

Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2018-01-09

Sumário

Prefácio	9
Agradecimentos	9
1 Introdução	11
1.1 Motivação	11
1.2 Objetivos do Livro	14
1.3 Estrutura do Livro	15
1.4 Laboratório de R do Capítulo 1.	17
1.5 Laboratório de R do Capítulo 1 - Extra.	19
2 Referencial para Inferência	21
2.1 Modelagem - Primeiras Ideias	21
2.2 Fontes de Variação	21
2.3 Modelos de Superpopulação	21
2.4 Planejamento Amostral	21
2.5 Planos Amostrais Informativos e Ignoráveis	21
3 Estimação Baseada no Plano Amostral	23
3.1 Estimação de Totais	24
3.2 Por que Estimar Variâncias	24
3.3 Linearização de Taylor para Estimar variâncias	24
3.4 Método do Conglomerado Primário	24
3.5 Métodos de Replicação	24
3.6 Laboratório de R	24
3.7 faixa	24
3.8 0,0118	24
3.9 Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv_se_plan)	24
3.10 Ratios=	24
3.11 faixa	24
3.12 analf.faixa 0,119	24
3.13 SEs=	24
3.14 faixa	24
3.15 analf.faixa 0,0118	24
3.16 Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_jkn)	24
3.17 Ratios=	24
3.18 faixa	24
3.19 analf.faixa 0,119	24
3.20 SEs=	24
3.21 [,1]	24
3.22 [1,] 0,0118	24
3.23 Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_boot)	24
3.24 Ratios=	24

3.25	faixa	24
3.26	analf.faixa 0,119	24
3.27	SEs=	24
3.28	[,1]	24
3.29	[1,] 0,0129	24
3.30	[1] “svyrep.design”	24
3.31	[1] “repweights” “pweights” “type”	24
3.32	[4] “rho” “scale” “rscales”	24
3.33	[7] “call” “combined.weights” “selfrep”	24
3.34	[10] “mse” “variables” “degf”	24
3.35	[1] 8903	24
3.36	[1] 276	24
3.37	num [1:8903, 1:276] 0 0 1,06 1,06 1,06	24
3.38	[1] 0,0118	24
3.39	theta SE	24
3.40	[1,] 0,119 0,01	24
3.41	theta SE	24
3.42	[1,] 0,504 0,05	24
3.43	nlcon SE	24
3.44	contrast 0,504 0,05	24
4	Efeitos do Plano Amostral	25
4.1	Introdução	26
4.2	Efeito do Plano Amostral (EPA) de Kish	26
4.3	Efeito do Plano Amostral Ampliado	26
4.4	245,188/1719,979=0,143	26
4.5	0,401/1,207=0,332	26
4.6	244,176/1613,3=0,151	26
4.7	0,435/1,188=0,366	26
4.8	Intervalos de Confiança e Testes de Hipóteses	26
4.9	Efeitos Multivariados de Plano Amostral	26
4.10	Laboratório de R	26
4.11	Média das estimativas pontuais para as 500 amostras aas	26
4.12	[1] 163,50 4,17	26
4.13	Média das estimativas pontuais para as 500 amostras aes	26
4.14	sal rec	26
4.15	78,07 2,06	26
4.16	[,1] [,2]	26
4.17	[1,] 1720,0 26,78	26
4.18	[2,] 26,8 1,21	26
4.19	sal rec	26
4.20	sal 245,19 3,172	26
4.21	rec 3,17 0,401	26
4.22	Matriz de covariância populacional	26
4.23	Matriz de covariância considerando o plano amostral	26
4.24	verdadeiro	26
4.25	estimativa de efeitos generalizados do plano amostral	26
4.26	[1] 706 14	26
4.27	[1] “setor” “np” “domic” “sexo” “renda” “lrenda” “raca”	26
4.28	[8] “estudo” “idade” “na” “peso” “domtot” “pesol” “pesof”	26
4.29	raca renda se	26
4.30	1 1 110406 11262	26
4.31	2 2 73560 8207	26
4.32	sexo renda se	26

4.33	1 1 108746 11696	26
4.34	2 2 40039 4042	26
4.35	26
4.36	Design-based t-test	26
4.37	26
4.38	data: renda ~ sexo	26
4.39	t = -6, df = 20, p-value = 0,000005	26
4.40	alternative hypothesis: true difference in mean is not equal to 0	26
4.41	sample estimates:	26
4.42	difference in mean	26
4.43	-68707	26
4.44	26
4.45	Design-based t-test	26
4.46	26
4.47	data: renda ~ raca	26
4.48	t = -4, df = 20, p-value = 0,0006	26
4.49	alternative hypothesis: true difference in mean is not equal to 0	26
4.50	sample estimates:	26
4.51	difference in mean	26
4.52	-36846	26
5	Ajuste de Modelos Paramétricos	27
5.1	Introdução	27
5.2	Método de Máxima Verossimilhança (MV)	27
5.3	Ponderação de Dados Amostrais	27
5.4	Método de Máxima Pseudo-Verossimilhança	27
5.5	Robustez do Procedimento MPV	27
5.6	Desvantagens da Inferência de Aleatorização	27
5.7	Laboratório de R	27
6	Modelos de Regressão	29
6.1	Modelo de Regressão Linear Normal	30
6.2	Modelo de Regressão Logística	30
6.3	Teste de Hipóteses	30
6.4	Laboratório de R	30
6.5	[1] "stra" "psu" "pesopes" "informal" "sx" "id"	30
6.6	[7] "ae" "ht" "re" "um"	30
6.7	stra psu pesopes informal sx id ae	30
6.8	"numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"	30
6.9	ht re um	30
6.10	"numeric" "numeric" "numeric"	30
6.11	Wald test for ht:re	30
6.12	in svyglm(formula = informal ~ sx + ae + ht + id + re + sx * id +	30
6.13	sx * ht + ae * ht + ht * id + ht * re, design = pnad.des,	30
6.14	family = quasibinomial())	30
6.15	F = 6,74 on 4 and 616 df: p= 0,00003	30
7	Testes de Qualidade de Ajuste	31
7.1	Introdução	32
7.2	Teste para uma Proporção	32
7.3	Teste para Várias Proporções	32
7.4	Laboratório de R	32
7.5	[,1]	32
7.6	[,1] 5,74	32

7.7	[,1]	32
7.8	[1,] 0,219	32
7.9	[,1]	32
7.10	[1,] 11,5	32
7.11	[,1]	32
7.12	[1,] 0,021	32
8	Testes em Tabelas de Duas Entradas	33
8.1	Introdução	34
8.2	Tabelas 2x2	34
8.3	Tabelas de Duas Entradas (Caso Geral)	34
8.4	Laboratório de R	34
8.5	[1] “stra” “psu” “pesopes” “informal” “sx” “id”	34
8.6	[7] “ae” “ht” “re” “um”	34
8.7	stra psu pesopes informal sx id ae	34
8.8	“numeric” “numeric” “numeric” “numeric” “numeric” “numeric” “numeric”	34
8.9	ht re um	34
8.10	“numeric” “numeric” “numeric”	34
8.11	mean SE	34
8.12	sx1 0,657 0,01	34
8.13	sx2 0,343 0,01	34
8.14	mean SE	34
8.15	re1 0,155 0,01	34
8.16	re2 0,584 0,01	34
8.17	re3 0,261 0,01	34
8.18	mean SE	34
8.19	ae1 0,313 0,01	34
8.20	ae2 0,320 0,01	34
8.21	ae3 0,367 0,01	34
8.22	ht1 ht2 ht3	34
8.23	0,210 0,615 0,175	34
8.24	\$names	34
8.25	[1] “ht1” “ht2” “ht3”	34
8.26		34
8.27	\$var	34
8.28	ht1 ht2 ht3	34
8.29	ht1 0,00003666 -0,0000332 -0,00000344	34
8.30	ht2 -0,00003323 0,0000676 -0,00003436	34
8.31	ht3 -0,00000344 -0,0000344 0,00003780	34
8.32		34
8.33	\$statistic	34
8.34	[1] “mean”	34
8.35		34
8.36	\$class	34
8.37	[1] “svystat”	34
8.38	ht1 ht2 ht3	34
8.39	ht1 0,00003666 -0,0000332 -0,00000344	34
8.40	ht2 -0,00003323 0,0000676 -0,00003436	34
8.41	ht3 -0,00000344 -0,0000344 0,00003780	34
8.42	ht1 ht2 ht3	34
8.43	ht1 0,00003666 -0,0000332 -0,00000344	34
8.44	ht2 -0,00003323 0,0000676 -0,00003436	34
8.45	ht3 -0,00000344 -0,0000344 0,00003780	34
8.46	mean SE DEff	34

8.47	re1 0,15546 0,00690 2,36	34
8.48	re2 0,58356 0,00820 1,80	34
8.49	re3 0,26098 0,00961 3,12	34
8.50	re	34
8.51	sx 1 2 3	34
8.52	1 0,073 0,388 0,196	34
8.53	2 0,082 0,195 0,065	34
8.54	sx re1 re2 re3 se.re1 se.re2 se.re3	34
8.55	1 1 0,111 0,591 0,298 0,00573 0,0103 0,0111	34
8.56	2 2 0,240 0,570 0,190 0,01250 0,0119 0,0111	34
8.57	mean SE DEff	34
8.58	I((sx == 1 & re == 1) * 1) 0,07299 0,00383 1,42	34
8.59	re	34
8.60	sx 1 2 3	34
8.61	1 0,0730 0,3882 0,1959	34
8.62	2 0,0825 0,1953 0,0651	34
8.63	re	34
8.64	sx 1 2 3	34
8.65	1 7,30 38,82 19,59	34
8.66	2 8,25 19,53 6,51	34
8.67	[1] 1,64	34
8.68	34
8.69	Pearson's Chi-squared test	34
8.70	34
8.71	data: tab.amo	34
8.72	X-squared = 200, df = 2, p-value <0,0000000000000002	34
9	Estimação de densidades	35
9.1	Introdução	35
10	Modelos Hierárquicos	37
10.1	Introdução	37
11	Não-Resposta	39
11.1	Introdução	39
12	Diagnóstico de ajuste de modelo	41
12.1	Introdução	41
13	Agregação vs. Desagregação	43
13.1	Introdução	43
13.2	Modelagem da Estrutura Populacional	43
13.3	Modelos Hierárquicos	43
13.4	Análise Desagregada: Prós e Contras	43
14	Pacotes para Analisar Dados Amostrais	45
14.1	Introdução	45
14.2	Pacotes Computacionais	45
15	Placeholder	47

Prefácio

Agradecimentos

Capítulo 1

Introdução

1.1 Motivação

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

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

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

- i.) **probabilidades distintas de seleção das unidades;**
- ii.) **conglomeramento das unidades;**
- iii.) **estratificação;**
- iv.) **calibração ou imputação para não-resposta e outros ajustes.**

As estimativas pontuais de parâmetros descritivos da população ou de modelos são influenciadas por pesos distintos das observações. Além disso, as estimativas de variância (ou da precisão dos estimadores) são influenciadas pela conglomeramento, estratificação e pesos, ou no caso de não resposta, também por eventual imputação de dados faltantes ou reponderação das observações disponíveis. Ao ignorar estes aspectos, os pacotes tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

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

Exemplo 1.1. Distribuição dos pesos da amostra da PPV

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

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

Os dados deste exemplo são relativos à distribuição dos pesos na amostra da Pesquisa sobre Padrões de Vida (PPV), realizada pelo IBGE nos anos 1996-97. (Albieri and Bianchini, 1997) descrevem resumidamente a PPV, que foi realizada nas Regiões Nordeste e Sudeste do País.

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

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

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

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

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

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

No cálculo dos pesos amostrais foram consideradas as probabilidades de inclusão dos elementos na amostra, bem como correções para compensar a não-resposta. Contudo, a grande variabilidade dos pesos amostrais da PPV é devida, principalmente, à variabilidade das probabilidades de inclusão na amostra, ilustrando desta forma o ponto i) citado anteriormente nesta seção. Tal variabilidade foi provocada pela decisão de alocar a amostra de forma igual entre os estratos geográficos, cujos totais populacionais são bastante distintos.

Na análise de dados desta pesquisa, deve-se considerar que há elementos da amostra com pesos muito distintos.

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

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

Por exemplo, a razão entre o maior e o menor peso é cerca de 40 vezes. Os pesos também variam bastante entre as regiões, com mediana 3,5 vezes maior na região Sudeste quando comparada com a região Nordeste, em função da alocação igual mas desproporcional da amostra nas regiões.

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

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

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

Para exemplificar o impacto de ignorar os pesos e o plano amostral ao estimar quantidades descritivas populacionais, tais como totais, médias, proporções e razões, calculamos estimativas de quantidades desses diferentes tipos usando a amostra da PPV juntamente com estimativas das respectivas variâncias. Tais estimativas de variância foram calculadas sob duas estratégias: a) considerando amostragem aleatória simples (portanto ignorando o plano amostral efetivamente adotado na pesquisa), e b) considerando o plano amostral da pesquisa e os pesos diferenciados das unidades.

A razão entre as estimativas de variância obtidas sob o plano amostral verdadeiro e sob amostragem aleatória simples foi calculada para cada uma das estimativas consideradas usando o pacote **survey** do R (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 do pacote **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 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;

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

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

7. % de analfabetos de mais de 14 anos;
8. Total de mulheres de 12 a 49 anos que tiveram filhos;
9. Total de mulheres de 12 a 49 anos que tiveram filhos vivos;
10. Total de mulheres de 12 a 49 anos que tiveram filhos mortos;
11. Número médio de filhos tidos por mulheres de 12 a 49 anos;
12. Razão de dependência.

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

Note que as variáveis e parâmetros cujas estimativas são apresentadas na Tabela 1.3 não foram escolhidas de forma a acentuar os efeitos ilustrados, mas tão somente para representar distintos parâmetros (totais, médias, proporções, razões) e variáveis de interesse. Os resultados apresentados para as estimativas de EPA ilustram bem o cenário típico em pesquisas amostrais complexas: o impacto do plano amostral sobre a inferência varia conforme a variável e o tipo de parâmetro de interesse. Note ainda que, à exceção dos dois menores valores (1,26 e 1,99), todas as demais estimativas de EPA apresentaram valores superiores a 2.

1.2 Objetivos do Livro

Este livro tem três objetivos principais:

- 1) **Ilustrar e analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise de dados quando estes são provenientes de pesquisas amostrais complexas;**
- 2) **Apresentar uma coleção de métodos e recursos computacionais disponíveis para análise de dados amostrais complexos, equipando o analista para trabalhar com tais dados, reduzindo assim o risco de inferências incorretas;**
- 3) **Ilustrar o potencial analítico de muitas das pesquisas produzidas por agências de estatísticas oficiais para responder questões de interesse, mediante uso de ferramentas de**

análise estatística agora já bastante difundidas, aumentando assim o valor adicionado destas pesquisas.

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

A comparação dos resultados de análises feitas das duas formas (considerando ou ignorando o plano amostral) permite avaliar o impacto de não se considerar os pontos i) a iv) anteriormente citados. O ponto iv) não é tratado de forma completa neste texto. O leitor interessado na análise de dados sujeitos a não-resposta pode consultar (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 Trabalho da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1990, onde ajustamos um modelo de regressão logística. Neste exemplo, são feitas comparações entre resultados de ajustes obtidos

através de um programa especializado, o pacote `survey` (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 com ênfase na adaptação dos testes clássicos para proporções, de independência e de homogeneidade em tabelas de contingência, para dados provenientes de pesquisas amostrais complexas. Apresentamos correções das estatísticas clássicas e também a estatística de Wald baseada no plano amostral. As referências básicas usadas nesses capítulos foram os o capítulo 4 em (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 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 multiníveis) têm sido bastante utilizados para explorar situações em que as relações entre variáveis de interesse em uma certa população de unidades elementares (por exemplo, crianças em escolas, pacientes em hospitais, empregados em empresas, moradores em regiões, etc.) são afetadas por efeitos de grupos determinados ao nível de unidades conglomeradas (os grupos). Ajustar e interpretar tais modelos é tarefa mais difícil que o mero ajuste de modelos lineares, mesmo em casos onde os dados são obtidos de forma exaustiva, mas ainda mais complicada quando se trata de dados obtidos através de pesquisas amostrais complexas. Várias alternativas de métodos para ajuste de modelos hierárquicos estão disponíveis, e este capítulo apresenta uma revisão de tais abordagens, ilustrando com aplicações a dados de pesquisas amostrais de escolares.

O Capítulo 11 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 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., 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 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. Alguns destes programas

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

Uma das características que procuramos dar ao livro foi o emprego de exemplos com dados reais, retirados principalmente da experiência do IBGE com pesquisas amostrais complexas. Embora a experiência de fazer inferência analítica com dados desse tipo seja ainda incipiente no Brasil, acreditamos ser fundamental difundir essas ideias para alimentar um processo de melhoria do aproveitamento dos dados das inúmeras pesquisas realizadas pelo IBGE e instituições congêneres, que permita ir além da tradicional estimação de totais, médias, proporções e razões. Esperamos com esse livro fazer uma contribuição a esse processo.

Uma dificuldade em escrever um livro como este vem do fato de que não é possível começar do zero: é preciso assumir algum conhecimento prévio de ideias e conceitos necessários à compreensão do material tratado. Procuramos tornar o livro acessível para um estudante de fim de curso de graduação em Estatística. Por essa razão, optamos por não apresentar provas de resultados e, sempre que possível, apresentar os conceitos e ideias de maneira intuitiva, juntamente com uma discussão mais formal para dar solidez aos resultados apresentados. As provas de vários dos resultados aqui discutidos se restringem a material disponível apenas em artigos em periódicos especializados estrangeiros e portanto, são de acesso mais difícil. Ao leitor em busca de maior detalhamento e rigor, sugerimos consultar diretamente as inúmeras referências incluídas ao longo do texto. Para um tratamento mais profundo do assunto, os livros de (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 livros de (Lehtonen and Pahkinen, 1995) e (Heeringa et al., 2010) podem ser opções interessantes.

1.4 Laboratório de R do Capítulo 1.

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

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

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

## [1] 19409    13

names(ppv_dat)

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

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

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

```
# Adiciona variáveis ao arquivo da ppv_dat
ppv_dat <- transform(ppv_dat,
  anal1 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >= 7 & v02a08 <= 14)) * 1,
```

```

analf2 = ((v04a01 == 2 | v04a02 == 2) & (v02a08 >14)) * 1,
faixa1 = (v02a08 >= 7 & v02a08 <= 14) *1,
faixa2 = (v02a08 > 14) * 1)
#str(ppv_dat)

```

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

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

O passo fundamental para utilização do pacote `survey` (Lumley, 2017) é criar um objeto que guarde as informações relevantes sobre a estrutura do plano amostral junto dos dados. Isso é feito por meio da função `svydesign()`. As variáveis que definem estratos, conglomerados e pesos na PPV são `estrato`, `nsetor` e `pesof` respectivamente. O objeto de desenho amostral, `ppv_plan` incorpora as informações da estrutura do plano amostral adotado na PPV.

```

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

```

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

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

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

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

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

```
##          total      SE DEff
## analf1 1174220 127982 2,05
```

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

```
##          total      SE DEff
## analf2 4792344 318877 3,32
```

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

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

```
## Ratio estimator: svyratio.survey.design2(~analf1, ~faixa1, ppv_se_plan)
## Ratios=
##          faixa1
## analf1  0,119
## SEs=
##          faixa1
## analf1 0,0118
```

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

```
## Ratio estimator: svyratio.survey.design2(~analf2, ~faixa2, ppv_se_plan)
## Ratios=
##      faixa2
## analf2 0,109
## SEs=
##      faixa2
## analf2 0,00673
```

Uma alternativa para obter estimativa por domínios é utilizar a função `svyby()` do pacote `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_plan, svytotal, deff = TRUE)
```

```
##      regiao  analf1      se DEff.analf1
## Nordeste Nordeste 3512866 352620      9,66
## Sudeste  Sudeste 1174220 127982      2,05
```

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

1.5 Laboratório de R do Capítulo 1 - Extra.

Exemplo 1.3. Exemplo anterior usando o pacote `srvyr`

- Carrega o pacote `srvyr`:

```
library(srvyr)
```

- Cria objeto de desenho:

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

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

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

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

```
result1 <- ppv_plan %>%
  group_by(regiao) %>%
  summarise(
    taxa_analf1 = 100*survey_ratio(analf1, faixa1),
    taxa_analf2 = 100*survey_ratio(analf2, faixa2)
```

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
Nordeste	42,3	3,1	33,6	1,6
Sudeste	11,9	1,2	10,9	0,7

```

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

```

Capítulo 2

Referencial para Inferência

2.1 Modelagem - Primeiras Ideias

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

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

3.7 faixa

3.8 0,0118

3.9 Ratio estimator: `svyratio.survey.design2(~analf.faixa, ~faixa, ppv__se__plan)`

3.10 Ratios=

3.11 faixa

3.12 `analf.faixa` 0,119

3.13 SEs=

3.14 faixa

Capítulo 4

Efeitos do Plano Amostral

4.1 Introdução

4.2 Efeito do Plano Amostral (EPA) de Kish

4.3 Efeito do Plano Amostral Ampliado

4.4 $245,188/1719,979=0,143$

4.5 $0,401/1,207=0,332$

4.6 $244,176/1613,3=0,151$

4.7 $0,435/1,188=0,366$

4.8 Intervalos de Confiança e Testes de Hipóteses

4.9 Efeitos Multivariados de Plano Amostral

4.10 Laboratório de R

4.11 Média das estimativas pontuais para as 500 amostras aas

4.12 [1] 163,50 4,17

4.13 Média das estimativas pontuais para as 500 amostras aes

4.14 sal rec

4.15 78,07 2,06

4.16 [,1] [,2]

Capítulo 5

Ajuste de Modelos Paramétricos

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

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

6.5 [1] “stra” “psu” “pesopes” “informal” “sx” “id”

6.6 [7] “ae” “ht” “re” “um”

6.7 stra psu pesopes informal sx id ae

6.8 “numeric” “numeric” “numeric” “numeric” “numeric” “numeric” “numeric”

6.9 ht re um

6.10 “numeric” “numeric” “numeric”

6.11 Wald test for ht:re

6.12 in svyglm(formula = informal ~ sx + ae + ht + id + re + sx * id +

Capítulo 7

Testes de Qualidade de Ajuste

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

7.5 [1]

7.6 [1,] 5,74

7.7 [1]

7.8 [1,] 0,219

7.9 [1]

7.10 [1,] 11,5

7.11 [1]

7.12 [1,] 0,021

Capítulo 8

Testes em Tabelas de Duas Entradas

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

8.5 [1] “stra” “psu” “pesopes” “informal” “sx” “id”

8.6 [7] “ae” “ht” “re” “um”

8.7 stra psu pesopes informal sx id ae

8.8 “numeric” “numeric” “numeric” “numeric” “numeric” “numeric” “numeric”

8.9 ht re um

8.10 “numeric” “numeric” “numeric”

Capítulo 9

Estimação de densidades

9.1 Introdução

Capítulo 10

Modelos Hierárquicos

10.1 Introdução

Capítulo 11

Não-Resposta

11.1 Introdução

Capítulo 12

Diagnóstico de ajuste de modelo

12.1 Introdução

Capítulo 13

Agregação vs. Desagregação

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

14.1 Introdução

14.2 Pacotes Computacionais

Capítulo 15

Placeholder

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.
- 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). *Hierarchical 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.
- 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.
- 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.