Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2018-05-01

Sumário

Prefácio

Placeholder

Agradecimentos

Capítulo 1

Introdução

1.1 Motivação

Este livro trata de questões e ideias de grande importância para os analistas de dados obtidos através de pesquisas amostrais, tais como as conduzidas por agências produtoras de informações estatísticas oficiais ou públicas. Tais dados são comumente utilizados em análises descritivas envolvendo a obtenção de estimativas para totais, médias, proporções e razões. Nessas análises, em geral, são devidamente incorporados os pesos distintos das observações e a estrutura do plano amostral empregado para obter os dados considerados.

Nas últimas décadas tornou-se muito mais frequente um outro tipo de uso de dados de pesquisas amostrais. Tal uso, denominado secundário e/ou analítico, envolve a construção e ajuste de modelos, geralmente feito por analistas que trabalham fora das agências produtoras dos dados. Neste caso, o foco da análise busca estabelecer a natureza de relações ou associações entre variáveis ou testar hipóteses. Para tais fins, a estatística clássica conta com um vasto arsenal de ferramentas de análise, já incorporadas aos principais sistemas estatísticos disponíveis (tais como MINITAB, R, SAS, SPSS, etc).

Muitas ferramentas de análise convencionais disponíveis nesses sistemas estatísticos geralmente partem de hipóteses básicas sobre as amostras disponíveis que só são válidas quando os dados foram obtidos através de Amostras Aleatórias Simples Com Reposição (AASC). Por exemplo, a hipótese de observações Independentes e Identicamente Distribuídas (IID) corresponde justamente ao caso de observações selecionadas por AASC de uma população especificada. Tais hipóteses são geralmente inadequadas para modelar observações provenientes pesquisas amostrais de populações finitas, pois desconsideram os seguintes aspectos relevantes dos planos amostrais usualmente empregados nessas pesquisas:

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

ii.) **conglomeração das unidades**;

iii.) **estratificação**;

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

Em amostragem de populações finitas, a abordagem probabilística emprega pesos para as observações amostrais que dependem das probabilidades de seleção das unidades, que podem ser desiguais. Em consequência, as estimativas pontuais de parâmetros descritivos da população ou mesmo de parâmetros de modelos são influenciadas por pesos distintos das observações.

Além disso, as estimativas de variância (ou da precisão dos estimadores) são influenciadas pela conglomeração, estratificação e pesos, ou no caso de não resposta, também por eventual imputação de dados faltantes ou reponderação das observações disponíveis para compensar a não resposta. Ao ignorar estes aspectos, as ferramentas convencionais dos sistemas estatísticos tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

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

Pesquisa TIC Domicílios 2019 do NIC.br

Os dados deste exemplo são relativos à distribuição dos pesos de domicílios na amostra da Pesquisa TIC Domicílios 2019 do NIC.br (TICDOM 2019), realizada pelo NIC.br. (de Informação e Coordenação do Ponto BR, 2020) apresenta os resultados da pesquisa, e seu capítulo intitulado 'Relatório Metodológico' descreve os métodos e o plano amostral empregado na pesquisa, que foi estratificado e conglomerado em múltiplos estágios, com alocação desproporcional da amostra nos estratos.

As Unidades Primárias de Amostragem (UPAs) foram municípios ou setores censitários da Base Operacional Geográfica do IBGE conforme usada para o Censo Demográfico de 2010. A seleção de municípios quando estes eram UPAs foi feita usando Amostragem Sistemática com Probabilidades Proporcionais ao Tamanho (AS-PPT) - ver a seção 10.6 de (Silva et al., 2020). A seleção dos setores dentro de cada município também foi feita com AS-PPT. Dentro cada setor censitário selecionado, quinze domicílios foram selecionados por amostragem aleatória simples sem reposição, após a atualização do cadastro de domicílios do setor.

A amostra da pesquisa foi planejada e dimensionada visando ao fornecimento de estimativas com precisão adequada para as cinco macrorregiões do Brasil. Os tamanhos da amostra planejada de setores e domicílios para as macrorregiões são apresentados na tabela 1.1.

Tabela 1.1: Tamanhos da amostra de setores e domicílios por macrorregião

Macrorregião	Setores	Domicílios
Norte	201	3.015
Nordeste	617	9.255
Sudeste	863	12.945
Sul	337	5.055
Centro-Oeste	196	2.940
Total	2.214	33.210

Tabela 1.2: Resumos da distribuição dos pesos de domicílios por macrorregião

Macrorregião	Mínimo	Quartil1	Mediana	Quartil3	Máximo
Norte	1,8	1.957	2.898	4.359	82.627
Nordeste	103,8	1.283	2.057	3.314	40.118
Sudeste	36,0	1.814	2.583	3.583	27.993
Sul	20,0	1.028	1.756	2.706	118.715
Centro-Oeste	140,8	1.153	2.401	3.640	29.029
Total	1,8	1.546	2.470	3.636	118.715

A Tabela 1.2 apresenta um resumo das distribuições dos pesos amostrais dos domicílios pesquisados na TICDOM 2019 para as macrorregiões separadamente, e também para o conjunto da amostra da pesquisa.

No cálculo dos pesos amostrais foram consideradas as probabilidades de inclusão dos domicílios na amostra, bem como as correções de calibração para compensar a não-resposta. Contudo, a grande variabilidade dos pesos amostrais da TICDOM 2019 é devida, principalmente, à variabilidade das probabilidades de inclusão na amostra, ilustrando desta forma o ponto i) citado anteriormente nesta seção. Tal variabilidade é devida à alocação desproporcional da amostra entre os estratos geográficos, e ao emprego de contagens defasadas de domicílios nos setores para definir probabilidades de seleção dos mesmos.

Nas análises de dados desta pesquisa, deve-se considerar que há domicílios com pesos muito diferentes. Por exemplo, a razão entre o maior e o menor peso é cerca de 66 mil vezes. Os pesos também variam bastante entre as regiões, com mediana 1,65 vezes maior na região Sudeste quando comparada com a região Norte, em função da alocação desproporcional da amostra nas regiões. Os maiores pesos são também muito maiores que os pesos medianos, com essa razão sendo 48 vezes para o conjunto da amostra.

Tais pesos são utilizados para *expandir* os dados, multiplicando-se cada observação pelo seu respectivo peso. Assim, por exemplo, para *estimar* quantos

domicílios *da população* pertencem a determinado conjunto (*domínio*), basta somar os pesos dos domicílios da amostra que pertencem a este conjunto. É possível ainda incorporar os pesos, de maneira simples e natural, quando se quer estimar medidas descritivas simples da população, tais como totais, médias, proporções, razões, etc. Os métodos para estimação de parâmetros descritivos da população como os aqui citados são cobertos com maior detalhe em (Silva et al., 2020).

Por outro lado, quando se quer utilizar a amostra para estudos analíticos, as opções padrão disponíveis nos sistemas estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações IID. Por exemplo, os procedimentos padrão disponíveis para estimar a média populacional permitem utilizar pesos distintos das observações amostrais, mas tratariam tais pesos como se fossem frequências de observações repetidas na amostra, e portanto interpretariam a soma dos pesos como tamanho amostral, situação que, na maioria das vezes, geraria inferências incorretas sobre a precisão das estimativas resultantes. Isto ocorre porque o tamanho da amostra é muito menor que a soma dos pesos amostrais usualmente encontrados nos arquivos de microdados de pesquisas disseminados por agências de estatísticas oficiais ou públicas, como é o caso da pesquisa TICDOM 2019 aqui considerada. Em tais pesquisas, a opção mais frequente é disseminar pesos que, quando somados, estimam o total de unidades *da população*.

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

Para exemplificar o impacto de ignorar os pesos e o plano amostral ao estimar quantidades descritivas populacionais, tais como totais e proporções, calculamos estimativas de quantidades desses diferentes tipos usando a amostra da TICDOM 2019 juntamente com estimativas das respectivas variâncias. Tais estimativas de variância foram calculadas sob duas estratégias:

a) **considerando Amostragem Aleatória Simples (AAS)** , e portanto ignorando o plano amostral efetivamente adotado na pesquisa; e

b) **considerando o plano amostral da pesquisa e os pesos diferenciados das unidades.**

A razão entre as estimativas de variância obtidas sob o plano amostral verdadeiro (de fato usado na pesquisa) e sob AAS foi calculada para cada uma das estimativas consideradas usando o pacote `survey` do R (Lumley, 2017). Essa razão fornece uma medida do efeito de ignorar o plano amostral. Os resultados das estimativas pontuais e das correspondentes variâncias são apresentados na Tabela **??**, juntamente com as medidas dos Efeitos de Plano Amostral (EPA).

Exemplos de utilização do pacote `survey` para obtenção de estimativas apresentadas na **??** estão na Seção 4. As outras estimativas da Tabela **??** podem ser obtidas de maneira análoga.

Na Tabela **??** apresentamos as estimativas dos seguintes parâmetros populacionais:

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

Como se pode observar da quarta coluna da Tabela **??**, 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 **??** não foram escolhidas de forma a acentuar os efeitos ilustrados, mas tão somente para representar distintos parâmetros (totais, médias, proporções, razões) e variáveis de interesse. Os resultados apresentados para as estimativas de EPA ilustram bem o cenário típico em pesquisas amostrais complexas: o impacto do plano amostral sobre a inferência varia conforme a variável e o tipo de parâmetro de interesse. Note ainda que, à exceção dos dois menores valores (1,26 e 1,99), todas as demais estimativas de EPA apresentaram valores superiores a 2.

1.2 Objetivos do Livro

Este livro tem três objetivos principais:

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

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

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

Para alcançar tais objetivos, adotamos uma abordagem fortemente ancorada na apresentação de exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando os recursos do pacote estatístico R (http://www.r-project.org/).

A comparação dos resultados de análises feitas das duas formas (considerando ou ignorando o plano amostral) permite avaliar o impacto de não se considerar os pontos i) a iv) anteriormente citados. O ponto iv) não é tratado de forma completa neste texto. O leitor interessado na análise de dados sujeitos a não-resposta pode consultar (Kalton, 1983), (Little and Rubin, 2002), (Rubin, 1987), (Särndal et al., 1992), ou (Schafer, 1997), por exemplo.

1.3 Estrutura do Livro

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

No segundo capítulo, procuramos dar uma visão das diferentes abordagens utilizadas na análise estatística de dados de pesquisas amostrais complexas. Apresentamos um referencial para inferência com ênfase no *Modelo de Superpopulação* que incorpora, de forma natural, tanto uma estrutura estocástica para descrever a geração dos dados populacionais (modelo) como o plano amostral efetivamente utilizado para obter os dados amostrais (plano amostral). As referências básicas para seguir este capítulo são o capítulo 2 em (Nascimento Silva, 1996), o capítulo 1 em (Skinner et al., 1989) e os capítulos 1 e 2 em (Chambers and Skinner, 2003).

Esse referencial tem evoluído ao longo dos anos como uma forma de permitir a incorporação de ideias e procedimentos de análise e inferência usualmente associados à Estatística Clássica à prática da análise e interpretação de dados provenientes de pesquisas amostrais. Apesar dessa evolução, sua adoção não é livre de controvérsia e uma breve revisão dessa discussão é apresentada no Capítulo 2.

No Capítulo 3 apresentamos uma revisão sucinta, para recordação, de alguns resultados básicos da Teoria de Amostragem, requeridos nas partes subsequentes do livro. São discutidos os procedimentos básicos para estimação de totais considerando o plano amostral, e em seguida revistas algumas técnicas para estimação de variâncias que são necessárias e úteis para o caso de estatísticas complexas, tais como razões e outras estatísticas requeridas na inferência analítica com dados amostrais. As referências centrais para este capítulo são os capítulos 2 e 3 em (Särndal et al., 1992), (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 estrutura dos dados populacionais ou do plano amostral sobre a estimativa da variância de um estimador. Para isso, comparamos o estimador da variância apropriado para dados obtidos por Amostragem Aleatória Simples (hipótese de AAS) com o valor esperado deste mesmo estimador sob a distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado (plano amostral verdadeiro). Aqui a referência principal foi o livro (Skinner et al., 1989), complementado com o texto de (Lehtonen 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 de interesse mais comuns, como também das variâncias dessas estimativas. As referências básicas utilizadas nesse capítulo foram (Skinner et al., 1989), (Pfeffermann, 1993), (Binder, 1983) e o capítulo 6 em (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 sobre Trabalho da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1990, onde ajustamos um modelo de regressão logística. Neste exemplo, foram feitas comparações entre resultados de ajustes obtidos através de um programa especializado, o pacote `survey` (Lumley, 2017), e através de um programa de uso geral, a função *glm* do R. As referências centrais são o capítulo 6 em (Nascimento Silva, 1996) e (Binder, 1983), além de (Pessoa et al., 1997).

Os Capítulos 7 e 8 tratam da análise de dados categóricos, dando ênfase à adaptação dos testes clássicos para proporções, de independência e de homogeneidade em tabelas de contingência, para lidar com dados provenientes de pesquisas amostrais complexas. Apresentamos correções das estatísticas clássicas e também a estatística de Wald baseada no plano amostral. As referências básicas usadas nesses capítulos foram os o capítulo 4 em (Skinner et al., 1989) e o capítulo 7 (Lehtonen and Pahkinen, 1995). Também são apresentadas as ideias básicas de como efetuar ajuste de modelos log-lineares a dados de frequências

em tabelas de múltiplas entradas.

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

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

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

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

O Capítulo 13 discute algumas formas alternativas de analisar dados de pesquisas amostrais 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., 1998).

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 formulado para descrever propriedades da população de interesse (de onde foi extraída a amostra a ser analisada), 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 amostrais*, seu ajuste prossegue por métodos convencionais tais como *Máxima Verossimilhança (MV)* ou mesmo *Markov Chain Monte Carlo (MCMC)*.

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. Alguns destes programas podem ser adquiridos gratuitamente via *internet*, nos endereços fornecidos de seus produtores. Com isto, pretendemos indicar aos leitores o caminho mais curto para permitir a implementação prática das técnicas e métodos aqui discutidos.

Uma das características que procuramos dar ao livro foi o emprego de exemplos com dados reais, retirados principalmente da experiência do IBGE com pesquisas amostrais complexas. Sem prejuízo na concentração de exemplos que se utilizam de dados de pesquisas do IBGE, incluímos também alguns exemplos que consideram aplicações a dados de pesquisas realizadas por outras instituições. Nas duas décadas desde a primeira edição deste livro foram muitas as iniciativas de realizar pesquisas por amostragem em várias áreas, tendo a educação e a saúde como as mais proeminentes. Para facilitar a localização e replicação dos exemplos pelos leitores, estes foram em sua maioria introduzidos em seções denominadas *Laboratório* ao final de cada um dos capítulos. Os códigos em R dos exemplos são todos fornecidos, o que torna simples a replicação dos mesmos pelos leitores. Optamos pelo emprego do sistema R que, por ser de acesso livre e gratuito, favorece o amplo acesso aos interessados em replicar nossas análises e também em usar as ferramentas disponíveis para implementar suas próprias análises de interesse com outros conjuntos de dados.

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

Uma dificuldade em escrever um livro como este vem do fato de que não é possível começar do zero: é preciso assumir algum conhecimento prévio de ideias e conceitos necessários à compreensão do material tratado. Procuramos tornar o livro acessível para um estudante de fim de curso de graduação em Estatís-

tica. Por essa razão, optamos por não apresentar provas de resultados e, sempre que possível, apresentar os conceitos e ideias de maneira intuitiva, juntamente com uma discussão mais formal para dar solidez aos resultados apresentados. As provas de vários dos resultados aqui discutidos se restringem a material disponível apenas em artigos em periódicos especializados estrangeiros e portanto, são de acesso mais difícil. Ao leitor em busca de maior detalhamento e rigor, sugerimos consultar diretamente as inúmeras referências incluídas ao longo do texto. Para um tratamento mais profundo do assunto, os livros de (Skinner et al., 1989) e (Chambers and Skinner, 2003) são as referências centrais a consultar. Para aqueles querendo um tratamento ainda mais prático que o nosso, os livro de (Lehtonen and Pahkinen, 1995) e (Heeringa et al., 2010) podem ser opções interessantes.

Capítulo 2

Referencial para Inferência

Placeholder

2.1 Modelagem - Primeiras Ideias

2.1.1 Abordagem 1 - Modelagem Clássica

2.1.2 Abordagem 2 - Amostragem Probabilística

2.1.3 Discussão das Abordagens 1 e 2

2.1.4 Abordagem 3 - Modelagem de Superpopulação

2.2 Fontes de Variação

2.3 Modelos de Superpopulação

2.4 Planejamento Amostral

2.5 Planos Amostrais Informativos e Ignoráveis

Capítulo 3

Estimação Baseada no Plano Amostral

Placeholder

3.1 Estimação de Totais

3.2 Por que Estimar Variâncias

3.3 Linearização de Taylor para Estimar variâncias

3.4 Método do Conglomerado Primário

3.5 Métodos de Replicação

3.6 Laboratório de R

Capítulo 4

Efeitos do Plano Amostral

Placeholder

4.1 Introdução

4.2 Efeito do Plano Amostral (EPA) de Kish

4.3 Efeito do Plano Amostral Ampliado

4.4 Intervalos de Confiança e Testes de Hipóteses

4.5 Efeitos Multivariados de Plano Amostral

4.6 Laboratório de R

Capítulo 5

Ajuste de Modelos Paramétricos

Placeholder

5.1 Introdução

5.2 Método de Máxima Verossimilhança (MV)

5.3 Ponderação de Dados Amostrais

5.4 Método de Máxima Pseudo-Verossimilhança

5.5 Robustez do Procedimento MPV

5.6 Desvantagens da Inferência de Aleatorização

5.7 Laboratório de R

Capítulo 6

Modelos de Regressão

Placeholder

6.1 Modelo de Regressão Linear Normal

6.1.1 Especificação do Modelo

6.1.2 Pseudo-parâmetros do Modelo

6.1.3 Estimadores de MPV dos Parâmetros do Modelo

6.1.4 Estimação da Variância de Estimadores de MPV

6.2 Modelo de Regressão Logística

6.3 Teste de Hipóteses

6.4 Laboratório de R

Capítulo 7

Testes de Qualidade de Ajuste

Placeholder

7.1 Introdução

7.2 Teste para uma Proporção

7.2.1 Correção de Estatísticas Clássicas

7.2.2 Estatística de Wald

7.3 Teste para Várias Proporções

7.3.1 Estatística de Wald Baseada no Plano Amostral

7.3.2 Situações Instáveis

7.3.3 Estatística de Pearson com Ajuste de Rao-Scott

7.4 Laboratório de R

Capítulo 8

Testes em Tabelas de Duas Entradas

Placeholder

8.1 Introdução

8.2 Tabelas 2x2

8.2.1 Teste de Independência

8.2.2 Teste de Homogeneidade

8.2.3 Efeitos de Plano Amostral nas Celas

8.3 Tabelas de Duas Entradas (Caso Geral)

8.3.1 Teste de Homogeneidade

8.3.2 Teste de Independência

8.3.3 Estatística de Wald Baseada no Plano Amostral

8.3.4 Estatística de Pearson com Ajuste de Rao-Scott

8.4 Laboratório de R

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

Placeholder

13.1 Introdução

13.2 Modelagem da Estrutura Populacional

13.3 Modelos Hierárquicos

13.4 Análise Desagregada: Prós e Contras

Capítulo 14

Pacotes para Analisar Dados Amostrais

Placeholder

14.1 Introdução

14.2 Pacotes Computacionais

Referências Bibliográficas

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarquical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park.

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

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

de Informação e Coordenação do Ponto BR, N. N. (2020). *Pesquisa Sobre O Uso Das Tecnologias Da Informação E Da Comunicação No Brasil*.

Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis.

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

Lehtonen, R. and Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley and Sons, Chichester.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with missing data*. John Wiley and Sons, Nova Iorque.

Lumley, T. (2017). *survey: Analysis of Complex Survey Samples*. R package version 3.32-1.

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.

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. (1998). Parametric distributions of complex survey data under informative probability survey. *Statistica Sinica*, 8:1087–1114.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Nova Iorque.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall / CRC.

Silva, P. L. d. N., Bianchini, Z. M., and Dias, A. J. R. (2020). *Amostragem: teoria e prática usando R*.

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.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, Nova Iorque.