
##
## data:  renda ~ raca
## t = -3.9714, df = 23, p-value = 0.000604
## alternative hypothesis: true difference in mean is not equal to 0
## sample estimates:
## difference in mean
##          -36846.09
```

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 and Doksum, 1977) e (Garthwaite et al., 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 do 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 et al., 1992). Já os métodos requeridos para inferências analíticas só recentemente foram consolidados em livro ((Skinner et al., 1989)). Este capítulo apresenta um dos métodos centrais disponíveis para ajuste de modelos paramétricos regulares considerando dados amostrais complexos, baseado no trabalho de (Binder et al., 1987). 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 subseqüente. 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 et al., 1995).

5.2 Método de Máxima Verossimilhança (MV)

Seja $\mathbf{y}_i = (y_{i1}, \ldots, y_{iR})'$ 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 $i = 1, \ldots, n$, onde n é o tamanho da amostra. Suponha que os vetores aleatórios $\mathbf{Y}_i$, para $i = 1, \ldots, n$, são independentes e identicamente distribuídos (IID) com distribuição comum $f(\mathbf{y}; \theta)$, onde $\theta = (\theta_1, \ldots, \theta_K)'$ é um vetor $K \times 1$ de parâmetros desconhecidos de interesse. Sob essas hipóteses, a verossimilhança amostral é dada por

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

e a correspondente log-verossimilhança por

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

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

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

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

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

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

e um estimador consistente dessa variância é dado por

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

onde

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

e

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

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 Amostra 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.

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

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

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

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 $\widehat{Y}_R = \widehat{Y}_\pi \, X \, / \, \widehat{X}_\pi = \sum_{i \in s} w_i^R y_i$ (tal

como definidos por (3.15), com pesos dados por $w_i^R = \pi_i^{-1} X \,/\, \widehat{X}_\pi$ (veja (3.17), onde X é o total da população no estrato obtido por métodos demográficos de projeção, utilizado como variável auxiliar, e $\widehat{X}_\pi$ e $\widehat{Y}_\pi$ são os estimadores π-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

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

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 et al., 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:

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

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

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

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 $\mathbf{W}_s = diag\,\{w_i; i \in s\}$ é a matriz diagonal com os pesos amostrais.

Não é possível justificar o estimador $\widehat{\beta}_w$ 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 $\widehat{\beta}_w$ 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. Nes-te 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 $\widehat{\beta}_w$ acoplado a estimadores de variância que incorporam o planejamento amostral e os pesos, tal como se verá mais adiante.

O estimador $\widehat{\beta}_w$ foi proposto formalmente por (Fuller, 1975), que o concebeu como uma função de estimadores de totais populacionais. A mesma ideia subsidiou vários outros autores que estudaram a estimação de coeficientes de regressão partindo de dados amostrais complexos, tais como (Nathan and Holt, 1980), (Pfeffermann and 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 cap. 6, (Nascimento Silva, 1996).

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 $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ são IID com densidade $f(\mathbf{y}, \theta)$. Se todos os elementos da população finita U fossem conhecidos, as funções de verossimilhança e de log-verossimilhança `populacionais` seriam dadas respectivamente por

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

e

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

As equações de verossimilhança `populacionais` correspondentes são dadas por

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

onde

$$\mathbf{u}_i(\theta) = \partial \log\left[f(\mathbf{y}_i;\theta)\right]/\partial\theta \tag{5.6}$$

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

Sob condições de regularidade (Cox and Hinkley, 1974), p. 281, a solução θ_U deste sistema é o `Estimador de Máxima Verossimilhança` de θ no caso de um `censo`. Podemos considerar θ_U como uma `Quantidade Descritiva Populacional Correspondente (QDPC)` a θ, no sentido definido por (Pfeffermann, 1993), sobre a qual se deseja fazer inferências com base em informações da amostra. Essa definição da QDPC θ_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 θ_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, $\theta_U - \theta = o_p(1)$. Como em pesquisas por amostragem o tamanho da população é geralmente grande, um estimador adequado para θ_U será geralmente adequado também para θ.

Seja $\mathbf{T} = \sum_{i \in U} \mathbf{u}_i(\theta)$ 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 $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\theta)$ (veja Capítulo 2.4) onde w_i são pesos propriamente definidos. Com essa notação, podemos agora obter um estimador para θ_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 (MPV) $\hat{\theta}_{MPV}$ de θ_U (e consequentemente de θ) será a solução das equações dePseudo-Verossimilhança dadas por

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

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 $\hat{\theta}_{MPV}$ e seu estimador correspondente, dados respectivamente por:

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

e

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

onde

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

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

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

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

o que fornece uma base para a inferência sobre θ_U (ou θ) 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 no apêndice de (Binder, 1983) 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 π-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 $f(\mathbf{y};\theta)$ pelo método da Máxima Pseudo-Verossimilhança seria dado pelos passos indicados a seguir.

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

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

onde

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

e

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

3. Usar $\hat{\theta}_\pi$ e $\hat{V}_p\left(\hat{\theta}_\pi\right)$ 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 θ.

Observação. 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. 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 π_i e os π_{ij} respectivamente.

Observação. Sob probabilidades de seleção iguais, os pesos π_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 $f(\mathbf{y};\theta)$. 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 θ_U da população finita permanece um alvo válido para inferência, mesmo quando o modelo especificado por $f(\mathbf{y};\theta)$ 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 $N \to \infty$ a QDPC θ_U converge para o parâmetro θ, isto é, $\theta_U - \theta \to \mathbf{0}$ em probabilidade, segundo a distribuição de probabilidades do modelo M. Se $\hat{\theta}_{MPV}$ for consistente, então quando $n \to \infty$ temos que $\hat{\theta}_{MPV} - \theta_U \to \mathbf{0}$ 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

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

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

Nesse caso, o parâmetro θ do modelo não tem interpretação substantiva significante, porém a QDPC θ_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 θ_U, a inferência baseada no procedimento MPV segue válida para este pseudo-parâmetro, independente da inadequação do modelo para a população. (Skinner, 989b), p. 81, discute essa situação, mostrando que θ_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. 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 $n \to \infty$. 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 $n \to \infty$ e $N \to \infty$. 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 et al., 1992), p. 166, (Brewer, 1979), (Isaki and Fuller, 1982), (Robinson and Särndal, 1983), (Hájek, 1960) e (Skinner et al., 1989), p. 18-19.

5.6 Desvantagens da Inferência de 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.

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

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

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;

$N_a/\widehat{N}_a-$ 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` $\widehat{N}_k$ 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 χ^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.

5.7 Laboratório de R

Usar função svymle da library `survey` (Lumley, 2017) para incluir exemplo de estimador MPV?

Possibilidade: explorar o exemplo 2.1?

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 $\mathbf{z}_i = (z_{i1}, \ldots, z_{iP})^{'}$ 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 predizer 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 $\left(Y_i, \mathbf{Z}_i^{'}\right)^{'}$, $i \in U$, são vetores aleatórios independentes e identicamente distribuídos tais que

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

onde $\beta = (\beta_1, \ldots, \beta_P)^{'}$ 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 β e σ_e e sua interpretação com respeito ao modelo agregado (6.1).

Modelos como (6.1) já foram considerados por vários autores, por exemplo (Holt et al., 1980b), (Nathan and Holt, 1980), pág. 81 de (Skinner, 989b) , (Chambers, 1986), (Chambers, 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 β e σ_e correspondentes ao modelo (6.1) podem ser facilmente obtidas como

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

e

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

6.1.2 Pseudo-parâmetros do Modelo

Se todos os elementos da população tivessem sido pesquisados, os EMVs de β e σ_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

$$\sum_{i\in U}\mathbf{u}_i\left(\mathbf{B}\right)=\sum_{i\in U}\mathbf{z}_i\left(y_i-\mathbf{z}_i'\beta\right)=\mathbf{z}_U^{'}\mathbf{y}_U-\left(\mathbf{z}_U^{'}\mathbf{z}_U\right)\mathbf{B}=\mathbf{0} \tag{6.3}$$

e

$$\sum_{i\in U}u_i\left(S_e\right)=\sum_{i\in U}\left[\left(y_i-\mathbf{z}_i'\mathbf{B}\right)^2-S_e\right]=\left(\mathbf{y}_U-\mathbf{z}_U'\mathbf{B}\right)^{'}\left(\mathbf{y}_U-\mathbf{z}\mathbf{z}_U'\mathbf{B}\right)-NS_e=0 \tag{6.4}$$

onde $\mathbf{z}_U=\left(\mathbf{z}_1,\ldots,\mathbf{z}_N\right)^{'}$ e $\mathbf{y}_U=\left(y_1,\ldots,y_N\right)^{'}$.

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

$$\mathbf{B}=\left(\mathbf{z}_U^{'}\mathbf{z}_U\right)^{-1}\mathbf{z}_U^{'}\mathbf{y}_U \tag{6.5}$$

e

$$S_e=N^{-1}\sum_{i\in U}\left(y_i-\mathbf{z}_i'\mathbf{B}\right)^2=N^{-1}\left(\mathbf{y}_U-\mathbf{z}_U'\mathbf{B}\right)^{'}\left(\mathbf{y}_U-\mathbf{z}_U'\mathbf{B}\right)\ . \tag{6.6}$$

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 β_2 (igual a β excluído o primeiro componente), β_1 e σ_e são dados respectivamente por

$$\mathbf{B}_2=\mathbf{S}_{\mathbf{z}}^{-1}\mathbf{S}_{\mathbf{z}y}\ , \tag{6.7}$$

$$B_1 = \bar{Y} - \bar{\mathbf{Z}}^{'}\mathbf{B}_2 \ , \tag{6.8}$$

e

$$S_e = N^{-1}\sum_{i\in U}\left(y_i - B_1 - \mathbf{z}_i^{'}\mathbf{B}_2\right)^2 = N^{-1}\sum_{i\in U} e_i^2 \ , \tag{6.9}$$

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

Os EMVs do censo dados em (6.1) a (6.9) coincidem com os estimadores de mínimos quadrados ordinários, sob as hipóteses mais fracas do modelo dadas por (6.10) a seguir (ver Nathan e Holt, 1980), onde se dispensou a hipótese de normalidade dos erros, isto é

$$\begin{aligned} E_M\left(Y_i|\,\mathbf{z}_i = \mathbf{z}_i\right) &= \beta_1 + \mathbf{z}_i^{'}\beta_2 \\ V_M\left(Y_i|\,\mathbf{z}_i = \mathbf{z}_i\right) &= \sigma_e \\ COV_M\left(Y_i, Y_j|\,\mathbf{z}_i = \mathbf{z}_i, \mathbf{z}_j = \mathbf{z}_j\right) &= 0 \quad \forall i \neq j \in U. \end{aligned} \tag{6.10}$$

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 β e σ_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

$$\begin{aligned} \sum_{i\in s} w_i\mathbf{u}_i\left(\hat{\mathbf{B}}_w\right) &= \sum_{i\in s} w_i\mathbf{z}_i\left(y_i - \mathbf{z}_i^{'}\hat{\mathbf{B}}_w\right) \\ &= \mathbf{z}_s^{'}\mathbf{W}_s\mathbf{y}_s - \left(\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{y}_s\right)\hat{\mathbf{B}}_w = \mathbf{0} \end{aligned} \tag{6.11}$$

e

$$\begin{aligned} \sum_{i\in s} w_i u_i\left(s_e^w\right) &= \sum_{i\in s} w_i\left[\left(y_i - \mathbf{z}_i^{'}\hat{\mathbf{B}}_w\right)^2 - s_e^w\right] \\ &= \left(\mathbf{y}_s - \mathbf{z}_s\hat{\mathbf{B}}_w\right)^{'}\mathbf{W}_s\left(\mathbf{y}_s - \mathbf{z}_s\hat{\mathbf{B}}_w\right) - \left(\mathbf{1}_s^{'}\mathbf{W}_s\mathbf{1}_s\right)s_e^w = 0 \end{aligned} \tag{6.12}$$

onde $\mathbf{z}_s$ e $\mathbf{y}_s$ são os análogos amostrais de $\mathbf{z}_U$ e $\mathbf{y}_U$, respectivamente, $\mathbf{W}_s = diag\left[(w_{i_1}, \ldots, w_{i_n})\right]$ 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 β e σ_e respectivamente.

Supondo que $\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{z}_s$ é não-singular e resolvendo (6.11) e (6.12) em $\hat{\mathbf{B}}_w$ e s_e^w obtemos as seguintes expressões para os estimadores MPV dos parâmetros do modelo:

$$\widehat{\mathbf{B}}_w = \left(\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{z}_s\right)^{-1}\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{y}_s \tag{6.13}$$

e

$$\begin{aligned} s_e^w &= \left(\mathbf{1}_s^{'}\mathbf{W}_s\mathbf{1}_s\right)^{-1}\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}_w\right)^{'}\mathbf{W}_s\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}_w\right) \\ &= \left(\mathbf{1}_s^{'}\mathbf{W}_s\mathbf{1}_s\right)^{-1}\mathbf{y}_s^{'}\left[\mathbf{W}_s - \mathbf{W}_s\mathbf{z}_s\left(\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{z}_s\right)^{-1}\mathbf{z}_s^{'}\mathbf{W}_s\right]\mathbf{y}_s \end{aligned} \tag{6.14}$$

sendo a segunda expressão para s_e^w obtida mediante substituição do valor de $\widehat{\mathbf{B}}_w$ em (6.13) na primeira linha de (6.14).

Observe que a hipótese de não-singularidade de $\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{z}_s$ 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 $\mathbf{z}_s^{'}\mathbf{W}_s\mathbf{z}_s$ é não-singular.

Estimadores pontuais dos parâmetros do modelo podem ser derivados a partir de (6.13) e (6.14) 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:

$$\widehat{\mathbf{B}} = \left(\mathbf{z}_s^{'}\mathbf{z}_s\right)^{-1}\mathbf{z}_s^{'}\mathbf{y}_s \tag{6.15}$$

e

$$s_e = n^{-1}\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}\right)^{'}\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}\right) . \tag{6.16}$$

Substituindo $\mathbf{W}_s$ em (6.13) e (6.14) por $diag\left(\pi_i : i \in s\right) = \mathbf{\Pi}_s^{-1}$, onde os π_i em geral não são todos iguais, obtemos estimadores, chamados de mínimos quadrados $\pi-$ponderados, dados por:

$$\widehat{\mathbf{B}}_\pi = \left(\mathbf{z}_s^{'}\mathbf{\Pi}_s^{-1}\mathbf{z}_s\right)^{-1}\mathbf{z}_s^{'}\mathbf{\Pi}_s^{-1}\mathbf{y}_s \tag{6.17}$$

e

$$s_e^\pi = \left(\mathbf{1}_s^{'}\mathbf{\Pi}_s^{-1}\mathbf{1}_s\right)^{-1}\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}_\pi\right)^{'}\mathbf{\Pi}_s^{-1}\left(\mathbf{y}_s - \mathbf{z}_s\widehat{\mathbf{B}}_\pi\right) . \tag{6.18}$$

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 β. As expressões a seguir são obtidas por aplicação direta dos resultados gerais fornecidos na Seção 5.4, observando-se que os escores correspondentes a β 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 $e_i = \left(y_i - \bar{Y}\right) - \left(\mathbf{z}_i - \bar{\mathbf{Z}}\right)' \mathbf{B}$ para $i \in U$, com o Jacobiano correspondente dado por

$$\begin{aligned} J(\mathbf{B}) &= \sum\nolimits_{i\in U} \partial \mathbf{z}_i (y_i - \mathbf{z}_i'\beta)/\partial\beta\Big|_{\beta=\mathbf{B}} \\ &= \partial(\mathbf{z}_U'\mathbf{y}_U - \mathbf{z}_U'\mathbf{z}_U\beta)/\partial\beta|_{\beta=\mathbf{B}} = -\mathbf{z}_U'\mathbf{z}_U \ . \end{aligned} \tag{6.19}$$

Substituindo em (6.7) e (6.8) os valores dos escores, do jacobiano e dos estimadores π-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

$$V_p\left(\widehat{\mathbf{B}}_\pi\right) = \left(\mathbf{z}_U'\mathbf{z}_U\right)^{-1} V_p\left(\sum_{i\in s} \pi_i^{-1}\mathbf{z}_i e_i\right)\left(\mathbf{z}_U'\mathbf{z}_U\right)^{-1} \tag{6.20}$$

e

$$\hat{V}_p\left(\widehat{\mathbf{B}}_\pi\right) = \left(\mathbf{z}_s'\mathbf{\Pi}_s^{-1}\mathbf{z}_s\right)^{-1} \hat{V}_p\left(\sum_{i\in s} \pi_i^{-1}\mathbf{z}_i e_i\right)\left(\mathbf{z}_s'\mathbf{\Pi}_s^{-1}\mathbf{z}_s\right)^{-1} \ , \tag{6.21}$$

onde

$$V_p\left(\sum_{i\in s} \pi_i^{-1}\mathbf{z}_i e_i\right) = \sum_{i\in U}\sum_{j\in U} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} e_i\mathbf{z}_i\mathbf{z}_j' e_j \ , \tag{6.22}$$

$$\hat{V}_p\left(\sum_{i\in s} \pi_i^{-1}\mathbf{z}_i \hat{e}_i\right) = \sum_{i\in s}\sum_{j\in s} \left(\pi_i^{-1}\pi_j^{-1} - \pi_{ij}^{-1}\right)\hat{e}_i\mathbf{z}_i\mathbf{z}_j'\hat{e}_j \ , \tag{6.23}$$

e $\hat{e}_i = y_i - \mathbf{z}_i'\widehat{\mathbf{B}}_\pi$ 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

$$f(y_i|\mathbf{z}_i,\beta) = \left[p\left(\mathbf{z}_i'\beta\right)\right]^{y_i}\left[1 - p\left(\mathbf{z}_i'\beta\right)\right]^{1-y_i}, \tag{6.24}$$

onde,

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

A função escore de β é

$$\mathbf{u}_i\left(\beta\right) = \partial \log(y_i|\mathbf{z}_i, \beta)/\partial\beta = \left[y_i - p\left(\mathbf{z}_i'\beta\right)\right]\mathbf{z}_i \tag{6.25}$$

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

$$\sum\nolimits_{i\in U} \mathbf{u}_i\left(\beta\right) = \sum\nolimits_{i\in U}\left[y_i - p\left(\mathbf{z}_i'\beta\right)\right]\mathbf{z}_i = \mathbf{0} \ . \tag{6.26}$$

O estimador de MPV do vetor de coeficientes β no modelo (6.24) é a solução da equação

$$\sum\nolimits_{i\in s} w_i\mathbf{u}_i\left(\beta\right) = \sum\nolimits_{i\in s} w_i\left[y_i - p\left(\mathbf{z}_i'\beta\right)\right]\mathbf{z}_i = \mathbf{0}, \tag{6.27}$$

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

A matriz de covariância do estimador de MPV de β pode ser obtida conforme indicado na Seção 5.4, bastando substituir os valores dos escores $\mathbf{u}_i\left(\beta\right) = \left[y_i - p\left(\mathbf{z}_i'\beta\right)\right]\mathbf{z}_i$ 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 14.

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.

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.

Tabela 6.1: Descrição das variáveis explicativas

Fatores	Níveis	Descrição dos níveis
Sexo(sx)	sx(1)sx(2)	HomensMulheres
Anos de estudo(ht)	ae(1)ae(2)ae(3)	Até 4De 5 a 89 ou mais
Horas trabalhadas(ht)	ht(1)ht(2)ht(3)	Menos de 40De 40 a 48Mais de 48
Idade em anos completos(id)	id(1)id(2)id(3)id(4)	Até 17De 18 a 25De 26 a 4950 ou mais
Rendimento médio mensal(re)	re(1)re(2)re(3)	Menos de 1De 1 a 5Mais de 5

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

$$\begin{aligned} \log\left(\frac{p_{ijklm}}{1-p_{ijklm}}\right) &= \mu+\beta_i^{sx}+\beta_j^{ae}+\beta_k^{ht}+\beta_l^{id}+\beta_m^{re} \\ &\quad +\beta_{ij}^{sx.id}+\beta_{ik}^{sx.ht}+\beta_{jk}^{ae.ht}+\beta_{kl}^{ht.id}+\beta_{km}^{ht.re}, \end{aligned} \tag{6.28}$$

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 sequencialmente 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 função `svyglm()`da library `survey` (Lumley, 2017).

As estimativas calculadas pela função `svyglm` são feitas pelo `Método de Máxima Pseudo-Verossimilhança`, resolvendo a equação (6.27). As estimativas dos desvios padrões são obtidas das variâncias calculadas pelo método de linearização descrito na Seção 5.4, equação (5.5), considerando os escores tal como apresentados na equação (6.25). 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, conforme descrito na Seção .

Tabela 6.2: Estmativas dos efeitos e respectivos erros padrões obtidos pela library survey do R

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.515	0.260	-1.978	0.048
sx1	0.148	0.222	0.666	0.506
ae1	0.745	0.165	4.528	0.000
ae2	0.496	0.156	3.176	0.002
ht1	-0.377	0.317	-1.187	0.236
ht2	-0.697	0.275	-2.531	0.012
id1	-0.239	0.540	-0.442	0.659
id2	-0.729	0.302	-2.412	0.016
id3	0.227	0.231	0.982	0.327
re1	0.286	0.277	1.032	0.302
re2	0.065	0.144	0.451	0.652
sx1:id1	0.878	0.348	2.521	0.012
sx1:id2	0.300	0.231	1.296	0.195
sx1:id3	-0.259	0.190	-1.363	0.173
sx1:ht1	-0.736	0.206	-3.572	0.000
sx1:ht2	-0.089	0.185	-0.480	0.631
ae1:ht1	0.792	0.240	3.294	0.001
ae2:ht1	0.739	0.227	3.261	0.001
ae1:ht2	0.026	0.197	0.132	0.895
ae2:ht2	0.089	0.183	0.488	0.626
ht1:id1	-1.420	0.605	-2.345	0.019
ht2:id1	-0.413	0.506	-0.817	0.414
ht1:id2	-0.124	0.355	-0.351	0.726
ht2:id2	-0.109	0.279	-0.391	0.696
ht1:id3	-0.220	0.248	-0.888	0.375
ht2:id3	-0.537	0.205	-2.619	0.009
ht1:re1	1.529	0.356	4.293	0.000
ht2:re1	0.338	0.320	1.056	0.292
ht1:re2	0.490	0.233	2.100	0.036
ht2:re2	-0.115	0.183	-0.629	0.530

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

Contraste	gl_num	gl_den	Estatística_F	valor_p
ht:re	4	616	6.743	0.000
ht:id	6	616	3.545	0.002
sx:id	3	616	6.996	0.000
sx:ht	2	616	9.540	0.000
ae:ht	4	616	4.722	0.001

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

ht	varia_ae	raz_vantagem
1	1 para 2	0.739
1	2 para 3	0.291
2	1 para 2	0.830
2	2 para 3	0.557
3	1 para 2	0.779
3	2 para 3	0.609

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 Tabela6.3, obtida pela função `svyglm()` da library survey, utiliza a estatística de Wald baseada no plano amostral com correção.

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 pela função `svyglm()` da library survey 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;

Tabela 6.5: Intervalos de confiança de 95de ht

ht	varia_ae	LIC	LSC
1	1 para 2	0.516	1.059
1	2 para 3	0.212	0.398
2	1 para 2	0.694	0.994
2	2 para 3	0.452	0.687
3	1 para 2	0.412	0.753
3	2 para 3	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	2308

2) dados da tabela estimada considerando os pesos e

3) dados individuais com pesos normalizados.

Em todas 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 $\beta = (\beta_1, \ldots, \beta_P)'$ o vetor de parâmetros num modelo de regressão. Como é sabido, para testar a hipótese $H_0 : \beta_j = 0$, para algum $j \in \{1, \ldots, P\}$, usamos um teste t, e para para testar a hipótese $H_0 : (\beta_{j_1}, \ldots, \beta_{j_R})' = \mathbf{0}_R$, onde $(j_1, \ldots, j_R) \subset (1, \ldots, P)$ 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 hipotético 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

$$H_0 : \mathbf{C}\beta = \mathbf{c}, \tag{6.29}$$

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 $H_0 : \beta_2 = \mathbf{0}_R$, onde $\beta' = (\beta_1', \beta_2')$, com β_1 de dimensão $Q \times 1$ e β_2 de dimensão $R \times 1$,

$\mathbf{C} = \left[\begin{array}{cc} \mathbf{0}_{R\times Q} & \mathbf{I}_R \end{array} \right]$ e $c = \mathbf{0}_R$, 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 (6.29) é definida por

$$X_W^2 = \left(\mathbf{C}\widehat{\beta} - \mathbf{c}\right)' \left(\mathbf{C}\widehat{\mathbf{V}}\left(\hat{\beta}\right)\mathbf{C}'\right)^{-1} \left(\mathbf{C}\widehat{\beta} - \mathbf{c}\right), \tag{6.30}$$

onde os estimadores $\widehat{\beta}$ e $\widehat{\mathbf{V}}\left(\hat{\beta}\right)$ 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, $\hat{\beta}$ pela estimativa MPV $\widehat{\mathbf{B}}_\pi$ de β dada em (6.17), e $\widehat{\mathbf{V}}\left(\hat{\beta}\right)$pela estimativa da matriz de covariância do estimador de MPV $\hat{V}_p\left(\widehat{\mathbf{B}}_\pi\right)$ dada em (6.21). Tais estimativas consideram os pesos diferentes das observações e o plano amostral efetivamente utilizado. A normalidade assintótica do estimador de MPV de β 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 $\mathbf{V}\left(\hat{\beta}\right)$. Uma alternativa é usar a aproximação

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

onde $v = m - H$ é o número de UPAs da amostra menos o n'{u}mero 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^2_W, 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.

6.4 Laboratório de R

Usar exemplo da amolim ou conseguir exemplo melhor? Reproduzir usando a survey os resultados do Exemplo 6.1???

```
library(survey)
library(anamco)
names(pnadrj90)
```

```
##  [1] "stra"     "psu"      "pesopes"  "informal" "sx"       "id"
##  [7] "ae"       "ht"       "re"       "um"
```

Preparação dos dados: Variáveis explicativas são fatores. Ver tipo de variável:

```
unlist(lapply(pnadrj90, mode))
```

```
##      stra       psu   pesopes  informal        sx        id        ae
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        ht        re        um
## "numeric" "numeric" "numeric"
```

Transformar variáveis para fatores e mudar o nível básico do fator (último)

```
pnadrj90$sx<-as.factor(pnadrj90$sx)
pnadrj90$sx<-relevel(pnadrj90$sx,ref="2")
pnadrj90$id<-as.factor(pnadrj90$id)
pnadrj90$id<-relevel(pnadrj90$id,ref="4")
pnadrj90$ae<-as.factor(pnadrj90$ae)
pnadrj90$ae<-relevel(pnadrj90$ae,ref="3")
pnadrj90$ht<-as.factor(pnadrj90$ht)
pnadrj90$ht<-relevel(pnadrj90$ht,ref="3")
pnadrj90$re<-as.factor(pnadrj90$re)
pnadrj90$re<-relevel(pnadrj90$re,ref="3")
##transformar variável de resposta para 0,1:
pnadrj90$informal<-ifelse(pnadrj90$informal==1,1,0)
```

Cria objeto de desenho

```
pnad.des<-svydesign(id=~psu,strata=~stra,weights=~pesopes,data=pnadrj90,nest=TRUE)
```

Ajusta modelo de regressão logística na Tabela 6.2 Comparar resultado com o da página 106 de Pessoa e Silva (1998)

```
inf.logit<-svyglm(informal~sx+ae+ht+id+re+sx*id+sx*ht+ae*ht+ht*id+ht*re,
  design=pnad.des, family=quasibinomial())
```

```
knitr::kable(summary(inf.logit)$coefficients,booktabs=TRUE, digits= c(3,3,3,2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.515	0.260	-1.978	0.05
sx1	0.148	0.222	0.666	0.51
ae1	0.745	0.165	4.528	0.00
ae2	0.496	0.156	3.176	0.00
ht1	-0.377	0.317	-1.187	0.24
ht2	-0.697	0.275	-2.531	0.01
id1	-0.239	0.540	-0.442	0.66
id2	-0.729	0.302	-2.412	0.02
id3	0.227	0.231	0.982	0.33
re1	0.286	0.277	1.032	0.30
re2	0.065	0.144	0.451	0.65
sx1:id1	0.878	0.348	2.521	0.01
sx1:id2	0.300	0.231	1.296	0.20
sx1:id3	-0.259	0.190	-1.363	0.17
sx1:ht1	-0.736	0.206	-3.572	0.00
sx1:ht2	-0.089	0.185	-0.480	0.63
ae1:ht1	0.792	0.240	3.294	0.00
ae2:ht1	0.739	0.227	3.261	0.00
ae1:ht2	0.026	0.197	0.132	0.90
ae2:ht2	0.089	0.183	0.488	0.63
ht1:id1	-1.420	0.605	-2.345	0.02
ht2:id1	-0.413	0.506	-0.817	0.41
ht1:id2	-0.124	0.355	-0.351	0.73
ht2:id2	-0.109	0.279	-0.391	0.70
ht1:id3	-0.220	0.248	-0.888	0.37
ht2:id3	-0.537	0.205	-2.619	0.01
ht1:re1	1.529	0.356	4.293	0.00
ht2:re1	0.338	0.320	1.056	0.29
ht1:re2	0.490	0.233	2.100	0.04
ht2:re2	-0.115	0.183	-0.629	0.53

Teste de Wald para a hipótese $H_0 : ht : re = 0$

```
regTermTest(inf.logit,"ht:re")
```

```
## Wald test for ht:re
##  in svyglm(formula = informal ~ sx + ae + ht + id + re + sx * id +
##     sx * ht + ae * ht + ht * id + ht * re, design = pnad.des,
##     family = quasibinomial())
## F =  6.742662  on  4  and  616  df: p= 2.58e-05
```

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 Exemplo @ref{ex:exebin). 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 α 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 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^2_P 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.

Tabela 7.1: Frequências observadas e proporções hipotéticas

Categoria	j	n_j	p_{0j}
Cobertos por planos de saúde	1	840	0.8
Não cobertos	2	160	0.2
Todos empregados	-	1000	1.0

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 (π_j, digamos) no contexto de um modelo de superpopulação.

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

$$X^2_P = \sum_{j=1}^{2} \left(n_j - n\,p_{0j}\right)^2 / \left(n\,p_{0j}\right) = n \sum_{j=1}^{2} \left(\widehat{p}_j - p_{0j}\right)^2 / p_{0j}, \tag{7.1}$$

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

```
p0 <- c(.8, .2)
obs<- c(840, 160)
n<- sum(obs)
phat <- obs/n
# Estatística de Pearson
x2p <- sum((obs-n*p0)^2/(n*p0))
x2p
```

```
## [1] 10
```

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

$$X^2_P = n \sum_{j=1}^{2} \left(\widehat{p}_j - p_{0j}\right)^2 / p_{0j} = \frac{\left(\widehat{p} - p_0\right)^2}{p_0\left(1 - p_0\right)/n} = Z^2_{bin} \tag{7.2}$$

onde $\widehat{p} = \widehat{p}_1$ 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 (χ^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 $\widehat{p}$ da proporção p. O efeito de plano amostral nesse caso é dado por $d = V_p(\widehat{p})/V_{bin}(\widehat{p})$.

Para uma amostra de empregados selecionada por amostragem aleatória simples, teríamos $d = 1$ pois $V_p(\widehat{p})$ e $V_{bin}(\widehat{p})$ 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 $E_p\left(X_P^2\right) = d$. Como $E_p\left(X_P^2/d\right) = E\left(\chi^2(1)\right) = 1$, 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,

$$X_P^2(d) = X_P^2/d, \tag{7.3}$$

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

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

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

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

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

```
p <- p0[1]
m <- 50
# Efeito do plano amostral
d <- (p*(1-p)/m)/ (p*(1-p)/n)
d
```

```
## [1] 20
```

```
# Estatística de Peason corrigida
x2pd <- x2p/d
x2pd
```

```
## [1] 0.5
```

```
# valor-p do teste
pchisq(x2pd, 1, lower.tail = F)
```

```
## [1] 0.4795001
```

$$\begin{array}{ccccc} d & = & V_p\left(\widehat{p}\right)/V_{bin}\left(\widehat{p}\right) & = & \frac{p(1-p)/m}{p(1-p)/n} \\ & = & \frac{0{,}0032}{0{,}00016} & = & 20 \end{array}$$

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

```
# Estatística da RV
x2rv <- 2*n*sum(phat*log(phat/p0))
x2rv
```

```
## [1] 10.56154
```

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

com $p$valor $0,0012$, e

```
# Estatística da RV corrigida
x2rvd <- x2rv/d
x2rvd
```

```
## [1] 0.528077
```

```
# valor-p do teste
pchisq(x2rvd, 1, lower.tail = FALSE)
```

```
## [1] 0.4674165
```

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

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 $\widehat{d}$ correspondentes são maiores que um. Para conglomerados de tamanho médio igual a 20 ($m = 20$) como neste exemplo, os valores típicos de $\widehat{d}$ são menores que 3, tendo em correspondência correlações intraclasse estimadas positivas $\widehat{\rho} < 0,1$.

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^2_N de Neyman. No caso de duas celas, ela se reduz a

$$X^2_N = n\sum_{j=1}^{2}\left(\hat{p}_j - p_{0j}\right)^2/\hat{p}_j = \frac{\left(\widehat{p} - p_0\right)^2}{\hat{p}\left(1-\hat{p}\right)/n} \quad . \tag{7.6}$$

Note que a expressão de X^2_N em (7.6) pode ser obtida substituindo-se no denominador de X^2_P 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

$$X^2_W = \left(\widehat{p} - p_0\right)^2/\hat{V}_p\left(\widehat{p}\right), \tag{7.7}$$

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

O efeito do termo $\hat{V}_p\left(\widehat{p}\right)$, que aparece no denominador de X^2_W, 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 $\widehat{V}_p\left(\widehat{p}\right)$ a variância $\widehat{V}_{bin}\left(\widehat{p}\right) = \widehat{p}\left(1-\widehat{p}\right)/n$. Neste caso, estatística resultante X^2_{bin} coincide com a estatística X^2_N de Neyman.

Para o plano amostral de conglomerados considerado no Exemplo 4.4, a estatística X^2_W, 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

```
x2n <- 1000* sum((phat-p0)^2/phat)
vhatp <- (phat[1]*phat[2])/m # variância usa número efetivo
# Estatística de Wald
x2w <- ((phat[1]-p0[1])^2)/vhatp
x2w
```

Tabela 7.2: Valores observados e valores-p de estatísticas de testes para os dados do Exemplo 4.4

Estatística	gl	vobs	valorp
Pearson	1	10.00	0.0016
Pearson ajustada	1	0.50	0.4795
RV	1	10.56	0.0012
RV ajustada	1	0.53	0.4674
Wald	1	0.60	0.4404

```
## [1] 0.5952381
```

```
pchisq(x2w,1, lower.tail = FALSE)
```

```
## [1] 0.4404007
```

$$X^2_W = (0,84 - 0,80)^2 / 0,002743 = 0,583 \ .$$

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.

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 $H_0 : p_j = p_{0j}$, $j = 1, \ldots, J$, 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 $\mathbf{p} = (p_1, \ldots, p_{J-1})'$ é o vetor de proporções populacionais desconhecidas e $\mathbf{p}_0 = (p_{01}, \ldots, p_{0\,J-1})'$ é o vetor de proporções hipotéticas.

O vetor de estimativas consistentes das proporções das celas, baseado em n observações, é denotado por $\widehat{\mathbf{p}}=(\widehat{p}_1,\ldots,\widehat{p}_{J-1})^{'}$, onde $\widehat{p}_j=\widehat{n}_j/n$. Os $\widehat{n}_j$ 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 $\sum_{j=1}^{J}\widehat{n}_j=n$. Se n não for fixado de antemão, os $\widehat{p}$ 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 $\widehat{\mathbf{p}}$, 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^2_W, 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

$$X^2_W=(\widehat{\mathbf{p}}-\mathbf{p}_0)^{'}\,\widehat{\mathbf{V}}_p^{-1}\,(\widehat{\mathbf{p}}-\mathbf{p}_0)\ , \tag{7.8}$$

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

Sob a hipótese nula H_0, a estatística X^2_W 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^2_W 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 $\widehat{\mathbf{V}}_p$, 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

$$F_{1.p}=\frac{f-J+2}{f\,(J-1)}X^2_W\ \ , \tag{7.9}$$

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

$$F_{2.p}=\frac{X^2_W}{(J-1)}, \tag{7.10}$$

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 **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 and 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

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

onde $\mathbf{P}_0 = diag\left(\mathbf{p}_0\right) - \mathbf{p}_0\mathbf{p}_0^{'}$ e $\mathbf{P}_0/n$ é a matriz $(J-1)\times(J-1)$ de covariância multinomial de $\widehat{\mathbf{p}}$ sob a hipótese nula, e $diag\left(\mathbf{p}_0\right)$ 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 $\widehat{\mathbf{V}}_p$. No caso de apenas duas celas, X_P^2 reduz-se à fórmula simples antes considerada $X_P^2 = \left(\widehat{p}_1 - p_{01}\right)^2/\left[p_{01}\left(1-p_{01}\right)/n\right]$, 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 $\delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_{J-1}W_{J-1}$ de $J-1$ variáveis aleatórias independentes W_j , cada uma tendo distribuição qui-quadrado com um grau de liberdade. Os pesos δ_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 $\widehat{\mathbf{p}}$ 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 δ_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 χ^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 δ_j tendem a ser maiores que um, e assim a distribuição assintótica da variável aleatória $\sum_{j=1}^{J-1}\delta_j W_j$ será 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 and 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 $\widehat{\mathbf{\Delta}}$.

Inicialmente, consideramos um `ajuste simples de EPA médio` à estatística X_P^2, devido a (Fellegi, 1980) e (Holt et al., 1980a), e o ajuste de primeira ordem de Rao-Scott. Estes ajustes são úteis nos casos em que não é possível obter uma estimativa adequada $\hat{\mathbf{V}}_p$ 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 $\hat{d}_{\cdot}$ dos efeitos univariados de plano amostral:

$$X_P^2\left(\hat{d}_{\cdot}\right) = X_P^2/\hat{d}_{\cdot} \tag{7.12}$$

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

Estimamos os efeitos do plano amostral por $\hat{d}_j = \hat{V}_p\left(\hat{p}_j\right) / \left(\hat{p}_j\left(1-\hat{p}_j\right)/n\right)$, onde $\hat{V}_p\left(\hat{p}_j\right)$ 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 $\hat{d}_{\cdot}$ 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 $E\left(X_P^2\right) = \sum_{j=1}^{J-1} \delta_j$, logo $E\left(X_P^2/\bar{\delta}\right) = E\left(\chi^2\left(J-1\right)\right) = J-1$, onde a média dos autovalores é $\bar{\delta} = \sum_{j=1}^{J-1} \delta_j/\left(J-1\right)$. Este raciocínio conduz ao ajuste de primeira ordem de Rao-Scott para X_P^2, dado por

$$X_P^2\left(\hat{\delta}_{\cdot}\right) = X_P^2/\hat{\delta}_{\cdot} \ , \tag{7.13}$$

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

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

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

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

A estatística ajustada $X_P^2\left(\hat{\delta}_{\cdot}\right)$ 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 $X_P^2\left(\hat{\delta}_{\cdot}\right)$ só são necessários os efeitos multivariados de plano amostral dos $\hat{p}_j$ que aparecem na diagonal da matriz $\hat{\mathbf{\Delta}}$. 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 $X_P^2\left(\hat{\delta}_.\right)$ é mais exato do que o ajuste do EPA médio da estatística $X_P^2\left(\hat{d}_.\right)$, que é considerada uma alternativa conservadora de $X_P^2\left(\hat{\delta}_.\right)$.

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

$$X_P^2\left(\hat{\delta}_., \hat{a}^2\right) = X_P^2\left(\hat{\delta}_.\right) / \left(1 + \hat{a}^2\right), \tag{7.14}$$

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

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

Um estimador da soma dos quadrados dos autovalores é dado por

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

onde $\hat{V}_p\left(\hat{p}_j, \hat{p}_k\right)$ 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 $X_P^2\left(\hat{\delta}_., \hat{a}^2\right)$ é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite dados por $gl_S = (J-1) / \left(1 + \hat{a}^2\right)$.

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

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

A estatística $FX_P^2\left(\hat{\delta}_.\right)$ tem distribuição de referência F com $J-1$ e f graus de liberdade. (Thomas and 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.

```
library(survey)
ppv1 <- readRDS("~\\GitHub\\adac\\data\\ppv.rds")
# cria idade categorizada
```

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	prop_contagem	frequência	prop_est_ppv	epa
0-14	0.2842	2516	0.2845	2.2862
15-29	0.2747	2360	0.2678	2.1867
30-44	0.2263	2018	0.2225	2.2542
45-59	0.1261	1177	0.1316	1.9906
60+	0.0860	832	0.0935	3.1632

```
ppv1<-transform(ppv1,idatab=cut(v02a08,
c(0,14,29,44,59,200), include.lowest=T), one=1)

# Objeto de desenho da PPV
ppv.des<-svydesign(id=~nsetor,strat=~estratof,weights=~pesof,
data=ppv1,nest=TRUE)
# Considera região sudeste
ppv.se.des<-subset(ppv.des,regiao==2)
# estima proporções nas classes de IDATAB
ppv.id<-svymean(~idatab,ppv.se.des,deff=T)
vhat<-vcov(ppv.id)
ppv.id <- data.frame(ppv.id)
freq <- svyby(~one, ~idatab, ppv.se.des, unwtd.count)$counts
# matriz de variância-covariância estimada
idad_tab <- c("0-14", "15-29", "30-44","45-59","60+")
tab73 <- data.frame(idade= idad_tab, prop_contagem= c(0.2842, 0.2747, 0.2263, 0.1261, 0.0860), frequênc
knitr::kable(tab73, booktabs= TRUE, digits=c(0,4,0,4,4), align = "lcccc",
 caption = "Vetores de proporções por classes de idade da PPV 96/97 e Contagem
 96 e EPAs calculados para a PPV - Região Sudeste")
```

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 $\widehat{\mathbf{V}}_p$ pela aplicação do método de linearização de Taylor descrito na Seção 3.3 através da fórmula (3.22) obtendo

	0-14	15-29	30-44	45-59	60+
0-14	52.274	-3.899	-5.672	-19.292	-23.411
15-29	-3.899	48.164	-29.346	-3.399	-11.520
30-44	-5.672	-29.346	43.799	-8.226	-0.556
45-59	-19.292	-3.399	-8.226	25.551	5.366
60+	-23.411	-11.520	-0.556	5.366	30.120

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 $\mathbf{P}_0/n = \frac{diag(\mathbf{p}_0)-\mathbf{p}_0\mathbf{p}_0'}{8.903}$, resultando em

22.850	-8.855	-7.224	-4.025	-2.745
-8.855	22.515	-7.051	-3.929	-2.680
-7.224	-7.051	19.666	-3.205	-2.186
-4.025	-3.929	-3.205	12.378	-1.218
-2.745	-2.680	-2.186	-1.218	8.829

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:

```
dhat <- tab73$epa
dhatdot <- mean(dhat)
dhatdot
```

```
## [1] 2.376168
```

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

```
phat <- tab73$prop_est_ppv
deltahatdot <- sum(phat*(1-p0)*dhat/(4*p0))
deltahatdot
```

```
## [1] 2.459607
```

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

```
nppv <- sum (freq)
raz <- matrix(NA,5,5)
for(i in 1:5){
  for(j in 1:5){
    raz[i,j]<- vhat[i,j]^2/(p0[i]*p0[j])
  }
}
ahat2_1 <- nppv^2* sum(raz)/(4*deltahatdot^2)
ahat2_1
```

```
## [1] 1.249524
```

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

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

```
x2p <-nppv* sum((phat -p0)^2/p0)
x2p
```

```
## [1] 11.50719
```

$$X_P^2 = 11,64$$

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

```
x2p/dhatdot
```

```
## [1] 4.842751
```

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

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

```
x2p/deltahatdot
```

```
## [1] 4.678467
```

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

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

```
J <- length(phat)
x2p/ ((J-1)*deltahatdot)
```

```
## [1] 1.169617
```

$$FX_P^2\left(\hat{\delta}_{.}\right) = 4,74/4 = 1,85$$

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

```
(x2p/deltahatdot)/ahat2_1
```

```
## [1] 3.744199
```

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

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

```
phatv <- matrix(phat[-5], ncol=1)
p0v <- matrix(p0[-5], ncol=1)
vhat0 <- vhat[-5,-5]
x2w <-  t(phatv-p0v)%*% solve(vhat0)%*%(phatv-p0v)
x2w
```

```
##          [,1]
## [1,] 5.742022
```

$$X^2_W = 5,691$$

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

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

```
# número de estratos da PPV-sudeste
nestrat <- length(unique(ppv1$estratof[ppv1$regiao == 2]))
# número de setores da PPV-sudeste
nsetor <- length(unique(ppv1$nsetor[ppv1$regiao == 2]))
f <- nsetor -nestrat
f1p <- ((f-J+2)/(f*(J-1)))*x2w
f1p
```

```
##          [,1]
## [1,] 1.419005
```

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

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

```
f2p <- x2w/(J-1)
f2p
```

```
##          [,1]
## [1,] 1.435505
```

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

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.

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

Estatística	Tipo	Valor	Distribuição	valor-p
X^2_P	Adequada para IID	11.640	$\chi^2(4)$	0.020
$X^2_P\left(\hat{d}_.\right)$	Ajustes e	4.901	$\chi^2(4)$	0.298
$X^2_P\left(\hat{\delta}_.\right)$	correções da	4.740	$\chi^2(4)$	0.315
$FX^2_P\left(\hat{\delta}_.\right)$	Estatística	1.850	$F\,(4;261)$	0.318
$X^2_P\left(\hat{\delta}_., \hat{a}^2\right)$	X^2_P	3.784	$\chi^2(3,19)$	0.314
X^2_W	Baseadas no	5.691	$\chi^2(4)$	0.223
$F_{1.p}$	plano	1.406	$F\,(4;259)$	0.232
$F_{2.p}$	amostral	1.423	$F\,(4;261)$	0.228

Tabela 7.5: Estimativas das proporções nas classes

idade	mean	SE	deff
0 a 14 anos	0.2845	0.0072	2.286
15 a 29 anos	0.2678	0.0069	2.187
30 a 44 anos	0.2225	0.0066	2.254
45 a 59 anos	0.1316	0.0051	1.991
60 anos e mais	0.0935	0.0055	3.163

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.

7.4 Laboratório de R

Exemplo 7.1 pode ser substituído por: Criar variável ITAB (Não aparece)

```
ppv.des<-svydesign(id=~nsetor,strat=~estratof,weights=~pesof,
data=ppv1,nest=TRUE)
ppv.se.des<-subset(ppv.des,regiao==2)
```

```
ppv.id<-svymean(~idatab,ppv.se.des,deff=T)
vhvat<-vcov(ppv.id)
```

```
library(xtable)
fr_ppv_id<- data.frame(ppv.id)
row.names(fr_ppv_id) <- NULL
fr_ppv_id <- cbind(idade= c("0 a 14 anos","15 a 29 anos", "30 a 44 anos","45 a 59 anos", "60 anos e mais
knitr::kable(fr_ppv_id, booktabs= TRUE, digits=c(0,4,4,3), caption="Estimativas das proporções nas class
```

Estatística de Wald calculada a partir da fórmula (7.8)

```
#Vetor de proporções estimadas
phat<-coefficients(ppv.id)
# Vetor de proporções obtido na Contagem Populacional de 1996
p0<-c(.2842,.2774,.2263,.1261,.086)
# Estatística de Wald
x2_w<-matrix((phat-p0)[-5],nrow=1)%*%solve(vhat[-5,-5])%*%
matrix((phat-p0)[-5],ncol=1)
x2_w
```

```
##          [,1]
## [1,] 5.742022
```

```
#Cálculo do p-valor
round(pchisq(x2_w,4,lower.tail=FALSE),digits=3)
```

```
##       [,1]
```

```
## [1,] 0.219
```

Estatística de Pearson calculada a partir da fórmula 7.11

```
n<-8903
P0<-diag(p0)-matrix(p0,ncol=1)%*%matrix(p0,nrow=1)
x2_p<-n*matrix((phat-p0)[-5],nrow=1)%*%solve(P0[-5,-5])%*%
matrix((phat-p0)[-5],ncol=1)
x2_p
```

```
##          [,1]
## [1,] 11.50719
```

Cálculo do valor-p:

```
round(pchisq(x2_p,4,lower.tail=FALSE),digits=3)
```

```
##       [,1]
## [1,] 0.021
```

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 $\chi^2\left((L-1)(C-1)\right)$. 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 intraclasse 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.

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 2×2, 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 $p_{+c} = \sum_l p_{lc}$.

Tabela 8.1: Tabela 2x2 de proporções.

		var2		
		1	2	Total
Var1	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

8.2.1 Teste de Independência

A hipótese de independência corresponde a

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

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

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

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, $\hat{p}_{l+} = \sum_c \widehat{p}_{lc}$ 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 2×2, queremos testar a hipótese

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

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:

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

onde $n_{l+} = \sum_c n_{lc}$ para $l = 1, 2$ e $\widehat{p}_{lc} = n_{lc}/n_{l+}$ 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 multinomiais 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 $X^2_P(I)$ e $X^2_P(H)$ 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 $X^2_P(I)$ e $X^2_P(H)$ 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 $X^2_P(I)$ e $X^2_P(H)$ 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 2×2. Mas, é importante observar que os efeitos de plano amostral e as correções a serem considerados são distintos para as duas estatísticas $X^2_P(I)$ e $X^2_P(H)$.

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 $X^2_P(I)$, se baseia no EPA médio das estimativas das proporções nas celas $\hat{p}_{lc} = \hat{n}_{lc}/n$, enquanto que para a estatística $X^2_P(H)$ ele se baseia no EPA médio das estimativas das proporções nas linhas $\hat{p}_{lc} = \hat{n}_{lc}/n_{l+}$.

Os valores das estatísticas $X^2_P(I)$ e $X^2_P(H)$ são iguais no caso IID, mas para planos amostrais complexos, as estatísticas corrigidas pelo EPA médio são distintas, apesar de terem, para tabelas 2×2, 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 da proporções nas linhas na tabela de proporções são iguais a 1. A tabela para a população é da forma da Tabela 8.5.

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_{l1}	...	p_{lc}	...	p_{lC}	1
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
L	p_{L1}	p_{L2}	...	p_{Lc}	...	p_{LC}	1
Total	p_{+1}	p_{+2}	...	p_{+c}	...	p_{+C}	1

Note que aqui as proporções que aparecem nas linhas da tabela são proporções calculadas em relação à freqüência 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 $l = 1, \ldots, L$ e $c = 1, \ldots, C$.

Vamos considerar o caso em que $L = 2$ regiões devem ser comparadas. Seja $\mathbf{p}_l = (p_{l1}, \ldots, p_{l\,C-1})'$ 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 $\mathbf{p}_0 = (p_{+1}, \ldots p_{+\,C-1})'$ o vetor comum de proporções sob H_0, desconhecido. Denotemos por $\widehat{\mathbf{p}}_l = (\hat{p}_{l1}, \ldots, \hat{p}_{l\,C-1})'$ os vetores de proporções estimadas ($l = 1, 2$), baseados em amostras independentes para as diferentes regiões, onde $\hat{p}_{lc} = \widehat{N}_{lc}/\widehat{N}_{l+}$ é um estimador consistente da proporção p_{lc} na população correspondente, e $\widehat{N}_{lc}$ e $\widehat{N}_{l+}$ são estimadores ponderados das frequências nas celas e nas marginais de linha da tabela, respectivamente, de modo que $\sum_{c=1}^{C} \widehat{N}_{lc} = \widehat{N}_{l+}$. 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 $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1)$ e $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)$ estimadores consistentes das matrizes de variância de aleatorização dos vetores $\widehat{\mathbf{p}}_1$ e $\widehat{\mathbf{p}}_2$, respectivamente. A estatística de Wald baseada no plano amostral $X^2_W(H)$ para efetuar o teste de homogeneidade no caso de duas regiões ($L = 2$) é dada por

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

pois as amostras são disjuntas e supostas independentes.

No caso, a estatística de Wald $X^2_W(H)$ 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 $X^2_W(H)$ para testar a hipótese de igualdade das pirâmides etárias estimadas pela Pesquisa sobre 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 $\left[\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)\right]$, onde m e H são os números totais de conglomerados e de estratos nas amostras das duas regiões, respectivamente. As correções F da estatística $X^2_W(H)$ são dadas por

$$F_{1.p} = \frac{f-(C-1)+1}{f\,(C-1)} X^2_W\,(H)\,, \tag{8.2}$$

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

$$F_{2.p} = X^2_W\,(H)\,/\,(C-1) \tag{8.3}$$

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

$$X^2_P\,(H) = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)^{'} \left(\hat{\mathbf{P}}/\widehat{n}_{1+} + \hat{\mathbf{P}}/\widehat{n}_{2+}\right)^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2) \quad , \tag{8.4}$$

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

Neste caso, $\hat{\mathbf{P}}/\widehat{n}_{1+}$ é o estimador da matriz de covariância de $\hat{\mathbf{p}}_0$ na primeira região e $\hat{\mathbf{P}}/\widehat{n}_{2+}$ 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 $X^2_P\,(H)$, 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 $X^2_P\,(H)$ é qui-quadrado com $(C-1)$ graus de liberdade.

Para introduzir em $X^2_P\,(H)$ 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

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

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

A matriz estimada de efeito multivariado de plano amostral é

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

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

$$X^2_P\left(H;\hat{d}_{\cdot}\right) = X^2_P\,(H)\,/\hat{d}_{\cdot}, \tag{8.7}$$

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

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

$$X_P^2\left(H;\hat{\delta}_{\cdot}\right) = X_P^2\left(H\right)/\hat{\delta}_{\cdot}, \tag{8.8}$$

onde

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

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

As distribuições de referência, tanto de $X_P^2\left(H;\hat{d}_{\cdot}\right)$ como de $X_P^2\left(H;\hat{\delta}_{\cdot}\right)$, são qui-quadrado com $(C-1)$ graus de liberdade. Estes ajustes corrigem a estatística $X_P^2\left(H\right)$ 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

$$X_P^2\left(H;\hat{\delta}_{\cdot},\hat{a}^2\right) = X_P^2\left(H;\hat{\delta}_{\cdot}\right)/\left(1+\hat{a}^2\right) \tag{8.9}$$

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

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

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

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

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

Quando as estimativas $\hat{\mathbf{V}}_p\left(\widehat{\mathbf{p}}_1\right)$ e $\hat{\mathbf{V}}_p\left(\widehat{\mathbf{p}}_2\right)$ 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

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

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 $p_{lc} = N_{lc}/N$.

Tabela 8.3: Proporções por cela na população.

				Var2			
Var 1	1	2	...	c	...	C	Total
1	p_{11}	p_{12}	...	p_{1c}	...	p_{1C}	p_{1+}
2	p_{21}	p_{22}	...	p_{2c}	...	p_{2C}	p_{2+}
⋮	⋮	⋮	.	⋮	.	⋮	⋮
l	p_{l1}	p_{l2}	...	p_{lc}	...	p_{lC}	p_{l+}
⋮	⋮	⋮	.	⋮	.	⋮	⋮
L	p_{L1}	p_{L2}	...	p_{Lc}	...	p_{LC}	p_{L+}
Total	p_{+1}	p_{+2}	...	p_{+c}	...	p_{+c}	1

Estamos interessados em testar a hipótese de independência

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

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:

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

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

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

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 $\hat{\mathbf{f}} = \left(\hat{f}_{11}, \ldots, \hat{f}_{1\,C-1}, \ldots, \hat{f}_{L-1\,1}, \ldots, \hat{f}_{L-1\,C-1}\right)'$, 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 $X^2_W(I)$, para o teste de independência, tem a forma da expressão (8.8), 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 $\hat{\mathbf{V}}_{\mathbf{f}}$ 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

$$X^2_W(I) = \hat{\mathbf{f}}'\hat{\mathbf{V}}_{\mathbf{f}}^{-1}\hat{\mathbf{f}} \ , \tag{8.10}$$

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

A estimativa $\hat{\mathbf{V}}_{\mathbf{f}}$ 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, $\mathbf{f} = \mathbf{g}(\mathbf{p}) = \mathbf{g}(p_{11}, \ldots, p_{1\,C-1}, \ldots, p_{L-1\,1}, \ldots, p_{L-1\,C-1})$. Assim, a matriz de covariância de $\hat{\mathbf{f}}$ pode ser estimada por

$$\hat{\mathbf{V}}_{\mathbf{f}} = \boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})\,\hat{\mathbf{V}}_p^{-1}\boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})^{'}, \tag{8.11}$$

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

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

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

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

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

$$X_P^2(I) = n\,\hat{\mathbf{f}}'\,\widehat{\mathbf{P}}_{0\mathbf{f}}\,\hat{\mathbf{f}}, \tag{8.12}$$

onde

$$\widehat{\mathbf{P}}_{0\mathbf{f}} = \boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})\,\hat{\mathbf{P}}_0\boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})', \tag{8.13}$$

$$\hat{\mathbf{P}}_0 = diag(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0\hat{\mathbf{p}}_0^{'},$$

$\widehat{\mathbf{P}}_0/n$ 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 $\hat{p}_{l+}\,\hat{p}_{+c}$, e $diag(\hat{\mathbf{p}}_0)$ representa a matriz diagonal com elementos $\hat{p}_{l+}\,\hat{p}_{+c}$ 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 $\hat{d}_{\cdot}$ dos efeitos univariados de plano amostral:

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

onde $\hat{d}_{\cdot}=\sum_{c=1}^{C}\sum_{l=1}^{L}\hat{d}_{lc}/(LC)$ 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

$$\hat{\mathbf{\Delta}}=n\,\hat{\mathbf{P}}_{0\mathbf{f}}^{-1}\,\hat{\mathbf{V}}_{\mathbf{f}}, \tag{8.14}$$

onde $\hat{\mathbf{V}}_{\mathbf{f}}$ foi definido em (8.11) e $\hat{\mathbf{P}}_{0\mathbf{f}}$ definido em (8.13).

O ajuste de Rao-Scott de primeira ordem para $X_P^2(I)$ é dado por

$$X_P^2\left(I;\hat{\delta}_{\cdot}\right)=X_P^2(I)/\hat{\delta}_{\cdot}, \tag{8.15}$$

onde $\hat{\delta}_{\cdot}$ é um estimador da média $\bar{\delta}$ dos autovalores desconhecidos da matriz $\mathbf{\Delta}$ 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

$$\begin{aligned}\hat{\delta}_{\cdot} &= \frac{1}{(L-1)(C-1)}\sum_{l=1}^{L}\sum_{c=1}^{C}\frac{\hat{p}_{lc}(1-\hat{p}_{lc})}{\hat{p}_{l+}\hat{p}_{+c}}\hat{d}_{lc}\\ &\quad-\sum_{l=1}^{L}(1-\hat{p}_{l+})\,\hat{d}_{l+}-\sum_{c=1}^{C}(1-\hat{p}_{+c})\,\hat{d}_{+c},\end{aligned}$$

sem precisar calcular a matriz de efeitos multivariados de plano amostral. A distribuição assintótica de $X_P^2\left(I;\hat{\delta}_{\cdot}\right)$, sob H_0, é qui-quadrado com $(L-1)\times(C-1)$ graus de liberdade.

O ajuste de Rao-Scott de segunda ordem é definido por

$$X_P^2\left(I;\hat{\delta}_{\cdot};\hat{a}^2\right)=X_P^2(I)/\left(\hat{\delta}_{\cdot}\left(1+\hat{a}^2\right)\right),$$

Tabela 8.4: Frequências Amostrais por celas na PNADRJ90

	1	2	3	Sum
1	476	2527	1273	4276
2	539	1270	422	2231
Sum	1015	3797	1695	6507

onde $\hat{\delta}_{\cdot}$ é um estimador da média dos autovalores de $\hat{\mathbf{\Delta}}$, dado por

$$\hat{\delta}_{\cdot} = \frac{tr\left(\hat{\mathbf{\Delta}}\right)}{(L-1)(C-1)}$$

e $\hat{a}^2$ é um estimador do quadrado do coeficiente de variação dos autovalores desconhecidos de $\mathbf{\Delta}$, δ_k, $k = 1, \ldots, (L-1)(C-1)$, dado por

$$\hat{a}^2 = \sum_{k=1}^{(L-1)(C-1)} \hat{\delta}_k^2 / \left((L-1)(C-1)\hat{\delta}_{\cdot}^2\right) - 1.$$

Um estimador da soma dos quadrados dos autovalores é

$$\sum_{k=1}^{(L-1)(C-1)} \hat{\delta}_k^2 = tr\left(\hat{\mathbf{\Delta}}^2\right).$$

A estatística $X_P^2\left(I; \hat{\delta}_{\cdot}; \hat{a}^2\right)$ é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite $gl_S = (L-1)(C-1)/\left(1+\hat{a}^2\right)$.

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

$$FX_P^2\left(\hat{\delta}_{\cdot}\right) = X_P^2\left(\hat{\delta}_{\cdot}\right)/(L-1)(C-1) \quad . \tag{8.16}$$

A estatística (8.16) tem distribuição de referência F com $(L-1) \times (C-1)$ 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, 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.

No teste de homogeneidade das distribuições de renda, consideramos fixadas as marginais `4276` e `2231` da variável Sexo na tabela de frequências amostrais. Usando a library survey, 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.

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, para cada sexo,

Tabela 8.5: Proporções nas linha, desvios padrões e EPAs de re para cada nivel de sx

Est	Sexo	r1	r2	r3
prop_lin	1	0.111	0.591	0.298
SE_prop_lin	1	57.269	102.576	111.213
Def_prop_lin	1	1.422	1.863	2.531
prop_lin	2	0.240	0.570	0.190
SE_prop_lin	2	125.026	119.375	111.410
Def_prop_lin	2	1.911	1.298	1.802

Tabela 8.6: Proporções nas linha, desvios padrões e EPAs de `re` na população

	mean	SE	deff
re1	0.155	68.977	2.361
re2	0.584	82.001	1.803
re3	0.261	96.130	3.122

as estimativas: da proporção na linha, do desvio-padrão da estimativa da proporção na linha (×10.000), e do efeito de plano amostral da estimativa de proporção na linha.

As mesmas estimativas para a tabela marginal de `Rendas` são dadas por:

```
marg_re_pop <-as.data.frame(svymean(~re,pnad.des, deff=TRUE ))
marg_re_pop<- transform(marg_re_pop, SE=10000*SE )
knitr::kable(marg_re_pop ,booktabs=TRUE, digits=3,
  caption="Proporções nas linha, desvios padrões e EPAs de `re` na população")
```

Vamos calcular, a título de ilustração, o efeito do plano amostral da estimativa na cela (1,1) da Tabela 8.5, apresentado na cela (1,7) dessa tabela.

```
# cria dummy para re=1
pnad.des <- update(pnad.des,
  ind_rend1= as.numeric(re==1))
# proporção de re=1 quando  sx= 1:
prop_re1_des <- svymean(~ind_rend1, subset(pnad.des, sx==1))
# objeto de desenho AAS
pnad.des_AAS<-svydesign(id=~1,data=pnadrj90)
# dummy para re=1
pnad.des_AAS <- update(pnad.des_AAS,
  ind_rend1= as.numeric(re==1))
# proporção de re=1 quando sx=1
prop_re1_AAS <- svymean(~ind_rend1, subset(pnad.des_AAS, sx==1))
round(attr(prop_re1_des,"var")/attr(prop_re1_AAS,"var"),2)
```

```
##           ind_rend1
## ind_rend1      1.42
```

A estimativa do efeito médio de plano amostral para corrigir a estatística $X^2_P\,(H)$ é $\hat{d}_{\cdot} = 1,802$, 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 8.7 contém, em cada cela, as estimativas: da proporção na cela, do desvio-padrão da estimativa da proporção na cela (×10.000), e do efeito de plano amostral da estimativa de proporção na cela.

Tabela 8.7: Proporções nas cela, desvios padrões e EPAs

Est	Sexo	r1	r2	r3
prop_lin	1	0.073	0.388	0.196
SE_prop_lin	1	38.343	80.435	71.772
Def_prop_lin	1	1.416	1.775	2.131
prop_lin	2	0.082	0.195	0.065
SE_prop_lin	2	44.401	51.582	40.219
Def_prop_lin	2	1.698	1.103	1.731

Tabela 8.8: Proporções nas linha, desvios padrões e EPAs de sx na população

	mean	SE	deff
sx1	0.657	55.814	0.901
sx2	0.343	55.814	0.901

Tabela de proporções de `sx` para a população inteira:

```
marg_sx_pop <-data.frame(svymean(~sx,pnad.des, deff=TRUE ))
marg_sx_pop <- transform(marg_sx_pop, SE = 10000*SE)
knitr::kable(marg_sx_pop, digits=3,
  caption="Proporções nas linha, desvios padrões e EPAs de sx na população")
```

Vamos calcular, a título de ilustração, o efeito de plano amostral na cela (1,1) da Tabela 8.7 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: $0,073 \times (1-0,073)/6.507$. 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 $\hat{d}_{\cdot} = 1,640$, calculada tomando a média dos EPAs das celas correspondentes aos níveis 1 e 2 da variável `sx`.

Calculando as estatísticas $X_P^2(I)$ e $X_P^2(H)$ para os testes clássicos de independência e homogeneidade a partir da Tabela 8.7, 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 `sx` 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 $X_P^2\left(I;\hat{d}_{\cdot}\right) = 137,117$ e $X_P^2\left(H;\hat{d}_{\cdot}\right) = 124,742$, 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 8.7, 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.

8.4 Laboratório de R

Vamos reproduzir alguns resultados usando dados da PNAD90 para o Rio de Janeiro, descritos no Exemplo 6.1.

Exemplo 8.2. Estimativas de medidas descritivas em tabelas

```
library(survey)
library(anamco) #carrega dados
names(pnadrj90)
```

```
##  [1] "stra"     "psu"      "pesopes"  "informal" "sx"       "id"
##  [7] "ae"       "ht"       "re"       "um"
```

```
n <- nrow (pnadrj90)
```

- Transformação em fatores:

```
unlist(lapply(pnadrj90, mode))
```

```
##      stra       psu   pesopes  informal        sx        id        ae
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        ht        re        um
## "numeric" "numeric" "numeric"
```

```
pnadrj90<-transform(pnadrj90,sx=factor(sx),id=factor(id),ae=factor(ae),ht=factor(ht),re=factor(re))
```

Definição do objeto de desenho:

```
pnad.des<-svydesign(id=~psu,strata=~stra,weights=~pesopes,data=pnadrj90,nest=TRUE)
```

- Estimativas de proporções:

```
svymean(~sx,pnad.des)          #estimativa de proporção para sx
```

```
##        mean     SE
## sx1 0.65708 0.0056
## sx2 0.34292 0.0056
```

```
svymean(~re,pnad.des)          #estimativa de proporção para re
```

```
##        mean     SE
## re1 0.15546 0.0069
## re2 0.58356 0.0082
## re3 0.26098 0.0096
```

```
svymean(~ae,pnad.des)          #estimativa de proporção para ae
```

```
##        mean     SE
## ae1 0.31304 0.0095
## ae2 0.31972 0.0071
## ae3 0.36725 0.0105
```

```
ht.mean<-svymean(~ht,pnad.des)
```

- Exemplos de funções extratoras e atributos:

```
coef(ht.mean)                          #estimativas das proporções
```

```
##       ht1       ht2       ht3
## 0.2103714 0.6148881 0.1747405
```

```
attributes(ht.mean)                          #ver atributos
```

```
## $names
## [1] "ht1" "ht2" "ht3"
##
## $var
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
##
## $statistic
## [1] "mean"
##
## $class
## [1] "svystat"
```

```
vcov(ht.mean)                          #estimativas de variâncias e covariâncias
```

```
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
```

```
attr(ht.mean, "var")
```

```
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
```

- Estimativas de proporções nas classes de renda por domínios definidos pela variável `sx`:

Proporções por sexo

sx

re1

re2

re3

se.re1

se.re2

se.re3

1

0.11

0.59

0.30

0.01

0.01

0.01

2

0.24

0.57

0.19

0.01

0.01

0.01

- Estimativas das proporções nas classes de renda e a tabela cruzada das variáveis sexo e renda:

```
svymean(~re,pnad.des,deff=T)
```

```
##          mean        SE   DEff
## re1 0.1554555 0.0068977 2.3611
## re2 0.5835630 0.0082001 1.8027
## re3 0.2609815 0.0096130 3.1216
```

```
round(svytable(~sx+re,pnad.des,Ntotal=1),digits=3)
```

```
##    re
## sx     1     2     3
##   1 0.073 0.388 0.196
##   2 0.082 0.195 0.065
```

```
svyby(~re,~sx,pnad.des,svymean,keep.var=T)
```

```
##   sx       re1       re2       re3      se.re1     se.re2     se.re3
## 1  1 0.1110831 0.5908215 0.2980955 0.005726888 0.01025759 0.01112131
## 2  2 0.2404788 0.5696548 0.1898663 0.012502636 0.01193753 0.01114102
```

```
svymean(~I((sx==1&re==1)*1), pnad.des, deff=T)
```

```
##                                  mean        SE   DEff
## I((sx == 1 & re == 1) * 1) 0.0729904 0.0038343 1.4156
```

```
#proporções nas celas
svytable(~sx+re,pnad.des,Ntotal=1)
```

```
##    re
## sx           1          2          3
##   1 0.07299044 0.38821684 0.19587250
##   2 0.08246505 0.19534616 0.06510900
```

```
# porcentagens nas celas
svytable(~sx+re,pnad.des,Ntotal=100)
```

```
##    re
## sx          1         2         3
##   1  7.299044 38.821684 19.587250
##   2  8.246505 19.534616  6.510900
```

```
# produz se e deff
sx.re_mean <- data.frame(svymean(~interaction(sx,re), pnad.des, deff=T))
```

```
# média de epas para correção de testes
mean(sx.re_mean$deff)
```

```
## [1] 1.642161
```

Exemplo 8.3. Testes de Hipóteses

- Teste de independência e homogeneidade baseado nos dados da amostra sem considerar o plano amostral:

```
attach(pnadrj90)
tab.amo <- table(sx,re)
chisq.test(tab.amo)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab.amo
## X-squared = 227.03, df = 2, p-value < 2.2e-16
```

Observação. Identificar resultados dos testes obtidos pela library `survey` (Lumley, 2017) com as fórmulas do texto:

```
n <- nrow (pnadrj90)
pearson <- chisq.test( pnadrj90$sx, pnadrj90$re, correct = FALSE )

# teste Chi-quadrado para ponderado pelo pesos
pearsonPond <- chisq.test(svytable(~ sx+re , pnad.des, Ntotal = n), correct = FALSE)


# teste Chi-quadrado de Pearson com ajuste de Rao-Scott
pearsonF <-svychisq( ~ sx+re , pnad.des, statistic = "F", na.rm=TRUE)
# teste Chi-quadrado de Pearson com ajuste de Rao-Scott
pearsonChisq <-svychisq( ~ sx+re , pnad.des, statistic = "Chisq", na.rm=TRUE)
# teste de Wald baseado no desenho amostral
pearsonWald <- svychisq( ~ sx+re , pnad.des, statistic = "Wald", na.rm=TRUE)
# teste de Wald com ajuste
pearsonAdjWald <-svychisq( ~ sx+re , pnad.des, statistic = "adjWald", na.rm=TRUE)
# teste Chi-quadrado de Pearson: distribuição assintótica exata
```

```
result <- data.frame(
Metodo = c("AASR", "AASRPOND", "RAO-SCOTT", "RAO.SCOTT.F", "WALD","ADJWALD" ), Estatistica = c(pearson$
Valorp = c(pearson$p.value, pearsonPond$p.value, pearsonChisq$p.value, pearsonF$p.value, pearsonWald$p.

)
knitr::kable(result,digits= c(0,3, 5),booktabs=TRUE )
```

Metodo	Estatistica	Valorp
AASR	227.025	0
AASRPOND	224.848	0
RAO-SCOTT	224.848	0
RAO.SCOTT.F	108.337	0
WALD	68.410	0
ADJWALD	68.304	0

Capítulo 9

Estimação de densidades

9.1 Introdução

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

Capítulo 10

Modelos Hierárquicos

10.1 Introdução

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

Capítulo 11

Não-Resposta

11.1 Introdução

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

Capítulo 12

Diagnóstico de ajuste de modelo

12.1 Introdução

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

Capítulo 13

Agregação vs. Desagregação

13.1 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 análise desagregada, em contraposição aos procedimentos indicados nos capítulos anteriores. Para um exame mais detalhado do tema, o leitor deve consultar cap. 10 a 13 de (Skinner, 989a) e (Bryk and Raudenbush, 1992).

13.2 Modelagem da Estrutura Populacional

Para introduzir a abordagem de análise desagregada, vamos considerar um modelo simples de regressão linear, definido por

$$E_M\left(Y_i \,|X_i = x_i\,\right) = \beta_0 + \beta_1 x_i \tag{13.1}$$

onde β_0 e β_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 β_0 e β_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

$$E_M\left(Y_i\,|X_i = x_i\,, Z_i = z_i\,\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \tag{13.2}$$

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 β_2 e β_3 .

Se $\beta_3 = 0$, o efeito do estrato é modificar o intercepto de β_0 para $\beta_0 + \beta_1$, 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 β_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 (13.1) em favor de um modelo desagregado da forma (13.2), 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 (13.1) de maneira simples, bastando para isso considerar um modelo da forma

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \tag{13.3}$$

no qual j denota conglomerado e i denota indivíduo no conglomerado.

Em dados de pesquisas amostrais, os erros ε_{ij} não satisfazem, em geral, a hipótese de IID. Além disso, no modelo (13.3), β_0 e β_1 não variam para os diferentes conglomerados. Pode ser adequado supor que β_0 e β_1 variam entre conglomerados. Isto pode ser obtido substituindo β_0 e β_1 em (13.3) por coeficientes aleatórios, que dependem dos conglomerados, isto é, adotando-se o modelo

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \\ \beta_{0j} = \beta_0 + \eta_{0j} \\ \beta_{1j} = \beta_1 + \eta_{1j} \end{cases} \tag{13.4}$$

com β_0 e β_1 fixos e desconhecidos e ε_{ij}, η_{0j} e η_{1j} variáveis aleatórias, satisfazendo

$$\begin{aligned} E_M(\varepsilon_{ij}) &= E_M(\eta_{0j}) = E_M(\eta_{1j}) = 0 \\ V_M(\varepsilon_{ij}) &= \sigma^2,\ V_M(\eta_{0j}) = \sigma_0^2,\ V_M(\eta_{1j}) = \sigma_1^2\ , \\ COV_M\left(\varepsilon_{ij}, \eta_{0j'}\right) &= COV_M\left(\varepsilon_{ij}, \eta_{1j'}\right) = 0\ , \\ COV_M\left(\varepsilon_{ij}, \varepsilon_{i'j'}\right) &= 0\ , \quad j \neq j^{'} \text{ ou } i \neq i^{'}\ , \end{aligned} \tag{13.5}$$

e

$$COV_M\left(\eta_{0j}, \eta_{1j'}\right) = \begin{cases} \sigma_{01} & j = j^{'} \\ 0 & j \neq j^{'} \end{cases}\ . \tag{13.6}$$

Podemos juntar as expressões em (13.4) e reescrever o modelo como

$$\begin{aligned} Y_{ij} &= (\beta_0 + \eta_{0j}) + (\beta_1 + \eta_{1j})\, x_{ij} + \varepsilon_{ij} \\ &= \beta_0 + \beta_1 x_{ij} + \eta_{0j} + \eta_{1j} x_{ij} + \varepsilon_{ij}\ . \end{aligned} \tag{13.7}$$

Em (13.7), os coeficientes β_0 e β_1 são fixos e os coeficientes η_{0j} e η_{1c} são aleatórios, sendo o modelo denominado de efeitos mistos: fixos e aleatórios (veja por exemplo (Longford, 1993), (Diggle et al., 1994) e (Bryk and Raudenbush, 1992).

Em (13.5) e (13.6) os valores de σ_0^2, σ_1^2, σ_{01} e σ^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 σ_0^2, σ_1^2 e talvez reduzir a covariância σ_{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

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01} a_j + \eta_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} a_j + \eta_{1j} \end{cases}\ . \tag{13.8}$$

Mais uma vez o objetivo básico da inferência se altera, pois agora está centralizado nos parâmetros $\left(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}, \sigma_0^2, \sigma_1^2, \sigma^2, \sigma_{01}\right)$, com intervalos de confiança e testes de hipóteses relativos a estes

parâmetros. O modelo (13.8) é de efeitos mistos, com efeitos fixos $(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$ e efeitos aleatórios (η_{0j}, η_{1j}).

Modelos de efeitos mistos da forma (13.8) 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 (13.2); 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 (13.8) 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.

13.3 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, 989a), 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:

$$ESC_{ij} = \beta_{0j} + \beta_{1j} SEX_{ij} + \beta_{2j} CSA_{ij} + \varepsilon_{ij}, \tag{13.9}$$

onde $i = 1, \ldots, n_j$ denota o aluno dentro da turma e $j = 1, \ldots, J$ denota a turma.

é possível que os coeficientes β_{0j}, β_{1j} e β_{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):

$$\left\{\begin{array}{l} \beta_{0j} = \gamma_{00} + \gamma_{01} CST_j + \gamma_{02} EXP_j + \eta_{0j} \ , \\ \beta_{1j} = \gamma_{10} + \gamma_{11} CST_j + \gamma_{12} EXP_j + \eta_{1j} \ , \\ \beta_{2j} = \gamma_{20} + \gamma_{21} CST_j + \gamma_{22} EXP_j + \eta_{2j} \ , \end{array}\right. \tag{13.10}$$

onde η_{0j}, η_{1j} e η_{2j} são erros no nível 2 satisfazendo as condições em (13.5). As equações (13.9) e (13.10) definem um modelo hierárquico, que pode ser escrito de forma equivalente como

$$\begin{array}{rcl} ESC_{ij} & = & \gamma_{00} + \gamma_{01} CST_j + \gamma_{02} EXP_j \\ & & + \left(\gamma_{10} + \gamma_{11} CST_j + \gamma_{12} EXP_j\right) SEX_{ij} \\ & & + \left(\gamma_{20} + \gamma_{21} CST_j + \gamma_{22} EXP_j\right) CSA_{ij} \\ & & + \eta_{0j} + \eta_{1j} SEX_{ij} + \eta_{2j} CSA_{ij} + \varepsilon_{ij} \quad . \end{array} \tag{13.11}$$

A presença dos erros aleatórios η_{0j}, η_{1j} e η_{2j} (de nível 2), torna (13.11) um modelo misto. Se os erros fossem suprimidos em (13.10), 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 (13.10) 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:

$$\begin{array}{rcl} ESC_{ij} & = & \gamma_{00} + \gamma_{01} CST_j + \gamma_{02} EXP_j \\ & & + \beta_{1j} SEX_{ij} + \beta_{2j} CSA_{ij} + \left(\eta_{0j} + \varepsilon_{ij}\right) \quad . \end{array} \tag{13.12}$$

Além da parte fixa, o modelo contém uma `parte residual`, om os erros aleatórios com média zero: η_{0j}, que representa o desvio da média dos indivíduos da turma j com relação à média total, e ε_{ij}, que é o desvio do i-ésimo aluno com relação à média da turma j. Vamos supor, ainda, que os η_{0j} e os ε_{ij} são independentes entre si e os η_{0j} são independentes dos ε_{ij} , com

$$E_M\left(\eta_{0j}\right) = E_M\left(\varepsilon_{ij}\right) = 0, \ V_M\left(\eta_{0j}\right) = \sigma_0^2, \ V_M\left(\varepsilon_{ij}\right) = \sigma^2, \ \forall i, j \ . \tag{13.13}$$

A parte aleatória do modelo (13.12) é o termo $\eta_{0j} + \varepsilon_{ij}$, com distribuição tendo parâmetros σ_0^2 e σ^2 a serem estimados.

O modelo (13.12) 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 β_{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 (13.11), 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

$$COV_M\left(ESC_{ij}, ESC_{ij'}\right) = COV_M\left[\left(\eta_{0j} + \varepsilon_{ij}\right), \left(\eta_{0j} + \varepsilon_{ij'}\right)\right] = \sigma_0^2, \tag{13.14}$$

pois supusemos que ε_{ij} e $\varepsilon_{ij'}$ são independentes. Por outro lado, condicionando na parte fixa do modelo, tem-se que

$$V_M\left(ESC_{ij}\right) = V_M\left(ESC_{ij'}\right) = \sigma_0^2 + \sigma^2,$$

e portanto

$$CORR_M\left(ESC_{ij}, ESC_{ij'}\right) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}. \tag{13.15}$$

A expressão (13.15) define a `correlação intraclasse` 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 (13.11), 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 and 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., 1998b) apresentam uma forma de incorporar pesos no ajuste de modelos hierárquicos para compensar diferentes probabilidades de inclusão das unidades na amostra.

Exemplo 13.1. Plano amostral de pesquisa educacional ((Lehtonen and 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

Tabela 13.1: Análise do modelo (9.16) sob hipótese de observações IID, ignorando complexidades do plano amostral

Variável	Coeficiente	Desvio_Padrão	Teste_t	pvalor	EPA
Intercepto	2.912	0.427	6.82	0.000	1
TEMP	-0.894	0.174	-5.14	0.000	1
EXP	0.254	0.127	2.00	0.045	1

Tabela 13.2: Análise agregada do modelo (9.16), usando MPV e estatísticas de Wald (PROC LOGISTIC do SUDAAN)

Variável	Coeficiente	Desvio_Padrão	Teste_t	pvalor	EPA
Intercepto	2.899	0.578	5.02	0.00	1.83
TEMP	-0.906	0.211	-4.29	0.00	1.47
EXP	0.271	0.181	1.50	0.14	2.03

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 (TEMP)_j + \beta_2 \left(EXP\right)_j, \tag{13.16}$$

no qual β_0, β_1 e β_2 são coeficientes a serem estimados, e o índice j se refere a um domínio de estudo, $j = 1, \ldots, 9$ (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 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 13.1.

Os *p*valores da Tabela 13.1 indicam que os coeficientes são significantemente 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 13.3.

A coluna de *p*valores da Tabela 13.3 indica que o coeficiente de $TEMP$ é significantemente 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 9×9 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 13.2, conforme se pode verificar consultando os valores da Tabela 13.3. 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;

Tabela 13.3: Análise agregada do modelo (9.16), usando estatísticas de Wald corrigidas (programa PC CARP)

Variável	Coeficiente	Desvio_Padrão	Teste_t	pvalor	EPA
Intercepto	2.899	0.597	4.86	0.00	1.95
TEMP	-0.906	0.219	-4.14	0.00	1.58
EXP	0.271	0.186	1.46	0.15	2.14

Tabela 13.4: 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
TEMP	-0.927	0.179	-5.18	0.00	1.06
EXP	0.254	0.188	1.36	0.19	2.19

- 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 (13.16) para incluir o efeito da turma empregando

$$\log\left(\frac{p_{jk}}{1-p_{jk}}\right) = \beta_0 + \beta_1 TEMP_{jk} + \beta_2 EXP_j + u_k. \tag{13.17}$$

O erro aleatório $u_k \sim N\left(0, \sigma_u^2\right)$ representa a variação aleatória no nível 2. A variação no nível 1, entre alunos, é introduzida da seguinte forma. Sob a hipótese binomial, em cada domínio, a proporção de alunos atingindo o nível adequado tem variância $p_{jk}\left(1-p_{jk}\right)/n_{jk}$. Vamos supor que a variação residual no nível 1 é denotada por σ_e^2, e também que a variabilidade entre alunos acarreta uma variação extra-binomial

$$\frac{p_{jk}\left(1-p_{jk}\right)\sigma_e^2}{n_{jk}} \quad . \tag{13.18}$$

Os resultados desta análise são apresentados na Tabela 13.4.

A coluna de $p$valores da Tabela 13.4 indica novamente que o coeficiente de EXP não é significantemente 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 significante 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 13.4, 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 13.2. ((Bryk and 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

$$y_{ij} = \text{Eficiência do Professor;}$$

- Variáveis preditoras, relativas à Escola (unidade de nível 2)

$$x_{1j} = \text{Experiência Acadêmica Média antes da Escola Secundária;}$$

$$x_{2j} = \text{Status Sócio-Econômico Médio;}$$

$$x_{3j} = \text{Proporção Alta de Minorias;}$$

$$x_{4j} = \text{Tamanho;}$$

$$x_{5j} = \text{Mistura étnica;}$$

$$x_{6j} = \text{Mistura de Status Sócio-Econômico;}$$

$$x_{7j} = \text{Grau de Organização Comunitária } (Comunit).$$

A ideia 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:

$$\hat{\sigma}^2 = 0,915;\ \hat{\tau}_{00} = 0,084 \text{ e}$$

$$\text{Correlação Intra-Escola} = \frac{\hat{\tau}_{00}}{\hat{\sigma}^2 + \hat{\tau}_{00}} = 0,092.$$

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

Tabela 13.5: Efeitos da escola na eficácia do professor (Modelo II)

Coeficiente	Estimativa	Desvio_Padrão
gamma_01	0.044	0.020
gamma_02	0.133	0.023
gamma_03	0.031	0.046
gamma_04	-0.066	0.027
gamma_05	-0.014	0.019
gamma_06	-0.028	0.023
sigma_2	0.915	NA
tau_00	0.055	NA

modelo, vamos introduzir no modelo algumas variáveis explicativas referentes à escola, a saber as variáveis $(x_1, \ldots, x_6)$.

Modelo II: Hierárquico com dois níveis, usando as variáveis $(x_1, \ldots, x_6)$, 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 13.5 apresenta as estimativas dos parâmetros para este modelo.

Com essas estimativas, a proporção da variação total entre escolas do parâmetro β_{0j} (nível médio da eficácia dos professores por escola) explicada pelas variáveis $(x_1, \ldots, x_6)$ aumentou para

$$\frac{0,084 - 0,055}{0,084} = 35.$$

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 $(x_1, \ldots, x_7)$, 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 13.6 apresenta as estimativas dos coeficientes do Modelo III e seus respectivos desvios padrões.

Tabela 13.6: Efeitos da escola na eficácia do professor (Modelo III)

Coeficiente	Estimativa	Desvio_Padrão
gamma_01	0.038	0.017
gamma_02	0.015	0.022
gamma_03	-0.055	0.040
gamma_04	0.061	0.026
gamma_05	-0.014	0.016
gamma_06	-0.001	0.020
gamma_07	0.504	0.045
sigma_2	0.915	NA
tau_00	0.031	NA

Tabela 13.7: Efeitos da escola na eficácia do professor (Modelo IV)

Coeficiente	Estimativa	Desvio_Padrão
gamma_01	0.040	0.013
gamma_02	0.015	0.017
gamma_03	-0.056	0.031
gamma_04	0.062	0.021
gamma_05	-0.014	0.013
gamma_06	-0.002	0.016
gamma_07	0.507	0.035

A proporção da variação total entre escolas do parâmetro β_{0j} (nível médio da eficácia dos professores por escola) explicada pelas variáveis $(x_1, \ldots, x_7)$ 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 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 $(x_{1j}, \ldots, x_{7j})$ 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 13.7 apresenta as estimativas de Mínimos Quadrados Ponderados de $(\gamma_0, \gamma_1, \ldots \gamma_7)$ 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 13.6 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 $\left(\tau_{00}+\sigma^2\right)$ no denominador. No modelo hierárquico correspondente (Modelo III) este denominador é τ_{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 $y_{ij}=\beta_{0j}+\mathbf{x}_j^T\beta+v_{ij}$, onde os coeficientes de regressão β 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 β também são aleatórios. O Modelo V a seguir é desse tipo.

Modelo V: Modelo de Coeficientes Aleatórios

$$Y_{ij}=\mathbf{x}_{ij}^T\beta_j+v_{ij},$$

onde

$$\beta_j=\mathbf{x}_j^T\gamma+\delta_j,$$

com

$\delta_j(Q\times 1)$ e $E_M\left(\delta_j\right)=\mathbf{0}$, $E_M\left(v_{ij}\right)=0$, $V_M\left(v_{ij}\right)=\sigma_1^2$, $V_M\left(\delta_j\right)=\mathbf{\Delta}$, $\mathbf{\Delta}$ 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.

13.4 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 and 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 arregimentados 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 mai-o-ria 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., 1998b).

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, 989a), pág. 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 sensitivas individuais, e requer o uso de técnicas apropriadas.

Capítulo 14

Pacotes para Analisar Dados Amostrais

14.1 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.

14.2 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 prático 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.

A seguir transcrevemos do artigo (Lepkowski and 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

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

internet: 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 commandos **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

internet: 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

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 utlizado 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 variaç% ão, 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

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 freqüências 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

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, propor% ções 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 varia% ção, dos efeitos de plano amostral e tamanhos de amostras considerados nos cálculos (freqüências não expandidas), bem como estimativas 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

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 freqüências não expandidas, proporções e médias expandidas, desvios padrões, limites de confiança de 95% e efeitos de plano amostral.

Library survey do R

Estimativas e suas precisões podem ser obtidas por meio da library `survey` do R, (Lumley, 2017). As funções da library `survey` produzem estimativas que incorporam as características do plano amostral utilizado na coleta dos dados.

A library `survey` contém funções para estimar:

- Médias (svymean);
- Totais (svytotal);
- Razões (svyratio);
- Quantis (svyquantile);
- Tabelas de contingência (svytable);
- Modelos lineares generalizados (svyglm)
- Curvas de sobrevivência (svycoxph);
- Testes de postos (svyranktest).

Para a amostra inteira e para domínios.

As variâncias podem ser obtidas por linearização de Taylor or por pesos replicados (BRR, jackknife, bootstrap, multistage bootstrap, ou fornecido pelo usuário).

Mais detalhes estão no site da library survey

Referências Bibliográficas

Albieri, S. and Bianchini, Z. M. (1997). Aspectos de amostragem relativos à pesquisa domiciliar sobre padrões de vida. Technical report, IBGE, Departamento de Metodologia, Rio de Janeiro.

Bickel, P. J. and Doksum, K. A. (1977). Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, San Francisco.

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

Binder, D. A., Kovar, J. G., Kumar, S., Paton, D., and Baaren, A. V. (1987). Analytic uses of survey data: a review. In MacNeil, I. B. and Umphrey, G. J., editors, Applied Probability, Stochastic Processes and Sampling Theory, pages 243--264. John Wiley.

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

Brewer, K. W. R. (1979). A class of robust sampling designs for large scale surveys. Journal of the American Statistical Association, 74:911--915.

Bryk, A. S. and Raudenbush, S. W. (1992). Hierarquical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park.

Carlson, B. L. (1998). Software for statistical analysis of sample survey data. In Encyclopaedia of Biostatistics. John Wiley.

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

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

Chambers, R. L. (1986). Design-adjusted parameter estimation. Journal of the Royal Statistical Society, 149:161--173.

Chambers, R. L. (1995). Regression analysis with sample survey data. Manuscrito inédito cedido por cortesia do autor, 30:70--87.

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

Cox, D. R. and Hinkley, D. V. (1974). Theoretical Statistics. Chapman & Hall, Londres.

Damico, A. J. (2016). lodown: locally download and prepare publicly-available microdata. R package version 0.1.0.

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

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). Analysis of Longitudinal Data. Clarendon Press, Oxford.

Fellegi, I. P. (1980). Approximate tests of independence and goodness-of-fit based on stratified multistage samples. Journal of the American Statistical Association, 75:261--268.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankhyā C, 37:117--132.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. Survey Methodology, 10:97--118.

Garthwaite, P. H., Jollife, I. T., and Jones, B. (1995). Statistical Inference. Prentice Hall, Nova Iorque.

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

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). Sample Survey Methods and Theory. John Wiley and Sons, Nova Iorque.

Hájek, J. (1960). Limiting distributions in simple random sampling from finite populations. Pub.Math. Inst. Hung. Acad. Sci., 5:361--374.

Holt, D., Scott, A., and Ewings, P. D. (1980a). Chi-squared tests with survey data. Journal of the Royal Statistical Society A, 143:303--320.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980b). Regression analysis of data from complex surveys. Journal of the Royal Statistical Society, A, 143:474--487.

IBGE (1981). Metodologia da Pesquisa Nacional por Amostra de Domicílios na Década de 70. Série Relatórios Metodológicos 1, IBGE, Rio de Janeiro.

IBGE (1985). Amostra de Uso Público do Censo Demográfico de 1980 - Metodologia e Manual do Usuário. Technical report, IBGE, Rio de Janeiro.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77:89--96.

Johnson, R. A. and Wichern, D. W. (1988). Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, New Jersey.

Kalton, G. (1983a). Compensating for missing survey data. Technical report, The University of Michigan, Institute for Social Research, Survey Research Center, Ann Arbor, Michigan.

Kalton, G. (1983b). Models in the practice of survey sampling. International Statistical Review, 51:175--188.

Kish, L. (1965). Survey Sampling. John Wiley and Sons, Nova Iorque.

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

Leote, R. M. D. (1996). Um perfil sócio-econômico das pessoas ocupadas no setor informal na área urbana do Rio de Janeiro. Technical Report 2, IBGE, Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Lepkowski, J. and Bowles, J. (1996). Sampling error software for personal computers. The Survey Statistician, 35:10--17.

Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with missing data. John Wiley and Sons, Nova Iorque.

Longford, N. (1993). Random Coefficient Models. Clarendon Press, Oxford.

Lumley, T. (2017). survey: Analysis of Complex Survey Samples. R package version 3.32-1.

Mahalanobis, P. C. (1939). A sample survey of the acreage under jute in bengal. Sankhya, 4:511--531.

Mahalanobis, P. C. (1944). On large-scale sample surveys. Philosophical Transactions of the Royal Society of London B, 231:329--451.

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

Nascimento Silva, P. L. D. (1996). Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys. PhD thesis, University of Southampton, Department of Social Statistics.

Nascimento Silva, P. L. D. and Moura, F. A. S. (1990). Efeitos de conglomeração da malha setorial do censo demográfico 80. Série Textos para Discussão 32, IBGE, Diretoria de Pesquisas, Rio de Janeiro.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. Journal of the Royal Statistical Society B, 42:377--386.

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

of Labor Statistics, U. B. (1984). Bls handbook of methods - volume ii - the consumer price index. Bls bulletin 2134-2, Washington DC.

Pessoa, D. G. C., Nascimento Silva, P. L. D., and Duarte, R. P. N. (1997). Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. Revista Brasileira de Estatística, 33:44--57.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. International Statistical Review, 61:317--337.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability survey. Statistica Sinica, 8:1087--1114.

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from complex samples. Journal of the American Statistical Association, 76:p. 681--689.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998b). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society B, 60:23--40.

Quenoille, M. H. (1949). Problems in plane sampling. Annals of Mathematical Statistics, 20:p. 355--375.

Quenoille, M. H. (1956). Notes on bias in estimation. Biometrika, 43:353--360.

Rao, J. N. K. and 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, 76:221--230.

Robinson, P. M. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. Sankhyā B, 45:240--248.

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

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63:581--592.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons, Nova Iorque.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics, 2:110--114.

Shah, B. V., Folsom, R. E., LaVange, L. M., Wheeless, S. C., Boyle, K. E., and Williams, R. L. (1993). Statistical methods and mathematical algorithms used in sudaan. Technical report.

Skinner, C. J. (1989a). Introduction to Part A. In Analysis of Complex Surveys, pages 23--57. John Wiley and Sons, Chichester.

Skinner, C. J. (1989b). Domain means, regression and multivariate analysis. In Analysis of Complex Surveys, pages 59--87. John Wiley and Sons, Chichester.

Skinner, C. J., Holt, D., and Smith, T. M. F., editors (1989). Analysis of Complex Surveys. John Wiley and Sons, Chichester.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). Model Assisted Survey Sampling. Springer-Verlag, Nova Iorque.

Stata (1997). Stata User's Guide, Release 5. College Station, Texas: Stata Press.

Sudman, S. (1976). Applied Sampling. Academic Press, Nova Iorque.

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

Thomas, D. R. and 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, 82:630--636.

Tillé, Y. and Matei, A. (2016). sampling: Survey Sampling. R package version 2.8.

Westat (1996). A User's Guide to WesVarPc, version 2.0. Westat, Inc., Mariland.

Wickham, H. and Chang, W. (2016). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 2.2.1.

Wolter, K. M. (1985). Introduction to Variance Estimation. Springer-Verlag, Nova Iorque.