Técnicas de Estimação no Âmbito da Pós-estratificação

por

Ana Cristina Marinho da Costa

Dissertação apresentada como requisito parcial para a obtenção do grau de

Mestre em Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão de Informação

da

Universidade Nova de Lisboa

Lisboa, Dezembro de 2000

Agradecimentos

Ao Professor Doutor José Ferreira Machado, orientador deste trabalho, pelo seu apoio permanente, disponibilidade, dedicação e orientação preciosa.

Ao Instituto Nacional de Estatística, em particular ao Departamento de Estatísticas das Empresas, que proporcionou a minha colaboração no Projecto "Mudanças de Estrato no Inquérito às Empresas / Harmonizado" que motivou a elaboração deste trabalho e disponibilizou os dados dos exemplos práticos.

À Dra. Helena Guerra, pelas suas sugestões competentes e sensatas, pela sua paciência e apoio nos bons e maus momentos.

Aos meus amigos e a todos aqueles que, no ISEGI, me têm incentivado e apoiado.

Ao João, pela inspiração e força que me transmitiu durante a realização deste trabalho.

Aos meus pais e irmão, pelas palavras de encorajamento, pela paciência e apoio que me têm dado na vida.

Resumo

Neste trabalho abordam-se os problemas da base de sondagem e a ocorrência de não respostas nos inquéritos por amostragem. Estes erros não amostrais têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram. Analisam-se alguns métodos de estimação, na abordagem "clássica" da Teoria das Sondagens, que visam o tratamento dos problemas em apreço, destacando-se os métodos de pós-estratificação. As propriedades teóricas dos estimadores de pós-estratificação requerem ainda alguma investigação para planos de sondagem complexos, embora estes métodos sejam frequentemente utilizados. É então abordada a metodologia *Bootstrap* para a estimação da variância dos estimadores. Apresentam-se também alguns exemplos de aplicação, dos métodos de pós-estratificação e do algoritmo *Without Replacement Bootstrap* (*BWO*), proposto por Sitter (1992b), aos dados do Inquérito às Empresas/Harmonizado de 1996, conduzido pelo Instituto Nacional de Estatística.

Palavras Chave: Pós-estratificação; problemas da base de sondagem; não resposta; reponderação; métodos de ajustamento; inferência condicional; Bootstrap.

Abstract

This thesis approaches issues related with frame problems and nonresponse in surveys. These nonsampling errors affect the accuracy of the estimates whereas the estimators become biased and less precise. We analyse some estimation methods that deal with those problems, in the design-based perspective, and give an especial focus to the poststratification procedures. For complex sampling designs the theoretical properties of the poststratification estimators need further research, although these methods are often used in practice. We then address the Bootstrap methodology for variance estimation. Some practical examples of the poststratification estimators and the *Without Replacement Bootstrap* (*BWO*) algorithm, proposed by Sitter (1992b), are also presented, using data from the 1996 Annual Business Survey, conducted by Portugal's National Statistics Institute.

Keywords: Poststratification; frame problems; nonresponse; reweighting; adjustment methods; conditional inference; Bootstrap.

ÍNDICE

1 INTRODUÇÃO 1

2 TÓPICOS DE SONDAGENS 4

2.1 Introdução 4

2.1.1 Métodos de sondagem empíricos 5

2.1.2 Métodos de sondagem probabilísticos 5

2.1.3 Erros de amostragem e erros não amostrais 6

2.1.4 Planeamento e implementação de uma sondagem 7

2.2 Considerações gerais, definições e notação 10

2.2.1 Parâmetros de interesse na população 10

2.2.2 Propriedades desejáveis e critérios de comparação dos estimadores 10

2.2.3 Intervalos de confiança 15

2.2.4 Consistência e não enviesamento assimptótico 20

2.2.5 Probabilidades de inclusão 21

2.3 Sondagem aleatória simples 24

2.3.1 Sondagem aleatória simples com reposição (SASCR) 25

2.3.2 Sondagem aleatória simples sem reposição (SASSR) 28

2.3.3 Comparação entre os estimadores SASCR e SASSR 30

2.4 Sondagem aleatória com probabilidades desiguais 32

2.4.1 Estimação de τ 33

2.4.2 Estimação de μ 36

2.4.3 Pesos de inclusão 37

2.5 Sondagem aleatória estratificada 38

2.5.1 Relações e notação 40

2.5.2 Estimação de μ e τ 44

2.5.3 Comparação com a sondagem aleatória simples 48

2.5.4 Eventuais problemas na estimação 54

2.6 Estimação da variância pelo método de linearização de Taylor 56

2.6.1 Estimadores de Horvitz-Thompson para várias variáveis de estudo 56

2.6.2 Método de linearização de Taylor 61

2.7 Estimação da variância por métodos de Bootstrap 67

2.7.1 Introdução ao Bootstrap 67
2.7.2 Sondagem aleatória simples sem reposição 69
2.7.3 Sondagem aleatória estratificada 71

3 ESTIMAÇÃO NA PRESENÇA DE ERROS NÃO AMOSTRAIS 74

3.1 Introdução 74
3.2 Estimação na presença de erros na base de sondagem 76
3.2.1 O problema das mudanças de estrato 78
3.3 Métodos básicos de estimação pelo quociente 80
3.3.1 Estimação de um quociente 80
3.3.2 Estimação pelo quociente, na presença de informação auxiliar 88
3.4 Métodos básicos de estimação em domínios 95
3.4.1 Notação 96
3.4.2 Alguns métodos de estimação em domínios 97
3.5 Estimadores de pós-estratificação 106
3.5.1 Algumas abordagens à pós-estratificação 107
3.5.2 Sondagem aleatória simples sem reposição 111
3.5.3 Sondagem aleatória estratificada 121
3.6 Estimação na presença de não respostas 129
3.6.1 Introdução aos métodos de ajustamento das não respostas 131
3.6.2 Método de ajustamento por ponderação em classes 132
3.6.3 Métodos de ajustamento por pós-estratificação 135
3.6.4 Sondagem aleatória simples sem reposição 138

4 APLICAÇÕES PRÁTICAS 143

4.1 Introdução 143
4.2 Inquérito às Empresas / Harmonizado (IEH) 144
4.2.1 Especificações metodológicas 144
4.2.2 Alguns dados do IEH96 148
4.2.3 Variáveis de estudo 150
4.3 Apresentação dos resultados 151
4.3.1 Metodologia dos exemplos práticos 151
4.3.2 Exemplo I 154
4.3.3 Exemplo II 157

5 CONCLUSÃO 162

6 REFERÊNCIAS 165

ANEXO 1 - ABREVIATURAS E NOTAÇÃO 170

A1.1 Abreviaturas 171
A1.2 Notação 171
A1.2.1 Notação geral 171
A1.2.2 Notação referente à população 173
A1.2.3 Notação referente à amostra 173

ANEXO 2 – DEMONSTRAÇÕES 174

A2.1 Resultados da secção 2.4 175
A2.1.1 Estimação de τ numa sondagem aleatória com probabilidades desiguais 175
A2.2 Resultados da secção 2.5 177
A2.2.1 Sondagem aleatória estratificada 177
A2.3 Resultados da secção 3.4 179
A2.3.1 Estimação em domínios numa sondagem aleatória estratificada 179
A2.3.2 Estimação em domínios numa sondagem aleatória simples sem reposição (SASSR) 185
A2.4 Resultados da secção 3.5 191
A2.4.1 Estimador da variância do estimador de pós-estratificação, proposto por Rao (1985) 191

ANEXO 3 – CLASSIFICAÇÃO PORTUGUESA DAS ACTIVIDADES ECONÓMICAS CAE - REV. 2 197

A3.1 Designações da CAE – Rev. 2, por secções 198
A3.2 Designações da CAE – Rev. 2, por divisões 199

ANEXO 4 – VARIÁVEIS DE ESTRATIFICAÇÃO DO IEH 202

A4.1 Escalões de NUTS II (*ENUT*) 203
A4.2 Escalões de número de pessoas ao serviço (*ENPS*) 203
A4.3 Escalões de forma jurídica (*EFJR*) 203
A4.4 Escalões de volume de vendas (*EVVN*) 204
A4.5 Escalões de Classificação Portuguesa das Actividades Económicas CAE – Rev. 2 204

ANEXO 5 – HISTOGRAMAS DAS RÉPLICAS BOOTSTRAP 206

ANEXO 6 – INSTRUMENTOS DE NOTAÇÃO DO IEH 216

LISTA DE QUADROS

Quadro 2.2.1 – Probabilidade de cobertura P_0 como função de $BR(\hat{\theta})$ 19
Quadro 2.3.1 – Propriedades do estimador de μ, para a SAS..... 30
Quadro 2.3.2 – Propriedades do estimador de τ, para a SAS 31
Quadro 2.4.1 – Propriedades do estimador de Horvitz-Thompson para τ,..... 36
Quadro 2.5.1 – Propriedades do estimador de μ, para a sondagem estratificada 48
Quadro 2.5.2 – Propriedades do estimador de τ, para a sondagem estratificada..... 48
Quadro 2.6.1 – Propriedades do estimador de Horvitz-Thompson para τ..... 57
Quadro 2.6.2 – Propriedades do estimador de Horvitz-Thompson para o vector de totais $t = (\tau_1, \ldots, \tau_g, \ldots, \tau_G)^T$ 61
Quadro 3.3.1 – Propriedades do estimador usual do quociente $R = \tau_y/\tau_x$..... 86
Quadro 3.3.2 – Propriedades do estimador "weighted sample mean" de μ 88
Quadro 3.3.3 – Propriedades do estimador pelo quociente usual de $\tau_y = \tau_x R$, ... 91
Quadro 3.4.1 – Propriedades do estimador usual de τ_d (Horvitz-Thompson) 101
Quadro 3.4.2 – Propriedades dos estimadores de τ_d e μ_d..... 101
Quadro 4.2.1 – Instrumentos de notação do Inquérito às Empresas / Harmonizado..... 146
Quadro 4.2.2 – Código de situação de Instrumento de Notação (CSV)..... 146
Quadro 4.2.3 – Código de situação da empresa perante a actividade (STA) 147
Quadro 4.2.4 – Número de meses de actividade 147
Quadro 4.2.5 – Situação de apuramento (SA)..... 148
Quadro 4.2.6 – Resumo das condições e situação de apuramento 148
Quadro 4.2.7 – Dimensões do universo e da amostra e número de estratos,..... 149
Quadro 4.2.8 – Dimensões do universo e da amostra e número de estratos..... 149
Quadro 4.2.9 – Situação de apuramento (SA) das empresas da amostra..... 150
Quadro 4.3.1 – Escalões de número de pessoas ao serviço (ENPS)..... 152
Quadro 4.3.2 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Nº médio de pessoas ao serviço* (*Q20001*) 156
Quadro 4.3.3 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Vendas* (*Q4160*) 156

Quadro 4.3.4 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Prestações de serviços* (*Q4190*) 156
Quadro 4.3.5 – Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Nº médio de pessoas ao serviço* (*Q20001*) 157
Quadro 4.3.6 – Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Vendas* (*Q4160*) 158
Quadro 4.3.7 – Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Prestações de serviços* (*Q4190*)..... 158

1 INTRODUÇÃO

Pretende-se com este trabalho fazer uma abordagem aos métodos de estimação por pós-estratificação em inquéritos por amostragem e apresentar alguns exemplos de aplicação ao Inquérito às Empresas/Harmonizado, conduzido pelo Instituto Nacional de Estatística.

A escolha do assunto que preside a esta dissertação assenta, essencialmente, na necessidade que os investigadores e outros utilizadores de dados provenientes de inquéritos por amostragem, sentem em desenvolver métodos de estimação que lidem com os problemas da base de amostragem.

Não só a existência de erros na base de amostragem, como também a ocorrência de não respostas nos inquéritos, têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram. Os estimadores de pós-estratificação têm como principal objectivo lidar com o primeiro problema apontado e, sob determinados pressupostos, podem também ser utilizados para reduzir o enviesamento dos estimadores provocado pela não resposta.

A base de amostragem, ou base de sondagem, é uma lista actualizada de todos os elementos da população alvo. Idealmente, a base de sondagem deveria permitir identificar a população alvo na totalidade. Mas, muitas vezes, não é possível garantir uma representação completa, perfeita e actualizada da população alvo, uma vez que a constituição e actualização de uma base de amostragem é um processo complexo e difícil de realizar.

Para lidar com este problema têm sido desenvolvidas diversas técnicas de estimação que utilizam informação auxiliar que se encontre presente na base de sondagem durante a fase de estimação ou informação proveniente de outras fontes. Algumas das técnicas mais utilizadas são os métodos de pós-estratificação. Na abordagem “clássica” da Teoria das Sondagens sobre os métodos de pós-estratificação, são de referir os trabalhos de Williams (1962), Holt e Smith (1979), Rao (1985), Särndal, Swensson e Wretman (1992) e Valliant (1993).

Outro problema da maioria das sondagens consiste na falta de obtenção, total ou parcial, de resposta aos questionários. Na presença de não respostas os estimadores usuais são enviesados.

Os estimadores de pós-estratificação inserem-se numa classe de métodos de tratamento de não respostas usualmente designados por métodos de recomposição ou métodos de ajustamento. Estes procedimentos consistem em reponderar a amostra, isto é, ajustar os coeficientes de extrapolação, por forma a que os pesos ajustados tenham em consideração as não respostas (Little, 1986). De um modo geral, estes métodos são utilizados no tratamento das não respostas totais.

Existem outros métodos que procuram lidar com o problema da não resposta, tanto na fase de planeamento e recolha dos dados, como na fase de estimação, por exemplo, os métodos de imputação (Lessler e Kalsbeek, 1992). A abordagem a estes métodos encontra-se fora do âmbito deste trabalho.

Quando se utilizam dados provenientes de inquéritos para inferir sobre parâmetros da população, é importante considerar os erros de amostragem. Alguns dos estimadores apresentados neste trabalho não são lineares, pelo que as expressões exactas do enviesamento e da variância são muito difíceis de obter, se não mesmo impossíveis. No caso particular dos estimadores de pós-estratificação, as suas propriedades são também difíceis de investigar, principalmente quando se consideram planos de sondagem complexos. Neste trabalho são abordados alguns dos métodos mais utilizados para contornar estes problemas, nomeadamente o *método de linearização de Taylor* e os *métodos de Bootstrap* introduzidos por Efron (1979).

Com o desenvolvimento dos computadores e das aplicações informáticas, os métodos de *Bootstrap*, entre outros métodos de re-amostragem (*resampling*), têm merecido especial atenção, revelando-se promissores para a estimação correcta da variância dos estimadores e a obtenção de intervalos de confiança válidos [Rao e Wu (1988), Sitter (1992a, 1992b), Chen e Sitter (1993) e Shao e Tu (1995)]. No entanto, as propriedades teóricas dos *estimadores Bootstrap* requerem ainda investigação, quando se consideram estimadores e planos de sondagem complexos.

As referências bibliográficas incluem as principais obras e artigos de referência. Contudo, não foi possível aceder a alguns trabalhos, por não se encontrarem

disponíveis em Portugal. Ainda assim, procurou-se contornar esta limitação, contactando diversos autores, alguns dos quais, amavelmente, nos forneceram cópia dos trabalhos solicitados.

Outra limitação deste trabalho deriva do facto de não ter sido possível obter alguns dados, de algumas variáveis do Inquérito às Empresas/Harmonizado, que possibilitariam a apresentação de mais alguns exemplos práticos da aplicação das técnicas de pós-estratificação. Estas aplicações seriam, não só pertinentes para este estudo, mas também extremamente interessantes de analisar.

O texto desta dissertação encontra-se estruturado em cinco capítulos e anexos. O capítulo um, que terminamos com a organização geral da tese, pretende fazer um enquadramento do estudo, justificar a importância do tema proposto e apresentar não só os objectivos deste trabalho, como também as limitações do mesmo.

No segundo capítulo faz-se o enquadramento teórico necessário à compreensão da metodologia que é apresentada nos capítulos seguintes. Para tal, introduz-se a notação e as definições essenciais da teoria das sondagens, apresentam-se alguns planos de sondagem aleatória e aborda-se o método de linearização de Taylor e os métodos de Bootstrap.

No terceiro capítulo é apresentada a fundamentação teórica que serviu de apoio à formulação dos objectivos bem como à definição da metodologia utilizada nas aplicações práticas. Esta fundamentação teórica teve por base a revisão de literatura específica relacionada não só com a descrição dos métodos de pós-estratificação, mas também com os erros não amostrais (erros na base de sondagem e ocorrência de não respostas no inquérito) que motivaram a investigação dessa metodologia.

No capítulo quatro, apresentam-se alguns exemplos práticos de aplicação dos métodos de pós-estratificação. É então referida a metodologia subjacente ao Inquérito às Empresas / Harmonizado de 1996, cujos dados serviram de base às aplicações práticas, e descrita a metodologia utilizada. São também apresentados e discutidos os resultados obtidos.

No quinto e último capítulo, apresentam-se as principais conclusões do trabalho e fazem-se algumas sugestões para futuras investigações.

2 TÓPICOS DE SONDAGENS

2.1 Introdução

A observação de todos os elementos ou indivíduos da população (recenseamento) é, na maioria das situações, impossível de efectuar, quer por questões de tempo e custos, quer por questões operacionais de implementação.

Para fazer face à crescente necessidade de informação, tanto por parte das empresas e instituições, como por parte dos particulares, surgiu a necessidade de desenvolver métodos estatísticos que permitissem recolher essa informação a partir da observação de apenas uma parte da população.

De um modo geral, o termo **sondagem** é utilizado para designar um conjunto de técnicas estatísticas que permitem inferir sobre determinadas características ou parâmetros da população ou universo, a partir de um conjunto limitado dos seus elementos (**amostra**).

O método de selecção dos elementos da amostra permite agrupar os métodos de sondagem em duas grandes categorias:

- os métodos **probabilísticos**
- os métodos **empíricos**

Nas subsecções que se seguem faz-se uma descrição resumida destes métodos, apresentam-se os vários tipos de erros associados às sondagens e referem-se as principais etapas para implementação de uma sondagem probabilística.

Este capítulo tem por objectivo apresentar o enquadramento teórico necessário para a compreensão da metodologia apresentada em capítulos posteriores. Os métodos de sondagem empíricos encontram-se fora do âmbito da dissertação, pelo que serão apresentados em mais detalhe alguns métodos de sondagem probabilísticos.

Na secção 2.2 introduz-se a notação e as definições essenciais da teoria das sondagens. Em seguida apresentam-se alguns planos de sondagem: a sondagem aleatória simples (secção 2.3); a sondagem aleatória com probabilidades desiguais (secção 2.4) e, em mais detalhe, a sondagem aleatória estratificada (secção 2.5).

Alguns dos estimadores apresentados neste trabalho não são lineares, pelo que as expressões exactas do enviesamento e da variância são muito difíceis de obter, se não mesmo impossíveis. O *método de linearização de Taylor*, também designado na literatura por *método-δ*, e os *métodos de Bootstrap* permitem contornar este problema e são abordados nas secções 2.6 e 2.7, respectivamente.

2.1.1 Métodos de sondagem empíricos

Os métodos de sondagem empíricos, também designados na literatura por métodos de escolha judiciosa, são especialmente utilizados em sondagens de opinião e estudos de mercado e caracterizam-se pelo facto de não ser possível *a priori* determinar a probabilidade de um elemento pertencer à amostra (Gomes, 1998, p. 21). A facilidade de implementação destes métodos e a flexibilidade de selecção dos elementos da amostra permite reduzir os custos e efectuar mais rapidamente a sondagem. No entanto, os métodos empíricos têm a grande desvantagem de não ser possível avaliar a qualidade dos resultados.

2.1.2 Métodos de sondagem probabilísticos

O princípio de base dos métodos probabilísticos é que a probabilidade de se seleccionar um elemento da população para a amostra é conhecida. Särndal, Swensson e Wretman (1992, p. 8) apresentam quatro condições necessárias para a obtenção de uma amostra probabilística de uma determinada população:

1. Ser possível definir o conjunto de todas as amostras, $S = \{s_1, s_2, ..., s_m\}$, que se podem obter através do procedimento de amostragem.
2. Ser conhecida a probabilidade *p(s)* de seleccionar a amostra *s*, do conjunto de amostras possíveis.
3. Ser não nula a probabilidade de seleccionar cada elemento da população para a amostra.

4. O processo de selecção dos elementos da amostra ser aleatório, i.e. não ser baseado em julgamentos empíricos, e tal que cada amostra *s* que se pode obter tenha exactamente a probabilidade *p(s)*.

São de salientar duas hipóteses fundamentais que estão subjacentes a estes métodos: a dimensão, *N*, da população é conhecida e é fixada a dimensão da amostra (*n*).

Neste contexto, designa-se formalmente por **plano de amostragem** ou **plano de sondagem** a função *p(.)* que define a distribuição de probabilidade sobre o conjunto $S = \{s_1, s_2, ..., s_m\}$. O plano de amostragem irá determinar as propriedades estatísticas dos estimadores (por exemplo, o valor esperado e a variância) que permitem avaliar a qualidade das estimativas obtidas.

Ao longo da dissertação, iremos utilizar os termos plano de amostragem, plano de sondagem ou desenho da amostra para referir genericamente a forma como a amostra foi seleccionada da população.

2.1.3 Erros de amostragem e erros não amostrais

Os erros que derivam de uma sondagem são essencialmente de dois tipos: os erros devidos à amostragem e os erros que não se devem à amostragem.

O erro que resulta de não se observar toda a população é designado erro de amostragem. No caso das sondagens probabilísticas é possível apresentar medidas da exactidão ou precisão (i.e. da qualidade) das estimativas obtidas a partir da amostra (veja-se a secção 2.2.2).

Os erros que não estão relacionados com o processo de amostragem designam-se erros não amostrais e podem ocorrer em qualquer fase da implementação da sondagem. Alguns exemplos deste tipo de erros são: os erros da base de sondagem (problemas de cobertura, informação auxiliar incorrecta ou desactualizada, ...); erros na recolha de informação (defeitos do questionário, erros no registo das respostas, não resposta total ou parcial, ...); erros no processamento dos dados (edição, codificação, análise, ...).

A qualidade dos resultados de uma sondagem depende, assim, da qualidade com que todas as suas etapas são implementadas.

2.1.4 Planeamento e implementação de uma sondagem

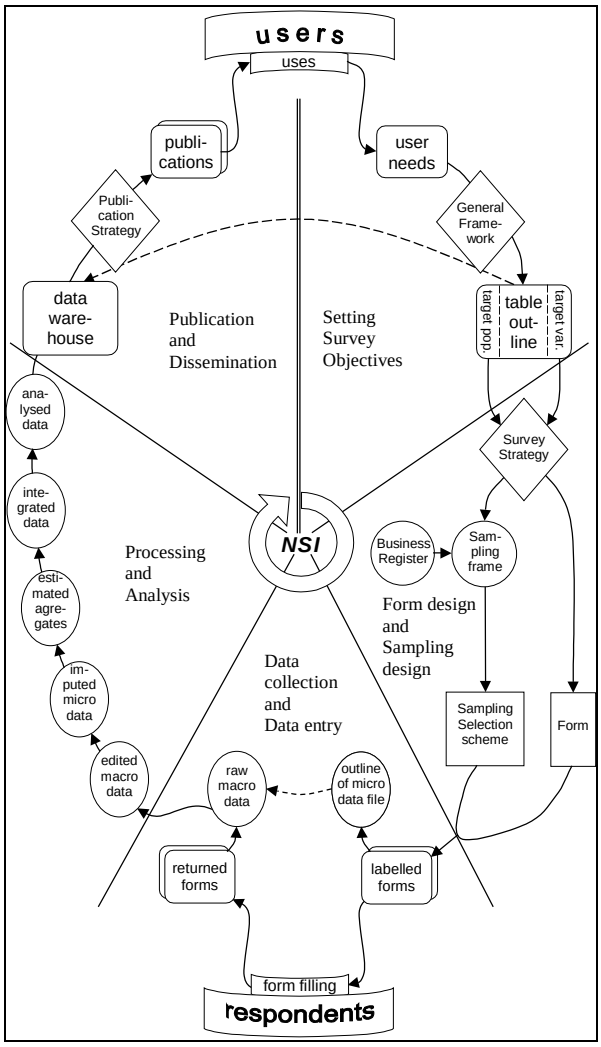
A concepção de uma sondagem é um processo que envolve diversas fases interdependentes, sendo importante encontrarem-se claramente definidos os conceitos, métodos e procedimentos.

Särndal, Swensson e Wretman (1992, p. 18) apresentam as principais etapas para a concepção e implementação de uma sondagem:

1. Especificação do objectivo da sondagem.
2. Tradução do problema em estudo num problema de sondagem.
3. Especificação da população alvo, variáveis de interesse, variáveis auxiliares disponíveis e parâmetros a estimar.
4. Construção ou obtenção da base de sondagem.
5. Inventariação dos recursos disponíveis em termos orçamentais, humanos, técnicos, de equipamentos, entre outros.
6. Especificação de requisitos a que a sondagem deve obedecer, como por exemplo a calendarização e a precisão das estimativas.
7. Especificação do método de recolha dos dados, incluindo a elaboração do questionário.
8. Especificação do desenho da amostra (plano de amostragem), mecanismo de selecção da amostra e determinação da sua dimensão.
9. Especificação dos métodos de processamento dos dados, incluindo a edição e imputação.
10. Especificação da forma dos estimadores e das medidas de precisão.
11. Treino dos recursos humanos e organização do trabalho de campo.
12. Alocação de recursos às diferentes operações da sondagem.
13. Alocação de recursos ao controlo e avaliação.

Relativamente à implementação de sondagens pelos Institutos Oficiais de Estatística (*NSI – National Statistical Institutes*), Koeijers e Willeboordse (1995) apresentam um manual de referência sobre o planeamento e implementação de inquéritos às empresas, onde as principais etapas podem ser apresentadas sob a forma de um ciclo ininterrupto (Figura 1).

Figura 1 – (Re) desenho e implementação de uma sondagem



Fonte: Koeijers e Willeboordse (1995)

Das diversas fases de implementação de uma sondagem, tem particular relevância para este trabalho aquela se refere à definição da base de sondagem, uma vez que a metodologia apresentada em capítulos posteriores é essencialmente motivada pela possível ocorrência de erros nessa etapa.

A base de amostragem, ou base de sondagem, é uma lista actualizada de todos os elementos da população alvo[1]. Idealmente, a base de sondagem deveria permitir identificar a população alvo na totalidade. A possível existência de erros na base de

[1] A população alvo ou universo de referência é o conjunto de elementos que, de acordo com os objectivos da sondagem, verificam as especificações estabelecidas. A população alvo

amostragem tem repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram (para mais detalhes veja-se a secção 3.2).

consiste em unidades amostrais, no sentido estatístico (por exemplo: indivíduos, empresas, famílias, ...).

2.2 Considerações gerais, definições e notação

2.2.1 Parâmetros de interesse na população

A população alvo ou universo de referência denota-se por *U* e considera-se que tem dimensão finita conhecida (salvo indicação em contrário) *N*. A cada elemento de U pode então ser associado um índice (k = 1, 2, …, N). Por uma questão de simplicidade de notação, vamos denotar o *k-ésimo* elemento da população pelo respectivo índice *k*:

(2.2.1) $U = \{1, \ldots, k, \ldots, N\}$

Assim, denotam-se por $y_1, y_2, \ldots, y_N$ os valores da variável de estudo *Y* na população *U*. Ao longo deste trabalho, vamos considerar como parâmetros de interesse na população o total e a média da variável *Y* (que designaremos simplesmente por total e média da população) que se denotam, respectivamente, por τ e μ, ou τ_y e μ_y, caso se pretenda deixar explicito a que variável estas quantidades se referem.

O total da população τ é a soma dos valores da variável de interesse *Y* para todos os elementos da população:

(2.2.2) $$\tau = \sum_{k \in U} y_k$$

A média da população μ corresponde à média dos valores da variável de interesse *Y* para todos os elementos da população:

(2.2.3) $$\mu = \tau/N = \frac{1}{N} \sum_{k \in U} y_k$$

2.2.2 Propriedades desejáveis e critérios de comparação dos estimadores

Quando se pretende inferir da amostra para a população e se dispõem de diversas técnicas de estimação, há que optar pelo estimador mais adequado, no sentido de

que este deverá fornecer estimativas que se aproximem o mais possível do valor do parâmetro desconhecido da população (i.e. o erro amostral deverá ser o menor possível). Ao analisar a qualidade dos estimadores recorre-se, geralmente, a duas propriedades fundamentais: o enviesamento e a dispersão.

2.2.2.1 Não enviesamento

Seja $\hat{\theta}$ um estimador do parâmetro $\theta = g(y_1, \ldots, y_N)$, onde *g* é uma função dos *N* valores da variável de estudo *Y* na população.

O **enviesamento** do estimador é definido pela diferença entre o valor esperado, ou esperança matemática, do estimador e o verdadeiro valor do parâmetro:

(2.2.4) $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Naturalmente, uma propriedade desejável para o estimador $\hat{\theta}$ é que este seja **centrado** ou **não enviesado**, ou seja, que $B(\hat{\theta}) = 0$.

2.2.2.2 Precisão

Para avaliar a dispersão da distribuição amostral do estimador é usual utilizar-se a **variância** ou o **desvio padrão** que se definem, respectivamente, por:

(2.2.5) $V(\hat{\theta}) = E[\,(\hat{\theta} - E(\hat{\theta}))^2\,]$

(2.2.6) $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$

Como uma medida do erro amostral utiliza-se, geralmente, o desvio padrão de uma estimativa não enviesada.

O quociente entre o desvio padrão do estimador e o seu valor esperado designa-se **coeficiente de variação** do estimador:

(2.2.7) $CV(\hat{\theta}) = \dfrac{\sigma_{\hat{\theta}}}{E(\hat{\theta})}$

Särndal, Swensson e Wretman (1992, p. 42) referem que, na prática, se designa por *coeficiente de variação* a quantidade (2.2.8) expressa em percentagem, sendo utilizada como um indicador da precisão obtida na sondagem, quando se utiliza um estimador $\hat{\theta}$ centrado ou quase centrado.

(2.2.8) $$cv(\hat{\theta}) = \frac{\hat{\sigma}_{\hat{\theta}}}{\hat{\theta}}$$

Mais formalmente, Särndal, Swensson e Wretman (1992, p. 42) descrevem o valor obtido através de (2.2.8) como uma estimativa, enviesada, do coeficiente de variação "teórico" (2.2.7).

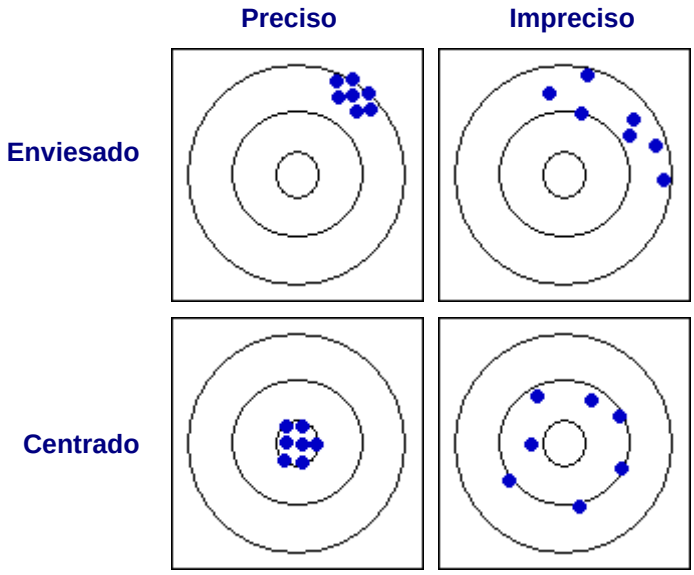
A precisão, sendo uma medida da proximidade esperada entre o estimador e o verdadeiro valor do parâmetro, pode ser medida através do **erro quadrático médio**, definido por:

(2.2.9) $$EQM(\hat{\theta}) = E[\,(\hat{\theta} - \theta)^2\,] = V(\hat{\theta}) + [\,B(\hat{\theta})\,]^2$$

A Figura 2 ilustra claramente a razão pela qual, regra geral, se opta pela utilização de estimadores centrados. No entanto, nalgumas situações, a utilização de um estimador com um enviesamento moderado é preferível, pelos seguintes motivos (Särndal, Swensson e Wretman, 1992, p. 164):

- muitos parâmetros têm uma estrutura formal que dificulta a determinação de um estimador centrado;
- um estimador com um enviesamento moderado pode muitas vezes ter variância e erro quadrático médio inferior a um estimador centrado.

Figura 2 – Ilustração do enviesamento e da precisão, sendo o verdadeiro valor o centro da menor circunferência[1]



Um critério de escolha entre um estimador enviesado $\hat{\theta}_1$ e um estimador centrado $\hat{\theta}_2$ é dado pela comparação entre o erro quadrático médio do primeiro estimador e a variância do segundo. Ou seja, se

(2.2.10) $\quad EQM(\hat{\theta}_1) < V(\hat{\theta}_2)$

então, o estimador enviesado $\hat{\theta}_1$ é preferível ao estimador centrado $\hat{\theta}_2$.

2.2.2.3 Efeito de sondagem

Uma das medidas de comparação entre dois estimadores centrados, designa-se por **efeito de sondagem** (*design effect*) e foi inicialmente definido por Kish (1965) como sendo o quociente de duas variâncias onde, no numerador, figurava a variância do estimador sob o plano de sondagem utilizado (*s*) e, no denominador, tinha-se como referência a variância do estimador correspondente a um plano de sondagem aleatória simples sem reposição (SASSR), para uma amostra com a mesma

[1] Esta figura baseia-se na ilustração apresentada na Internet por EASTON, V. J. e MCCOLL, J. H. (1998). *Statistics Glossary*. http://www.stats.gla.ac.uk/steps/glossary/sampling.html, Steps, vers. 1.1.

dimensão fixa *n*. Neste caso, o efeito de sondagem para o estimador do total da população (τ) sob o plano utilizado *s*, seria definido por:

$$(2.2.11) \qquad DEFF(\hat{\tau}) = \frac{V_S(\hat{\tau})}{V_{SASSR}(\hat{\tau})}$$

Mas, por vezes, tem interesse utilizar como referência a variância de outro estimador, que não o da sondagem aleatória simples, para analisar comparativamente a eficiência de um dado estimador. Assim, o efeito de sondagem pode ser definido de uma forma mais geral como se segue.

Seja $\hat{\theta}_S$ um estimador centrado do parâmetro $\theta = g(y_1, \ldots, y_N)$, sob um determinado plano de amostragem *s*. Dizemos que um método de sondagem s_1 é mais preciso do que outro método s_2 se, para a mesma dimensão amostral *n*, a medida de precisão de $\hat{\theta}_{s_1}$ for melhor do que para $\hat{\theta}_{s_2}$.

Neste contexto, vamos definir o efeito de sondagem dos planos s_1 e s_2 para os respectivos estimadores de θ da seguinte forma:

$$(2.2.12) \qquad DEFF(\hat{\theta}_{s_1} | \hat{\theta}_{s_2}) = \frac{V(\hat{\theta}_{s_1})}{V(\hat{\theta}_{s_2})}$$

Assim, se $DEFF(\hat{\theta}_{s_1} | \hat{\theta}_{s_2}) < 1$ concluímos que o método s_1 é mais preciso do que o método s_2.

Obtém-se um estimador do efeito de sondagem através do quociente:

$$(2.2.13) \qquad deff(\hat{\theta}_{s_1} | \hat{\theta}_{s_2}) = \frac{\hat{V}(\hat{\theta}_{s_1})}{\hat{V}(\hat{\theta}_{s_2})}$$

2.2.3 Intervalos de confiança

Nesta secção apresentam-se, de forma abreviada, as ideias básicas subjacentes ao conceito de intervalo de confiança para um parâmetro desconhecido de uma população finita. A exposição segue de perto a apresentada por Särndal, Swensson e Wretman (1992, p. 55-56, 163-166).

Seja $\hat{\theta}$ um estimador do parâmetro $\theta = g(y_1, \ldots, y_N)$, onde *g* é uma função dos *N* valores da variável de estudo *Y* na população. Tal como na teoria geral da inferência estatística, um intervalo de confiança é uma realização de um intervalo aleatório que tem uma determinada probabilidade de conter o verdadeiro valor (desconhecido) do parâmetro. Um **intervalo de confiança para θ** é dado por:

(2.2.14) $IC(s) = [\theta_{inf}(s) , \theta_{sup}(s)]$

onde $\theta_{inf}(s)$ e $\theta_{sup}(s)$ são duas estatísticas tais que $\theta_{inf}(s) \leq \theta_{sup}(s)$ para qualquer amostra aleatória *s*.

A probabilidade do intervalo (2.2.14) conter o verdadeiro valor de θ, designa-se **nível de confiança** ou **probabilidade de cobertura**. Pretende-se geralmente que essa probabilidade esteja próxima de 1 para uma dada amplitude.

Neste intervalo a aleatoriedade é introduzida pela amostra aleatória *s* seleccionada. No contexto da teoria das sondagens, um intervalo de confiança é interpretado com relação ao plano de sondagem *p(s)*, que define a distribuição de probabilidade sobre o conjunto de todas as amostras possíveis, como se passa a expor.

Seja $S_0 \subseteq S$ o conjunto de todas as amostras *s* tal que *p(s)*>0 e seja S_{0c} o subconjunto de S_0 tal que, para cada amostra $s \in S_0$, o intervalo (2.2.14) contém o verdadeiro valor de θ. Denote-se por S'_{0c} o complementar de S_{0c} em S_0. Nestas condições, a probabilidade de cobertura do intervalo de confiança (2.2.14) é dada por:

(2.2.15) $P[\ IC(s) \ni \theta\ ] = 1 - \alpha$

onde,

(2.2.16) $$\alpha = \sum_{s \in S'_{0c}} p(s)$$

Ou seja, α é a probabilidade acumulada das amostras *s* para as quais o intervalo não inclui θ.

As estatísticas $\theta_{inf}(s)$ e $\theta_{sup}(s)$ devem permitir obter intervalos com os níveis de confiança $1-\alpha$ desejados (por exemplo, os usuais 95%). No entanto, os estimadores típicos da teoria das sondagens permitem atingir este objectivo apenas de forma aproximada e sob determinadas condições.

Seja $\hat{\theta}$ um estimador centrado de θ e $z_{1-\alpha/2}$ o quantil de probabilidade $1-\alpha/2$ da distribuição Normal Standard, N(0, 1). Utiliza-se frequentemente o seguinte intervalo de confiança:

(2.2.17) $$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}$$

Este intervalo conterá o parâmetro desconhecido θ, para uma proporção aproximada de $1-\alpha$ de amostras *s* (obtidas segundo um determinado plano de sondagem), se as duas condições seguintes se verificarem:

1. A distribuição amostral de $\hat{\theta}$ é aproximadamente uma distribuição Normal de valor médio θ e variância $V(\hat{\theta})$.
2. Existe um estimador $\hat{V}(\hat{\theta})$ de $V(\hat{\theta})$ consistente.

Antes de se comentarem estas duas condições, note-se que:

(2.2.18) $$\frac{\hat{\theta}-\theta}{\sqrt{\hat{V}(\hat{\theta})}} = \frac{\hat{\theta}-\theta}{\sqrt{V(\hat{\theta})}}\sqrt{\frac{V(\hat{\theta})}{\hat{V}(\hat{\theta})}}$$

A primeira condição é essencialmente equivalente à aplicação do Teorema Limite Central. Ou seja, sob a primeira condição,

(2.2.19) $$\frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}}$$

tende para a distribuição N(0, 1), quando a dimensão da amostra aumenta.

Sob a segunda condição tem-se essencialmente que, quando a amostra é suficientemente grande, é também grande a probabilidade de

(2.2.20) $$\sqrt{\frac{\hat{V}(\hat{\theta})}{V(\hat{\theta})}}$$

estar próximo de 1.

Conclui-se assim que se pode tratar a variável (2.2.18) como tendo aproximadamente distribuição N(0, 1) se a amostra for suficientemente grande. Pelo que se justifica a utilização do quantil da distribuição Normal, $z_{1-\alpha/2}$, no intervalo (2.2.17)[1].

Särndal, Swensson e Wretman (1992, p. 57) apresentam referências bibliográficas relevantes para uma análise detalhada da validade teórica e empírica dos intervalos de confiança obtidos através de (2.2.17). Em seguida, comenta-se o efeito do enviesamento dos estimadores sobre os intervalos de confiança e, em particular, sobre (2.2.17).

2.2.3.1 Efeito do enviesamento dos estimadores

Uma propriedade desejável dos estimadores é, sem dúvida, o não enviesamento. No entanto, utilizam-se, por vezes, estimadores *aproximadamente* não enviesados (veja-se a secção 2.2.2). Neste caso, uma medida da precisão de um estimador $\hat{\theta}$, com enviesamento $B(\hat{\theta}) = E(\hat{\theta}) - \theta$, é o erro quadrático médio:

---

[1] Observe-se que, se a variância do estimador for conhecida, o valor de $V(\hat{\theta})$ deverá ser utilizado no intervalo (2.2.17) em vez de $\hat{V}(\hat{\theta})$, uma vez que a aproximação à distribuição Normal poderá ser mais rápida, neste caso. Assim, é de esperar que o intervalo que se obtém utilizando $V(\hat{\theta})$, para uma dada dimensão amostral, tenha uma probabilidade de cobertura mais próxima de $1-\alpha$ do que o intervalo (2.2.17).

(2.2.21) $EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [B(\hat{\theta})]^2$

Para além de ser desejável que o $EQM(\hat{\theta})$ seja pequeno, também é conveniente que o enviesamento do estimador seja pequeno relativamente ao desvio padrão $[V(\hat{\theta})]^{1/2}$. Estas considerações são importantes uma vez que condicionam a validade dos intervalos de confiança, como se verá, resumidamente, em seguida.

Antes de mais, considere-se o quociente (*bias ratio*):

(2.2.22) $$BR(\hat{\theta}) = \frac{B(\hat{\theta})}{\sqrt{V(\hat{\theta})}}$$

Särndal, Swensson e Wretman (1992, p. 164) referem que, ainda que um determinado estimador seja enviesado, desde que $BR(\hat{\theta})$ seja pequeno, os intervalo de confiança que se obtêm não terão um erro muito grande.

Para simplificar, suponhamos que por hipótese:

(2.2.23) $$Z = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{V(\hat{\theta})}} \sim N(0, 1)$$

ou seja, que Z tem distribuição N(0, 1).

Nestas condições, considere-se o seguinte intervalo:

(2.2.24) $$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{V(\hat{\theta})}$$

O nível de confiança ou probabilidade de cobertura deste intervalo é:

(2.2.25) $$P_0 = P\{\hat{\theta} - z_{1-\alpha/2}\sqrt{V(\hat{\theta})} < \theta < \hat{\theta} + z_{1-\alpha/2}\sqrt{V(\hat{\theta})}\} =$$

$$= P\{-z_{1-\alpha/2} - BR(\hat{\theta}) < Z < z_{1-\alpha/2} - BR(\hat{\theta})\}$$

onde Z é a variável aleatória definida em (2.2.23).

Sob a hipótese da normalidade da variável Z e supondo que $V(\hat{\theta})$ é conhecido, a probabilidade de cobertura do intervalo (2.2.24) é $1-\alpha$ apenas se $BR(\hat{\theta})$ for zero. Assim, o efeito do enviesamento sobre o nível de confiança será pequeno apenas se $BR(\hat{\theta})$ for próximo de zero.

O Quadro 2.2.1 apresenta as probabilidades de cobertura P_0 para alguns valores de $BR(\hat{\theta})$, quando se toma $1-\alpha = 95\%$.

Quadro 2.2.1 - Probabilidade de cobertura P_0 como função de $BR(\hat{\theta})$

$\lvert BR(\hat{\theta})\rvert$	P_0
0.00	0.9500
0.05	0.9497
0.10	0.9489
0.30	0.9396
0.50	0.9210
1.00	0.8300

Fonte: Särndal, Swensson e Wretman (1992, p. 165)

Este quadro tem por objectivo dar uma ideia aproximada do efeito de $BR(\hat{\theta})$ sobre as probabilidades de cobertura dos intervalos. Na prática, $BR(\hat{\theta})$ é desconhecido e, geralmente, a hipótese formulada em (2.2.23) corresponde apenas a uma aproximação, quando se utilizam amostras grandes. Fica, no entanto, clara a importância da relação entre do enviesamento e a variância dos estimadores para a obtenção de intervalos com os níveis de cobertura desejados.

Uma vez que, geralmente, tanto o enviesamento como a variância do estimador são desconhecidos, utiliza-se frequentemente o seguinte intervalo:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{EQM}(\hat{\theta})} \qquad (2.2.26)$$

onde $\hat{EQM}(\hat{\theta})$ é um estimador do erro quadrático médio do estimador enviesado $\hat{\theta}$.

As propriedades de cobertura do intervalo (2.2.26) podem ser analisadas de forma análoga à apresentada para o intervalo (2.2.24), supondo que EQM($\hat{\theta}$) é conhecido.

2.2.4 Consistência e não enviesamento assimptótico

As definições de consistência e não enviesamento assimptótico, da teoria geral da inferência estatística, não podem ser aplicadas directamente aos estimadores definidos sobre amostras de uma população finita. Se for *N* a dimensão da população *U* e *s* uma amostra de dimensão *n*, uma vez que n≤N, naturalmente, não é possível calcular limites com $n \to \infty$.

A utilização destes conceitos no âmbito da teoria das sondagens requer ferramentas matemáticas mais complexas, pelo que não serão aqui apresentados. No entanto, referir-se-ão algumas considerações essenciais, pelo que, seguidamente, se relembram as definições da teoria geral da inferência estatística. Referências bibliográficas relevantes sobre este tema podem ser encontradas em Särndal, Swensson e Wretman (1992).

Seja θ um parâmetro desconhecido na população e considere-se o estimador $\hat{\theta}_n$ definido como uma função de *n* variáveis aleatórias $\xi_1, \xi_2, \ldots, \xi_n$ independentes e identicamente distribuídas. Nestas condições, diz-se que $\hat{\theta}_n$ é um estimador assimptoticamente centrado de θ se

$$\text{(2.2.27)} \qquad \lim_{n \to \infty} E(\hat{\theta}_n) = \theta$$

e, diz-se que $\hat{\theta}_n$ é consistente para θ se, para qualquer valor $\varepsilon>0$ fixo:

$$\text{(2.2.28)} \qquad \lim_{n \to \infty} P\left\{ \left| \hat{\theta}_n - \theta \right| > \varepsilon \right\} = 0$$

Os conceitos de consistência e não enviesamento assimptótico de um estimador são importantes, no âmbito da teoria das sondagens, essencialmente pelos seguintes motivos. Se for possível mostrar que um estimador é assimptoticamente centrado então, pode-se considerar que é *aproximadamente* não enviesado se a dimensão da amostra for suficientemente grande. Por outro lado, se o estimador for consistente,

pode-se considerar que a sua distribuição amostral se concentra muito próximo do verdadeiro valor do parâmetro, quando as amostras são de dimensão suficientemente grande.

Consideram-se por vezes, neste trabalho, estimadores assimptoticamente centrados (*aproximadamente* não enviesados) ou consistentes sem demonstração formal, uma vez que estas não se encontram no âmbito da dissertação. Recorrer-se-á então a referências bibliográficas.

2.2.5 Probabilidades de inclusão

Para inferir da amostra para a população é fundamental determinar as probabilidades de inclusão. A sua definição formal é a seguinte:

Definição 1

Designa-se por **probabilidade de inclusão de 1ª ordem** a probabilidade de um indivíduo da população ser seleccionado para a amostra, ou seja:

$$\pi_i = P(i \in s), \qquad i \in U$$

Definição 2

A **probabilidade de inclusão de 2ª ordem** corresponde à probabilidade de dois indivíduos da população serem seleccionados para a amostra:

$$\pi_{ij} = P(i \in s \cap j \in s), \qquad i \neq j, \qquad i, j \in U.$$

Naturalmente, estas probabilidades dependem da forma como os elementos são seleccionados e, portando, do plano de sondagem adoptado.

Vamos agora estudar as propriedades de duas variáveis indicatriz que estão estreitamente relacionadas com as probabilidades de inclusão. Estas variáveis irão ser fundamentais para a demonstração de alguns dos resultados apresentados em posteriores secções; devendo-se a Cornfield (1944) a sugestão da sua utilização na demonstração dos principais resultados da sondagem aleatória simples sem reposição.

Considere-se a **variável indicatriz** $\mathbb{1}_{i\in s}$, também designada **variável de Cornfield**, definida por:

$$(2.2.29) \qquad \mathbb{1}_{i\in s} = \begin{cases} 1 & \text{se } i\in s \ (i\in U) \\ 0 & \text{se } i\notin s \ (i\in U) \end{cases}$$

Uma vez que a variável $\mathbb{1}_{i\in s}$ segue uma distribuição de Bernoulli de parâmetro π_i, tem-se:

$$(2.2.30) \qquad E(\mathbb{1}_{i\in s}) = \pi_i \ , \qquad i\in U$$

$$(2.2.31) \qquad V(\mathbb{1}_{i\in s}) = \pi_i (1 - \pi_i) \ , \qquad i\in U$$

Seja $\mathbb{1}_{i,j\in s}$ a **variável indicatriz** definida por:

$$(2.2.32) \qquad \mathbb{1}_{i,j\in s} = \begin{cases} 1 & \text{se } i\in s \text{ e } j\in s \ , \ i\neq j \ (i,j\in U) \\ 0 & \text{caso contrário} \end{cases}$$

Analogamente, a variável $\mathbb{1}_{i,j\in s}$ segue uma distribuição de Bernoulli de parâmetro π_{ij}, donde,

$$(2.2.33) \qquad E(\mathbb{1}_{i,j\in s}) = \pi_{ij} \ , \qquad i, j\in U$$

$$(2.2.34) \qquad V(\mathbb{1}_{i,j\in s}) = \pi_{ij} (1 - \pi_{ij}) \ , \qquad i, j\in U$$

Um resultado extremamente útil resulta do facto de, para $i \neq j$, o produto $\mathbb{1}_{i\in s} \times \mathbb{1}_{j\in s}$ tomar sempre o valor 1, excepto quando os elementos *i* e *j* não pertencem simultaneamente à amostra. Ou seja,

$$(2.2.35) \qquad \mathbb{1}_{i\in s} \times \mathbb{1}_{j\in s} = \mathbb{1}_{i,j\in s} \ , \ i\neq j$$

Assim, facilmente se deduz a expressão da covariância entre essas duas variáveis:

$$(2.2.36) \qquad \begin{aligned} \mathrm{Cov}(\mathbb{1}_{i\in s}, \mathbb{1}_{j\in s}) &= E(\mathbb{1}_{i\in s} \times \mathbb{1}_{j\in s}) - E(\mathbb{1}_{i\in s}) \, E(\mathbb{1}_{j\in s}) = \\ &= E(\mathbb{1}_{i,j\in s}) - E(\mathbb{1}_{i\in s}) \, E(\mathbb{1}_{j\in s}) = \pi_{ij} - \pi_i \, \pi_j \end{aligned}$$

que denotaremos por Δ_{ij}, ou seja,

(2.2.37) $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j , \quad i \neq j, \quad i, j \in U$

$\Delta_{ii} = \pi_i (1 - \pi_i)$

Esta notação permitirá simplificar a forma de algumas expressões apresentadas em posteriores secções.

2.3 Sondagem aleatória simples

Diz-se que o desenho de uma amostra de dimensão *n*, retirada de uma população de *N* elementos, corresponde a uma **sondagem aleatória simples** quando todos os elementos da população têm a mesma probabilidade de serem escolhidos para fazer parte da amostra. Ou seja, qualquer combinação de *n* elementos da população tem a mesma probabilidade de corresponder à amostra seleccionada.

Este plano de sondagem é muito utilizado devido ao facto de a sua implementação ser mais simples e geralmente com menores custos, do que outros planos mais complexos. Outro factor a seu favor é o facto de não ser necessária muita informação sobre a população que se pretende estudar, comparativamente a outras técnicas de sondagens. No entanto, a sondagem aleatória simples deve ser utilizada apenas quando a população é homogénea; o que se torna limitativo quando se pretende estudar múltiplos atributos da população. Se for esse o caso, este plano só é adequado se a dimensão da amostra for razoavelmente grande, devendo-se optar, de preferência, por outros planos de sondagens, como por exemplo a sondagem aleatória estratificada.

A razão fundamental pela qual introduzimos este tipo de sondagem prende-se com o facto de grande parte dos princípios da amostragem serem explicados em termos de sondagem aleatória simples e depois adaptados a desenhos mais complexos. Por outro lado, a teoria da sondagem aleatória simples pode, sob certas condições, fornecer um guia de comparação da precisão que se espera obter quando são utilizados desenhos mais complexos.

No caso da sondagem aleatória simples há que destinguir dois planos de amostragem: o caso em que as tiragens são efectuadas com reposição (SASCR) e sem reposição (SASSR).

Como veremos nas próximas secções, a sondagem aleatória simples sem reposição é mais eficiente do que a sondagem com reposição, sendo, portanto, a mais utilizada. Assim, a SASSR será apresentada mais detalhadamente e será este um dos planos que utilizaremos para efectuar comparações relativamente à precisão que se espera obter quando são utilizados desenhos mais complexos.

As demonstrações das propriedades apresentadas nesta secção podem ser consultadas em diversos livros de texto como, por exemplo, em Grosbras (1987). As que se referem à SASSR podem também ser facilmente derivadas a partir dos resultados do estimador de Horvitz-Thompson (veja-se a secção 2.4).

2.3.1 Sondagem aleatória simples com reposição (SASCR)

Quando as tiragens são efectuadas com reposição, o mesmo indivíduo pode ser seleccionado mais do que uma vez para a amostra. Neste caso, há N^n amostras possíveis e, consequentemente, a probabilidade associada a cada amostra, *s*, de dimensão fixa *n* é:

$$(2.3.1) \qquad p(s) = \frac{1}{N^n}$$

A probabilidade de inclusão de 1ª ordem corresponde, neste caso, à probabilidade do indivíduo *i* ser seleccionado pelo menos uma vez para fazer parte da amostra. Como as tiragens são independentes e a probabilidade de um indivíduo nunca pertencer à amostra é $\left(1-\frac{1}{N}\right)^n$, tem-se:

$$(2.3.2) \qquad \pi_i = 1 - \left(1-\frac{1}{N}\right)^n , \qquad i = 1, 2, \ldots, N$$

De forma análoga se deduz a probabilidade:

$$\begin{aligned} P(i\in s \cap j\in s) &= 1 - P[(i\notin s) \vee (j\notin s)] = \\ &= 1 - \{ P(i\notin s) + P(j\notin s) - P[(i\notin s) \wedge (j\notin s)] \} = \\ &= 1 - \left[\left(1-\frac{1}{N}\right)^n + \left(1-\frac{1}{N}\right)^n - \left(1-\frac{2}{N}\right)^n\right] = \\ &= 1 - 2\left(1-\frac{1}{N}\right)^n + \left(1-\frac{2}{N}\right)^n , \quad i, j = 1, 2, \ldots, N \; ; \; i \neq j \end{aligned}$$

e, portanto, a probabilidade de inclusão de 2ª ordem é dada por

(2.3.3) $$\pi_{ij} = 1 - 2\left(1-\frac{1}{N}\right)^n + \left(1-\frac{2}{N}\right)^n \quad , \quad i, j = 1, 2, \ldots, N \; ; \; i \neq j$$

Note-se que se *n* for pequeno relativamente a *N* (n << N) tem-se:

$$\pi_i \approx \frac{n}{N} \; , \qquad i = 1, 2, \ldots, N$$

$$\pi_{ij} \approx \frac{n(n-1)}{N^2} \; , \qquad i, j = 1, 2, \ldots, N \; ; \; i \neq j$$

2.3.1.1 Estimação de μ

A média da população μ corresponde à média dos valores da variável de interesse *Y* para todos os elementos da população:

(2.3.4) $$\mu = \frac{1}{N}\sum_{i \in U} y_i$$

A média amostral, $\bar{y}$, é um estimador centrado de μ,

(2.3.5) $$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i \in s} y_i \; ,$$

sendo a sua variância

(2.3.6) $$V(\bar{y}) = \frac{\sigma^2}{n} \; ,$$

com

(2.3.7) $$\sigma^2 = \frac{1}{N}\sum_{i \in U} (y_i - \mu)^2 \; .$$

A variância de $\bar{y}$ pode ser estimada sem enviesamento por:

(2.3.8) $$\hat{V}(\bar{y}) = \frac{s^2}{n} ,$$

sendo s^2 a variância amostral corrigida, dada por:

(2.3.9) $$s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2 ,$$

uma vez que s^2 é um estimador não enviesado de σ^2, no caso da SASCR.

2.3.1.2 Estimação de τ

O total da população τ não é mais do que a soma dos valores da variável de interesse *Y* para todos os elementos da população:

(2.3.10) $$\tau = \sum_{i\in U} y_i$$

Uma vez que $\tau = N\mu$, o estimador natural deste parâmetro será

(2.3.11) $$\hat{\tau} = N\bar{y} ,$$

sendo imediato que $E(\hat{\tau}) = \tau$, ou seja, o estimador é centrado e tem variância

(2.3.12) $$V(\hat{\tau}) = N^2 \frac{\sigma^2}{n} ,$$

sendo

(2.3.13) $$\hat{V}(\hat{\tau}) = N^2 \frac{s^2}{n} ,$$

um estimador centrado de $V(\hat{\tau})$.

2.3.2 Sondagem aleatória simples sem reposição (SASSR)

Quando as tiragens são efectuadas sem reposição, o mesmo indivíduo só pode ser seleccionado uma única vez para a amostra. Neste caso há $\binom{N}{n}$ amostras distintas possíveis, sendo a probabilidade associada a cada amostra *s* de dimensão fixa *n* dada por:

$$(2.3.14) \qquad p(s) = \frac{1}{\binom{N}{n}}$$

Dado que o número de amostras que incluem um dado elemento *i* da população é $\binom{N-1}{n-1}$, a probabilidade de inclusão de 1ª ordem é dada por:

$$(2.3.15) \qquad \pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}\ , \qquad i = 1, 2, \ldots, N$$

De forma análoga, dado que o número de amostras que contêm simultaneamente os elementos *i* e *j* ($i \neq j$) é $\binom{N-2}{n-2}$ e o número de amostras distintas possíveis é $\binom{N}{n}$, a probabilidade de inclusão de 2ª ordem é

$$(2.3.16) \qquad \pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}\ , \qquad i, j = 1, 2, \ldots, N\ ;\ i \neq j$$

2.3.2.1 Estimação de μ

A média amostral $\bar{y}$ é um estimador centrado da média da população μ:

$$(2.3.17) \qquad \hat{\mu} = \bar{y}$$

Utilizando a seguinte identidade

(2.3.18) $$\bar{y} = \frac{1}{n}\sum_{i\in S} y_i = \frac{1}{n}\sum_{i\in U} y_i \mathbb{I}_{i\in S}$$

e as propriedades da variável indicatriz, já apresentada, facilmente se demonstra que $E(\bar{y}) = \mu$:

(2.3.19) $$E(\bar{y}) = \frac{1}{n}\sum_{i\in U} y_i E(\mathbb{I}_{i\in S}) = \frac{1}{n}\sum_{i\in U} y_i \pi_i = \frac{1}{n}\sum_{i\in U} y_i \frac{n}{N} = \mu$$

A variância de $\bar{y}$ é

(2.3.20) $$V(\bar{y}) = (1-f)\frac{S^2}{n},$$

sendo S^2 a variância corrigida da população e *f* a taxa de sondagem:

(2.3.21) $$S^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \mu)^2$$

(2.3.22) $$f = \frac{n}{N}$$

A demonstração da propriedade (2.3.20) pode ser encontrada, por exemplo, em Hansen, Hurwitz e Madow (1953b, p. 92-96) ou em Grosbras (1987, p. 16-18).

No caso da sondagem aleatória simples sem reposição, a variância de $\bar{y}$ é agora estimada, sem enviesamento, por:

(2.3.23) $$\hat{V}(\hat{\mu}) = (1-f)\frac{s^2}{n},$$

uma vez que a variância amostral corrigida, s^2, é um estimador não enviesado de S^2 (variância corrigida da população), no caso da SASSR.

2.3.2.2 Estimação de τ

Pelo que foi exposto anteriormente, um estimador centrado de τ será, obviamente,

(2.3.24) $\hat{\tau} = N\bar{y}$

com variância

(2.3.25) $V(\hat{\tau}) = N^2(1-f)\frac{S^2}{n}$,

que pode ser estimada sem enviesamento por:

(2.3.26) $\hat{V}(\hat{\tau}) = N^2(1-f)\frac{s^2}{n}$.

2.3.3 Comparação entre os estimadores SASCR e SASSR

Comparando os estimadores SASCR e SASSR através do efeito de sondagem, definido por (2.2.12), conclui-se facilmente, como veremos, que o estimador da sondagem aleatória simples sem reposição é mais preciso do que o da amostragem com reposição.

Os quadros seguintes apresentam um resumo dos resultados relativos aos planos de sondagem aleatória simples com reposição e sem reposição, no que se refere aos estimadores de μ (Quadro 2.3.1) e τ (Quadro 2.3.2).

Quadro 2.3.1 – Propriedades do estimador de μ, para a SAS

SASCR	SASSR
$\hat{\mu} = \bar{y}$	$\hat{\mu} = \bar{y}$
$E(\hat{\mu}) = \mu$	$E(\hat{\mu}) = \mu$
$V(\hat{\mu}) = \frac{\sigma^2}{n}$	$V(\hat{\mu}) = (1-f)\frac{S^2}{n}$
$\hat{V}(\hat{\mu}) = \frac{s^2}{n}$	$\hat{V}(\hat{\mu}) = (1-f)\frac{s^2}{n}$

Quadro 2.3.2 – Propriedades do estimador de τ, para a SAS

SASCR	SASSR
$\hat{\tau} = N\bar{y}$	$\hat{\tau} = N\bar{y}$
$E(\hat{\tau}) = \tau$	$E(\hat{\tau}) = \tau$
$V(\hat{\tau}) = N^2\frac{\sigma^2}{n}$	$V(\hat{\tau}) = N^2(1-f)\frac{S^2}{n}$
$\hat{V}(\hat{\tau}) = N^2\frac{s^2}{n}$	$\hat{V}(\hat{\tau}) = N^2(1-f)\frac{s^2}{n}$

Considere-se como parâmetro de interesse, por exemplo, $\theta = \mu$. O efeito de sondagem para os planos de sondagem aleatória simples com e sem reposição é dado por:

$$(2.3.27) \qquad DEFF(\hat{\mu}_{SASSR} \mid \hat{\mu}_{SASCR}) = \frac{(1-f)\frac{S^2}{n}}{\frac{\sigma^2}{n}} = \frac{\frac{N-n}{N-1}\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} = \frac{N-n}{N-1}$$

ou seja,

$$(2.3.28) \qquad DEFF(\hat{\mu}_{SASSR} \mid \hat{\mu}_{SASCR}) \approx 1 - f < 1, \qquad \text{para } n > 1$$

Note-se que o ganho de precisão é fraco se *n* for pequeno relativamente a *N*, visto que, dessa forma, $f \approx 0$.

2.4 Sondagem aleatória com probabilidades desiguais

Os métodos de sondagem aleatória para os quais nem todos os elementos da população têm a mesma probabilidade de serem incluídos na amostra designam-se métodos de sondagem com probabilidades desiguais.

A utilização das probabilidades de inclusão desiguais pode resultar implicitamente do desenho escolhido para a amostra (por exemplo, no caso da amostragem estratificada), ou pode resultar de uma decisão tomada propositadamente com o intuito de incluir na amostra determinados indivíduos da população com maior (ou menor) probabilidade em virtude de eles serem mais (ou menos) importantes para o objectivo da sondagem (Thompson, 1992, p. 46). Coelho (1995, p. 32) refere que este método de sondagem tem também interesse quando as probabilidades de tiragem estão correlacionadas com o fenómeno em estudo.

Sejam quais forem as razões que levem à utilização de probabilidades de inclusão desiguais, estas têm que ser tomadas em consideração na forma dos estimadores, de modo a que as suas propriedades não se deteriorem e, se possível, se obtenham ganhos significativos de precisão.

Também neste caso, podemos distinguir dois tipos de amostragem: com e sem reposição. Hansen e Hurwitz (1943) introduziram um estimador para a abordagem com reposição. As propriedades deste estimador, bem como a descrição do plano de sondagens com probabilidades desiguais com reposição, podem ser facilmente encontradas na literatura (veja-se, por exemplo, Hansen, Hurwitz e Madow (1953b); Grosbras (1987) e Thompson (1992)).

A teoria geral da amostragem com probabilidades desiguais (com e sem reposição) foi desenvolvida por Horvitz e Thompson (1952), tendo sido crucial para o desenvolvimento de métodos de estimação em sondagens aleatórias, pelo que apresentaremos apenas esta abordagem. Para uma descrição detalhada do estimador de Horvitz-Thompson e das suas propriedades consulte-se, por exemplo, Gourieroux (1987) ou Särndal, Swensson e Wretman (1992).

2.4.1 Estimação de τ

Dadas as probabilidades de inclusão de 1ª ordem ($\pi_i > 0, i \in U$), o estimador do total da população proposto por Horvitz e Thompson é

$$(2.4.1) \qquad \hat{\tau}_{HT} = \sum_{i=1}^{\upsilon} \frac{y_i}{\pi_i}$$

onde, υ corresponde ao número de unidades distintas na amostra[1].

No caso em que as tiragens são efectuadas sem reposição, o estimador de Horvitz-Thompson é dado por:

$$(2.4.2) \qquad \hat{\tau}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Trata-se obviamente de um estimador centrado, como se demonstra facilmente:

$$(2.4.3) \qquad E(\hat{\tau}_{HT}) = E\Big[\sum_{i \in U} \frac{y_i}{\pi_i} \mathbb{I}_{i \in s}\Big] = \sum_{i \in U} \frac{y_i}{\pi_i} E[\mathbb{I}_{i \in s}] = \sum_{i \in U} \frac{y_i}{\pi_i} \pi_i = \tau$$

A variância do estimador é:

$$(2.4.4) \qquad V_1(\hat{\tau}_{HT}) = \sum_{i \in U} \left(\frac{1 - \pi_i}{\pi_i} \right) y_i^2 + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j$$

A demonstração deste resultado é efectuada utilizando as propriedades das variáveis indicatriz, introduzidas anteriormente (c.f. secção 2.2.5):

$$(2.4.5) \qquad V_1(\hat{\tau}_{HT}) = V\Big[\sum_{i \in U} \frac{y_i}{\pi_i} \mathbb{I}_{i \in s}\Big] =$$

[1] Para mais detalhes veja-se, por exemplo, Thompson (1992)

$$= \sum_{i\in U} \frac{y_i^2}{\pi_i^2} V(\mathbb{I}_{i\in s}) + \sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov(\mathbb{I}_{i\in s}, \mathbb{I}_{j\in s}) =$$

$$= \sum_{i\in U} \frac{y_i^2}{\pi_i^2} \pi_i(1-\pi_i) + \sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i\pi_j)$$

Se todas as probabilidades de inclusão conjuntas forem maiores que zero ($\pi_{ij} > 0$), a variância do estimador de Horvitz-Thompson pode ser estimada sem enviesamento por:

(2.4.6) $$\hat{V}_1(\hat{\tau}_{HT}) = \sum_{i\in s} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i\in s} \sum_{\substack{j\in s \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

A demonstração deste resultado é efectuada utilizando também as propriedades das variáveis indicatriz. Assim, desde que $\pi_{ij} > 0 \ \forall i,j \in U$, o estimador (2.4.6) pode ser escrito como:

(2.4.7) $$\hat{V}_1(\hat{\tau}_{HT}) = \sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 \mathbb{I}_{i\in s} + \sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \mathbb{I}_{i\in s} \mathbb{I}_{j\in s}$$

Conclui-se então o resultado pretendido, usando as propriedades das variáveis indicatriz:

(2.4.8) $$E[\hat{V}_1(\hat{\tau}_{HT})] = \sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 E(\mathbb{I}_{i\in s}) + \sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} E(\mathbb{I}_{i\in s} \mathbb{I}_{j\in s})$$

$$= \sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 \pi_i + \sum_{i\in U} \sum_{\substack{j\in U \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \pi_{ij}$$

$$= V_1(\hat{\tau}_{HT})$$

Se a dimensão da amostra, *n*, for fixa, pode-se também considerar a formulação alternativa para a variância de $\hat{\tau}_{HT}$ devida a Sen, Yates e Grundy:

$$(2.4.9) \qquad V_2(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U \\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Se todas as probabilidades de inclusão conjuntas forem maiores que zero ($\pi_{ij} > 0$), um estimador centrado de $V_2(\hat{\tau}_{HT})$ é:

$$(2.4.10) \qquad \hat{V}_2(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

A demonstração do resultado (2.4.9) encontra-se no Anexo 2, secção A2.1.1, sendo a demonstração de (2.4.10) análoga. O primeiro estimador da variância (2.4.6) é atribuído a Horvitz e Thompson (1952). O estimador (2.4.10) deve-se a Yates e Grundy (1953) e Sen (1953).

Observe-se que $\hat{V}_1(\hat{\tau}_{HT})$ poderá ser inferior ou igual a zero. Uma condição suficiente para que $\hat{V}_2(\hat{\tau}_{HT})$ seja superior ou igual a zero é (Condição de Yates-Grundy):

$$(2.4.11) \qquad \pi_i\pi_j - \pi_{ij} \geq 0, \qquad \forall\ i, j \in U \ \ (i \neq j)$$

Devido à dificuldade de implementação de qualquer um destes estimadores, Thompson (1992) cita vários autores que procuraram outras formas de estimar a variância (Hájek, 1981; Brewer e Hanif, 1983; Kott, 1988).

O Quadro 2.4.1 apresenta um resumo dos resultados relativos ao plano de sondagem aleatória com probabilidades desiguais, quando as tiragens são efectuadas sem reposição, no que se refere ao estimador de Horvitz-Thompson para τ.

Quadro 2.4.1 – Propriedades do estimador de Horvitz-Thompson para τ, para a sondagem aleatória com probabilidades desiguais (tiragens sem reposição)

Estimação de τ
$\hat{\tau}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$
$E(\hat{\tau}_{HT}) = \tau$
$V_1(\hat{\tau}_{HT}) = \sum_{i \in U} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j$
$\hat{V}_1(\hat{\tau}_{HT}) = \sum_{i \in s} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$
$E[\hat{V}_1(\hat{\tau}_{HT})] = V_1(\hat{\tau}_{HT})$ se $\pi_{ij} > 0, \forall\, i,j \in U, \quad i \neq j$
Se *s* tiver dimensão fixa:
$V_2(\hat{\tau}_{HT}) = \frac{1}{2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$
$\hat{V}_2(\hat{\tau}_{HT}) = \frac{1}{2} \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$
$E[\hat{V}_2(\hat{\tau}_{HT})] = V_2(\hat{\tau}_{HT})$ se $\pi_{ij} > 0, \forall\, i,j \in U, \quad i \neq j$

2.4.2 Estimação de μ

Naturalmente, se o parâmetro de interesse for a média da população, obtêm-se as seguintes propriedades, para o estimador de Horvitz-Thompson:

(2.4.12) $$\hat{\mu}_{HT} = \frac{1}{N} \hat{\tau}_{HT}$$

(2.4.13) $$V(\hat{\mu}_{HT}) = \frac{1}{N^2} V(\hat{\tau}_{HT})$$

(2.4.14) $$\hat{V}(\hat{\mu}_{HT}) = \frac{1}{N^2}\hat{V}(\hat{\tau}_{HT})$$

2.4.3 Pesos de inclusão

A forma do estimador de Horvitz-Thompson evidencia a importância das probabilidades de inclusão para inferir da amostra para a população. O inverso de π_i é usualmente designado por **coeficiente de extrapolação** ou **peso de inclusão** (*design weight*) do elemento *i*, e é denotado por $w_i = 1/\pi_i$. O termo peso ou ponderador refere-se geralmente aos coeficientes de extrapolação.

Numa sondagem por amostragem aleatória simples sem reposição, as probabilidades de inclusão são $\pi_i = n/N$. Neste caso, os estimadores para o total (2.3.24) e para a média da população (2.3.17) são equivalentes ao estimador de Horvitz-Thompson.

2.5 Sondagem aleatória estratificada

De um modo geral, as técnicas de amostragem estratificada são utilizadas quando a população é heterogénea e é possível identificar determinados grupos homogéneos, sendo de esperar que o parâmetro de interesse varie entre as diferentes sub-populações. A sondagem aleatória simples (SAS) é mais adequada quando toda a população é homogénea.

A estratificação da população em sub-populações é uma técnica muito popular onde se utiliza informação auxiliar na fase de selecção da amostra. Suponhamos que se dispõe de informação auxiliar adequada que permita dividir a população em H sub-populações, ou estratos, mutuamente exclusivos, de dimensões $N_1, ..., N_h, \ldots, N_H$. Seja n a dimensão escolhida para a amostra. A amostragem aleatória estratificada consiste em seleccionar, de forma independente, uma amostra em cada estrato. As amostras em cada estrato têm dimensão pré-fixada: $n_1, ..., n_h, \ldots, n_H$, tal que $\sum n_h = n$. Em cada estrato poderá ser utilizada uma técnica de sondagem diferente. No entanto, é usual utilizar o mesmo tipo de sondagem em todos os estratos.

Existem várias razões que levam à adopção de um plano de sondagem estratificada:

- quando efectuada correctamente, assegura-se, não só a obtenção de estimativas centradas dos parâmetros da população, mas também dos parâmetros de interesse em subgrupos da população;
- se os estratos forem homogéneos, sendo a dimensão da amostra fixa (n), a amostragem estratificada pode fornecer estimativas mais precisas do que a sondagem aleatória simples, para os parâmetros de interesse na população;
- o custo por observação da sondagem pode ser reduzido (por exemplo, quando a estratificação é do tipo geográfico);
- para um determinado custo de implementação pode haver um aumento de precisão;
- é simples de implementar quando comparada com outras técnicas de sondagens alternativas.

O delineamento dos estratos depende, geralmente, de diversos factores:

- da variabilidade das características de interesse na população – ou seja, se forem identificadas sub-populações com maiores ou menores valores médios e maior ou menor variabilidade, relativamente a outras sub-populações;
- dos objectivos do estudo – ou seja, se há interesse em determinar estimativas para cada sub-população;
- da facilidade de implementação – ou seja, se há factores que facilitem a gestão do esforço de amostragem pela utilização de mais de uma amostra (por exemplo: questões geográficas, logísticas ou custos).

Consoante os objectivos da sondagem, a informação auxiliar disponível e as características da população, esta pode ser estratificada relativamente a mais do que uma característica.

Dos factores apresentados decorre, naturalmente, que os critérios de estratificação são subjectivos e devem ser definidos pelo estatístico em cooperação com os especialistas do problema em estudo. Como Hansen, Hurwitz e Madow (1953a, p. 229) referem:

> "*Statistical theory does not provide a general series of procedures or steps for determining the one best set of strata. It does provide some guiding principles and gives a method for comparing and choosing among alternatives*".

Ao subdividir a população em estratos deve-se procurar maximizar a precisão do estimador do parâmetro de interesse, tentando-se garantir que os estratos sejam o mais homogéneos possível. Como veremos posteriormente, os estratos devem ser determinados por forma a que:

- as médias dos estratos sejam o mais dessemelhantes possível, e
- a variância de cada estrato seja a menor possível.

O problema da definição dos estratos e respectivas dimensões amostrais não se encontra no âmbito desta tese pelo que se optou por apresentar apenas algumas considerações gerais. No entanto, não podemos deixar de referir dois casos particulares notáveis: a **amostragem estratificada proporcional**, sugerida por Bowley (1926), que consiste em determinar as dimensões dos estratos, n_h, por forma

a que as taxas de sondagem sejam constantes ($f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$) em todas as sub-populações; e a **amostragem estratificada óptima**[1] atribuída a Neyman (1934) que consiste em determinar os n_h de modo a que a variância do estimador seja mínima, para um determinado custo total fixo ($C = c_0 + nc$, onde *c* é o custo associado a cada elemento da amostra). Para mais detalhes sobre critérios de estratificação e sobre estes métodos de repartição da amostra, em particular, veja-se Hansen, Hurwitz e Madow (1953a), Cochran (1977), Barnett (1991) e Hedayat e Sinha (1991).

Na secção 2.5.1 introduz-se a notação referente à população e à amostra que será utilizada na análise dos estimadores da sondagem aleatória estratificada (secção 2.5.2). Na secção 2.5.3 faz-se uma breve discussão da escolha deste plano de amostragem em alternativa a um plano de sondagem aleatória simples e comparam-se os respectivos estimadores quanto à eficiência. Na secção 2.5.4 consideram-se algumas situações que podem conduzir a problemas na estimação.

2.5.1 Relações e notação

2.5.1.1 Notação referente à população

U	população de dimensão finita
H	número de estratos
U_h	sub-população correspondente ao estrato *h*, $h = 1, \ldots, H$
N_h	número de elementos no estrato *h*, $h = 1, \ldots, H$
$N = \sum_{h=1}^{H} N_h$	dimensão da população

[1] Cochran (1977, p. 99) refere que, após se ter vulgarizado a designação "*repartição óptima de Neyman*", foi descoberta uma demonstração desse método em Tschuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2**, 461-493, 646-683.

y_{hi} valor assumido pelo elemento *i* (i = 1, …, N_h), pertencente ao estrato *h* (h = 1, …, H) para a variável *Y*

Totais, médias e variâncias:

$\tau_h = \sum_{i=1}^{N_h} y_{hi}$ total da variável *Y* no estrato *h*, h = 1, …, H

$\tau = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}$ total da variável *Y*

$\tau = \sum_{h=1}^{H} \tau_h$

$\mu_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}$ média da variável *Y* no estrato *h*, h = 1, …, H

$\mu_h = \frac{\tau_h}{N_h}$

$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}$ média da variável *Y*

$\mu = \sum_{h=1}^{N_h} \frac{N_h}{N}\mu_h$

$\mu = \frac{\tau}{N}$

$\sigma_h^2 = \sum_{i=1}^{N_h} \frac{(y_{hi} - \mu_h)^2}{N_h}$ variância da variável *Y* no estrato *h*, h = 1, …, H

$\sigma^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2$ variância da variável *Y*

$\sigma^2 = \sigma^2_{intra} + \sigma^2_{inter}$

$\sigma^2_{intra} = \sum_h \frac{N_h}{N}\sigma^2_h$ *variância intra-estratos*

$\sigma^2_{inter} = \sum_h \frac{N_h}{N}(\mu_h-\mu)^2$ *variância inter-estratos*

$S^2_h = \sum_{i=1}^{N_h}\frac{(y_{hi}-\mu_h)^2}{N_h-1}$ variância corrigida da variável *Y* no estrato *h*, h = 1, …, H

$S^2_h = \frac{N_h}{N_h-1}\sigma^2_h$

$S^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2$ variância corrigida da variável *Y*

$S^2 = \frac{N}{N-1}\sigma^2$

$S^2 = \sum_h \frac{N_h-1}{N-1}S^2_h + \sum_h \frac{N_h}{N-1}(\mu_h-\mu)^2$

$S^2 \approx S^2_{intra} + \sigma^2_{inter}$

$S^2_{intra} = \sum_h \frac{N_h}{N}S^2_h$ *variância corrigida intra-estratos*

2.5.1.2 Notação referente à amostra

s conjunto de unidades da amostra

s_h	conjunto de unidades da amostra pertencentes ao estrato *h*, h = 1, …, H
n_h	número de elementos do estrato *h*, h = 1, …, H
$n = \sum_{h=1}^{N_h} n_h$	dimensão da amostra
$f = \frac{n}{N}$	taxa de sondagem
$f_h = \frac{n_h}{N_h}$	taxa de sondagem do estrato *h*, h = 1, …, H

Totais, médias e variâncias:

$t_h = \sum_{i \in s_h} y_{hi}$	total amostral da variável *Y* no estrato *h*, h = 1, …, H
$t = \sum_{h=1}^{H} t_h$	total amostral da variável *Y*
$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$	média amostral da variável *Y* no estrato *h*, h = 1, …, H
$\bar{y}_h = \frac{t_h}{n_h}$	
$\bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h$	média amostral da variável *Y*
$\bar{y} = \frac{t}{n}$	
$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$	variância amostral corrigida da variável *Y* no estrato *h*, h = 1, …, H

$s^2 = \frac{1}{n-1}\sum_{h=1}^{H}\sum_{i\in s_h}(y_{hi} - \bar{y})^2$ variância amostral da variável *Y*

2.5.2 Estimação de μ e τ

Na amostragem estratificada, a forma dos estimadores depende do plano de sondagem utilizado em cada sub-população e corresponde, geralmente, a somas ponderadas de estimadores individuais dos estratos, onde os pesos são, naturalmente, os pesos dos estratos, N_h/N (Lehtonen e Pahkinen, 1996, p. 68). Para que seja possível determinar os pesos dos estratos com precisão, assume-se que as dimensões dos estratos N_h são conhecidas (Barnett, 1991, p. 109).

Denotaremos por $\hat{\mu}_{STR}$ e $\hat{\tau}_{STR}$ os estimadores de μ e τ, respectivamente, para o plano de sondagem estratificada. Se $\hat{\mu}_h$ for um estimador centrado da média do estrato *h* então, o estimador da média da população dado por (2.5.1) é também não enviesado relativamente a μ.

(2.5.1) $$\hat{\mu}_{STR} = \sum_{h=1}^{H}\frac{N_h}{N}\hat{\mu}_h$$

De forma análoga, se $\hat{V}(\hat{\mu}_h)$ for um estimador centrado de $V(\hat{\mu}_h)$ para o plano de sondagem considerado, então $\hat{V}(\hat{\mu}_{STR})$ dado por (2.5.3) será também um estimador não enviesado de $V(\hat{\mu}_{STR})$, uma vez que as amostras são retiradas de forma independente em cada estrato.

(2.5.2) $$V(\hat{\mu}_{STR}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 V(\hat{\mu}_h)$$

(2.5.3) $$\hat{V}(\hat{\mu}_{STR}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \hat{V}(\hat{\mu}_h)$$

O estimador do total da população e as suas propriedades podem ser deduzidas a partir do estimador da média:

$$(2.5.4) \qquad \hat{\tau}_{STR} = N\hat{\mu}_{STR} = \sum_{h=1}^{H} N_h \hat{\mu}_h$$

$$(2.5.5) \qquad V(\hat{\tau}_{STR}) = N^2 V(\hat{\mu}_{STR}) = \sum_{h=1}^{H} N_h{}^2 V(\hat{\mu}_h)$$

$$(2.5.6) \qquad \hat{V}(\hat{\tau}_{STR}) = N^2 \hat{V}(\hat{\mu}_{STR}) = \sum_{h=1}^{H} N_h{}^2 \hat{V}(\hat{\mu}_h)$$

Na secção que se segue, apresentam-se os estimadores, referentes ao plano de sondagem aleatória estratificada, para o caso em que se utiliza um plano de sondagem aleatória simples em cada estrato.

2.5.2.1 Sondagem aleatória simples em cada estrato

Nesta secção apresentam-se os estimadores, referentes ao plano de sondagem aleatória estratificada, para os casos em que se utiliza um plano de sondagem aleatória simples, com e sem reposição, em cada estrato. A situação em que se utiliza a SASSR em cada estrato será analisada mais detalhadamente, uma vez que a dedução dos resultados referentes a tiragens com reposição efectua-se de forma análoga.

Suponhamos que as amostras são retiradas, dentro de cada estrato, através de um plano SASSR. Ou seja, retiram-se as amostras s_h $(h = 1, \ldots, H)$ de cada sub-população independentemente umas das outras, de forma equiprovável e através de tiragens sem reposição.

As probabilidades de inclusão dos indivíduos da população têm agora que ter em consideração os estratos a que estes pertencem. Uma vez que estamos a supor que os elementos de cada estrato *h* são seleccionados para a amostra através de um plano SASSR, pelo resultado (2.3.15) conclui-se que a probabilidade de inclusão de 1ª ordem do indivíduo *i* do estrato *h* é dada por:

(2.5.7) $\pi_{(i,\,h)} = \dfrac{n_h}{N_h}$, $\qquad i \in U_h$; h = 1, …, H

Para as probabilidades de inclusão de 2ª ordem, há que distinguir duas situações: o caso em que ambos os indivíduos pertencem ao mesmo estrato (2.5.8) e o caso em que pertencem a estratos diferentes (2.5.9). Analogamente, pelo resultado (2.3.16), tem-se:

(2.5.8) $\pi_{(i,\,h)(j,\,h)} = \dfrac{n_h(n_h-1)}{N_h(N_h-1)}$, $\qquad i, j \in U_h$, $i \neq j$; h = 1, …, H

(2.5.9) $\pi_{(i,\,h)(j,\,k)} = \dfrac{n_h}{N_h}\dfrac{n_k}{N_k}$, $\qquad i \in U_h$, $j \in U_k$; h, k = 1, …, H , $h \neq k$

O estimador da média da população e as suas propriedades obtêm-se a partir dos resultados apresentados nesta secção, tendo em consideração que os estimadores individuais dos estratos são os estimadores do plano SASSR (c f. secção 2.3.2):

(2.5.10) $\hat{\mu} = \sum_{h=1}^{H} \dfrac{N_h}{N}\bar{y}_h$

(2.5.11) $V(\hat{\mu}) = \sum_{h=1}^{H}\left(\dfrac{N_h}{N}\right)^2 \dfrac{N_h - n_h}{N_h}\dfrac{S_h^2}{n_h}$

(2.5.12) $\hat{V}(\hat{\mu}) = \sum_{h=1}^{H}\left(\dfrac{N_h}{N}\right)^2 \dfrac{N_h - n_h}{N_h}\dfrac{s_h^2}{n_h}$

Uma expressão equivalente a (2.5.10), para o estimador de μ , é

(2.5.13) $\hat{\mu} = \dfrac{1}{N}\sum_{h=1}^{H}\dfrac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} = \dfrac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\dfrac{y_{hi}}{n_h/N_h}$

e, pela probabilidade de inclusão de 1ª ordem (2.5.7), tem-se

(2.5.14) $$\hat{\mu} = \frac{1}{N} \sum_{(i,h) \in s} \frac{y_{hi}}{\pi_{(i,h)}}$$

Pelo que foi exposto relativamente ao estimador da média, obtêm-se os seguintes resultados para o estimador do total da população.

O estimador do total da população, τ, é dado por:

(2.5.15) $$\hat{\tau} = \sum_{h=1}^{H} N_h \bar{y}_h$$

A variância deste estimador é:

(2.5.16) $$V(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

e pode ser estimada por:

(2.5.17) $$\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}$$

Pelo resultado (2.5.14) conclui-se que, para um plano de amostragem aleatória estratificada em se utiliza a SASSR em cada estrato, os estimadores da média (2.5.10) e do total (2.5.15) da população são formalmente equivalentes ao estimador de Horvitz-Thompson (c.f. secção 2.4).

Os quadros seguintes apresentam um resumo dos resultados relativos à utilização dos planos SASCR (c.f. secção 2.3.1) e SASSR em cada estrato, no que se refere aos estimadores de μ (Quadro 2.5.1) e τ (Quadro 2.5.2).

Quadro 2.5.1 – Propriedades do estimador de μ, para a sondagem estratificada

SASCR em cada estrato	SASSR em cada estrato
$\hat{\mu} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$	$\hat{\mu} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$
$E(\hat{\mu}) = \mu$	$E(\hat{\mu}) = \mu$
$V(\hat{\mu}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}$	$V(\hat{\mu}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$
$\hat{V}(\hat{\mu}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}$	$\hat{V}(\hat{\mu}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}$

Quadro 2.5.2 – Propriedades do estimador de τ, para a sondagem estratificada

SASCR em cada estrato	SASSR em cada estrato
$\hat{\tau} = \sum_{h=1}^{H} N_h \bar{y}_h$	$\hat{\tau} = \sum_{h=1}^{H} N_h \bar{y}_h$
$E(\hat{\tau}) = \tau$	$E(\hat{\tau}) = \tau$
$V(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{\sigma_h^2}{n_h}$	$V(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$
$\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}$	$\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}$

2.5.3 Comparação com a sondagem aleatória simples

Como já foi referido, intuitivamente tem-se a percepção de que a sondagem aleatória estratificada é mais apropriada do que a sondagem aleatória simples se os estratos forem homogéneos e o parâmetro de interesse variar entre as diferentes sub-populações. A sondagem aleatória simples será mais adequada quando toda a população é homogénea. Para confirmar esta intuição, vamos começar por analisar a expressão da variância total da população.

O resultado que se segue apresenta a variância total da população como a soma de duas parcelas: a primeira, designa-se por *variância intra-estratos* e corresponde a uma média ponderada da dispersão no interior dos estratos; a segunda, designa-se por *variância inter-estratos* e corresponde a uma média ponderada dos quadrados dos desvios entre a média de cada estrato e a média da população, pelo que pode ser interpretada como uma medida da dispersão entre os estratos.

(2.5.18) $$\sigma^2 = \sigma^2_{\text{int ra}} + \sigma^2_{\text{int er}}$$

onde

(2.5.19) $$\sigma^2_{\text{int ra}} = \sum_h \frac{N_h}{N} \sigma^2_h$$

(2.5.20) $$\sigma^2_{\text{int er}} = \sum_h \frac{N_h}{N} (\mu_h - \mu)^2$$

A demonstração de (2.5.18) encontra-se no Anexo 2, secção A2.2.1. A partir deste resultado vamos analisar duas situações extremas: na primeira, supõe-se que a média da população é constante em todos os estratos (i.e., a população é homogénea) e, na segunda, que a variância é nula dentro de cada estrato.

Se $\mu_h = \mu, \ \forall h$ então, por (2.5.20) tem-se que $\sigma^2_{\text{int er}} = 0$ e, por (2.5.18) conclui-se que $\sigma^2 = \sigma^2_{\text{int ra}}$. Ou seja, a variância entre estratos é nula e, no interior de cada estrato, a variância é σ^2. Nesta situação extrema, verifica-se que para conhecer a população *U*, basta estudar todos os indivíduos de apenas uma das sub-populações U_h (Figura 3).

Figura 3 – *População homogénea*

U_1	U_2	…	U_H
μ	μ	…	μ

Se $\sigma_h^2 = 0$ então, por (2.5.19) tem-se $\sigma_{intra}^2 = 0$ e conclui-se que a variância total da população é, neste caso, $\sigma^2 = \sigma_{inter}^2$. Nesta situação, para conhecer a população *U*, basta estudar um elemento de cada uma das sub-populações U_h (Figura 4).

Figura 4 – *População homogénea em cada estrato*

U_1	U_2	…	U_H
μ_1	μ_2	…	μ_H

Pelo resultado (2.5.18), conclui-se assim que a dispersão *intra-estratos* é tanto menor quanto maior for a dispersão *inter-estratos*, dado que σ^2 é constante. Esta análise vai de encontro à nossa intuição. Interessa agora verificar em que condições a amostragem estratificada fornece estimativas mais precisas do que a sondagem aleatória simples.

2.5.3.1 Precisão relativa da sondagem estratificada e da sondagem aleatória simples

Como já foi referido, quando um plano de sondagem aleatória estratificada é implementado de forma adequada, as estimativas que se obtêm são, na maioria das situações, mais precisas do que as que seriam obtidas por sondagem aleatória simples. No entanto, tal como Cochran (1977, p. 99) também salienta, não é verdade que tal sucede com *qualquer* amostra estratificada. Sob determinadas condições a sondagem aleatória simples poderá ser mais eficiente.

Para analisar esta questão, discute-se em seguida a eficiência dos estimadores de μ para cada um destes planos amostrais, supondo que se pretende retirar uma amostra de dimensão fixa *n* de uma população com *N* elementos.

Seja $\hat{\mu}_{STR}$ o estimador de μ, dado por (2.5.10), num plano de sondagem estratificada no qual se utiliza a SASSR em cada estrato. Pelo resultado (2.5.11) a sua variância é dada por:

(2.5.21) $$V(\hat{\mu}_{STR}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2$$

Para um plano de sondagem aleatória simples sem reposição, o estimador de μ, que denotaremos por $\hat{\mu}_{SAS}$, é a média amostral e, pelo resultado (2.3.20), a sua variância é dada por:

(2.5.22) $$V(\hat{\mu}_{SAS}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

sendo S^2 a variância corrigida da população.

Utilizando as relações apresentadas na secção 2.5.1.1 e o resultado (2.5.18), tem-se:

$$(N-1)S^2 = N\sigma_{intra}^2 + N\sigma_{inter}^2$$

$$\Leftrightarrow (N-1)S^2 = \sum_{h=1}^{H} N_h\sigma_h^2 + \sum_{h=1}^{H} N_h(\mu_h-\mu)^2$$

(2.5.23) $$\Leftrightarrow (N-1)S^2 = \sum_{h=1}^{H}(N_h-1)S_h^2 + \sum_{h=1}^{H} N_h(\mu_h-\mu)^2$$

Supondo que N é suficientemente grande tal que $N/(N-1) \approx 1$ e que as dimensões dos estratos N_h são, ou suficientemente grandes, ou quase constantes, tal que $N_h/(N_h-1) \approx N/(N-1)$, a expressão de S^2 apresentada em (2.5.23) pode ser aproximada por:

(2.5.24) $$S^2 \approx \sum_{h=1}^{H}\frac{N_h}{N}S_h^2 + \sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2$$

Note-se que a segunda parcela deste resultado corresponde à variância *inter-estratos*. Assim, substituindo (2.5.24) em (2.5.22) e simplificando a expressão correspondente à diferença entre as variâncias dos estimadores $\hat{\mu}_{SAS}$ e $\hat{\mu}_{STR}$, obtém-se:

$$(2.5.25)\qquad V(\hat{\mu}_{SAS}) - V(\hat{\mu}_{STR}) \approx \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{H}\frac{N_h}{N}S_h^2 + \left(\frac{1}{n}-\frac{1}{N}\right)\sigma_{inter}^2 - \\ - \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{S_h^2}{n_h} + \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2$$

e, simplificando este resultado, conclui-se:

$$(2.5.26)\qquad V(\hat{\mu}_{SAS}) - V(\hat{\mu}_{STR}) \approx \left(\frac{1}{n}-\frac{1}{N}\right)\sigma_{inter}^2 + \sum_{h=1}^{H}\frac{N_h}{N^2}S_h^2\left(\frac{N}{n}-\frac{N_h}{n_h}\right)$$

Do resultado (2.5.26), é imediata a observação de que o primeiro termo é sempre não negativo e toma o valor zero se e só se $\sigma_{inter}^2 = 0$. O segundo termo também poderá ser não negativo se a dimensão dos estratos da amostra, n_h, for determinada de forma adequada. Por exemplo, se $n_h = n\frac{N_h}{N}$ para h = 1, …, H, a segunda parcela anula-se. Isto é, a já referida **amostragem proporcional** poderá conduzir a um estimador pelo menos tão eficiente como o estimador de uma sondagem aleatória simples sem reposição.

Neste caso, as probabilidades de inclusão de 1ª ordem reduzem-se a

$$(2.5.27)\qquad \pi_i = n/N\ ,\ \ i = 1, \ldots, N$$

e portanto, da expressão geral do estimador de Horvitz-Thompson para μ (2.5.14), conclui-se que o <u>estimador de μ para a amostragem proporcional</u> não é mais do que a média amostral:

$$(2.5.28)\qquad \hat{\mu}_{prop} = \bar{y}$$

A variância deste estimador obtém-se substituindo $n_h = n\frac{N_h}{N}$ em (2.5.11):

(2.5.29) $$V(\hat{\mu}_{prop}) = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2$$

Note-se que, no caso da amostragem proporcional, as taxas de sondagem são iguais em todos os estratos, i.e., $f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$.

Assim, utilizando as aproximações já consideradas, $N/(N-1) \approx 1$ e $N_h/(N_h - 1) \approx N/(N-1)$, conclui-se que

(2.5.30) $$V(\hat{\mu}_{SAS}) - V(\hat{\mu}_{prop}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_{inter}^2$$

Este resultado indica, aproximadamente, o ganho de precisão que se obtém com a utilização da sondagem estratificada com repartição proporcional dos estratos, relativamente à sondagem aleatória simples sem reposição, numa amostra de dimensão fixa *n*. Naturalmente, quanto maior for a dispersão entre os estratos, maior será o ganho de precisão da amostragem proporcional.

As comparações efectuadas entre $V(\hat{\mu}_{SAS})$ e $V(\hat{\mu}_{prop})$, baseadas no resultado (2.5.30), podem não ser válidas para populações pequenas, ou seja, a sondagem aleatória simples poderá ser mais eficiente do que a sondagem estratificada com repartição proporcional dos estratos, embora na prática tal seja difícil de ocorrer (veja-se Cochran 1977).

Para uma discussão mais detalhada da precisão dos estimadores da amostragem proporcional, da amostragem óptima e da sondagem aleatória simples veja-se Hansen, Hurwitz e Madow (1953a, 1953b), Cochran (1977) e Barnettt (1991).

Da análise apresentada, conclui-se que os critérios de estratificação são fundamentais para que haja um aumento de precisão significativo relativamente à sondagem aleatória simples. Como Barnett (1991, p. 137) refere:

> "*Population characteristics can be more efficiently estimated from a stratified sample than from an overall s.r.* [simple random] *sample* if strata means differ widely, and within-strata variation is low. *The greater this*

effect the greater the efficiency of the corresponding estimators. With freedom of choice of strata, the aim should be to construct strata with these characteristics."

2.5.4 Eventuais problemas na estimação

Para que seja possível obter estimativas da variância dos estimadores a partir da amostra, é necessário que esta tenha dimensão superior ou igual a dois em todos os estratos, ou seja, que $n_h \geq 2$ (h = 1, …, H).

Hansen, Hurwitz e Madow (1953a, p. 438) afirmam que quando existe apenas uma unidade na amostra de um determinado estrato, é necessário agrupar os estratos para que seja possível estimar a variância a partir da amostra. Algumas considerações sobre esse processo e respectiva descrição podem ser obtidas em Hansen, Hurwitz e Madow (1953a; Ch. 9, Sec. 15; Ch. 10, Sec. 13). Hartley, Rao e Kiefer (1969) apresentam também uma solução para o problema da estimação da variância quando n_h=1 para algum *h*.

Cochran (1977, p. 138-140) apresenta dois métodos que permitem lidar com o caso extremo de existir, em todos os estratos, apenas um elemento na amostra. Este autor apresenta uma técnica para agrupar os estratos que designa por "*method of collapsing strata*" e cita uma outra abordagem desenvolvida por Fuller (1970).

2.5.4.1 Efeitos de erros nas dimensões dos estratos

Os ganhos de precisão das estimativas obtidas por estratificação dependem do grau de homogeneidade que é atingido no interior de cada estrato que, por sua vez, depende do grau de precisão com que os estratos foram definidos (Hansen, Hurwitz e Madow, 1953a, p. 183). Ora, os totais dos estratos N_h podem não ser conhecidos de forma exacta por estarem desactualizados, por exemplo, devido a informação incorrecta na base de sondagem.

Cochran (1977, p. 117) apresenta, em termos gerais, as consequências da utilização de pesos dos estratos, N_h/N, incorrectos:

1. Os estimadores são enviesados.

2. O enviesamento permanece constante à medida que a dimensão da amostra aumenta. Consequentemente, as estimativas obtidas são menos precisas do que as obtidas por amostragem aleatória simples, perdendo-se todo o ganho de precisão da estratificação.
3. A variância do estimador subestima o verdadeiro valor do erro, uma vez que não contém a contribuição do enviesamento para o erro.

Stephan (1945, citado por Hansen, Hurwitz e Madow 1953a, p. 233) fornece algumas considerações adicionais para este problema.

2.6 Estimação da variância pelo método de linearização de Taylor

Quando se consideram estimadores não lineares, é muitas vezes impossível obter as expressões exactas do enviesamento e da variância. Uma das técnicas mais utilizadas para contornar este problema é o *método de linearização de Taylor* ou *método-δ*. Esta técnica é utilizada há longa data (pelo menos desde o tempo de Gauss) em vários campos da estatística. A aplicação dos princípios básicos do método de linearização de Taylor a estimadores não lineares, definidos sob planos de sondagem complexos, deve-se a Keyfitz (1957) e Tepping (1968). Outra referência fundamental da sua aplicação no âmbito da teoria das sondagens é Woodruff (1971).

O método de linearização de Taylor consiste em utilizar a aproximação ao primeiro termo da expansão em série de Taylor, da função que define o estimador não linear $\hat{\theta}$. Esta técnica permite então obter uma expressão aproximada para a variância de $\hat{\theta}$, bem como, sob determinadas condições, um estimador dessa variância.

Na secção 2.6.1, relembram-se os resultados relativos ao estimador de Horvitz-Thompson (apresentado na secção 2.4), introduzindo-se alguma notação adicional, e considera-se o estimador de Horvitz-Thompson para um vector de totais de várias variáveis de estudo. Na secção 2.6.2 apresenta-se, em termos genéricos, o método de linearização de Taylor. A exposição desta técnica segue de perto a apresentação efectuada por Särndal, Swensson e Wretman (1992, p. 172-176). Os métodos de estimação da variância por *Bootstrap* são considerados resumidamente na secção 2.7. A aplicação destes métodos será exemplificada em secções posteriores.

2.6.1 Estimadores de Horvitz-Thompson para várias variáveis de estudo

Em seguida, relembram-se os resultados relativos ao estimador de Horvitz-Thompson (veja-se a secção 2.4), utilizando-se alguma notação adicional que se passa a apresentar.

No que se segue, e em posteriores secções, denotar-se-á por

(2.6.1) $$\sum\sum_{U} a_{ij}$$

a expressão:

(2.6.2) $$\sum_{i\in U}\sum_{j\in U} a_{ij} = \sum_{i\in U} a_{ii} + \sum_{i\in U}\sum_{\substack{j\in U \\ j\neq i}} a_{ij}$$

Relembre-se ainda a notação considerada na secção 2.2.5:

(2.6.3) $$\Delta_{ij} = \pi_{ij} - \pi_i\pi_j \,, \qquad i\neq j, \qquad i,j\in U$$
$$\Delta_{ii} = \pi_i(1-\pi_i)$$

que representa a covariância entre duas variáveis de Cornfield, sendo π_i e π_{ij} as probabilidades de inclusão de 1ª e 2ª ordem, respectivamente, associadas a um determinado plano de sondagem.

Esta notação permite resumir os resultados referentes ao estimador de Horvitz-Thompson do total da variável *Y* na população (veja-se o Quadro 2.4.1), que se denotará por $\hat{\tau}_\pi$, como se apresenta no Quadro 2.6.1.

Quadro 2.6.1 - Propriedades do estimador de Horvitz-Thompson para τ

Estimação do total da variável *Y* na população
$\hat{\tau}_\pi = \sum_{i\in s} \frac{y_i}{\pi_i}$
$E(\hat{\tau}_\pi) = \tau$
$V_1(\hat{\tau}_\pi) = \sum\sum_{U} \Delta_{ij} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$
$\hat{V}_1(\hat{\tau}_\pi) = \sum\sum_{s} \frac{\Delta_{ij}}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$
$E[\hat{V}_1(\hat{\tau}_\pi)] = V_1(\hat{\tau}_\pi)$ se $\pi_{ij}>0$, $\forall\ i,j\in U,\ i\neq j$

Quadro 2.6.1 (*continuação*) **- Propriedades do estimador de Horvitz-Thompson para τ**

Estimação do total da variável *Y* na população
Se *s* tiver dimensão fixa:
$V_2(\hat{\tau}_\pi) = -\frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U \\ j\neq i}}\Delta_{ij}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$
$\hat{V}_2(\hat{\tau}_\pi) = -\frac{1}{2}\sum_{i\in S}\sum_{\substack{j\in S \\ j\neq i}}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$
$E[\hat{V}_2(\hat{\tau}_\pi)] = V_2(\hat{\tau}_\pi)$ se $\pi_{ij}>0, \forall\ i,j\in U, i\neq j$

Suponhamos agora que se pretende estimar as G componentes de um vector **t** composto pelos totais, desconhecidos na população, das variáveis de estudo $Y_1, \ldots, Y_g, \ldots, y_G$:

(2.6.4) $\mathbf{t} = (\tau_1, \ldots, \tau_g, \ldots, \tau_G)^T$

onde,

(2.6.5) $\tau_g = \sum_U y_{gi}$, $g = 1, \ldots, G$

e $y_{g1}, \ldots, y_{gi}, \ldots, y_{gN}$ são os N valores da variável de estudo Y_g ($g = 1, \ldots, G$) na população.

Seja *s* uma amostra aleatória, obtida através de um determinado plano de sondagem *p(s)*, e sejam π_i e π_{ij} as probabilidades de inclusão de 1ª e 2ª ordem, respectivamente, associadas a *p(s)*. Supondo que se pode observar em *s* o vector,

(2.6.6) $\mathbf{y_i} = (y_{1i}, \ldots, y_{gi}, \ldots, y_{Gi})^T$, $i\in s$

os estimadores de Horvitz-Thompson para cada um dos totais desconhecidos das variáveis $Y_1, \ldots, Y_g, \ldots, y_G$ (ou seja, $\tau_1, \ldots, \tau_g, \ldots, \tau_G$) são:

(2.6.7) $\hat{\tau}_{g\pi} = \sum_s \frac{y_{gi}}{\pi_i}$, g = 1, …, G

Obtém-se então o vector de estimadores:

(2.6.8) $\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}} = (\hat{\tau}_{1\pi}, \ldots, \hat{\tau}_{g\pi}, \ldots, \hat{\tau}_{G\pi})^T$

É imediato que:

(2.6.9) $E(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}) = \mathbf{t}$

uma vez que $\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}$ é um vector de estimadores centrados.

A matriz de variâncias-covariâncias associada a $\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}$,

(2.6.10) $V(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}) = E\{(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}} - \mathbf{t})(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}} - \mathbf{t})^T\}$

é, obviamente, uma matriz simétrica tal que o *g-ésimo* elemento da diagonal principal corresponde à variância de $\hat{\tau}_{g\pi}$ que é dada por (veja-se o Quadro 2.6.1):

(2.6.11) $V(\hat{\tau}_{g\pi}) = \sum\sum_U \Delta_{ij} \frac{y_{gi}}{\pi_i} \frac{y_{gj}}{\pi_j}$, g = 1, …, G

Naturalmente, também pelos resultados referentes ao estimador de Horvitz-Thompson, os elementos da diagonal principal da matriz (2.6.10) podem ser estimados sem enviesamento por:

(2.6.12) $\hat{V}(\hat{\tau}_{g\pi}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_{gi}}{\pi_i} \frac{y_{gj}}{\pi_j}$, g = 1, …, G

O elemento *gg'* que se encontra fora da diagonal principal (i.e., g≠g'), corresponde à covariância entre $\hat{\tau}_{g\pi}$ e $\hat{\tau}_{g'\pi}$. A sua expressão obtém-se utilizando as variáveis de Cornfield:

(2.6.13) $\quad \mathrm{Cov}(\hat{\tau}_{g\pi},\ \hat{\tau}_{g'\pi}) = \mathrm{Cov}\left[\sum_U \frac{y_{gi}}{\pi_i}\mathbb{I}_{i\in s},\ \sum_U \frac{y_{g'i}}{\pi_i}\mathbb{I}_{i\in s}\right]$

Pelas propriedades da covariância e utilizando a notação apresentada anteriormente, conclui-se que

(2.6.14) $\quad \mathrm{Cov}(\hat{\tau}_{g\pi},\ \hat{\tau}_{g'\pi}) = \sum\sum_U \mathrm{Cov}\,(\mathbb{I}_{i\in s},\ \mathbb{I}_{j\in s})\frac{y_{gi}}{\pi_i}\frac{y_{g'j}}{\pi_j}$

ou seja,

(2.6.15) $\quad \mathrm{Cov}(\hat{\tau}_{g\pi},\ \hat{\tau}_{g'\pi}) = \sum\sum_U \Delta_{ij}\frac{y_{gi}}{\pi_i}\frac{y_{g'j}}{\pi_j}$, $\quad g, g' = 1, \ldots, G,\ g\neq g'$

Assim, analogamente aos resultados apresentados para o estimador de Horvitz-Thompson, verifica-se que o elemento *gg'* (g≠g') da matriz de variâncias-covariâncias de $\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}$, dada por (2.6.10), pode ser estimado sem enviesamento através de:

(2.6.16) $\quad \hat{\mathrm{Cov}}\,(\hat{\tau}_{g\pi},\ \hat{\tau}_{g'\pi}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}}\frac{y_{gi}}{\pi_i}\frac{y_{g'j}}{\pi_j}$, $\quad g, g' = 1, \ldots, G,\ g\neq g'$

O Quadro 2.6.2 apresenta um resumo dos resultados referentes à estimação do vector **t**, composto pelos totais das variáveis de estudo Y_1, …, Y_g, …, y_G, quando se considera o estimador de Horvitz-Thompson de cada um desses totais.

Os resultados apresentados nesta secção são fundamentais para a exposição do método de linearização de Taylor. Esta técnica é apresentada, em termos genéricos, na secção seguinte e, em posteriores secções, será aplicada a diversos estimadores não lineares.

Quadro 2.6.2 – Propriedades do estimador de Horvitz-Thompson para o vector de totais $\mathbf{t} = (\tau_1, \ldots, \tau_g, \ldots, \tau_G)^T$

Estimação do total das variáveis $Y_1, \ldots, Y_g, \ldots, Y_G$ na população

$\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}} = (\hat{\tau}_{1\pi}, \ldots, \hat{\tau}_{g\pi}, \ldots, \hat{\tau}_{G\pi})^T$, $\quad \hat{\tau}_{g\pi} = \sum_s \frac{y_{gi}}{\pi_i}$, g = 1, …, G

$E(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}) = \mathbf{t}$

$V(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}})$ – matriz de variâncias-covariâncias associada a $\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}}$

- elementos *gg* da diagonal principal:

$$V(\hat{\tau}_{g\pi}) = \sum\sum_U \Delta_{ij} \frac{y_{gi}}{\pi_i} \frac{y_{gj}}{\pi_j} \quad , \qquad g = 1, \ldots, G$$

- elementos gg' fora da diagonal principal:

$$Cov(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi}) = \sum\sum_U \Delta_{ij} \frac{y_{gi}}{\pi_i} \frac{y_{g'j}}{\pi_j} \quad , \quad g, g' = 1, \ldots, G, \; g \neq g'$$

$\hat{V}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}})$ – estimador centrado da matriz $V(\hat{\boldsymbol{\tau}}_{\boldsymbol{\pi}})$

- estimador dos elementos *gg* da diagonal principal:

$$\hat{V}(\hat{\tau}_{g\pi}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_{gi}}{\pi_i} \frac{y_{gj}}{\pi_j} \quad , \qquad g = 1, \ldots, G$$

- estimador dos elementos gg' fora da diagonal principal:

$$Côv(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_{gi}}{\pi_i} \frac{y_{g'j}}{\pi_j} \quad , \quad g, g' = 1, \ldots, G, \; g \neq g'$$

2.6.2 Método de linearização de Taylor

Suponhamos que se pretende estimar um parâmetro θ que pode ser expresso como uma função de *G* totais da população:

(2.6.17) $\quad \theta = f(\tau_1, \ldots, \tau_g, \ldots, \tau_G)$, $\quad \tau_g = \sum_U y_{gi}$, g = 1, …, G

onde $y_{1i}, \ldots, y_{gi}, \ldots, y_{Gi}$ são, respectivamente, os valores correspondentes ao *i-ésimo* elemento das variáveis $Y_1, \ldots, Y_g, \ldots, Y_G$ na população.

Cada total desconhecido pode ser estimado sem enviesamento pelo respectivo estimador de Horvitz-Thompson:

$$(2.6.18) \qquad \hat{\tau}_{g\pi} = \sum_s \frac{y_{gi}}{\pi_i} \ , \quad g = 1, \ldots, G$$

quando se pode observar na amostra o vector $(y_{1i}, \ldots, y_{gi}, \ldots, y_{Gi})^T$ para todos os elementos $i \in s$.

Um estimador de θ é então:

$$(2.6.19) \qquad \hat{\theta} = f(\hat{\tau}_{1\pi}, \ldots, \hat{\tau}_{g\pi}, \ldots, \hat{\tau}_{G\pi})$$

Se *f* for uma função linear, as propriedades de $\hat{\theta}$ deduzem-se facilmente a partir dos resultados apresentados na secção precedente. Neste caso, o estimador (2.6.19) é centrado e pode-se escrever na forma:

$$(2.6.20) \qquad \hat{\theta} = a_0 + \sum_{g=1}^{G} a_g \hat{\tau}_{g\pi}$$

onde, $\hat{\tau}_{g\pi}$ é o estimador de Horvitz-Thompson de τ_g (g = 1, …, G) dado por (2.6.18).

A variância do estimador (2.6.20) é

$$(2.6.21) \qquad V(\hat{\theta}) = \sum_{g=1}^{G} \sum_{g'=1}^{G} a_g a_{g'} Cov\left(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi}\right)$$

que pode ser estimada sem enviesamento por

$$(2.6.22) \qquad \hat{V}(\hat{\theta}) = \sum_{g=1}^{G} \sum_{g'=1}^{G} a_g a_{g'} \hat{Cov}\left(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi}\right)$$

onde, as covariâncias $\text{Cov}(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi})$ são obtidas através do resultado (2.6.15) e os respectivos estimadores obtêm-se através de (2.6.16); quando *g=g'*, obviamente, $\hat{\text{Cov}}(\hat{\tau}_{g\pi}, \hat{\tau}_{g'\pi})$ corresponde ao estimador $\hat{V}(\hat{\tau}_{g\pi})$ apresentado em (2.6.12).

Neste caso, em que *f* é uma função linear, estes resultados podem ser apresentados através de expressões alternativas, mais simples de implementar. Ou seja, uma vez que o estimador (2.6.20) pode ser escrito como:

$$(2.6.23) \qquad \hat{\theta} = a_0 + \sum_s \frac{u_i}{\pi_i}$$

onde,

$$(2.6.24) \qquad u_i = \sum_{g=1}^{G} a_g y_{gi} \quad , \quad i \in s$$

A variância (2.6.21) pode então ser escrita como

$$(2.6.25) \qquad V(\hat{\theta}) = \sum\sum_U \Delta_{ij} \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}$$

e o respectivo estimador, apresentado em (2.6.22), vem:

$$(2.6.26) \qquad \hat{V}(\hat{\theta}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}$$

Caso o parâmetro $\theta = f(\tau_1, \ldots, \tau_g, \ldots, \tau_G)$ tenha uma expressão não linear, é muitas vezes impossível obter as expressões do enviesamento e da variância do estimador (2.6.19). Uma das formas de contornar este problema é a utilização do método de linearização de Taylor, que consiste em determinar um *pseudo-estimador*[1] $\hat{\theta}_0$,

[1] $\hat{\theta}_0$ depende, geralmente, de certas quantidades desconhecidas pelo que nem sempre será um verdadeiro estimador.

através da aproximação da função *f* à 1ª ordem da expansão em série de Taylor, em torno do ponto $(\tau_1, \ldots, \tau_g, \ldots, \tau_G)$, ignorando-se os restantes termos.

Neste caso, como $\hat{\theta}_0$ é uma função linear, determina-se facilmente $V(\hat{\theta}_0)$ como uma aproximação de $V(\hat{\theta})$, bem como um estimador de $V(\hat{\theta}_0)$.

O método de linearização de Taylor consiste, então, em determinar as propriedades de $\hat{\theta}_0$ que se obtém da seguinte forma:

$$(2.6.27) \qquad \hat{\theta} \doteq \hat{\theta}_0 = \theta + \sum_{g=1}^{G} a_g(\hat{\tau}_{g\pi} - \tau_g)$$

com

$$(2.6.28) \qquad a_g = \left.\frac{\partial f}{\partial \hat{\tau}_{g\pi}}\right|_{(\hat{\tau}_{1\pi},\ldots,\hat{\tau}_{G\pi})=(\tau_1,\ldots,\tau_G)}, \quad g = 1, \ldots, G$$

quando as derivadas parciais dadas por a_g existem e não são conjuntamente nulas.

Para amostras grandes[1], o estimador $\hat{\theta}$ comporta-se aproximadamente como $\hat{\theta}_0$. No que se segue, assume-se que o enviesamento e a variância de $\hat{\theta}$ podem ser aproximados pelas correspondentes expressões referentes à estatística linear $\hat{\theta}_0$.

Utilizando as expressões alternativas para estimadores lineares, apresentadas em (2.6.23) e (2.6.24), conclui-se que $\hat{\theta}_0$ pode ser escrito como:

$$(2.6.29) \qquad \hat{\theta}_0 = \left[\theta - \sum_{g=1}^{G} a_g \tau_g\right] + \sum_s \frac{u_i}{\pi_i}$$

onde,

---

[1] Ou seja, quando $\hat{\tau}_{1\pi},\ldots,\hat{\tau}_{G\pi}$ tomam valores próximos de $\tau_1, \ldots, \tau_G$, respectivamente, com grande probabilidade (Särndal, Swensson e Wretman, 1992, p. 174).

(2.6.30) $$u_i = \sum_{g=1}^{G} a_g y_{gi} \quad , \quad i \in s$$

e as expressões dos a_g (g = 1, …, G) são dadas pelas equações (2.6.28).

Denote-se por $AV(\hat{\theta})$ a *variância aproximada* de $\hat{\theta}$ e que corresponde à variância exacta de $\hat{\theta}_0$, isto é,

(2.6.31) $$AV(\hat{\theta}) = V(\hat{\theta}_0)$$

Note-se que as equações (2.6.21) e (2.6.25) são equivalentes, pelo que a variância de $\hat{\theta}_0$ pode ser obtida através de

(2.6.32) $$V(\hat{\theta}_0) = V\left(\sum_{g=1}^{G} a_g \hat{\tau}_{g\pi} \right) = V\left(\sum_s \frac{u_i}{\pi_i} \right)$$

e conclui-se, portanto, que a *variância aproximada*[1] de $\hat{\theta}$ é

(2.6.33) $$AV(\hat{\theta}) = V(\hat{\theta}_0) = \sum\sum_U \Delta_{ij} \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}$$

O estimador de (2.6.33) não se obtém de forma imediata, uma vez que os valores dos a_g (g = 1, …, G) dependem de totais desconhecidos na população e, portanto, as quantidades u_i ($i \in s$) são também desconhecidas. Para contornar este problema, é usual substituir-se cada um desses totais pelos respectivos estimadores de Horvitz-Thompson. Obtêm-se, desta forma, estimadores $\hat{a}_g$ de a_g que permitem calcular, para qualquer $i \in s$, a variável:

(2.6.34) $$\hat{u}_i = \sum_{g=1}^{G} \hat{a}_g y_{gi} \quad , \quad i \in s$$

[1] Observe-se que, uma vez que $E(\hat{\theta}_0) = \theta$ se conclui que $EQM(\hat{\theta}) \doteq EQM(\hat{\theta}_0) = V(\hat{\theta}_0)$.

Obtém-se então o seguinte estimador de $V(\hat{\theta}_0)$:

(2.6.35) $$\hat{V}(\hat{\theta}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{u}_i}{\pi_i} \frac{\hat{u}_j}{\pi_j}$$

onde, $\hat{u}_i$ ($i \in s$) é dado por (2.6.34).

Relativamente a esta expressão, Särndal, Swensson e Wretman (1992, p. 175) referem que:

> "*The justification for the procedure is that* $\hat{u}_k$*, being a (possibly nonlinear) function of* π *estimators, is consistent for* u_k*. Now,* $\hat{V}(\hat{\theta})$ *is a function of the consistent estimators* $\hat{u}_k$ *and should in large samples behave as if it had been based on the true (unknown)* u_k*. Thus,* $\hat{V}(\hat{\theta})$ *can be assumed to be consistent for* $V(\hat{\theta})$*.*"

A expressão (2.6.35) corresponde, estritamente falando, a um estimador de $AV(\hat{\theta})$ e não de $V(\hat{\theta})$. No entanto, uma vez que para amostras grandes a variância aproximada $AV(\hat{\theta})$ "concorda bem" com $V(\hat{\theta})$, pode-se considerar que, para essas amostras, (2.6.35) é um bom estimador de $V(\hat{\theta})$. Särndal, Swensson e Wretman (1992, p. 175) referem que tal já foi demonstrado em diversos estudos por simulação.

Quando o desenho da amostra é de dimensão fixa, um estimador da variância alternativo (veja-se, por exemplo, o Quadro 2.6.1) é dado por:

(2.6.36) $$\hat{V}(\hat{\theta}) = -\frac{1}{2} \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \frac{\Delta_{ij}}{\pi_{ij}} \left(\frac{\hat{u}_i}{\pi_i} - \frac{\hat{u}_j}{\pi_j} \right)^2$$

2.7 Estimação da variância por métodos de Bootstrap

Alternativamente aos estimadores da variância que se obtêm pelo método de linearização de Taylor, podem-se considerar outros métodos de estimação baseados em técnicas de re-amostragem (*resampling*). Estes métodos baseiam-se na ideia de que a amostra obtida é representativa da população alvo, podendo extrair-se novas e repetidas amostras a partir da amostra original, com o objectivo de estimar variâncias ou intervalos de confiança. Alguns exemplos destes métodos são o *Jackknife*, proposto por Quenouille (1949), e o *Bootstrap*, introduzido por Efron (1979). Referências bibliográficas relevantes sobre as extensões destes métodos a dados de sondagens podem ser encontradas em Shao e Tu (1995).

Referências fundamentais sobre os métodos de Bootstrap são Efron (1979, 1982) e Efron e Tibshinari (1993). Quando se pretende lidar com o problema da não resposta através de métodos de imputação, a estimação da variância dos estimadores requer algum cuidado. Neste contexto, é de salientar a extensão dos métodos de Bootstrap proposta por Shao e Sitter (1996).

Os métodos de Bootstrap permitem obter estimativas válidas para a variância de estimadores lineares e não lineares, ao contrário dos métodos de Jackknife. Por este motivo, e também devido à simplicidade de implementação, vamos nos restringir aos métodos de Bootstrap e, em particular, às suas aplicações aos planos de sondagem aleatória simples sem reposição (secção 2.7.2) e sondagem aleatória estratificada sem reposição (secção 2.7.3). Na secção que se segue, apresenta-se uma breve introdução à metodologia Bootstrap.

2.7.1 Introdução ao Bootstrap

Efron e Tibshinari (1993, p. 45-49) apresentam as ideias base subjacentes ao método de Bootstrap, no âmbito da inferência clássica da estatística, como se segue.

Seja $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ uma amostra aleatória obtida a partir de uma população com função de distribuição (desconhecida) *F*. Seja $\hat{F}$ a função de distribuição empírica associada à amostra obtida, tal que a cada valor observado x_i (i=1, …, n), atribui a probabilidade 1/n.

Uma ***amostra bootstrap*** é uma amostra $\mathbf{x^*} = (x_1^*, x_2^*, \ldots, x_n^*)$ obtida de forma aleatória e com reposição a partir da amostra inicial $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, também designada ***população bootstrap***. Suponhamos que são seleccionadas *B* amostras bootstrap que se denotam por $\mathbf{x^{*1}}$, $\mathbf{x^{*2}}$, …, $\mathbf{x^{*B}}$.

Para cada amostra bootstrap $\mathbf{x^*}$, calcula-se o estimador pretendido (i.e., calcula-se uma ***réplica bootstrap do estimador***):

(2.7.1) $\hat{\theta}^* = g(\mathbf{x^*})$

onde, *g(.)* é a função que define o estimador $\hat{\theta}$ de θ.

A estimativa bootstrap da variância do estimador obtém-se através da variância dos valores $g(\mathbf{x^{*1}}), \ldots, g(\mathbf{x^{*B}})$, que denotaremos por $\hat{V}_B^*(\hat{\theta})$.

A metodologia apresentada pode ser implementada utilizando-se o algoritmo apresentado por Efron e Tibshinari (1993, p. 47):

1º – Seleccionar B amostras bootstrap independentes, $\mathbf{x^{*1}}, \ldots, \mathbf{x^{*b}}, \ldots, \mathbf{x^{*B}}$, onde $\mathbf{x^{*b}} = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$, b = 1, …, B, de forma aleatória e com reposição a partir da população bootstrap $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

2º – Calcular a réplica bootstrap do estimador para cada uma das amostras bootstrap:

(2.7.2) $\hat{\theta}^{*b} = g(\mathbf{x^{*b}})$, b = 1, …, B

3º – Calcular a estimativa bootstrap da variância do estimador, através da variância dos valores obtidos no 2º passo:

(2.7.3) $$\hat{V}_B^*(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\theta}^{*b} - \hat{\theta}^*(\cdot)\right]^2$$

onde,

(2.7.4) $$\hat{\theta}^{*}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$$

Efron e Tibshinari (1993, p. 47) referem que o limite da variância bootstrap, quando *B* tende para infinito, é a estimativa ideal da variância do estimador de θ, i.e. de $V(\hat{\theta})$:

(2.7.5) $$\lim_{B\to\infty} \hat{V}_B^{*}(\hat{\theta}) = V^{*}(\hat{\theta}^{*})$$

onde, V* denota a variância relativamente à amostra bootstrap.

Aqueles autores referem ainda que, na prática, o número de réplicas bootstrap *B* varia entre 25 e 200 quando se pretende estimar a variância. No entanto, quando se pretende construir intervalos de confiança o número de réplicas deverá ser superior a 200.

A aplicação da metodologia Bootstrap a dados de sondagens requer algumas alterações ao algoritmo apresentado, como se verá posteriormente. Nas secções que se seguem apresenta-se de forma abreviada a aplicação desta metodologia à SASSR e à sondagem aleatória estratificada.

2.7.2 Sondagem aleatória simples sem reposição

Seja $(y_1, y_2, \ldots, y_n)$ uma amostra aleatória obtida através de um plano SASSR de uma população com N elementos. Suponhamos que o parâmetro de interesse é a média da população, $\theta = \mu$, e se utiliza como estimador de μ a média amostral:

(2.7.6) $$\hat{\theta} = g(y_1, y_2, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Seja $(y_1^{*}, y_2^{*}, \ldots, y_n^{*})$ a amostra bootstrap que se obtém, a partir de $(y_1, y_2, \ldots, y_n)$, através de *n* tiragens aleatórias com reposição e probabilidades iguais. Denote-se por $\hat{\theta}^{*} = \bar{y}^{*} = g(y_1^{*}, y_2^{*}, \ldots, y_n^{*})$ a réplica bootstrap da média amostral, ou seja,

(2.7.7) $$\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^*$$

Deville (1987, p. 161) apresenta as propriedades de (2.7.7):

(2.7.8) $$\hat{V}^*(\bar{y}^*) = \frac{1}{n}\frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)$$

Este resultado permite concluir que

(2.7.9) $$E[\hat{V}^*(\bar{y}^*)] = \left(1-\frac{1}{n}\right)\frac{S^2}{n}$$

onde,

(2.7.10) $$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

Relembre-se que, para este plano de sondagem, se tem

(2.7.11) $$V(\bar{y}) = (1 - f)\frac{S^2}{n}$$

onde, f = n/N e S^2 é dado por (2.7.10).

Verifica-se então que (2.7.9) é igual a (2.7.11) apenas quando f = 1/n, ou seja, quando n = $\sqrt{N}$. Conclui-se, assim, que $\hat{V}^*(\bar{y}^*)$ não é um estimador consistente de $V(\bar{y})$.

Para ultrapassar este problema, é necessário alterar o algoritmo “puro” (também designado na literatura por bootstrap *naïf* ou *naive*) apresentado na secção anterior. Os primeiros trabalhos nesta área devem-se a Gross (1980) e a Chao e Lo (1985). O método proposto por estes autores para a SASSR designa-se, geralmente, na

literatura por ***Bootstrap Without Replacement*** ou ***Without Replacement Bootstrap*** (***BWO***). Bickel e Freedman (1984) e Sitter (1992b) propuseram uma extensão do método à sondagem aleatória estratificada.

O algoritmo BWO para a SASSR consiste em construir uma pseudo-população (i.e., uma população artificial) de dimensão N, replicando-se *k* vezes a amostra original ($N = k \times n$) e em retirar amostras aleatórias sem reposição, de dimensão n, a partir dessa população. Estas amostras são, então, as amostras bootstrap a partir das quais se obtêm as estimativas pretendidas, de forma análoga ao algoritmo "puro" (para mais detalhes veja-se, por exemplo, Deville 1987).

Para planos de sondagem complexos a extensão deste método não é imediata, como se verá para o caso da sondagem aleatória estratificada, na secção seguinte.

2.7.3 Sondagem aleatória estratificada

Para a sondagem aleatória estratificada (SASSR no estratos) foram propostos três algoritmos que apresentam estimativas bootstrap, da variância de estimadores lineares, correctas: o *Without Replacement Bootstrap* (*BWO*) (Bickel e Freedman, 1984; Sitter, 1992b); o *Rescaling Bootstrap* (*RB*) (Rao e Wu, 1988) e o *Mirror-Match Bootstrap* (*MMB*) (Sitter, 1992a). Veja-se também Chen e Sitter (1993).

Qualquer um destes métodos constitui uma metodologia válida para se obterem estimadores bootstrap da variância, de estimadores lineares, para o plano de sondagem em apreço. No entanto, o método BWO proposto por (Sitter, 1992b) parece ser o mais promissor quando se considera o caso não linear. Por este motivo, apresenta-se apenas esse método.

O método BWO proposto por Bickel e Freedman (1984) é bastante limitado pelo que, como refere Sitter (1992b), nem sempre é aplicável. Apresenta-se, então, apenas a extensão do método BWO proposta por Sitter (1992b) para a sondagem aleatória estratificada[1].

O **algoritmo BWO proposto por Sitter (1992b)** é o seguinte:

[1] Sitter (1992b) apresenta ainda extensões do método BWO a outros planos de sondagem mais complexos.

1º – Construir de forma independente para cada estrato *h* (h = 1, …, H) uma pseudo-população, replicando-se os elementos da amostra original desse estrato k_h vezes, sendo

(2.7.12) $$k_h = \frac{N_h}{n_h}\left(1-\frac{1-f_h}{n_h}\right) , \qquad h = 1, \ldots, H$$

(2.7.13) $$f_h = n_h/N_h , \qquad h = 1, \ldots, H$$

2º – Seleccionar n'_h unidades de cada estrato *h* através de tiragens aleatórias sem reposição, sendo

(2.7.14) $$n'_h = n_h - (1 - f_h) , \qquad h = 1, \ldots, H$$

por forma a obter-se uma amostra bootstrap. A partir da amostra bootstrap, calcular a réplica bootstrap do estimador, $\hat{\theta}^*$.

3º – Repetir o 2º passo um grande número de vezes, *B*, por forma a obterem-se $\hat{\theta}^{*1}$, …, $\hat{\theta}^{*b}$, …, $\hat{\theta}^{*B}$ réplicas bootstrap do estimador.

4º – Estimar $V(\hat{\theta})$ por

(2.7.15) $$\hat{V}^*_{BWO} = E^*[\hat{\theta}^* - E^*(\hat{\theta}^*)]^2$$

onde, E* denota o valor esperado relativamente às amostras bootstrap; ou, pela aproximação de Monte Carlo

(2.7.16) $$\hat{V}^*_{BWO} \approx \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^*(\cdot)\right)^2$$

onde,

(2.7.17) $$\hat{\theta}^{*}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$$

Note-se que este método pressupõe que k_h e n'_h sejam inteiros. No entanto, tal não ocorre a menos que f_h seja igual a “0” (zero) ou igual a 1. Sitter (1992b) sugere um procedimento que designa por *randomization between bracketing integers* para contornar este problema e refere as propriedades do estimador bootstrap, subjacente a este algoritmo, para o caso linear. Neste caso, o método conduz a estimadores bootstrap consistentes. No caso não linear, o método parece ser promissor, como verificou Sitter (1992b) através de estudos por simulação. As propriedades teóricas dos estimadores bootstrap requerem ainda investigação, quando se consideram estimadores e planos de sondagem complexos.

3 ESTIMAÇÃO NA PRESENÇA DE ERROS NÃO AMOSTRAIS

3.1 Introdução

Uma prática comum, no tratamento e análise de dados provenientes de sondagens probabilísticas, consiste em associar a cada observação da amostra um **peso** ou **ponderador**. Suponhamos que foi retirada uma determinada amostra, de acordo com um certo desenho ou plano de sondagem.

Como referem Koeijers e Willeboordse (1995, p. 87) os objectivos da ponderação são:

1. Extrapolar a amostra para a população (i.e, inferir).
2. Lidar com a não-resposta.
3. Aumentar a precisão através da utilização de informação auxiliar.
4. Obter consistência com os dados de outras fontes.

No que se segue, faremos uma distinção entre **ponderar** e **reponderar** (*weighting* e *reweighting*). De um modo geral, a ponderação está associada ao primeiro objectivo e depende do plano de amostragem escolhido, pelo que pode ser efectuada antes da recolha dos dados. O termo peso ou ponderador refere-se geralmente ao coeficiente de extrapolação, ou seja, ao inverso da probabilidade de inclusão (c.f. secção 2.4.3). A reponderação é efectuada depois de os dados terem sido recolhidos e prende-se com os restantes objectivos.

Kalton e Kasprzyk (1986, p. 4, citados por Lundström e Särndal 1999, p. 307) descrevem claramente este processo:

> ”*A common approach is initially to determine the sample weights needed to compensate for unequal selection probabilities, next to revise these weights to compensate for unequal response rates in different sample weighting classes* (…)*, and finally to revise the weights again to make the weighted sample distribution for certain characteristics (e.g., age/sex) conform to the known population distribution for those characteristics.*”

De um modo geral, o processo de reponderação baseia-se em informação auxiliar que se encontre presente na base de sondagem durante a fase de estimação ou em informação proveniente de outras fontes. Existe abundante literatura sobre métodos de estimação que utilizam informação auxiliar; veja-se por exemplo: Wright (1983); Rao (1988); Skinner, Holt e Smith (1989); Särndal, Swensson e Wretman (1992); Thompson, (1992); Hedlin *et al.* (1998); Lundström e Särndal (1999). Algumas das técnicas mais utilizadas são os métodos de *pós-estratificação*, *estimação pelo quociente* e *estimação pela regressão*.

Ao longo deste capítulo, procuraremos apresentar alguns estimadores que, face à informação auxiliar disponível durante a fase de estimação, permitam melhorar as estimativas obtidas por ponderação. A apresentação destes métodos é essencialmente motivada pela possível existência certos erros na base de sondagem e ocorrência de não respostas no inquérito, como veremos nas secções 3.2 e 3.6, respectivamente.

Embora fosse extremamente interessante analisar e testar estimadores pela regressão, estes encontram-se fora do âmbito da dissertação. Pelo mesmo motivo, a estimação pelo quociente e a estimação em domínios são aqui apresentadas apenas de uma forma abreviada. Estes métodos serão introduzidos essencialmente com o intuito de enquadrar teoricamente alguns conceitos utilizados na análise dos estimadores de pós-estratificação.

Na secção 3.3 serão introduzidos os métodos básicos de estimação pelo quociente. Na secção 3.4 apresentaremos os conceitos elementares da estimação em domínios. Como veremos posteriormente, estes métodos permitem também solucionar certos tipos de imperfeições da base sondagem. Na secção 3.5 analisaremos mais detalhadamente alguns estimadores de pós-estratificação, considerando-se a ausência de não respostas e, na secção 3.6, estes métodos são apresentados como uma das formas de lidar com a ocorrência de não respostas e com os problemas das bases de sondagem.

3.2 Estimação na presença de erros na base de sondagem

Durante a fase de desenho da amostra, é fundamental definir a população alvo ou universo de referência. No que diz respeito às especificações da população alvo, ocorrem por vezes algumas imperfeições. É de salientar que os resultados do inquérito podem ficar comprometidos se o conjunto de elementos do universo de referência for definido ambiguamente.

A base de amostragem, ou base de sondagem, sendo uma lista actualizada de todos os elementos da população alvo deveria, idealmente, permitir identificar a população alvo na totalidade. A constituição e actualização de uma base de amostragem é, no entanto, um processo complexo e difícil de realizar. Por um lado, as características e a composição da população alteram-se constantemente e, por outro lado, as fontes de informação sobre essas alterações, ainda que sejam acessíveis, são muitas vezes imperfeitas, tanto no que se refere à exactidão como à actualização.

Tipicamente, não existem ficheiros que possam garantir uma representação completa, perfeita e actualizada da população alvo. Lessler e Kalsbeek (1992) efectuam uma discussão detalhada sobre seis imperfeições comuns às bases de sondagem, das quais vamos salientar quatro, particularmente relevantes para a fase de estimação:

1. Subcobertura (não inclusão de unidades da população alvo).
2. Sobrecobertura (inclusão de unidades que não pertencem à população).
3. Registos duplicados ou múltiplos.
4. Informação auxiliar incorrecta (dimensão, actividade, etc.).

A possível existência de erros na base de amostragem tem repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram (veja-se Lessler e Kalsbeek 1992 para mais detalhes). De entre os quatro problemas apontados anteriormente, a subcobertura é talvez o problema mais sério, uma vez que não é possível detectá-lo, nem a partir da base de amostragem, nem a partir da amostra. Dado que uma parte da população não pode ser observada, a amostra obtida conduz a estimativas enviesadas.

Um problema óbvio de sobrecobertura corresponde à situação em que a base de sondagem contém unidades mortas. Designam-se por unidades mortas os elementos da população que, de alguma forma, deixaram de existir (por exemplo, cessação da actividade de empresas, falecimento de indivíduos, etc.). É usual as bases de sondagem utilizadas em inquéritos a empresas conterem mais de 10% de unidades mortas porque as fontes externas não as conseguem identificar (Koeijers e Willeboordse, 1995, p. 44).

A utilização de informação auxiliar incorrecta reduz a precisão das estimativas da sondagem (Lessler e Kalsbeek, 1992, p. 77). Este tipo de erros pode conduzir a sobrecobertura. O inverso poderá também ocorrer, ou seja, serem excluídos elementos erradamente, conduzindo a um problema de subcobertura. Se os efeitos de sobrecobertura e subcobertura tendem, ou não, a anular-se depende das variáveis de estudo (Koeijers e Willeboordse, 1995, p. 90).

Designam-se por **domínios** as sub-populações para as quais são necessárias estimativas pontuais separadas. Um domínio pode ser qualquer sub-população ou toda a população. Lessler e Kalsbeek (1992, p. 69) e Särndal, Swensson e Wretman (1992, p. 543) apontam os métodos de estimação em domínios como uma solução para os problemas de sobrecobertura, uma vez que, neste caso, a população alvo é um domínio da base de amostragem. Estes métodos podem ser aplicados quando, após o inquérito ter sido efectuado, se verifica a existência de elementos que não pertencem à população. Se for possível reconhecer esses elementos, excluindo-os da amostra, obtém-se uma amostra *s'* de dimensão aleatória. Nestas condições, o estimador de Horvitz-Thompson é centrado para as estimativas do total da população obtidas a partir de *s'* (para mais detalhes, veja-se a secção 3.4).

Relativamente à subcobertura, Särndal, Swensson e Wretman (1992, p. 544) sugerem a escolha de um ajustamento pelo quociente como forma de reduzir o enviesamento do estimador de Horvitz-Thompson provocado por este problema.

Uma forma de minimizar os problemas de sobrecobertura e subcobertura, originados por deficiente informação na base de sondagem, é a utilização de métodos de pós-estratificação, estimação pelo quociente e estimação pela regressão. Estas técnicas de reponderação têm por objectivo melhorar as estimativas obtidas, podendo utilizar no momento da estimação, informação auxiliar mais actualizada.

Por outro lado, não podemos deixar de referir que a base de amostragem não deve ter apenas um tipo de imperfeição, tal como sucede na maioria das sondagens, e não é possível fornecer uma solução geral que resolva todos os seus problemas simultaneamente (Koeijers e Willeboordse, 1995, p. 90).

Nunca será demais lembrar que é preferível prevenir eventuais problemas, procurando que haja o menor número possível de erros em todas as etapas da implementação da sondagem, do que procurar soluções *a posteriori*[1].

3.2.1 O problema das mudanças de estrato

Suponhamos que o desenho da sondagem corresponde a um plano de amostragem aleatória estratificada. Os estratos, sendo homogéneos na sua constituição, podem conter elementos com comportamentos muito diferenciados. Por exemplo, nos inquéritos às empresas, é comum utilizar-se o tipo de actividade como variável de estratificação, mas verifica-se por vezes que as respostas obtidas no inquérito sugerem que determinadas empresas não se mantêm nos estratos iniciais.

As mudanças de estrato são resultado de uma deficiente informação no ficheiro de base. Ou seja, a informação auxiliar que consta da base de sondagem e que permitiu efectuar a estratificação da população, encontra-se de alguma forma incorrecta ou desactualizada. Nesta situação deixa de existir uma correspondência exacta entre os estratos na base de sondagem e na população.

Por outro lado, quando a base de sondagem possui informação incorrecta, os totais dos estratos na população podem não ser conhecidos de forma exacta. Na secção 2.5.4 apresentámos, em termos gerais, as consequências da utilização de ponderadores incorrectos.

Como vimos anteriormente, o problema de sobrecobertura pode ser tratado, na fase de estimação, através de métodos de estimação em domínios. O problema das mudanças de estrato pode também ser tratado desta forma. Por exemplo, quando se pretende estimar o total de empresas por actividade económica e se observaram

[1] Como refere Coelho (1995, p. 158-159): "*a habitual tentação de definir e testar os estimadores num momento posterior à definição dos planos de amostragem poderá comprometer o processo, impedindo a aproximação a planos óptimos, ou mesmo dando origem a incompatibilidades que mais tarde não podem ser evitadas e que acabam por remeter para piores soluções.*"

mudanças de actividade (ou seja, de estrato), as estimativas podem ser obtidas por domínios, uma vez que cada actividade económica define um domínio (sub-população). Como já foi referido, essas estimativas são centradas; no entanto, a variância dessas estimativas é geralmente maior, do que a variância das estimativas que seriam obtidas se todas as empresas tivessem sido correctamente estratificadas, quando a base de sondagem foi construída (Lessler e Kalsbeek, 1992, p. 77).

Embora os eventuais procedimentos que permitam evitar a ocorrência de mudanças de estrato, bem como efectuar uma manutenção adequada da base de sondagem, se encontrem fora do âmbito deste trabalho, não podemos deixar de referir alguns trabalhos desenvolvidos nesta área ou que abordem esta questão.

Koop (1988) apresenta uma técnica que procura controlar e reduzir os erros de não amostragem, entre os quais os problemas de cobertura da base de sondagem e erros derivados da não resposta. Särndal, Swensson e Wretman (1992) apresentam algumas questões ligadas à manutenção da base de sondagem. Rivest (1999) propõe, para a fase de desenho da sondagem, alguns algoritmos de estratificação que têm em consideração a possível ocorrência de mudanças de estrato.

3.3 Métodos básicos de estimação pelo quociente

Como se referiu anteriormente, uma forma de tratar os problemas da base de sondagem e, de um modo geral, aumentar a precisão das estimativas das variáveis de interesse, consiste em utilizar informação auxiliar que se encontre presente na base de sondagem durante a fase de estimação ou informação proveniente de outras fontes. Uma das técnicas mais utilizadas, para atingir estes objectivos, são os métodos de estimação pelo quociente.

Estes procedimentos têm sido investigados, pelo menos, desde a década de 30 (Rao, 1988), existindo, assim, abundante literatura sobre este tema. Nesta secção, apresentam-se resumidamente alguns métodos de estimação pelo quociente. Uma referência mais detalhada sobre estes métodos pode ser encontrada em Cochran (1977), Rao (1988) e Särndal, Swensson e Wretman (1992).

Em seguida considera-se o caso em que o parâmetro de interesse corresponde ao rácio de duas quantidades desconhecidas na população e, na secção 3.3.2, o estimador pelo quociente usual.

3.3.1 Estimação de um quociente

Sejam *Y* e *X* duas variáveis quantitativas. Suponhamos que se pretende estimar o rácio entre a média (total) da variável *Y* e a média (total) de *X* na população U, de dimensão N. Ou seja, a quantidade que se pretende estimar é:

$$(3.3.1) \qquad R = \frac{\mu_y}{\mu_x} = \frac{\tau_y}{\tau_x}$$

Se os dois totais desconhecidos forem estimados por $\hat{\tau}_y$ e $\hat{\tau}_x$, respectivamente, então o estimador de R é dado por:

$$(3.3.2) \qquad \hat{R} = \frac{\hat{\mu}_y}{\hat{\mu}_x} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$$

O estimador $\hat{R}$ não é linear e, de um modo geral, não é centrado, ainda que os estimadores de τ_y e τ_x o sejam.

Sejam $\hat{\mu}_y$ e $\hat{\mu}_x$ dois estimadores centrados de μ_y e μ_x, respectivamente. Nestas condições, o enviesamento de $\hat{R}$ é dado por:

$$(3.3.3) \qquad B(\hat{R}) = \frac{-\mathrm{Cov}(\hat{R},\hat{\mu}_x)}{\mu_x}$$

Este resultado foi demonstrado por Hartley e Ross (1954) para o caso de um plano de sondagem aleatória simples sem reposição. Para um plano de sondagem arbitrário, a demonstração de (3.3.3) é bastante simples, como se pode verificar em seguida:

Uma vez que a expressão (3.3.2) se pode escrever como $\hat{R}\,\hat{\mu}_x = \hat{\mu}_y$, tem-se:

$$(3.3.4) \qquad \mathrm{Cov}(\hat{R}\,,\,\hat{\mu}_x) = E[\hat{R}\,\hat{\mu}_x] - E[\hat{R}]\,E[\hat{\mu}_x] = E[\hat{\mu}_y] - E[\hat{R}]\,E[\hat{\mu}_x]$$

E, como por hipótese, $\hat{\mu}_y$ e $\hat{\mu}_x$ são estimadores centrados, obtém-se:

$$(3.3.5) \qquad \mathrm{Cov}(\hat{R}\,,\,\hat{\mu}_x) = \mu_y - E[\hat{R}]\,\mu_x$$

Por outro lado, a expressão do parâmetro de interesse R, dada por (3.3.1), também se pode escrever na forma $\mu_y = R\mu_x$, pelo que se obtém:

$$(3.3.6) \qquad \mathrm{Cov}(\hat{R}\,,\,\hat{\mu}_x) = R\mu_x - E[\hat{R}]\,\mu_x = \mu_x\{R - E[\hat{R}]\}$$

Sendo o enviesamento de $\hat{R}$ dado por $B(\hat{R}) = E[\hat{R}] - R$, conclui-se que:

$$(3.3.7) \qquad \mathrm{Cov}(\hat{R}\,,\,\hat{\mu}_x) = -\mu_x\,B(\hat{R})$$

Obtendo-se então o resultado (3.3.3).

Analogamente, tem-se também:

(3.3.8) $$B(\hat{R}) = \frac{-\mathrm{Cov}(\hat{R},\hat{\tau}_X)}{\tau_X}$$

Pelo resultado (3.3.3) e dado que o quadrado do coeficiente de correlação (ρ) é sempre inferior ou igual a 1, obtém-se a seguinte desigualdade:

(3.3.9) $$\frac{\left[B(\hat{R})\right]^2}{V(\hat{R})} \leq \frac{V(\hat{\mu}_X)}{\mu_X{}^2}$$

como se passa a demonstrar.

Elevando ambos os membros da equação (3.3.3) ao quadrado, tem-se:

(3.3.10) $$[B(\hat{R})]^2 = \frac{\left[\mathrm{Cov}(\hat{R},\hat{\mu}_X)\right]^2}{\mu_X{}^2}$$

Uma vez que o quadrado do coeficiente de correlação é dado por:

(3.3.11) $$\rho^2 = \frac{\left[\mathrm{Cov}(\hat{R},\hat{\mu}_X)\right]^2}{V(\hat{R})V(\hat{\mu}_X)}$$

a expressão (3.3.10) pode ser escrita de forma equivalente como:

$$[B(\hat{R})]^2 = \frac{\rho^2 V(\hat{R})V(\hat{\mu}_X)}{\mu_X{}^2}$$

(3.3.12) $$\Leftrightarrow \frac{\left[B(\hat{R})\right]^2}{V(\hat{R})} = \rho^2 \frac{V(\hat{\mu}_X)}{\mu_X{}^2}$$

E, uma vez que $\rho^2 \leq 1$, conclui-se o resultado (3.3.9).

Analogamente se verifica que:

(3.3.13) $$\frac{\left[B(\hat{R})\right]^2}{V(\hat{R})} \leq \frac{V(\hat{\tau}_X)}{\tau_X{}^2}$$

A desigualdade (3.3.9), ou (3.3.13), permite tirar algumas conclusões interessantes relativamente à validade dos intervalos de confiança que se obtêm através do estimador $\hat{R}$. Relembre-se que a quantidade (*bias ratio*) definida por (2.2.22):

$$(3.3.14) \qquad BR(\hat{R}) = \frac{B(\hat{R})}{\sqrt{V(\hat{R})}}$$

é fundamental para a obtenção de intervalos de confiança válidos, quando os estimadores são enviesados (c.f. secção 2.2.3).

A desigualdade (3.3.9) permite então concluir que:

$$(3.3.15) \qquad [BR(\hat{R})]^2 \leq \frac{V(\hat{\mu}_x)}{\mu_x{}^2}$$

Ou seja, se $\sqrt{V(\hat{\mu}_x)} \,/\, |\mu_x|$ se aproximar de zero, quando a dimensão da amostra aumenta, então $BR(\hat{R})$ também se aproximará de zero. Como já foi referido, esta é uma condição importante para a validade dos intervalos de confiança.

Os totais de *Y* e *X* são usualmente estimados pelos respectivos estimadores de Horvitz-Thompson. Ou seja, o ***estimador usual do quociente*** $R=\tau_y/\tau_x$ é

$$(3.3.16) \qquad \hat{R} = \frac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}}$$

O método de linearização de Taylor, apresentado na secção 2.6, permite obter a *variância aproximada* do estimador (3.3.16) e um estimador dessa variância. Torna-se então possível determinar intervalos de confiança para o parâmetro não linear R, quando as amostras são grandes.

Uma vez que $\hat{R}$ é uma função (não linear) de $\hat{\tau}_{y\pi}$ e $\hat{\tau}_{x\pi}$, o primeiro passo do método consiste em determinar as derivadas parciais:

(3.3.17) $$\frac{\partial \hat{R}}{\partial \hat{\tau}_{y\pi}} = \frac{1}{\hat{\tau}_{x\pi}}$$

(3.3.18) $$\frac{\partial \hat{R}}{\partial \hat{\tau}_{x\pi}} = -\frac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}^{2}}$$

Calculando estas derivadas no ponto (τ_y, τ_x) obtêm-se as expressões de a_1 e a_2 dadas por (2.6.28):

(3.3.19) $$a_1 = \frac{1}{\tau_x}$$

$$a_2 = -\frac{\tau_y}{\tau_x^{2}} = -\frac{R}{\tau_x}$$

Para amostras grandes, $\hat{R}$ comporta-se aproximadamente como a estatística linear (veja-se o resultado (2.6.29)):

(3.3.20) $$\hat{R}_0 = [R - (a_1\tau_y + a_2\tau_x)] + \sum_s \frac{u_i}{\pi_i} = R + \sum_s \frac{u_i}{\pi_i}$$

onde, as expressões de u_i ($i \in s$) são dadas por (2.6.30):

(3.3.21) $$u_i = a_1 y_i + a_2 x_i \,, \qquad i \in s$$

ou seja,

(3.3.22) $$u_i = \frac{1}{\tau_x}(y_i - R x_i) \,, \qquad i \in s$$

Substituindo (3.3.22) em (3.3.20), conclui-se que

(3.3.23) $$\hat{R} \doteq \hat{R}_0 = R + \frac{1}{\tau_x}\sum_s \frac{y_i - R x_i}{\pi_i}$$

Sob a aproximação (3.3.23), $\hat{R}$ é *aproximadamente não enviesado* e, por (2.6.33), a *variância aproximada* de $\hat{R}$ é dada por:

$$(3.3.24) \qquad AV(\hat{R}) = V(\hat{R}_0) = \frac{1}{\tau_x^2} \sum\sum_U \Delta_{ij} \frac{y_i - Rx_i}{\pi_i} \frac{y_j - Rx_j}{\pi_j}$$

Uma vez que a_1 e a_2 podem ser estimados, respectivamente, por:

$$(3.3.25) \qquad \hat{a}_1 = \frac{1}{\hat{\tau}_{x\pi}}$$

$$\hat{a}_2 = -\frac{\hat{R}}{\hat{\tau}_{x\pi}}$$

pelo resultado (2.6.35), obtém-se o estimador:

$$(3.3.26) \qquad \hat{V}(\hat{R}) = \frac{1}{\hat{\tau}_{x\pi}^2} \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{R}x_i}{\pi_i} \frac{y_j - \hat{R}x_j}{\pi_j}$$

Note-se que, sob a aproximação (3.3.23), o enviesamento de $\hat{R}$, embora não seja nulo, aproxima-se de zero, isto é,

$$(3.3.27) \qquad E(\hat{R}) \doteq E(\hat{R}_0) = R$$

Assim, tem-se que

$$(3.3.28) \qquad EQM(\hat{R}) \doteq EQM(\hat{R}_0) = V(\hat{R}_0)$$

Mas, quando a aproximação não é boa (por exemplo, nos casos em que a amostra é pequena), o método de linearização de Taylor tem tendência a subestimar o erro quadrático médio do estimador. Nestes casos, deve-se procurar obter uma aproximação melhor, incluindo-se os termos de 2ª ordem da expansão em série de Taylor.

No Quadro 3.3.1 apresenta-se um resumo das propriedades do estimador usual do quociente (3.3.16).

Quadro 3.3.1 – Propriedades do estimador usual do quociente $R = \tau_y/\tau_x$

Estimação do quociente entre os totais de *Y* e *X* na população
$\hat{R} = \frac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}} = \frac{\sum_s \frac{y_i}{\pi_i}}{\sum_s \frac{x_i}{\pi_i}}$
$AV(\hat{R}) = \frac{1}{\tau_x^2} \sum\sum_U \Delta_{ij} \frac{y_i - Rx_i}{\pi_i} \frac{y_j - Rx_j}{\pi_j}$
$\hat{V}(\hat{R}) = \frac{1}{\hat{\tau}_{x\pi}^2} \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{R}x_i}{\pi_i} \frac{y_j - \hat{R}x_j}{\pi_j}$

Interessa agora analisar o caso particular de $\hat{R}$ em que se considera $\tau_x = N$ que, como se verá, tornará mais clara a forma de alguns estimadores apresentados em posteriores secções.

3.3.1.1 Caso particular: estimação da média da população

Suponhamos que se pretende estimar a média de uma variável *Y* na população (μ_y). Atendendo a que $\mu_y = \frac{\tau_y}{N}$, pode-se utilizar o estimador do quociente (3.3.2) fazendo $\tau_x = N$. Ou seja, supõe-se que a variável *X* toma o valor 1 para todos os elementos da população.

Um estimador do quociente para μ_y é dado por:

(3.3.29) $$\hat{\mu}_y = \frac{\hat{\tau}_y}{\hat{N}}$$

Sendo π_i a probabilidade de inclusão de 1ª ordem do indivíduo *i*, podem-se considerar os seguintes estimadores de Horvitz-Thompson:

(3.3.30) $$\hat{\tau}_{y\pi} = \sum_s \frac{y_i}{\pi_i}$$

(3.3.31) $$\hat{N} = \sum_s \frac{1}{\pi_i}$$

Obtém-se, desta forma, um estimador do quociente para μ_y, também designado na literatura por "***weighted sample mean***":

(3.3.32) $$\hat{\mu}_{yw} = \frac{1}{\sum_s \frac{1}{\pi_i}} \sum_s \frac{y_i}{\pi_i}$$

Observe-se que este estimador pode ser sempre utilizado, quer a dimensão da população, *N*, seja conhecida, ou não. Se a dimensão da população for conhecida e se o desenho da amostra corresponder, por exemplo, a um plano de sondagem aleatória simples sem reposição (SASSR) ou a um plano de sondagem aleatória estratificada (em que se utiliza a SASSR em cada estrato), este estimador é formalmente idêntico ao estimador usual da média (estimador de Horvitz-Thompson); numa sondagem com probabilidades desiguais, isto poderá não suceder. Nestes casos, pode-se optar entre o estimador de Horvitz-Thompson e o estimador do quociente (3.3.32). No entanto, a opção recai geralmente sobre este último, dado que este é muitas vezes o melhor estimador. Särndal, Swensson e Wretman (1992, p. 182-183) apresentam algumas situações que vão de encontro a esta intuição, uma vez que não é possível indicar condições exactas para que tal se verifique.

Relativamente às propriedades do estimador (3.3.32), pelos resultados genéricos apresentados anteriormente, (3.3.24) e (3.3.26), é imediato que a *variância aproximada* de $\hat{\mu}_{yw}$ e o respectivo estimador são:

(3.3.33) $$AV(\hat{\mu}_{yw}) = \frac{1}{N^2} \sum\sum_U \Delta_{ij} \frac{y_i - \mu_y}{\pi_i} \frac{y_j - \mu_y}{\pi_j}$$

(3.3.34) $$\hat{V}(\hat{\mu}_{yw}) = \frac{1}{\hat{N}^2} \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_{yw}}{\pi_i} \frac{y_j - \hat{\mu}_{yw}}{\pi_j}$$

No Quadro 3.3.2 apresenta-se um resumo das propriedades do estimador "*weighted sample mean*" de μ.

Quadro 3.3.2 - Propriedades do estimador "weighted sample mean" de μ

Estimação da média da variável *Y* na população
$\hat{\mu}_{yw} = \frac{\hat{\tau}_{y\pi}}{\hat{N}} = \frac{\sum_s \frac{y_i}{\pi_i}}{\sum_s \frac{1}{\pi_i}}$
$AV(\hat{\mu}_{yw}) = \frac{1}{N^2}\sum\sum_U \Delta_{ij}\frac{y_i-\mu_y}{\pi_i}\frac{y_j-\mu_y}{\pi_j}$
$\hat{V}(\hat{\mu}_{yw}) = \frac{1}{\hat{N}^2}\sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}}\frac{y_i-\hat{\mu}_{yw}}{\pi_i}\frac{y_j-\hat{\mu}_{yw}}{\pi_j}$

3.3.2 Estimação pelo quociente, na presença de informação auxiliar

Seja *Y* uma variável de interesse e considere-se uma variável auxiliar *X* da qual se conhece o seu total τ_x na população U (de dimensão *N*) e, portanto, também a sua média na população. Esta informação é utilizada nos estimadores pelo quociente do total e da média da variável de interesse, τ_y e μ_y respectivamente.

Observe-se que o total da população (e de forma análoga a média da população) pode ser escrito como:

$$\tau_y = \tau_x \frac{\tau_y}{\tau_x} = \tau_x R \qquad (3.3.35)$$

Assim, sendo $\hat{\tau}_y$ e $\hat{\tau}_x$ os usuais estimadores de Horvitz-Thompson, R pode ser estimado por $\hat{R} = \frac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}}$ e obtém-se então o **estimador pelo quociente usual de τ_y**:

$$\hat{\tau}_{yQ} = \tau_x \hat{R} = \tau_x \frac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}} \qquad (3.3.36)$$

De forma análoga, o **estimador pelo quociente usual de μ_y** é dado por:

$$(3.3.37) \qquad \hat{\mu}_{yQ} = \mu_x \frac{\hat{\mu}_{y\pi}}{\hat{\mu}_{x\pi}}$$

Särndal, Swensson e Wretman (1992) referem que o total da variável auxiliar *X* deve ser conhecido com exactidão, uma vez que um valor impreciso de τ_x pode conduzir a um enviesamento não negligenciável do estimador pelo quociente. Estes autores salientam ainda que o estimador é muito preciso quando os pontos (y_k, x_k) da população se distribuem ao longo de uma recta que passa pela origem e tem um certo declive (desconhecido) R, podendo-se assim dizer que este modelo de regressão gera o estimador pelo quociente (veja-se Grosbras (1987) ou Särndal, Swensson e Wretman (1992) para mais detalhes). Daqui se conclui que os valores de *Y* devem ser proporcionais aos valores de *X* para que o método de estimação pelo quociente seja eficaz.

No que se refere à escolha da variável auxiliar, deve-se também ter em atenção que os valores da amostra que contribuem para a estimativa de $\hat{\mu}_x$, têm que pertencer à população da qual se conhece o verdadeiro valor de μ_x. Para deixar claro este pressuposto, Grosbras (1987, p. 130) fornece o seguinte exemplo: o consumo médio na indústria, μ_x, deve-se referir às empresas do *mesmo sector* das que foram utilizadas na amostra, ser calculado para o *mesmo período*, etc.

Relativamente às propriedades do estimador (3.3.36) (e, analogamente, para (3.3.37)), a dedução é imediata quando se consideram os resultados obtidos através da aplicação do método de linearização de Taylor ao estimador $\hat{R}$. Ou seja, uma vez que $\tau_y = \tau_x R$, por (3.3.23), conclui-se que:

$$(3.3.38) \qquad \hat{\tau}_{yQ} \doteq \tau_x R + \sum_s \frac{y_i - Rx_i}{\pi_i}$$

e, por (3.3.24),obtém-se a *variância aproximada* de $\hat{\tau}_{yQ}$:

(3.3.39) $$AV(\hat{\tau}_{yQ}) = \sum\sum_{U} \Delta_{ij} \frac{y_i - Rx_i}{\pi_i} \frac{y_j - Rx_j}{\pi_j}$$

Finalmente, como $\hat{\tau}_{yQ} = \tau_x \hat{R}$, por (3.3.26), obtém-se o estimador:

(3.3.40) $$\hat{V}(\hat{\tau}_{yQ}) = \left(\frac{\tau_x}{\hat{\tau}_{x\pi}}\right)^2 \sum\sum_{s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{R}x_i}{\pi_i} \frac{y_j - \hat{R}x_j}{\pi_j}$$

Estes resultados podem ser apresentados sob outra forma se atendermos a que a aproximação (3.3.38) pode ser escrita, de forma equivalente, como

(3.3.41) $$\hat{\tau}_{yQ} \doteq \tau_x R + (\hat{\tau}_{y\pi} - R\hat{\tau}_{x\pi})$$

Sob a aproximação (3.3.41), conclui-se que a variância aproximada de $\hat{\tau}_{yQ}$ é dada por:

(3.3.42) $$AV(\hat{\tau}_{yQ}) = V(\hat{\tau}_{y\pi}) + R^2 V(\hat{\tau}_{x\pi}) - 2RCov(\hat{\tau}_{y\pi}, \hat{\tau}_{x\pi})$$

e o respectivo estimador é

(3.3.43) $$\hat{V}(\hat{\tau}_{yQ}) = \hat{V}(\hat{\tau}_{y\pi}) + \hat{R}^2 \hat{V}(\hat{\tau}_{x\pi}) - 2\hat{R}\,\hat{Cov}(\hat{\tau}_{y\pi}, \hat{\tau}_{x\pi})$$

onde, os estimadores de $V(\hat{\tau}_{y\pi})$ [ou $V(\hat{\tau}_{x\pi})$] e $Cov(\hat{\tau}_{y\pi}, \hat{\tau}_{x\pi})$ são dados, respectivamente, por (2.6.12) e (2.6.16).

No Quadro 3.3.3 apresenta-se um resumo das propriedades do estimador pelo quociente usual do total da população, $\hat{\tau}_{yQ}$.

Quadro 3.3.3 - Propriedades do estimador pelo quociente usual de $\tau_y = \tau_x R$, com $R = \tau_y/\tau_x$

Estimação de τ_y, na presença de informação auxiliar
$\hat{\tau}_{yQ} = \tau_x \hat{R} = \tau_x \dfrac{\hat{\tau}_{y\pi}}{\hat{\tau}_{x\pi}}$, τ_x conhecido
$AV(\hat{\tau}_{yQ}) = \sum\sum_U \Delta_{ij} \dfrac{y_i - Rx_i}{\pi_i} \dfrac{y_j - Rx_j}{\pi_j}$
$\hat{V}(\hat{\tau}_{yQ}) = \left(\dfrac{\tau_x}{\hat{\tau}_{x\pi}}\right)^2 \sum\sum_s \dfrac{\Delta_{ij}}{\pi_{ij}} \dfrac{y_i - \hat{R}x_i}{\pi_i} \dfrac{y_j - \hat{R}x_j}{\pi_j}$

3.3.2.1 Estimação pelo quociente numa sondagem aleatória estratificada

Em seguida, apresentam-se dois métodos de estimação de μ_y pelo quociente (no caso do parâmetro de interesse ser τ_y, os resultados são análogos), para um plano de sondagem aleatória estratificada no qual foi utilizada a sondagem aleatória simples sem reposição (SASSR) em cada estrato. No primeiro método considera-se que a média da variável auxiliar *X* é conhecida na população (μ_x) e, no segundo, a média da variável auxiliar é conhecida em cada estrato *h* (h=1,…,H) da população ($\mu_{x,h}$).

♦ **<u>Método I</u> – Estimador combinado do quociente**

O estimador combinado do quociente deve-se a Hansen, Hurwitz e Gurney (1946) e obtém-se calculando os estimadores de Horvitz-Thompson de μ_x e μ_y para este plano de sondagem (c.f. secção 2.5) e recompondo em seguida esses estimadores através de (3.3.37). Utilizando a notação apresentada na secção 2.5, obtêm-se os resultados que se seguem.

$$\text{(3.3.44)} \qquad \hat{\mu}_{yQ_1} = \mu_x \frac{\hat{\mu}_{y\pi}}{\hat{\mu}_{x\pi}} = \mu_x \frac{\sum_h \frac{N_h}{N} \bar{y}_h}{\sum_h \frac{N_h}{N} \bar{x}_h}$$

Assim, este método pressupõe apenas o conhecimento da média da variável auxiliar *X* na população (μ_x).

O enviesamento deste estimador será pequeno se a dimensão da amostra for grande e o coeficiente de variação de $\hat{\mu}_x$ for pequeno (Rao, 1988). Para amostras grandes, a *variância aproximada* do estimador (3.3.44) é dada por:

$$(3.3.45) \qquad AV(\hat{\mu}_{yQ_1}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h}\left(S_{y,h}^2 - 2R\rho_h S_{y,h} S_{x,h} + R^2 S_{x,h}^2\right)$$

sendo ρ_h o coeficiente de correlação no estrato *h* e R o quociente $\frac{\mu_y}{\mu_x}$ na população; e, o respectivo estimador é:

$$(3.3.46) \qquad \hat{V}(\hat{\mu}_{yQ_1}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h}\left(s_{y,h}^2 - 2\hat{R} r_h s_{y,h} s_{x,h} + \hat{R}^2 s_{x,h}^2\right)$$

onde r_h é o estimador de ρ_h.

A demonstração destes resultados pode ser obtida utilizando os resultados do caso geral, apresentados anteriormente (para mais detalhes veja-se, por exemplo, Hansen, Hurwitz e Madow (1953a, 1953b) e Cochran (1977)). Rao (1988) fornece referências bibliográficas relevantes sobre formas alternativas de estimar a variância do estimador combinado do quociente.

♦ **<u>Método II</u>** – **Estimador separado do quociente**

O estimador separado do quociente obtém-se recompondo os estimadores das médias, estrato a estrato, e calculando-se em seguida o estimador usual da média para a sondagem aleatória estratificada. Utilizando a notação apresentada na secção 2.5, obtêm-se os resultados que se seguem.

O estimador separado do quociente é dado por:

$$(3.3.47) \qquad \hat{\mu}_{yQ_2} = \sum_h \frac{N_h}{N} \hat{\mu}_{yQ,h} = \sum_h \frac{N_h}{N} \mu_{x,h} \frac{\hat{\mu}_{y,h}}{\hat{\mu}_{x,h}}$$

Este método pressupõe que a média da variável auxiliar *X* é conhecida para cada estrato da população ($\mu_{x,h}$).

O enviesamento de $\hat{\mu}_{yQ_2}$ deduz-se facilmente tendo em consideração que

$$(3.3.48) \qquad B(\hat{\mu}_{yQ_2}) = \sum_h \frac{N_h}{N} B(\hat{\mu}_{yQ,h})$$

e, no estrato *h*, $\hat{\mu}_{yQ,h}$ não é mais do que o estimador pelo quociente usual para uma sondagem aleatória simples sem reposição.

O enviesamento será pequeno se as dimensões amostrais de cada estrato (n_h) forem grandes e os coeficientes de variação de X, em cada estrato, forem pequenos (Rao, 1988).

A *variância aproximada* do estimador separado do quociente é:

$$(3.3.49) \qquad AV(\hat{\mu}_{yQ_2}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h}\left[S_{y,h}^2 - 2R_h\rho_h S_{y,h}S_{x,h} + R_h^2 S_{x,h}^2\right]$$

sendo ρ_h o coeficiente de correlação e R_h o quociente $\frac{\mu_{y,h}}{\mu_{x,h}}$ na população, para cada estrato *h*.

Analogamente ao estimador combinado do quociente, um estimador da variância de (3.3.47) obtém-se substituindo: R_h por $\hat{R}_h$, as variâncias $S^2_{y,h}$ e $S^2_{x,h}$ pelas respectivas variâncias amostrais corrigidas ($s^2_{y,h}$ e $s^2_{x,h}$) e $\rho_h S_{y,h}S_{x,h}$ pela covariância amostral, em cada estrato *h* (Cochran, 1977).

♦ **<u>Comparação dos métodos</u>**

O estimador separado do quociente é preferível ao estimador combinado se os quocientes R_h forem distintos de estrato para estrato e, principalmente, se todos os n_h forem suficientemente grandes. Nestas condições, o enviesamento do estimador

será negligenciável e a expressão (3.3.49) será uma aproximação válida da variância de $\hat{\mu}_{yQ_2}$ (Cochran, 1977, p. 167; Grosbras, 1987, p. 138).

Assim, mesmo que a média da variável auxiliar *X* seja conhecida em cada estrato da população, o estimador combinado do quociente é preferível se a amostra não tiver um tamanho razoável em todos os estratos. Este estimador também é adequado se os R_h forem estáveis de estrato para estrato.

Para uma discussão mais detalhada sobre a escolha do estimador pelo quociente, numa sondagem aleatória estratificada, veja-se Hansen, Hurwitz e Madow (1953a).

3.4 Métodos básicos de estimação em domínios

Designa-se por domínio qualquer sub-população para a qual são necessárias estimativas pontuais separadas. Se for possível identificar domínios de estudo antes da amostra ser recolhida, então o plano de sondagem a adoptar deverá ter em consideração essas unidades. A dimensão da amostra em cada domínio deverá, então, garantir que as estimativas tenham uma precisão aceitável.

No entanto, nem sempre é possível implementar um plano de sondagem que contemple os domínios de estudo (por exemplo, por questões logísticas ou orçamentais ou por inexistência da informação auxiliar necessária) e, por vezes, a necessidade de estimativas separadas para certos domínios só se verifica depois da amostra ter sido recolhida. Nestes casos, é necessário recorrer a métodos de estimação que tirem o melhor partido possível da amostra que foi obtida e, eventualmente, da informação auxiliar que estiver disponível durante a fase de estimação. Estes procedimentos designam-se *métodos de estimação em domínios*.

O principal problema com que se defrontam os métodos de estimação em domínios prende-se com o facto da dimensão da amostra no interior dos domínios ser aleatória e, frequentemente, demasiado pequena para que seja possível obter estimativas de precisão aceitável.

Nos últimos anos foram desenvolvidas diversas técnicas que procuram lidar com o problema da *estimação em pequenos domínios*. Uma referência mais detalhada sobre estes métodos pode ser encontrada em Särndal (1984), Särndal e Hidiroglou (1989), Holt e Holmes (1994) e Coelho (1996).

Encontra-se fora do âmbito desta dissertação a apresentação detalhada desses métodos. Importa contudo apresentar as ferramentas básicas da estimação em domínios, não só por uma questão de enquadramento teórico de conceitos utilizados nos métodos de pós-estratificação, mas também como uma metodologia de estimação passível de resolver certos problemas de sobrecobertura das bases de sondagem (c. f. secção 3.2).

3.4.1 Notação

Seja U a população alvo da sondagem, de dimensão N, e considere-se uma partição de U em D sub-populações, U_1, …, U_d, …, U_D, designadas por domínios, de dimensões N_1, …, N_d, …, N_D, respectivamente. Ou seja,

$$(3.4.1) \qquad U = \bigcup_{d=1}^{D} U_d$$

$$(3.4.2) \qquad N = \sum_{d=1}^{D} N_d$$

O total e a média da variável de interesse *Y* no domínio U_d da população são, respectivamente:

$$(3.4.3) \qquad \tau_d = \sum_{i \in U_d} y_i \,, \qquad d = 1, \ldots, D$$

$$(3.4.4) \qquad \mu_d = \frac{\tau_d}{N_d} \qquad , d = 1, \ldots, D$$

sendo $\hat{\tau}_d$ e $\hat{\mu}_d$ a notação geral para os respectivos estimadores.

Suponhamos que foi retirada uma amostra aleatória *s*, de dimensão *n*, de acordo com um determinado plano de sondagem. Seja s_d o conjunto dos elementos de *s* que intersectam o domínio U_d, i.e.,

$$(3.4.5) \qquad s_d = s \cap U_d \qquad , d = 1, \ldots, D$$

e seja n_d a dimensão amostral de s_d. Tem-se, então:

$$(3.4.6) \qquad s = \bigcup_{d=1}^{D} s_d$$

(3.4.7) $$n = \sum_{d=1}^{D} n_d$$

Observe-se que n_d é aleatório e, por vezes, esta dimensão amostral é extremamente reduzida. Assume-se aqui que a probabilidade de s_d ter dimensão nula é negligenciável.

Considerando a variável indicatriz:

(3.4.8) $$\mathbb{I}_{i\in U_d} = \begin{cases} 1 & \text{se } i \in U_d \\ 0 & \text{caso contrário} \end{cases}$$

verifica-se que a dimensão amostral do domínio pode ser expressa como:

(3.4.9) $$n_d = \sum_U \mathbb{I}_{i\in U_d} \mathbb{I}_{i\in s} = \sum_{U_d} \mathbb{I}_{i\in s}$$

onde, $\mathbb{I}_{i\in s}$ é a variável indicatriz usual (variável de Cornfield, c.f. secção 2.2.5).

Desta forma, para um determinado plano de sondagem, com probabilidades de inclusão π_i, o valor esperado da dimensão amostral de s_d é:

(3.4.10) $$E(n_d) = \sum_U \mathbb{I}_{i\in U_d} \pi_i = \sum_{U_d} \pi_i$$

Verifica-se assim que as probabilidades de inclusão, definidas pelo plano de sondagem adoptado, são fundamentais para aquilo que se espera obter em termos da dimensão amostral do domínio.

3.4.2 Alguns métodos de estimação em domínios

O objectivo desta secção é apresentar os estimadores usuais do total (3.4.3) e da média (3.4.4) de um determinado domínio U_d da população, para um plano de sondagem genérico e para o caso particular de um plano de sondagem aleatória estratificada.

De um modo geral, a dimensão do domínio U_d na população é desconhecida. Neste caso, um estimador centrado do total de U_d é o estimador de Horvitz-Thompson:

(3.4.11) $$\hat{\tau}_{d\pi} = \sum_{i \in s_d} \frac{y_i}{\pi_i}$$

Observe-se que, utilizando a variável indicatriz $\mathbb{I}_{i \in U_d}$, definida por (3.4.8), se pode definir:

(3.4.12) $$y_{di} = y_i \mathbb{I}_{i \in U_d} = \begin{cases} y_i & \text{se } i \in U_d \\ 0 & \text{caso contrário} \end{cases}$$

e, portanto, o total do domínio U_d pode ser escrito como

(3.4.13) $$\tau_d = \sum_U y_{di}$$

e o estimador (3.4.11) vem equivalente a

(3.4.14) $$\hat{\tau}_{d\pi} = \sum_s \frac{y_{di}}{\pi_i}$$

Assim, a demonstração das propriedades do estimador (3.4.11) (e de outros estimadores em domínios, baseados no estimador de Horvitz-Thompson) são imediatas[1], quando se considera o caso geral do estimador de Horvitz-Thompson (veja-se a secção 2.6.1) e se toma como variável de estudo, a variável de interesse no domínio, Y_d, que toma na população U os valores y_{di}, dados por (3.4.12).

Utilizando a notação apresentada na secção 2.6.1, verifica-se então que a variância de $\hat{\tau}_{d\pi}$ é dada por

(3.4.15) $$V(\hat{\tau}_{d\pi}) = \sum\sum_{U_d} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

e é estimada sem enviesamento por

[1] Note-se que a utilização da variável indicatriz $\mathbb{I}_{i \in U_d}$ resolve o problema da aleatoriedade da dimensão amostral nos domínios.

(3.4.16) $$\hat{V}(\hat{\tau}_{d\pi}) = \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Quanto à estimação da média do domínio U_d, se a sua dimensão N_d for desconhecida então o parâmetro de interesse é um quociente entre duas quantidades desconhecidas ($\mu_d = \tau_d/N_d$). No entanto, supondo que $n_d \geq 1$, Särndal, Swensson e Wretman (1992, p. 391) referem que, ainda que se conheça o valor de N_d, dever-se-á utilizar sempre o estimador desse quociente (c. f. secção 3.3.1):

(3.4.17) $$\hat{\mu}_d = \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i}$$

onde,

(3.4.18) $$\hat{N}_d = \sum_{s_d} \frac{1}{\pi_i}$$

O estimador (3.4.17) é designado na literatura por "***weighted domain sample mean***". Pelos resultados apresentados na secção 3.3.1.1, é imediato que $\hat{\mu}_d$ é um estimador aproximadamente não enviesado e a variância aproximada é

(3.4.19) $$AV(\hat{\mu}_d) = \frac{1}{{N_d}^2} \sum\sum_{U_d} \Delta_{ij} \frac{y_i - \mu_d}{\pi_i} \frac{y_j - \mu_d}{\pi_j}$$

sendo o respectivo estimador dado por

(3.4.20) $$\hat{V}(\hat{\mu}_d) = \frac{1}{{\hat{N}_d}^2} \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_d}{\pi_i} \frac{y_j - \hat{\mu}_d}{\pi_j}$$

Quando a dimensão do domínio na população é conhecida, Särndal, Swensson e Wretman (1992, p. 391) referem que, o estimador a utilizar para estimar o total desse domínio é

(3.4.21) $$\hat{\tau}_{dw} = N_d \hat{\mu}_d$$

onde, $\hat{\mu}_d$ é o estimador "*wheighted domain sample mean*" dado por (3.4.17). Ou seja,

$$(3.4.22) \qquad \hat{\tau}_{dw} = \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i}$$

com $\hat{N}_d$ dado pela expressão (3.4.18).

Pelos resultados (3.4.19) e (3.4.20) e considerando, naturalmente, a forma do estimador (3.4.22), conclui-se que a sua variância aproximada é

$$(3.4.23) \qquad AV(\hat{\tau}_{dw}) = \sum\sum_{U_d} \Delta_{ij} \frac{y_i - \mu_d}{\pi_i} \frac{y_j - \mu_d}{\pi_j}$$

e pode ser estimada por

$$(3.4.24) \qquad \hat{V}(\hat{\tau}_{dw}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_d}{\pi_i} \frac{y_j - \hat{\mu}_d}{\pi_j}$$

No Quadro 3.4.1 apresenta-se um resumo das propriedades do estimador usual de τ_d (Horvitz-Thompson) e, no Quadro 3.4.2, um resumo das propriedades dos estimadores de τ_d e μ_d (*weighted domain sample mean*).

Na secção seguinte, apresentam-se os estimadores de domínios para um plano de sondagem aleatória estratificada (SASSR nos estratos). Alguns resultados referentes ao plano de sondagem aleatória simples sem reposição encontram-se no Anexo 2, secção A2.3.2.

Quadro 3.4.1 – Propriedades do estimador usual de τ_d (Horvitz-Thompson)

Estimação do total do domínio U_d
$\hat{\tau}_{d\pi} = \sum_{i \in s_d} \frac{y_i}{\pi_i}$
$E(\hat{\tau}_{d\pi}) = \tau_d$
$V(\hat{\tau}_{d\pi}) = \sum\sum_{U_d} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$
$\hat{V}(\hat{\tau}_{d\pi}) = \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$
$E(\hat{V}(\hat{\tau}_{d\pi})) = V(\hat{\tau}_{d\pi})$

Quadro 3.4.2 – Propriedades dos estimadores de τ_d e μ_d

(*weighted domain sample mean*)

Estimação da média do domínio U_d
$\hat{\mu}_d = \frac{\hat{\tau}_{d\pi}}{\hat{N}_d} = \frac{\sum_{s_d} \frac{y_i}{\pi_i}}{\sum_{s_d} \frac{1}{\pi_i}}$
$AV(\hat{\mu}_d) = \frac{1}{N_d^2} \sum\sum_{U_d} \Delta_{ij} \frac{y_i - \mu_d}{\pi_i} \frac{y_j - \mu_d}{\pi_j}$
$\hat{V}(\hat{\mu}_d) = \frac{1}{\hat{N}_d^2} \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_d}{\pi_i} \frac{y_j - \hat{\mu}_d}{\pi_j}$
Estimação do total do domínio U_d
$\hat{\tau}_{dw} = \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i}$
$AV(\hat{\tau}_{dw}) = \sum\sum_{U_d} \Delta_{ij} \frac{y_i - \mu_d}{\pi_i} \frac{y_j - \mu_d}{\pi_j}$
$\hat{V}(\hat{\tau}_{dw}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_d}{\pi_i} \frac{y_j - \hat{\mu}_d}{\pi_j}$

3.4.2.1 Estimação em domínios numa sondagem aleatória estratificada

Os resultados apresentados nesta secção referem-se ao caso em que foi utilizado um plano de sondagem aleatória estratificada tal que, em cada estrato *h* (h = 1, …, H), foi utilizado um plano de sondagem aleatória simples sem reposição.

Neste caso, o domínio U_d poderá atravessar os estratos definidos *a priori*, sendo, portanto, necessário definir alguma notação adicional.

Seja U_{dh} o conjunto de elementos da população da célula (d, h), definida pela intersecção do estrato inicial *h* com o domínio U_d, e seja N_{dh} o número de elementos de U_{dh}. O número de elementos do domínio U_d na população é então dado por:

$$(3.4.25) \qquad N_d = \sum_{h=1}^{H} N_{dh}$$

O total e a média da variável de interesse *Y* no domínio U_d da população são, respectivamente:

$$(3.4.26) \qquad \tau_d = \sum_{h=1}^{H} N_{dh}\mu_{dh}$$

e

$$(3.4.27) \qquad \mu_d = \sum_{h=1}^{H} \frac{N_{dh}}{N_d}\mu_{dh}$$

onde, μ_{dh} é a média da população em U_{dh}.

Seja s_{dh} o conjunto de elementos da amostra que intersectam U_{dh}. A dimensão, aleatória, de s_{dh} denota-se por n_{dh}.

Para este plano de sondagem, se N_d for desconhecido, dever-se-á utilizar o estimador de Horvitz-Thompson apresentado em (3.4.11), i.e.,

(3.4.28) $$\hat{\tau}_{d\pi_{str}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{dh}} y_i$$

A variância de $\hat{\tau}_{d\pi_{str}}$ e o estimador de $V(\hat{\tau}_{d\pi_{str}})$ são, respectivamente:

(3.4.29)
$$V(\hat{\tau}_{d\pi_{str}}) = \sum_{h=1}^{H} N_h{}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{N_h - 1} \left[\sum_{i \in U_{dh}} (y_i - \mu_{dh})^2 + N_{dh} \left(1 - \frac{N_{dh}}{N_h} \right) \mu_{dh}{}^2 \right]$$

(3.4.30)
$$\hat{V}(\hat{\tau}_{d\pi_{str}}) = \sum_{h=1}^{H} N_h{}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{n_h - 1} \left[\sum_{i \in s_{dh}} (y_i - \bar{y}_{s_{dh}})^2 + n_{dh} \left(1 - \frac{n_{dh}}{n_h} \right) \bar{y}_{s_{dh}}{}^2 \right]$$

onde,

(3.4.31) $$\bar{y}_{s_{dh}} = \frac{1}{n_{dh}} \sum_{i \in s_{dh}} y_i$$

Os resultados (3.4.29) e (3.4.30) obtêm-se a partir de (3.4.15) e de (3.4.16), respectivamente. Veja-se a demonstração detalhada de (3.4.30) no Anexo 2, secção A2.3.1 (a demonstração de (3.4.29) é análoga).

Para este plano de sondagem, o estimador (3.4.17) da média do domínio U_d é então dado por:

(3.4.32) $$\hat{\mu}_{d_{str}} = \frac{1}{\hat{N}_d} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{dh}} y_i$$

onde, N_h/n_h é o peso do estrato *h* (c. f. secção 2.6) e

(3.4.33) $$\hat{N}_d = \sum_{h=1}^{H} \frac{N_h}{n_h} n_{dh}$$

Naturalmente, $\hat{\mu}_{d_{str}}$ é um estimador enviesado da média do domínio μ_d, apesar de, como refere Gomes (1998, p. 123),

(3.4.34) $$\bar{y}_{s_{dh}} = \frac{1}{n_{dh}} \sum_{i \in s_{dh}} y_i$$

ser um estimador aproximadamente centrado de μ_{dh}. Por outro lado, este autor observa ainda que o enviesamento será tanto menor quanto mais semelhantes forem as médias, da variável de interesse no domínio, de estrato para estrato, visto que o enviesamento é nulo se $\mu_{dh} = \mu_d$, $\forall h$.

As expressões da variância aproximada de $\hat{\mu}_{d_{str}}$ e do respectivo estimador podem ser derivadas a partir dos resultados referentes ao estimador combinado do quociente apresentados na secção 3.3.2.1 (para mais detalhes veja-se Cochran 1977). Essas expressões podem também ser deduzidas, para este plano de sondagem, através de (3.4.19) e (3.4.20), respectivamente:

(3.4.35)

$$AV(\hat{\mu}_{d_{str}}) = \frac{1}{N_d^2} \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \frac{1}{N_h - 1} \left[\sum_{i \in U_{dh}} (y_i - \mu_{dh})^2 + N_{dh} \left(1 - \frac{N_{dh}}{N_h} \right) (\mu_{dh} - \mu_d)^2 \right]$$

(3.4.36)

$$\hat{V}(\hat{\mu}_{d_{str}}) = \frac{1}{\hat{N}_d^2} \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \frac{1}{n_h - 1} \left[\sum_{i \in s_{dh}} \left(y_i - \bar{y}_{s_{dh}} \right)^2 + n_{dh} \left(1 - \frac{n_{dh}}{n_h} \right) \left(\bar{y}_{s_{dh}} - \hat{\mu}_{d_{str}} \right)^2 \right]$$

onde, $\hat{N}_d$ é dado por (3.4.33), $\bar{y}_{s_{dh}}$ é a média da variável *Y* em s_{dh} é dada por (3.4.34) e $f_h = n_h/N_h$.

A demonstração detalhada do resultado (3.4.36) encontra-se no Anexo 2, secção A2.3.1 (a demonstração de (3.4.35) é análoga).

O termo $\left(\bar{y}_{s_{dh}} - \hat{\mu}_{d_{str}} \right)^2$ representa a contribuição das diferenças *inter-estratos* das médias do domínio. Quanto maiores forem as diferenças entre os μ_{dh}, de estrato

para estrato, maior será o ganho de precisão da estratificação relativamente à sondagem aleatória simples. No entanto, como já foi referido, as diferenças entre as médias da população em U_{dh} acentuam o enviesamento do estimador, pelo que, para reduzir esse efeito, o domínio deverá ter uma dimensão razoável em cada estrato.

Ou seja, tal como Durbin (1958, citado por Särndal, Swensson e Wretman 1992, p. 394) também observou, quando se pretendem estimativas para pequenos domínios da população, os ganhos de eficiência da estratificação perdem-se, dado que o enviesamento é reduzido apenas quando o domínio cobre uma parte significativa de cada estrato.

Relativamente à estimação do total do domínio U_d, neste plano de sondagem, as conclusões são imediatas se atendermos aos resultados anteriores. Caso N_d seja conhecido, atendendo a (3.4.21) e (3.4.32), o estimador proposto é

$$(3.4.37) \qquad \hat{\tau}_{d\,w_{str}} = N_d \hat{\mu}_{d_{str}} = \frac{N_d}{\hat{N}_d} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{dh}} y_i$$

onde $\hat{N}_d$ é dado por (3.4.33) e, portanto, a expressão (3.4.36) multiplicada por N_d^2 fornece um estimador de $AV(\hat{\tau}_{d\,w_{str}})$:

$$(3.4.38) \qquad \hat{V}(\hat{\tau}_{d\,w_{str}}) = N_d^2 \hat{V}(\hat{\mu}_{d_{str}})$$

3.5 Estimadores de pós-estratificação

Como o próprio nome indica, a pós-estratificação consiste em estratificar a amostra depois de esta ter sido recolhida, utilizando informação auxiliar que se encontre disponível na fase de estimação. Naturalmente, tal como nos planos de sondagem aleatória estratificada, os pós-estratos devem ser o mais homogéneos possível e, portanto, a variável que define os pós-estratos deverá estar fortemente correlacionada com as variáveis de interesse. Nos métodos de pós-estratificação, assume-se que as dimensões dos pós-estratos na população são conhecidas. Estes métodos consistem, então, em ajustar os pesos iniciais por forma a que a distribuição da amostra reponderada, para certas características da população, esteja de acordo com a distribuição conhecida do número de elementos da população com essas características.

Quando se recorre a duas ou mais variáveis auxiliares para pós-estratificar a amostra, podem ocorrer duas situações. Se a dimensão de todos os pós-estratos resultantes (do cruzamento dessas variáveis) for conhecida na população, o problema reduz-se ao caso em que se utiliza apenas uma variável de pós-estratificação e, portanto, os métodos de pós-estratificação são directamente aplicáveis.

No entanto, tal informação nem sempre se encontra disponível. Por vezes, dispõem-se apenas das dimensões marginais na população. Ou seja, a única informação auxiliar que existe diz respeito à dimensão da população nas categorias definidas por cada uma das variáveis, tomadas isoladamente. Para lidar com este problema, Deming e Stephan (1940) introduziram um método designado *raking ratio*. Posteriormente, Deville e Särndal (1992) desenvolveram uma família de *estimadores de calibração*, onde os pesos iniciais são ajustados através de um conjunto de *equações de calibração*. Por forma a que os pesos ajustados se aproximem o mais possível dos pesos de inclusão, é escolhida uma *função distância*. Uma extensão destes métodos, designada *generalized raking*, deve-se a Deville, Särndal e Sautory (1993) [1]. O método *raking ratio* proposto por Deming e Stephan constitui um caso particular destes métodos.

[1] Para mais detalhes sobre investigação efectuada recentemente nesta área veja-se Singh e Mohl (1996), Skinner (1998) e Lundström e Särndal (1999).

Lazzeroni e Little (1998, p. 61) expressam claramente as razões que nos levam a apresentar os estimadores de pós-estratificação em mais detalhe:

> "*Poststratification can reduce bias caused by problems in the sampling frame or unit nonresponse, and it can also increase the precision of estimates.*"

Lehtonen e Pahkinen (1996, p. 94) referem que, se os pós-estratos forem homogéneos internamente, a partição da amostra efectuada *a posteriori* pode capturar uma grande parte da variância total da variável de estudo, resultando numa diminuição da variância *design-based* do estimador, i.e. da variância introduzida pelo desenho da amostra. Holt e Smith (1979, p. 34) apontam um ponto fraco óbvio deste procedimento que é a falta de controlo sobre a localização da amostra que, em circunstâncias extremas pode conduzir a estratos de dimensão nula.

Outra vantagem dos métodos de pós-estratificação é que estes podem ser adaptados ao tratamento de *outliers* desde que seja possível ponderar os pós-estratos da população que contêm os indivíduos atípicos (Gomes, 1998, p. 90).

Na secção seguinte apresentam-se referências bibliográficas relevantes sobre algumas abordagens à pós-estratificação e o estimador de pós-estratificação, para um plano de sondagem genérico. Nas secções 3.5.2 e 3.5.3, respectivamente, apresentam-se em mais detalhe as situações em que foi utilizado um plano de sondagem aleatória simples sem reposição (SASSR) e um plano de sondagem aleatória estratificada (utilizando a SASSR em cada estrato), para retirar a amostra.

3.5.1 Algumas abordagens à pós-estratificação

A abordagem "clássica" da teoria das sondagens, aqui designada por *design based*, considera que as características da população são fixas e que a componente probabilística é introduzida quando se adopta um determinado plano de amostragem. Existe uma outra abordagem, baseada em modelos estatísticos, e que designaremos por *model based*. Na inferência *model based*, considera-se que os valores tomados pelos indivíduos da população ($Y_1, Y_2, \ldots, Y_N$), relativamente a uma determinada característica em estudo, correspondem a realizações de N variáveis aleatórias que seguem uma distribuição conjunta ξ. Um modelo de superpopulação,

como usualmente é designado, mais não é do que um conjunto de condições que permitem definir a classe de distribuições à qual ξ deverá pertencer (Dussaix, 1987, p. 67).

Existe uma outra abordagem, designada por *model-assisted*, onde se assume que a população não é realmente gerada pelo modelo ξ, ao contrário da inferência *model based*, e que, não sendo *design based*, utiliza também pressupostos do desenho da amostra, nomeadamente as probabilidades de inclusão subjacentes ao plano adoptado. Desta forma, as conclusões sobre os parâmetros da população finita são independentes das hipóteses formuladas sobre o modelo (Särndal, Swensson e Wretman, 1992, p. 227).

A escolha do tipo de inferência a utilizar (*design based* ou *model based*), no âmbito das sondagens, tem gerado grande controvérsia uma vez que são totalmente distintas e incompatíveis. Referências mais detalhadas sobre esta questão podem ser encontradas em Nathan (1988), Thomsen e Tesfu (1988) e Särndal, Swensson e Wretman (1992). Em particular, no que diz respeito aos estimadores de pós-estratificação, veja-se Skinner, Holt e Smith (1989) e Valliant (1993). Para mais detalhes sobre a abordagem *model assisted* veja-se Särndal, Swensson e Wretman (1992).

Os estimadores de pós-estratificação têm, então, sido analisados sob diversos pontos de vista por vários autores, veja-se por exemplo: Hansen, Hurwitz e Madow (1953a, 1953b); Williams (1962); Holt e Smith (1979); Rao (1985); Särndal, Swensson e Wretman (1992); Valliant (1993); Leonard *et al.* (1994) e Rao (1994).

Existem algumas abordagens recentes às técnicas de pós-estratificação que não se encontrando no âmbito deste trabalho seriam, no entanto, interessantes de investigar em mais pormenor. Neste contexto, são de salientar os trabalhos de Little (1993), Gelman e Little (1997) e Lazzeroni e Little (1998). Little (1993) considera a pós-estratificação através de uma versão *Bayesiana* da abordagem *model-based*; Gelman e Little (1997) apresentam um modelo de regressão logística hierárquica que é utilizado na obtenção das estimativas de uma variável binária; e Lazzeroni e Little (1998) utilizam uma abordagem *model-based* que pressupõe que as médias dos pós-estratos se distribuem sobre uma linha de regressão linear e que assume uma estrutura autoregressiva das covariâncias dessas médias.

Uma vez que estamos especialmente interessados em analisar os estimadores de pós-estratificação para a sondagem aleatória simples sem reposição e para a sondagem aleatória estratificada, iremos apresentar de forma abreviada a abordagem (*design-based*) considerada por Williams (1962) e Rao (1985), sobre a forma desses estimadores para um plano de sondagem genérico. Para os planos de sondagem em análise, os estimadores resultantes são formalmente idênticos ao da abordagem *model-assisted* (para mais detalhes veja-se, por exemplo, Särndal, Swensson e Wretman 1992).

Suponhamos que, na fase de estimação, se dispõe de informação auxiliar que permita dividir a amostra em L pós-estratos. Analogamente à notação utilizada para a sondagem aleatória estratificada, sejam n_1, …, n_i, …, n_L as dimensões amostrais dos pós-estratos e s_i o conjunto dos elementos da amostra que pertencem ao pós-estrato i (i=1, …, L). Uma vez que a estratificação da amostra é efectuada depois de esta ter sido recolhida, as dimensões amostrais dos pós-estratos são variáveis aleatórias, contrariamente à amostragem estratificada convencional. Nos métodos de pós-estratificação assume-se, também, que as dimensões N_1, …, N_i …, N_L dos pós-estratos na população são conhecidas.

Para um plano genérico de amostragem, um estimador de pós-estratificação do total da população, τ, é dado por:

$$(3.5.1) \qquad \hat{\tau}_{PS} = \sum_{i=1}^{L} N_i \frac{\hat{\tau}_i}{\hat{N}_i}$$

onde $\hat{\tau}_i$ e $\hat{N}_i$ são os usuais estimadores centrados de domínios do total e da dimensão do *i-ésimo* pós-estrato, respectivamente.

Ou seja, considerando que cada pós-estrato i (i=1, …, L) corresponde a um domínio na população, pelo resultado (3.4.11) tem-se

$$(3.5.2) \qquad \hat{\tau}_i = \hat{\tau}_{i\pi} = \sum_{k \in s_i} \frac{y_k}{\pi_k}$$

e, por (3.4.18),

(3.5.3) $$\hat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k}$$

E, portanto, o estimador de pós-estratificação (3.5.1) pode ser escrito como:

(3.5.4) $$\hat{\tau}_{PS} = \sum_{i=1}^{L} \hat{\tau}_{iw}$$

onde, $\hat{\tau}_{iw}$ é o estimador do total do domínio dado por (3.4.22):

(3.5.5) $$\hat{\tau}_{iw} = \frac{N_i}{\hat{N}_i} \sum_{k \in s_i} \frac{y_k}{\pi_k}$$

Sendo $w_k = 1/\pi_k$ o peso de inclusão (ou coeficiente de extrapolação) do indivíduo *k*, subjacente ao desenho da amostra (*design-weight*), observe-se que o estimador de pós-estratificação do total da população (3.5.1) também pode ser escrito da seguinte forma:

(3.5.6) $$\hat{\tau}_{PS} = \sum_{i=1}^{L} \sum_{k \in s_i} \frac{N_i}{\hat{N}_i} w_k y_k$$

Esta forma de apresentar o estimador de pós-estratificação permite-nos evidenciar o ajustamento pelo quociente dos pesos iniciais, w_k.

Para um plano genérico de amostragem, um estimador de pós-estratificação de μ é dado, obviamente, por:

(3.5.7) $$\hat{\mu}_{ps} = \frac{1}{N} \hat{\tau}_{PS}$$

Nas secções seguintes, apresentam-se com mais detalhe os estimadores de pós-estratificação para o plano de sondagem aleatória simples sem reposição (secção 3.5.2) e para o plano de sondagem aleatória estratificada, com SASSR em cada estrato (secção 3.5.3).

3.5.2 Sondagem aleatória simples sem reposição

Considere-se uma amostra aleatória, de dimensão *n*, obtida por sondagem aleatória simples sem reposição (SASSR) de uma população com dimensão conhecida *N*.

Nas condições apresentadas anteriormente, o algoritmo que se apresenta em seguida permite obter o estimador de pós-estratificação genérico (3.5.6), para o plano de sondagem em análise.

1º – Determinar os pesos de inclusão de cada indivíduo:

(3.5.8) $w_k = 1/\pi_k = N/n$, $k = 1, ..., n$

2º – Pós-estratificar a amostra.

Suponhamos que se obtiveram *L* pós-estratos com dimensões amostrais n_i para $i = 1, ..., L$. Sejam N_i as dimensões dos pós-estratos na população e w_{ik} o peso de inclusão do indivíduo *k* que se encontra no pós-estrato *i*.

(3.5.9) $w_{ik} = N/n$, $i = 1, ..., L$;
k pertencente ao pós-estrato *i*

3º – Calcular os pesos $N_i/\hat{N}_i$ para cada indivíduo da amostra.

(3.5.10) $$\frac{N_i}{\sum_{k=1}^{n_i} w_{ik}} = \frac{N_i}{\sum_{k=1}^{n_i} \frac{N}{n}} = \frac{N_i}{n_i \frac{N}{n}} \quad , \qquad i = 1, ..., L$$

k pertencente ao pós-estrato *i*

4º – Calcular os pesos ajustados.

(3.5.11) $$\frac{N_i}{\hat{N}_i} w_{ik} = \frac{N_i}{n_i \frac{N}{n}} \left(\frac{N}{n}\right) = \frac{N_i}{n_i} \quad , \qquad i = 1, ..., L$$

k pertencente ao pós-estrato *i*

5º – Calcular os **estimadores de pós-estratificação de τ e μ** para uma sondagem aleatória simples sem reposição.

(3.5.12) $$\hat{\tau}_{ps,sas} = \sum_{i=1}^{L} \sum_{k \in s_i} \frac{N_i}{n_i} y_k$$

(3.5.13) $$\hat{\mu}_{ps,sas} = \frac{1}{N} \sum_{i=1}^{L} \sum_{k \in s_i} \frac{N_i}{n_i} y_k$$

Conclui-se assim que, para este plano de sondagem, os estimadores resultantes são formalmente idênticos aos estimadores de Horvitz-Thompson para a amostragem estratificada. Ou seja, a estimativa do total da população obtém-se somando todas as observações da amostra multiplicadas pelos pesos ajustados N_i/n_i, isto é, pelos pesos dos pós-estratos.

Denotando por $\bar{y}_i$ a média amostral no pós-estrato *i*, o estimador de pós-estratificação da média da população, para este plano de sondagem, pode ser escrito como:

(3.5.14) $$\hat{\mu}_{ps,sas} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$$

Apresentando o estimador sob esta forma permite-nos observar que a média em cada pós-estrato é ponderada pela dimensão relativa desse pós-estrato na população. Desta forma, se uma amostra estiver desequilibrada para algumas características da população, o estimador de pós-estratificação corrige este desequilíbrio automaticamente (Holt e Smith, 1979, p. 34).

Relativamente às propriedades deste estimador, o facto de as dimensões amostrais nos estratos serem agora variáveis aleatórias, provoca alguma controvérsia sobre a forma mais adequada para a variância deste estimador. Em particular, surgem duas distribuições amostrais às quais pode estar associado:

1. a distribuição condicional sobre o vector das dimensões amostrais nos estratos efectivamente obtidos na amostra em estudo $\underset{\sim}{n} = (n_1, \ldots, n_L)$;

2. a distribuição não condicional determinada por todas as amostras possíveis de dimensão fixa *n*.

No que se refere ao enviesamento, ambas as abordagens conduzem à mesma conclusão, ou seja, os estimadores de pós-estratificação (3.5.12) e (3.5.13) são centrados. Em seguida, demonstra-se esta propriedade e, ainda, os resultados relativos à variância do estimador de pós-estratificação, para cada uma das abordagens.

Na abordagem condicional (1), supõe-se que é possível reconhecer que a amostra tem a configuração $\underset{\sim}{n} = (n_1, \ldots, n_L)$, definida por *L* estratos da população, cujos pesos N_i/N são conhecidos. O conjunto das amostras de referência (de dimensão fixa *n*) é agora o conjunto $S_{\underset{\sim}{n}}$ das $\prod \binom{N_i}{n_i}$ amostras que possuem a configuração efectiva $\underset{\sim}{n}$, uma vez que a distribuição de $\underset{\sim}{n}$ é completamente conhecida. Ou seja, a probabilidade associada a cada amostra *s* pertencente a $S_{\underset{\sim}{n}}$ é

$$(3.5.15) \qquad p(s \mid \underset{\sim}{n}) = \prod \binom{N_i}{n_i}^{-1}$$

e corresponde, portanto, a uma amostra aleatória estratificada $(s_1, \ldots, s_i, \ldots, s_L)$ com n_i $(n_i \geq 1)$ a dimensão dos estratos. Logo, pelos resultados apresentados na secção 2.5, conclui-se que

$$(3.5.16) \qquad E(\hat{\tau}_{ps,sas} \mid \underset{\sim}{n}) = \tau$$

$$(3.5.17) \qquad E(\hat{\mu}_{ps,sas} \mid \underset{\sim}{n}) = \mu$$

e que a **variância condicional** não é mais do que a variância usual para amostras estratificadas sem reposição. Ou seja, considerando uma notação

análoga à utilizada para a sondagem aleatória estratificada (veja-se a secção 2.5) obtém-se:

$$(3.5.18) \qquad V(\hat{\tau}_{ps,sas} \mid \underset{\sim}{n}) = \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}$$

$$(3.5.19) \qquad V(\hat{\mu}_{ps,sas} \mid \underset{\sim}{n}) = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}$$

desde que todos os $n_i > 1$.

Naturalmente, desde que todos os $n_i > 1$, os estimadores da variância condicional de $\hat{\tau}_{ps,sas}$ e $\hat{\mu}_{ps,sas}$ são, respectivamente:

$$(3.5.20) \qquad \hat{V}(\hat{\tau}_{ps,sas} \mid \underset{\sim}{n}) = \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}$$

$$(3.5.21) \qquad \hat{V}(\hat{\mu}_{ps,sas} \mid \underset{\sim}{n}) = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}$$

Särndal, Swensson e Wretman (1992, p. 288) salientam o facto de $s_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2$ ser um estimador condicionalmente centrado relativamente a $S_i^2 = \frac{1}{N_i - 1} \sum_{k=1}^{N_i} (Y_{ik} - \mu_i)^2$.

Relativamente à <u>abordagem não condicional</u> (2), há agora que considerar dois níveis de aleatoriedade:

- a introduzida pelas dimensões amostrais dos pós-estratos ($\underset{\sim}{n}$);
- a introduzida pelo plano de sondagem adoptado para seleccionar os elementos da amostra.

Em termos de cálculo do valor esperado e da variância considera-se, então, primeiro que as dimensões n_i são fixas e, em seguida, introduz-se a aleatoriedade dos n_i. Esquematicamente:

(3.5.22) $E(.) = E_{\underset{\sim}{n}}[E(.|\underset{\sim}{n})]$

(3.5.23) $V(.) = V_{\underset{\sim}{n}}[E(.|\underset{\sim}{n})] + E_{\underset{\sim}{n}}[V(.|\underset{\sim}{n})]$

Assim, o resultado (3.5.16) permite-nos obter o valor esperado não condicional do estimador de pós-estratificação do total da população:

(3.5.24) $E(\hat{\tau}_{ps,sas}) = E_{\underset{\sim}{n}}[E(\hat{\tau}_{ps,sas}|\underset{\sim}{n})] = E_{\underset{\sim}{n}}(\tau) = \tau$

Analogamente, tem-se:

(3.5.25) $E(\hat{\mu}_{ps,sas}) = \mu$

Quanto ao cálculo da variância não condicional (2), antes de mais, observe-se que

(3.5.26) $V_{\underset{\sim}{n}}[E(\hat{\tau}_{ps,sas}|\underset{\sim}{n})] = V_{\underset{\sim}{n}}(\tau) = 0$

e, portanto, por (3.5.23),

(3.5.27) $V(\hat{\tau}_{ps,sas}) = E_{\underset{\sim}{n}}[V(\hat{\tau}_{ps,sas}|\underset{\sim}{n})]$

Logo, por (3.5.18), conclui-se que a variância não condicional de $\hat{\tau}_{ps,sas}$ se obtém através de

(3.5.28) $V(\hat{\tau}_{ps,sas}) = \sum_{i=1}^{L} N_i^2 \left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i} \right] S_i^2$

Supondo que a probabilidade de $n_i=0$ é negligenciável ($\forall$ i=1, ..., L), obtém-se através da expansão em série de Taylor de $1/n_i$ a seguinte aproximação para $E(1/n_i)$ (Särndal, Swensson e Wretman, 1992, p. 286):

$$(3.5.29) \qquad E\left(\frac{1}{n_i}\right) \doteq \frac{1}{E(n_i)}\left[1+\frac{V(n_i)}{E(n_i)^2}\right]$$

Dado que a dimensão amostral n_i é uma variável aleatória com distribuição Hipergeométrica H(N, n, $\frac{N_i}{N}$) tal que $\sum_{i=1}^{L} n_i = n$, tem-se:

$$(3.5.30) \qquad E(n_i) = n\frac{N_i}{N}$$

$$(3.5.31) \qquad V(n_i) = \frac{N-n}{N-1}\, n\frac{N_i}{N}\left(1-\frac{N_i}{N}\right)$$

Logo, se *n* for suficientemente elevado (e N≈N–1) e se a probabilidade de $n_i=0$ é negligenciável, conclui-se:

$$(3.5.32) \qquad E\left(\frac{1}{n_i}\right) \doteq \frac{1}{n\frac{N_i}{N}}\left[1+\frac{(1-f)\left(1-N_i/N\right)}{n\frac{N_i}{N}}\right]$$

com f = n/N.

Substituindo a aproximação (3.5.32) em (3.5.28), obtém-se

$$(3.5.33) \qquad V(\hat{\tau}_{ps,sas}) \doteq \sum_{i=1}^{L}\frac{N_i}{f}\left[1+\frac{(1-f)(1-N_i/N)}{N_i f}\right]S_i^2 - \sum_{i=1}^{L} N_i\, S_i^2$$

Simplificando esta expressão, tem-se:

$$(3.5.34) \qquad V(\hat{\tau}_{ps,sas}) \doteq \left(\frac{1}{f}-1\right)\sum_{i=1}^{L} N_i\, S_i^2 \;+\; \sum_{i=1}^{L}\frac{1-f}{f^2}\left(1-\frac{N_i}{N}\right)S_i^2 =$$

$$= \left(\frac{N}{n} - 1\right) \sum_{i=1}^{L} N_i S_i^2 \; + \; \sum_{i=1}^{L} \left(\frac{N^2}{n^2} - \frac{N}{n}\right) \left(1 - \frac{N_i}{N}\right) S_i^2 =$$

$$= \frac{N}{n} \left(1 - \frac{n}{N}\right) \sum_{i=1}^{L} N_i S_i^2 \; + \; \sum_{i=1}^{L} \frac{N^2}{n^2} \left(1 - \frac{n}{N}\right) \left(1 - \frac{N_i}{N}\right) S_i^2$$

Conclui-se então que a **variância não condicional** de $\hat{\tau}_{ps,sas}$ e de $\hat{\mu}_{ps,sas}$ é, respectivamente,

(3.5.35) $$V(\hat{\tau}_{ps,sas}) \doteq N^2 \frac{1-f}{n} \sum_{i=1}^{L} \frac{N_i}{N} S_i^2 \; + \; N^2 \frac{1-f}{n^2} \sum_{i=1}^{L} \left(1 - \frac{N_i}{N}\right) S_i^2$$

(3.5.36) $$V(\hat{\mu}_{ps,sas}) \doteq \frac{1-f}{n} \sum_{i=1}^{L} \frac{N_i}{N} S_i^2 \; + \; \frac{1-f}{n^2} \sum_{i=1}^{L} \left(1 - \frac{N_i}{N}\right) S_i^2$$

Hansen, Hurwitz, e Madow (1953a, p. 232) defendem a utilização da variância não condicional (3.5.35), com S_i^2 estimado por s_i^2:

(3.5.37) $$\hat{V}(\hat{\tau}_{ps,sas}) = N^2 \frac{1-f}{n} \sum_{i=1}^{L} \frac{N_i}{N} s_i^2 \; + \; N^2 \frac{1-f}{n^2} \sum_{i=1}^{L} \left(1 - \frac{N_i}{N}\right) s_i^2$$

supondo que $n_i \geq 2$ para cada pós-estrato *i*.

Naturalmente, o respectivo estimador para a variância não condicional de $\hat{\mu}_{ps,sas}$ é, nas mesmas condições:

(3.5.38) $$\hat{V}(\hat{\mu}_{ps,sas}) = \hat{V}(\hat{\tau}_{ps,sas}) / N^2$$

Holt e Smith (1979) apresentam uma discussão detalhada sobre a abordagem condicional e não condicional. As suas conclusões são a favor da primeira quando se pretende fazer inferências depois da amostra ter sido retirada, e a favor da segunda apenas no momento da definição da amostra.

Little (1993, p. 1003) chama atenção para o facto de que a diferença entre (3.5.18) e 3.5.35 é de ordem n^{-2} e portanto não é crucial quando se consideram amostras grandes. Num estudo conduzido por Djerf (1997) verificou-se que a diferença entre as estimativas condicionadas e não condicionadas era de facto negligenciável. No entanto, com muitos pós-estratos e estimativas para sub-domínios da população, a diferença poderá não ser negligenciável.

Observe-se ainda que na expressão da variância não condicional (3.5.35), o primeiro termo corresponde à variância do estimador de τ para um plano de amostragem estratificada proporcional (c.f. resultado (2.5.29)). Assim, uma vez que o segundo termo é de ordem n^{-2}, podemos concluir que a SASSR com pós-estratificação é quase tão eficiente como a sondagem estratificada proporcional, quando a amostra é suficientemente grande[1].

É ainda de salientar que, o facto de S_i^2 surgir em ambas as parcelas da expressão da variância não condicional (3.5.35), nos leva mais uma vez a concluir que os pós-estratos devem ser o mais homogéneos possível.

3.5.2.1 Pós-estratos de dimensão inferior a dois

Observe-se que se $n_i=0$ para algum *i*, então nenhuma das variâncias pode ser aplicada directamente. Rao (1985) refere que se $n_i=1$ para algum *i*, então não é possível obter um estimador da variância centrado condicionalmente.

Se ocorrer $n_i=0$ para algum *i*, o estimador de pós-estratificação (3.5.14) reduz-se a:

$$(3.5.39) \qquad \hat{\mu}_{ps,sas} = \sum{}' \frac{N_i}{N} \bar{y}_i$$

onde $\sum'$ denota o somatório sobre todos os pós-estratos com $n_i \neq 0$.

Rao (1985, p. 21) refere que o estimador (3.5.39) é enviesado quer se considere a abordagem condicional quer a não condicional, podendo conduzir a uma séria subestimação do verdadeiro valor de μ.

[1] Cochran (1977, p. 134) refere que se deverá ter $n_i>20$, $\forall i$.

Um método muito utilizado para superar estas dificuldades consiste em agrupar estratos semelhantes (*method of collapsing strata*) por forma a garantir que $n_i>0$ para qualquer *i* no conjunto reduzido dos pós-estratos. Vários autores têm procurado desenvolver técnicas que permitam decidir quando e como se devem agrupar os estratos, veja-se por exemplo Little (1993) e os autores citados por Lazzeroni e Little (1998): Tremblay (1986) e Kalton e Maligalig (1991).

Fuller (1966) propõe vários estimadores, na perspectiva não condicional, para o caso particular em que se consideram apenas dois pós-estratos e ilustra um procedimento que permite generalizar esses resultados para qualquer número de estratos.

Rao (1985) faz uma análise bastante detalhada sobre o problema da existência de pós-estratos de dimensão amostral nula, apresentando diversos estimadores de pós-estratificação, entre os quais cita alguns estimadores propostos por Doss *et al.* (1979). Da discussão apresentada por Rao (1985) é de salientar a abordagem que passamos a expor.

Suponhamos que se dispõe de informação relativa a uma variável auxiliar X cujas médias de todos os pós-estratos na população, $\overline{X}_i$, são conhecidas e estão linearmente relacionadas com as respectivas médias $\overline{Y}_i$ da variável de interesse. Neste caso, é possível ajustar um modelo de regressão linear às médias observadas nos pós-estratos amostrais $\overline{y}_i$ e prever a média da população $\overline{Y}_i$ ($\overline{y}_i^*$) nos pós-estratos de dimensão nula:

(3.5.40) $$\overline{y}_i^* = \hat{\alpha} + \hat{\beta}\overline{X}_i$$

onde $\hat{\alpha}$ e $\hat{\beta}$ são estimados por mínimos quadrados ordinários.

Obtém-se assim outro estimador de pós-estratificação de μ:

(3.5.41) $$\hat{\mu}^*_{ps,sas} = \sum{}' \frac{N_i}{N}\overline{y}_i + \sum{}'' \frac{N_i}{N}\overline{y}_i^*$$

onde $\sum''$ denota o somatório sobre todos os pós-estratos com $n_i=0$.

Segundo Rao (1985, p. 22) este estimador deverá ter boas propriedades condicionais se o modelo de regressão ajustado for adequado.

3.5.2.2 Efeito de erros nas dimensões dos pós-estratos

Tal como sucede na sondagem aleatória estratificada (veja-se a secção 2.5.4), os estimadores de pós-estratificação correm o risco de enviesamento se o valor conhecido para a dimensão relativa de cada pós-estrato não for fiável.

Suponhamos então que os valores exactos das frequências associadas aos pós-estratos são N_i/N mas que se conhecem apenas as quantidades aproximadas N_{0i}/N_0. Então, o estimador (3.5.14) é dado por:

$$(3.5.42) \qquad \hat{\mu}_{ps,sas_0} = \sum_{i=1}^{L} \frac{N_{0i}}{N_0} \bar{y}_i$$

em vez de

$$(3.5.43) \qquad \hat{\mu}_{ps,sas} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$$

O enviesamento do estimador (3.5.14), $B(\hat{\mu}_{ps,sas})$, é facilmente determinado se atendermos a que:

$$(3.5.44) \qquad \hat{\mu}_{ps,sas_0} - \hat{\mu}_{ps,sas} = \sum_{i=1}^{L} \left(\frac{N_{0i}}{N_0} - \frac{N_i}{N} \right) \bar{y}_i$$

$$(3.5.45) \qquad E(\hat{\mu}_{ps,sas_0}) = \mu + \sum_{i=1}^{L} \left(\frac{N_{0i}}{N_0} - \frac{N_i}{N} \right) E(\bar{y}_i)$$

$$(3.5.46) \qquad B(\hat{\mu}_{ps,sas}) = \sum_{i=1}^{L} \left(\frac{N_{0i}}{N_0} - \frac{N_i}{N} \right) E(\bar{y}_i)$$

Este resultado permite-nos concluir que o enviesamento permanece constante à medida que a dimensão da amostra aumenta, perdendo-se assim o ganho de precisão da pós-estratificação que se obteria relativamente à SASSR quando a amostra é grande.

3.5.3 Sondagem aleatória estratificada

Seja *U* a população em estudo de dimensão *N* conhecida. Suponhamos que foi retirada de *U* uma amostra aleatória *s*, de dimensão *n*, através de um plano de sondagem aleatória estratificada, $s = (s_1, \ldots, s_h, \ldots, s_H)$, no qual foi utilizada a sondagem aleatória simples sem reposição (SASSR) em cada estrato.

Suponhamos que, na fase de estimação, se dispõe de informação auxiliar que permita dividir a amostra *s* em *L* pós-estratos, definidos por forma a que sejam o mais homogéneos possível. Como se referiu anteriormente, supõe-se que as dimensões dos pós-estratos na população são conhecidas.

Neste caso, os estratos iniciais podem cruzar os pós-estratos[1] e, portanto, as dimensões amostrais resultantes da intersecção dos estratos iniciais com os pós-estratos são aleatórias.

A pós-estratificação de uma amostra obtida através de uma sondagem estratificada pode colocar-se quando os estratos iniciais foram definidos, por exemplo, por razões de ordem operacional, tendo pouco poder explicativo, e existe uma variável auxiliar, fortemente relacionada com a variável de estudo *Y*, que permita efectuar a pós-estratificação.

Antes de apresentarmos o estimador de pós-estratificação genérico (3.5.1), para o plano de sondagem em análise, vamos considerar alguma notação adicional:

$N_{\bullet h}$ – dimensão do estrato inicial *h* na população (h = 1, …, H)

$N_{i\bullet}$ – dimensão do pós-estrato *i* na população (i = 1, …, L)

[1] Dependendo da forma como os estratos iniciais e os pós-estratos se relacionam, Särndal, Swensson e Wretman (1992) consideram quatro situações que podem ocorrer e apresentam algumas soluções na abordagem *model-assisted*.

$n_{\bullet h}$ – número de elementos da amostra pertencentes ao estrato inicial *h* (h = 1, …, H)

s_{ih} – conjunto de elementos da amostra que pertencem simultaneamente ao estrato inicial *h* (h = 1, …, H) e ao pós-estrato *i* (i = 1, …, L)

n_{ih} – dimensão (aleatória) de s_{ih}

Como vimos, o estimador de pós-estratificação genérico (3.5.1) pode ser escrito como:

$$(3.5.47) \qquad \hat{\tau}_{PS} = \sum_{i=1}^{L} \hat{\tau}_{iw}$$

onde,

$$(3.5.48) \qquad \hat{\tau}_{iw} = \frac{N_i}{\hat{N}_i} \sum_{k \in s_i} \frac{y_k}{\pi_k}$$

e

$$(3.5.49) \qquad \hat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k}$$

considerando-se que cada pós-estrato *i* (i=1, …, L) corresponde a um domínio na população.

Assim, com a devida adaptação da notação, pelo resultado (3.4.37) verifica-se que para uma sondagem aleatória estratificada (com SASSR em cada estrato) o estimador (3.5.48) é dado por:

$$(3.5.50) \qquad \hat{\tau}_{iw_{str}} = \frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} \sum_{k \in s_{ih}} y_k$$

onde,

$$(3.5.51) \qquad \hat{N}_{i\bullet} = \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} n_{ih}$$

E, portanto, o **estimador de pós-estratificação de τ**, para este plano de sondagem, é dado por:

$$(3.5.52) \qquad \hat{\tau}_{ps,str} = \sum_{i=1}^{L} \frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} \sum_{k \in s_{ih}} y_k$$

com $\hat{N}_{i\bullet}$ dado por (3.5.51).

Este estimador pode também ser escrito sob a forma

$$(3.5.53) \qquad \hat{\tau}_{ps,str} = \sum_{i=1}^{L} \sum_{h=1}^{H} \sum_{k \in s_{ih}} \frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \frac{N_{\bullet h}}{n_{\bullet h}} y_k$$

permitindo, portanto, evidenciar o ajustamento dos pesos iniciais pelo quociente $N_{i\bullet} / \hat{N}_{i\bullet}$. Assim, o algoritmo apresentado na secção 3.5.2, devidamente adaptado, fornece uma forma simples de implementar o estimador (3.5.52).

Naturalmente, uma vez que $\hat{\mu}_{ps} = \hat{\tau}_{ps} / N$, o **estimador de pós-estratificação de μ**, para este plano de sondagem, é dado por:

$$(3.5.54) \qquad \hat{\mu}_{ps,str} = \frac{1}{N} \sum_{i=1}^{L} \frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} \sum_{k \in s_{ih}} y_k$$

Rao (1985) apresenta um caso particular do estimador $\hat{\tau}_{ps,str}$ em que se consideram apenas H=2 estratos iniciais e L=2 pós-estratos, com o objectivo de ilustrar como é difícil investigar as propriedades condicionais do estimador (3.5.1) numa sondagem complexa. Mesmo para esta situação simples, Rao (1985) mostra que o valor esperado do estimador (3.5.1), i.e. o valor esperado do estimador (3.5.52),

condicionado sobre as dimensões amostrais observadas nos pós-estratos ($n_{1\bullet}$, $n_{2\bullet}$) não é tratável na abordagem condicional.

Williams (1962) sugeriu um estimador da variância do estimador de pós-estratificação genérico (3.5.1) que não revela boas propriedades na abordagem condicional, mesmo no caso em que o desenho da amostra corresponde a um plano SASSR, tal como demonstra Rao (1985). Este autor, propõe um estimador alternativo que, como veremos, pode ser preferível tanto na abordagem condicional, como não condicional.

Denote-se por $\hat{V}(\hat{\tau}_\pi) = \nu(y_k)$ a função que define o estimador da variância do estimador usual de τ. Ou seja, no caso da SASSR, $\nu(y_k)$ é dada por (c.f. secção 2.3.2):

$$(3.5.55) \qquad \nu(y_k) = N^2(1 - f)\frac{s^2}{n} = \hat{V}(\hat{\tau}_{\pi_{sas}})$$

onde,

$$(3.5.56) \qquad s^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \bar{y})^2$$

e, para a sondagem aleatória estratificada sem reposição, tem-se, com a devida adaptação de notação (c.f. secção 2.5.2):

$$(3.5.57) \qquad \nu(y_k) = \sum_{h=1}^{H} N_{\bullet h}^2\left(\frac{1}{n_{\bullet h}} - \frac{1}{N_{\bullet h}}\right)s_h^2 = \hat{V}(\hat{\tau}_{\pi_{str}})$$

onde,

$$(3.5.58) \qquad s_h^2 = \frac{1}{n_{\bullet h} - 1}\sum_{k\in s_h}(y_k - \bar{y}_h)^2$$

O estimador da variância de $\hat{\tau}_{PS}$ proposto por Rao (1985), para um plano de sondagem genérico, e que denotar-se-á por $\hat{V}_{rao}(\hat{\tau}_{ps})$, é dado por:

(3.5.59) $\hat{V}_{rao}(\hat{\tau}_{ps}) = \nu(z_k)$

onde, $\nu(z_k)$ se obtém a partir de $\hat{V}(\hat{\tau}_{\pi})$ substituindo-se y_k por:

(3.5.60) $z_k = \sum_i \frac{N_i}{\hat{N}_i}(y_{ik} - \frac{\hat{\tau}_i}{\hat{N}_i}\mathbb{I}_{k\in s_i}) = \sum_i \frac{N_i}{\hat{N}_i}(y_{ik} - \hat{\mu}_i\,\mathbb{I}_{k\in s_i})$

com,

(3.5.61) $\mathbb{I}_{k\in s_i} = \begin{cases} 1 & \text{se k pertence ao pós}-\text{estrato i} \\ 0 & \text{caso contrário} \end{cases}$

(3.5.62) $y_{ik} = y_k\,\mathbb{I}_{k\in s_i} = \begin{cases} y_k & \text{se k pertence ao pós}-\text{estrato i} \\ 0 & \text{caso contrário} \end{cases}$

e $\hat{\mu}_i = \hat{\tau}_i / \hat{N}_i$ sendo $\hat{\tau}_i$ e $\hat{N}_i$ dados, respectivamente, por (3.5.2) e (3.5.3).

Uma vez que as propriedades deste estimador são, também, difíceis de investigar, considere-se o caso mais simples do plano SASSR. Neste caso, tem-se:

(3.5.63) $\hat{\tau}_i = \sum_{k\in s_i} \frac{y_k}{\pi_k} = \sum_{k\in s_i} \frac{N}{n} y_k \;, \qquad i = 1, \ldots, L$

(3.5.64) $\hat{N}_i = \sum_{k\in s_i} \frac{1}{\pi_k} = \sum_{k\in s_i} \frac{N}{n} = n_i \frac{N}{n} \;, \qquad i = 1, \ldots, L$

e, portanto,

(3.5.65) $\hat{\mu}_i = \hat{\tau}_i / \hat{N}_i = \frac{1}{n_i} \sum_{k\in s_i} y_k \;, \qquad i = 1, \ldots, L$

Ou seja, $\hat{\mu}_i$ é a média amostral no pós-estrato *i* e denotar-se-á por $\bar{y}_{s_i}$.

Utilizando-se estes resultados e substituindo-se y_k por z_k (dado por (3.5.60)) na expressão do estimador da variância de $\hat{\tau}_{\pi_{sas}}$, i.e. em (3.5.55), conclui-se[1] que o estimador da variância de $\hat{\tau}_{ps,sas}$ proposto por Rao (1985) é

$$(3.5.66) \qquad \hat{V}_{rao}(\hat{\tau}_{ps,sas}) = (1-f)\sum_{i=1}^{L} N_i^2 \frac{n}{n-1}\frac{n_i-1}{n_i}\frac{s_i^2}{n_i}$$

onde, f = n/N e

$$(3.5.67) \qquad s_i^2 = \frac{1}{n_i-1}\sum_{k\in s_i}(y_k - \bar{y}_{s_i})^2$$

Observe-se que se utilizarmos a aproximação

$$(3.5.68) \qquad \frac{n}{n-1}\frac{n_i-1}{n_i} \approx 1$$

o estimador (3.5.66) vem dado por

$$(3.5.69) \qquad \hat{V}_{rao}(\hat{\tau}_{ps,sas}) = (1-f)\sum_{i=1}^{L} N_i^2 \frac{s_i^2}{n_i}$$

e, portanto, está aproximadamente de acordo com o estimador que se obteve na abordagem condicional, (3.5.20):

$$(3.5.70) \qquad \hat{V}(\hat{\tau}_{ps,sas} \mid \underset{\sim}{n}) = \sum_{i=1}^{L} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}$$

Note-se que se substituirmos $(1 - n_i/N_i)$ pelo seu valor médio $(1 - n/N)$ em (3.5.70), ou se ignorarmos estas correcções para população finita, obtém-se o estimador (3.5.69).

---

[1] A demonstração detalhada deste resultado encontra-se no Anexo 2, secção A2.4.

Conclui-se, assim, que o estimador (3.5.59) conduz a um estimador da variância condicionalmente válido, dadas as dimensões amostrais dos pós-estratos (Rao 1985, 1994). Särndal, Swensson e Wretman (1989, citados por Rao, 1994) justificam também o estimador (3.5.59) numa abordagem *model-assisted* adequada a planos de sondagem com uma etapa.

Assim, é de esperar que para planos de sondagem complexos o estimador (3.5.59) tenha também boas propriedades condicionais.

No caso da sondagem aleatória estratificada sem reposição, por (3.5.52) e (3.5.51), obtém-se, respectivamente:

$$(3.5.71) \qquad \hat{\tau}_i = \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} \sum_{k \in s_{ih}} y_k \ , \qquad i = 1, \ldots, L$$

$$(3.5.72) \qquad \hat{N}_{i\bullet} = \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} n_{ih} \ , \qquad i = 1, \ldots, L$$

e, portanto, denotando-se $\hat{\mu}_i$ por $\hat{\mu}_{i_{str}}$, no caso do plano de sondagem em análise, tem-se,

$$(3.5.73) \qquad \hat{\mu}_{i_{str}} = \frac{1}{\hat{N}_{i\bullet}} \sum_{h=1}^{H} \frac{N_{\bullet h}}{n_{\bullet h}} \sum_{k \in s_{ih}} y_k \ , \qquad i = 1, \ldots, L$$

Utilizando-se estes resultados e substituindo-se y_k por z_k (dado por (3.5.60)) na expressão do estimador da variância de $\hat{\tau}_{\pi_{str}}$, i.e. em (3.5.57), conclui-se[1] que o estimador da variância de $\hat{\tau}_{ps,str}$ proposto por Rao (1985) é dado por

$$(3.5.74) \qquad \hat{V}_{rao}(\hat{\tau}_{ps,str}) =$$

$$\sum_{i=1}^{L} \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \right)^2 \sum_{h=1}^{H} N_{\bullet h}^2 \left(\frac{1}{n_{\bullet h}} - \frac{1}{N_{\bullet h}} \right) \frac{1}{n_{\bullet h} - 1} \left[\sum_{k \in s_{ih}} (y_k - \bar{y}_{s_{ih}})^2 + n_{ih}(1 - n_{ih}/n_{\bullet h})(\bar{y}_{s_{ih}} - \hat{\mu}_{i_{str}})^2 \right]$$

[1] A demonstração detalhada deste resultado encontra-se no Anexo 2, secção A2.4.

onde $\hat{N}_{i\bullet}$ e $\hat{\mu}_{i_{str}}$ são dados, respectivamente, por (3.5.72) e (3.5.73) e $\bar{y}_{s_{ih}}$ é a média amostral dos elementos que pertencem simultaneamente ao estrato inicial *h* e ao pós-estrato *i*, ou seja,

(3.5.75) $$\bar{y}_{s_{ih}} = \frac{1}{n_{ih}} \sum_{k \in s_{ih}} y_k$$

Naturalmente, para estimar a variância dos estimadores é necessário que $n_{ih}>1$ em todos os pós-estratos *i* (i=1, …, L). Caso algum dos pós-estratos tenha dimensão inferior a dois, poder-se-á utilizar um dos métodos referidos na secção 3.5.2.1.

É de esperar que, para amostras grandes tais que os pós-estratos têm uma dimensão razoável em cada estrato inicial, o estimador (3.5.74) seja um bom estimador de $\hat{\tau}_{ps,str}$, na abordagem condicional.

Alternativamente ao estimador proposto por Rao (1985), poder-se-ão considerar métodos de re-amostragem (*resampling*), como o Bootstrap, para se estimar a variância de $\hat{\tau}_{ps,str}$. No entanto, como já foi referido, também esta área necessita ainda de alguma investigação teórica.

3.6 Estimação na presença de não respostas

Um problema da maioria das sondagens consiste na falta de obtenção, total ou parcial, de resposta aos questionários. A não resposta total (ou *unit nonresponse*) ocorre quando há ausência total de resposta ao questionário. Esta situação pode surgir, por exemplo, quando não é possível contactar a pessoa seleccionada para a amostra, ou quando esta se recusa a responder, ou quando se perdem questionários. A não resposta parcial (ou *item nonresponse*) ocorre quando há ausência de resposta apenas para uma parte do questionário.

Na presença de não respostas os estimadores usuais são enviesados, como se ilustra em seguida, através de um exemplo muito simples. Uma discussão detalhada sobre os efeitos estatísticos da não resposta pode ser obtida em Lessler e Kalsbeek (1992).

Suponhamos que a população U, de dimensão N, pode ser dividida em duas sub-populações: seja U_1 a sub-população, de dimensão N_1, correspondente aos elementos para os quais se obteria resposta se fossem seleccionados para a amostra; e seja U_0 a sub-população, de dimensão N_0, correspondente aos elementos de U para os quais não se obteria resposta se fossem seleccionados para a amostra.

Por uma questão de simplicidade, suponhamos que foi utilizado um plano SASSR para recolher a amostra *s* de dimensão $n = n_1 + n_0$; sendo n_1 o número (aleatório) de respondentes da amostra e n_0 o número de não respondentes. Denotando por μ_1 e μ_0 a média da população em U_1 e U_0, respectivamente, tem-se:

$$(3.6.1) \qquad \mu = \mu_1 + \mu_0 = \frac{N_1}{N}\mu_1 + \frac{N_0}{N}\mu_0$$

O estimador usual da média da população é a média amostral. Neste caso, a média dos não respondentes na amostra ($\bar{y}_0$) não é conhecida e, portanto, ao utilizar-se a média dos respondentes ($\bar{y}_1$) como estimador de μ, verifica-se que:

$$(3.6.2) \qquad E(\bar{y}_1) = E_{n_1}[E(\bar{y}_1 \mid n_1)] = E_{n_1}(\mu_1) = \mu_1$$

e que o enviesamento de $\bar{y}_1$ é dado por

$$(3.6.3) \qquad B(\bar{y}_1) = E(\bar{y}_1) - \mu = \mu_1 - \mu = \mu_1 - \left(\frac{N_1}{N}\mu_1 + \frac{N_0}{N}\mu_0\right) = \frac{N_0}{N}(\mu_1 - \mu_0)$$

Verifica-se, neste caso, que ainda que a taxa de não resposta (N_0/N) seja elevada, o enviesamento será pequeno se a média dos respondentes for próxima da média dos não respondentes. Naturalmente, uma vez que a amostra não fornece informação sobre μ_0 o enviesamento e, consequentemente, o erro quadrático médio, não podem ser estimados, a menos que haja outra fonte de informação.

Existem diversos métodos que permitem lidar com o problema da não resposta, tanto na fase de planeamento e recolha dos dados, como na fase de estimação. Como se referiu anteriormente, os métodos de pós-estratificação permitem, não só lidar com os problemas das bases de sondagem, mas também lidar com o problema das não respostas. A abordagem a outros métodos encontra-se fora do âmbito deste trabalho. Referências bibliográficas relevantes sobre os mesmos podem ser obtidas em Lessler e Kalsbeek (1992) e Azevedo (1999).

Os estimadores de pós-estratificação inserem-se numa classe de métodos de tratamento de não respostas usualmente designados por **métodos de recomposição** ou **métodos de ajustamento**. Estes procedimentos consistem em reponderar a amostra, i.e. ajustar os pesos de inclusão, por forma a que os pesos ajustados tenham em consideração as não respostas.

De um modo geral, estes métodos são utilizados no tratamento das não respostas totais. Podem também ser utilizados no caso das não respostas parciais apesar de exigirem mais trabalho, uma vez que é necessário calcular diferentes ponderadores para as diferentes variáveis de interesse.

Na secção 3.6.1 faz-se uma breve introdução às técnicas de ajustamento das não respostas e, na secção 3.6.2, apresenta-se o método de ajustamento por ponderação em classes. Este método está estreitamente relacionado com os métodos de pós-estratificação que se apresentam na secção 3.6.3.

3.6.1 Introdução aos métodos de ajustamento das não respostas

Considere-se, mais uma vez, a população U, de dimensão N, dividida em duas sub-populações, U_1 e U_0, nas condições definidas anteriormente; ou seja, U_1 corresponde à sub-população dos potenciais respondentes e U_0 corresponde à sub-população dos potenciais não respondentes. No que se segue, utiliza-se o índice *1* para designar os elementos respondentes (na população ou na amostra) e o índice *0* (zero) para designar os elementos não respondentes (na população ou na amostra).

Um conceito fundamental, para os métodos de ajustamento, é o de **probabilidade de resposta**, que se denota por p_k, e é dado por

$$p_k = P(\mathbb{I}_{k\in U_1} = 1) , \qquad k = 1, 2, \ldots, N \tag{3.6.4}$$

onde,

$$\mathbb{I}_{k\in U_1} = \begin{cases} 1 & \text{se } k \in U_1 \\ 0 & \text{se } k \in U_0 \end{cases} \tag{3.6.5}$$

Seja $w_k = 1/\pi_k$ o peso de inclusão, ou peso inicial, do elemento *k*. Na presença de não resposta, a amostra é constituída por $n_1<n$ elementos respondentes. Nestas condições, o estimador de Horvitz-Thompson

$$\hat{\tau}_{HT} = \sum_{k=1}^{n_1} w_k y_k \tag{3.6.6}$$

é enviesado. É possível obter-se um estimador centrado se o ponderador utilizado tiver em consideração a probabilidade de inclusão (π_k) e a probabilidade condicional de que o *k-ésimo* elemento torna-se respondente, se for seleccionado para a amostra, ou seja, quando o ponderador tem também em consideração a probabilidade de resposta ($p_k>0$, $\forall k=1,\ldots,N$) [Lessler e Kalsbeek 1992, p. 182]. Obtém-se, desta forma, o estimador centrado

$$\hat{\tau}^{*}_{HT} = \sum_{k=1}^{n_1} w^{*}_k y_k \tag{3.6.7}$$

onde, $w^*_k = 1/(\pi_k p_k)$.

Nargundkar e Joshi (1975, citados por Lessler e Kalsbeek 1992, p. 182) apresentam alguns aspectos teóricos deste estimador quando se supõe a ausência de outros erros não amostrais.

Os métodos de ajustamento das não respostas, na inferência clássica das sondagens (*design-based*), consistem então estabelecer estimadores que ajustam os pesos iniciais w_k, através de diferentes estimadores das probabilidades de resposta p_k (geralmente, desconhecidas). Para mais detalhes sobre os métodos de ajustamento, veja-se Little (1986) e Lessler e Kalsbeek (1992). Nas secções que se seguem, apresentam-se os métodos de ponderação em classes e de pós-estratificação.

3.6.2 Método de ajustamento por ponderação em classes

O método de ajustamento por ponderação em classes consiste em estimar as probabilidades de resposta através da divisão da amostra obtida (incluindo respondentes e não respondentes) em *H* subconjuntos mutuamente exclusivos e exaustivos, designados **classes** ou **células de ajustamento**. Assume-se que, em cada célula *h* (h=1,…,H), os elementos têm valores semelhantes para a variável de interesse *Y* e que todas as probabilidades de resposta são iguais (Lessler e Kalsbeek, 1992, p. 183).

Little (1986) analisa comparativamente alguns métodos de ajustamento da não resposta e considera alguns critérios de escolha para as células do ajustamento em classes. Lessler e Kalsbeek (1992, p. 188) referem que se o plano de sondagem for multi-etápico, é usual escolherem-se as unidades amostrais das primeiras etapas para definir as classes; no caso de uma sondagem aleatória estratificada, é usual utilizarem-se as variáveis de estratificação. Estes autores referem ainda que, idealmente, as variáveis que definem as células devem estar fortemente associadas à variável de interesse, mas não devem estar mutuamente associadas.

Seja $s_h = s_{1h} \cup s_{0h}$ o conjunto de elementos da amostra pertencentes à *h-ésima* célula de ajustamento (de dimensão amostral n_h); onde, s_{1h} é o subconjunto de s_h

correspondente aos elementos respondentes (de dimensão n_{1h}) e s_{0h} é o subconjunto constituído pelos elementos não respondentes (de dimensão n_{0h}).

Sejam ainda w_{hk} e p_{hk}, respectivamente, o peso de inclusão e a probabilidade de resposta do *k-ésimo* elemento da *h-ésima* célula de ajustamento.

O estimador genérico de p_{hk}, utilizado neste método, é dado por:

$$(3.6.8) \qquad \hat{p}_{hk} = \frac{\sum_{k=1}^{n_{1h}} w_{hk}}{\sum_{k=1}^{n_h} w_{hk}}, \qquad k \in s_h,\ h = 1, \ldots, H$$

O ponderador ajustado, a utilizar nos estimadores por ponderação em classes, é então

$$(3.6.9) \qquad w_{hk}^{(pc)} = \frac{\sum_{k=1}^{n_h} w_{hk}}{\sum_{k=1}^{n_{1h}} w_{hk}} w_{hk}, \qquad k \in s_h,\ h = 1, \ldots, H$$

Considerando a notação apresentada nas secções anteriores, o ponderador ajustado (3.6.9) pode ser escrito como

$$(3.6.10) \qquad w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}, \qquad k \in s_h,\ h = 1, \ldots, H$$

onde,

$$(3.5.11) \qquad \hat{N}_h = \sum_{k=1}^{n_h} w_{hk}$$

$$(3.5.12) \qquad \hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk}$$

Esta notação permite observar que o estimador de p_{hk} dado por (3.6.8) não é mais do que um estimador da proporção de respondentes na população, dentro da célula *h* (N_{1h}/N_h).

Um **estimador do total da população por ponderação em classes** é

$$(3.6.13) \qquad \hat{\tau}_{pc} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} w_{hk}^{(pc)} y_{hk} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} y_{hk}$$

onde, $\hat{N}_h$ e $\hat{N}_{1h}$ são dados, respectivamente, por (3.6.11) e (3.6.12) e y_{hk} é o valor da variável de interesse para o elemento *k* da *h-ésima* célula de ajustamento.

Um **estimador da média da população por ponderação em classes** é dado por

$$(3.6.14) \qquad \hat{\mu}_{pc} = \hat{\tau}_{pc} / N$$

Outro estimador da média da população é

$$(3.6.15) \qquad \hat{\mu}_{pc} = \hat{\tau}_{pc} / \hat{N}$$

onde,

$$(3.6.16) \qquad \hat{N} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} w_{hk}^{(pc)}$$

Demonstra-se facilmente que $\hat{N}$ dado por (3.6.16) pode ser obtido de forma equivalente através de

$$(3.6.17) \qquad \hat{N} = \sum_{h=1}^{H}\sum_{k=1}^{n_h} w_{hk}$$

como se verifica em seguida:

$$\hat{N} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} w_{hk}^{(pc)} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} = \sum_{h=1}^{H} \frac{\hat{N}_h}{\hat{N}_{1h}} \sum_{k=1}^{n_{1h}} w_{hk} =$$

$$= \sum_{h=1}^{H} \frac{\hat{N}_h}{\hat{N}_{1h}} \hat{N}_{1h} = \sum_{h=1}^{H}\sum_{k=1}^{n_h} w_{hk}$$

No caso da sondagem aleatória simples sem reposição e da sondagem aleatória estratificada sem reposição, o estimador (3.6.15) reduz-se ao caso (3.6.14) (veja-se a secção 3.3.1.1). Para planos de sondagem com probabilidades desiguais, tal poderá não suceder.

3.6.3 Métodos de ajustamento por pós-estratificação

Utilizando-se a notação apresentada anteriormente, suponhamos que a amostra pode ser pós-estratificada em *L* pós-estratos e se conhecem as dimensões $N_1,\ldots,N_i,\ldots,N_L$ dos pós-estratos na população.

Na secção 3.5.1 apresentou-se o estimador de pós-estratificação genérico com o intuito de lidar com os problemas da base de sondagem, na ausência de não respostas. Neste caso, o peso inicial, w_{ik}, do elemento *k* pertencente ao pós-estrato *i*, era ajustado por $N_i/\hat{N}_i$, com

$$(3.6.18) \qquad \hat{N}_i = \sum_{k=1}^{n_i} w_{ik}\,, \qquad i = 1, \ldots, L$$

Quando se pretende lidar simultaneamente com os erros da base de sondagem e com o enviesamento provocado pela presença de não respostas, podem-se combinar os métodos de pós-estratificação e de ponderação em classes.

Uma forma de combinar esses dois métodos consiste em assumir-se que os pós-estratos correspondem exactamente às células de ajustamento (do método de ponderação em classes). Obtém-se, assim, um ponderador ajustado dado por

$$(3.6.19) \qquad w_{ik}^{(ps)} = \frac{N_i}{\hat{N}_i} w_{ik}^{(pc)} = \frac{N_i}{\hat{N}_i} \frac{\hat{N}_i}{\hat{N}_{1j}} w_{ik} = \frac{N_i}{\hat{N}_{1j}} w_{ik}, \qquad k \in s_i,\ i = 1, \ldots, L$$

com, $\hat{N}_{1j}$ dado por (3.6.12), ou seja,

$$(3.6.20) \qquad \hat{N}_{1j} = \sum_{k=1}^{n_{1j}} w_{ik}$$

Lessler e Kalsbeek (1992, p. 184) apresentam outro ponderador ajustado para o método de pós-estratificação:

$$(3.6.21) \qquad w_{ik}^{*} = \frac{\hat{N}}{N} \frac{N_i}{\hat{N}_{1j}} w_{ik}, \qquad k \in s_i,\ i = 1, \ldots, L$$

com,

$$(3.6.22) \qquad \hat{N} = \sum_{i=1}^{L} \sum_{k=1}^{n_i} w_{ik}$$

que, como já foi referido, no caso dos planos SASSR e sondagem aleatória estratificada sem reposição se reduz ao ponderador $w_{ik}^{(ps)}$ dado por (3.6.19). Utilizando-se estes pesos ajustados, um **estimador de pós-estratificação do total da população**, na presença de não respostas, é dado por

$$(3.6.23) \qquad \hat{\tau}_{ps} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1j}} w_{ik}^{(ps)} y_{ik} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1j}} \frac{N_i}{\hat{N}_{1j}} w_{ik} y_{ik}$$

onde, $\hat{N}_{1j}$ é dado por (3.6.20) e y_{ik} é o valor da variável de interesse para o elemento *k* do *i-ésimo* pós-estrato (célula de ajustamento).

Naturalmente, um **estimador de pós-estratificação da média da população**, na presença de não respostas, é

$$(3.6.24) \qquad \hat{\mu}_{ps} = \hat{\tau}_{ps} / N$$

Observe-se que os estimadores por ponderação em classes requerem que os pesos iniciais sejam conhecidos, tanto para os respondentes, como para os não respondentes, em cada célula. Os estimadores de pós-estratificação requerem apenas esse conhecimento ao nível dos respondentes. Por outro lado, nestes métodos, a dimensão dos pós-estratos na população, N_i, tem que ser conhecida.

Até ao momento, assumiu-se que as células de ajustamento da não resposta correspondem exactamente aos pós-estratos. No entanto, uma das abordagens mais utilizadas, para lidar com a não resposta total, consiste em obter os ponderadores ajustados pelo método de ajustamento em classes e, em seguida, ajustar esses ponderadores através da pós-estratificação. Ou seja, usualmente, as células de ajustamento são definidas separadamente para cada um dos métodos de ajustamento.

Assim, o primeiro passo consiste em ajustar os pesos iniciais nas células de ajustamento *h* do método de ponderação em classes, através de (3.6.10):

$$w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}, \qquad k \in s_h,\ h = 1, \ldots, H \tag{3.6.25}$$

com, $\hat{N}_h$ e $\hat{N}_{1h}$ definidos, respectivamente, por (3.6.11) e (3.6.12); e w_{hk} o peso inicial do indivíduo *k* pertencente à *h-ésima* célula de ajustamento da não resposta.

Em seguida, estes ponderadores são ajustados novamente através da pós-estratificação da amostra:

$$w_{ik}^{(pc,ps)} = \frac{N_i}{\hat{N}_{1i}^{*}} w_{ik}^{(pc)}, \qquad k \in s_i,\ i = 1, \ldots, L \tag{3.6.26}$$

onde, $w_{ik}^{(pc)}$ é o peso ajustado por ponderação em classes, (3.6.25), do elemento *k* pertencente ao pós-estrato *i*; e $\hat{N}_{1i}^{*}$ é agora dado por

$$\hat{N}_{1i}^{*} = \sum_{k=1}^{n_{1i}} w_{ik}^{(pc)} \tag{3.6.27}$$

Um **estimador de pós-estratificação do total da população**, com **ajustamento da não resposta por ponderação em classes**, é dado por

$$(3.6.28) \qquad \hat{\tau}_{pc,ps} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} w_{ik}^{(pc,ps)} y_{ik}$$

onde, $w_{ik}^{(pc,ps)}$ é o ponderador ajustado definido por (3.6.26).

Para planos de sondagem complexos, os estimadores dos métodos de ajustamento por ponderação em classes e por pós-estratificação são difíceis de analisar. Assim, apresenta-se em seguida uma breve discussão das propriedades desses estimadores para um plano de sondagem aleatória simples sem reposição.

3.6.4 Sondagem aleatória simples sem reposição

Suponhamos que foi recolhida uma amostra aleatória de dimensão *n* através de um plano SASSR. Neste caso, os pesos iniciais são iguais para todos os elementos da população: $w_k = N/n$ (k=1,…,N).

A forma genérica do estimador do total da população por ponderação em classes, é dada por (3.6.13):

$$(3.6.29) \qquad \hat{\tau}_{pc} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} y_{hk}$$

onde, w_{hk} é o peso inicial do elemento *k* pertencente à *h-ésima* célula de ajustamento. Para a SASSR, tem-se

$$(3.6.30) \qquad \hat{N}_h = \sum_{k=1}^{n_h} w_{hk} = \sum_{k=1}^{n_h} \frac{N}{n} = n_h \frac{N}{n}$$

$$(3.6.31) \qquad \hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk} = \sum_{k=1}^{n_{1h}} \frac{N}{n} = n_{1h} \frac{N}{n}$$

Assim, para o plano de sondagem em apreço, um estimador do total da população por ponderação em classes é dado por

$$(3.6.32) \qquad \hat{\tau}_{pc,sas} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{n_h \frac{N}{n}}{n_{1h}\frac{N}{n}} \frac{N}{n} y_{hk} = N \sum_{h=1}^{H} \frac{n_h}{n} \frac{1}{n_{1h}} \sum_{k=1}^{n_{1h}} y_{hk} =$$

$$= N \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{1h}$$

onde, $\bar{y}_{1h}$ é a média dos respondentes na *h-ésima* célula de ajustamento.

Um estimador da média da população por ponderação em classes é, naturalmente,

$$(3.6.33) \qquad \hat{\mu}_{pc,sas} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{1h}$$

Relativamente aos estimadores de ajustamento por pós-estratificação, no caso em que as células de ajustamento da não resposta são as mesmas que os pós-estratos, o estimador genérico apresentado em (3.6.23) é

$$(3.6.34) \qquad \hat{\tau}_{ps} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} \frac{N_i}{\hat{N}_{1i}} w_{ik} y_{ik}$$

onde, w_{ik} é o peso inicial do elemento *k* pertencente ao pós-estrato *i*; e, para a SASSR, tem-se

$$(3.6.35) \qquad \hat{N}_{1i} = \sum_{k=1}^{n_{1i}} w_{ik} = \sum_{k=1}^{n_{1i}} \frac{N}{n} = n_{1i} \frac{N}{n}$$

Desta forma, para o plano de sondagem em análise, um estimador do total da população por pós-estratificação é dado por

$$(3.6.36) \qquad \hat{\tau}_{ps,sas} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} \frac{N_i}{n_{1i}\frac{N}{n}} \frac{N}{n} y_{ik} = \sum_{i=1}^{L} N_i \frac{1}{n_{1i}} \sum_{k=1}^{n_{1i}} y_{ik} =$$

$$= \sum_{i=1}^{L} N_i \bar{y}_{1i}$$

onde, $\bar{y}_{1i}$ é a média dos respondentes no *i-ésimo* pós-estrato (célula de ajustamento).

Relativamente ao estimador da média da população, tem-se, neste caso,

(3.6.37) $$\hat{\mu}_{ps,sas} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_{1i}$$

O enviesamento e a variância dos estimadores da média da população, por ponderação em classes e por pós-estratificação, para um plano SASSR (dados, respectivamente, por (3.6.33) e (3.6.37)), foram considerados por Thomsen (1973, 1978, citado por Little 1986) e por Oh e Scheuren (1983, citados por Little 1986). No entanto, as comparações entre estes trabalhos não são imediatas uma vez que os pressupostos assumidos são diferentes, assim como a distribuição de referência.

Para deixar clara a diferença de abordagens, considere-se a seguinte notação: seja $\mathbf{y} = (y_1,\ldots,y_N)$ o vector dos valores da variável *Y* na população; $\mathbf{r} = (r_1,\ldots,r_N)$ o vector das variáveis indicatriz tais que $r_k = 1$ se o elemento *k* é respondente se for seleccionado para a amostra e $r_k = 0$, caso contrário; $\mathbf{s} = (s_1,\ldots,s_N)$ o vector das variáveis de Cornfield, ou seja, $s_k = 1$ se o elemento *k* pertencer à amostra e $s_k = 0$, caso contrário; $\mathbf{n} = (n_1,\ldots,n_H)$ o vector das dimensões da amostra nas *H* células de ajustamento; e $\mathbf{n_r} = (n_{1r},\ldots,n_{Hr})$ o vector das dimensões da amostra de respondentes nas células de ajustamento.

Little (1986) refere que Thomsen calculou o enviesamento e a variância de $\hat{\mu}_{pc,sas}$ e de $\hat{\mu}_{ps,sas}$ sob a distribuição de **s**, com **y** e **r** fixos; e Oh e Scheuren efectuaram os cálculos sobre a distribuição de **r** e **s** com (i) **y** fixo e (ii) **y**, **n** e $\mathbf{n_r}$ fixos.

Little (1986) apresenta vários argumentos que o levam a propor uma abordagem condicional sobre **y**, **r**, **n** e $\mathbf{n_r}$, por forma a que possam ser consideradas várias formas de definir as classes de ajustamento. Nestas condições, as expressões do

enviesamento e da variância do estimador de ponderação em classes, (3.6.33),são, respectivamente:

$$(3.6.38) \qquad B(\hat{\mu}_{pc,sas}) = \sum_{h=1}^{H}\left(\frac{n_h}{n} - \frac{N_h}{N}\right)\mu_{1h} + \sum_{h=1}^{H}\frac{N_h}{N}(\mu_{1h} - \mu_h)$$

$$(3.6.39) \qquad V(\hat{\mu}_{pc,sas}) = \sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2\left(1 - \frac{n_{1h}}{N_h}\right)\frac{1}{n_{1h}}S_{1h}^2$$

e, para o estimador de pós-estratificação (3.6.37), tem-se

$$(3.6.40) \qquad B(\hat{\mu}_{ps,sas}) = \sum_{i=1}^{L}\frac{N_i}{N}(\mu_{1i} - \mu_i)$$

$$(3.6.41) \qquad V(\hat{\mu}_{ps,sas}) = \sum_{i=1}^{L}\left(\frac{N_i}{N}\right)^2\left(1 - \frac{n_{1i}}{N_i}\right)\frac{1}{n_{1i}}S_{1i}^2$$

onde, μ_{1h} [ou μ_{1i}] é a média dos potenciais respondentes na população que pertencem à classe de ajustamento *h* [pós-estrato *i*] e S_{1h}^2 [ou S_{1i}^2] é a variância corrigida na população na *h-ésima* classe de ajustamento [*i-ésimo* pós-estrato].

Uma vez que o estimador $\hat{\mu}_{pc,sas}$ não está definido se ocorrer $n_i > 0$ e $n_{1i} = 0$ para alguma das células de ajustamento, assume-se ainda que tal não ocorre nos cálculos condicionais sobre **n** e $\mathbf{n_r}$.

Observe-se que à medida que a proporção n_h/n converge para a proporção análoga na população, N_h/N, o primeiro termo do enviesamento de $\hat{\mu}_{pc,sas}$ tende para zero; e, a segunda parcela será nula se $\mu_{1h} = \mu_h \ \forall h$.

Relativamente ao estimador $\hat{\mu}_{ps,sas}$, basta que esta última condição se verifique para que o enviesamento seja nulo (relembre-se que o estimador de pós-estratificação considerado pressupõe que as classes de ajustamento e os

pós-estratos sejam idênticos, pelo que a segunda parcela da expressão de $B(\hat{\mu}_{pc,sas})$ corresponde exactamente ao enviesamento de $\hat{\mu}_{ps,sas}$).

O estudo por simulação conduzido por Little (1986) sugere que, nesta abordagem, o estimador de pós-estratificação deverá ter erro quadrático médio inferior ao do estimador por ponderação em classes. Esta conclusão está de acordo com as que Holt e Smith (1979) obtiveram numa análise semelhante para o caso de ausência de não respostas.

Por outro lado, na abordagem considerada por Kalton (1983, citado por Lessler e Kalsbeek 1992) o enviesamento dos dois estimadores é igual e tem-se $V(\hat{\mu}_{ps,sas}) < V(\hat{\mu}_{pc,sas})$, pelo que o erro quadrático médio de $\hat{\mu}_{ps,sas}$ é menor do que o de $\hat{\mu}_{pc,sas}$.

4 Aplicações práticas

4.1 Introdução

Neste capítulo apresentam-se algumas aplicações das técnicas de pós-estratificação aos dados do Inquérito às Empresas / Harmonizado de 1996 (IEH96), conduzido pelo Instituto Nacional de Estatística (INE).

O desenho do IEH corresponde a um plano de amostragem aleatória estratificada sem reposição. A base de amostragem é constituída a partir do Ficheiro Geral de Unidades Estatísticas (FGUE) do INE.

A aplicação dos métodos de pós-estratificação ao IEH é essencialmente motivada pelo problema das *mudanças de estrato*. As respostas obtidas no inquérito sugerem que determinadas empresas não se mantêm nos estratos iniciais. Este problema resulta da informação auxiliar que consta do FGUE, e que serviu de base à estratificação, se encontrar desactualizada ou incorrecta. Por outro lado, o IEH96 apresenta também não resposta total. Como foi referido anteriormente, os problemas da base de sondagem e das não respostas têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram.

Os estimadores de pós-estratificação ajustam o coeficiente de extrapolação de cada elemento da amostra, por forma a que esta reflicta a estrutura actual da população e tenha também em conta a ocorrência de não respostas. Como consequência, espera-se que, com estes métodos, seja possível melhorar as estimativas das diversas variáveis de interesse.

Na secção que se segue faz-se uma introdução ao Inquérito às Empresas / Harmonizado, onde se apresenta a respectiva metodologia, alguns dados referentes ao inquérito de 1996 e se indicam as variáveis de interesse. Na secção 4.3 apresentam-se os resultados obtidos em dois exemplos práticos da aplicação, aos dados do IEH96, dos métodos de ajustamento das não respostas e do método de Bootstrap BWO proposto por Sitter (1992b).

4.2 Inquérito às Empresas / Harmonizado (IEH)

O Instituto Nacional de Estatística (1997, p. 1) apresenta claramente os objectivos do Inquérito às Empresas / Harmonizado (IEH):

> "*O Inquérito às Empresas / Harmonizado tem como principal objectivo estabelecer um quadro comum de recolha, compilação e transmissão de dados sobre a estrutura e actividade das empresas.*
>
> *Pretende-se disponibilizar informação estatística que permita designadamente analisar:*
>
> - *A estrutura e evolução da actividade das empresas*
> - *Os factores de produção utilizados e outros elementos que permitam medir a actividade, os resultados e a competitividade das empresas*
> - *O desenvolvimento regional, nacional, comunitário e internacional das empresas*
> - *As relações com os mercados externos*
> - *As características das pequenas e médias empresas*
> - *As particularidades das empresas face às especificidades dos sectores em que desenvolvem a sua actividade*".

O inquérito tem cobertura nacional e é realizado anualmente por via postal. Em alguns casos, o envio postal é complementado com entrevista telefónica e/ou recolha directa. A unidade estatística de observação é a empresa.

O Ficheiro Geral de Unidades Estatísticas (FGUE), do INE, é o instrumento base para constituição do universo de referência e selecção da amostra do IEH.

4.2.1 Especificações metodológicas

O universo de referência do IEH inclui as empresas que, simultaneamente, respeitem um determinado conjunto de regras, das quais se destacam:

- empresas que, de acordo com a Classificação Portuguesa das Actividades Económicas CAE - REV. 2, se encontram classificadas com actividade principal nas Secções A, B, C, D, E, F, G, H, I, K, M, N, O (veja-se o Anexo 3, secção A3.1). São também consideradas as empresas que, em termos de

CAE - Rev. 2, desenvolvem actividades secundárias no âmbito da secção D - Indústrias transformadoras;

- empresas cuja data de constituição é inferior ou igual ao ano do inquérito;
- empresas com localização da sede no Continente e nas regiões Autónomas dos Açores e da Madeira;
- empresas em actividade ou com actividade sazonal;
- do universo do IEH devem ser excluídas as empresas que, simultaneamente, apresentem zero pessoas ao serviço e ausência de volume de vendas.

O universo é estratificado pelos escalões definidos pelas seguintes variáveis (a descrição dos escalões das variáveis de estratificação encontra-se no Anexo 4):

- *ENUT* – Escalões de NUTS II (Nomenclatura das Unidades Territoriais para Fins Estatísticos)
- *ECAE* – Escalões de Classificação Portuguesa das Actividades Económicas CAE - REV. 2
- *ENPS* – Escalões de número de pessoas ao serviço
- *EFJR* – Escalões de forma jurídica
- *EVVN* – Escalões de volume de vendas

O Instituto Nacional de Estatística (1997, p. 37) refere que "*a variável* Volume de vendas *não pode ser utilizada como variável de apuramento. A sua inclusão como variável de estrato visou, apenas, assegurar uma maior homogeneidade entre as empresas dos diversos estratos do universo*".

O inquérito é realizado por amostragem e de forma exaustiva de acordo segundo os seguintes critérios:

- Amostragem – unidades estatísticas com menos de 100 pessoas ao serviço
- Exaustivo – unidades estatísticas com 100 e mais pessoas ao serviço

Os instrumentos de notação do IEH (veja-se o Anexo 6) são compostos por um *Módulo Comum* – Modelos A e B – e por *Anexos Específicos* (Quadro 4.2.1).

Quadro 4.2.1 – Instrumentos de notação do Inquérito às Empresas / Harmonizado

	Modelo A	Modelo B
Módulo Comum	Unidades estatísticas com EFJR = 1, 2, 3 e ENPS ≥ 3	Unidades estatísticas com EFJR = 1, 2, 3 e ENPS ≤ 2
Anexos Específicos	*Indústria* *Construção* Empresas com 20 e mais pessoas ao serviço Empresas com 100 e mais pessoas ao serviço *Comércio* *Educação* *Saúde*	*Indústria* *Construção* *Comércio* *Educação* *Saúde*

4.2.1.1 Condições de apuramento

Para efeitos de apuramento em termos de variáveis de estrato considera-se sempre a *situação inicial* da empresa. A *situação inicial* refere-se à classificação sobre a qual recaiu a selecção da amostra, constante do Ficheiro de Lançamento, correspondente à informação existente no FGUE. A *situação final* deriva da avaliação da resposta da empresa ao Inquérito, podendo ou não coincidir com a inicial.

A passagem de uma empresa para apuramento está condicionada por três parâmetros:

- *CSV* – Código de situação de Instrumento de Notação (Quadro 4.2.2)
- *STA* – Código de situação da empresa perante a actividade (Quadro 4.2.3)
- *Número de meses de actividade* (Quadro 4.2.4)

Quadro 4.2.2 – Código de situação de Instrumento de Notação (CSV)

CSV	Descrição
0	Não lançado
1	Lançado não recebido
2	Recebido
3	Recebido pendente
4	Registado com erros fatais
5	Registado com erros de aviso
6	Registado correcto

De acordo com o Quadro 4.2.2, são passíveis de apuramento os Instrumentos de Notação que, após crítica, registo e validação, se apresentem com CSV = 5 e 6. São apuráveis com tratamento de não respostas os Instrumentos de Notação com CSV = 1, 2, 3 e 4. As empresas com CSV = 2 só devem ser consideradas para efeitos de resultados antecipados.

Quadro 4.2.3 – Código de situação da empresa perante a actividade (STA)

STA	**Descrição**
00	Situação indefinida
01	Aguardando início de actividade
02	Em actividade
03	Actividade suspensa
04	Cessação definitiva por outras razões
05	Cessação definitiva por dissolução ou extinção
06	Cessação definitiva por incorporação
07	Cessação definitiva por fusão
08	Pendente / Inquirição suspensa
09	Actividade sazonal
10	Pendente / Inquirição suspensa / CTT
99	Empresas que estão fora do âmbito do inquérito (mudança de actividade - CAE)

Para efeitos de apuramento (em termos de valores resposta, valores a zero ou não resposta) é considerada a *situação final* da empresa perante a actividade. As condições de apuramento relativas à situação da empresa perante a actividade (*STA*) encontram-se resumidas no Quadro 4.2.6.

Quadro 4.2.4 – Número de meses de actividade

Código	**Descrição**
0	Sem significado
1	Com significado

"*Não se define à partida um número mínimo de funcionamento. Sempre que o número de meses em actividade, indicado pela empresa, se apresente como insuficiente, será feita uma análise casuística, em função da coerência global da resposta, cruzada com as especificidades do sector onde a empresa desenvolve a sua actividade*" (Instituto Nacional de Estatística 1997, p. 42).

O código “0” (zero) equivale ao apuramento com valores iguais a zero e só se aplica às empresas com situação final perante a actividade STA = 02 (em actividade).

A passagem da empresa para apuramento é definida de acordo com o Quadro 4.2.5. No Quadro 4.2.6 apresenta-se um resumo das condições e situação de apuramento.

Quadro 4.2.5 – Situação de apuramento (*SA*)

SA	Descrição
0	Não apurável
1	Apurável com valores resposta
2	Apurável com valores a zero
3	Apurável com tratamento de não respostas

Quadro 4.2.6 – Resumo das condições e situação de apuramento

CSV	*STA*	Apuramento	*SA*
1, 2, 3, 4	00	Tratamento de não respostas(*)	3
5 / 6	08 e 99	Tratamento de não respostas(*)	3
	02 e 09	Apurado com valores de registo	1
	01 03 04 05 06 07 10	Apurado com valores a zero	2

(*) O método de tratamento de não respostas utilizado pelo INE é o "Hot-Deck - Imputação aleatória dentro do estrato[1]".

4.2.2 Alguns dados do IEH96

O Departamento de Estatísticas das Empresas do INE forneceu dois ficheiros, com dados relativos ao Inquérito às Empresas / Harmonizado de 1996 (IEH96): um ficheiro com as respostas de algumas variáveis do inquérito (com 84519 registos) e

[1] Este método traduz-se na substituição de cada valor em falta por um valor escolhido aleatoriamente entre o conjunto dos respondentes do estrato.

um ficheiro com as dimensões dos estratos na população e na amostra (com 13191 registos). Os dados destes ficheiros encontram-se protegidos pela *Lei do Segredo Estatístico*, pelo que não é possível apresentar um excerto dos mesmos.

Apesar do ficheiro com as respostas ao IEH96 conter uma variável referente aos coeficientes de extrapolação, procedeu-se novamente ao cálculo desses valores, através da combinação dos dois ficheiros, por forma a evitarem-se erros de arredondamento.

Designem-se por *Pequenas e médias empresas* as empresas consideradas para inquirição com recurso à teoria de amostragem e por *Grandes empresas* as empresas consideradas para inquirição exaustiva (vejam-se os respectivos critérios na secção 4.2.1). No Quadro 4.2.7 apresentam-se as dimensões da população e da amostra e o número de estratos referentes às *Pequenas e médias empresas* e às *Grandes empresas*.

Quadro 4.2.7 – Dimensões do universo e da amostra e número de estratos, por dimensão da empresa

Dimensão da empresa	Universo	Amostra	Número de estratos
Pequenas e médias empresas	736353	66937	5411
Grandes empresas	17582	17582	7780

No presente estudo consideram-se apenas as empresas do Continente inquiridas por amostragem, ou seja, as *Pequenas e médias empresas* do Continente (Quadro 4.2.8). É de salientar que, neste caso, não existem estratos de dimensão inferior a 2, na amostra.

Quadro 4.2.8 – Dimensões do universo e da amostra e número de estratos no Continente, por dimensão da empresa

Dimensão da empresa	Universo	Amostra	Número de estratos
Pequenas e médias empresas	712642	62846	4927
Grandes empresas	14882	14882	6418

Relativamente à situação de apuramento (*SA*), verifica-se que cerca de 29% das empresas da amostra (*Pequenas e médias empresas* do Continente) foram apuradas para efeitos de tratamento de não respostas (veja-se Quadro 4.2.9), segundo os critérios definidos pelo INE na metodologia do IEH (c.f. secção 4.2.1.1).

Quadro 4.2.9 – Situação de apuramento (*SA*) das empresas da amostra (*Pequenas e médias empresas* do Continente)

SA	Número de empresas	Percentagem
1	35817	57.0
2	9016	14.3
3	18013	28.7

Eliminando-se da amostra (*Pequenas e médias empresas* do Continente) as empresas em que se verifica não resposta total (*SA=3*), verifica-se que a amostra, constituída apenas pelas empresas consideradas respondentes, tem dimensão 44833.

4.2.3 Variáveis de estudo

As variáveis de interesse sobre as quais incidiu o estudo são comuns aos dois módulos do questionário e são as seguintes:

- *Q20001* – número médio de pessoas ao serviço – total (remunerado e não remunerado)
- *Q4160* – vendas
- *Q4190* – prestações de serviços

Para estas variáveis não se verifica a ocorrência de não respostas parciais (quando se eliminam da amostra as empresas identificadas com não resposta total).

4.3 Apresentação dos resultados

Foram realizados exemplos práticos[1] da aplicação, aos dados do IEH96, dos métodos de ajustamento das não respostas (secção 3.6) e do método de Bootstrap BWO proposto por Sitter (1992b), apresentado na secção 2.7.

Os estimadores considerados foram: o estimador de ponderação em classes, o estimador de pós-estratificação e o estimador de pós-estratificação com ajustamento das não respostas por ponderação em classes.

Na secção que segue, apresenta-se em mais detalhe a metodologia utilizada nos exemplos práticos, cujos resultados são analisados nas secções 4.3.2 e 4.3.3.

4.3.1 Metodologia dos exemplos práticos

No método de ajustamento por ponderação em classes consideraram-se os estratos iniciais como sendo as células de ajustamento das não respostas. Supõe-se então que, em cada estrato, os elementos têm valores semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais.

O estimador do total da população por ponderação em classes utilizado foi:

$$(4.3.1) \qquad \hat{\tau}_{pc} = \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} w_{hk}^{(pc)} y_{hk}$$

onde,

$$(4.3.2) \qquad w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}\,, \qquad k \in s_h,\ h = 1, \ldots, H$$

$$(4.3.3) \qquad \hat{N}_h = \sum_{k=1}^{n_h} w_{hk}$$

[1] Para se realizarem as aplicações práticas utilizou-se o produto informático de estatística *Statistical Analysis System*, *SAS* versão 6.12, com o módulo *STAT*.

(4.3.4) $$\hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk}$$

e w_{hk} é o peso inicial do indivíduo *k* pertencente à *h-ésima* célula de ajustamento (estrato inicial).

No método de ajustamento por pós-estratificação a amostra foi pós-estratificada em L = 5 pós-estratos, segundo a variável *ENPS – Escalões de número de pessoas ao serviço* (veja-se Quadro 4.3.1) que se supõe estreitamente relacionada com as variáveis de interesse.

Quadro 4.3.1 – Escalões de número de pessoas ao serviço (*ENPS*)

Valor	Descrição
0	0 pessoas ao serviço
1	1 a 9 pessoas ao serviço
2	10 a 19 pessoas ao serviço
3	20 a 49 pessoas ao serviço
4	50 a 99 pessoas ao serviço

Neste método assume-se que os pós-estratos correspondem a células de ajustamento da não resposta. Supõe-se então que os indivíduos têm valores semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais, em cada pós-estrato.

O estimador do total da população por pós-estratificação utilizado foi:

(4.3.5) $$\hat{\tau}_{ps} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} w_{ik}^{(ps)} y_{ik}$$

onde,

(4.3.6) $$w_{ik}^{(ps)} = \frac{N_i}{\hat{N}_{1i}} w_{ik}\,, \qquad k \in s_i,\ i = 1, \ldots, L$$

(4.3.7) $$\hat{N}_{1i} = \sum_{k=1}^{n_{1i}} w_{ik}$$

sendo, w_{ik} o peso inicial do elemento *k* pertencente ao *i–ésimo* pós-estrato e N_i a dimensão (conhecida) do pós-estrato *i* na população.

No método de pós-estratificação com ajustamento da não resposta por ponderação em classes consideraram-se os estratos iniciais como sendo as classes de ajustamento da não resposta. Supõe-se, portanto, que os elementos têm valores semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais, em cada estrato inicial.

Também neste caso, a amostra foi pós-estratificada em L = 5 pós-estratos, segundo a variável *ENPS – Escalões de número de pessoas ao serviço* (veja-se Quadro 4.3.1).

O estimador de pós-estratificação do total da população com ajustamento da não resposta por ponderação em classes obteve-se através da reponderação dos elementos da amostra, da seguinte forma.

Em primeiro lugar, obtiveram-se os pesos ajustados $w_{hk}^{(pc)}$ por ponderação em classes de não resposta (estratos iniciais), através de (4.3.2). Em seguida, estes ponderadores foram novamente ajustados através da pós-estratificação da amostra. Desta forma, obtiveram-se os ponderadores finais:

(4.3.8) $$w_{ik}^{(pc,ps)} = \frac{N_i}{\hat{N}_{1i}^{*}} w_{ik}^{(pc)}, \qquad k \in s_i,\ i = 1, \ldots, L$$

onde, $w_{ik}^{(pc)}$ é o peso ajustado, na etapa anterior, do elemento *k* pertencente ao pós-estrato *i*, e

(4.3.9) $$\hat{\tau}_{pc,ps} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} w_{ik}^{(pc,ps)} y_{ik}$$

No primeiro exemplo prático (secção 4.3.2), considerou-se também o estimador da variância do estimador de pós-estratificação proposto por Rao (1985) para o caso da ausência de não respostas, para um plano de sondagem estratificado (para mais detalhes veja-se a secção 3.5):

(4.3.10) $\hat{V}_{rao}(\hat{\tau}_{ps,str}) =$

$$\sum_{i=1}^{L}\sum_{h=1}^{H}\left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 N_{\bullet h}^2\left(\frac{1}{n_{\bullet h}}-\frac{1}{N_{\bullet h}}\right)\frac{1}{n_{\bullet h}-1}\left[\sum_{k\in s_{ih}}(y_k-\bar{y}_{s_{ih}})^2+n_{ih}(1-n_{ih}/n_{\bullet h})(\bar{y}_{s_{ih}}-\hat{\mu}_{i_{str}})^2\right]$$

Este estimador poderá ter boas propriedades, na abordagem condicional, na ausência de não respostas. No caso em que ocorrem não respostas, as propriedades do estimador são ainda mais difíceis de analisar. No entanto, poderá também ter boas propriedades numa abordagem condicional se os elementos tiverem valores semelhantes para as variáveis de interesse e as probabilidades de resposta forem iguais, em cada pós-estrato (veja-se a análise efectuada para a SASSR, na secção 3.6.4.

Relativamente aos restantes estimadores propostos, não foi possível encontrar na literatura estimadores da variância para o plano de sondagem subjacente ao IEH (sondagem aleatória estratificada sem reposição). Para contornar este problema, foi utilizado o método de Bootstrap BWO, proposto por Sitter (1992b), que conduz a estimadores bootstrap da variância válidos, no caso de estimadores lineares. No caso não linear, o método também parece ser promissor, como se referiu anteriormente. Os resultados apresentados no segundo exemplo (secção 4.3.3) foram, então, obtidos através da aplicação desta metodologia (c.f. secção 2.7.3).

Em ambos os exemplos práticos, a amostra considerada é a que se refere às *pequenas e médias empresas do Continente*.

4.3.2 Exemplo I

Para as variáveis de interesse indicadas na secção 4.2.3, procedeu-se ao cálculo dos totais e das médias estimadas através dos seguintes métodos: ponderação em

classes, pós-estratificação e pós-estratificação com ajustamento da não resposta por ponderação em classes.

Obtiveram-se também estimativas do total e da média através do estimador de Horvitz-Thompson que, apesar de neste caso ser enviesado, não deixa de ser uma referência.

Nos dois métodos de pós-estratificação considerados, utilizou-se a variável *ENPS - Escalões de número de pessoas ao serviço* (veja-se Quadro 4.3.1) para pós-estratificar a amostra. Os escalões desta variável foram também utilizados na estratificação inicial. A sua escolha, como variável de pós-estratificação, prende-se, por um lado, com o facto de se dispor dos valores da variável *Q20001 - número médio de pessoas ao serviço* na amostra e das dimensões dos pós-estratos na população. Por outro lado, suspeitava-se que os estratos iniciais, sendo homogéneos na sua constituição, continham empresas com comportamentos muito diferenciados, pelo que poderiam pôr em causa essa homogeneidade. No que se refere a esta variável, verificou-se que 13090 empresas mudaram de estrato.

A variável *EVVN - Escalões de volume de vendas* (ou, uma combinação desta com a *ENPS*[1]) seria talvez a melhor candidata a variável de pós-estratificação, uma vez que o INE (1997) refere que esta variável assegura uma maior homogeneidade entre as empresas dos diversos estratos do universo. No entanto, não foi utilizada pelo facto de não ter sido possível obter os valores relativos às dimensões dos escalões de volume de vendas na população.

Apresentam-se em seguida os resultados referentes às variáveis *Q20001 – Número médio de pessoas ao serviço* (Quadro 4.3.2); *Q4160 – Vendas* (Quadro 4.3.3) e *Q4190 – Prestações de serviços* (Quadro 4.3.4).

[1] Lessler e Kalsbeek (1992, p. 188) referem que parece ser preferível definir as células de ajustamento da não resposta a partir do cruzamento de várias variáveis aceitáveis, do que formar o mesmo número de células a partir de uma divisão mais fina de apenas uma delas.

Quadro 4.3.2 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Nº médio de pessoas ao serviço* (Q20001)

Estimador	Estimativa do total	Estimativa da média	Estimativa da variância do estimador da média	Coeficiente de variação da média estimado (%)
HT	1510168.60	2.12	0.0003	0.81
PC	2038545.58	2.86	-	-
PS	2006329.83	2.82	0.0002	0.51
PC, PS	2025145.38	2.84	-	-

Quadro 4.3.3 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Vendas* (*Q4160*)

Estimador	Estimativa do total	Estimativa da média	Estimativa da variância do estimador da média	Coeficiente de variação da média estimado (%)
HT	14779153197	20738.54	199658.51	2.15
PC	19981217698	28038.23	-	-
PS	19569497969	27460.49	371786.16	2.17
PC, PS	19810036906	27798.02	-	-

Quadro 4.3.4 – Estimativas obtidas para os estimadores: Horvitz-Thompson (HT), ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Prestações de serviços* (*Q4190*)

Estimador	Estimativa do total	Estimativa da média	Estimativa da variância do estimador da média	Coeficiente de variação da média estimado (%)
HT	3938549690.4	5526.69	17910.48	2.42
PC	5229277518.8	7337.87	-	-
PS	5554052079	7793.61	21056.35	1.98
PC, PS	5528976885.8	7758.42	-	-

Os resultados apresentados permitem observar que a média estimada apresenta-se muito semelhante para os três métodos de ajustamento considerados e é superior à média estimada pelo estimador de Horvitz-Thompson, nas três variáveis de interesse. A estimativa da variância deste estimador subestima o verdadeiro valor do erro, uma vez que não contém a contribuição do enviesamento.

Sob as hipóteses formuladas, é de esperar que os estimadores por ajustamento sejam aproximadamente centrados e que o estimador da variância do estimador de pós-estratificação (PS) proposto por Rao (1985) tenha boas propriedades, numa abordagem condicional (à semelhança da SASSR – veja-se a secção 3.6.4).

No exemplo prático que se segue apresentam-se os resultados relativos à aplicação do método Bootstrap BWO, proposto por Sitter (1992b), para os três estimadores por ajustamento em análise.

4.3.3 Exemplo II

O procedimento utilizado ao longo deste exemplo foi o algoritmo Bootstrap BWO, proposto por Sitter (1992b), apresentado na secção 2.7.3. As estimativas da variância obtiveram-se através das aproximações de Monte Carlo. Para tal, foram retiradas 600 amostras bootstrap, da população bootstrap construída segundo o referido algoritmo. Os resultados que se obtiveram são apresentados em seguida.

Quadro 4.3.5 - Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Nº médio de pessoas ao serviço* (Q20001)

Estimador	Estimativa bootstrap do total	Estimativa bootstrap da média	Estimativa bootstrap da variância do estimador da média	Coeficiente de variação bootstrap estimado do estimador da média (%)
PC	1887715.58	2.79	0.0005	0.79
PS	2038626.07	3.02	0.0003	0.53
PC, PS	2027816.36	3.00	0.0002	0.52

Quadro 4.3.6 - Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Vendas* (*Q4160*)

Estimador	Estimativa bootstrap do total	Estimativa bootstrap da média	Estimativa bootstrap da variância do estimador da média	Coeficiente de variação bootstrap estimado do estimador da média (%)
PC	18380211769	27194.85	334765.33	2.13
PS	19963467795	29537.39	416107.28	2.18
PC, PS	19831732906	29342.48	454619.21	2.30

Quadro 4.3.7 - Estimativas obtidas através do método Bootstrap BWO, com 600 réplicas dos estimadores: ponderação em classes (PC), pós-estratificação (PS) e pós-estratificação com ajustamento da não resposta por ponderação em classes (PC,PS), para a variável *Prestações de serviços* (*Q4190*)

Estimador	Estimativa bootstrap do total	Estimativa bootstrap da média	Estimativa bootstrap da variância do estimador da média	Coeficiente de variação bootstrap estimado do estimador da média (%)
PC	5255307304	7775.61	45661.52	2.75
PS	5258899182	7780.92	22760.55	1.94
PC, PS	5567635301	8237.72	34973.91	2.27

Relativamente às estimativas apresentadas, retiram-se algumas ilações:

- Como era de esperar, as estimativas bootstrap da média são semelhantes às que foram obtidas no Exemplo I, para todas as variáveis de interesse.
- As estimativas bootstrap da média são semelhantes para os dois estimadores de pós-estratificação considerados, principalmente quando se tomam como variáveis de interesse a *Q20001 – Nº médio de pessoas ao serviço* [$\hat{\mu}^{*}_{ps}(\cdot)=3.02$, $\hat{\mu}^{*}_{pc,ps}(\cdot)=3$] e a *Q4160 – Vendas* [$\hat{\mu}^{*}_{ps}(\cdot)=29537.39$, $\hat{\mu}^{*}_{pc,ps}(\cdot)=29342.48$]. Para estas variáveis, o estimador de ponderação em classes apresenta estimativas bootstrap com valores inferiores às obtidas para os outros dois estimadores.

- Para a variável *Q4190 – Prestações de serviços*, as estimativas bootstrap da média são semelhantes para os estimadores de ponderação em classes e de pós-estratificação [$\hat{\mu}^{*}_{pc}(\cdot)$=7775.61, $\hat{\mu}^{*}_{ps}(\cdot)$=7780.92]. No entanto, a estimativa bootstrap da variância do estimador de ponderação em classes é quase o dobro da que foi obtida para o estimador de pós-estratificação [$\hat{V}^{*}_{BWO}(\hat{\mu}_{pc})$=45661.52, $\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})$=22760.55].

- As estimativas da variância do estimador de pós-estratificação, obtidas no Exemplo I, são inferiores às estimativas bootstrap da variância desse estimador, em todas as variáveis de interesse. Quando se comparam os respectivos desvios padrão os resultados são os seguintes:

  - *Nº médio de pessoas ao serviço*: $\sqrt{\hat{V}_{rao}(\hat{\mu}_{ps})}$ = 0.01; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 0.02

  - *Vendas*: $\sqrt{\hat{V}_{rao}(\hat{\mu}_{ps})}$ = 609.74; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 645.06

  - *Prestações de serviços*: $\sqrt{\hat{V}_{rao}(\hat{\mu}_{ps})}$ = 145.11; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 150.87

- As estimativas bootstrap da variância dos dois estimadores de pós-estratificação são semelhantes, para as variáveis de interesse consideradas:

  - *Nº médio de pessoas ao serviço*: $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 0.02; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{pc,ps})}$ = 0.01

  - *Vendas*: $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 645.06; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{pc,ps})}$ = 674.25

  - *Prestações de serviços*: $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{ps})}$ = 150.87; $\sqrt{\hat{V}^{*}_{BWO}(\hat{\mu}_{pc,ps})}$ = 187.01

Além dos resultados apresentados, construíram-se também os histogramas das réplicas bootstrap dos estimadores em estudo, para cada uma das variáveis de interesse (veja-se o Anexo 5). A observação dos histogramas permite constatar que as distribuições das réplicas aparentam ter uma forma aproximadamente Normal, o que concorda com a teoria (Shao e Tu, 1995).

Com base nos resultados obtidos e nas evidências teóricas apresentadas, parece razoável concluir que, dos métodos de ajustamento das não respostas considerados, os métodos de pós-estratificação sejam os mais adequados. Aliás, como foi referido anteriormente, é de esperar que os métodos de pós-estratificação tenham um erro quadrático médio inferior ao do estimador de ponderação em classes. Por outro lado, os métodos de pós-estratificação permitem também lidar com os problemas da base de sondagem, através da reponderação dos elementos da amostra, por forma a que esta reflicta a estrutura da população.

Interessa também referir que o esforço computacional necessário à obtenção de estimativas bootstrap poderá não se justificar, quando se opta pelos estimadores de pós-estratificação (por exemplo, quando se pretendem resultados antecipados ou preliminares, quando há um número elevado de variáveis de interesse ou quando há um calendário muito exigente relativamente à disponibilização de resultados). Esta análise resulta essencialmente do facto de ambos os estimadores apresentarem estimativas semelhantes e de se dispor de uma forma mais simples, em termos de implementação, de se obterem estimativas da variância do estimador de pós-estratificação: através do estimador da variância proposto por Rao (1985); apesar de existirem evidências de que essas estimativas subestimam o valor da verdadeira variância, estas não diferem muito das que se obtiveram por Bootstrap.

Com este exemplo não se pretende comparar a precisão dos estimadores propostos mas, simplesmente, apresentar um método que poderá ser adequado para estimar a variância desses estimadores. Relembre-se que os estimadores apresentados estão sujeitos a certos pressupostos que são essenciais ao seu desempenho e que para medir a precisão dos estimadores seria necessário conhecer os respectivos erros quadráticos médios na população.

Como se referiu anteriormente, a variável *EVVN - Escalões de volume de vendas* (ou, uma combinação desta com a *ENPS - Escalões de número de pessoas ao serviço*) seria talvez a melhor candidata a variável de pós-estratificação. É também de esperar que, caso esta variável fosse utilizada, as estimativas que se obteriam para os dois estimadores de pós-estratificação fossem menos semelhantes. Esta suposição baseia-se nas evidências de que as variáveis de estratificação utilizadas pelo plano de sondagem são adequadas para definir as células de ajustamento da não resposta (Lessler e Kalsbeek, 1992, p. 188) e a variável *Escalões de volume de*

vendas poderá garantir a homogeneidade dos pós-estratos e reflectir, de forma mais adequada, a estrutura da população. Teria sido, portanto, extremamente interessante analisar a utilização destas técnicas de pós-estratificação no IEH.

Ainda no âmbito da pós-estratificação, os métodos *generalized raking* (referidos na secção 3.5), propostos por Deville, Särndal e Sautory (1993), apresentam-se também como uma metodologia muito promissora, quando a base de sondagem apresenta problemas e se dispõe de alguma informação auxiliar sobre a população, pelo que o estudo destes métodos teria sido enriquecedor.

5 CONCLUSÃO

Com este trabalho pretendeu-se estudar métodos de estimação por pós-estratificação em inquéritos por amostragem e evidenciar os efeitos dos erros não amostrais na estimação, em particular a existência de erros nas bases de sondagem e a ocorrência de não respostas aos inquéritos.

De um modo geral, as bases de amostragem não garantem uma representação completa, perfeita e actualizada da população alvo. Dos problemas que podem ocorrer nas bases de sondagem são de salientar: a subcobertura, a sobrecobertura, os registos duplicados ou múltiplos e a informação auxiliar incorrecta. Todos estes problemas têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram.

A subcobertura conduz a estimativas enviesadas uma vez que uma parte da população não pode ser observada. Este é talvez o problema mais sério dada a impossibilidade de detectá-lo, quer a partir da amostra, quer a partir da base de sondagem. Neste caso, uma forma de reduzir o enviesamento do estimador de Horvitz-Thompson é a utilização de um ajustamento pelo quociente (Särndal, Swensson e Wretman, 1992).

Quando ocorrem problemas de sobrecobertura, a população alvo é um domínio da base de amostragem e, portanto, os métodos de estimação em domínios revelam-se como uma metodologia adequada para tratar este problema.

A utilização de informação auxiliar incorrecta reduz a precisão das estimativas da sondagem (Lessler e Kalsbeek, 1992). Este tipo de erros pode conduzir tanto a problemas de sobrecobertura, como de subcobertura.

Outro problema que pode surgir, quando se utiliza informação auxiliar incorrecta ou desactualizada para implementar um plano de sondagem estratificada, é a ocorrência de mudanças de estrato; ou seja, as respostas ao inquérito podem sugerir que não existe uma correspondência exacta entre os estratos na base de sondagem e na população. Neste caso, as estimativas por estrato podem ser obtidas por métodos de estimação em domínios.

Os métodos de estimação pelo quociente e, em especial, os métodos de pós-estratificação são apresentados na literatura como uma forma de lidar com os problemas originados por deficiente informação na base de sondagem. Estas técnicas de reponderação têm por objectivo melhorar as estimativas obtidas, podendo utilizar, no momento da estimação, informação auxiliar mais actualizada.

Os estimadores de pós-estratificação ajustam o coeficiente de extrapolação de cada elemento da amostra, por forma a que esta reflicta a estrutura actual da população. Assim, se uma amostra estiver desequilibrada para algumas características da população, o estimador de pós-estratificação corrige este desequilíbrio automaticamente (Holt e Smith, 1979).

Outro erro não amostral que ocorre na maioria das sondagens é a não resposta total, ou parcial, ao inquérito. Nesta situação, os estimadores usuais são enviesados. Os estimadores de pós-estratificação inserem-se numa classe de métodos de estimação que ajustam os coeficientes de extrapolação, por forma a que os pesos obtidos tenham em consideração as não respostas. Estes métodos são, geralmente, utilizados no tratamento das não respostas totais.

Apesar dos métodos de pós-estratificação serem muito utilizados, a pesquisa bibliográfica efectuada revela alguma insuficiência de referências, no que diz respeito às propriedades teóricas destes estimadores, quando se consideram planos de sondagem complexos; sendo de salientar a abordagem condicional efectuada por Rao (1985). Procurou-se, então, superar esta dificuldade através da apresentação de métodos Bootstrap para a estimação da variância dos estimadores.

Em alguns dos exemplos práticos, foi utilizado o método Bootstrap BWO, proposto por Sitter (1992b). Este procedimento de replicação procura captar as dimensões importantes da variância na selecção original da amostra e os ajustamentos efectuados durante a fase de determinação dos ponderadores.

Com base nos resultados obtidos e nas evidências teóricas apresentadas, parece razoável concluir que, dos métodos de ajustamento das não respostas considerados, os métodos de pós-estratificação são os mais adequados. No caso do IEH, este era o resultado esperado uma vez que, para além do problema das não respostas totais, o IEH apresenta alguns problemas na base de sondagem: por um

lado, a informação auxiliar que serviu de base à estratificação estava desactualizada ou incorrecta (o que conduziu ao problema das mudanças de estrato); por outro lado, a base de sondagem tinha problemas de cobertura (a sobrecobertura é evidente quando se analisam as *condições de apuramento* definidas pelo INE).

É também de salientar que as estimativas da variância do estimador de pós-estratificação, obtidas através do estimador proposto por Rao (1985), não diferem muito das obtidas por Bootstrap, para as variáveis analisadas. No entanto, tal poderá não ocorrer se forem consideradas outras variáveis (de análise ou de pós-estratificação) ou se os pressupostos assumidos não se verificarem, dado que há evidências de que as estimativas obtidas através do estimador proposto por Rao (1985) subestimam o valor da verdadeira variância. Assim, teria sido muito interessante analisar e comparar a utilização de outras variáveis de pós-estratificação no IEH.

Ainda no âmbito da pós-estratificação, fica como sugestão para futuras investigações, o estudo dos métodos *generalized raking*, propostos por Deville, Särndal e Sautory (1993); dado que estes métodos permitem utilizar variáveis de pós-estratificação para as quais a única informação auxiliar disponível diz respeito à dimensão da população nas categorias definidas por cada uma das variáveis, tomadas isoladamente.

É ainda de referir que os métodos de pós-estratificação desenvolvidos no âmbito da inferência *model-based* seriam também interessantes de analisar.

6 Referências

AZEVEDO, Áurea Sofia Pimenta (1999). *Estimação na Presença de Não Respostas – Aplicação ao Inquérito às Empresas (Harmonizado) do Instituto Nacional de Estatística*. Dissertação de Mestrado, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa.

BARNETT, V. (1991). *Sample Survey - Principles and Methods*. 4th edition, Edward Arnold, London.

BICKEL, P. J. e FREEDMAN, D. A. (1984). Asymptotic normality and the Bootstrap in stratified sampling. *Annals of Statistics* 12, 470-482.

BOWLEY, A. L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute* 22, 1-62.

CHAO, M. T. e LO, S. H. (1985). A Bootstrap method for finite populations. *Sankhyä* A 47, 399-405.

CHEN, J. e SITTER, R. R. (1993). Edgeworth expansion and the Bootstrap for stratified sampling without replacement from a finite population. *The Canadian Journal of Statistics* 21, No. 4, 347-357.

COCHRAN, W. G. (1977). *Sampling Techniques.* 3rd Edition, A Wiley publication in Applied Statistics, John Wiley & Sons, New York.

COELHO, Pedro Miguel Pereira Simões (1995). *Avaliação de Imagem Institucional - Uma Sondagem de Opinião no Mercado Segurador*. Dissertação de Mestrado, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa.

COELHO, Pedro Miguel Pereira Simões (1996). Estimadores combinados para pequenos domínios. *Revista de Estatística* 2, 23-43.

CORNFIELD, J. (1944). On samples from finite populations. *Journal of the American Statistical Association* 39, 236-239.

DEMING, W. E. e STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11, 427-444.

DEVILLE, J. C. (1987). "Replications d'échantillons : Demi-Echantillons, Jackknife et Bootstrap." in *Les Sondages*, eds. J.-J. Droesbeke , B. Fichet e P. Tassi, Association pour la Statistique et les Utilisations, Economica, Paris, 147-171.

DEVILLE, J. C. e SÄRNDAL, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, No. 418, 376-382.

DEVILLE, J. C., SÄRNDAL, C. E. e SAUTORY, O. (1993). Generalised raking procedures in survey sampling. *Journal of the American Statistical Association* 88, No. 423, 1013-1020.

DJERF, K. (1997). Effects of post-stratification on estimates of the Finnish Labour Force Survey. *Journal of Official Statistics* 13, No. 1, 29-39.

DUSSAIX, A. M. (1987). "Modèles de surpopulation." in *Les Sondages*, eds. J.-J. Droesbeke , B. Fichet e P. Tassi, Association pour la Statistique et les Utilisations, Economica, Paris, 67-88.

EFRON, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Mathematical Statistics* 7, 1-26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

EFRON, B. e TIBSHINARI, R. J. (1993). *Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57, Chapman & Hall.

FULLER, W. A. (1966). Estimation employing post strata. *Journal of the American Statistical Association* 61, 1172-1183.

GELMAN, A. e LITTLE, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* 23, No. 2, 127-135.

GOMES, Paulo (1998). *Tópicos de Sondagens*. VI Congresso Anual, Sociedade Portuguesa de Estatística, Tomar, Junho de 1998.

GOURIEROUX, C. (1987). "Sondages sans biais." in *Les Sondages*, eds. J.-J. Droesbeke , B. Fichet e P. Tassi, Association pour la Statistique et les Utilisations, Economica, Paris, 43-66.

GROSBRAS, Jean-Marie (1987). *Methodes Statistiques des Sondages.* Collection Économie et Statistiques Avancées, Economica, Paris.

GROSS, S. (1980). “Median estimation in sample surveys.” Proceedings of the Section *on* Survey Research Methods, American Statistical Association, 181-184.

HANSEN, M. H. e HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

HANSEN, M. H., HURWITZ, W. N. e GURNEY, M. (1946). Problems and methods of the sample surveys of business. *Journal of the American Statistical Association* 41, 173-189.

HANSEN, M. H., HURWITZ, W. N. e MADOW, W. G. (1953a). *Sample Survey Methods and Theory.* Vol. I - Methods and Aplications, Wiley Classics Library Edition, John Wiley & Sons, New York.

HANSEN, M. H., HURWITZ, W. N. e MADOW, W. G. (1953b). *Sample Survey Methods and Theory.* Vol. II - Theory, Wiley Classics Library Edition, John Wiley & Sons, New York.

HARTLEY, H. O., RAO, J. N. K. e KIEFER, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association* 64, 841-851.

HARTLEY, H. O. e ROSS, A. (1954). Unbiased ratio estimators. *Nature* 174, 270-271.

HEDAYAT, A. S. e SINHA, B. K. (1991). *Design and Inference in Finite Population Sampling.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

HEDLIN, D., FALVEY, H., CHAMBERS, R., KOKIC, P. (1998). "The effective use of auxiliary information in a business survey." Paper presented at NTTS'98 – Seminar on New Techniques & Technologies for Statistics, Sorrento, Italy, 4-6 Nov. 1998.

HOLT, D. e HOLMES, D. J. (1994). Small domain estimation for unequal probability survey designs. *Survey Methodology* 20, No. 1, 23-31.

HOLT, D. e SMITH, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society* A 142, 33-46.

HORVITZ, D. G. e THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-685.

INSTITUTO NACIONAL DE ESTATÍSTICA (1997). *Inquérito ás Empresas / Harmonizado – Dossier Global do Projecto*. Departamento de Estatísticas das Empresas, INE/DEE, Junho 1997.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association* 52, 503-510.

KISH, Leslie (1965). *Survey Sampling.* John Wiley, New York.

KOEIJERS, Elly e WILLEBOORDSE, Ad (1995). *Reference manual on design and implementation of business surveys.* Statistics Netherlands, First Draft, March 1995.

KOOP, J. C. (1988). "The technique of replicated or interpenetrating samples." in *Handbook of Statistics*, Vol. 6, eds. P. R. Krishnaiah e C. R. Rao, Elsevier Science Publishers B. V, 333-368.

LAZZERONI, L. C. e LITTLE, R. J. A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics* 14, No. 1, 61-78.

LEHTONEN, R. e PAHKINEN, E. J. (1996). *Practical Methods for Design and Analysis of Complex Surveys.* Revised Edition, November 1996, Statistics in Practice, John Wiley & Sons, Chichester.

LEONARD, K. A., *et al.* (1994). "Approximating the variance of the survey regression estimator using poststratification." Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 222-227.

LESSLER, J. T. e KALSBEEK, W. D. (1992). *Nonsampling Error in Surveys*. Wiley Series in probability and Mathematical Statistics, John Wiley & Sons, New York.

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54, No. 2, 139-157.

LITTLE, R. J. A. (1993). Post-stratification: A Modeler's Perspective. *Journal of the American Statistical Association* 88, No. 423, 1001-1012.

LUNDSTRÖM, S. e SÄRNDAL, C. E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* 15, No. 2, 305-327.

NATHAN, G. (1988). "Inference Based on Data from Complex Sample Designs." in *Handbook of Statistics*, Vol. 6, eds. P. R. Krishnaiah e C. R. Rao, Elsevier Science Publishers B. V, 247-266

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, 558-625.

QUENOUILLE, M. (1949). Approximate tests of correction in time series. *Journal of the Royal Statistical Society* B 11, 18-44.

RAO, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology* 11, No. 1, 15-31.

RAO, J. N. K. (1994). “Resampling methods for complex surveys.” Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 35-41.

RAO, J. N. K. e WU, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* 83, 231-241.

RAO, P. S. R. S. (1988). "Ratio and regression estimators." in *Handbook of Statistics*, Vol. 6, eds. P. R. Krishnaiah e C. R. Rao, Elsevier Science Publishers B. V, 449-468.

RIVEST, Louis-Paul (1999). “Stratum jumpers: can we avoid them?.” Paper presented at the 1999 Joint Statistical Meetings, Survey Research Methods Section, Baltimore, Maryland, EUA, 7-12 Ag. 1999.

SÄRNDAL, C. E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association* 79, 624-631.

SÄRNDAL, C. E. e HIDIROGLOU, M. A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* 84, 266-275.

SÄRNDAL, C. E., SWENSSON, B. e WRETMAN, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

SEN, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 5, 119-127.

SHAO, J. e SITTER, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* 91, No. 435, 1278-1288.

SHAO, J. e TU, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SINGH, A. C. e MOHL, C. A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology* 22, No. 2, 107-115.

SITTER, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association* 87, No. 419, 755-765.

SITTER, R. R. (1992b). Comparing three Bootstrap methods for survey data. *The Canadian Journal of Statistics* 20, No. 2, 135-154.

SKINNER, C. (1998). “Calibration weighting and non-sampling errors.” Paper presented at NTTS’98 – Seminar on New Techniques & Technologies for Statistics, Sorrento, Italy, 4-6 Nov. 1998.

SKINNER, C. J., HOLT, D. e SMITH, T. M. F. (1989). *Analysis of Complex Surveys.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Chichester.

TEPPING, B. J. (1968). “Variance estimation in complex surveys.” Proceedings of the Social Statistics Section, American Statistical Association, 11-18

THOMPSON, Steven K. (1992). *Sampling.* A Wiley Interscience Publication, John Wiley & Sons, New York.

THOMSEN, Ib e TESFU, D. (1988). "On the Use of Models in Sampling from Finite Populations." in *Handbook of Statistics*, Vol. 6, eds. P. R. Krishnaiah e C. R. Rao, Elsevier Science Publishers B. V, 369-397

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* 88, nº 421, 89-96.

WILLIAMS, W. H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association* 57, 622-627.

WOODRUFF, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 66, No. 334, 411-414.

WRIGHT, R. L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association* 78, 879-884.

YATES, F. e GRUNDY, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* B 15, 235-261.

ANEXO 1 – Abreviaturas e notação

A1.1 Abreviaturas

BWO	*Bootstrap Without Replacement* ou *Without Replacement Bootstrap*
FGUE	Ficheiro geral de Unidades Estatísticas
IEH, IEH96	Inquérito às Empresas / Harmonizado (1996)
INE	Instituto Nacional de Estatística
MMB	*Mirror-Match Bootstrap*
NSI	*National Statistical Institutes*
NUTS	Nomenclatura das Unidades Territoriais para Fins estatísticos
RB	*Rescaling Bootstrap*
SASCR	Sondagem aleatória simples com reposição
SAS, SASSR	Sondagem aleatória simples sem reposição

A1.2 Notação

A1.2.1 Notação geral

Y	variável de interesse
y_i	*i-ésima* observação da variável Y
X	variável auxiliar
x_i	*i-ésima* observação da variável X
θ	notação genérica para parâmetro
$\hat{\theta}$	notação genérica para estimador do parâmetro θ
$E(\hat{\theta})$	valor esperado de $\hat{\theta}$
$B(\hat{\theta})$	enviesamento de $\hat{\theta}$
$V(\hat{\theta})$	variância de $\hat{\theta}$
$AV(\hat{\theta})$	*variância aproximada* de $\hat{\theta}$, obtida pelo método de linearização de Taylor
$\hat{V}(\hat{\theta})$	notação genérica para estimador da variância de $\hat{\theta}$
$\hat{V}^{*}_{BWO}(\hat{\theta})$	estimador da variância de $\hat{\theta}$ do algoritmo BWO

$\sigma_{\hat{\theta}}$	desvio padrão de $\hat{\theta}$
$CV(\hat{\theta})$	coeficiente de variação de $\hat{\theta}$
$EQM(\hat{\theta})$	erro quadrático médio de $\hat{\theta}$
$BR(\hat{\theta})$	*bias ratio* de $\hat{\theta}$ (quociente entre o enviesamento e o desvio padrão de $\hat{\theta}$)
π_i	probabilidade de inclusão de 1ª ordem do elemento *i*, sob um determinado plano de sondagem
π_{ij}	probabilidade de inclusão de 2ª ordem dos elementos *i* e *j* ($i \neq j$), sob um determinado plano de sondagem
w_i	peso de inclusão ou coeficiente de extrapolação do elemento *i*
$\mathbb{I}_{i\in\zeta}$	variável indicatriz do conjunto ζ (variável que toma o valor 1 se o elemento pertence ao conjunto ζ e toma o valor zero, caso contrário)
$\sum\sum_U a_{ij}$, $\sum_{i\in U}\sum_{j\in U} a_{ij}$	$\sum_{i\in U} a_{ii} + \sum_{i\in U}\sum_{\substack{j\in U \\ j\neq i}} a_{ij}$
$\hat{\theta}_{SAS}$, $\hat{\theta}_{SASSR}$	estimador de θ sob um plano SASSR
$\hat{\theta}_{SASCR}$	estimador de θ sob um plano SASCR
$\hat{\theta}_{HT}$, $\hat{\theta}_{\pi}$	estimador de Horvitz-Thompson de θ
$\hat{\theta}_{STR}$	estimador de θ sob um plano de sondagem aleatória estratificada
$\hat{\theta}_{prop}$	estimador de θ sob um plano de sondagem aleatória estratificada proporcional
$\hat{\mu}_w$	estimador do quociente para μ (*weighted sample mean*)
$\hat{\theta}_Q$	estimador pelo quociente usual de θ
$\hat{\theta}_{ps}$	estimador de pós-estratificação de θ, na ausência de não respostas; estimador de pós-estratificação de θ, na presença de não respostas, quando as classes de ajustamento são os pós-estratos
$\hat{V}_{rao}(\hat{\theta}_{ps})$	estimador da variância de $\hat{\theta}_{ps}$ proposto por Rao (1985)
$\hat{\theta}_{pc}$	estimador por ponderação em classes de ajustamento da não resposta de θ

	resposta de θ
$\hat{\theta}_{pc,ps}$	estimador de pós-estratificação de θ, na presença de não respostas, quando as classes de ajustamento são diferentes dos pós-estratos

A1.2.2 Notação referente à população

U	população alvo ou universo de referência
U_d	sub-população (domínio) de U
N	dimensão da população
N_d	dimensão da sub-população (domínio) U_d
τ, τ_y [τ_x]	total da variável Y [X]
τ_d	total da variável Y na sub-população (domínio) U_d
μ, μ_y [μ_x]	média da variável Y [X]
μ_d	média da variável Y na sub-população (domínio) U_d
R	quociente entre os totais (médias) de duas variáveis
σ^2, σ_y^2	variância da variável Y
S^2, S_y^2	variância corrigida da variável Y

A1.2.3 Notação referente à amostra

s	conjunto dos elementos da amostra
s_d	conjunto dos elementos da amostra que pertencem à sub-população (domínio) U_d
n	dimensão de *s*
n_d	dimensão de s_d
f	taxa de sondagem
$\bar{y}$ [$\bar{x}$]	média amostral da variável Y [X]
s^2, s_y^2	variância amostral corrigida

ANEXO 2 – Demonstrações

A2.1 Resultados da secção 2.4

A2.1.1 Estimação de τ numa sondagem aleatória com probabilidades desiguais

- Demonstração do resultado (2.4.9)

$$\hat{V}_2(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2, \quad \text{se a dimensão da amostra, } n\text{, for fixa}$$

apresentado na secção 2.4.1.

Para um plano de sondagem aleatória com probabilidades desiguais sem reposição, pretende-se demonstrar que a expressão da variância do estimador de Horvitz-Thompson do total da população (2.4.9), devida a Sen, Yates e Grundy, é equivalente à expressão da variância devida a Horvitz-Thompson (2.4.4), se a dimensão da amostra, *n*, for fixa.

(A2.1.1)
$$\hat{V}_2(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 =$$

$$= \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left(\frac{y_i^2}{\pi_i^2} - 2\frac{y_iy_j}{\pi_i\pi_j} + \frac{y_j^2}{\pi_j^2}\right)(\pi_i\pi_j - \pi_{ij}) =$$

$$= \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i^2}{\pi_i^2}(\pi_i\pi_j - \pi_{ij}) - \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_iy_j}{\pi_i\pi_j}(\pi_i\pi_j - \pi_{ij}) +$$

$$+ \frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_j^2}{\pi_j^2}(\pi_i\pi_j - \pi_{ij}) =$$

Uma vez que a primeira e a última parcela são iguais, tem-se:

(A2.1.2) $$\hat{V}_2(\hat{\tau}_{HT}) = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{y_i^2}{\pi_i^2}(\pi_i\pi_j - \pi_{ij}) - \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{y_i y_j}{\pi_i\pi_j}(\pi_i\pi_j - \pi_{ij}) =$$

$$= \sum_{i\in U}\frac{y_i^2}{\pi_i^2}\left[\pi_i \sum_{\substack{j\in U\\ j\neq i}}\pi_j - \sum_{\substack{j\in U\\ j\neq i}}\pi_{ij}\right] + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{y_i y_j}{\pi_i\pi_j}(\pi_{ij} - \pi_i\pi_j)$$

Se *n* é fixo, tem-se:

(A2.1.3) $$\sum_{\substack{j\in U\\ j\neq i}}\pi_j = n - \pi_i$$

(A2.1.4) $$\sum_{\substack{j\in U\\ j\neq i}}\pi_{ij} = (n-1)\,\pi_i$$

Logo, substituindo-se estas expressões em (A2.1.2), obtém-se o resultado pretendido:

(A2.1.5) $$\hat{V}_2(\hat{\tau}_{HT}) = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{y_i y_j}{\pi_i\pi_j}(\pi_{ij} - \pi_i\pi_j) = \hat{V}_1(\hat{\tau}_{HT})$$

A2.2 Resultados da secção 2.5

A2.2.1 Sondagem aleatória estratificada

- Demonstração do resultado (2.5.18):

$$\sigma^2 = \sigma^2_{intra} + \sigma^2_{inter} = \sum_h \frac{N_h}{N}\sigma_h^2 + \sum_h \frac{N_h}{N}(\mu_h - \mu)^2$$

apresentado na secção 2.5.3.

A variância da variável *Y* na população pode ser escrita como (c.f. secção 2.5):

(A2.2.1) $$N\sigma^2 = \sum_h \sum_i (y_{hi} - \mu)^2 \Leftrightarrow$$

$$\Leftrightarrow N\sigma^2 = \sum_h \sum_i (y_{hi} - \mu_h + \mu_h - \mu)^2 \Leftrightarrow$$

$$\Leftrightarrow N\sigma^2 = \sum_h \sum_i (y_{hi} - \mu_h)^2 + 2\sum_h \sum_i (y_{hi} - \mu_h)(\mu_h - \mu) + \sum_h \sum_i (\mu_h - \mu)^2$$

Nesta expressão, a segunda parcela é igual a zero, como se verifica facilmente:

(A2.2.2) $$\sum_h \sum_i (y_{hi} - \mu_h)(\mu_h - \mu) = \sum_h \left[(\mu_h - \mu) \sum_i (y_{hi} - \mu_h) \right] = 0$$

dado que,

(A2.2.3) $$\sum_i (y_{hi} - \mu) = \sum_{i=1}^{N_h} y_{hi} - N_h\mu_h = \sum_{i=1}^{N_h} y_{hi} - \sum_{i=1}^{N_h} y_{hi} = 0$$

Assim, por (A2.2.2), conclui-se que a expressão de σ^2 (A2.2.1) se decompõe em duas parcelas:

(A2.2.4) $$\sigma^2 = \frac{1}{N}\sum_h\sum_i (y_{hi} - \mu_h)^2 + \frac{1}{N}\sum_h\sum_i (\mu_h - \mu)^2 =$$

$$= \frac{1}{N}\sum_h\sum_i (y_{hi} - \mu_h)^2 + \frac{1}{N}\sum_h N_h\,(\mu_h - \mu)^2 =$$

$$= \sigma^2_{intra} + \sigma^2_{inter}$$

Como se pretendia demonstrar.

A2.3 Resultados da secção 3.4

A2.3.1 Estimação em domínios numa sondagem aleatória estratificada

- Demonstração do resultado (3.4.30)

$$\hat{V}(\hat{\tau}_{d\pi_{str}}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \frac{1}{n_h - 1} \left[\sum_{i \in s_{dh}} (y_i - \bar{y}_{s_{dh}})^2 + n_{dh} \left(1 - \frac{n_{dh}}{n_h}\right) \bar{y}_{s_{dh}}^2 \right]$$

apresentado na secção 3.4.2.1.

Para um plano de sondagem genérico, o resultado (3.4.16) fornece um estimador de $V(\hat{\tau}_{d\pi})$:

(A2.3.1) $$\hat{V}(\hat{\tau}_{d\pi}) = \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Como já foi observado, este resultado obtém-se a partir das propriedades do estimador de Horvitz-Thompson (veja-se a secção 2.4), nomeadamente através de

(A2.3.2) $$\hat{V}(\hat{\tau}_{\pi}) = \sum\sum_{s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

utilizando-se os valores da variável Y_d definidos por (3.4.12)

(A2.3.3) $$y_{di} = y_i \mathbb{I}_{i \in U_d} = \begin{cases} y_i & \text{se } i \in U_d \\ 0 & \text{caso contrário} \end{cases}$$

no lugar dos valores y_i.

Analogamente, quando se pretende obter o estimador (A2.3.1) para um plano de sondagem aleatória estratificada (SASSR nos estratos), basta substituir os

valores y_i por y_{di} na expressão do estimador (A2.3.2) para o plano de sondagem em apreço, ou seja, em (c.f. secção 2.5):

$$\text{(A2.3.4)} \qquad \hat{V}(\hat{\tau}_{\pi_{str}}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_h^2$$

onde,

$$\text{(A2.3.5)} \qquad s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_i - \bar{y}_h)^2$$

e

$$\text{(A2.3.6)} \qquad \bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$$

Vamos começar por desenvolver o somatório da expressão (A2.3.5), por forma a poder-se substituir y_i pelos valores y_{di}, definidos por (A2.3.3).

$$\text{(A2.3.7)} \qquad \sum_{i \in s_h} (y_i - \bar{y}_h)^2 = \sum_{s_h} (y_i^2 - 2\bar{y}_h y_i + \bar{y}_h^2) =$$

$$= \sum_{s_h} y_i^2 - 2n_h \bar{y}_h^2 + n_h \bar{y}_h^2 =$$

$$= \sum_{s_h} y_i^2 - n_h \bar{y}_h^2 =$$

$$= \sum_{s_h} y_i^2 - \frac{1}{n_h} \left(\sum_{s_h} y_i \right)^2$$

Substituindo y_i por y_{di} em (A2.3.7) e denotando $\bar{y}_{s_{dh}} = \frac{1}{n_{dh}} \sum_{s_{dh}} y_i$ (dado por 3.4.31) por $\bar{y}_{dh}$, obtém-se:

$$\text{(A2.3.8)} \qquad \sum_{s_h} y_{di}^2 - \frac{1}{n_h} \left(\sum_{s_h} y_{di} \right)^2 = \sum_{s_{dh}} y_i^2 - \frac{1}{n_h} \left(\sum_{s_{dh}} y_i \right)^2 =$$

$$= \sum_{s_{dh}} y_i^2 - 2n_{dh}\bar{y}_{dh}^2 + n_{dh}\bar{y}_{dh}^2 + n_{dh}\bar{y}_{dh}^2 - \frac{1}{n_h} n_{dh}^2 \bar{y}_{dh}^2 =$$

$$= \sum_{s_{dh}} (y_i^2 - 2\bar{y}_{dh} y_i + \bar{y}_{dh}^2) + n_{dh}\left(1 - \frac{n_{dh}}{n_h}\right)\bar{y}_{dh}^2 =$$

$$= \sum_{s_{dh}} (y_i - \bar{y}_{dh})^2 + n_{dh}\left(1 - \frac{n_{dh}}{n_h}\right)\bar{y}_{dh}^2$$

Substituindo (A2.3.8) em (A2.3.4) obtém-se o estimador de $V(\hat{\tau}_{d\pi})$, apresentado em (3.4.30), para um plano de sondagem estratificada (SASSR em cada estrato):

(A2.3.9) $$\hat{V}(\hat{\tau}_{d\pi_{str}}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \frac{1}{n_h - 1} \left[\sum_{s_{dh}} (y_i - \bar{y}_{dh})^2 + n_{dh}\left(1 - \frac{n_{dh}}{n_h}\right)\bar{y}_{dh}^2\right]$$

- Demonstração do resultado (3.4.36)

$$\hat{V}(\hat{\mu}_{d_{str}})=\frac{1}{\hat{N}_d^{\,2}}\sum_{h=1}^{H}N_h^{\,2}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\left[\sum_{s_{dh}}(y_i-\bar{y}_{s_{dh}})^2+\left(n_{dh}-\frac{n_{dh}^{\,2}}{n_h}\right)(\bar{y}_{s_{dh}}-\hat{\mu}_{d_{str}})^2\right]$$

apresentado na secção 3.4.2.1.

Para um plano de sondagem genérico, o resultado (3.4.16) fornece um estimador de $V(\hat{\tau}_{d\pi})$:

(A2.3.10) $$\hat{V}(\hat{\tau}_{d\pi})=\sum\sum_{s_d}\frac{\Delta_{ij}}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

e o resultado (3.4.30) apresenta este estimador para um plano de sondagem aleatória estratificada (SASSR nos estratos):

(A2.3.11)
$$\hat{V}(\hat{\tau}_{d\pi_{str}})=\sum_{h=1}^{H}N_h^{\,2}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\left[\sum_{i\in s_{dh}}(y_i-\bar{y}_{s_{dh}})^2+n_{dh}\left(1-\frac{n_{dh}}{n_h}\right)\bar{y}_{s_{dh}}^{\,2}\right]$$

Para um plano de sondagem genérico, o estimador de $AV(\hat{\mu}_d)$ é dado por (3.4.20):

(A2.3.12) $$\hat{V}(\hat{\mu}_d)=\frac{1}{\hat{N}_d^{\,2}}\sum\sum_{s_d}\frac{\Delta_{ij}}{\pi_{ij}}\frac{y_i-\hat{\mu}_d}{\pi_i}\frac{y_j-\hat{\mu}_d}{\pi_j}$$

Assim, para se obter o estimador (A2.3.12) para um plano de sondagem aleatória estratificada (SASSR nos estratos), $\hat{V}(\hat{\mu}_{d_{str}})$, substitui-se em (A2.3.11) y_i por $y_i - \hat{\mu}_{d_{str}}$ e multiplica-se esse resultado por $(1/\hat{N}_d)^2$, com $\hat{N}_d$ dado por (3.4.33). Uma vez que os passos seguintes envolvem uma certa complexidade algébrica, vamos começar por fazer a substituição em (A2.3.8),

(A2.3.13) $$\sum_{s_{dh}} y_i^2 - \frac{1}{n_h}\left(\sum_{s_{dh}} y_i\right)^2$$

que corresponde à expressão entre parentesis rectos de (A2.3.11):

(A2.3.14) $$\sum_{s_{dh}} (y_i - \hat{\mu}_{d_{str}})^2 - \frac{1}{n_h}\left[\sum_{s_{dh}} \left(y_i - \hat{\mu}_{d_{str}}\right)\right]^2 =$$

$$= \sum_{s_{dh}} y_i^2 - 2\hat{\mu}_{d_{str}} \sum_{s_{dh}} y_i + n_{dh}\hat{\mu}_{d_{str}}{}^2 - \frac{1}{n_h}\left[\sum_{s_{dh}} y_i - n_{dh}\hat{\mu}_{d_{str}}\right]^2$$

Sendo $\bar{y}_{dh} = \frac{1}{n_{dh}}\sum_{s_{dh}} y_i$, a expressão (A2.3.14) vem igual a:

(A2.3.15) $$\sum_{s_{dh}} y_i^2 - 2\hat{\mu}_{d_{str}}\, n_{dh}\bar{y}_{dh} + n_{dh}\hat{\mu}_{d_{str}}{}^2 -$$

$$- \frac{1}{n_h}(n_{dh}{}^2\bar{y}_{dh}{}^2 - 2n_{dh}{}^2\bar{y}_{dh}\,\hat{\mu}_{d_{str}} + n_{dh}{}^2\hat{\mu}_{d_{str}}{}^2)$$

Esta expressão não se altera se somarmos $(-2n_{dh}\bar{y}_{dh}{}^2 + n_{dh}\bar{y}_{dh}{}^2 + n_{dh}\bar{y}_{dh}{}^2)$. Assim, (A2.3.15) vem igual a:

(A2.3.16) $$\sum_{s_{dh}} y_i^2 - 2n_{dh}\bar{y}_{dh}{}^2 + n_{dh}\bar{y}_{dh}{}^2 + n_{dh}\bar{y}_{dh}{}^2 - 2n_{dh}\hat{\mu}_{d_{str}}\,\bar{y}_{dh} + n_{dh}\hat{\mu}_{d_{str}}{}^2 -$$

$$- \frac{n_{dh}{}^2}{n_h}(\bar{y}_{dh}{}^2 - 2\bar{y}_{dh}\,\hat{\mu}_{d_{str}} + \hat{\mu}_{d_{str}}{}^2) =$$

$$= (\sum_{s_{dh}} y_i^2 - 2\bar{y}_{dh}\sum_{s_{dh}} y_i + n_{dh}\bar{y}_{dh}{}^2) +$$

$$+ n_{dh}(\bar{y}_{dh}{}^2 - 2\hat{\mu}_{d_{str}}\,\bar{y}_{dh} + \hat{\mu}_{d_{str}}{}^2) - \frac{n_{dh}{}^2}{n_h}(\bar{y}_{dh} - \hat{\mu}_{d_{str}})^2 =$$

$$= \sum_{s_{dh}} (y_i^2 - 2\bar{y}_{dh}y_i + \bar{y}_{dh}{}^2) + n_{dh}(\bar{y}_{dh} - \hat{\mu}_{d_{str}})^2 -$$

$$- \frac{n_{dh}{}^2}{n_h}(\bar{y}_{dh} - \hat{\mu}_{d_{str}})^2 =$$

$$= \sum_{s_{dh}} (y_i - \bar{y}_{dh})^2 + \left(n_{dh} - \frac{n_{dh}{}^2}{n_h}\right)(\bar{y}_{dh} - \hat{\mu}_{d_{str}})^2$$

Substituindo (A2.3.16) em (A2.3.11) e multiplicando toda a expressão por $(1/\hat{N}_d)^2$, obtém-se o estimador de AV($\hat{\mu}_{d_{str}}$), dado por (3.4.20), para um plano de sondagem estratificada (SASSR em cada estrato). Ou seja, obtém-se o resultado (3.4.36) que se pretendia demonstrar:

(A2.3.17)

$$\hat{V}(\hat{\mu}_{d_{str}}) = \frac{1}{\hat{N}_d^{\,2}} \sum_{h=1}^{H} N_h^{\,2} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{n_h - 1} \left[\sum_{s_{dh}} (y_i - \bar{y}_{dh})^2 + \left(n_{dh} - \frac{n_{dh}^{\,2}}{n_h} \right) (\bar{y}_{dh} - \hat{\mu}_{d_{str}})^2 \right]$$

A2.3.2 Estimação em domínios numa sondagem aleatória simples sem reposição (SASSR)

Nesta secção, apresentam-se alguns resultados relativos aos estimadores em domínios (c.f. secção 3.4.2) para um plano de sondagem aleatória simples sem reposição.

O estimador de Horvitz-Thompson do total do domínio U_d,

(A2.3.18) $\tau_d = \sum_U y_{di} = \sum_{U_d} y_i$, $\quad y_{di} = y_i \mathbb{I}_{i \in U_d} = \begin{cases} y_i & \text{se } i \in U_d \\ 0 & \text{caso contrário} \end{cases}$

é dado por

(A2.3.19) $\hat{\tau}_{d\pi} = \sum_s \frac{y_{di}}{\pi_i} = \sum_{s_d} \frac{y_i}{\pi_i}$

Pelos resultados apresentados na secção 2.3 conclui-se que, para uma sondagem aleatória simples sem reposição, este estimador é dado por:

(A2.3.20) $\hat{\tau}_{d\pi_{sas}} = \frac{N}{n} \sum_s y_{di} = \frac{N}{n} \sum_{s_d} y_i$

Trata-se, obviamente, de um estimador centrado (c.f. secção 2.4). A variância do estimador (A2.3.19) é

(A2.3.21) $V(\hat{\tau}_{d\pi_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} [(N_d - 1) S_{U_d}^2 + N_d(1 - N_d/N)\mu_d^2]$

onde f = 1/N,

(A2.3.22) $S_{U_d}^2 = \frac{1}{N_d - 1} \sum_{U_d} (y_i - \mu_d)^2$

é a variância da variável de interesse *Y* no domínio U_d e

(A2.3.23) $\mu_d = \frac{1}{N_d} \sum_{U_d} y_i = \tau_d / N_d$

é a média de *Y* no domínio U_d.

A variância (A2.3.21) pode ser estimada sem enviesamento por:

(A2.3.24) $\hat{V}(\hat{\tau}_{d\pi_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{n-1} [(n_d - 1) s_{s_d}^2 + n_d(1 - n_d/n) \bar{y}_{s_d}^2]$

onde,

(A2.3.25) $s_{s_d}^2 = \frac{1}{n_d - 1} \sum_{s_d} (y_i - \bar{y}_{s_d})^2$

(A2.3.26) $\bar{y}_{s_d} = \frac{1}{n_d} \sum_{s_d} y_i$

Outro estimador do total do domínio U_d (c.f. secção 3.4.2) é

(A2.3.27) $\hat{\tau}_{dw} = \frac{N_d}{\hat{N}_d} \sum_{s_d} \frac{y_i}{\pi_i}$, $\hat{N}_d = \sum_{s_d} \frac{1}{\pi_i}$

Para um plano de sondagem aleatória simples sem reposição (SASSR), este estimador é dado por:

(A2.3.28) $\hat{\tau}_{dw_{sas}} = \frac{N_d}{\sum_{s_d} \frac{N}{n}} \sum_{s_d} \frac{N}{n} y_i = \frac{N_d}{n_d} \sum_{s_d} y_i$

Pelo resultado (3.4.23), verifica-se que a variância aproximada de $\hat{\tau}_{dw}$ é dada por:

(A2.3.29) $AV(\hat{\tau}_{dw}) = \sum \sum_{U_d} \Delta_{ij} \frac{y_i - \mu_d}{\pi_i} \frac{y_j - \mu_d}{\pi_j}$

e pode ser estimada através de (veja-se resultado (3.4.24))

$$\text{(A2.3.30)} \quad \hat{V}(\hat{\tau}_{dw}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum\sum_{s_d} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i - \hat{\mu}_d}{\pi_i} \frac{y_j - \hat{\mu}_d}{\pi_j}$$

Para uma sondagem aleatória simples sem reposição, o resultado (A2.3.29) obtém-se a partir de (A2.3.21) substituindo-se y_i por $y_i - \mu_d$; e, para esse plano de sondagem, o estimador (A2.3.30) obtém-se a partir de (A2.3.24), substituindo-se y_i por $y_i - \hat{\mu}_d$ e multiplicando-se a expressão que se obtém por $(N_d / \hat{N}_d)^2$. Conclui-se, desta forma, que

$$\text{(A2.3.31)} \quad AV(\hat{\tau}_{dw_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} (N_d - 1) S_{U_d}^2$$

$$\text{(A2.3.32)} \quad \hat{V}(\hat{\tau}_{dw_{sas}}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 N^2 \frac{1-f}{n} \frac{1}{n-1} (n_d - 1) s_{s_d}^2$$

onde, $S_{U_d}^2$ é dado por (A2.3.22), $s_{s_d}^2$ é dado por (A2.3.25) e

$$\text{(A2.3.33)} \quad \hat{N}_d = n_d \frac{N}{n}$$

Obviamente, o cálculo de (A2.3.32) requer que $n_d > 1$.

Em seguida, apresentam-se as demonstrações dos resultados (A2.3.21) e (A2.3.31) (os resultados (A2.3.24) e (A2.3.32) obtém-se, respectivamente, de forma análoga).

- <u>Demonstração do resultado</u> (A2.3.21)

$$V(\hat{\tau}_{d\pi_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} [(N_d - 1) S_{U_d}^2 + N_d(1 - N_d/N)\mu_d^2]$$

A expressão de $V(\hat{\tau}_{d\pi_{sas}})$ obtém-se a partir do resultado (2.3.25):

(A2.3.34) $$V(\hat{\tau}_{\pi_{sas}}) = N^2 \frac{1-f}{n} S^2$$

onde,

(A2.3.35) $$S^2 = \frac{1}{N-1} \sum_U (y_i - \mu)^2$$

(A2.3.36) $$\mu = \frac{1}{N} \sum_U y_i$$

substituindo-se os valores y_i pelos valores y_{di} dados por (c.f. secção 3.4.1):

(A2.3.37) $$y_{di} = y_i \mathbb{I}_{i \in U_d} = \begin{cases} y_i & \text{se } i \in U_d \\ 0 & \text{caso contrário} \end{cases}$$

Antes de se efectuar essa substituição, vamos começar por simplificar o somatório da expressão (A2.3.35):

(A2.3.38) $$\sum_U (y_i - \mu)^2 = \sum_U y_i^2 - N\mu^2 = \sum_U y_i^2 - \frac{1}{N} \left(\sum_U y_i \right)^2$$

Substituindo y_i por y_{di} em (A2.3.38), vem:

(A2.3.39) $$\sum_U y_{di}^2 - \frac{1}{N} \left(\sum_U y_{di} \right)^2 = \sum_{U_d} y_i^2 - \frac{1}{N} \left(\sum_{U_d} y_i \right)^2 =$$

$$= \sum_{U_d} y_i^2 - \frac{1}{N_d} \left(\sum_{U_d} y_i \right)^2 + \frac{1}{N_d} \left(\sum_{U_d} y_i \right)^2 - \frac{1}{N} \left(\sum_{U_d} y_i \right)^2 =$$

$$= \sum_{U_d} y_i^2 - \frac{2}{N_d} \left(\sum_{U_d} y_i \right)^2 + \frac{1}{N_d} \left(\sum_{U_d} y_i \right)^2 + \left(\frac{1}{N_d} - \frac{1}{N} \right) \left(\sum_{U_d} y_i \right)^2 =$$

$$= \sum_{U_d} y_i^2 - 2\mu_d \sum_{U_d} y_i + N_d \mu_d^2 + \left(\frac{1}{N_d} - \frac{1}{N} \right) N_d^2 \mu_d^2 =$$

$$= \sum_{U_d} (y_i^2 - 2\mu_d y_i + \mu_d^2) + \left(\frac{1}{N_d} - \frac{1}{N}\right) N_d^2 \mu_d^2 =$$

$$= \sum_{u_d} (y_i - \mu_d)^2 + N_d \left(1 - \frac{N_d}{N}\right) \mu_d^2$$

Substituindo esta expressão em (A2.3.35) e em (A2.3.34), obtém-se o resultado pretendido:

(A2.3.40) $V(\hat{\tau}_{d\pi_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} [(N_d - 1) S_{U_d}^2 + N_d(1 - N_d/N)\mu_d^2]$

- Demonstração do resultado (A2.3.31)

$$AV(\hat{\tau}_{dw_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} (N_d - 1) S_{U_d}^2$$

Para se obter o resultado (A2.3.31), substitui-se y_i por $y_i - \mu_d$ em (A2.3.21), ou seja, em:

(A2.3.41) $V(\hat{\tau}_{d\pi_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} [(N_d - 1) S_{U_d}^2 + N_d(1 - N_d/N)\mu_d^2]$

Por uma questão de simplicidade de apresentação, vamos efectuar esta substituição em (A2.3.39), uma vez que este resultado fornece uma expressão simplificada das parcelas que se encontram entre parêntesis rectos de (A2.3.41):

(A2.3.42) $$\sum_{u_d} (y_i - \mu_d)^2 + N_d \left(1 - \frac{N_d}{N}\right) \mu_d^2 = \sum_{U_d} y_i^2 - \frac{1}{N} \left(\sum_{U_d} y_i\right)^2$$

Efectuando a substituição, obtém-se, então,

(A2.3.43) $$\sum_{U_d} (y_i - \mu_d)^2 - \frac{1}{N} \left(\sum_{U_d} (y_i - \mu_d)\right)^2 =$$

$$= \sum_{U_d} (y_i - \mu_d)^2 - \frac{1}{N}\left(\sum_{U_d} y_i - N_d\mu_d)\right)^2 =$$

$$= \sum_{U_d} (y_i - \mu_d)^2 - \frac{1}{N}(N_d\mu_d - N_d\mu_d)^2 =$$

$$= \sum_{U_d} (y_i - \mu_d)^2$$

Substituindo a expressão que se encontra entre parêntesis rectos de (A2.3.41), por (A2.3.43), obtém-se o resultado pretendido:

(A2.3.44) $$AV(\hat{\tau}_{dw_{sas}}) = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_{U_d} (y_i - \mu_d)^2 = N^2 \frac{1-f}{n} \frac{1}{N-1} (N_d - 1) S^2_{U_d}$$

A2.4 Resultados da secção 3.5

A2.4.1 Estimador da variância do estimador de pós-estratificação, proposto por Rao (1985)

- Demonstração do resultado (3.5.66)

$$\hat{V}_{rao}(\hat{\tau}_{ps,sas}) = (1-f)\sum_{i=1}^{L} N_i^2 \frac{n}{n-1}\frac{n_i-1}{n_i}\frac{s_i^2}{n_i}$$

apresentado na secção 3.5.3.

Para um plano de sondagem aleatória simples sem reposição, o estimador usual da variância dos estimador do total da população (c.f. secção 2.3.2) é dado por:

(A2.4.1) $$\hat{V}(\hat{\tau}_{\pi_{sas}}) = N^2(1-f)\frac{s^2}{n}$$

onde, f = n/N e

(A2.4.2) $$s^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \bar{y})^2$$

Vamos começar por desenvolver o somatório da expressão de s^2 por forma a substituir-se, em seguida, y_k por $z_k = \frac{N_i}{\hat{N}_i}(y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_i \mathbb{I}_{k\in s_i})$ (veja-se o resultado 3.5.59).

(A2.4.3) $$(n-1)s^2 = \sum_{k\in s}(y_k - \bar{y})^2 =$$

$$= \sum_{k\in s} y_k^2 - 2n\bar{y}^2 + n\bar{y}^2 =$$

$$= \sum_{k\in s} y_k^2 - n\left(\frac{1}{n}\sum_{k\in s} y_k\right)^2 =$$

$$= \sum_{k\in s} y_k^{\,2} - \frac{1}{n}\left(\sum_{k\in s} y_k\right)^2$$

Efectuando-se a referida substituição na expressão que se encontra à direita do sinal de igual de (A2.4.3), obtém-se:

(A2.4.4) $$\sum_{k\in s}\left[\frac{N_i}{\hat{N}_i}(y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_i \mathbb{I}_{k\in s_i})\right]^2 - \frac{1}{n}\left[\sum_{k\in s}\left(\frac{N_i}{\hat{N}_i}(y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_i \mathbb{I}_{k\in s_i})\right)\right]^2 =$$

$$= \left(\frac{N_i}{\hat{N}_i}\right)^2 \left\{\sum_{k\in s}\left[(y_k - \hat{\mu}_i)^2 (\mathbb{I}_{k\in s_i})^2\right] - \frac{1}{n}\left[\sum_{k\in s}(y_k - \hat{\mu}_i)\mathbb{I}_{k\in s_i}\right]^2\right\} =$$

$$= \left(\frac{N_i}{\hat{N}_i}\right)^2 \left\{\sum_{k\in s_i}(y_k - \hat{\mu}_i)^2 - \frac{1}{n}\left[\sum_{k\in s_i}(y_k - \hat{\mu}_i)\right]^2\right\} =$$

$$= \left(\frac{N_i}{\hat{N}_i}\right)^2 \left\{\sum_{k\in s_i}(y_k - \hat{\mu}_i)^2 - \frac{1}{n}\left[\sum_{k\in s_i} y_k - n_i\hat{\mu}_i\right]^2\right\} =$$

Uma vez que, no caso SASSR, se tem $\hat{\mu}_i = \frac{1}{n_i}\sum_{k\in s_i} y_k = \bar{y}_{s_i}$, conclui-se que a última parcela desta expressão é igual a zero e, portanto, (A2.4.4) simplifica-se para:

(A2.4.5) $$\left(\frac{N_i}{\hat{N}_i}\right)^2 \sum_{k\in s_i}(y_k - \bar{y}_{s_i})^2$$

Por outro lado, uma vez que para a SASSR se tem

(A2.4.6) $$\hat{N}_i = \sum_{k\in s_i}\frac{N}{n} = n_i\frac{N}{n}$$

a expressão (A2.4.5) vem igual a

(A2.4.7) $$N_i^2 \left(\frac{1}{n_i}\frac{n}{N}\right)^2 (n_i - 1)s_i^2$$

onde,

(A2.4.8) $$s_i^2 = \frac{1}{n_i - 1}\sum_{k \in s_i}(y_k - \bar{y}_{s_i})^2$$

Substituindo-se (A2.4.7) em (A2.4.1), obtém-se o resultado que se pretendia demonstrar:

(A2.4.9) $$\hat{V}_{rao}(\hat{\tau}_{ps,sas}) = \sum_{i=1}^{L} N^2 (1 - f)\frac{1}{n}\frac{1}{n-1} N_i^2 \left(\frac{1}{n_i}\frac{n}{N}\right)^2 (n_i - 1)s_i^2 =$$

$$= (1 - f) \sum_{i=1}^{L} N_i^2 \frac{n}{n-1}\frac{n_i - 1}{n_i}\frac{s_i^2}{n_i}$$

- Demonstração do resultado (3.5.74)

$$\hat{V}_{rao}(\hat{\tau}_{ps,str}) =$$

$$\sum_{i=1}^{L}\sum_{h=1}^{H}\left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 N_{\bullet h}^{\,2}\left(\frac{1}{n_{\bullet h}}-\frac{1}{N_{\bullet h}}\right)\frac{1}{n_{\bullet h}-1}\left\{\sum_{k\in s_{ih}}(y_k-\bar{y}_{s_{ih}})^2+(n_{ih}-\frac{n_{ih}^{\,2}}{n_{\bullet h}})(\bar{y}_{s_{ih}}-\hat{\mu}_{i_{str}})^2\right\}$$

apresentado na secção 3.5.3.

Utilizando-se a notação apresentada na secção 3.5.3, o estimador usual da variância dos estimador do total da população, para um plano de sondagem aleatória estratificada (SASSR nos estratos), c.f. secção 2.5.2, é dado por:

(A2.4.10) $$\hat{V}(\hat{\tau}_{\pi_{str}}) = \sum_{h=1}^{H} N_{\bullet h}^{\,2}\left(\frac{1}{n_{\bullet h}}-\frac{1}{N_{\bullet h}}\right)s_h^2$$

onde,

(A2.4.11) $$s_h^2 = \frac{1}{n_{\bullet h}-1}\sum_{k\in s_h}(y_k-\bar{y}_h)^2$$

(A2.4.12) $$\bar{y}_h = \frac{1}{n_{\bullet h}}\sum_{k\in s_h} y_k$$

Vamos começar por desenvolver o somatório da expressão (A2.4.11) por forma a substituir-se, em seguida, y_k por $z_k = \frac{N_{i\bullet}}{\hat{N}_{i\bullet}}(y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_{i_{str}} \mathbb{I}_{k\in s_i})$ (veja-se o resultado 3.5.59). Por A2.3.7 obtém-se

(A2.4.13) $$(n_{\bullet h}-1)s_h^2 = \sum_{k\in s_h}(y_k-\bar{y}_h)^2 = \sum_{k\in s_h} y_k^{\,2} - \frac{1}{n_{\bullet h}}\left[\sum_{k\in s_h} y_k\right]^2$$

Efectuando-se a referida substituição na expressão que se encontra à direita do sinal de igual de (A2.4.13), obtém-se:

$$\text{(A2.4.14)}\quad \sum_{k\in s_h}\left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 (y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_{i_{str}} \mathbb{I}_{k\in s_i})^2 - \frac{1}{n_{\bullet h}}\Big[\frac{N_{i\bullet}}{\hat{N}_{i\bullet}} \sum_{k\in s_h}(y_k \mathbb{I}_{k\in s_i} - \hat{\mu}_{i_{str}} \mathbb{I}_{k\in s_i})\Big]^2 =$$

$$= \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \Big\{ \sum_{k\in s_{ih}}(y_k - \hat{\mu}_{i_{str}})^2 - \frac{1}{n_{\bullet h}}\Big[\sum_{k\in s_{ih}} y_k - \sum_{k\in s_{ih}} \hat{\mu}_{i_{str}} \Big]^2 \Big\} =$$

$$= \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \Big\{ \sum_{k\in s_{ih}} y_k^{\,2} - 2\hat{\mu}_{i_{str}} \sum_{k\in s_{ih}} y_k + n_{ih}\hat{\mu}_{i_{str}}^{\,2} - \frac{1}{n_{\bullet h}}\Big[\sum_{k\in s_{ih}} y_k - n_{ih}\hat{\mu}_{i_{str}}\Big]^2\Big\} =$$

Sendo $\bar{y}_{s_{ih}} = \frac{1}{n_{ih}}\sum_{s_{ih}} y_k$, a expressão (A2.4.14) vem igual a:

$$\text{(A2.4.15)}\quad \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \Big\{ \sum_{k\in s_{ih}} y_k^{\,2} - 2n_{ih}\bar{y}_{s_{ih}}\,\hat{\mu}_{i_{str}} + n_{ih}\hat{\mu}_{i_{str}}^{\,2} - \frac{1}{n_{\bullet h}}\Big[n_{ih}^{\,2}\,\bar{y}_{s_{ih}}^{\,2} - 2n_{ih}^{\,2}\bar{y}_{s_{ih}}\,\hat{\mu}_{i_{str}} + n_{ih}^{\,2}\hat{\mu}_{i_{str}}^{\,2}\Big]\Big\}$$

Esta expressão não se altera se somarmos $(-2n_{ih}\bar{y}_{s_{ih}}^{\,2} + n_{ih}\bar{y}_{s_{ih}}^{\,2} + n_{ih}\bar{y}_{s_{ih}}^{\,2})$ e, portanto, (A2.4.15) vem igual a:

$$\text{(A2.4.16)}\quad \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \Big\{ \sum_{k\in s_{ih}} y_k^{\,2} - 2n_{ih}\bar{y}_{s_{ih}}^{\,2} + n_{ih}\bar{y}_{s_{ih}}^{\,2} + n_{ih}\bar{y}_{s_{ih}}^{\,2} - 2n_{ih}\bar{y}_{s_{ih}}\,\hat{\mu}_{i_{str}} + n_{ih}\hat{\mu}_{i_{str}}^{\,2} - \frac{n_{ih}^{\,2}}{n_{\bullet h}}\Big[\bar{y}_{s_{ih}}^{\,2} - 2\bar{y}_{s_{ih}}\,\hat{\mu}_{i_{str}} + \hat{\mu}_{i_{str}}^{\,2}\Big]\Big\} =$$

$$= \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \Big\{ \sum_{k\in s_{ih}}(y_k^{\,2} - 2\bar{y}_{s_{ih}}\,y_k + \bar{y}_{s_{ih}}^{\,2}) + n_{ih}(\bar{y}_{s_{ih}}^{\,2} - 2\bar{y}_{s_{ih}}\,\hat{\mu}_{i_{str}} + \hat{\mu}_{i_{str}}^{\,2}) - \frac{n_{ih}^{\,2}}{n_{\bullet h}}(\bar{y}_{s_{ih}} - \hat{\mu}_{i_{str}})^2 \Big\} =$$

$$= \left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 \left\{ \sum_{k\in s_{ih}} (y_k - \bar{y}_{s_{ih}})^2 + (n_{ih} - \frac{n_{ih}^2}{n_{\bullet h}})(\bar{y}_{s_{ih}} - \hat{\mu}_{i_{str}})^2 \right\}$$

Substituindo-se (A2.4.16) em (A2.4.10), obtém-se o resultado que se pretendia demonstrar:

(A2.4.17) $\hat{V}_{rao}(\hat{\tau}_{ps,str}) =$

$$\sum_{i=1}^{L}\sum_{h=1}^{H}\left(\frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\right)^2 N_{\bullet h}^2\left(\frac{1}{n_{\bullet h}}-\frac{1}{N_{\bullet h}}\right)\frac{1}{n_{\bullet h}-1}\left\{\sum_{k\in s_{ih}}(y_k-\bar{y}_{s_{ih}})^2+(n_{ih}-\frac{n_{ih}^2}{n_{\bullet h}})(\bar{y}_{s_{ih}}-\hat{\mu}_{i_{str}})^2\right\}$$

ANEXO 3 – Classificação Portuguesa das Actividades Económicas CAE - REV. 2

A3.1 Designações da CAE – Rev. 2, por secções

Secção	Designação
A	Agricultura, produção animal, caça e silvicultura
B	Pesca
C	Indústrias extractivas
D	Indústrias transformadoras
E	Produção e distribuição de electricidade, de gás e de água
F	Construção
G	Comércio por grosso e a retalho; reparação de veículos automóveis, motociclos e de bens de uso pessoal e doméstico
H	Alojamento e restauração (restaurantes e similares)
I	Transportes, armazenagem e comunicações
J	Actividades financeiras
K	Actividades imobiliárias, alugueres e serviços prestados às empresas
L	Administração pública, defesa e segurança social obrigatória
M	Educação
N	Saúde e acção social
O	Outras actividades de serviços colectivos, sociais e pessoais
P	Famílias com empregados domésticos
Q	Organismos internacionais e outras instituições extra-territoriais

A3.2 Designações da CAE – Rev. 2, por divisões

Divisão	Designação	Secção
01	Agricultura, produção animal, caça e actividades dos serviços relacionados	A
02	Silvicultura, exploração florestal e actividades dos serviços relacionados	A
05	Pesca, aquacultura e actividades dos serviços relacionados	B
10	Extracção de hulha, linhite e turfa	C
11	Extracção de petróleo bruto, gás natural e actividades dos serviços relacionados excepto a prospecção	C
12	Extracção de minérios e urânio e de tório	C
13	Extracção e preparação de minérios metálicos	C
14	Outras indústrias extractivas	C
15	Indústrias alimentares e de bebidas	D
16	Indústria do tabaco	D
17	Fabricação de têxteis	D
18	Indústria do vestuário; preparação, tingimento e fabricação de artigos e peles com pelo	D
19	Curtimento e acabamento de peles sem pelo; fabricação de artigos de viagem, marroquinaria, artigos de correeiro, seleiro e calçado	D
20	Indústrias da madeira e da cortiça e suas obras, excepto mobiliário; fabricação de obras de cestaria e de espartaria	D
21	Fabricação de pasta, de papel e cartão e seus artigos	D
22	Edição, impressão e reprodução de suportes de informação gravados	D
23	Fabricação de coque, produtos petrolíferos refinados e tratamento de combustível nuclear	D
24	Fabricação de produtos químicos	D
25	Fabricação de artigos de borracha e de matérias plásticas	D
26	Fabricação de outros produtos minerais não metálicos	D
27	Indústrias metalúrgicas de base	D
28	Fabricação de produtos metálicos, excepto máquinas e equipamento	D
29	Fabricação de máquinas e equipamento, N.E.	D
30	Fabricação de máquinas de escritório e de equipamento para o tratamento automático da informação	D
31	Fabricação de máquinas e aparelhos eléctricos, N.E.	D
32	Fabricação de equipamento e de aparelhos de rádio, televisão e comunicação	D

Divisão	Designação	Secção
33	Fabricação de aparelhos e instrumentos médico-cirúrgicos, ortopédicos, de precisão, de óptica e de relojoaria	D
34	Fabricação de veículos automóveis, reboques e semi-reboques	D
35	Fabricação de outro material de transporte	D
36	Fabricação de mobiliário; outras indústrias transformadoras, N.E.	D
37	Reciclagem	D
40	Produção e distribuição de electricidade, gás e água	E
41	Captação, tratamento e distribuição de água	E
45	Construção	F
50	Comércio, manutenção e reparação de veículos automóveis, motociclos; comércio a retalho de combustíveis para veículos	G
51	Comércio por grosso e agentes do comércio, excepto de veículos automóveis e de motociclos	G
52	Comércio a retalho (excepto de veículos automóveis, motociclos e combustíveis para veículos), reparação de bens pessoais e domésticos	G
55	Alojamento e restauração (restaurantes e similares)	H
60	Transportes terrestres; transportes por oleodutos ou gasodutos (pipelines)	I
61	Transportes por água	I
62	Transportes aéreos	I
63	Actividades anexas e auxiliares dos transportes; agentes de viagem e de turismo	I
64	Correios e telecomunicações	I
65	Intermediação financeira, excepto seguros e fundos de pensões	J
66	Seguros, fundos de pensões e de outras actividades complementares de segurança social	J
67	Actividades auxiliares de intermediação financeira	J
70	Actividades imobiliárias	K
71	Aluguer de máquinas e de equipamentos sem pessoal e de bens pessoais e domésticos	K
72	Actividades informáticas e conexas	K
73	Investigação e desenvolvimento	K
74	Outras actividades de serviços prestados principalmente às empresas	K
75	Administração pública, defesa e segurança social obrigatória	L

Divisão	Designação	Secção
80	Educação	M
85	Saúde e acção social	N
90	Saneamento, higiene pública e actividades similares	O
91	Actividades associativas diversas, N.E.	O
92	Actividades recreativas, culturais e desportivas	O
93	Outras actividades de serviços	O
95	Famílias com agregados domésticos	P
99	Organismos internacionais e outras instituições extra-territoriais	Q

ANEXO 4 – Variáveis de estratificação do IEH

A4.1 Escalões de NUTS II (*ENUT*)

Valor	Descrição
101	Norte
102	Centro
103	Lisboa e Vale do Tejo
104	Alentejo
105	Algarve
201	Açores
301	Madeira

A4.2 Escalões de número de pessoas ao serviço (*ENPS*)

Valor	Descrição
0	0 pessoas ao serviço
1	1 a 9 pessoas ao serviço
2	10 a 19 pessoas ao serviço
3	20 a 49 pessoas ao serviço
4	50 a 99 pessoas ao serviço
5	100 a 249 pessoas ao serviço
6	250 a 499 pessoas ao serviço
7	500 e mais pessoas ao serviço

A4.3 Escalões de forma jurídica (*EFJR*)

Valor	Descrição
1	Empresas do sector público
2	Empresas privadas
3	Empresário em nome individual

A4.4 Escalões de volume de vendas (*EVVN*)

Valor	Descrição
1	VVN ≤ 30 000 mil escudos
2	VVN > 30 000 mil escudos

A4.5 Escalões de Classificação Portuguesa das Actividades Económicas CAE – Rev. 2

Só são consideradas no Universo e Amostra do IEH as empresas que no FGUE se encontrem classificadas ao nível máximo de desagregação da CAE – Rev. 2. A representatividade por actividade tem em conta os seguintes escalões:

01100	15310	15892	17510	21120	24200	26250	27510
01200	15320	15893	17520	21210	24300	26260	27520
01300	15331	15911	17530	21220	24410	26300	27530
01400	15332	15912	17541	21230	24420	26400	27540
01500	15333	15913	17542	21240	24510	26510	28110
02010	15334	15920	17543	21250	24520	26520	28120
02020	15335	15931	17544	22110	24610	26530	28210
05010	15411	15932	17600	22120	24620	26610	28220
05020	15412	15940	17710	22130	24630	26620	28300
10100	15413	15950	17720	22140	24640	26630	28400
10200	15420	15960	18100	22150	24650	26640	28510
10300	15430	15970	18210	22210	24660	26650	28520
11100	15510	15981	18220	22220	24700	26660	28610
11200	15520	15982	18230	22230	25110	26700	28620
12000	15611	16000	18240	22240	25120	26810	28630
13100	15612	17110	18300	22250	25130	26820	28710
13200	15613	17120	19100	22310	25210	27100	28720
14110	15620	17130	19200	22320	25220	27210	28730
14120	15710	17140	19301	22330	25230	27220	28740
14130	15720	17150	19302	23100	25240	27310	28750
14210	15810	17160	20100	23200	26110	27320	29110
14220	15820	17170	20200	23300	26120	27330	29120
14300	15830	17210	20300	24110	26130	27340	29130
14400	15840	17220	20400	24120	26140	27350	29140
14500	15850	17230	20511	24130	26150	27410	29210
15110	15860	17240	20512	24140	26210	27420	29220
15120	15870	17250	20521	24150	26220	27430	29230
15130	15880	17300	20522	24160	26230	27440	29240
15200	15891	17400	21110	24170	26240	27450	29310

29320	36632	51380	55210	74820
29400	36633	51390	55220	74830
29510	36634	51410	55230	74840
29520	36635	51420	55300	80101
29530	36636	51430	55400	80102
29540	37100	51440	55510	80211
29550	37200	51450	55520	80212
29561	40100	51460	60100	80220
29562	40200	51470	60210	80300
29563	40300	51510	60220	80410
29564	41000	51520	60230	80420
29600	45110	51530	60240	85110
29710	45120	51540	60300	85120
29720	45211	51550	61100	85130
30010	45212	51560	61200	85140
30020	45220	51570	62100	85200
31100	45230	51610	62200	85300
31200	45240	51620	62300	90000
31300	45250	51630	63100	91100
31400	45310	51640	63210	91200
31500	45320	51650	63220	91300
31610	45330	51660	63230	92100
31620	45340	51700	63300	92200
32100	45410	52111	63400	92310
32200	45420	52112	64110	92320
32300	45430	52120	64120	92330
33100	45440	52210	64200	92340
33200	45450	52220	70100	92400
33300	45500	52230	70200	92500
33400	50100	52240	70300	92600
33500	50200	52250	71100	92710
34100	50300	52260	71210	92720
34200	50400	52270	71220	93010
34300	50500	52310	71230	93020
35110	51110	52320	71300	93030
35120	51120	52330	71400	93040
35200	51130	52410	72100	93050
35300	51140	52420	72200	
35410	51150	52431	72300	
35420	51160	52432	72400	
35430	51170	52440	72500	
35500	51180	52450	72600	
36110	51190	52460	73000	
36120	51210	52470	74110	
36130	51220	52480	74120	
36140	51230	52500	74130	
36150	51240	52610	74140	
36210	51250	52620	74150	
36220	51310	52630	74200	
36300	51320	52710	74300	
36400	51330	52720	74400	
36500	51340	52730	74500	
36610	51350	52740	74600	
36620	51360	55110	74700	
36631	51370	55120	74810	

ANEXO 5 – Histogramas das réplicas bootstrap

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador Ponderação em classes da média
Variável: Q20001

M20001PC

Midpoint		Freq
2.730	,	2
2.742	,*	5
2.754	,*******	34
2.766	,**********	61
2.778	,**********	110
2.790	,**********	118
2.802	,**********	120
2.814	,**********	79
2.826	,**********	51
2.838	,***	14
2.850	,*	3
2.862	,*	3

```
          ,
          Šfffˆfffˆfffˆfffˆfffˆfffˆ
              20  40  60  80 100 120

                    Frequency
```

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador Ponderação em classes da média
Variável: Q4160

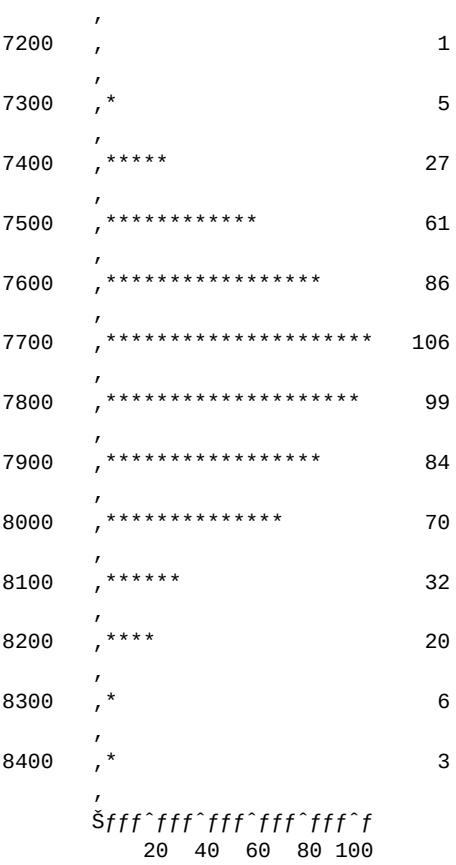
M4160PC
Midpoint Freq

```
        ,
  25800 ,                          1
        ,
  26100 ,***                      15
        ,
  26400 ,***********              60
        ,
  26700 ,*********************   111
        ,
  27000 ,***********************  115
        ,
  27300 ,***********************  114
        ,
  27600 ,*****************         84
        ,
  27900 ,*********                 45
        ,
  28200 ,********                  38
        ,
  28500 ,**                        10
        ,
  28800 ,*                          5
        ,
  29100 ,                          1
        ,
  29400 ,                          0
        ,
  29700 ,                          1
        ,
        Šfffˆfffˆfffˆfffˆfffˆfff
            20  40  60  80 100

                Frequency
```

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador Ponderação em classes da média
Variável: Q4190

M4190PC
Midpoint Freq



HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média, por ENPS
Variável: Q20001

M20001PS

Midpoint		Freq
2.968	,	1
2.976	,*	3
2.984	,****	20
2.992	,********	38
3.000	,**************	72
3.008	,********************	99
3.016	,************************	120
3.024	,**********************	103
3.032	,****************	80
3.040	,*********	43
3.048	,***	13
3.056	,*	5
3.064	,*	3

```
          ,
          Šfff^fff^fff^fff^fff^fff^
              20  40  60  80 100 120

                    Frequency
```

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média, por ENPS
Variável: Q4160

M4160PS

Midpoint		Freq
27900	*	3
28200	**	9
28500	******	32
28800	**********	66
29100	**********	89
29400	**********	129
29700	**********	101
30000	**********	66
30300	**********	49
30600	******	31
30900	***	15
31200	*	7
31500		2
31800		0
32100		1

20 40 60 80 100 120

Frequency

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média, por ENPS
Variável: Q4190

M4190PS

Midpoint		Freq
7400	*	7
7480	****	19
7560	********	39
7640	**********	86
7720	************************	125
7800	**********************	116
7880	********************	101
7960	**********	61
8040	******	31
8120	**	9
8200	*	4
8280		2

20 40 60 80 100 120

Frequency

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média com ponderação em classes, por ENPS
Variável: Q20001

M201PCPS

Midpoint		Freq
2.956	,	1
2.964	,**	10
2.972	,*****	24
2.980	,**********	59
2.988	,**********	79
2.996	,**********	138
3.004	,**********	105
3.012	,**********	89
3.020	,**********	56
3.028	,*****	27
3.036	,**	10
3.044	,	0
3.052	,	2

```
          ,
          Šfff^fff^fff^fff^fff^fff^fff^
              20  40  60  80 100 120 140

                     Frequency
```

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média com ponderação em classes, por ENPS
Variável: Q4160

M416PCPS

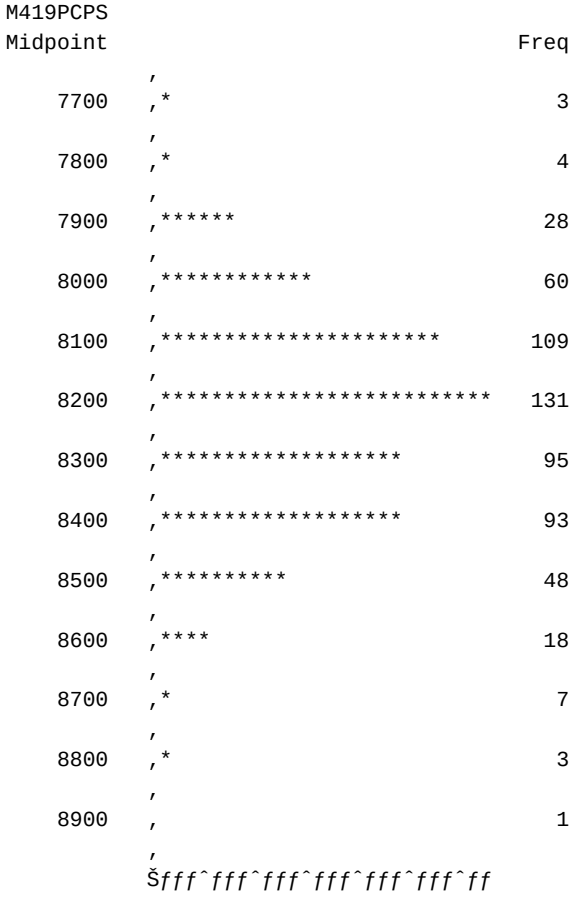
Midpoint		Freq
27800	,*	5
28200	,****	33
28600	,**********	93
29000	,******************	141
29400	,*****************	131
29800	,**************	102
30200	,********	57
30600	,***	22
31000	,*	11
31400	,	2
31800	,	2
32200	,	1

Šfffˆfffˆfffˆfffˆfff
30 60 90 120

Frequency

HISTOGRAMA DAS RÉPLICAS BOOTSTRAP
Estimador de Pós-estratificação da média com ponderação em classes, por ENPS
Variável: Q4190

M419PCPS



ANEXO 6 – Instrumentos de Notação do IEH