Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Uma revisão sobre o uso analítico de dados provenientes de amostras com estruturas complexas

Gislaine Rocha Pereira

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2016

Gislaine Rocha Pereira
Licenciada em Matemática

Uma revisão sobre o uso analítico de dados provenientes de amostras com estruturas complexas

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof^{a}. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2016

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Pereira, Gislaine Rocha
Uma revisão sobre o uso analítico de dados provenientes de amostras com estruturas complexas / Gislaine Rocha Pereira. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.
89 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Amostragem 2. Análise de regressão 3. Tabelas de contingência I. Título

CDD 519.52
P426r

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

DEDICATÓRIA

Aos meus pais, pelo apoio e suporte contínuos em todas as minhas escolhas, dedico.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por ter me abençoado e me permitido chegar até aqui colocando sempre um dos seus anjos ao meu lado e me mostrando o caminho nos momentos incertos.

À minha família pelo amor, paciência e incentivo. Agradeço especialmente à minha mãe, uma mulher extraordinária, uma incansável trabalhadora e um exemplo de vida.

Aos docentes e funcionários do Departamento de Ciências Exatas da Esalq por toda atenção e pelos ensinamentos que me acompanharão por toda vida. Agradeço em especial à minha orientadora Profª Sônia Maria De Stefano Piedade por aceitar trabalhar comigo e ao Profº Idemauro Rodrigues Lara por me ajudar com a sugestão do tema.

Aos grandes amigos que fiz nesse período, em especial aos amigos de departamento Maíra, Viviane, César, Vinícius, Christopher, Alessandra, Renan e Altemir, que me ajudaram grandemente tanto nas disciplinas quanto na escrita deste trabalho.

À Capes pelo apoio financeiro.

A todos, muito obrigada.

"Tente uma, duas, três vezes e se possível tente a quarta, a quinta e quantas vezes for necessário. Só não desista nas primeiras tentativas. A persistência é amiga da conquista. Se você quer chegar onde a maioria não chega, faça o que a maioria não faz."

Bill Gates

SUMÁRIO

RESUMO 11
ABSTRACT 13
LISTA DE TABELAS 15
1 INTRODUÇÃO 17
2 REVISÃO BIBLIOGRÁFICA 21
2.1 Planos Amostrais 21
2.1.1 Amostragem Aleatória Simples 21
2.1.1.1 Amostragem Aleatória Simples com Reposição 22
2.1.1.2 Amostragem Aleatória Simples sem Reposição 22
2.1.2 Amostragem Estratificada 23
2.1.3 Amostragem por Conglomerados 24
2.2 Efeitos do Plano Amostral 25
2.2.1 Efeito do Plano Amostral (EPA) de Kish 25
2.2.2 Efeito do Plano Amostral Ampliado 26
2.2.3 Efeito Multivariado de Plano Amostral 29
2.2.4 Estimação de Variâncias 29
2.2.4.1 Método do Conglomerado Primário 30
2.2.4.2 Método da Linearização de Taylor 31
2.3 Modelos de Regressão Incorporando o Esquema Amostral 33
2.3.1 Método da Máxima Verossimilhança 35
2.3.2 Método de Máxima Pseudo-Verossimilhança 36
2.3.3 Ajuste de Modelos de Regressão 40
2.3.3.1 Regressão Linear Normal 40
2.3.3.2 Regressão Logistíca 42
2.3.3.3 Teste de Hipóteses 43
2.3.4 Aplicação em Regressão Linear Normal 43
2.4 Análise em Tabelas de Contingência 48
2.4.1 Tabelas de uma entrada 50
2.4.1.1 Estatística χ^2 de Pearson 50
2.4.1.2 Estatística de Wald 53
2.4.1.3 Análise de Resíduos 54
2.4.1.4 Uma aplicação aos dados da Mini-Finland Health Survey 54
2.4.2 Tabelas de duas entradas 58
2.4.2.1 Teste de Homogeneidade 60
2.4.2.1.1 Análise de Resíduos 62
2.4.2.2 Teste de Independência 63

2.4.2.2.1 Análise de Resíduos 66
2.4.2.3 Uma aplicação do teste de homogeneidade aos dados da Occupational Health Care Survey 66
3 DESENVOLVIMENTO 71
3.1 Exemplo 1 71
3.2 Exemplo 2 74
3.3 Exemplo 3 77
4 CONSIDERAÇÕES FINAIS 81
REFERÊNCIAS 83

RESUMO

Uma revisão sobre o uso analítico de dados provenientes de amostras com estruturas complexas

Neste trabalho foi realizada uma revisão bibliográfica acerca das metodologias encontradas na literatura de como são aplicados os métodos para o uso analítico de dados provenientes de pesquisas que envolvem esquemas amostrais complexos. Objetivou-se mostrar e discutir alguns estudos que avaliam o impacto de ignorar o plano amostral na análise dos dados. Foi feito também um levantamento de artigos com o objetivo de fazer um estudo de trabalhos publicados em jornais, revistas ou periódicos, cujos assuntos abordados tratam da incorporação da estrutura complexa da amostra na análise. Essa revisão evidenciou que os métodos clássicos de análise, ou seja, aqueles que supõem que os dados provém de uma amostragem aleatória simples, podem levar a resultados incorretos produzindo conclusões errôneas ou equivocadas quando os dados provém de esquemas amostrais complexos.

Palavras-chave: Amostragem; Análise de regressão; Tabelas de contingência

ABSTRACT

A review about the analytic use of data from complex structures

This work was carried out a literature review about the methodologies found in the literature of how the methods for data analytical use from research involving complex sampling schemes are applied. It was aimed to show and discuss some studies that assess the impact of ignoring the sampling scheme in the data analysis. It was also made a survey of articles in order to make a study of works published in newspapers, magazines or periodicals, which addressed issues dealing with the incorporation of the complex structure of the sample in the analysis. This review shown that the classical methods of analysis, i.e. those who assume that the data comes from a simple random sampling can lead to incorrect results producing quite erroneous and misleading conclusions when the data come from complex sample schemes.

Keywords: Sampling; Regression analysis; Contingency tables

LISTA DE TABELAS

Tabela 2.1 – Distribuição de frequência relativa dos jovens por situação ocupacional . . . 28
Tabela 2.2 – Estimativas dos coeficientes dos logitos relacionados às categorias inatividade e desemprego . . . 29
Tabela 2.3 – Fragmento do conjunto de dados com variáveis auxiliares . . . 44
Tabela 2.4 – Descrição das variáveis . . . 45
Tabela 2.5 – Modelo completo . . . 47
Tabela 2.6 – Modelo final considerando AAS . . . 48
Tabela 2.7 – Modelo final considerando o plano amostral . . . 49
Tabela 2.8 – Distribuíção da idade, proporções (em %) de pessoas com doenças crônicas e expostas a riscos de saúde física no trabalho, e pressão arterial em um subgrupo da pesquisa MFH com homens entre 30 e 64 anos de idade . . . 55
Tabela 2.9 – Vetores de proporções, efeitos do plano amostral e resíduos padronizados por classes de idade na MFH . . . 55
Tabela 2.10–Valores e p-valores de estatísticas alternativas de teste . . . 57
Tabela 2.11–Tabela (2×2) de proporções . . . 58
Tabela 2.12–Proporções de linhas em uma tabela $L \times C$ para um teste de homogeneidade. 60
Tabela 2.13–Proporções por cela em uma tabela $L \times C$ para um teste de independência . 63
Tabela 2.14–Dados da pesquisa OHC categorizados por sexo, idade e proporções $(\%)$ de pessoas com doenças crônicas e de pessoas expostas a riscos para a saúde física de trabalho . . . 67
Tabela 2.15–Proporções na classes da variável “Riscos no Trabalho” em empresas públicas e em empresas privadas, com o efeito do plano amostral nas celas entre parenteses . . . 68
Tabela 2.16–Resíduos padronizados baseados no plano amostral . . . 70
Tabela 3.1 – Níveis de significância estimados $(\%)$ em oposição a um nível de significância nominal de 5% no teste de bondade de ajuste para os dados da GHS . . . 78
Tabela 3.2 – Níveis de significância estimados $(\%)$ em oposição a um nível de significância nominal de 5% no teste de bondade de ajuste para os dados da BES . . . 79
Tabela 3.3 – Níveis de significância estimados $(\%)$ em oposição a um nível de significância nominal de 5% no teste de homogeneidade para os dados da GHS . . . 79
Tabela 3.4 – Níveis de significância estimados $(\%)$ em oposição a um nível de significância nominal de 5% no teste de homogeneidade para o desenho amostral B com várias populações . . . 80
Tabela 3.5 – Níveis de significância estimados $(\%)$ em oposição a um nível de significância nominal de 5% no teste de independência para os dados da GHS . . . 80

1 INTRODUÇÃO

Pesquisas por amostragem são comumente utilizadas em diversas áreas como agricultura, economia, educação, política, saúde dentre outras. Organizações governamentais nacionais e internacionais e agências privadas coletam informações por meio dessas pesquisas para avaliar a eficácia de medicamentos no tratamento de doenças, a relação entre incidência de doenças e fatores de risco, a intenção de votos durante o período eleitoral, entre outros interesses.

Segundo Scheaffer *et al.* (2011), o objetivo de pesquisas por amostragem é fazer inferências sobre a população a partir de informações contidas em uma amostra selecionada dentro desta população. A grande vantagem em utilizá-las é que, quando realizadas com um plano amostral representativo, ou seja, quando a amostra selecionada tem as mesmas propriedades que a população de interesse, elas podem fornecer informações importantes para problemas práticos com boa precisão.

Diante da dificuldade de avaliar se um plano amostral é representativo sem conhecer todas as informações populacionais de interesse, é preferível trabalhar com o conceito de plano amostral probabilístico, que envolve procedimentos em que cada possível amostra tem uma probabilidade conhecida de ocorrer. Särndal, Swensson e Wretman (2003) definem amostragem probabilística como uma abordagem para a seleção de amostras que satisfaz:

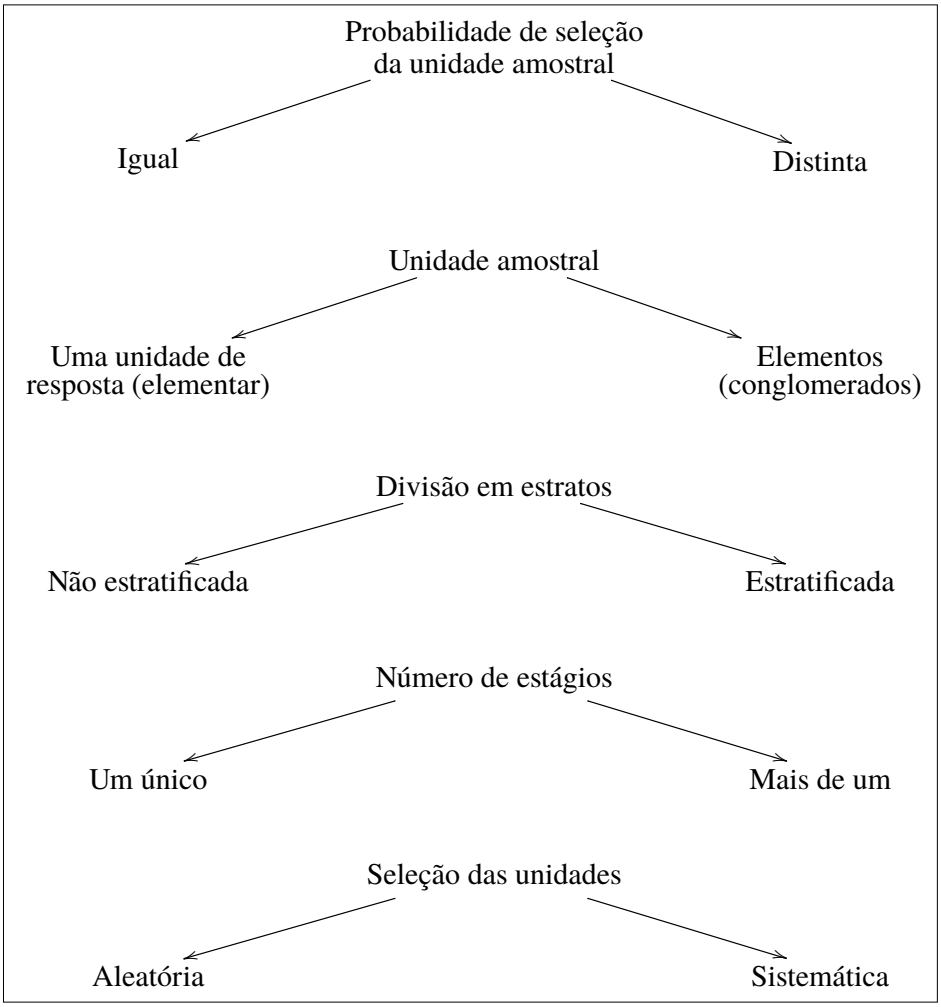
i) Pode-se definir um conjunto $S(\mathcal{U}) = \{s_1, s_2, \ldots, s_M\}$ de todas as amostras que são possíveis de obter com o processo de amostragem;

ii) Uma probabilidade conhecida de seleção $p(s)$ está associada a cada amostra possível s;

iii) Cada elemento da população tem uma probabilidade diferente de zero de ser selecionado;

iv) A amostra é selecionada por um mecanismo aleatório de forma que cada possível amostra s recebe exatamente a probabilidade $p(s)$;

O diagrama a seguir, encontrado em Bolfarine e Bussab (2005), resume os critérios propostos por Kish (1965) para classificar amostras probabilísticas. Usando, por exemplo, as primeiras opções de cada critério tem-se o caso mais simples de amostragem probabilística que ocorre quando cada unidade elementar é sorteada com igual probabilidade, individualmente, sem estratificação, e um único estágio e seleção aleatória. Neste caso, diz-se que a amostra foi sorteada por Amostragem Aleatória Simples.



Segundo Bolfarine e Bussab (2005), o sorteio das unidades com igual probabilidade é apenas uma estratégia que facilita bastante o desenvolvimento das propriedades matemáticas associadas ao plano, mas às vezes é conveniente sortear as unidade com probabilidades desiguais. Neste último caso, e se ainda não for feita reposição, os modelos de análise tornam-se bastante difíceis de serem derivados.

Um plano amostral com probabilidades desiguais de seleção amplamente utilizado é Amostragem Aleatória Estratificada que consiste, segundodefinição de Bolfarine e Bussab (2005), na divisão de uma população em grupos (estratos) segundo alguma(s) característica(s) conhecida(s) na população sob estudo, e de cada um desses estratos são selecionadas amostras em proporções convenientes. A grande vantagem em utilizar este método é o aumento da eficiência e precisão das estimativas, e a desvantagem é que para isso ocorrer é necessário conhecer os resultados de uma pesquisa anterior com informações das características que se deseja pesquisar na população de interesse.

Quando essas informações sobre a população estão desatualizadas ou não estão dispo-

níveis, a amostragem pode ser facilitada dividindo a população em alguns grupos, chamados de conglomerados, segundo algum critério conveniente. Geralmente, os conglomerados são escolhidos como regiões geográficas. Este tipo de plano amostral com probabilidades de seleção desiguais, denominada Amostragem Aleatória por Conglomerados, consiste na divisão da população em conglomerados e no sorteio de alguns desses conglomerados para a mostra. Posteriormente, dentro de cada conglomerado escolhido, pode ser feita a coleta das informações desejadas em todas as unidades da população ou pode ser realizado um novo sorteio para determinar as unidades da população pertencentes aos conglomerados que responderão à pesquisa.

A análise de dados provenientes de pesquisas por amostragem que incorpora níveis de complexidade como os descritos anteriormente - estratificação, conglomeração (em um ou mais níveis), probabilidades desiguais de seleção - é denominada análise de dados amostrais complexos. Em sua maioria, estes dados são utilizados para análises descritivas como, por exemplo, o cálculo de médias, totais, taxas, razões, proporções populacionais etc, nas quais são devidamente considerados os pesos distintos das observações e o esquema amostral usado na pesquisa.

Porém, como explicam Pessoa e Silva (1998), outras vezes 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, abordagens clássicas de análise utilizando pacotes computacionais estatísticos padrões disponíveis para a seleção e ajuste de modelos.

Contudo, o grande problema desta abordagem é o fato de que normalmente esses dados, como dito anteriormente, provém de pesquisas amostrais complexas. Por este motivo, não é aconselhável utilizar os procedimentos clássicos de análise disponíveis nesses pacotes, pois seu uso se faz sob condições que não refletem a complexidade envolvida nas pesquisas amostrais de populações finitas. Vale lembrar que, em geral, estes pacotes tratam os dados como se fossem observações independentes e identicamente distribuídas, isto é, ignoram o método de seleção da amostra.

De acordo com Silva, Pessoa e Lila (2002), as estimativas pontuais de medidas descritivas da população são influenciadas pelos pesos distintos das observações. Já as estimativas de variância e desvio padrão, ou seja, medidas de precisão dos estimadores, e as estimativas de parâmetros para ajustes de alguns tipos de modelos são influenciadas pela estratificação, conglomeração e ponderação. Quando esses aspectos são ignorados, as técnicas e sistemas de análise comuns podem produzir resultados errôneos tanto para as estimativas pontuais como para os respectivos desvios padrões e níveis de significância, o que pode comprometer a qualidade do ajuste de modelos e a interpretação dos resultados obtidos.

O objetivo deste trabalho é apresentar uma revisão bibliográfica das metodologias

encontradas na literatura, especialmente de Skinner, Holt e Smith (1989), Silva (1996) e Lehtonen e Pahkinen (2004), de como são aplicados os métodos para o uso analítico de dados provenientes de pesquisas com esquemas amostrais complexos. Objetiva-se também, mostrar e discutir sobre alguns estudos que avaliam o impacto de ignorar o plano amostral na análise dos dados.

Primeiramente, é feita uma revisão bibliográfica dos principais planos amostrais, dando ênfase no cálculo dos pesos de cada elemento contido na amostra. Aqui, os desenhos amostrais são tratados separadamente, mas no caso das grandes pesquisas, inclusive as que serão tratadas e discutidas ao longo desse trabalho, é comum encontrar combinações de esquemas complexos, tornando ainda mais complicado o desenho amostral.

Depois, é abordado o conceito de Efeito do Plano Amostral, que permite avaliar o impacto de ignorar a estrutura complexa do desenho amostral sobre a estimativa da variância de um estimador durante o planejamento da pesquisa. Outro conceito abordado é o de Efeito do Plano Amostral Ampliado, proposto por Skinner, Holt e Smith (1989), que também permite avaliar o impacto de ignorar a estrutura complexa do desenho amostral sobre a estimativa da variância de um estimador, porém, no estágio da análise dos dados.

Também é abordado o método da Máxima Pseudo-Verossimilhança, que é um dos métodos usados para incorporar os pesos e o plano amostral na estimativa de parâmetros de modelos de regressão e na estimativa da variância desses parâmetros. Essa metodologia é aplicada para a obtenção de estimativas em modelos de regressão linear normal e regressão logística. Posteriormente, é abordado a análise de dados complexos em tabelas de uma e duas entradas mostrando como testes de homogeneidade e independência, por exemplo, podem ser modificados a fim de corrigir as estatísticas de teste.

Posteriormente, são discutidos alguns dos principais e mais relevantes artigos encontrados nessa área de pesquisa na opinião da autora do presente trabalho. O objetivo é mostrar, segundo o trabalho de alguns autores, o impacto que a pressuposição de que os elementos contidos numa amostra proveniente de um plano amostral complexo são independentes pode trazer.

Por fim, são feitas algumas considerações e sintetizadas algumas conclusões sobre a revisão e as metodologias mostradas para incorporar o esquema amostral na análise analítica dos dados.

2 REVISÃO BIBLIOGRÁFICA

2.1 Planos Amostrais

De acordo com Scheaffer *et al.* (2011), amostragem probabilística é a base fundamental para toda pesquisa de opinião de grande porte. O princípio básico é que, selecionada aleatoriamente, uma pequena amostra ou porcentagem de uma população pode representar suas atitudes, opiniões ou projetar o comportamento de todas as pessoas, se for selecionada corretamente.

O método mais simples para a seleção de uma amostra é a Amostragem Aleatória Simples (AAS), porém, em muitos casos, a estrutura populacional não permite o uso deste método de seleção. Em outras palavras, a estimação dos parâmetros fica comprometida, produzindo resultados e conclusões errôneas sobre a população. Assim, nesses casos, é conveniente o uso de outros métodos de seleção de amostras.

Nesta seção, serão mostrados alguns dos principais planos amostrais utilizados para a seleção de uma amostra representativa da população. Em cada plano, serão abordados alguns dos tópicos teóricos mais importantes para o desenvolvimento deste trabalho. Uma teoria mais detalhada, bem como um estudo mais profundo do conteúdo, podem ser encontrados em: Bellhouse (1988), que fala sobre a história da amostragem; Cochran (1977), Hansen, Hurwitz e Madow (1953), Kish (1965) e Murthy *et al.* (1967), que abordam a amostragem de forma teórica; FAO (1989), Hendricks (1956) e Sukhatme (1957), que tratam da amostragem na agricultura, floresta e pesca; Thompson (1992), que tratam da amostragem na geologia e biologia; Deming (1960), que aborda a amostragem no contexto dos negócios; e Fienberg e Tanur (1983), Moser, Kalton *et al.* (1971), Sudman (1976), que mostram o uso da amostragem em sociologia.

2.1.1 Amostragem Aleatória Simples

Considerando, primeiramente, um sistema de referências completo descrevendo cada unidade elementar, ou seja, um universo bem listado da forma

$$\mathcal{U} = \{1, 2, \ldots, N\},$$

a Amostragem Aleatória Simples é o processo de amostragem probabilística em que as combinações de n diferentes elementos, dos N que compõem a população, possuem igual probabilidade de estar na amostra sorteada (COCHRAN, 1977).

Além disso, a amostragem aleatória simples pode ser feita com reposição, caso em que cada elemento da população pode entrar mais do que uma vez na amostra, denotada por AASc, ou sem reposição, caso em que cada elemento da população só pode entrar uma vez na amostra, denotada por AASs.

2.1.1.1 Amostragem Aleatória Simples com Reposição

Seguindo a notação de Bolfarine e Bussab (2005), sejam $f_i(s)$ a variável que indica o número de vezes, ou seja, a frequência que a i-ésima unidade populacional aparece na amostra s, π_i a probabilidade de o i-ésimo elemento de $\mathcal{U}$ pertencer à amostra e π_{ij} a probabilidade do i-ésimo e do j-ésimo elementos de $\mathcal{U}$ pertencerem simultâneamente à amostra.

Para o plano amostral AASc, a variável f_i tem distribuição binomial (MORETTIN; BUSSAB, 2004) com parêmetros n e $p = 1/N$, denotada por

$$f_i \sim B\left(n; \frac{1}{N}\right)$$

de modo que

$$E[f_i] = np = \frac{n}{N},$$

$$Var[f_i] = np(1-p) = \frac{n}{N}\left(1 - \frac{1}{N}\right),$$

$$\pi_i = P(f_i \neq 0) = 1 - P(f_i = 0) = 1 - \begin{pmatrix} n \\ 0 \end{pmatrix}\left(1 - \frac{1}{N}\right)^n = 1 - \left(1 - \frac{1}{N}\right)^n,$$

e

$$\begin{aligned}
\pi_{ij} &= P(f_i \neq 0 \cap f_j \neq 0) \\
&= 1 - P(f_i = 0 \cup f_j = 0) \\
&= 1 - [P(f_i = 0) + P(f_j = 0) - P(f_i = 0 \cap f_j = 0)] \\
&= 1 - \left[\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n\right] \\
&= 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n.
\end{aligned}$$

2.1.1.2 Amostragem Aleatória Simples sem Reposição

Para o plano amostral AASs, a variável f_i segue a distribuição de Bernoulli (MORETTIN; BUSSAB, 2004) com probabilidade de sucesso $p = n/N$, denotada por

$$f_i \sim Bernoulli\left(\frac{n}{N}\right)$$

satisfazendo

$$P(f_i = 1) = \frac{n}{N} \quad e \quad P(f_i = 0) = 1 - \frac{n}{N}$$

de modo que

$$E[f_i] = p = \frac{n}{N},$$

$$Var[f_i] = p(1-p) = \frac{n}{N}\left(1 - \frac{n}{N}\right),$$

$$\begin{aligned}
\pi_i &= P(f_i \neq 0) = 1 - P(f_i = 0) \\
&= 1 - \binom{1}{0}\left(1 - \frac{n}{N}\right)^0 \left(1 - \frac{n}{N}\right)^1 \\
&= 1 - \left(1 - \frac{n}{N}\right) \\
&= \frac{n}{N}
\end{aligned}$$

e

$$\begin{aligned}
\pi_{ij} &= P(f_i \neq 0 \cap f_j \neq 0) \\
&= 1 - P(f_i = 0 \cup f_j = 0) \\
&= 1 - [P(f_i = 0) + P(f_j = 0) - P(f_i = 0 \cap f_j = 0)] \\
&= \frac{n}{N}\frac{n-1}{N-1}.
\end{aligned}$$

2.1.2 Amostragem Estratificada

Quando é possível identificar subpopulações que variam muito entre si no que diz respeito a variável em estudo, mas que variam pouco dentro de si, uma amostra estratificada pode fornecer resultados mais precisos do que uma AAS extraída da população. De acordo com Murthy *et al.* (1967), um dos principais motivos para o uso da estratificação é o fato de que ela pode proporcionar um aumento significativo de precisão nas estimativas dos parâmetros populacionais e o resultado será mais eficaz quanto maior for a habilidade do pesquisador em identificar estratos homogêneos.

Sendo assim, este método consiste em dividir a população em grupos relativamente homogêneos e mutuamente exclusivos, chamados estratos, e em selecionar amostras aleatórias simples e independentes dentro de cada estrato. Para cada amostra, usam-se estimadores convenientes para os parâmetros do estrato. Posteriormente, usa-se para a população um estimador combinando os estimadores de cada estrato.

Se o número de elementos de cada amostra for proporcional ao tamanho do estrato, as observações podem ser misturadas para se obter os resultados globais. Se, no entanto, todas as amostras tiverem o mesmo número de elementos, os resultados de cada estrato tem que ser ponderados pela proporção desse estrato na população.

Em termos de notação, e seguindo a que foi proposta por Bolfarine e Bussab (2005), considere novamente uma população bem descrita por um sistema de referência

$$\mathcal{U} = \{1, 2, \ldots, N\}$$

de tal forma que exista uma partição $\mathcal{U}_1, \ldots, \mathcal{U}_H$ de $\mathcal{U}$ tal que

$$\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h \quad e \quad \mathcal{U}_h \bigcap \mathcal{U}_{h'} = \emptyset$$

para $h \neq h' = 1, \ldots, H$, e que cada subconjunto $\mathcal{U}_h$, bem determinado, é identificado por duplas ordenadas, do seguinte modo:

$$\mathcal{U}_h = \{(h,1),(h,2),\ldots,(h,N_h)\}.$$

em que H é o número de estratos e N_h é o número de elementos dentro de cada estrato.

Assim, o universo todo pode ser escrito por

$$\mathcal{U} = \{(1,1),\ldots,(1,N_1),\ldots,(h,1),\ldots,(h,i),\ldots,(h,N_h),\ldots,(H,1),\ldots,(H,N_H)\}$$

de modo a facilitar a identificação do estrato e do elemento dentro dele.

Agora, sejam N_h e h como definidos acima e n_h o tamanho da amostra selecionada dentro do estrato h. Se o método de seleção usado para retirar as amostras dos estratos for a amostragem aleatória simples, então a probabilidade de inclusão da unidade i do estrato h e a probabilidade das unidades i e j na amostra são dadas, respectivamente, por

$$\pi_{ih} = 1 - \left(1 - \frac{1}{N_h}\right)_h^n, \quad i = 1, \ldots, N_h$$

e

$$\pi_{ijh} = 1 - 2\left(1 - \frac{1}{N_h}\right)_h^n + \left(1 - \frac{2}{N_h}\right)_h^n, \quad i,j \in \{1,\ldots,N_h\}$$

no caso da amostragem aleatória simples com reposição, e por

$$\pi_{ih} = \frac{n_h}{N_h}, \quad i = 1, \ldots, N_h$$

e

$$\pi_{ijh} = \frac{n_h}{N_h}\frac{n_h - 1}{N_h - 1}, \quad i \neq j \in \{1,\ldots,N_h\}$$

no caso da amostragem aleatória simples sem reposição.

Exemplos de pesquisas utilizando este tipo de planejamento amostral incluem Benedetti *et al.* (2008), Garcia (1986), Netto e Sanquetta (1994) e Tavares *et al.* (2007).

2.1.3 Amostragem por Conglomerados

Os dois planos amostrais vistos anteriormente sorteavam unidades elementares diretamente da população ou de estratos dessa mesma população. Porém, em algumas situações o sistema de referência não possui uma estrutura que permite o uso desses planos. Nestes casos, é conveniente tomar como unidade elementar grupos, chamados de conglomerados.

Hansen, Hurwitz e Madow (1953) definem Amostragem por Conglomerados como um plano amostral em que: (a) os elementos populacionais a serem amostrados são agrupados dentro de conglomerados, de tal forma que cada elemento populacional está associado a um, e

somente um, conglomerado; (b) no planejamento amostral, os conglomerados, que passam a ser as unidades amostrais, são selecionados por Amostragem Aleatória Simples.

Uma das inconveniências para o uso da amostragem de conglomerados prende-se ao fato de que as unidades, dentro de um mesmo conglomerado, tendem a ter valores parecidos em relação às variáveis que estão sendo pesquisadas, e isso torna estes planos menos eficientes, pois um bom conglomerado deve ser uma micro representação do universo, ou seja, ter todo tipo de participante e não do mesmo tipo. Segundo Kish (1965), se a amostragem de conglomerados for comparada com a amostragem de elementos de mesmo tamanho, notar-se-á que na primeira: (1) o custo por elemento é menor; (2) a variância é maior; (3) o custo e os problemas com as análises estatísticas são maiores.

Dado que a eficiência de um conglomerado depende do grau de similaridade de seus elementos, faz sentido o uso de uma medida que indique esse grau de eficiência. Na literatura existem várias propostas para tais medidas, sendo que a mais comum entre elas é o coeficiente de correlação intraclasse, que pode assumir valores entre 1 (caso de homogeneidade máxima dentro dos conglomerados) e $1/(1-B)$, em que B é o tamanho do conglomerado, caso todos eles tenham tamanho iguais (BOLFARINE; BUSSAB, 2005).

De acordo com Bolfarine e Bussab (2005), quando o coeficiente de correlação intraclasse for alto, isto é, quando os conglomerados são homogêneos, torna-se menos recomendável o uso da amostragem por conglomerados completos, ou seja, a coleta de todas as suas unidades. Dado que as unidades são parecidas, pode-se dizer que elas trarão o mesmo tipo de informação, aumentando a variação amostral. Logo, uma alternativa para elevar a eficiência, sem aumentar o tamanho amostral, é sub sortear elementos dos conglomerados selecionados. Procedendo dessa forma, tem-se uma amostragem em dois estágios: no primeiro é feito o sorteio dos conglomerados, unidades primárias, e no segundo o sorteio dos elementos.

Numa amostragem por conglomerados em dois estágios, por exemplo, pode ocorrer que a seleção das unidades amostrais no primeiro estágio ocorra por amostragem aleatória simples com reposição e no segundo estágio sem reposição, ou outras combinações possíveis. Assim, a probabilidade de inclusão da unidade i na amostra dependerá de como foram realizadas as seleções dentro de cada estágio.

Exemplos de pesquisas utilizando este tipo de planejamento amostral incluem Barata *et al.* (2005), Sousa e Silva (2003), Filho *et al.* (2001) e Cordeiro (2001).

2.2 Efeitos do Plano Amostral

2.2.1 Efeito do Plano Amostral (EPA) de Kish

O impacto ao se fazer estimativas pontuais e construir intervalos de confiança para essas estimativas, por exemplo, pode ser muito grande quando não incorporamos o esquema amostral

na análise. Neste contexto, Kish (1965) propôs uma medida chamada Efeito do Plano Amostral (EPA) para medir o efeito do planejamento amostral sobre a variância de um estimador no estágio de planejamento da pesquisa.

O EPA de Kish é uma razão entre variâncias de um estimador calculadas para dois planos amostrais diferentes para os quais pretende-se fazer a comparação de ganhos ou perdas de precisão. Na maioria das vezes, o plano amostral de referência é a AAS. Sendo assim, o EPA de Kish para um estimador $\hat{\theta}$ é definido como:

$$EPA_{Kish}(\hat{\theta}) = \frac{V_{VERD}(\hat{\theta})}{V_{AAS}(\hat{\theta})}$$

em que $V_{VERD}(\hat{\theta})$ é a variância do estimador $\hat{\theta}$ calculada baseada no plano amostral complexo (verdadeiro) e $V_{AAS}(\hat{\theta})$ é a variância do estimador $\hat{\theta}$ calculada baseada em uma Amostragem Aleatória Simples.

O cálculo do EPA de Kish é usado para planejar pesquisas amostrais pois permite comparar e antecipar o impacto do uso de alguns esquemas amostrais alternativos sobre a precisão de estimadores de várias variáveis relevantes à pesquisa.

2.2.2 Efeito do Plano Amostral Ampliado

O cálculo do EPA de Kish tem pouca utilidade para estudos analíticos, pois seu uso se dá no estágio de planejamento da pesquisa. Logo, uma vez que a amostra já está selecionada, não faz sentido comparar a eficiência do plano amostral utilizado com outro.

Para resolver este problema, uma nova abordagem foi proposta por Skinner, Holt e Smith (1989), que tem por finalidade avaliar a tendência de um estimador consistente da variância de $\hat{\theta}$, calculado sob hipótese de que as observações são independentes (ou equivalentemente provenientes de uma AASc), subestimar ou superestimar a variância verdadeira do estimador pontual.

Esta medida é chamada de Efeito do Plano Amostral Ampliado e é definida por:

$$EPA(\hat{\theta}, v_0) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)} \tag{2.1}$$

em que $v_0 = \hat{V}_{IID}(\hat{\theta})$ é um estimador consistente da variância de $\hat{\theta}$ assumindo a hipótese de que as observações são independentes e identicamente distribuídas, $V_{VERD}(\hat{\theta})$ é a variância do estimador sobre o plano amostral utilizado e $E_{VERD}(v_0)$ é a esperança do estimador consistente sob o plano amostral utilizado. Quanto mais afastado de 1 for o valor de $EPA(\hat{\theta}, v_0)$, mais incorreta será considerada a especificação do plano amostral ou do modelo.

Segundo Pessoa e Silva (1998), são esperadas, em geral, as seguintes consequências sobre o EPA ao ignorar o plano amostral efetivamente adotado e admitir que a seleção da amostra foi AAS:

i) Ignorar os pesos em v_0 pode inflacionar o EPA;

ii) Ignorar conglomeração em v_0 pode inflacionar o EPA;

iii) Ignorar estratificação em v_0 pode reduzir o EPA;

Com o objetivo de ilustrar a diferença entre $EPA_{Kish}(\hat{\theta})$ e $EPA(\hat{\theta}, v_0)$, Skinner, Holt e Smith (1989) consideraram uma população com conglomerados de tamanho 2 (exemplos de população com essa estrutura são pares de irmão gêmeos, casais e observações emparelhadas da aplicação de um tratamento).

Suponha que uma variável de pesquisa y medida nesses pares de unidades tem média θ, variância σ^2 e correlação intraclasse ρ. Além disso, suponha também que um conglomerado é selecionado aleatoriamente e são anotados os valores de y_1 e y_2 das duas unidades dentro do conglomerado.

Então, temos:

$$E(y_i) = \theta, \quad V(y_i) = \sigma^2 \quad e \quad CORR(y_1, y_2) = \rho, \quad i = 1, 2.$$

Aqui, o objetivo é verificar o impacto de ignorar a conglomeração da estrutura dos dados. Sendo assim, considere, primeiramente, um estimador não viciado para a média θ. Este estimador é dado por

$$\hat{\theta} = \frac{y_1 + y_2}{2}$$

e, baseado na falsa suposição de que o plano amostral seja AAS, tem variância

$$Var(\hat{\theta}) = \frac{\sigma^2}{2}$$

com um estimador não viciado dado por

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

Porém, a variância verdadeira (considerando a conglomeração do plano amostral) de $\hat{\theta}$ é

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

e o valor esperado do estimador de variância v_0 sob o plano conglomerado é

$$E_{VERD}(v_0) = \frac{\sigma^2(1-\rho)}{2}.$$

Assim, tem-se que

$$EPA_{kish} = \frac{\frac{\sigma^2(1+\rho)}{2}}{\frac{\sigma^2}{2}} = 1 + \rho$$

e

$$EPA(\hat{\theta}, v_0) = \frac{\frac{\sigma^2(1+\rho)}{2}}{\frac{\sigma^2(1-\rho)}{2}} = \frac{1+\rho}{1-\rho}.$$

Se, por exemplo, $\rho = 0,8$, então $EPA_{kish} = 1,8$, mostrando que a verdadeira variância de $\hat{\theta}$ é 80% maior do que teria sido de o plano amostral AAS fosse usado, ou seja, o efeito da especificação incorreta do plano amostral pode ser grande e trazer consequências graves.

Neste exemplo, foi usado como parâmetro a média estimada $\hat{\theta}$. Porém, outra abordagem é considerar como parâmetros as estimativas dos coeficientes de um modelo.

Cunha, Araújo e Lima (2015) realizaram um estudo cujo objetivo foi verificar que fatores são mais relevantes para explicar o desemprego e a inatividade de jovens entre 16 e 29 anos de idade no Brasil metropolitano. Para isso, ajustaram um modelo de logitos generalizados considerando as três categorias em que o jovem poderia se encontrar: inativo, ativo e empregado e ativo e desempregado. Os dados utilizados são da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007, cuja caracteristíca é incorporar todos os aspectos que definem um plano amostral complexo: estratificação das unidades de amostragem, conglomeração (seleção da amostra em vários estágios, com unidades compostas de amostragem) e probabilidades desiguais de seleção em um ou mais estágios.

As características individuais que podem afetar as probabilidades de desemprego e inatividade, consideradas neste estudo, foram idade, raça, gênero, escolaridade, experiência, renda familiar e condição do indivíduo na família.

A Tabela 2.1 apresenta a distribuição dos jovens conforme a situação ocupacional e a Tabela 2.2 mostra as estimativas dos coeficiêntes do modelo e do efeito do plano amostral. A categoria utilizada como base nesse ajuste foi a de empregados.

Tabela 2.1 – Distribuição de frequência relativa dos jovens por situação ocupacional

Situação ocupacional	Número de jovens	Frequência relativa
Inativo	8.930	0,26
Empregado	20.609	0,60
Desempregado	4,809	0,14
Total	34.348	1,00

Fonte: Cunha, Araújo e Lima (2015)

As estimativas do efeito do plano amostral EPA_{Kish} e $EPA(\hat{\theta}, v_0)$ indicaram que sua correta especificação foi fundamental para a obtenção de estimativas robustas e não viciadas. No caso da variável Renda, por exemplo, $EPA(\hat{\theta}, v_0) = 3,0234$, indicando que a estimativa da variância do estimador da variável Renda baseada na hipótese IID superestimou em aproximadamente 200% a verdadeira estimativa da variância. Enquanto isso, para essa mesma

Tabela 2.2 – Estimativas dos coeficientes dos logitos relacionados às categorias inatividade e desemprego

	Inatividade			Desemprego		
	Coeficiente	EPA_{Kish}	$EPA(\hat{\theta}, v_0)$	Coeficiente	EPA_{Kish}	$EPA(\hat{\theta}, v_0)$
Intercepto	13,7050	1,5199	1,5145	4,2676	1,3602	1,3478
Homem	-0,4986	1,5277	1,3846	-0,2718	1,4903	1,4885
Raça	0,0005	1,4920	1,4319	-0,1018	1,4638	1,4279
Idade	-1,1804	1,5104	1,5249	-0,5158	1,3575	1,3501
Idade2	0,0291	1,4978	1,5485	0,0146	1,3595	1,3732
Estudante	0,6530	1,5390	1,3614	0,1363	1,5233	1,5079
Escolaridade	-0,1331	1,3183	1,4122	0,0960	1,2980	1,5574
Escolaridade2	-0,0015	1,3239	1,4317	-0,0077	1,3037	1,5207
Experiência	-0,4837	1,4985	2,7154	-0,3408	1,3734	2,0550
Filho	-0,0065	1,3126	1,4074	0,2398	1,3984	1,4996
Cônjuge	0,9836	1,3641	1,5933	0,4816	1,3954	1,5607
Outros	-0,0244	1,2102	1,2468	0,2050	1,3602	1,4730
Renda	0,0000	1,2846	1,5733	-0,0008	1,6442	3,0234

Fonte: Cunha, Araújo e Lima (2015)

variável, $EPA_{Kish} = 1,6442$, expressando o efeito que o uso de um plano amostral diferente de AASc tem na variância do estimador, ou ainda o erro que se comete ao usar um estimador "ingênuo" da variância do estimador sob um plano amostral diferente de AASc.

2.2.3 Efeito Multivariado de Plano Amostral

Nas sessões anteriores, $\hat{\theta}$ diz respeito ao estimador de um parâmetro univariado. Considerando, agora, $\hat{\boldsymbol{\theta}}$ o estimador do vetor de p parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ e $\boldsymbol{V}_0$ um estimador da matriz $(p \times p)$ de covariância de $\hat{\boldsymbol{\theta}}$ baseado na suposição de que as observações são IID, pode-se definir uma generalização da equação (2.1), denominado Efeito Multivariado do Plano Amostral de $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{V}_0$, por (SKINNER; HOLT; SMITH, 1989):

$$EMPA(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_0) = \boldsymbol{D} = E_{VERD}(\boldsymbol{V}_0)^{-1}\boldsymbol{V}_{VERD}(\hat{\boldsymbol{\theta}})$$

em que $E_{VERD}(\boldsymbol{V}_0)$ é o valor esperado de $\boldsymbol{V}_0$ e $\boldsymbol{V}_{VERD}(\hat{\boldsymbol{\theta}})$ é a matriz de covariância de $\hat{\boldsymbol{\theta}}$, ambas calculadas com respeito à distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado.

Os autovalores $\lambda_1 \geq \ldots \geq \lambda_p$ de $\boldsymbol{D}$ são chamados de efeitos generalizados do plano amostral e terão grande importância na teoria apresentada na Sessão 2.4.

2.2.4 Estimação de Variâncias

Numa pesquisa por amostragem é inerente a existência de um erro amostral pelo fato de estar analisando apenas parte da população do estudo. Assim, é recomendável que se calcule,

além das estimativas dos parâmetros de interesse, uma medida de precisão dessas estimativas. Logo, a estimação de variâncias tem suma importância na abordagem inferencial adotada, pois se não houver estimativas de variâncias, não há como obter uma indicação da precisão (e portanto, da qualidade) das estimativas de interesse.

A estimação de variâncias é amplamente tratada nos livros de amostragem, porém, um detalhe é que geralmente são estimadores de variância de estimadores de totais, de médias, de proporções etc, isto é, quando os estimadores são lineares nas observações amostrais, não viciados e todas as probabilidades de inclusão conjuntas são não nulas. Entretanto, para alguns planos amostrais utilizados na prática, as probabilidades de inclusão conjuntas podem ser nulas, caso da amostragem sistemática, ou difíceis de calcular, caso de alguns esquemas de seleção com probabilidades desiguais. Além disso, em outros casos, os parâmetros de interesse são não-lineares como, por exemplo, coeficientes de regressão, fazendo com que as estatísticas usadas para estimar tais parâmetros também sejam não-lineares. Assim, é necessário o emprego de técnicas especiais para a estimação de $V_{VERD}(\hat{\theta})$ em (2.1), por exemplo.

Duas dessas técnicas são o Método do Conglomerado Primário, o qual é largamente utilizado para estimação de variância de estimativas provenientes de amostras com planos amostrais complexos, e o Método de Linearização de Taylor, que é empregado para obter estimativas aproximadas de variância para estimadores não-lineares de quantidades populacionais de interesse.

2.2.4.1 Método do Conglomerado Primário

De acordo com Lila (2002), o termo Conglomerado Primário, proposto por Hansen, Hurwitz e Madow (1953), é utilizado para representar o conjunto de unidades incluídas na amostra contidas em uma mesma unidade primária de amostragem. A definição de conglomerado primário torna-se útil quando são empregados planos amostrais com vários estágios de seleção.

A ideia central deste método é considerar apenas a variação entre informações disponíveis ao nível das unidades primárias de amostragem (UPAs), isto é, ao nível dos conglomerados primários e admitir que estes teriam sido selecionados com reposição da população. Embora algumas vezes a seleção das UPAs seja feita sem reposição, devido ao fato de ser mais eficiente que em planos com reposição de igual tamanho, o estimador de conglomerados primários pode fornecer uma aproximação razoável da correspondente variância de aleatorização.

Pessoa e Silva (1998) explicam que essa é uma ideia simples, entretanto muito poderosa, pois permite acomodar uma grande variedade de planos amostrais, envolvendo estratificação e seleção com probabilidades desiguais tanto das unidades primárias como das demais unidades de amostragem. Os requisitos fundamentais para a aplicação deste método é que estejam disponíveis estimadores não viciados dos totais da variável de interesse para cada um dos conglomerados primários selecionados, e que pelo menos dois destes sejam selecionados em cada estrato (se a

amostra for estratificada no primeiro estágio).

De acordo com Hansen, Hurwitz e Madow (1953), esta abordagem fornece uma estimativa consistente da variância para amostras que possuem duas ou mais unidades de primeiro estágio selecionadas em cada estrato.

Considere um plano amostral em vários estágios, no qual n_h unidades primárias são selecionadas no estrato h, $h = 1, 2, \ldots, H$. Denotando por π_{hi} a probabilidade de inclusão na amostra da unidade primária de amostragem (conglomerado primário) i do estrato h, e por $\hat{Y}_{hi}$ um estimador não viciado do total Y_{hi} da variável de pesquisa y no i-ésimo conglomerado primário do estrato h. Usando o Método do Conglomerado Primário, segue que um estimador não viciado do total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$ da variável de pesquisa y na população é dado por

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

e um estimador não viciado da variância de aleatorização correspondente por

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

em que o índice p se refere ao operador de variância referente à distribuição de probabilidades $p(s)$ induzida pelo planejamento amostral, e $\hat{Y}_h = \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{\pi_{hi}}$ para $h = 1, 2, \ldots, H$.

2.2.4.2 Método da Linearização de Taylor

Em pesquisas por amostragem, deseja-se, em algumas ocasiões, utilizar estimadores não-lineares para estimar certas quantidades de interesse. Razões, diferenças de razões, coeficientes de correlação e coeficientes de regressão são exemplos comuns de parâmetros que requerem emprego de estimadores não-lineares. Porém, expressões exatas para as variâncias de estimadores amostrais destas quantidades são difíceis de obter, além de não existirem estimadores exatamente não-tendenciosos de variância.

O Método da Linearização de Taylor é um método baseado na expansão em séries de Taylor (GUIDORIZZI, 1995) frequentemente usado para a estimação da variância de um estimador não-linear, cujo objetivo é definir uma estatística linearizada e aproximar a variância da estatística amostral de interesse pela variância da estatística linearizada sob o plano amostral adotado. Além disso, este método combinado com o Método do Conglomerado Primário permite a estimação da variância de estimadores não-lineares desde que estes possam ser escritos como função de estimadores não-viciados de totais populacionais.

Para apresentar o método, considere o vetor $\mathbf{Y} = \sum_{i=1}^{N} \mathbf{y}_i$ de totais das r variáveis da pesquisa na população, a partir de uma amostra observada s. Um estimador para o total $\mathbf{Y}$

é o estimador π-ponderado, devido a Horvitz e Thompson (1952), dado por (SÄRNDAL; SWENSSON; WRETMAN, 2003), que pode ser escrito como

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

e sua variância de aleatorização é dada por

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

com estimador dado por

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

Agora, considere um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ escrito da forma $\boldsymbol{\theta} = g(\mathbf{Y})$. Um estimador para $\boldsymbol{\theta}$ poderia ser dado por

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

Em notação matricial, $\boldsymbol{\theta}$ e $\hat{\boldsymbol{\theta}}$ podem ser escritos como

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

respectivamente, em que $\mathbf{A}$ é uma matriz $(k \times r)$ de constantes. Este estimador é não-viciado e tem variância

$$V_p(\hat{\boldsymbol{\theta}}) = \mathbf{A} V_p(\hat{\mathbf{Y}}_\pi) \mathbf{A}'. \tag{2.2}$$

Porém, quando g é não-linear, a ideia é considerar a expansão de $g(\hat{\mathbf{Y}}_\pi)$ em torno de $\mathbf{Y}$ até o termo de primeira ordem desprezando o resto, dada por

$$\hat{\boldsymbol{\theta}} \simeq \hat{\boldsymbol{\theta}}_{LT} = g(\mathbf{Y}) + \Delta g(\mathbf{Y})(\hat{\mathbf{Y}}_\pi - \mathbf{Y}), \tag{2.3}$$

em que $\Delta g(\mathbf{Y})$ é a matriz Jacobiana $k \times r$ cuja c-ésima coluna é $\dfrac{\partial g(\mathbf{Y})}{\partial Y_c}$, para $c = 1, \ldots, r$ (MASSAGO, 2010).

Tomando as variâncias de aleatorização dos dois lados em (2.3) e notando que no lado direito o único termo que tem variância de aleatorização $\Delta g(\mathbf{Y})(\hat{\mathbf{Y}}_\pi - \mathbf{Y})$ é uma função linear de $\hat{\mathbf{Y}}_\pi$, segue que (PESSOA; SILVA, 1998)

$$V_p(\hat{\boldsymbol{\theta}}) \simeq \Delta g(\mathbf{Y}) V_p(\hat{\mathbf{Y}}_\pi) \Delta g(\mathbf{Y})'$$

e

$$\hat{V}_p(\hat{\boldsymbol{\theta}}) = \Delta g(\hat{\mathbf{Y}}) \hat{V}_p(\hat{\mathbf{Y}}_\pi) \Delta g(\hat{\mathbf{Y}})'.$$

Uma das vantagens da aplicação deste método é que se as derivadas parciais são conhecidas, a linearização quase sempre fornece uma estimativa de variância que pode ser aplicada em todos os tipos de planos amostrais. Já as desvantagens devem-se ao fato de que algumas vezes os cálculos podem ser muito trabalhosos e o método pode ser complicado quando for aplicado a funções complexas. Além disso, nem todos os tipos de estatísticas podem ser expressos como funções de totais populacionais.

Mais detalhes sobre este método, bem como vários outros utilizados para estimar variâncias em pesquisas amostrais complexas, podem ser encontrados em Wolter (1985) e Korn e Graubard (2011).

2.3 Modelos de Regressão Incorporando o Esquema Amostral

Diante do exposto anteriormente, fica claro que há medidas para avaliar o impacto do plano amostral antes ou depois da coleta de dados. Porém, essas medidas não solucionam o problema de como incorporar o desenho amostral utilizado nas inferências analíticas. Uma maneira de contornar esse problema seria incorporar informações populacionais auxiliares na modelagem. Duarte (1999) explica que essas informações auxiliares podem ser incorporadas de duas maneiras no ajuste do modelo linear.

Sejam $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})'$ um vetor de variáveis explicativas com p elementos, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})'$ um vetor de variáveis dependentes com r elementos e $\mathbf{x}_i$ um vetor que contém q variáveis auxiliares para o elemento i da amostra, $i \in \{1, \ldots, N\}$. Além disso, considere Y_i, Z_i e X_i como sendo as variáveis aleatórias que geram $\mathbf{y}_i$, $\mathbf{z}_i$ e $\mathbf{x}_i$, respectivamente.

A primeira maneira de incorporar informações auxiliares no modelo é a Abordagem Desagregada que incorpora as variáveis auxiliares diretamente, especificando-se um modelo supondo que a esperança condicional de Y_i dadas as variáveis explicativas Z_i e as variáveis auxiliares X_i é da forma

$$E_\xi(Y_i|Z_i = z_i, X_i = x_i) = \beta_0 + z_i'\beta + x_i'\eta. \tag{2.4}$$

A segunda maneira é a chamada Abordagem Agregada, na qual especifica-se um modelo supondo que a esperança condicional de Y_i dadas as variáveis explicativas Z_i é da forma

$$E_\xi(Y_i|Z_i = z_i) = \beta_0 + z_i'\beta. \tag{2.5}$$

Neste caso, as variáveis auxiliares são usadas na estimação dos parâmetros e não incorporadas diretamente no modelo. Silva (1996) faz algumas considerações sobre a abordagem (2.4) e algumas de suas dificuldades.

Dentro da abordagem (2.5) ainda existem dois tipos de inferência: a descritiva e a analítica. Na inferência descritiva, o interesse é estimar algum parâmetro que seja uma função conhecida de valores relacionados às N unidades da população, e pode ser subdividida em outras duas frentes.

A primeira é a abordagem clássica, que baseia-se exclusivamente na distribuição amostral, sendo que a única relação existente entre as unidades amostrais e as de fora da amostra é a existência de uma probabilidade positiva de que todas as observações pertençam à amostra.

A segunda é baseada numa teoria que emprega um modelo estocástico para representar a estrutura da população e as inferências são feitas baseadas na distribuição de probabilidades especificada ou baseada no modelo. Um problema encontrado nessa abordagem, estudada por Fuller (1975), é o fato de que as inferências são feitas baseadas no modelo, não levando em conta o desenho amostral da pesquisa.

Na inferência analítica, de acordo com Duarte (1999), o parâmetro a ser estimado não é função dos valores de todas as unidades populacionais, logo não é fácil formular uma teoria satisfatória empregando apenas a distribuição baseada no desenho amostral. Precisa-se de uma hipótese que relacione a população finita com o parâmetro analítico de interesse. Geralmente, assume-se que os valores da população finita correspondem a uma amostra aleatória de uma superpopulação (infinita). Assim, como o tamanho N dessa população finita é, na maioria das vezes, muito grande, um estimador da população finita baseado nos N valores seria bem próximo do parâmetro desconhecido da superpopulação. Logo, considera-se o estimador da população finita como um parâmetro da população finita que poderá ser estimado usando a distribuição baseada no desenho amostral.

Em outras palavras, como explica Ravines (2003), o processo de inferência estatística a partir de uma amostra compreende um conjunto de princípios e procedimentos que podem envolver, por exemplo, o conhecimento de algum processo aleatório que possa ter gerado o verdadeiro valor desconhecido da característica de interesse, y, para cada unidade da população. Esse processo é representado por um modelo que é utilizado como base para a realização de inferências. Esta abordagem é denominada Modelos de Superpopulação.

O termo superpopulação refere-se, então, ao modelo $f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ que especifica a distribuição conjunta dos valores da variável de interesse $\mathbf{y}$ na população, isto é, a distribuição conjunta de $\mathbf{y} = (y_1, y_2, \ldots, y_N)$. A inferência analítica em dados amostrais refere-se à inferência sobre o vetor de parâmetros $\boldsymbol{\theta}$ do modelo de superpopulação. Autores que utilizaram a abordagem de modelos de superpopulação incluem Brewer (1963), Godambe (1955), Isaki (1970), Madow e Madow (1944) e Yates *et al.* (1960).

Aqui, será tratado o problema de como fazer inferências analíticas com dados de pesquisas que envolvem esquemas amostrais complexos. Mais precisamente, mostra como estruturas complexas, como estratificação, conglomeração e probabilidades desiguais de seleção podem "entrar" no ajuste de um modelo de regressão.

A abordagem que será utilizada no presente trabalho para incorporar essa estrutura no modelo de regressão é a abordagem agregada, através do método de Máxima Pseudo-Verossimilhança (Binder (1983); Godambe e Thompson (1986)). Outros métodos disponíveis

na literatura para fazer inferências analíticas com dados de amostras complexas são Inferência Bayesiana (Little (2003)), Método de Verossimilhança da Amostra (Krieger e Pfeffermann (1992); Krieger e Pfeffermann (1997); Pfeffermann, Krieger e Rinott (1998a)) e Método de Verossimilhança com Informação Completa (Breckling *et al.* (1994)).

2.3.1 Método da Máxima Verossimilhança

O Método da Máxima Verossimilhança (MMV) foi introduzido por Ronald A. Fisher em 1912. O forte apelo intuitivo, sua aplicabilidade em muitos casos e a existência de propriedades desejáveis para os estimadores resultantes do método são razões para o seu amplo uso dentre os métodos de estimação.

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias com função densidade de probabilidade $f(x_i, \boldsymbol{\theta})$, em que $\boldsymbol{\theta} \in \Theta$ é um vetor de parâmetros desconhecidos de interesse. A função de verossimilhança é definida por (SAHOO, 2013)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}).$$

A ideia deste método é "olhar" para $f(x_i, \boldsymbol{\theta})$ não mais como função da amostra, mas sim como função do vetor de parâmetros desconhecidos $\boldsymbol{\theta}$, considerando a amostra observada $x_1, x_2, \ldots, x_n$ fixa. Essa função passa a ser chamada de função de verossimilhança. O Método da Máxima Verossimilhança procura estabelecer o valor mais provável para $\boldsymbol{\theta}$ que resultou nas observações $x_1, x_2, \ldots, x_n$.

Seja $L(\boldsymbol{\theta})$ a função de verossimilhança para as variáveis aleatórias $X_1, X_2, \ldots, X_n$. Se $\hat{\boldsymbol{\theta}} = \phi(x_1, x_2, \ldots, x_n)$ é um valor de $\boldsymbol{\theta}$, em Θ, que maximiza $L(\boldsymbol{\theta})$, então $\hat{\Theta} = \phi(X_1, X_2, \ldots, X_n)$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ e $\hat{\boldsymbol{\theta}}$ é a estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ para a amostra $x_1, x_2, \ldots, x_n$ (NOJOSA, 2006).

A verossimilhança expressa a plausibilidade para diferentes valores de $\boldsymbol{\theta}$ e informa a preferência por um dentre diversos valores possíveis para o parâmetro. O valor que resulta na maior verossimilhança é denominado estimativa de máxima verossimilhança.

Muitas funções de verossimilhança satisfazem condições de regularidade que permitem obter o respectivo máximo por derivação. Nestes casos, a estimativa de máxima verossimilhança resulta de

$$\frac{\partial L(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{2.6}$$

A derivação dos produtórios em (2.6) torna trabalhosa a obtenção do máximo para $L(\boldsymbol{\theta})$. Para contornar essa dificuldade, usa-se a função logarítmica que é uma função monótona e crescente, logo $L(\boldsymbol{\theta})$ e $l(\boldsymbol{\theta}) = logL(\boldsymbol{\theta})$, que é denominada função de log-verossimilhança,

tem seus máximos para o mesmo valor de $\boldsymbol{\theta}$. Assim, pode-se obter o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ através da equação

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0}, \tag{2.7}$$

em que

$$\mathbf{u}_i(\boldsymbol{\theta}) = \frac{\partial log[f(x_i, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}$$

é o vetor dos escores da unidade i da amostra. A função escore, denotada por $\mathbf{u}(\boldsymbol{\theta})$ é definida como a primeira derivada da função de log-verossimilhança com respeito a $\boldsymbol{\theta}$, isto é,

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial log[f(\mathbf{x}, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}.$$

A variância assintótica de $\hat{\boldsymbol{\theta}}$ é dada por

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

e um estimador consistente dessa variância é

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

em que

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

é denomidada função de informação, e

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

A estimação detalhada de parâmetros em modelos paramétricos bem como outros métodos de estimação e suas propriedades também é tratada em Casella e Berger (2002), Garthwaite, Jolliffe e Jones (2002), Mood (1950) e Hoel *et al.* (1954).

2.3.2 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 $i \in \mathcal{U}$, em que $\mathcal{U} = \{1, \ldots, N\}$ é o conjunto de rótulos dos elementos da população. Além disso, suponha também que $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ são vetores aleatórios independentes e identicamente distribuídos com função densidade de probabilidade comum $f(\mathbf{y}; \boldsymbol{\theta})$ tal que

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_N; \boldsymbol{\theta}) = \prod_{i \in \mathcal{U}} f(\mathbf{y}_i; \boldsymbol{\theta})$$

em que $\boldsymbol{\theta} \in \Theta$ é um vetor $(k \times 1)$ de parâmetros desconhecidos.

Se todos os elementos da população finita $\mathcal{U}$ fossem conhecidos, a função de verossimilhança e de log-verossimilhança populacionais seriam dadas,respectivamente, por

$$L_{\mathcal{U}} = \prod_{i \in \mathcal{U}} f(\mathbf{y}_i; \boldsymbol{\theta})$$

e

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

Igualando-se as derivadas parciais de $l_{\mathcal{U}}$ em relação a cada componente de $\boldsymbol{\theta}$ a zero, temos as equações de verossimilhança populacionais

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

em que

$$\mathbf{u}_i(\boldsymbol{\theta}) = \frac{\partial log[f(\mathbf{y}_i, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}$$

é o vetor $(k \times 1)$ dos escores do elemento $i \in \mathcal{U}$.

A solução $\boldsymbol{\theta}_{\mathcal{U}}$ deste sistema é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ no caso de um censo, mas $\boldsymbol{\theta}_{\mathcal{U}}$ também pode ser interpretado como uma quantidade desconhecida da população finita para a qual se pretende fazer inferências com base em informações da amostra. Sob algumas condições de regularidade, que são satisfeitas, por exemplo, pelos modelos lineares generalizados usuais (COX; HINKLEY, 1979), tem-se que $\boldsymbol{\theta}_{\mathcal{U}} - \boldsymbol{\theta} = \mathbf{0}$. Como em pesquisas por amostragem o tamanho populacional é geralmente grande, um bom estimador para $\boldsymbol{\theta}_{\mathcal{U}}$ também será bom para $\boldsymbol{\theta}$.

Seja $\mathbf{T} = \sum_{i \in \mathcal{U}} \mathbf{u}_i(\boldsymbol{\theta})$ a soma dos vetores de escores da população, que é um vetor de totais populacionais. Para estimar este vetor de totais, pode-se usar um estimador linear ponderado da forma $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta})$, em que w_i são pesos definidos apropriadamente. Binder (1983) propôs uma abordagem para obter um estimador para $\boldsymbol{\theta}_{\mathcal{U}}$ resolvendo o sistema de equações obtido igualando o estimador $\hat{\mathbf{T}}$ do total $\mathbf{T}$ a zero. Este estimador, chamado de estimador de máxima pseudo-verossimilhança, denotado por $\hat{\boldsymbol{\theta}}_{MPV}$, é a solução das equações de pseudo-verossimilhança dadas por

$$\hat{\mathbf{T}}(\hat{\boldsymbol{\theta}}_{MPV}) = \sum_{i \in s} w_i \mathbf{u}_i(\hat{\boldsymbol{\theta}}_{MPV}) = \mathbf{0}. \tag{2.8}$$

Através da linearização de Taylor pode-se obter a variância assintótica, sob o plano amostral, do estimador $\hat{\boldsymbol{\theta}}_{MPV}$ e seu estimador correspondente, dados respectivamente por

$$V_p(\hat{\boldsymbol{\theta}}_{MPV}) \simeq [I(\boldsymbol{\theta}_{\mathcal{U}})]^{-1} V_p \left[\sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta}_{\mathcal{U}}) \right] [I(\boldsymbol{\theta}_{\mathcal{U}})]^{-1} \tag{2.9}$$

e

$$\hat{V}_p(\hat{\boldsymbol{\theta}}_{MPV}) \simeq [I(\hat{\boldsymbol{\theta}}_{MPV})]^{-1}\hat{V}_p\left[\sum_{i\in s} w_i\mathbf{u}_i(\hat{\boldsymbol{\theta}}_{MPV})\right][I(\hat{\boldsymbol{\theta}}_{MPV})]^{-1} \tag{2.10}$$

em que

$$I(\boldsymbol{\theta}_{\mathcal{U}}) = \left.\frac{\partial T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U} = \sum_{i\in\mathcal{U}} \left.\frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U}, \tag{2.11}$$

$$I(\hat{\boldsymbol{\theta}}_{MPV}) = \left.\frac{\partial \hat{T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}} = \sum_{i\in\mathcal{U}} \left.\frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}},$$

$$V_p\left[\sum_{i\in s} w_i\mathbf{u}_i(\boldsymbol{\theta}_{\mathcal{U}})\right]$$

é a matriz de variância do estimador do total populacional dos escores e

$$\hat{V}_p\left[\sum_{i\in s} w_i\mathbf{u}_i(\hat{\boldsymbol{\theta}}_{MPV})\right]$$

é um estimador consistente para esta variância.

Uma quantidade vasta de modelos paramétricos podem ser ajustados (variando-se os planos amostrais e usando estimadores de totais diferentes) pelo método de máxima pseudo-verossimilhança satisfeitas algumas condições de regularidades discutidas em Binder (1983), como, por exemplo, tamanho da amostra grande e estimadores de totais de funções das variáveis de pesquisa baseados no desenho, assintoticamente não viciados e normalmente distribuídos. Além disso, os estimadores de máxima pseudo-verossimilhança não são únicos, pois existem diversas maneiras de se definir os pesos w_i.

Os pesos mais utilizados, por sua simplicidade, são os do estimador de Horwitz-Thompson para totais, dados pelo inverso da probabilidade de inclusão do indivíduo $i \in s$, ou seja, $w_i = \pi_i^{-1}$. Neste caso, o estimador de totais dado em (2.8) ficaria da forma:

$$\hat{\mathbf{T}}(\hat{\boldsymbol{\theta}}_\pi) = \sum_{i\in s} \pi_i^{-1}\mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi) = \mathbf{0}.$$

Analogamente, a variância assintótica de $\hat{\boldsymbol{\theta}}_\pi$ sob o plano amostral e seu estimador são obtidos substituindo w_i pelos pesos π_i^{-1} em (2.9) e (2.10):

$$V_p(\hat{\boldsymbol{\theta}}_\pi) \simeq [I(\boldsymbol{\theta}_{\mathcal{U}})]^{-1}V_p\left[\sum_{i\in s} \pi_i^{-1}\mathbf{u}_i(\boldsymbol{\theta}_{\mathcal{U}})\right][I(\boldsymbol{\theta}_{\mathcal{U}})]^{-1}$$

e

$$\hat{V}(\hat{\boldsymbol{\theta}}_\pi) \simeq [I(\hat{\boldsymbol{\theta}}_\pi)]^{-1}\hat{V}_\pi\left[\sum_{i\in s} \pi_i^{-1}\mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi)\right][I(\hat{\boldsymbol{\theta}}_\pi)]^{-1}$$

em que $I(\boldsymbol{\theta}_{\mathcal{U}})$ é dado em (2.11),

$$V_p\left[\sum_{i\in s}\pi_i^{-1}\mathbf{u}_i(\boldsymbol{\theta}_{\mathcal{U}})\right]=\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left[\mathbf{u}_i(\boldsymbol{\theta}_{\mathcal{U}})\right]\left[\mathbf{u}_j(\boldsymbol{\theta}_{\mathcal{U}})\right]',$$

$$I(\hat{\boldsymbol{\theta}}_\pi)=\sum_{i\in s}\pi_i^{-1}\frac{\partial\mathbf{u}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\pi},$$

e

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

Em suma, os passos de um procedimento padrão para ajustar um modelo paramétrico regular $f(\mathbf{y};\boldsymbol{\theta})$ pelo método da máxima pseudo-verossimilhança pode ser dados por:

i. Resolver $\sum_{i\in s}\pi_i^{-1}\mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi)=\mathbf{0}$ e calcular o estimador pontual $\hat{\boldsymbol{\theta}}_\pi$ do parâmetro $\boldsymbol{\theta}$ no modelo $f(\mathbf{y};\boldsymbol{\theta})$.

ii. Calcular a matriz de variância estimada $\hat{V}(\hat{\boldsymbol{\theta}}_\pi)\simeq[I(\hat{\boldsymbol{\theta}}_\pi)]^{-1}\hat{V}_\pi\left[\sum_{i\in s}\pi_i^{-1}\mathbf{u}_i(\hat{\boldsymbol{\theta}}_\pi)\right][I(\hat{\boldsymbol{\theta}}_\pi)]^{-1}$.

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

Silva (1996) e Duarte (1999) discutem vantagens e desvantagens do método de máxima pseudo-verossimilhança. Uma das vantagens é que este método proporciona estimativas baseadas no plano amostral para a variância assintótica dos estimadores dos parâmetros, as quais são razoavelmente simples de calcular. Além disso, mesmo quando o estimador pontual de $\boldsymbol{\theta}$ coincide com o estimador usual de máxima verossimilhança, a estimativa da variância obtida pelo procedimento de máxima pseudo-verossimilhança é diferente e pode ser preferível aos estimadores usuais da variância baseados no modelo, que ignoram o plano amostral.

Dentre as desvantagens está o fato de que 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. Um outro problema é o fato de que os métodos usuais de diagnóstico de ajuste de modelos, tais como gráficos de resíduos, e outros procedimentos da inferência clássica, como testes estatísticos de razão de verossimilhanças, não podem ser utilizados, embora hajam recursos alternativos para diagnóstico que consideram os efeitos dos diferentes aspectos do desenho amostral complexo empregado (Eltinge (1999); Korn e Graubard (2011)).

2.3.3 Ajuste de Modelos de Regressão

A teoria exposta anteriormente será estendida à classe de modelos de regressão linear normal e logístico, e está detalhada em Silva (1996) e Pessoa e Silva (1998).

2.3.3.1 Regressão Linear Normal

Sendo $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ um modelo de regressão com o objetivo de explicar os valores de uma variável de interesse y através da matriz de p variáveis explicativas $\mathbf{Z}$ (por exemplo, Shaper e Pocock (1985) ajustaram um modelo de regressão utilizando dados da **National Health and Nutrition Examination Survey** (MCDOWELL, 1981) considerando pressão arterial como variável resposta e níveis de chumbo no sangue, idade e outras variáveis de saúde como variáveis regressoras), para cada observação i está associado um valor y_i e um vetor $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$.

Sejam Y_i e $\mathbf{X}_i$ as notações para a variável e o vetor aleatório que geram y_i e $\mathbf{x}_i$, respectivamente, em que $(Y_i, \mathbf{X}_i)'$, $i \in \mathcal{U}$, são vetores aleatórios e identicamente distribuídos de modo que

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right] \tag{2.12}$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ e σ^2 são parâmetros desconhecidos do modelo.

As funções escores para $\boldsymbol{\beta}$ e σ^2 são dadas por:

$$\mathbf{u}_{i_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \sigma^2) = \frac{\partial log[f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)]}{\partial \boldsymbol{\beta}} = \frac{\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sigma^2} \propto \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$$

e

$$\mathbf{u}_{i_{\sigma^2}}(\boldsymbol{\beta}, \sigma^2) = \frac{\partial log[f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)]}{\partial \sigma^2} = \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2}{\sigma^4} \propto (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2.$$

No caso de um censo, ou seja, se todos os elementos da população forem pesquisados, os estimadores de $\boldsymbol{\beta}$ e σ são as soluções das equações de verossimilhança do censo, dadas por

$$\sum_{i \in \mathcal{U}} \mathbf{u}_i(\boldsymbol{\beta}_N) = \sum_{i \in \mathcal{U}} \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}_N) = \mathbf{x}_{\mathcal{U}}'\mathbf{y}_{\mathcal{U}} - (\mathbf{x}_{\mathcal{U}}'\mathbf{x}_{\mathcal{U}})\boldsymbol{\beta}_N = \mathbf{0}$$

e

$$\sum_{i \in \mathcal{U}} \mathbf{u}_i(\sigma_N^2) = \sum_{i \in \mathcal{U}} [(y_i - \mathbf{x}_i'\boldsymbol{\beta}_N)^2 - \sigma_N^2] = (\mathbf{y}_{\mathcal{U}} - \mathbf{x}_{\mathcal{U}}'\boldsymbol{\beta}_N)'(\mathbf{y}_{\mathcal{U}} - \mathbf{x}_{\mathcal{U}}\mathbf{x}_{\mathcal{U}}') - N\sigma_N^2 = 0.$$

Isto é,

$$\boldsymbol{\beta}_N = (\mathbf{x}_{\mathcal{U}}'\mathbf{x}_{\mathcal{U}})^{-1}(\mathbf{x}_{\mathcal{U}}'\mathbf{y}_{\mathcal{U}})$$

e

$$\sigma_N^2 = N^{-1} \sum_{i \in \mathcal{U}} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = N^{-1}(\mathbf{y}_{\mathcal{U}} - \mathbf{x}_{\mathcal{U}}'\boldsymbol{\beta}_N)'(\mathbf{y}_{\mathcal{U}} - \mathbf{x}_{\mathcal{U}}'\boldsymbol{\beta}_N),$$

em que os índices N e $\mathcal{U}$ são apenas indicadores de que são estimadores para o censo e que todos os elementos da população foram observados.

Quando apenas uma amostra de tamanho n é observada, os pesos w_i são introduzidos para obter os estimadores de máxima pseudo-verossimilhança de $\boldsymbol{\beta}$ e σ^2, que são encontrados resolvendo as seguintes equações de máxima pseudo-verossimilhança:

$$\sum_{i\in s} w_i \mathbf{u}_{i_{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}_s; \hat{\sigma}_s^2) = \sum_{i\in s} w_i \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_s) = 0$$

e

$$\sum_{i\in s} w_i \mathbf{u}_{i_{\sigma^2}}(\hat{\boldsymbol{\beta}}_s; \hat{\sigma}_s^2) = \sum_{i\in s} w_i [(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_s)^2 - \hat{\sigma}_s^2] = 0.$$

Neste caso, as expressões para os estimadores de máxima pseudo-verossimilhança dos parâmetros do modelo são dadas por:

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{x}_s' \mathbf{W}_s \mathbf{x}_s)^{-1} \mathbf{x}_s' \mathbf{W}_s \mathbf{y}_s \tag{2.13}$$

e

$$\hat{\sigma}_s^2 = (\mathbf{1}_s' \mathbf{W}_s \mathbf{1}_s)^{-1} \mathbf{y}_s' [\mathbf{W}_s - \mathbf{W}_s \mathbf{x}_s (\mathbf{x}_s' \mathbf{W}_s \mathbf{x}_s)^{-1} \mathbf{x}_s' \mathbf{W}_s] \mathbf{y}_s,$$

em que $\mathbf{W}_s = diag(w_{i1}, \ldots, w_{in})$ é uma matriz diagonal nxn que contém os pesos dos elementos que pertencem à amostra na diagonal principal, e o índice s um indicador de que são estimadores baseados na amostra.

Se todos os elementos da amostra tiverem o mesmo peso, isto é, $w_i = \bar{w}$ e $\mathbf{W}_s = \bar{w}\mathbf{I}_n$, os estimadores pontuais não dependem de $\bar{w}$. Neste caso, eles seriam equivalentes aos estimadores de mínimos quadrados ordinários.

Para o modelo de regressão linear, as expressões para a variância assintótica de aleatorização do estimador $\hat{\boldsymbol{\beta}}_s$ e seu estimador consistente podem ser obtidos substituindo os valores dos escores e do jacobiano correspondentes em (2.9) e (2.10):

$$V(\hat{\boldsymbol{\beta}}_s) = (\mathbf{x}_{\mathcal{U}}' \mathbf{x}_{\mathcal{U}})^{-1} V_p(\sum_{i\in s} w_i \mathbf{x}_i \boldsymbol{\epsilon}_i)(\mathbf{x}_{\mathcal{U}}' \mathbf{x}_{\mathcal{U}})^{-1},$$

em que

$$V_p(\sum_{i\in s} w_i \mathbf{x}_i \boldsymbol{\epsilon}_i) = \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{U}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \boldsymbol{\epsilon}_i \mathbf{x}_i \mathbf{x}_j' \boldsymbol{\epsilon}_j$$

e $\boldsymbol{\epsilon}_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}_N$ é o vetor de resíduos.

O estimador de $V(\hat{\boldsymbol{\beta}}_s)$ é:

$$\hat{V}(\hat{\boldsymbol{\beta}}_s) = (\mathbf{x}_s' \mathbf{W}_s \mathbf{x}_s)^{-1} \hat{V}_p(\sum_{i\in s} w_i \mathbf{x}_i \hat{\boldsymbol{\epsilon}}_i)(\mathbf{x}_s' \mathbf{W}_s \mathbf{x}_s)^{-1}, \tag{2.14}$$

em que

$$\hat{V}_p(\sum_{i\in s} w_i \mathbf{x}_i \hat{\boldsymbol{\epsilon}}_i) = \sum_{i\in s} \sum_{j\in s} (\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}) \hat{\boldsymbol{\epsilon}}_i \mathbf{x}_i \mathbf{x}_j' \hat{\boldsymbol{\epsilon}}_j$$

e $\hat{\boldsymbol{\epsilon}}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_s$ é um estimador do vetor de resíduos.

2.3.3.2 Regressão Logistíca

Regressão logística, também chamada de modelo logit ou modelo logístico, analisa a relação entre diversas variáveis regressoras e uma variável resposta categórica, e estima a probabilidade de ocorrência de um evento através do ajuste dos dados a uma curva logística (PARK, 2013). Existem dois modelos de regressão logística: regressão logística binária e regressão logística multinomial. Regressão logística binária é normalmente utilizada quando a variável dependente é dicotômica e as variáveis independentes são contínuas ou categóricas. Quando a variável dependente não é dicotômica e é composta por mais de duas categorias, uma regressão logística multinomial pode ser utilizada.

Como na seção anterior, considerando que para cada observação i da amostra está associado um valor y_i e um vetor $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ com p variáveis explicativas. Sejam Y_i e $\mathbf{X}_i$ as notações para a variável e o vetor aleatório que geram y_i e $\mathbf{x}_i$, em que $(Y_i, \mathbf{X}_i)'$, $i \in \mathcal{U}$, são vetores aleatórios e identicamente distribuídos de modo que o modelo de superpopulação é dado por:

$$f(y_i|\mathbf{x}_i; \boldsymbol{\beta}) = [p(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - p(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}$$

em que

$$p(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

e $\boldsymbol{\beta} = (\beta_i, \ldots, \beta_p)'$ é um vetor de parâmetros desconhecidos do modelo.

A função escore para $\boldsymbol{\beta}$ é dada por

$$\mathbf{u}_i(\boldsymbol{\beta}) = \frac{\partial log[f(y_i|\mathbf{x}_i; \boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = [y_i - p(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i.$$

No caso de um censo, o estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ é a solução do sistema de equações de verossimilhança dadas por:

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

Quando apenas uma amostra de tamanho n é observada, os pesos w_i são introduzidos para obter os estimadores de máxima pseudo-verossimilhança, que são encontrados resolvendo

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

Neste caso, assim como é feito na modelagem clássica, a obtenção do estimador $\hat{\boldsymbol{\beta}}$ só pode ser feita através de métodos númericos interativos (Paula (2004); Cordeiro e Demétrio (2010); McCullagh e Nelder (1989)).

2.3.3.3 Teste de Hipóteses

Após o ajuste de um modelo de regressão, testam-se hipóteses relativas aos parâmetros desse modelo. Assim como o método de estimação dos parâmetros sofreu ajustes para incorporar o esquema amostral, é intuitivo pensar em adaptar também as estatísticas dos testes de hipóteses.

Seja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ o vetor de parâmetros de um modelo de regressão. Quando há interesse em testar $H_0 : \beta_j = 0$ para algum $j \in \{1, \ldots, p\}$, usa-se o teste t, e quando o interesse é testar $H_0 : (\beta_{j1}, \ldots, \beta_{jr})' = \mathbf{0}$, usa-se o teste F. Ambos os testes, sob as hipóteses do modelo clássico de regressão com erros normais, tratam-se de testes de Razão de Máxima Verossimilhança. Porém, em pesquisas amostrais complexas, a função de verossimilhança usual não representa a distribuição conjunta das observações.

Neste caso, são aplicados testes baseados na estatística de Wald, devido a Wald (1943), que mede a distância entre o estimador pontual e o valor hipotético do parâmetro numa métrica definida pela matriz de covariância do estimador. Esta estatística permite que os pesos e o plano amostral sejam incorporados sem dificuldade.

Chambers e Skinner (2003) e Pessoa e Silva (1998) consideraram o problema de testar a hipótese linear geral:

$$H_0 : \quad \mathbf{C}\boldsymbol{\beta} = \mathbf{c} \tag{2.15}$$

em que $\mathbf{C}$ é uma matriz $(r \times p)$ de posto completo, $\boldsymbol{\beta}$ o vetor de parâmetros desconhecidos e $\mathbf{c}$ um vetor $(r \times 1)$.

A estatística de Wald clássica para testar a hipótese nula (2.15) é definida por

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

em que $\hat{\boldsymbol{\beta}}$ e $\hat{V}(\hat{\boldsymbol{\beta}})$ são obtidos pelo método dos mínimos quadrados ordinários.

No caso de dados de pesquisas amostrais complexas, a correção é feita substituindo $\hat{\boldsymbol{\beta}}$ e $\hat{V}(\hat{\boldsymbol{\beta}})$ pela estimativa de máxima pseudo-verossimilhança de $\boldsymbol{\beta}$ dada em (2.13) e pela estimativa da matriz de covariância do estimador de máxima pseudo-verossimilhança de $V(\hat{\boldsymbol{\beta}})$ dada em (2.14), respectivamente, em X_W^2.

Segundo Binder (1983), a estatística X_W^2 corrigida tem distribuição assintótica qui-quadrado com r graus de liberdade.

2.3.4 Aplicação em Regressão Linear Normal

Esta aplicação, cujos dados são provenientes de uma parte do Censo Experimental de Limeira, realizado em 1988, encontra-se em Figueiredo (2004). O conjunto de dados contém informações de 706 questionários preenchidos pelos chefes de domicílios durante o censo. Na Tabela 2.3 encontram-se algumas informações referentes ao domicílio incluído na amostra

pesquisada e outras informações auxiliares calculadas posteriormente. Na Tabela 2.4 encontram-se as variáveis disponíveis neste conjunto de dados.

Tabela 2.3 – Fragmento do conjunto de dados com variáveis auxiliares

SETOR	DOMIC	NP	NA	PESO1	PESO2	PESOF	DOMTOT
57	17	274	24	6,38919708	11,41666667	72,94333333	43766
57	27	274	24	6,38919708	11,41666667	72,94333333	43766
57	37	274	24	6,38919708	11,41666667	72,94333333	43766
57	47	274	24	6,38919708	11,41666667	72,94333333	43766
57	57	274	24	6,38919708	11,41666667	72,94333333	43766
57	67	274	24	6,38919708	11,41666667	72,94333333	43766
57	77	274	24	6,38919708	11,41666667	72,94333333	43766
57	87	274	24	6,38919708	11,41666667	72,94333333	43766
57	97	274	24	6,38919708	11,41666667	72,94333333	43766
57	107	274	24	6,38919708	11,41666667	72,94333333	43766
57	117	274	24	6,38919708	11,41666667	72,94333333	43766
57	127	274	24	6,38919708	11,41666667	72,94333333	43766
57	137	274	24	6,38919708	11,41666667	72,94333333	43766
57	147	274	24	6,38919708	11,41666667	72,94333333	43766
57	157	274	24	6,38919708	11,41666667	72,94333333	43766
57	177	274	24	6,38919708	11,41666667	72,94333333	43766
57	187	274	24	6,38919708	11,41666667	72,94333333	43766
57	217	274	24	6,38919708	11,41666667	72,94333333	43766
57	227	274	24	6,38919708	11,41666667	72,94333333	43766
57	237	274	24	6,38919708	11,41666667	72,94333333	43766
57	247	274	24	6,38919708	11,41666667	72,94333333	43766
57	257	274	24	6,38919708	11,41666667	72,94333333	43766
57	267	274	24	6,38919708	11,41666667	72,94333333	43766
57	287	274	24	6,38919708	11,41666667	72,94333333	43766
84	4	197	15	8,886497462	13,13333333	116,7093333	43766
84	14	197	15	8,886497462	13,13333333	116,7093333	43766
84	34	197	15	8,886497462	13,13333333	116,7093333	43766
84	44	197	15	8,886497462	13,13333333	116,7093333	43766
84	54	197	15	8,886497462	13,13333333	116,7093333	43766
84	64	197	15	8,886497462	13,13333333	116,7093333	43766
84	74	197	15	8,886497462	13,13333333	116,7093333	43766
84	84	197	15	8,886497462	13,13333333	116,7093333	43766
84	104	197	15	8,886497462	13,13333333	116,7093333	43766
84	144	197	15	8,886497462	13,13333333	116,7093333	43766
84	154	197	15	8,886497462	13,13333333	116,7093333	43766
84	164	197	15	8,886497462	13,13333333	116,7093333	43766
84	184	197	15	8,886497462	13,13333333	116,7093333	43766
84	194	197	15	8,886497462	13,13333333	116,7093333	43766
84	214	197	15	8,886497462	13,13333333	116,7093333	43766

Fonte: Figueiredo(2004)

Tabela 2.4 – Descrição das variáveis

Variável	Código	Descrição
SETOR	Numérico	Número do setor
DOMIC	Numérico	Número do domicílio no setor
NP	Numérico	Número de domicílios no setor na população
NA	Numérico	Número de pessoas na amostra do setor
SEXO	1	Masculino
	2	Feminino
RAÇA	1	Pessoa de raça branca
	2	Pessoa de outras raças
ESTUDO	1	Até 4 anos completos de estudo
	2	De 5 a 8 anos completos de estudo
	3	Mais de 8 anos completos de estudo
IDADE	1	Até 25 anos completos de idade
	2	De 26 a 40 anos completos de idade
	3	De 41 a 60 anos completos de idade
	4	Mais de 60 anos de idade
RENDA	Numérico	Valor da renda mensal (em unidade monetária vigente na época do Censo de Limeira)
LRENDA	Numérico	Logaritmo do valor da renda mensal
PESO1	Numérico	Peso do setor (1º estágio)
PESO2	Numérico	Peso do domicílio no setor (2º estágio)
PESOF	Numérico	Peso final do domicílio
DOMTOT	Numérico	Número total de domicílios no município

Fonte: Figueiredo(2004)

A amostra foi selecionada em duas etapas. Primeiramente, selecionaram-se 25 setores censitários com probabilidades proporcionais ao número de domicílios. Posteriormente, selecionaram-se aproximadamente 10% dos domicílios de cada um dos setores, segundo uma amostragem sistemática simples sem reposição.

Na Tabela 2.3 aparecem 2 dos 25 setores censitários selecionados na amostra. Por exemplo, o 57° setor continha 274 dos 43.766 domicílios do município de Limeira. Logo, a probabilidade de inclusão do 57° setor no primeiro estágio de amostragem é calculada como $\pi_{57(1^\circ)}$ = $n\times$(Total de domicílios no 57° setor)/(Total de domicílios em Limeira), em que n é o número de setores selecionados. Assim,

$$\pi_{57(1^\circ)} = 25 \times \frac{NP}{DOMTOT} = 25 \times \frac{274}{43766} = 0,1565.$$

Como o peso é o inverso da probabilidade de inclusão do setor 57 na amostra, o peso no primeiro estágio é dado por $\pi^{-1}_{57(1^\circ)} = 6,389$, e no segundo estágio por

$$\pi^{-1}_{57(2^\circ)} = \frac{274}{24} = 11,4167.$$

Portanto, o peso final de cada domicílio é:

$$\pi^{-1}_{57(1^\circ)} \times \pi^{-1}_{57(2^\circ)} = 6,389 \times 11,4167 = 72,94$$

Figueiredo (2004) ajustou um modelo de regressão linear para explicar o logaritmo da renda (variável LRENDA) em função do sexo (SEXO), idade (IDADE), anos de estudo (ESTUDO) e raça (RAÇA). O logaritmo da renda foi utilizado para normalizar a variável RENDA.

Como forma de verificar, descritivamente, quais interações de primeira ordem deveriam ser incluídas no modelo, foram construídos gráficos de perfis de médias. Feito esta análise, o primeiro modelo a ser ajustado foi:

$$\begin{aligned} LRENDA &= \beta_0 + \beta_1 SEXO + \beta_2 IDADE + \beta_3 ESTUDO + \beta_4 RACA + \beta_5 (IDADE * \\ &\quad SEXO) + \beta_6 (IDADE * ESTUDO) + \beta_7 (ESTUDO * SEXO) + \\ &\quad \beta_8 (ESTUDO * RACA). \end{aligned}$$

Nesta análise, Figueiredo (2004) aplicou o método que incorpora o esquema amostral no processo de estimação dos parâmetros por meio do pacote computacional SUDAAN (SUrvey DAta ANalysis). Segundo ele, este é um pacote que consiste de uma família de procedimentos usados para análise de dados provenientes de pesquisas amostrais complexas, ou outros estudos experimentais ou observacionais que envolvam dados correlacionados, possibilitando realizar operações de grande porte.

As estimativas dos parâmetros e de seus respectivos desvios padrões para o modelo considerando o plano amostral e considerando AAS estão na Tabela 2.5. Feito este ajuste, foi realizada uma seleção de variáveis baseada na estatística de Wald a fim de verificar cada efeito de interação (permanência/saída de cada coeficiente).

O resultado desta análise levou a dois modelos finais distintos, pois, como mostram as Tabelas 2.6 e 2.7, as estimativas dos parâmetros são diferentes quando consideramos AAS e quando consideramos o plano amostral efetivamente usado. Além disso, os dois modelos também diferem pela inclusão da interação IDADE*SEXO no modelo considerando AAS.

Modelo considerando o plano:

LRENDA = 11,35 + 0,65 SEXO - 0,74 IDADE1 - 0,02 IDADE2 + 0,10 IDADE3 - 1,87 ESTUDO1 - 1,39 ESTUDO2 + 0,21 RACA + 1,44 (ESTUDO1*IDADE1) + 0,84 (ESTUDO1*IDADE2) + 0,51 (ESTUDO1*IDADE3) + 1,02 (ESTUDO2*IDADE1) + 0,71 (ESTUDO2*IDADE2) + 0,58 (ESTUDO2*IDADE3)

Modelo considerando AAS:

LRENDA = 11,43 + 0,34 SEXO - 0,85 IDADE1 - 0,16 IDADE2 + 0,12 IDADE3 - 1,70 ESTUDO1 - 1,16 ESTUDO2 + 0,20 RACA + 1,41 (ESTUDO1*IDADE1) + 0,66 (ESTUDO1*IDADE2) + 0,36 (ESTUDO1*IDADE3) + 0,88 (ESTUDO2*IDADE1) + 0,45 (ESTUDO2*IDADE2)

Tabela 2.5 – Modelo completo

Variáveis	Considerando o plano Coeficiente	d.p.	p-valor	Considerando AAS Coeficiente	DP	p-valor
Intercepto	11,27	0,29	<0,0001	11,16	0,38	<0,0001
SEXO						
1	0,24	0,32	0,4559	0,22	0,27	0,4134
IDADE						
1	-0,57	0,33	0,0954	-0,71	0,45	0,1165
2	-0,03	0,25	0,9066	-0,02	0,34	0,9480
3	-0,01	019	0,9498	-0,01	0,36	0,9806
ESTUDO						
1	-1,57	0,37	0,0003	-1,43	0,40	0,0004
2	-0,99	0,45	0,0376	-0,92	0,50	0,0644
RACA						
1	0,42	0,22	0,0658	0,52	0,23	0,0252
SEXO*IDADE						
1,1	0,15	0,32	0,6454	0,22	0,29	0,4500
1,2	0,32	0,33	0,3485	0,37	0,24	0,1249
1,3	0,40	0,26	0,1398	0,40	0,22	0,0722
ESTUDO*SEXO						
1,1	0,13	0,22	0,5637	0,13	0,23	0,5909
2,1	0,28	0,37	0,4602	0,36	0,33	0,2732
ESTUDO*IDADE						
1,1	1,17	0,37	0,0046	1,00	0,43	0,0030
1,2	0,60	0,21	0,0092	1,00	0,34	0,1139
1,3	0,29	0,19	0,1405	0,00	0,35	0,4408
2,1	0,65	0,39	0,1082	1,00	0,42	0,1114
2,2	0,38	0,21	0,0837	0,00	0,35	0,3791
2,3	0,31	0,33	0,3615	0,00	0,39	0,3422
ESTUDO*RACA						
1,1	-0,20	0,25	0,4155	0,00	0,24	0,1865
2,1	-0,39	0,25	0,1372	-1,00	0,27	0,0569

Fonte: Figueiredo(2004)

+ 0,45 (ESTUDO2*IDADE3) + **0,22 (SEXO*IDADE1) + 0,39 (SEXO*IDADE2) + 0,43 (SEXO*IDADE3)**

A Tabela 2.7 também mostra as estimativas dos EPA's para cada coeficiente que permaneceu no modelo. Os valores baixos do EPA realçam a importância de considerar o plano amostral verdadeiro ao estimar variâncias e desvios padrões associados às estimativas dos parâmetros. Neste caso, as estimativas de variância baseadas na hipótese de AAS superestimaram substancialmente as variâncias corretas.

Nesta sessão, foram abordados ajustes de modelos de regressão linear normal e regressão logística. Contudo, é possível incorporar o esquema amostral em outras classes de modelos

Tabela 2.6 – Modelo final considerando AAS

Variáveis	Coeficiênte	d.p.	p-valor
Intercepto	11,43	0,28	<0,0001
SEXO			
1	0,34	0,18	0,0508
IDADE			
1	-0,85	0,41	0,400
2	-0,16	0,32	0,6167
3	-0,12	0,32	0,7020
ESTUDO			
1	-1,70	0,28	<0,0001
2	-1,16	0,29	0,0001
RACA			
1	0,20	0,06	0,0007
SEXO*IDADE			
1,1	0,22	0,28	0,4338
1,2	0,39	0,23	0,0982
1,3	0,43	0,22	0,0487
ESTUDO*IDADE			
1,1	1,41	0,39	0,0003
1,2	0,66	0,30	0,0268
1,3	0,36	0,31	0,2490
2,1	0,88	0,38	0,0214
2,2	0,45	0,30	0,1412
2,3	0,45	0,35	0,1992

Fonte: Figueiredo(2004)

também. Por exemplo, modelos de logitos generalizados como feito em Cunha, Araújo e Lima (2015) (Seção 2.2.2), Leite e Silva (2013) e Araújo *et al.* (2011), adequado para os casos em que a variável resposta é qualitativa, nominal e politômica. E também modelo de logitos cumulativos como feito em Moraes, Moreira e Luiz (2012), adequado para os casos em que a variável resposta é qualitativa, nominal e politômica ordinal.

2.4 Análise em Tabelas de Contingência

Kish e Frankel (1974) chamaram a atenção para os problemas que surgem quando os métodos estatísticos padrões, com base na suposição de que as observações são independentes, são aplicados para o levantamento de dados. Na maioria dos estudos, a suposição de independência está longe de ser realista. Qualquer pesquisa em larga escala envolverá amostragem estratificada em múltiplos estágios, e as correlações entre as unidades no mesmo conglomerado (ou estrato) pode ter um impacto substancial.

A teoria exposta nesta seção é baseada em Lehtonen e Pahkinen (2004), Rao e Scott (1981)

Tabela 2.7 – Modelo final considerando o plano amostral

Variáveis	Coeficiênte	d.p.	p-valor	EPA
Intercepto	11,35	0,19	<0,0001	0,20
SEXO				
1	0,65	0,06	<0,0001	0,44
IDADE				
1	-0,74	0,31	0,0235	0,42
2	-0,02	0,22	0,9161	0,25
3	0,10	0,13	0,4722	0,09
ESTUDO				
1	-1,87	0,19	0,0000	0,20
2	-1,39	0,17	0,0000	0,04
RACA				
1	0,21	0,07	0,0062	1,14
ESTUDO*IDADE				
1,1	1,44	0,40	0,0014	0,60
1,2	0,84	0,26	0,0034	0,35
1,3	0,51	0,17	0,0073	0,14
2,1	1,02	0,33	0,0051	0,16
2,2	0,71	0,22	0,0039	0,08
2,3	0,58	0,27	0,0395	0,11

Fonte: Figueiredo(2004)

e Rao e Scott (1984), e aborda a questão do efeito desta estrutura sobre o comportamento dos testes qui-quadrado comuns para os testes de bondade de ajuste, homogeneidade e independência em tabelas de contingência.

Exemplos do uso dos testes qui-quadrado em pesquisas amostrais complexas são abundantes. Um pesquisador pode querer, por exemplo, comparar as proporções das amostras obtidas para categorias de variáveis tais como idade e sexo com proporções populacionais conhecidas ou obtidas em anos anteriores. De modo geral, pode-se fazer comparações entre várias pesquisas diferentes da mesma população ou entre diferentes regiões do país. Neste caso, mais especificamente, um teste qui-quadrado usual de homogeneidade seria útil, exceto se não fosse a complexidade das populações amostradas.

É bastante comum na análise de dados amostrais complexos, tabelas contendo proporções relativas a uma ou mais variáveis categóricas (tabelas de uma entrada) ou constituídas por celas de uma classificação cruzada de duas ou mais variáveis categóricas (tabelas de duas entradas ou mais). No primeiro caso, o objetivo é testar hipóteses relativas a bondade de ajuste. No segundo caso, o interesse é testar a independência ou a homogeneidade entre as variáveis. O teste de independência é adequado para estudar a associação entre duas variáveis categóricas, enquanto o teste de homogeneidade é apropriado para estudar a igualdade das distribuições condicionais de uma variável resposta categórica.

Para hipóteses de homogeneidade e independência em uma tabela ($r \times c$), é comum, na prática, ignorar a complexidade da pesquisa e proceder como se os testes qui-quadrado padrões se comportassem da mesma forma que se comportam sob amostragem aleatória simples, ou seja, proceder como se a estatística qui-quadrado de Pearson fosse assintoticamente χ^2 com $(r-1)(c-1)$ graus de liberdade. Contudo, esta forma de conduzir a pesquisa não é válida para amostras complexas. Uma correlação intra-classe positiva das variáveis, por exemplo, faz com que o teste seja excessivamente liberal em relação aos níveis de significância nominais. Portanto, os valores observados para a estatística do teste pode ser muito grande, o que pode levar a inferências errôneas.

Para que as inferências sejam válidas, algumas correções nas estatísticas clássicas precisam ser feitas para terem a mesma distribuição assintótica de referência que a obtida para o caso de amostragem aleatória simples.

2.4.1 Tabelas de uma entrada

2.4.1.1 Estatística χ^2 de Pearson

A estatística clássica de Pearson para testar a hipótese

$$H_0 : p_k = p_{0k}, \quad k = 1, 2 \tag{2.16}$$

é dada por

$$X_P^2 = \sum_{k=1}^{2} \frac{n(\hat{p_k} - p_{0k})^2}{p_{0k}} \tag{2.17}$$

em que $\hat{p_k} = n_k/n$ são as estimativas amostrais usuais para as proporções populacionais p_k e p_{0k} são as proporções hipotéticas das celas. Sob a hipótese de observações IID, a distribuição assintótica de X_P^2 é qui-quadrado e, neste caso, com um grau de liberdade.

Rao e Scott (1981, 1984) propuseram um método para corrigir a estatística qui-quadrado de Pearson em estudos com esquema amostral complexo, estimando pesos como função do efeito do delineamento amostral e usando os mesmos para corrigir esta estatística. Eles mostraram que a distribuição assintótica da estatística qui-quadrado é uma soma ponderada de variáveis aleatórias qui-quadrado independentes, em que os pesos são funções do efeito do delineamento amostral.

No caso de uma tabela com apenas duas celas, a distribuição assintótica de X_P^2 é a mesma distribuição da variável aleatória dW, em que W tem distribuição qui-quadrado com um grau de liberdade e d é o efeito do plano amostral da estimativa $\hat{p}$ de p. Neste caso, $d = \dfrac{V_{VERD}(\hat{p})}{V_{IID}(\hat{p})}$.

Como em planos complexos, geralmente, $d \neq 1$, então X_P^2 não tem distribuição $\chi^2_{(1)}$. Logo, é necessária uma correção em X_P^2. A esperança assintótica de X_P^2 é $E(X_P^2) = d$ que é diferente de 1 nos casos de correlação intra-classe positiva e como $E(X_P^2/d) = E(\chi^2_{(1)}) = 1$, a

correção em (2.17) pode ser dada por:

$$X_P^2(d) = \frac{X_P^2}{d} \sim \chi^2_{(1)}.$$

Um procedimento de teste válido para testar (2.16) nos casos de tabelas com mais de duas celas é mais complicado que o simples caso de duas celas. No caso de $J \geq 2$ celas, podemos escrever a hipótese nula (2.16) da seguinte forma:

$$H_0: \quad \boldsymbol{p} = \boldsymbol{p}_0 \tag{2.18}$$

em que $\boldsymbol{p} = (p_1, \ldots, p_{K-1})'$ é o vetor de proporções populacionais desconhecidas e $\boldsymbol{p}_0 = (p_{01}, \ldots, p_{0(K-1)})'$ é o vetor de proporções hipotéticas. O vetor de estimadores para as proporções das celas $\boldsymbol{p}$ é denotado por $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{(K-1)})'$, em que $\hat{p}_k = n_k/n$.

Apenas $(J-1)$ elementos são inclusos nos vetores $\boldsymbol{p}$, $\boldsymbol{p}_0$ e $\hat{\boldsymbol{p}}$. Isso é devido ao fato da soma das proporções ser igual a um. Logo, por exemplo, $\hat{p}_K = 1 - \sum_{k=1}^{K-1} \hat{p}_k$.

Neste caso, a estatística qui-quadrado de Pearson (2.17) pode ser escrita como:

$$X_P^2 = n \sum_{k=1}^{K} \frac{(\hat{p_k} - p_{0k})^2}{p_{0k}} = n(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)' \mathbf{P}_0^{-1} (\hat{\boldsymbol{p}} - \boldsymbol{p}_0) \tag{2.19}$$

em que $\mathbf{P}_0 = diag(\boldsymbol{p}_0) - \boldsymbol{p}_0 \boldsymbol{p}_0'$ e $\mathbf{P}_0/n$ é a matriz $[(K-1) \times (K-1)]$ de covariância multinomial de $\hat{\boldsymbol{p}}$ sob a hipótese nula, e $diag(\boldsymbol{p}_0)$ representa uma matriz diagonal com elementos p_{0k} na diagonal.

Para examinar a distribuição de (2.19), Lehtonen e Pahkinen (2004), baseados nos resultados de Rao e Scott (1981), generalizaram os resultados anteriores para o caso geral $K \geq 2$. Neste caso, X_P^2 é distribuído assintóticamente como uma soma ponderada $\lambda_1 W_1 + \lambda_2 W_2 + \ldots + \lambda_{K-1} W_{K-1}$ de $(K-1)$ variáveis aleatórias independentes W_k com distribuição $\chi^2_{(1)}$.

Os pesos λ_k são os autovalores da matriz de efeito multivariado $\mathbf{D} = \mathbf{P}_0^{-1} \boldsymbol{V}_p$, em que $\boldsymbol{V}_p/n$ é a matriz de covariância do estimador $\hat{\boldsymbol{p}}$ do vetor de proporções $\boldsymbol{p}$ baseada no plano amostral verdadeiro. Esses autovalores também são conhecidos como efeitos generalizados do plano amostral.

No caso de amostragem aleatória simples, $\mathbf{P}_0$ e $\boldsymbol{V}_p$ se coincidem. Assim, $\mathbf{D} = \mathbf{I}$, matriz identidade. Logo, os pesos λ_k são todos iguais a um. Neste caso, $\sum_{k=1}^{K-1} \lambda_k W_k$ se reduz a $\sum_{k=1}^{K-1} W_k$, cuja distribuição é χ^2 com $(K-1)$ graus de liberdade. Dessa forma, X_P^2 precisa ser corrigida para ter a mesma distribuição assintótica de referência que a obtida para o caso de amostragem aleatória simples.

Dois dos possíveis ajustes são as correções de primeira e segunda ordem de Rao-Scott. A correção de primeira ordem parte do princípio de que, sob a hipótese nula, a esperança assintótica

de X_P^2 é $E(X_P^2) = \sum_{k=1}^{K-1} \lambda_k$. Logo, $E(X_P^2/\bar{\lambda}) = E(\chi^2_{(K-1)}) = K-1$, em que $\bar{\lambda} = \sum_{k=1}^{K-1} \frac{\lambda_k}{K-1}$ é a média dos autovalores λ_k.

Este raciocínio conduz ao ajuste de primeira ordem para X_P^2, dado por:

$$X_P^2(\hat{\lambda}_*) = \frac{X_P^2}{\hat{\lambda}_*} \tag{2.20}$$

em que $\hat{\lambda}_*$ é um estimador da média $\bar{\lambda}$ dos autovalores desconhecidos da matriz $\mathbf{D}$ que pode ser estimado por

$$(K-1)\hat{\lambda}_* = \sum_{k=1}^{K} \frac{\hat{p}_k}{p_{ok}}(1-\hat{p}_{0k})\hat{d}_k.$$

Pela equação acima, $\hat{\lambda}_*$ pode ser estimado sem estimar os próprios autovalores. Assim, essa estatística de teste é preferida quando não é possível obter uma estimativa adequada $\hat{\boldsymbol{V}}_p$ para a matriz $\boldsymbol{V}_p$. Quando for possível obter essa estimativa, deve-se usar o ajuste de segunda ordem, que é mais preciso. Uma outra observação é que $X_P^2(\hat{\lambda}_*)$ só tem distribuição $\chi^2_{(K-1)}$ se os autovalores λ_k forem iguais. Na prática isso raramente acontece, mas essa estatística é usada quando a variação dos autovalores é pequena, o que é aceitável pois a correção de primeira ordem tem por objetivo tornar a média assintótica da estatística ajustada igual ao número de graus de liberdade da distribuição de referência.

Para os casos em que essa variação for grande, deve-se usar a correção de segunda ordem baseada no método de Satterthwaite (1946) dada por:

$$X_P^2(\hat{\lambda}_*, \hat{a}^2) = \frac{X_P^2(\hat{\lambda}_*)}{(1+\hat{a}^2)} \tag{2.21}$$

em que

$$\hat{a}^2 = \sum_{k=1}^{K-1} \frac{\hat{\lambda}_k^2}{\left[(K-1)\hat{\lambda}_*\right] - 1}$$

é um estimador do quadrado do coeficiente de variação a^2 dos autovalores desconhecidos.

Um estimador da soma dos quadrados dos autovalores é dada por:

$$\sum_{k=1}^{K-1} \hat{\lambda}_*^2 = n^2 \sum_{k=1}^{K} \sum_{i=1}^{K} \frac{\hat{V}_p^2(\hat{p}_k, \hat{p}_i)}{p_{0k}p_{0i}}$$

em que $\hat{V}_p^2(\hat{p}_k, \hat{p}_i)$ são os estimadores das covariâncias de aleatorização de $\hat{p}_k$ e $\hat{p}_i$.

Os graus de liberdade também precisam ser corrigidos. A estatística (2.21) é assintoticamente χ^2 com número de graus de liberdade com ajuste de Satterthwaite dado por

$$gl = \frac{K-1}{1+\hat{a}^2}.$$

2.4.1.2 Estatística de Wald

Como uma segunda opção para testar (2.16), existe a estatística de Wald, que difere da estatística qui-quadrado de Pearson por incorporar automaticamente o plano amostral e é dada, no caso de duas celas, por:

$$X_W^2 = \frac{(\hat{p} - p_0)^2}{\hat{V}_{VERD}(\hat{p})} \quad \sim \quad \chi^2_{(1)}. \tag{2.22}$$

Um procedimento de teste válido para testar (2.16) nos casos de tabelas com mais de duas celas usando a estatística de Wald também é mais complicado que o simples caso de duas celas, como foi visto na estatística de Pearson com ajuste de Rao-Scott baseadas no plano amostral. No caso de $K \geq 2$ celas, considerando novamente a hipótese nula (2.16) escrita da seguinte forma

$$H_0: \quad \boldsymbol{p} = \boldsymbol{p}_0$$

em que $\boldsymbol{p} = (p_1, \ldots, p_{K-1})'$ é o vetor de proporções populacionais desconhecidas e $\boldsymbol{p}_0 = (p_{01}, \ldots, p_{0(K-1)})'$ é o vetor de proporções hipotéticas, o vetor de estimadores pra as proporções das celas $\boldsymbol{p}$ é denotado por $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{(K-1)})'$, em que $\hat{p}_k = n_k/n$.

A estatística de Wald baseada no plano amostral neste caso é dada por:

$$X_W^2 = (\hat{\boldsymbol{p}} - \boldsymbol{p}_0)' \hat{\boldsymbol{V}}_p^{-1} (\hat{\boldsymbol{p}} - \boldsymbol{p}_0) \tag{2.23}$$

em que $\hat{\boldsymbol{V}}_p^{-1}$ é um estimador consistente da matriz de covariância $\boldsymbol{V}_p$ do estimador $\hat{\boldsymbol{p}}$ do vetor de proporções $\boldsymbol{p}$, e pode ser obtido pelo Método de Linearização de Taylor, por exemplo. Sob H_0, $X_W^2 \sim \chi^2_{(K-1)}$ (SHUSTER; DOWNING, 1976).

Porém, existem algumas situações instáveis que dependendo do desenho amostral utilizado e, por exemplo, se o número de celas do vetor $\boldsymbol{p}$ for pequeno, (2.23) pode não funcionar adequadamente devido ao pequeno número de graus de liberdade disponíveis para o cálculo de $\hat{\boldsymbol{V}}_p$. Neste caso, existem duas alternativas de correção, que são as chamadas estatísticas de Wald F-corrigidas, dadas por

$$F_{W1} = \frac{f - K + 2}{f(K-1)} X_W^2 \tag{2.24}$$

que tem distribuição assintótica F com $(K-1)$ e $(f - K + 2)$ graus de liberdade, e

$$F_{W2} = \frac{X_W^2}{J - 1} \tag{2.25}$$

que também tem distribuição assintótica F, porém com $(K-1)$ e f graus de liberdade, com $f = m - H$, em que m e H são os números de conglomerados e estratos do desenho amostral, respectivamente.

Da mesma maneira, a estatística (2.20) pode ser corrigida nessas situações de instabilidade. Neste caso, a estatística F-corrigida é dada por:

$$FX_P^2(\hat{\lambda}_*) = \frac{X_P^2}{(K-1)\hat{\lambda}_*} \tag{2.26}$$

Na literatura existem diversos estudos sobre as estatísticas apresentadas neste trabalho, bem como o comportamento de cada uma nos mais diversos casos, que incluem Bhapkar (1966), Kiang-khang (1984), Larntz (1978) e Thomas e Rao (1987).

2.4.1.3 Análise de Resíduos

Quando uma hipótese do tipo $H_0 : \mathbf{p} = \mathbf{p}_0$ é testada e posteriormente rejeitada, conclui-se que $p_k \neq p_{0k}$ para pelo menos algum k. Neste caso, é possível fazer uma análise de resíduos para ter ideia de onde está a diferença, ou seja, estudar os desvios com relação à H_0. Para uma amostra aleatória simples, os resíduos padronizados são da forma:

$$\hat{e}_k = \frac{\hat{p}_k - p_{0k}}{ep_{IID}(\hat{p}_k)}, \quad k = 1, \ldots, K \tag{2.27}$$

em que $ep_{IID}(\hat{p}_k)$ é a raíz quadrada do k-ésimo elemento da diagonal da matriz de covariância multinomial estimada $\mathbf{P}/n$, com $\hat{\mathbf{P}} = diag(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}'$.

Um valor absoluto grande para $\hat{e}_k$ indica desvios em relação à hipótese nula. Porém, em amostras complexas, esse resíduos padronizados podem ser muito grandes pois os erros-padrões baseados na distribuição multinomial tendem a ser menores que os verdadeiros erros-padrões. Logo, é necessário um ajuste em (2.27) dado por

$$\hat{e}_k = \frac{\hat{p}_k - p_{0k}}{ep_p(\hat{p}_k)}, \quad k = 1, \ldots, K, \tag{2.28}$$

que são chamados de resíduos padronizados baseados no plano amostral.

Se as estimativas $\hat{d}_k$ forem maiores que 1, os resíduos dados em (2.28) serão menores que os dados em (2.27), que por sua vez podem ser tomados como variáveis normais padronizadas sob a hipótese nula, logo podem ser comparados aos valores críticos da distribuição $\mathcal{N}(0, 1)$.

2.4.1.4 Uma aplicação aos dados da Mini-Finland Health Survey

A **Mini-Finland Health Survey** é uma pesquisa finlandesa da área de saúde e foi projetada para obter uma visão abrangente da saúde e da necessidade de cuidados em adultos finlandeses, e desenvolver métodos para monitorar a saúde da população. O plano de amostragem da pesquisa pertence à classe de duas fases de amostragem estratificada por conglomerados. Uma variedade de métodos de coleta de dados foram utilizados e um dos objetivos da pesquisa foi comparar a confiabilidade destes vários métodos (HELIÖVAARA *et al.*, 1993). O público alvo do estudo foi a população finlandesa com idade superior a 30 anos, e o desenho amostral possui um total de 24 estratos e 48 conglomerados (LEHTONEN *et al.*, 1986).

Uma pequena demostração do conjunto de dados da MFH está presente na Tabela 2.8, que consiste de homens com idades entre 30 e 64 anos avaliados na pesquisa e de três variáveis respostas, sendo elas duas binárias (doença crônica e homens que sofrem ou sofreram de riscos para a saúde física no trabalho) e uma contínua (pressão arterial).

Tabela 2.8 – Distribuíção da idade, proporções (em %) de pessoas com doenças crônicas e expostas a riscos de saúde física no trabalho, e pressão arterial em um subgrupo da pesquisa MFH com homens entre 30 e 64 anos de idade

Idade	Amostra n	Amostra %	Doença Crônica	Riscos no Trabalho	Pressão Arterial (em média)
30-34	508	18.8	13.8	12.8	134.0
35-39	384	14.2	21.4	17.4	136.2
40-44	437	16.2	28.4	18.8	138.5
45-49	395	14.6	44.8	18.5	141.9
50-54	379	14.0	52.2	17.4	144.7
55-59	336	12.4	68.5	21.4	151.2
60-64	260	9,6	73.8	21.2	154.3
Total	2699	100	39.8	17.8	141.8

Fonte: Lehtonen e Pahkinen (2004)

Lehtonen e Pahkinen (2004) consideram um teste de bondade de ajuste das distribuição das idades da pesquisa MFH para os subgrupos de homens com idade de 30-44, 45-54 e 55-64 anos de idade. O número de estratos e conglomerados utilizados na pesquisa é de $H = 24$ e $m = 48$, respectivamente. Portanto, $f = m - H = 24$. O objetivo é, além de exemplificar o uso dos testes, mostrar como um número pequeno de conglomerados pode afetar os resultados. As informações utilizadas estão na Tabela 2.9, incluindo os resíduos padronizados baseados no plano.

Tabela 2.9 – Vetores de proporções, efeitos do plano amostral e resíduos padronizados por classes de idade na MFH

Idade	n_k	Proporção Estimada ($\hat{p}_k$)	Proporção Hipotética ($\hat{p}_{0k}$)	EPA_{Kish} ($\hat{d}_k$)	Resíduos ($\hat{e}_k$)
30-44	1329	0.492	0.521	1.51	-2.45
45-54	774	0.287	0.277	1.70	0.88
55-64	596	0.221	0.202	0.43	3.64
Total	2699	1	1		

Fonte: Lehtonen e Pahkinen (2004)

A hipótese nula que está sendo testada é $H_0 : p_k = p_{0k}$, com $k = 1, 2, 3$, e serão calculados os valores da estatística clássica de Pearson e suas correções, bem como a estatística de Wald baseada no plano. Para isso, deve-se ter $\hat{\boldsymbol{V}}_p$, matriz estimada para as proporções estimadas. No texto deste exemplo não foi citado qual método de linearização foi utilizado, mas ela é da forma:

$$\hat{\boldsymbol{V}}_p(\hat{\mathbf{p}}) = 10^{-5} \times \left[\begin{array}{cc|c} 13.9481 & -12.0731 & -1.8750 \\ -12.0731 & 12.9158 & -0.8427 \\ \hline -1.8750 & -0.8427 & 2.7177 \end{array}\right]$$

.

Para obter a estatística de Pearson (2.19), é preciso o cálculo da matriz de covariância populacional e uma estimativa dessa matriz de covariância sob a suposição de distribuição multinomial dada por $\mathbf{P}_0/n = \dfrac{diag(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0'}{2699}$. Este cálculo resulta em:

$$\hat{\mathbf{P}}/n = 10^{-5} \times \left[\begin{array}{cc|c} 9.2464 & -5.3471 & -3.8993 \\ -5.3471 & 7.4202 & -2.0731 \\ \hline -3.8993 & -2.0731 & 5.9724 \end{array}\right]$$

.

E para obter as diferentes correções dessa estatística, precisa-se dos valores dos EPAs listados na quinta coluna da Tabela 2.9 e as seguintes quantidades:

$$\hat{\lambda}_* = \sum_{k=1}^{3} \hat{p}_k p_{0k}^{-1}(1 - \hat{p}_k)\hat{d}_k/2 = 1.17,$$

$$1 + \hat{a}^2 = 2699^2 \sum_{k=1}^{3}\sum_{i=1}^{3} \frac{\hat{V}_p^2(\hat{p}_k\hat{p}_i)/p_{0k}p_{0i}}{2 \times 1.17^2} = 1.37,$$

$$df_S = \frac{K-1}{1+\hat{a}^2} = 1.46.$$

As linhas desenhadas nas matrizes servem para notar que para o cálculo de X_W^2 e X_P^2 dadas, respectivamente, em (2.19) e (2.23), são necessárias apenas as submatrizes (2×2) de $\hat{\boldsymbol{V}}_p(\hat{\mathbf{p}})$ e $\hat{\mathbf{P}}/n$, correspondentes aos dois elementos dos vetores $\hat{\boldsymbol{p}}$ e $\boldsymbol{p}_0$.

Dessa forma, a estatística de Pearson (2.19) resulta em

$$X_P^2 = 10.15$$

com 2 graus de liberdade e p-valor 0.006.

A estatística de Pearson com ajuste de Rao-Scott de primeira ordem, dada em (2.20), resulta em

$$X_P^2(\hat{\lambda}_*) = \frac{10.15}{1.17} = 8.66$$

com 2 graus de liberdade e p-valor 0.013.

O ajuste de Rao-Scott de primeira ordem F-corrigido para a estatística de Pearson, dado em (2.26), é

$$FX_P^2(\hat{\lambda}_*) = \frac{10.15}{2 \times 1.17} = 4.33$$

com 2 e 24 graus de liberdade, e p-valor 0.025.

O ajuste de Rao-Scott de segunda ordem para a estatística de Pearson, dado em (2.21), resulta em

$$X_P^2(\hat{\lambda}_*, \hat{a}^2) = \frac{8.66}{1.37} = 6.30$$

com $2/1.37$ graus de liberdade e p-valor 0.023.

A estatística de Wald baseada no plano amostral, dada em (2.23), é

$$X_W^2 = 15,28$$

com 2 graus de liberdade e p-valor 0.001.

As estatísticas de Wald F-corrigidas, definidas em (2.24) e (2.25), resultam em

$$F_{W1} = \frac{24 - 3 + 2}{24 \times 2} \times 15.28 = 7.32$$

com 2 e 23 graus de liberdade e p-valor 0.003, e

$$F_{W2} = \frac{15.28}{2} = 7.64$$

com 2 e 24 graus de liberdade, e p-valor 0.003.

A Tabela 2.10 resume os valores das diversas estatísticas de testes calculadas, bem como informações comparativas com as respectivas distribuições de referência.

Tabela 2.10 – Valores e p-valores de estatísticas alternativas de teste

Estatística	Tipo	Valor	Distribuição	p-valor
X_P^2	Adequada para IID	11.640	$\chi^2_{(2)}$	0.020
$X_P^2(\hat{\lambda}_*)$	Correções da	8.66	$\chi^2_{(2)}$	0.013
$FX_P^2(\hat{\lambda}_*)$	estatística	4.33	$F_{(2;24)}$	0.025
$X_P^2(\hat{\lambda}_*, \hat{a}^2)$	X_P^2	6.30	$\chi^2_{(1.46)}$	0.023
X_W^2	Baseadas	15.28	$\chi^2_{(2)}$	0.001
F_{W1}	no plano	7.32	$F_{(2;23)}$	0.003
F_{W2}	amostral	7.64	$F_{(2;24)}$	0.003

Dentre os testes calculados, e levando em consideração o fato da instabilidade, poderia ser esperado que a correção de segunda ordem de Rao-Scott ($X_P^2(\hat{\lambda}_*, \hat{a}^2)$) e as estatísticas de

Wald F-corrigidas (F_{W1} e F_{W2}) fornecessem os resultados mais adequados. Porém, pode-se perceber que a estatística de Wald baseada no plano é liberal, muito provavelmente devido ao pequeno número de graus de liberdade disponíveis para a estimação de $\hat{\boldsymbol{V}}_p$, e suas correções, aparentemente, não trazem melhorias neste caso.

A estatística de Pearson com ajuste de primeira ordem de Rao-Scott ($X_P^2(\hat{\lambda}_*)$) pode ser desconsiderada, pois deve ser preferida apenas quando não é possível obter $\hat{\boldsymbol{V}}_p$. Nos casos em que é possível obter essa estimativa, a correção de segunda ordem sempre deve ser a escolhida, pois é mais precisa.

Portanto, analisando os resultados obtidos para este caso, a estatística de Pearson com ajuste de segunda ordem de Rao-Scott $X_P^2(\hat{\lambda}_*, \hat{a}^2)$ deve ser preferida. Se apenas as estimativas do efeito do plano amostral $\hat{d}_k$ fossem fornecidas, e não $\hat{\boldsymbol{V}}_p$, o ajuste de Rao-Scott de primeira ordem F-corrigido para a estatística de Pearson $FX_P^2(\hat{\lambda}_*)$ deveria ser o escolhido, pois diminui o efeito de instabilidade.

2.4.2 Tabelas de duas entradas

Os principais testes em tabelas de duas entradas são os testes de homogeneidade e independência. O teste de homogeneidade é apropriado para estudar a igualdade das distribuições condicionais de uma variável resposta categórica correspondente a diferentes níveis de uma variável preditora também categórica. O teste de independência é adequado para estudar a associação entre duas variáveis categóricas. Segundo Pessoa e Silva (1998), 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, e apesar de conceitualmente distintas, as duas hipóteses podem ser testadas, no caso de amostragem aleatória simples, utilizando a mesma estatística de teste. Já nos casos envolvendo amostragem complexa, as estatísticas de testes para as duas hipóteses devem ser corrigidas, e de maneira diferente uma da outra.

Como preliminar, considere a Tabela 2.11.

Tabela 2.11 – Tabela (2×2) de proporções

	Variável 1		
Variável 2	1	2	Total
1	p_{11}	p_{12}	p_{1+}
2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

A hipótese de independência, cujo objetivo é verificar se as duas variáveis estão associadas ou não, é dada por

$$H_0: \quad p_{lc} = p_{+c}p_{l+}, \quad l, c = 1, 2. \tag{2.29}$$

em que p_{lc} representa a proporção populacional de unidades no nível l da variável 1 e no nível c da variável 2, ou seja, $p_{lc} = \dfrac{N_{lc}}{N}$, $p_{+c} = \sum_l p_{lc}$ e $p_{l+} = \sum_c p_{lc}$ são as proporções marginais na tabela, e $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

No caso de amostragem aleatória simples, a estatística de teste de Pearson é dada por

$$X_P^2(I) = n \sum_{l=1}^{2} \sum_{c=1}^{2} \frac{(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c})^2}{\hat{p}_{l+}\hat{p}_{+c}} \tag{2.30}$$

em que $\hat{p}_{lc} = n_{lc}/n$, $\hat{p}_{+c} = \sum_l \hat{p}_{lc}$ e $\hat{p}_{l+} = \sum_c \hat{p}_{lc}$, e cuja distribuição assintótica é qui-quadrado com 1 grau de liberdade.

Já a hipótese de homogeneidade, cujo objetivo é verificar se uma variável aleatória se comporta de modo homogêneo em várias subpopulações, é dada por

$$H_0: \quad p_{1c} = p_{2c}, \quad c = 1, 2 \tag{2.31}$$

em que p_{lc} representa a proporção da linha l de unidades na coluna c, ou seja, agora as proporções na linha somam 1, isto é, $p_{11} + p_{12} = p_{21} + p_{22} = 1$.

No caso de amostragem aleatória simples, a estatística de teste de Pearson é dada por

$$X_P^2(H) = \sum_{l=1}^{2} \sum_{c=1}^{2} \frac{n_{l+}(\hat{p}_{lc} - \hat{p}_{+c})^2}{\hat{p}_{+c}} \tag{2.32}$$

em que $n_{l+} = \sum_c n_{lc}$ e $\hat{p}_{lc} = n_{lc}/n_{l+}$, e cuja distribuição assintótica também é qui-quadrado com um grau de liberdade.

Os testes de independência e homogeneidade são definidos sobre o vetor de proporções de distribuições multinomiais. No caso de independência, tem-se uma distribuição multinomial com vetor de probabilidades $(p_{11}, p_{12}, p_{21}, p_{22})$, e no caso do teste de homogeneidade, tem-se duas multinomiais com vetores de probabilidades (p_{11}, p_{12}) e (p_{21}, p_{22}). O processo de contagem que gera essas multinomiais pressupõe que cada observação seja independente uma da outra e com distribuições iguais. Porém, Pessoa e Silva (1998) explicam que em uma amostragem complexa, essas pressuposições deixam de ser válidas e as frequências nas celas são estimadas de acordo com o plano amostral utilizado.

Denotando por $\hat{N}_{lc}$ o estimador do número de observações na cela (l, c) na população e por $\hat{n}_{lc} = \left(\dfrac{\hat{N}_{lc}}{\hat{N}}\right) \times n$ o valor padronizado de $\hat{N}_{lc}$ de modo que $\sum_{l=1}^{2} \sum_{c=1}^{2} \hat{n}_{lc} = n$, tem-se os estimadores das proporções nas celas dados por $\hat{p}_{lc} = \hat{n}_{lc}/n$ no caso do teste de independência, e por $\hat{p}_{lc} = \hat{n}_{l+}/n$ no caso do teste de homogeneidade. As estatísticas $X_P^2(I)$ e $X_P^2(H)$ calculadas com $\hat{n}_{lc}$ no lugar de n_{lc} não tem distribuição assintótica qui-quadrado com 1 grau de liberdade, portanto devem ser corrigidas.

2.4.2.1 Teste de Homogeneidade

Na literatura de análise de dados amostrais complexos, o teste de homogeneidade é geralmente usado para comparar distribuições de uma variável categórica para um conjunto de regiões não superpostas, a partir de amostras independentes obtidas através de um plano amostral com vários estágios.

Assumindo que as colunas são formadas pelas classes da variável resposta e as linhas são as regiões, a tabela para a população é da forma da Tabela 2.12.

Tabela 2.12 – Proporções de linhas em uma tabela $L \times C$ para um teste de homogeneidade.

	Variável Resposta						
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
:							
l	p_{l1}	p_{l2}	...	p_{lc}	...	p_{lC}	1
:							
L	p_{L1}	p_{L2}	...	p_{Lc}	...	p_{LC}	1

Casos em que se pretende comparar três ou mais regiões são bastante complexos e são tratados em Rao e Thomas (1988). Nos casos envolvendo apenas duas regiões, a hipótese nula para o teste de homogeneidade pode ser escrita como

$$H_0: \quad \boldsymbol{p}_1 = \boldsymbol{p}_2$$

em que $\boldsymbol{p}_l = (p_{l1}, \ldots, p_{l,c-1})'$ denota o vetor de proporções da l-ésima região. Cada vetor é composto, então, de $(c-1)$ elementos, pois a somatória dos c elementos deve ser igual a 1 independente da região.

Sejam $\boldsymbol{p}_0 = (p_{+1}, \ldots, p_{+,c-1})'$ o vetor comum de proporções sob H_0, desconhecido, e $\hat{\boldsymbol{p}}_l = (\hat{p}_{l1}, \ldots, \hat{p}_{lc-1})'$ os vetores de proporções regionais estimadas baseados em amostras independentes para as diferentes regiões, em que $\hat{p}_{lc} = \hat{N}_{lc}/\hat{N}_{l+}$ é um estimador consistente da proporção p_{lc} na população correspondente, e $\hat{N}_{lc}$ e $\hat{N}_{l+}$ são estimadores ponderados das frequências nas celas e nas marginais de linha da tabela, respectivamente. Estes estimadores levam em consideração as probabilidades desiguais de inclusão na amostra.

Sejam também $\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_1)$ e $\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_2)$ estimadores consistentes das matrizes de variância de aleatorização dos vetores $\hat{\boldsymbol{p}}_1$ e $\hat{\boldsymbol{p}}_2$, respectivamente. A estatística de Wald baseada no plano amostral é dada, no caso de $L = 2$ regiões, por

$$X_W^2(H) = (\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2)'[\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_1) + \hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_2)]^{-1}(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) \tag{2.33}$$

cuja distribuição assintótica é qui-quadrado com $[(2-1)(C-1) = (C-1)]$ graus de liberdade.

Nos casos de instabilidade, ou seja, quando há um pequeno número de graus de liberdade disponíveis para o cálculo de $\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_1)$ e $\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_2)$, $X^2_W(H)$ pode não funcionar adequadamente. Nestes casos, usa-se, alternativamente, as estatísticas de Wald F-corrigidas dada por

$$F_{W1} = \frac{f-(C-1)+1}{f(C-1)} X^2_W(H) \tag{2.34}$$

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

$$F_{W2} = \frac{X^2_W(H)}{C-1} \tag{2.35}$$

que tem distribuição de referência F com $(C-1)$ e f graus de liberdade. Estas estatísticas podem diminuir o efeito da instabilidade quando f não é grande em relação ao número de classes C da variável resposta.

A estatística de Pearson para o teste de homogeneidade, no caso de $L = 2$ regiões, é dada por

$$X^2_P(H) = (\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2)' \left[\frac{\hat{\boldsymbol{P}}}{\hat{n}_{1+}} + \frac{\hat{\boldsymbol{P}}}{\hat{n}_{2+}} \right]^{-1} (\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2), \tag{2.36}$$

em que $\hat{\boldsymbol{P}} = diag(\hat{\boldsymbol{p}}_0) - \hat{\boldsymbol{p}}_0\hat{\boldsymbol{p}}_0'$, $\hat{\boldsymbol{p}}_0$ é estimador do vetor de proporções sob a hipótese de homogeneidade e $\hat{\boldsymbol{P}}/\hat{n}_{1+}$ e $\hat{\boldsymbol{P}}/\hat{n}_{2+}$ são os estimadores da matriz de covariâncias de $\hat{\boldsymbol{p}}_0$ na primeira e na segunda região, respectivamente.

A estatística de Pearson $X^2_P(H)$ é válida somente para amostragem aleatória simples, dado que o estimador da matriz de covariância baseia-se nas hipóteses relativas à distribuição multinomial. Logo, tem distribuição qui-quadrado com $(C-1)$ graus de liberdade. Quando o desenho amostral é complexo, essa estatística requer ajustes, os quais são similares àqueles apresentados para o teste de bondade de ajuste, porém com fórmulas diferentes.

Para introduzir em $X^2_P(H)$ o ajuste de primeira ordem de Rao-Scott é preciso calcular as estimativas de efeitos do plano amostral das estimativas das proporções nas linhas em ambas as regiões, e para o ajuste de segunda ordem é necessária a matriz de efeito multivariado do plano amostral. As estimativas dos efeitos do plano amostral na região l são da forma

$$\hat{d}_{lc} = \hat{d}(\hat{p}_{lc}) = \frac{\hat{n}_{l+}\hat{V}_{lc}}{\hat{p}_{+c}(1-p_{+c})}, \quad l = 1,2 \quad e \quad c = 1,\ldots,C$$

em que $\hat{V}_{lc}$ é o c-ésimo elemento da diagonal de $\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_l)$. E a matriz estimada de efeito multivariado do plano amostral é da forma

$$\hat{\boldsymbol{D}} = \frac{\hat{n}_{1+} \times \hat{n}_{2+}}{\hat{n}_{1+} + \hat{n}_{2+}} \hat{\boldsymbol{P}}^{-1} (\hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_1) + \hat{\boldsymbol{V}}_p(\hat{\boldsymbol{p}}_2)).$$

A correção de primeira ordem de Rao-Scott é dada por

$$X^2_P(H, \hat{\lambda}_*) = \frac{X^2_P(H)}{\hat{\lambda}_*} \tag{2.37}$$

em que

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

é um estimador da média $\bar{\lambda}$ dos autovalores λ_c da matriz desconhecida $\boldsymbol{D}$ de efeito multivariado do plano amostral. Como pode ser notado, essa correção pode ser obtida sem ser necessário calcular os autovalores.

A distribuição assintótica de (2.37) é qui-quadrado com $(C-1)$ graus de liberdade e espera-se que esta estatística forneça resultados satisfatórios quando a variação das estimativas dos autovalores $\hat{\lambda}_*$ for pequena. Quando essa variação for significativa, é recomendado o ajuste de segunda ordem de Rao-Scott, dado por

$$X_P^2(H, \hat{\lambda}_*, \hat{a}^2) = \frac{X_P^2(H, \hat{\lambda}_*)}{1+\hat{a}^2} \tag{2.38}$$

em que $\hat{a}^2$ é o quadrado do coeficiente de variação dos quadrados das estimativas dos autovalores $\hat{\lambda}_c$, que é dado pela fórmula

$$\hat{a}^2 = \sum_{c=1}^{C}\frac{\hat{\lambda}_c^2}{(C-1)\hat{\lambda}_*^2} - 1$$

em que $\sum_{c=1}^{C}\hat{\lambda}_c^2 = tr(\hat{\boldsymbol{D}}^2)$. A estatística (2.38) tem distribuição assintótica qui-quadrado com graus de liberdade com ajuste de Satterthwaite dados por $gl_S = \dfrac{C-1}{1+\hat{a}^2}$.

Nos casos de instabilidade, é necessária uma correção ao ajuste de primeira ordem de Rao-Scott. Essa correção é dada por

$$FX_P^2(H, \hat{\lambda}_*) = \frac{X_P^2(H, \hat{\lambda}_*)}{C-1} \tag{2.39}$$

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

2.4.2.1.1 Análise de Resíduos

Quando a hipótese nula de homogeneidade for rejeitada, concluímos que há diferença significativa em pelo menos duas distribuições entre as distribuições das regiões com relação às categorias da variável resposta. Neste caso, a fim de detectar onde está essa diferença, os resíduos padronizados podem ser calculados para que se possa estudar os desvios com relação a H_0. Usando as estimativas de efeitos do plano amostral $\hat{d}_{lc}$, os resíduos padronizados baseados no plano amostral são dados por

$$\hat{e}_{lc} = \frac{\hat{p}_{lc} - \hat{p}_{+c}}{ep_p(\hat{p}_{lc} - \hat{p}_{+c})}, \quad l = 1, 2 \quad e \quad c = 1, \ldots, C$$

em que um estimador para $ep_p(\hat{p}_{lc} - \hat{p}_{+c})$ é a raiz quadrada de

$$\hat{V}_p(\hat{p}_{1c} - \hat{p}_{+c}) = \frac{\hat{n}_{2+}(\hat{n}_{2+}\hat{d}_{1c} + \hat{n}_{1+}\hat{d}_{2c})}{(\hat{n}_{1+} + \hat{n}_{2+})^2} \times \frac{\hat{p}_{+c}(1 - \hat{p}_{+c})}{\hat{n}_{1+}}, \quad c = 1, \ldots, C$$

para a primeira região, e

$$\hat{V}_p(\hat{p}_{2c} - \hat{p}_{+c}) = \frac{\hat{n}_{1+}(\hat{n}_{2+}\hat{d}_{1c} + \hat{n}_{1+}\hat{d}_{2c})}{(\hat{n}_{1+} + \hat{n}_{2+})^2} \times \frac{\hat{p}_{+c}(1 - \hat{p}_{+c})}{\hat{n}_{2+}}, \quad c = 1, \ldots, C$$

para a segunda região.

A partir dessas fórmulas, pode-se inferir que quando há uma correlação intra-classe positiva, os valores dos resíduos padronizados baseados no plano amostral são menores do que os valores obtidos para amostragem aleatória simples, em que $\hat{d}_{1c} = \hat{d}_{2c} = 1$. Quando as variáveis que constituem a tabela cruzada são supostas independentes, os termos $\hat{e}_{lc}$ tem distribuição assintótica normal padronizada, logo podem ser comparados aos valores críticos da distribuição $\mathcal{N}(0, 1)$.

2.4.2.2 Teste de Independência

O teste de independência é utilizado com o objetivo de verificar se há associação entre duas variáveis categóricas dentro de uma única população. A tabela para proporções dessa população não tem marginais fixas e é da forma da Tabela 2.13.

Tabela 2.13 – Proporções por cela em uma tabela $L \times C$ para um teste de independência

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

A hipótese de independência é dada por:

$$H_0 : p_{lc} = p_{l+}p_{+c}, \quad l = 1, \ldots, L-1 \quad e \quad c = 1, \ldots, C-1 \tag{2.40}$$

em que se tem $p_{lc} = N_{lc}/N$, $p_{l+} = \sum_{c=1}^{C} p_{lc}$, $p_{+c} = \sum_{l=1}^{L} p_{lc}$ e $\sum_{c=1}^{C}\sum_{l=1}^{L} p_{lc} = 1$.

Alternativamente, essa hipótese pode ser escrita como

$$H_0 : F_{lc} = p_{lc} - p_{l+}p_{+c} = 0, \quad l = 1, \ldots, L-1 \quad e \quad c = 1, \ldots, C-1, \tag{2.41}$$

em que os elementos F_{lc} são, as diferenças residuais entre as proporções desconhecidas nas celas e seus valores esperados sob a hipótese nula, a qual trata essas diferenças como sendo iguais a zero.

Sejam $\boldsymbol{F} = (F_{11}, \ldots, F_{1,C-1}, \ldots, F_{L-1,1}, \ldots, F_{L-1,C-1})$ o vetor formado pelos contrastes F_{lc} arranjados em ordem de linhas e seu estimador

$$\hat{\boldsymbol{F}} = (\hat{F}_{11}, \ldots, \hat{F}_{1,C-1}, \ldots, \hat{F}_{L-1,1}, \ldots, \hat{F}_{L-1,C-1}),$$

em que $\hat{F}_{lc} = \hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c}$ e $\hat{p}_{lc} = \hat{n}_{lc}/n$. A estatística de Wald baseada no plano amostral para o teste de independência é dada por:

$$X^2_W(I) = \hat{\boldsymbol{F}}'\hat{\boldsymbol{V}}_F^{-1}\hat{\boldsymbol{F}} \tag{2.42}$$

em que $\hat{\boldsymbol{V}}_F$ é a estimativa da matriz de covariância de $\hat{\boldsymbol{F}}$ e pode ser obtida pelo método de linearização de Taylor considerando o vetor de contrastes $\boldsymbol{F}$ como uma função não-linear do vetor $\boldsymbol{p}$, isto é, $\boldsymbol{F} = g(\boldsymbol{p}) = g(p_{11}, \ldots, p_{1,C-1}, \ldots, p_{L-1,1}, \ldots, p_{L-1,C-1})$. Assim, a matriz de covariância de $\hat{\boldsymbol{F}}$ pode ser estimada por

$$\hat{\boldsymbol{V}}_F = \Delta g(\hat{\boldsymbol{p}})\hat{\boldsymbol{V}}_p^{-1}\Delta g(\hat{\boldsymbol{p}})' \tag{2.43}$$

em que $\Delta g(\boldsymbol{p})$ é a matriz jacobiana de dimensão $[(L-1)(C-1) \times (L-1)(C-1)]$ dada por

$$\Delta g(\boldsymbol{p}) = \left[\frac{\partial g}{\partial p_{11}}, \ldots, \frac{\partial g}{\partial p_{1,C-1}}, \ldots, \frac{\partial g}{\partial p_{L-1,1}}, \ldots, \frac{\partial g}{\partial p_{L-1,C-1}}\right]$$

e $\hat{\boldsymbol{V}}_p$ é uma estimativa consistente da matriz de covariância de $\hat{\boldsymbol{p}}$ (PESSOA; SILVA, 1998).

Assim como no teste de bondade de ajuste e homogeneidade, a estatística de Wald para o teste de independência também pode ter problemas com instabilidade quando um número pequeno de graus de liberdade f estiver disponível para estimar $\hat{\boldsymbol{V}}_F$. As estatísticas de Wald F-corrigidas para estes casos são dadas por

$$F_{1W} = \frac{f - (L-1)(C-1) - 1}{f(L-1)(C-1)} X^2_W(I) \tag{2.44}$$

que tem distribuição F com $[(L-1)(C-1)]$ e $[f - (L-1)(C-1) - 1]$ graus de liberdade, e por

$$F_{2W} = \frac{X^2_W(I)}{(L-1)(C-1)} \tag{2.45}$$

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

A estatística de Pearson para o teste de independência é dada por:

$$X^2_P(I) = n\sum_{l=1}^{L}\sum_{c=1}^{C}\frac{(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c})^2}{\hat{p}_{l+}\hat{p}_{+c}} = n\hat{\boldsymbol{F}}'\hat{\boldsymbol{P}}_{0F}^{-1}\hat{\boldsymbol{F}} \tag{2.46}$$

em que

$$\hat{\boldsymbol{P}}_{0F} = \Delta g(\hat{\boldsymbol{p}})\hat{\boldsymbol{P}}_0\Delta g(\hat{\boldsymbol{p}})', \tag{2.47}$$

$\hat{\boldsymbol{P}}_0 = diag(\hat{\boldsymbol{p}}_0) - \hat{\boldsymbol{p}}_0\hat{\boldsymbol{p}}_0'$, $\hat{\boldsymbol{P}}_0/n$ é uma estimativa da matriz $[(L-1)(C-1) \times (L-1)(C-1)]$ da matriz de covariância multinomial de $\hat{\boldsymbol{p}}$ sob a hipótese nula, $\hat{\boldsymbol{p}}_0$ é o vetor com componentes $\hat{p}_{l+}\hat{p}_{+c}$, e $diag(\hat{\boldsymbol{p}}_0)$ é a matriz diagonal com elementos $\hat{p}_{l+}\hat{p}_{+c}$ na diagonal. Sob a hipótese nula, $X_P^2(I)$ tem distribuição assintótica qui-quadrado com $[(L-1)(C-1)]$ graus de liberdade.

A estatística (2.46) é utilizada quando o desenho amostral é composto por amostragem aleatória simples, porém, como nos testes anteriores, também pode ser corrigida de modo a ter a mesma distribuição assintótica. As correções propostas são os ajustes de primeira e segunda ordem de Rao-Scott que baseiam-se nos autovalores da matriz estimada de efeito multivariado do plano amostral, dada por

$$\hat{\boldsymbol{D}} = n\hat{\boldsymbol{P}}_{0F}^{-1}\hat{\boldsymbol{V}}_F \tag{2.48}$$

em que $\hat{\boldsymbol{V}}_F$ foi definido em (2.43) e $\hat{\boldsymbol{P}}_{0F}$ foi definido em (2.47).

A correção de primeira ordem de Rao-Scott é dada por:

$$X_P^2(I, \hat{\lambda}_*) = \frac{X_P^2(I)}{\hat{\lambda}_*} \tag{2.49}$$

em que

$$\hat{\lambda}_* = \frac{1}{(L-1)(C-1)}\sum_{l=1}^{L}\sum_{c=1}^{C}\frac{\hat{p}_{lc}(1-\hat{p}_{lc})}{\hat{p}_{l+}\hat{p}_{+c}}\hat{d}_{lc} - \sum_{l=1}^{L}(1-\hat{p}_{l+})\hat{d}_{l+} - \sum_{c=1}^{C}(1-\hat{p}_{+c})\hat{d}_{+c}$$

é um estimador da média $\bar{\lambda}$ dos autovalores desconhecidos da matriz $\boldsymbol{D}$ de efeitos multivariados do plano amostral e pode ser obtido sem calcular $\hat{\boldsymbol{D}}$. Sob a hipótese nula, $X_P^2(I, \hat{\lambda}_*)$ tem distribuição assintótica qui-quadrado com $[(L-1)(C-1)]$ graus de liberdade, e espera-se que essa estatística funcione adequadamente quando a variação dos autovalores estimados $\hat{\lambda}_l$ for pequena.

Nos casos de instabilidade, um ajuste é necessário em (2.49). Essa correção é dada por:

$$FX_P^2(I, \hat{\lambda}_*) = \frac{X_P^2(I, \hat{\lambda}_*)}{(L-1)(C-1)} \tag{2.50}$$

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

A correção de segunda ordem de Rao-Scott para a estatística de Pearson $X_P^2(I)$ para o teste de independência requer que esteja disponível a matriz de covariância estimada $\hat{\boldsymbol{V}}_p$ dos estimadores das proporções nas celas $\hat{p}_{lc}$, e é dada por

$$X_P^2(I, \hat{\lambda}_*, \hat{a}^2) = \frac{X_P^2(I)}{\hat{\lambda}_*(1+\hat{a}^2)} \tag{2.51}$$

em que

$$\hat{\lambda}_* = \frac{tr(\hat{\boldsymbol{D}})}{(L-1)(C-1)}$$

é um estimador da média dos autovalores $\hat{\lambda}_l$ da matriz estimada de efeito multivariado $\hat{\boldsymbol{D}}$ dada em (2.48), e

$$\hat{a}^2 = \sum_{k=1}^{(L-1)(C-1)} \frac{\hat{\lambda}_k^2}{(L-1)(C-1)\hat{\lambda}_*^2} - 1$$

é um estimador do quadrado do coeficiente de variação dos autovalores desconhecidos de $\boldsymbol{D}$, λ_k, $k = 1, \ldots, (L-1)(C-1)$.

Um estimador da soma dos quadrados dos autovalores é

$$\sum_{k=1}^{(L-1)(C-1)} \hat{\lambda}_k^2 = tr(\hat{\boldsymbol{D}}^2).$$

Sob a hipótese nula, a estatística $X_P^2(I, \hat{\lambda}_*, \hat{a}^2)$ tem distribuição assintótica qui-quadrado com graus de liberdade com ajuste de Satterthwaite, dados por

$$gl_S = \frac{(L-1)(C-1)}{1+\hat{a}^2}.$$

2.4.2.2.1 Análise de Resíduos

Se a hipótese nula de independência for rejeitada, os resíduos padronizados baseados no plano amostral podem ser úteis para verificar os desvios com relação à H_0. Esses resíduos são da forma

$$\hat{e}_{lc} = \frac{\hat{F}_{lc}}{ep(\hat{F}_{lc})} \tag{2.52}$$

em que $ep(\hat{F}_{lc})$ é a raiz quadrada da variância estimada em (2.43). Quando há uma correlação intra-classe positiva dentro dos conglomerados, estes resíduos tendem a ser menores que seus correspondentes calculados assumindo amostragem aleatória simples.

2.4.2.3 Uma aplicação do teste de homogeneidade aos dados da Occupational Health Care Survey

Na Finlândia, todo funcionário, público ou privado, é munido de um plano de saúde ocupacional, em que o foco é a prevenção de doenças e problemas provenientes do trabalho através do bem-estar físico, mental e social do indivíduo. Por lei, as empresas devem promover métodos para garantir a segurança de seus funcionários, através da prevenção de doenças. Neste contexto, a **Occupational Health Care Survey** é uma pesquisa por amostragem cujo objetivo é

fiscalizar e avaliar o funcionamento da lei. Para isso, diversas formas de estudos podem ser feitos, inclusive inferências analíticas dos dados obtidos por essa pesquisa.

A pesquisa OHC, na qual a população alvo tinha em torno de 2 milhões de funcionários e mais de 100 mil empresas/indústrias, foi focada em estabelecimentos de todos os setores, exceto da agricultura e da silvicultura, e seu desenho amostral é um exemplo de amostragem estratificada por conglomerados em um e dois estágios. Além disso, no desenho da pesquisa OHC, um número grande de conglomerados está disponível, os quais produzem efeitos perceptíveis para diversas variáveis resposta. A descrição detalhada do desenho amostral da pesquisa OHC pode ser encontrada em Lehtonen (1988).

Lehtonen e Pahkinen (2004) utilizaram os dados da pesquisa OHC para estudos com diversos fins analíticos, entre eles testes de homogeneidade e independência para algumas variáveis. A Tabela 2.14 apresenta estes dados categorizados por sexo e idade do pesquisado, e funcionários com doenças crônicas e expostos a riscos para a saúde física de trabalho. Estes dados provém de uma amostragem com 250 conglomerados e 5 estratos, e correspondem a um subgrupo composto por 7841 empregados de empresas/indústrias com no mínimo 10 funcionários.

Tabela 2.14 – Dados da pesquisa OHC categorizados por sexo, idade e proporções (%) de pessoas com doenças crônicas e de pessoas expostas a riscos para a saúde física de trabalho

Sexo	Idade	Amostra n	Amostra %	Doença Crônica	Riscos no Trabalho
Masculino		4485	57.2	29.3	46.0
Feminino		3356	42.8	29.2	19.4
Masculino	15-24	504	6.4	15.5	52.8
	25-34	1355	17.3	19.8	50.8
	35-44	1453	18.5	27.1	42.9
	45-54	847	10.8	44.2	41.9
	55-64	326	4.2	61.3	39.3
Feminino	15-24	418	5.3	16	19.1
	25-34	993	12.7	18.9	18.9
	35-44	1002	12.8	26.5	17.9
	45-54	681	8.7	43.5	18.5
	55-64	262	3.3	61.8	29.4
Ambos os Sexos	15-24	922	11.8	15.7	37.5
	25-34	2348	29.9	19.4	37.4
	35-44	2455	31.3	26.9	32.7
	45-54	1528	19.5	43.8	31.5
	55-64	588	7.5	61.6	34.9
Total		7841	100	29.2	34.6

Fonte: Lehtonen e Pahkinen (2004)

Com o objetivo de verificar se as distribuições de frequência para os dois grupos,

empresa pública e empresa privada, da váriavel categórica "Tipo de Empresa" se comportam de maneira similar com relação às classes da variável "Risco no Trabalho", foi testada a hipótese de homogeneidade para esses grupos. A variável "Risco no Trabalho" foi categorizada, como mostra a Tabela 2.15, em três classes com relação à seriedade dos riscos, em que a classe 3 refere-se aos casos mais graves.

Tabela 2.15 – Proporções na classes da variável "Riscos no Trabalho" em empresas públicas e em empresas privadas, com o efeito do plano amostral nas celas entre parenteses

	Riscos no Trabalho				Tamanho
Tipo de Empresa	1	2	3	Total	da Amostra
Empresa Pública	0.2939	0.3345	0.3716	1	1184
	(2.02)	(1.24)	(1.74)		
Empresa privada	0.3526	0.3216	0.3258	1	6657
	(1.73)	(1.23)	(1.57)		
Ambas	0.3437	0.3236	0.3327	1	7841

Fonte: Lehtonen e Pahkinen (2004)

A hipótese de homogeneidade de interesse é $H_0 : p_{1c} = p_{2c}, c = 1, 2, 3$. Para o cálculo das estatísticas de Wald são necessárias, primeiramente, as estimativas das matrizes de covariâncias completas dos vetores $\hat{\boldsymbol{p}}_1$ e $\hat{\boldsymbol{p}}_2$, que são dadas por

$$\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_1) = 10^{-5} \times \left[\begin{array}{cc|c} 35.3394 & -12.1408 & -23.1986 \\ -12.1408 & 23.3570 & -11.2161 \\ \hline -23.1986 & -11.2161 & 34.4148 \end{array}\right]$$

e

$$\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_2) = 10^{-5} \times \left[\begin{array}{cc|c} 5.9177 & -2.3978 & -3.5200 \\ -2.3978 & 4.0417 & -1.6439 \\ \hline -3.5200 & -1.6439 & 5.1639 \end{array}\right].$$

Como $C - 1 = 2$, são necessárias somente as primeiras submatrizes (2×2) das estimativas $\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_1)$ e $\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_2)$, respectivamente, para o cálculo da estatística de Wald e dos ajustes de Rao-Scott. A estatística de Wald baseada no plano, calculada como em (2.33), resulta em

$$X^2_W(H) = 8,62$$

com 2 graus de liberdade e p-valor igual a 0,0134.

As estatísticas de Wald F-corrigidas, definidas em (2.34) e (2.35), resultam em

$$F_{W1} = \frac{245 - 2 + 1}{245 \times 2} \times 8,62 = 4,29$$

com 2 e 244 graus de liberdade e p-valor gual a 0,0147, e

$$F_{W2} = \frac{8,62}{2} = 4,31$$

com 2 e 245 graus de liberdade e p-valor igual a 0,0144. Pode-se notar que as estatísticas corrigidas não causam grande efeito em $X^2_W(H)$. Isso se deve ao fato de não haver instabilidade, ou seja, o número de unidades primárias (conglomerados) é suficientemente grande para a estimação de $\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_1)$ e $\hat{\mathbf{V}}_p(\hat{\boldsymbol{p}}_2)$.

A estatística clássica de Pearson, dada em (2.36), resulta em

$$X^2_P(H) = 16,93$$

com p-valor igual a 0,0002, mostrando que esta estatística é bastante liberal com relação a $X^2_W(H)$, como esperado.

Para os diferentes ajustes de Rao-Scott, é necessário o cálculo das seguintes quantidades

$$\hat{\mathbf{D}} = \begin{bmatrix} 2.01374 & -0.03663 \\ 0.35554 & 1.23977 \end{bmatrix},$$

$\hat{\lambda}_* = \dfrac{tr(\hat{\boldsymbol{D}})}{2} = 1,627$, $\displaystyle\sum_{c=1}^{2} \hat{\lambda}^2_c = tr(\hat{\boldsymbol{D}}^2) = 5,566$ e $\hat{a}^2 = \displaystyle\sum_{c=1}^{2} \frac{\hat{\lambda}^2_c}{2 \times 1,627^2} - 1 = \frac{5,566}{5,294} = 0,051$.

Dessa forma, a estatística de Pearson com ajuste de primeira ordem de Rao-Scott, definida em (2.37), é

$$X^2_P(H, \hat{\lambda}_*) = \frac{16,93}{1,627} = 10,4$$

com p-valor igual a 0,0055.

O ajuste de Rao-Scott de primeira ordem F-corrigido para a estatística de Pearson, ddefinido em (2.39), é

$$FX^2_P(H, \hat{\lambda}_*) = \frac{10,4}{2} = 5,2$$

com 2 e 244 graus de liberdade, e p-valor igual a 0,0061, indicando, novamente, que as estatísticas sugeridas para os casos de instabilidade não produzem efeito significativo neste caso.

O ajuste de Rao-Scott de segunda ordem para a estatística de Pearson, definido em (2.38), resulta em

$$X^2_P(H, \hat{\lambda}_*, \hat{a}^2) = \frac{10,4}{1 + 0,051} = 9,89$$

com $gl_S = 2/(1 + 0,051) = 1,9$ e p-valor igual a 0,0063.

Estes resultados mostram que as estatísticas de Pearson corrigidas são um pouco liberais com relação à estatística de Wald baseada no plano amostral, indicando que esta estatística funciona mais adequadamente neste caso.

Tabela 2.16 – Resíduos padronizados baseados no plano amostral

Riscos no Trabalho	Empresa Pública $\hat{e}_{1c}$	Empresa Privada $\hat{e}_{2c}$
1	-2.79	2.79
2	0.78	-0.78
3	2.35	-2.35

Fonte: Lehtonen e Pahkinen (2004)

Como, ao nível de 5% de significância, os testes rejeitam a hipótese nula, o próximo passo para uma investigação completa é analisar os resíduos padronizados baseados no plano amostral para as três classes da variável "Riscos no Trabalho", que se encontram na Tabela 2.16.

Os valores críticos referentes a $\alpha = 0.01$ da distribuição $\mathcal{N}(0, 1)$ são -2,33 e 2,33. Como os resíduos referentes às primeira e terceira classes estão fora deste intervalo, pode-se concluir que, ao nível de 1% de significância, a hipótese nula foi rejeitada devido à essas duas classes.

3 DESENVOLVIMENTO

Como um dos objetivos deste trabalho, foi feita, após a revisão bibliográfica de algumas formas e métodos disponíveis na literatura para incorporar o esquema amostral em inferências analíticas, um levantamento de exemplos com o objetivo de fazer um estudo e uma discussão de trabalhos publicados em jornais, revistas ou periódicos, cujos assuntos abordados remetem à teoria exposta na sessão anterior como por exemplo: a questão do uso dos pesos amostrais na inferência dos parâmetros de um modelo de superpopulação; o impacto de ignorar o plano amostral na análise dos dados; e a performance de algumas estatísticas para análises de tabelas de contingência.

Foram escolhidos os trabalhos de Pfeffermann (1993), Holt, Smith e Winter (1980) e Holt, Scott e Ewings (1980).

3.1 Exemplo 1

Antes de discutir o estudo feito por Pfeffermann (1993), segue um breve resumo sobre as abordagens encontradas na literatura usadas para inferência em amostragem. São elas: inferência baseada no desenho (**design-based approach**), inferência baseada em modelos (**model-based approach**) e inferência baseada em modelos de superpopulação (**model-assisted approach**), que é uma junção das duas anteriores.

A inferência baseada no desenho (amostragem clássica) em pesquisas amostrais considera uma população finita $\mathcal{U} = \{1, 2, \ldots, N\}$ da qual é selecionada uma amostra $s = \{1, 2, \ldots, n\}$ segundo um plano amostral caracterizado por $p(s)$, probabilidade de ser selecionada a amostra s. Os valores $y_1, y_2, \ldots, y_N$ de uma variável de interesse Y na população finita são considerados fixos, entretanto desconhecidos. A partir dos valores $y_1, y_2, \ldots, y_n$ observados na amostra são feitas inferências a respeito de funções dos valores populacionais, como totais, médias etc. Nesta abordagem, a aleatoriedade é introduzida pelo pesquisador no processo para obtenção dos dados, através do planejamento amostral $p(s)$ utilizado e as distribuições das estatísticas de interesse são derivadas a partir dessa distribuição de aleatorização (p-**distribution**).

Uma alternativa à abordagem anterior é a abordagem baseada em modelos, em que um modelo (de superpopulação) serve como base para a inferência sobre os parâmetros da população, no contexto de pesquisas por amostragem. A diferença fundamental desta abordagem com a outra é que os valores $y_1, y_2, \ldots, y_N$ associados as N unidades da população são considerados como realizações das variáveis aleatórias $Y_1, Y_2, \ldots, Y_N$, tendo uma distribuição conjunta N-dimensional ξ (ξ-**distribution**) e, consequentemente, a população é uma realização de um processo aleatório, chamado modelo de superpopulação. O alvo da inferência, neste caso, é o parâmetro $\boldsymbol{\theta}$ do modelo de superpopulação $f(y_1, \ldots, y_N; \boldsymbol{\theta})$.

Na abordagem baseada em modelos, a inferência decorre do modelo de superpopulação e não a partir do desenho amostral. As propriedades dos estimadores são avaliadas através da distribuição de $\mathbf{Y}$ para uma dada amostra, e não através da distribuição de $\mathbf{Y}$ sobre todas as amostras possíveis. Ambas as abordagens podem ser encontradas com detalhes em Brus e Gruijter (1997), Cassel, Särndal e Wretman (1977), Gregoire (1998), Smith (1976) e Särndal *et al.* (1978).

Na terceira abordagem, um modelo superpopulação é usado para descrever a distribuição conjunta (ou condicional) das variáveis da pesquisa, levando em consideração as variáveis auxiliares. Este modelo é então utilizado para desenvolver estimadores que, no entanto, tem suas propriedades avaliadas no que diz respeito à distribuição de aleatorização $p(s)$ gerado pelo desenho da amostra (ξp-**distribution**).

Neste artigo, um dos principais objetivos de Pfeffermann (1993) foi, baseado em uma pesquisa rigorosa da literatura, responder à duas perguntas:

- É possível justificar o uso dos pesos na inferência baseada em modelos? Se sim, sob quais condições?
- É possível desenvolver diretrizes para saber como usar os pesos?

Primeiramente, Pfeffermann (1993) sugere verificar se o uso dos pesos amostrais podem interferir ou não em uma análise, propondo um teste cujo objetivo é verificar se um plano amostral é informativo ou não. Antes de descrever as características de um plano amostral informativo, considere, como feito ao longo deste trabalho, uma população finita de tamanho N na qual tem-se interesse em uma característica Y. Sejam $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ o vetor $(N \times 1)$ de dados completos correspondentes aos valores da característica Y das N unidades da população, $\mathbf{y}_s = (y_1, y_2, \ldots, y_n)'$ o vetor $(n \times 1)$ que representa o conjunto de dados observados ou medidos da característica Y para as n unidades que pertencem à amostra s e $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ o vetor $(N \times 1)$ formado pelas probabilidades dos indivíduos da população serem incluídos na amostra.

Segundo Ravines (2003), após propor um modelo, deve-se analisar se as probabilidades de seleção $\boldsymbol{\pi}$ dos elementos da população estão relacionadas com a variável resposta $\mathbf{y}$, condicionada às covariáveis $\mathbf{x}$ do modelo. Se essa relação existe, então, o desenho amostral é informativo. Se o desenho amostral é informativo, a distribuição dos valores na amostra, $f_s(y_i|\mathbf{x}, \boldsymbol{\theta})$, é diferente da sua distribuição na população, $f_p(y_i|\mathbf{x}, \boldsymbol{\theta})$. Quando $f_s(y_i|\mathbf{x}, \boldsymbol{\theta})$ e $f_p(y_i|\mathbf{x}, \boldsymbol{\theta})$ são iguais, os dados não observados não fornecem informação adicional ao modelo proposto e o desenho amostral é ignorável ou não informativo.

Exemplos de amostragem informativa podem ser encontrados em estudos ecológicos, sociais, da saúde pública e em pesquisas onde as unidades são selecionadas com probabilidades proporcionais aos seus valores. Na amostragem estratificada e na amostragem por conglomerados, por exemplo, as unidades da amostra final são, geralmente, selecionadas com probabilidades desiguais. Se essas probabilidades estão correlacionadas com as variáveis respostas, o desenho

torna-se informativo e o modelo apropriado para se ajustar aos dados amostrais é diferente do modelo para se ajustar aos dados populacionais (PFEFFERMANN; MOURA; SILVA, 2006).

Para testar se um plano amostral é ignorável, Pfeffermann (1993), baseado nos resultados de Hausman e Wise (1981), propõe um teste que, considerando um modelo com k parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, $\hat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ e $\hat{\boldsymbol{\theta}}_{MPV}$ o estimador de máxima pseudo-verossimilhança de $\boldsymbol{\theta}$, em que os pesos são incluídos na estimação, e supondo que $V(\hat{\boldsymbol{\theta}})$ e $V(\hat{\boldsymbol{\theta}}_{MPV})$ são as correspondentes variâncias de $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{MPV}$, respectivamente, tem-se sob a hipótese nula de que o plano é não-informativo, a estatística

$$T = (\hat{\boldsymbol{\theta}}_{MPV} - \hat{\boldsymbol{\theta}})[V(\hat{\boldsymbol{\theta}}_{MPV}) - V(\hat{\boldsymbol{\theta}})]^{-1}(\hat{\boldsymbol{\theta}}_{MPV} - \hat{\boldsymbol{\theta}})'$$

que tem distribuição qui-quadrado com k graus de liberdade.

Exemplos dos efeitos de ignorar planos informativos incluem os trabalhos de citeonlinedemets e Nathan e Holt (1980), que estudaram as propriedades dos estimadores de mínimos quadrados ordinários (MQO) quando a seleção da amostra é baseada nos valores de uma variável amostral Z que é correlacionada com a variável dependente Y, e os trabalhos de Hausman e Wise (1981) e Jewell (1985), que consideraram situações em que as probabilidades de seleção dependiam diretamente da variável dependente Y. Ambos os estudos indicaram que os estimadores de MQO dos coeficientes de regressão, os quais não incorporam os pesos, podem ser severamente viciados. Já Chambless e Boyle (1985) fizeram um estudo empírico sobre os efeitos de ignorar o desenho amostral na estimação de um modelo de regressão logística. Os autores usaram dados da **Lipid Research Clinics Program Prevalence Study** (LRCPPS). A LRCPPS usa uma amostra aleatória estratificada desproporcional nos estratos definidos pela raça e pela zona de lípidios. A variável dependente na análise é novamente definida pelos níveis da zona de lípidios. Os autores descobriram que quando o modelo logístico é estimado ignorando o desenho (isto é, partindo do princípio de amostragem aleatória simples), a estimativa do intercepto tem um vício substancial.

Os estudos citados anteriormente são exemplos de que a inferência analítica deve levar em consideração o fato do plano amostral ser informativo e que os pesos amostrais tem um papel relevante na análise. Mas de que forma o uso dos pesos amostrais protegem contra planos informativos e má especificação do modelo?

Na Seção 2.3.2, foi definido que a solução $\boldsymbol{\theta}_{\mathcal{U}}$ das equações de verossimilhança populacional é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ no caso de um censo. Pfeffermann (1993) define $\boldsymbol{\theta}_{\mathcal{U}}$ como uma Quantidade Descritiva Populacional Correspondente (QPDC) a $\boldsymbol{\theta}$, sobre a qual se deseja fazer inferências com base em informações da amostra. Na abordagem de modelos de superpopulação para fazer inferências analíticas, postula-se um modelo como na abordagem clássica e a inferência é direcionada aos parâmetros do modelo. Porém, em vez de achar um estimador ótimo sob o modelo (estimador de mínimos quadrados ordinários), acha-se um estimador na classe dos estimadores consistentes para a QDPC, em que a consistência é referida à distribuição de aleatorização do estimador. A definição de QDPC dada por Pfeffermann

(1993) é semelhante às definições dadas por Binder (1983), Godambe e Thompson (1986) e Skinner, Holt e Smith (1989).

Mas porque usar a QDPC? Segundo Pfeffermann (1993), a resposta é robustez. Pessoa e Silva (1998) explicam o porquê essa abordagem oferece maior robustez considerando dois casos.

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

Então, quando $N \to \infty$, a QDPC $\boldsymbol{\theta}_{\mathcal{U}}$ converge para o parâmetro $\boldsymbol{\theta}$, isto é, $\boldsymbol{\theta}_{\mathcal{U}} - \boldsymbol{\theta} \to \mathbf{0}$ em probabilidade, segundo a distribuição de probabilidades do modelo ξ. Se $\boldsymbol{\theta}_{MPV}$ for consistente, então quando $n \to \infty$, $\boldsymbol{\theta}_{MPV} - \boldsymbol{\theta}_{\mathcal{U}} \to \mathbf{0}$ em probabilidade, segundo a distribuição de aleatorização p. Juntando essas condições, $\boldsymbol{\theta}_{MPV} \to \boldsymbol{\theta}$ em probabilidade segundo a mistura ξp.

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

Neste caso, o parâmetro $\boldsymbol{\theta}$ não tem interpretação significante, porém a QDPC $\boldsymbol{\theta}_{\mathcal{U}}$ é definida na população finita (real) com interpretação clara, independente da validade do modelo. Como $\boldsymbol{\theta}_{MPV}$ é consistente para a QDPC $\boldsymbol{\theta}_{\mathcal{U}}$, a inferência baseada no procedimento MPV segue válida para este parâmetro, independente da inadequação do modelo para a população.

Assim, a principal conclusão desse estudo é que os pesos amostrais podem ter uma função crucial em dois aspectos diferentes no processo de modelagem. São eles:

- Proteger contra planos amostrais informativos e não-ignoráveis, que poderiam introduzir ou causar vícios;
- Proteger contra a má especificação do modelo;

Outras referências nessa linhagem de estudo são Asparouhov e Muthen (2007), Brewer (1999), Chambers (2003), Little (2011), Pfeffermann (2011), Pfeffermann, Krieger e Rinott (1998b) e Sugden e Smith (1984).

3.2 Exemplo 2

Neste artigo, a abordagem utilizada por Holt, Smith e Winter (1980) é a baseada em modelos, e discute a principal forma que um analista, desconhecendo a estrutura da população (ou até ciente disto, mas desconhecendo a forma de análise correta), usaria para ajustar um modelo de regressão linear normal, que é o método dos mínimos quadrados ordinários. Alternativamente, propõe um estimador de máxima verossimilhança que leva em consideração os efeitos da seleção através dos valores de uma variável auxiliar Z, conhecida para todos os elementos da população, e sua relação com as variáveis do modelo.

O caso de usar o método dos mínimos quadrados ordinários e ignorar a estrutura da população baseia-se nas seguintes pressuposições. Assume-se que uma população finita de N unidades é uma amostra aleatória simples de tamanho N de uma população infinita. Os dados populacionais podem ser ajustados por uma regressão da forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

em que $\mathbf{Y}$ é um vetor $N \times 1$, $\mathbf{X}$ é uma matriz $N \times p$, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros a ser estimados e o vetor de erros $\boldsymbol{\epsilon}$ tem a propriedade $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$. Uma amostra de tamanho n é retirada da população segundo um plano amostral $p(s)$. O planejamento amostral não está relacionado à variável $\mathbf{Y}$ e assume-se que esse fato pode ser ignorado na inferência do modelo (3.1). Se $V(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$, o método dos mínimos quadrados é apropriado para a análise. Porém, se $V(\boldsymbol{\epsilon}|\mathbf{X}) \neq \sigma^2\mathbf{I}$, o método dos mínimos quadrados generalizados deve ser empregado.

Contudo, Holt, Smith e Winter (1980) explicam que, uma vez que o plano amostral foi escolhido e a amostra desenhada, as probabilidades de seleção não constam na inferência baseada em modelos, mas a seleção da amostra desempenha um papel fundamental. A maneira mais óbvia na qual a seleção da amostra afeta a inferência é em termos de eficiência.

Em pesquisas amostrais, as unidades populacionais serão, geralmente, conglomerados em que os elementos dentro desses conglomerados são correlacionados. O efeito desse tipo de correlação foi estudado por Kish e Frankel (1974) e a conclusão foi que para uma correlação intra conglomerados positiva, as variâncias amostrais são infladas.

Uma das pressuposições do modelo (3.1) para que o método dos mínimos quadrados seja aplicado é que $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$. Neste caso, o plano amostral é não-informativo. Entretanto, se a seleção da amostra for baseada na variável dependente Y ou na variável independente X, este método não será satisfatório. Por exemplo, considerando uma regressão linear simples de Y em X e supondo que a amostra é selecionada com base em valores conhecidos da variável dependente Y escolhendo todas as unidades de forma que $Y > Y^*$, para dado Y^*, então, uma unidade em que $X = X^*$ estará na amostra somente se $Y > Y^*$, isto é, $\beta X^* + \epsilon > Y^*$ ou $\epsilon > Y^* - \beta X^*$. Assim, $E(\epsilon|X^*) \neq 0$ em geral por causa do truncamento envolvido na seleção.

Uma situação intermediária entre uma seleção baseada em Y e uma baseada em X ocorre quando a seleção é baseada em uma variável auxiliar Z correlacionada com Y e X mas que não aparece na equação de regressão. Novamente, este fato leva a $E(\boldsymbol{\epsilon}|\mathbf{X}) \neq 0$. Existem casos em que a variável Z deve ser incluída no modelo como uma variável regressora, levando ao modelo

$$Y = \beta_{YX \cdot Z} X + \beta_{YZ \cdot X} Z + \epsilon,$$

mas para esses casos os métodos padrões se aplicam. O interesse de Holt, Smith e Winter (1980) é somente na regressão de Y em X, sendo Z apenas uma variável conhecida na população disponível para o planejamento, mas que não exerce nenhum papel na estrutura causal do modelo.

Por exemplo, ao estudar a relação entre nível de escolaridade e renda, os impostos pagos no ano anterior podem estar disponíveis mas não seriam usados na regressão

$$\mathbf{Y} = \boldsymbol{\beta}_{YX}\mathbf{X} + \boldsymbol{\epsilon},$$

que é o modelo de interesse.

Os valores da variável Z podem ser escritos como uma matriz $\mathbf{Z}$, $N \times p$, a qual resume a informação disponível antes da coleta e é usada para determinar o planejamento amostral $p(s)$ que, neste caso, pode ser escrito como $p(s|\mathbf{Z})$. O plano amostral levará informações sobre as variáveis da pesquisa através da sua relação com as variáveis auxiliares.

Seguindo Pearson (1903), Holt, Smith e Winter (1980) assumem que estão lidando com uma distribuição normal multivariada, em que a população finita é uma amostra aleatória dessa distribuição. As variáveis podem ser divididas dentro de três grupos: as variáveis respostas $\mathbf{Y}$, as variáveis regressoras $\mathbf{X}$ e as variáveis auxiliares $\mathbf{Z}$. O vetor de médias e a matriz de covariâncias da superpopulação normal multivariada podem ser particionadas como

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_X \\ \mu_Z \end{bmatrix} \quad e \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} \\ \Sigma_{XY} & \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZY} & \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}.$$

O parâmetro a ser estimado é a matriz de coeficientes $\boldsymbol{\beta}_{YX}$ da regressão multivariada de Y em X dada por $\boldsymbol{\beta}_{YX} = \Sigma_{YX}\Sigma_{XX}^{-1}$. Dada uma amostra de n unidades selecionada usando $p(s|\mathbf{Z})$, Holt, Smith e Winter (1980) derivaram um estimador de máxima-verossimilhança para $\boldsymbol{\beta}_{YX}$. Se $n\mathbf{S}_{jk}$ é a soma de quadrados não-ponderada correspondente a $\boldsymbol{\Sigma}_{jk}$ e $N\hat{\boldsymbol{\Sigma}}_{XX}$ é a soma de quadrados baseada em todas as N observações, então o estimador de máxima verossimilhança de $\boldsymbol{\beta}_{YX}$ é

$$\hat{\boldsymbol{\beta}}_{YX} = \{\mathbf{S}_{YX} + \mathbf{S}_{YZ}\mathbf{S}_{ZZ}^{-1}(\hat{\Sigma}_{ZZ}\mathbf{S}_{ZZ}^{-1} - \mathbf{I})\mathbf{S}_{ZX}\}\{\mathbf{S}_{XX} + \mathbf{S}_{XZ}\mathbf{S}_{ZZ}^{-1}(\hat{\Sigma}_{ZZ}\mathbf{S}_{ZZ}^{-1} - \mathbf{I})\mathbf{S}_{ZX}\}^{-1}.$$

Holt, Smith e Winter (1980) chamam atenção para o fato de que qualquer ajuste para a seleção da amostra ocorre através dos valores da variável auxiliar Z na população finita e na amostra, e não através do planejamento $p(s|\mathbf{Z})$. As inferências são analíticas e baseadas na distribuição especificada pelo modelo (**model-based approach**). Para fins de inferência, o plano amostral traz informações apenas através dos valores contidos na matriz $\mathbf{Z}$, que é conhecida pelo analista. Consequentemente, $p(s|\mathbf{Z})$ não traz nenhuma informação extra e é neste sentido que o planejamento amostral é não-informativo para a inferência baseada em modelos. Não é o planejamento amostral, mas sim a maneira que a variável auxiliar é usada que importa. Embora o planejamento amostral seja não-informativo, os valores de $\mathbf{Z}$ nas unidades selecionadas pelo plano amostral afetam a inferência através da sua relação com as variáveis do modelo.

Para investigar as propriedades dos dois precedimentos de análise, foi feito um estudo de simulação envolvendo seis planos amostrais diferentes. A principal conclusão de Holt, Smith

e Winter (1980) foi que a seleção da amostra exerce um papel fundamental na inferência em modelos de regressão. O estimador de mínimos quadrados ordinários é viesado quando o desenho amostral envolve probabilidade de seleção desigual. O viés pode ser grave sendo mais do que 10% do valor estimado e muito maior que o erro padrão do estimador. Em planos com probabilidade igual de seleção dos elementos, entretanto, o estimador de mínimos quadrados ordinários se comportou bem. O estimador de máxima-verossimilhança teve um comportamento satisfatório nas duas situações. Um estudo com conclusões semelhantes também pode ser encontrado em Nathan e Holt (1980).

3.3 Exemplo 3

Como tem sido visto ao longo deste trabalho, considerar observações de pesquisas complexas como sendo independentes é, quase que sempre, irrealista. Neste artigo, Holt, Scott e Ewings (1980) abordam o efeito de correlações entre os elementos da amostra sobre o desempenho dos testes qui-quadrado padrões utilizados para testar bondade de ajuste, homogeneidade e independência.

No caso do teste de bondade de ajuste, na Seção 2.4.1 foram apresentadas algumas possíveis correções para a estatística qui-quadrado. Porém, em Holt, Scott e Ewings (1980) encontra-se uma outra possibilidade de correção, mas que também aparecem nos trabalhos de Fellegi (1980) e Lehtonen e Pahkinen (2004). Trata-se de um ajuste simples de EPA médio à estatística qui-quadrado de Pearson (2.19) para bondade de ajuste. Esta correção, assim como a correção de primeira ordem de Rao-Scott, é bastante útil quando não é possível obter ou não está disponível uma estimativa completa de $\hat{\boldsymbol{V}}_p$, pois não é necessária nenhuma informação dessa matriz além da diagonal.

A correção de EPA médio baseia-se nos efeitos univariados de plano amostral estimados $\hat{d}_k$ das estimativas $\hat{p}_k$ e é obtida dividindo o valor observado da estatística (2.19) pela média dos efeitos univariados de plano amostral:

$$X_P^2(\hat{d}_*) = \frac{X_P^2}{\hat{d}_*} \tag{3.2}$$

em que $\hat{d}_* = \dfrac{\sum_{k=1}^{K} \hat{d}_k}{K}$ é um estimador da média $\bar{d}$ dos efeitos de plano amostral desconhecidos, e $\hat{d}_k = EPA_{kish}(\hat{p}_k)$.

Para investigar as performances das estatísticas na prática, realizou-se um estudo empírico com dados de duas pesquisas de grande porte, a primeira é a **General Household Survey** (GHS) e a segunda é a **British Election Survey** (BES), com o objetivo de investigar as performances de X_P^2 e $X_P^2(\hat{d}_*)$, e avaliar o tamanho real dos testes, ou seja, verificar qual nível de significância esses testes alcançam (em oposição a um nível de significância nominal de 5%).

As Tabelas (3.1) e (3.2) mostram os resultados para as variáveis da GHS e da BES, respectivamente. Estes resultados mostram claramente que o teste qui-quadrado baseado na suposição de que as observações são independentes pode gerar conclusões tão enganosas que parece haver utilidade nenhuma em realizá-lo sem fazer as devidas modificações. Mostram também que o tamanho real de cobertura é muito maior para variáveis como G1 que tem um EPA alto. A modificação é bastante satisfatória em todos os casos e parece substituir adequadamente o teste de Wald se as variâncias das celas estiverem disponíveis.

Tabela 3.1 – Níveis de significância estimados ($\%$) em oposição a um nível de significância nominal de 5% no teste de bondade de ajuste para os dados da GHS

Variável	g.l.	$\bar{d}$	X_P^2	$X_P^2(\hat{d}_*)$
G1	3	3,27	48	6
G2	2	3,42	41	5
G3	3	2,54	37	6
G4	3	2,17	30	6
G5	9	1,23	16	7
G6	3	1,48	16	6
G7	6	1,29	15	6
G8	9	1,19	14	6
G9	5	1,14	10	7
G10	2	1,26	10	5
G11	4	1,01	6	6
G12	1	1,13	6	5
G13	3	1,02	6	5

Fonte: Holt, Scott e Ewings (1980)

Com relação ao teste de homogeneidade, Holt, Scott e Ewings (1980) avaliaram as performances das estatísticas $X_P^2(H)$ e $X_P^2(H, \hat{\lambda}_*)$, apresentadas em (2.36) e (2.37), respectivamente, utilizando uma amostra de 1000 famílias da GHS em cinco desenhos amostrais diferentes. Os desenhos considerados foram:

- A: Amostragem em dois estágios com probabilidade proporcional ao tamanho com 50 unidades primárias de amostragem e 20 observações por unidade primária.
- B: Amostragem em dois estágios com probabilidade proporcional ao tamanho com 100 unidades primárias de amostragem e 10 observações por unidade primária.
- C: Amostragem em dois estágios com probabilidade proporcional ao tamanho com 200 unidades primárias de amostragem e 5 observações por unidade primária.
- D: Amostragem aleatória simples com 1000 observações.
- E: Amostragem estratificada com alocação proporcional usando os 141 estratos originais da GHS.

Tabela 3.2 – Níveis de significância estimados ($\%$) em oposição a um nível de significância nominal de 5% no teste de bondade de ajuste para os dados da BES

Variável	g.l.	$\bar{d}$	X_P^2	$X_P^2(\hat{d}_*)$
B1	3	3,38	50	6
B2	3	2,97	44	6
B3	3	1,93	26	6
B4	12	1,26	19	6
B5	4	1,43	17	6
B6	4	1,43	17	5
B7	3	1,55	17	5
B8	2	1,57	16	5
B9	2	1,34	12	5
B10	2	1,32	11	5
B11	3	1,19	10	5
B12	2	1,25	10	5
B13	10	1,13	10	6
B14	5	1,13	8	5
B15	1	1,29	8	5
B16	1	1,06	6	5

Fonte: Holt, Scott e Ewings (1980)

Os planos amostrais D e E foram inclusos apenas com o propósito de comparação, pois raramente são utilizados. Os resultados para a estatística modificada foram bastante satisfatórios com os níveis de significância variando sempre entre 5% e 6%. Os resultados para a estatística qui-quadrado padrão de algumas variáveis da GHS encontram-se na Tabela (3.3) para todos os pares possíveis de desenhos.

Tabela 3.3 – Níveis de significância estimados ($\%$) em oposição a um nível de significância nominal de 5% no teste de homogeneidade para os dados da GHS

	Desenhos														
Variável	AA	AB	AC	AD	AE	BB	BC	BD	BE	CC	CD	CE	DD	DE	EE
G4	51	43	38	34	32	32	26	19	18	17	11	10	5	5	4
G2	48	40	35	30	29	29	23	18	17	15	11	10	5	4	4
G7	38	32	27	23	22	22	17	13	13	12	8	8	5	5	4
G3	49	40	36	30	29	29	23	18	16	15	11	10	5	4	4
G11	30	23	21	18	17	17	14	10	10	10	7	7	5	5	5
G8	39	33	28	23	22	22	17	13	12	12	9	9	5	5	5
G6	30	24	22	18	18	17	14	12	11	11	8	8	5	5	4

Fonte: Holt, Scott e Ewings (1980)

Levando em consideração o plano amostral B, a Tabela (3.4) mostra o que acontece com o nível de significância do teste qui-quadrado padrão quando o mesmo desenho amostral é utilizado para l regiões (ou populações) diferentes.

Tabela 3.4 – Níveis de significância estimados (%) em oposição a um nível de significância nominal de 5% no teste de homogeneidade para o desenho amostral B com várias populações

			Número de populações				
	g.l.	$\hat{\lambda}_*$	2	3	4	5	10
G2	2	2,46	29	42	56	59	85
G3	3	2,13	29	45	52	61	86
G11	4	1,45	17	22	27	31	79

Fonte: Holt, Scott e Ewings (1980)

A principal característica dos resultados é a severa distorção que pode ocorrer com qualquer um dos planos complexos A, B e C. Assim, toda aplicação do teste de homogeneidade baseado na suposição de que as observações são independentes devem ser vistos com grande desconfiança. Os resultados tornam-se ainda piores quando o teste é feito em mais de duas populações. Neste caso, o nível sobe progressivamente à medida que o número de populações aumenta.

Tabela 3.5 – Níveis de significância estimados (%) em oposição a um nível de significância nominal de 5% no teste de independência para os dados da GHS

Variável	g.l.	$\hat{\lambda}_*$	$X_P^2(I)$	$X_P^2(I, \hat{\lambda}_*)$
G1 $\times$ G2	3	1,96	26	6
G1 $\times$ G3	3	2,04	28	5
G1 $\times$ G4	6	1,96	62	6
G1 $\times$ G8	6	1,57	23	6
G1 $\times$ G9	15	0,93	5	7
G1 $\times$ G10	6	0,92	4	5
G1 $\times$ G12	3	1,06	6	5
G1 $\times$ G13	9	0,87	3	5
G2 $\times$ G3	1	1,99	16	5
G2 $\times$ G4	2	1,97	22	5
G2 $\times$ G8	2	1,24	9	5
G2 $\times$ G9	5	0,91	4	5
G2 $\times$ G10	2	0,97	5	5
G2 $\times$ G12	1	0,99	5	5
G2 $\times$ G13	3	0,93	4	5

Fonte: Holt, Scott e Ewings (1980)

Com relação ao teste de independência, comparou-se, além do teste qui-quadrado usual, algumas estatísticas de teste modificadas, entre elas $X_P^2(I, \hat{\lambda}_*)$ definida em (2.49). Holt, Scott e Ewings (1980) mostram que os resultados obtidos são bem menos vulneráveis que os testes correspondentes para homogeneidade e bondade de ajuste, mas que, de qualquer forma, não deixam de merecer atenção. Uma parte dos resultados obtidos é mostrado na Tabela (3.5).

4 CONSIDERAÇÕES FINAIS

Neste trabalho foram estudados vários trabalhos que tratam da incorporação do esquema amostral em análises analíticas de dados provenientes de pesquisas de grande porte, em particular em análises de regressão e alguns testes de hipóteses para tabelas de contingência.

Foram apresentados dois conceitos que avaliam o impacto de ignorar o plano amostral na análise dos dados. O primeiro foi o Efeito do Plano Amostral, que mede o efeito de plano amostral sobre a variância de um estimador, e o segundo foi o Efeito do Plano Amostral Ampliado, que mede o efeito de ignorar o plano amostral no estimador da variância usado para medir a precisão do estimador após a seleção da amostra. Analisando o trabalho de Cunha, Araújo e Lima (2015), por exemplo, conclui-se que estes dois conceitos auxiliam bastante o pesquisador antes e/ou depois da coleta dos dados, mostrando o quanto a estimativa da variância de um estimador baseada na suposição de que a amostra é IID subestima ou superestima a verdadeira estimativa de variância.

No que diz respeito à classe de modelos de regressão, foi apresentado um método chamado Método de Máxima Pseudo-Verossimilhança, no qual os pesos amostrais são incorporados na análise através das equações de estimação dos parâmetros e da matriz de covariância dos estimadores. Neste método, o desenho amostral também é incorporado através da expressão para a variância do total dos escores sob o plano amostral, em que as propriedades do plano amostral estão contidas em π_i e π_{ij}.

Quanto às informações disponíveis para a análise, foi visto que elas exercem papel crucial na estimação. O método de análise escolhido pelo pesquisador analista depende das informações que ele tiver acesso. Analisando o estudo de Holt, Smith e Winter (1980), conclui-se que se o pesquisador não tiver acesso a nenhuma informação sobre o plano amostral utilizado na pesquisa e objetivar o ajuste de um modelo de regressão linear normal aos dados, ele pode até fazer o ajuste através do Método dos Mínimos Quadrados, porém deve estar ciente de que este método poderá levar a estimativas viciadas tanto para os coeficientes dos parâmetros quanto para seus respectivos estimadores de variância, produzindo conclusões errôneas. Se o pesquisador não possuir informações sobre o desenho amostral mas tiver informações sobre os pesos dos elementos contidos na amostra, ele pode optar por usar o Método dos Mínimos Quadrados Ponderados, o qual pode produzir boas estimativas dos coeficientes, porém o mesmo não ocorrerá com as estimativas de variância desses coeficientes.

No que diz respeito às análises em tabelas de contingência, foram apresentados alguns testes de hipóteses para bondade de ajuste, homogeneidade e independência encontrados na literatura que incorporam o esquema amostral nas estatísticas de teste.

Foi visto que o teste qui-quadrado clássico de cada uma das três hipóteses, ou seja, o

teste baseado na suposição de que as observações são independentes, pode ser muito enganoso. Analisando o estudo de Holt, Scott e Ewings (1980), conclui-se que o efeito de correlação entre as observações é bastante severo para os testes de bondade de ajuste e homogeneidade, e menos severo para o teste de independência.

Neste trabalho, foi feita uma revisão bibliográfica como forma de colaborar com os estudos nessa área de pesquisa. Espera-se que trabalhos futuros possam ser feitos aplicando uma ou algumas das técnicas apresentadas neste trabalho em dados agronômicos, o que é perfeitamente possível. Tais trabalhos também podem incluir: como fazer análise de resíduos para modelos de regressão, análise de modelos log-lineares em tabelas de contingência e inferência sobre modelos hierárquicos, todos com dados provenientes de amostras complexas.

REFERÊNCIAS

ARAÚJO, A. A.; LIMA, J. E.; LIMA, J. R. F.; GOMES, M. F. M.. Trabalho infantil no Brasil: análise dos principais determinantes. **Ensaios FEE**, Porto Alegre, v. 31, n. 2, p. 373–394, 2011.

ASPAROUHOV, T.; MUTHEN, B. Testing for informative weights and weights trimming in multivariate modeling with survey data. **Proceedings of the 2007 JSM meeting, Section on Survey Research Methods**, Salt Lake City, v. 3, p. 3394–3399, 2007.

BARATA, R. B.; MORAES, J. C.; ANTONIO, P. R. A.; DOMINGUEZ, M. Inquérito de cobertura vacinal: avaliação empírica da técnica de amostragem por conglomerados proposta pela Organização Mundial da Saúde. **Revista Panamericana de Salud Publica**, Washington, v. 17, n. 3, p. 184–190, 2005.

BELLHOUSE, D. R. A brief history of random sampling methods. **Handbook of statistics**, Amsterdam, v. 6, p. 1–14, 1988.

BENEDETTI, T. R. B.; BINOTTO, M. A.; PETROSKI, E. L.; GONÇALVES, L. H. T. Atividade física e prevalência de quedas em idosos residentes no sul do Brasil. **Revista Brasileira Geriatria e Gerontologia**, Rio de Janeiro, v. 11, n. 2, p. 145–154, 2008.

BHAPKAR, V. P. A note on the equivalence of two test criteria for hypotheses in categorical data. **Journal of the American Statistical Association**, New York, v. 61, n. 313, p. 228–235, 1966.

BINDER, D. A. On the variances of asymptotically normal estimators from complex surveys. **International Statistical Review**, Amsterdam, v. 51, n. 3, p. 279–292, 1983.

BOLFARINE, H.; BUSSAB, W. O. **Elementos de amostragem**. São Paulo Edgard Blücher, São Paulo, 2005. v. 1, 274 p.

BRECKLING, J. U.; CHAMBERS, R. L.; DORFMAN, A. H.; TAM, S. M.; WELSH, A. H. Maximum likelihood inference from sample survey data. **International Statistical Review**, Amsterdam, v. 62, n. 3, p. 349–363, 1994.

BREWER, K. R. W. Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. **Australian Journal of Statistics**, Camberra, v. 5, n. 3, p. 93–105, 1963.

BREWER, K. R. W. Design-based or prediction-based inference? stratified random vs stratified balanced sampling. **International Statistical Review**, Amsterdam, v. 67, n. 1, p. 35–47, 1999.

BRUS, D. J.; GRUIJTER, J.J. De. Random sampling or geostatistical modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). **Geoderma**, Amsterdam, v. 80, n. 1, p. 1–44, 1997.

CASELLA, G.; BERGER, R. L. **Statistical inference**. Pacific Grove: Duxbury, 2002. v. 2, 660 p.

CASSEL, C. M.; SÄRNDAL, C. E.; WRETMAN, J. H. **Foundations of inference in survey sampling**. Nova York: Wiley & Sons, 1977. 192 p.

CHAMBERS, R. L. Which sample survey strategy? a review of three different approaches. **Centre for Statistical and Survey Methodology**, Wollongong, v. 30, p. 1–30, 2003.

CHAMBERS, R. L.; SKINNER, C. J. **Analysis of survey data**. New York: John Wiley & Sons, 2003. 376 p.

CHAMBLESS, L. E.; BOYLE, K. E. Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. **Communications in Statistics-Theory and Methods**, Washington, v. 14, n. 6, p. 1377–1392, 1985.

COCHRAN, W. G. **Sampling techniques**. New York: Wiley & Sons, 1977. 428 p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados e extensões**. Recife: Universidade Federal Rural de Pernambuco, 2010. 249 p.

CORDEIRO, R. Efeito do desenho em amostragem de conglomerado para estimar a distribuição de ocupações entre trabalhadores. **Revista de Saúde Pública**, Universidade de São Paulo - Faculdade de Saúde Pública, São Paulo, p. 10–15, 2001.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. London: Chapman and Hall, 1979. 511 p.

CUNHA, D. A.; ARAÚJO, A. A.; LIMA, J. E. Determinantes do desemprego e inatividade de jovens no Brasil metropolitano. **Revista de Economia e Agronegócio**, Viçosa v. 9, n. 3, p. 369–391, 2015.

DEMETS, D.; HALPERIN, M. Estimation of a simple regression coefficient in samples arising from a sub-sampling procedure. **Biometrics**, Washington, v. 33, n. 1, p. 47–56, 1977.

DEMING, W. E. **Sample Designs in Business Research**. New York: Wiley & Sons, 1960. 544 p.

DUARTE, R. P. N. **Ajuste de modelos lineares usando estimadores de regressao para amostras complexas**. 1999. 78 p. Dissertação (Mestrado) – Universidade de São Paulo, São Paulo.

ELTINGE, J. L. Assessment of information capacity and sensitivity in the analysis of complex surveys. **Bulletin of the International Statistical Institute**, Washington, v. 3, p. 4–8, 1999.

FAO. **Sampling Methods for Agricultural Surveys**. Rome: Food and Agriculture Organization of the United Nations, 1989. 261 p.

FELLEGI, I. P. Approximate tests of independence and goodness of fit based on stratified multistage samples. **Journal of the American Statistical Association**, New York, v. 75, n. 370, p. 261–268, 1980.

FIENBERG, S. E.; TANUR, J. M. Large-scale social surveys: Perspectives, problems, and prospects. **Systems Research and Behavioral Science**, Washington, v. 28, n. 2, p. 135–153, 1983.

FIGUEIREDO, C. C. **Análise de regressão incorporando o esquema amostral**. 2004. 91 p. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo.

FILHO, C. C. B.; ASSUNÇÃO, R. M.; SILVA, B. F. A.; MARINHO, F. C.; REIS, I. A.; ALMEIDA, M. C. . Conglomerados de homicídios e o tráfico de drogas em Belo Horizonte, Minas Gerais, Brasil, de 1995 a 1999. **Caderno de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1163–1171, 2001.

FULLER, W. A. Regression analysis for sample survey. **Sankhya: The Indian Journal of Statistics**, Nova deli, v. 37, p. 117–132, 1975.

GARCIA, E. A. C. **Estudo técnico-econômico da pecuária bovina de corte do pantanal mato-grossense**. Corumbá: Embrapa-CPAP, Brasília, 1986. 150 p.

GARTHWAITE, P. H.; JOLLIFFE, I. T.; JONES, B. **Statistical inference**. London: Oxford University Press, 2002. 352 p.

GODAMBE, V. P. A unified theory of sampling from finite populations. **Journal of the Royal Statistical Society. Series B (Methodological)**, Malden, v. 17, n. 2, p. 269–278, 1955.

GODAMBE, V. P.; THOMPSON, M. E. Parameters of superpopulation and survey population: their relationships and estimation. **International Statistical Review**, Amsterdan, v. 54, n. 2, p. 127–138, 1986.

GREGOIRE, T. G. Design-based and model-based inference in survey sampling: appreciating the difference. **Canadian Journal of Forest Research**, Ottawa, v. 28, n. 10, p. 1429–1447, 1998.

GUIDORIZZI, H. L. **Um curso de cálculo**. São Paulo: Editora LTC, 1995. v. 1, 635 p.

HANSEN, M. H.; HURWITZ, W. N.; MADOW, W. G. **Sample survey methods and theory**. New York: John Wiley & Sons, 1953. 638 p.

HAUSMAN, J. A.; WISE, D. A. Stratification on endogenous variables and estimation: The gary income maintenance experiment. **Structural analysis of discrete data with econometric applications**, Cambridge: MIT Press, p. 365–391, 1981.

HELIÖVAARA, M.; AROMAA, A.; KLAUKKA, T.; KNEKT, P.; JOUKAMAA, M.; IMPIVAARA, O. Reliability and validity of interview data on chronic diseases the mini-finland health survey. **Journal of clinical epidemiology**, Elsevier, v. 46, n. 2, p. 181–191, 1993.

HENDRICKS, W. A. **The mathematical theory of sampling**. New Brunswick, N.J.: Scarecrow Press, 1956. 364 p.

HOEL, P. G. **Introduction to mathematical statistics**. New York: John Wiley & Sons, 1954. 2nd ed., 331 p.

HOLT, D.; SCOTT, A. J.; EWINGS, P. D. Chi-squared tests with survey data. **Journal of the Royal Statistical Society. Series A (General)**, Malden, v. 143, n. 3, p. 303–320, 1980.

HOLT, D.; SMITH, T. M. F.; WINTER, P. D. Regression analysis of data from complex surveys. **Journal of the Royal Statistical Society. Series A (General)**, Malden, v. 143, n. 4, p. 474–487, 1980.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. **Journal of the American Statistical Association**, New York, v. 47, n. 260, p. 663–685, 1952.

ISAKI, C. T. **Survey designs utilizing prior information**. Digital Repository@ Iowa State University, 1970. Disponivel em: <http://lib.dr.iastate.edu/>. Acesso em: 26 jul.2016.

JEWELL, N. P. Least squares regression with data arising from stratified samples of the dependent variable. **Biometrika**, London, v. 72, n. 1, p. 11–21, 1985.

KIANG-KHANG, P. **The effect of two-stage sampling on non-parametric tests**. 1984. 210 p. Tese (Doutorado) – University of Auckland, Auckland.

KISH, L. **Survey sampling**. New York: John Wiley & Sons, 1965. 643 p.

KISH, L.; FRANKEL, M. R. Inference from complex samples. **Journal of the Royal Statistical Society. Series B (Methodological)**, Malden, v. 36, n. 1, p. 1–37, 1974.

KORN, E. L.; GRAUBARD, B. I. **Analysis of health surveys**. New York: John Wiley & Sons, 2011. 382 p.

KRIEGER, A. M.; PFEFFERMANN, D. Maximum likelihood estimation from complex sample surveys. **Survey Methodology**, Waterloo, v. 18, n. 2, p. 225–239, 1992.

KRIEGER, A. M.; PFEFFERMANN, D. Testing of distribution functions from complex sample surveys. **Journal of Official Statistics**, Washington, v. 13, n. 2, p. 123, 1997.

LARNTZ, K. Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. **Journal of the American Statistical Association**, New York, v. 73, n. 362, p. 253–263, 1978.

LEHTONEN, R. The execution of the national occupational health care survey. **Helsinki: Publications of the Social Insurance Institution**, Helsinki, v. 64, p. 3–130, 1988.

LEHTONEN, R.; KUUSELA, V.; AROMAA, A.; HELIOVAARA, M.; IMPIVAARA, O.; KNEKT, P.; MAATELA, J. Statistical efficiency of the mini-finland health surveys sampling design. Part 5. **Helsinki and Turku: Publications of the Social Insurance Institution**, Helsink, v. 65, p. 3–94, 1986.

LEHTONEN, R.; PAHKINEN, E. **Practical methods for design and analysis of complex surveys**. New York: John Wiley & Sons, 2004. 344 p.

LEITE, P. G. P. G.; SILVA, D. B. N. Análise da situação ocupacional de crianças e adolescentes nas regiões sudeste e nordeste do brasil utilizando informações da pnad 1999. **Revista Brasileira de Estudos de População**, Rio de Janeiro, v. 19, n. 2, p. 47–63, 2013.

LILA, M. F. **Estimação de Variâncias em Pesquisas Amostrais Domiciliares**. 2002. 77 p. Dissertação (Mestrado) — Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

LITTLE, R. J. The bayesian approach to sample survey inference. Em: **Analysis of Survey Data**, New York: Wiley & Sons, p. 49–57, 2003.

LITTLE, R. J. To model or not to model? competing modes of inference for finite population sampling. **Journal of the American Statistical Association**, New York, v. 99, n. 466, p. 546–556, 2011.

MADOW, W. G.; MADOW, L. H. On the theory of systematic sampling, I. **The Annals of Mathematical Statistics**, Durham, v. 15, n. 1, p. 1–24, 1944.

MASSAGO, S. **Matriz Hessiana e Aplicações**. 2010. Disponível em: <www.dm.ufscar.br/~sadao>. Acesso em: 30 jul. 2016.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. London: New York Chapman and Hall, 1989. 511 p.

MCDOWELL, A. Plan and operation of the second national health and nutrition examination survey, 1976-1980. **Vital Health Stat**, Atlanta, n. 15, p. 1–144, 1981.

MOOD, A. M. **Introduction to the theory of statistics**. New York: McGraw-Hill, 1950. 433 p.

MORAES, J. R.; MOREIRA, J. P. L.; LUIZ, R. R. Efeito do plano amostral em modelo logístico ordinal: uma análise do estado de saúde autorreferido de adultos no brasil usando a pesquisa nacional por amostra de domicílios de 2008. **Caderno de Saúde Pública**, Rio de Janeiro, v. 28, p. 913–924, 2012.

MORETTIN, P. A.; BUSSAB, W. O. **Estatıstica básica**. São Paulo: Atual, 2004. 211 p.

MOSER, C. A.; KALTON, G. **Survey methods in social investigation.**, [s.l.], Heinemann Educational, 1971. 2nd ed., 549 p.

MURTHY, M. N. **Sampling theory and methods**, Calcutta: Statistical Publishing Society, Kolkata, 1967. 684 p.

NATHAN, G.; HOLT, D. The effect of survey design on regression analysis. **Journal of the Royal Statistical Society. Series B (Methodological)**, Malden, v. 42, n. 3, p. 377–386, 1980.

NETTO, S. P.; SANQUETTA, C. R. Determinação do número de estratos em estratificação volumétrica de florestas naturais e plantadas. **Revista Floresta**, Curitiba, v. 24, n. 12, p. 49–58, 1994.

NOJOSA, R. T. **Inferência estatística paramétrica**. 2006. Disponível em: <http://www.lce.esalq.usp.br/arquivos>. Acesso em: 30 jul. 2016.

PARK, H. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. **Journal of Korean Academy of Nursing**, Seul, v. 43, n. 2, p. 154–164, 2013.

PAULA, G. A. **Modelos de regressão:** com apoio computacional. São Paulo: IME-USP, 2004. 428 p.

PEARSON, K. Mathematical contributions to the theory of evolution. xi. on the influence of natural selection on the variability and correlation of organs. **Philosophical Transactions of the Royal Society A**, London, v. 200, p. 1–66, 1903.

PESSOA, D. G. C.; SILVA, P. L. N. **Análise de dados amostrais complexos**. São Paulo: Associação Brasileira de Estatística, 1998. 170 p.

PFEFFERMANN, D. The role of sampling weights when modeling survey data. **International Statistical Review**, Amsterdan, v. 61, n. 2, p. 317–337, 1993.

PFEFFERMANN, D. Modelling of complex survey data: Why model? why is it a problem? how can we approach it? **Survey Methodology**, Waterloo, v. 37, n. 2, p. 115–136, 2011.

PFEFFERMANN, D.; KRIEGER, A. M.; RINOTT, Y. Parametric distributions of complex survey data under informative probability sampling. **Statistica Sinica**, Taiwan, v. 8, n. 4, p. 1087–1114, 1998.

PFEFFERMANN, D.; KRIEGER, A. M.; RINOTT, Y. Parametric distributions of complex survey data under informative probability sampling. **Statistica Sinica**, Taiwan, v. 8, n. 4, p. 1087–1114, 1998.

PFEFFERMANN, D; MOURA, F. A. S.; SILVA, P. L. N. Multi-level modelling under informative sampling. **Biometrika**, London, v. 93, n. 4, p. 943–959, 2006.

RAO, J. N. K.; THOMAS, D. R. The analysis of cross-classified categorical data from complex sample surveys. **American Sociological Association**, Washington, v. 18, p. 213–269, 1988.

RAO, J. N. K.; SCOTT, A. J. 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**, New York, v. 76, n. 374, p. 221–230, 1981.

RAO, J. N. K.; SCOTT, A. J. On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. **The Annals of Statistics**, Durham, v. 2, n. 1, p. 46–60, 1984.

RAVINES, R. E. R. **Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos**. 2003. 116 p. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro.

SAHOO, P. **Probability and mathematical statistics**. Department of Mathematics - University of Louisville, Louisville, 2013. 698 p.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. **Model assisted survey sampling**. New York: Springer Science & Business Media, 2003. 694 p.

SÄRNDAL, C. E.; THOMSEN, I.; HOEM, J. M.; LINDLEY, D. V.; NIELSEN, O. B.; DALENIUS, T. Design-based and model-based inference in survey sampling [with discussion and reply]. **Scandinavian Journal of Statistics**, Scandinavian, v. 5, n. 1, p. 27–52, 1978.

SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. **Biometrics bulletin**, Washington, v. 2, n. 6, p. 110–114, 1946.

SCHEAFFER, R.; MENDENHALL, W.; OTT, R.; GEROW, K. **Elementary survey sampling**. Boston: Cengage Learning, 2011. 464 p.

SHAPER, A. G.; POCOCK, S. J. Blood lead and blood pressure. **British Medical Journal Publishing Group**, London, v. 291, n. 6503, p. 1147–1149, 1985.

SHUSTER, J. J.; DOWNING, D. J. Two-way contingency tables for complex sampling schemes. **Biometrika**, London, v. 63, n. 2, p. 271–276, 1976.

SILVA, P. L. N. **Utilizing auxiliary information in sample survey estimation and analysis.** 1996. 245 p. Tese (Doutorado) — University of Southampton, Southampton.

SILVA, P. L. N.; PESSOA, D. G. C.; LILA, M. F. Análise estatística de dados da pnad: incorporando a estrutura do plano amostral. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 7, n. 4, p. 659–670, 2002.

SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. **Analysis of complex surveys**. New York: John Wiley & Sons, 1989. 309 p.

SMITH, T. M. F. The foundations of survey sampling: a review. **Journal of the Royal Statistical Society. Series A (General)**, Malden, v. 139, n. 2, p. 183–204, 1976.

SOUSA, M. H.; SILVA, N. N. Estimativas obtidas de um levantamento complexo. **Revista de Saúde Pública**, Rio de Janeiro, v. 37, n. 5, p. 662–670, 2003.

SUDMAN, S. **Applied sampling**. New York: Academic Press, 1976. 249 p.

SUGDEN, R. A.; SMITH, T. M. F. Ignorable and informative designs in survey sampling inference. **Biometrika**, London, v. 71, n. 3, p. 495–506, 1984.

SUKHATME, P. V. **Sampling theory of surveys with applications**. Iowa State University Press, Nova Deli, 1957, 542 p.

TAVARES, D. M. S.; PEREIRA, G. A.; IWAMOTO, H. H,; MIRANZZI, S. S. C.; RODRIGUES, L. R.; MACHADO, A. R. M. Incapacidade funcional entre idosos residentes em um município do interior de minas gerais. **Texto & Contexto Enferm**, Florianopolis, v. 16, n. 1, p. 32–39, 2007.

THOMAS, D. R.; RAO, J. N. K. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. **Journal of the American Statistical Association**, New York, v. 82, n. 398, p. 630–636, 1987.

THOMPSON, S. K. **Sampling**. New York: Wiley & Sons, 1992. 472 p.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical Society**, Washington, v. 54, n. 3, p. 426–482, 1943.

WOLTER, K. **Introduction to variance estimation**. New York: Springer-Verlag, 1985. 427 p.

YATES, F. **Sampling methods for censuses and surveys**. London: Charles Griffin & Co. Ltd., 1960. 318 p.