REVISTA DE ESTATÍSTICA
3º QUADRIMESTRE DE 2001

PÓS-ESTRATIFICAÇÃO E WITHOUT REPLACEMENT BOOTSTRAP: APLICAÇÕES AO INQUÉRITO ÀS EMPRESAS/HARMONIZADO

POST-STRATIFICATION AND WITHOUT REPLACEMENT BOOTSTRAP APPROACHES IN A BUSINESS SURVEY

Autora: Ana Cristina M. Costa
- Assistente no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa

***RESUMO*:**

- Neste artigo analisam-se alguns métodos de estimação que visam o tratamento dos problemas da base de sondagem e da 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. Na inferência clássica das sondagens, são de destacar os estimadores de pós-estratificação. As propriedades teóricas destes estimadores 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 algumas aplicações, dos métodos de pós-estratificação e do algoritmo *Without Replacement Bootstrap* (*BWO*), proposto por Sitter (1992), aos dados do Inquérito às Empresas/Harmonizado (IEH) de 1997, conduzido pelo Instituto Nacional de Estatística. Os estimadores considerados são discutidos em termos de precisão, sendo efectuadas recomendações quanto às suas aplicações no âmbito do IEH.

***PALAVRAS-CHAVE*:**

- *Pós-estratificação; problemas da base de sondagem; não resposta; reponderação; métodos de ajustamento; Bootstrap*.

ABSTRACT:

- This paper 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 estimators need further research. We then address the Bootstrap methodology for variance estimation. Some applications of the poststratification estimators and the *Without Replacement Bootstrap* (*BWO*) algorithm, proposed by Sitter (1992), are also presented, using data from the 1997 Annual Business Survey, conducted by Portugal's National Statistics Institute. The precision of the analysed estimators is discussed and some recommendations are made regarding their applications under this survey.

KEY-WORDS:

- *Poststratification; frame problems; nonresponse; reweighting; adjustment methods; Bootstrap*.

1. INTRODUÇÃO

O Inquérito às Empresas / Harmonizado (IEH), conduzido pelo Instituto Nacional de Estatística de Portugal (INE), é realizado anualmente e tem cobertura nacional. O desenho do IEH corresponde a um plano de amostragem aleatória estratificada sem reposição. As estimativas dos totais das diversas variáveis de interesse têm sido obtidas através do estimador de Horvitz-Thompson. 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 constava do FGUE, e que serviu de base à estratificação, se encontrar desactualizada ou incorrecta. Por outro lado, o IEH apresenta também não resposta total. Os problemas da base de sondagem e também as não respostas têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram.

O presente artigo surge na sequência da investigação efectuada no âmbito de um projecto que teve por principal objectivo identificar métodos de estimação que permitissem lidar com o problema das *mudanças de estrato*.

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. Estes estimadores inserem-se numa classe de métodos de tratamento de não respostas usualmente designados por métodos de recomposição ou métodos de ajustamento. Espera-se, portanto, que com estes métodos seja possível melhorar as estimativas das diversas variáveis de interesse.

Na inferência clássica das sondagens (*design-based*), as propriedades teóricas dos estimadores de pós-estratificação requerem ainda alguma investigação para planos de sondagem complexos; sendo de salientar a abordagem condicional efectuada por Rao (1985). É então abordada a metodologia *Bootstrap* para a estimação da variância dos estimadores.

Apresentam-se e discutem-se os resultados obtidos através da aplicação, aos dados do IEH de 1997, de diversos métodos de ajustamento e do algoritmo *Without Replacement Bootstrap* (*BWO*), proposto por Sitter (1992). Os estimadores considerados são comparados em termos de precisão, sendo efectuadas recomendações quanto às suas aplicações no âmbito do IEH.

2. ENQUADRAMENTO TEÓRICO

2.1. 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). O método *raking ratio* proposto por Deming e Stephan constitui um caso particular destes métodos.

Os métodos de pós-estratificação têm sido analisados sob diversos pontos de vista por vários autores, veja-se por exemplo: Williams (1962); Holt e Smith (1979); Rao (1985); Valliant (1993); Leonard *et al.* (1994) e Rao (1994).

Em seguida, apresenta-se de forma abreviada a abordagem considerada por Williams (1962) e Rao (1985), sobre a forma dos estimadores de pós-estratificação para um plano de sondagem genérico e para o plano de sondagem aleatória estratificada.

Suponhamos que, na fase de estimação, se dispõe de informação auxiliar que permita dividir a amostra em L pós-estratos. Sejam $n_1, \ldots, n_i, \ldots, n_L$ 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 $N_1, \ldots, N_i \ldots, N_L$ 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:

$$\hat{\tau}_{PS} = \sum_{i=1}^{L} \sum_{k \in s_i} \frac{N_i}{\hat{N}_i} w_k y_k \qquad (1)$$

Sendo $w_k = 1/\pi_k$ o peso de inclusão (ou coeficiente de extrapolação) do indivíduo *k*, subjacente ao desenho da amostra (*design-weight*) e $\hat{N}_i$ o usual estimador centrado de domínios da dimensão do *i-ésimo* pós-estrato:

$$\hat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k} \qquad (2)$$

Esta forma de apresentar o estimador de pós-estratificação permite-nos evidenciar o ajustamento pelo quociente dos pesos iniciais w_k.

Sendo *N* a dimensão da população, um **estimador de pós-estratificação da média da população**, μ, é dado, obviamente, por:

$$\hat{\mu}_{ps} = \frac{1}{N} \hat{\tau}_{PS} \qquad (3)$$

Em seguida, apresentam-se em mais detalhe os estimadores de pós-estratificação para o plano de sondagem aleatória estratificada.

2.1.1. 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 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 e, portanto, as dimensões amostrais resultantes da intersecção dos estratos iniciais com os pós-estratos são aleatórias.

Antes de apresentarmos o estimador de pós-estratificação genérico (1), para o plano de sondagem em análise, vamos considerar alguma notação adicional. Seja $N_{\bullet h}$ a dimensão do estrato inicial *h* na população (h=1, …, H); $N_{i\bullet}$ a dimensão do pós-estrato *i* na população (i=1, …, L); $n_{\bullet h}$ o número de elementos da amostra pertencentes ao estrato inicial *h* (h=1, …, H); s_{ih} o conjunto de elementos da amostra que pertencem simultaneamente ao estrato inicial *h* (h=1, …, H) e ao pós-estrato *i* (i=1, …, L); e n_{ih} a dimensão (aleatória) de s_{ih}.

Um **estimador de pós-estratificação de τ**, para o plano de sondagem aleatória estratificada, é dado por:

$$\hat{\tau}_{ps,str} = \sum_{i=1}^{L}\sum_{h=1}^{H}\sum_{k\in s_{ih}} \frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\frac{N_{\bullet h}}{n_{\bullet h}} y_k \qquad (4)$$

com $\hat{N}_{i\bullet}$ dado por

$$\hat{N}_{i\bullet} = \sum_{h=1}^{H}\frac{N_{\bullet h}}{n_{\bullet h}} n_{ih} \qquad (5)$$

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 (1) numa sondagem complexa. Mesmo para esta situação simples, Rao (1985) demonstra que o valor esperado do estimador (1), i.e. o valor esperado do estimador (4), 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 (1) que não revela boas propriedades na abordagem condicional, mesmo no caso em que o desenho da amostra corresponde a um plano de sondagem aleatória simples sem reposição, tal como demonstra Rao (1985). Este autor propõe um estimador alternativo que pode ser preferível tanto na abordagem condicional, como não condicional.

Denote-se por $\hat{V}(\hat{\tau}_{\pi}) = v(y_k)$ a função que define o estimador da variância do estimador usual de τ (*estimador de Horvitz-Thompson* ou *estimador-π*). Ou seja, no caso da sondagem aleatória estratificada sem reposição, tem-se

$$v(y_k) = \sum_{h=1}^{H} N_{\bullet h}^2\left(\frac{1}{n_{\bullet h}} - \frac{1}{N_{\bullet h}}\right)s_h^2 = \hat{V}(\hat{\tau}_{\pi_{str}}) \qquad (6)$$

onde,

$$s_h^2 = \frac{1}{n_{\bullet h} - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2 \qquad (7)$$

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:

$$\hat{V}_{rao}(\hat{\tau}_{ps}) = v(z_k) \qquad (8)$$

onde, $v(z_k)$ se obtém a partir de $\hat{V}(\hat{\tau}_\pi)$ substituindo-se y_k por:

$$z_k = \sum_i \frac{N_i}{\hat{N}_i} (y_{ik} - \frac{\hat{\tau}_i}{\hat{N}_i} \mathbb{I}_{k \in s_i}) = \sum_i \frac{N_i}{\hat{N}_i} (y_{ik} - \hat{\mu}_i \mathbb{I}_{k \in s_i}) \qquad (9)$$

sendo, $\mathbb{I}_{k \in s_i}$ a variável indicatriz que toma o valor 1 se o elemento *k* pertence ao pós-estrato *i* e toma o valor zero no caso contrário;

$$y_{ik} = y_k \mathbb{I}_{k \in s_i} = \begin{cases} y_k, & \text{se k pertence ao pós} - \text{estrato i} \\ 0, & \text{caso contrário} \end{cases} \qquad (10)$$

e $\hat{\mu}_i = \hat{\tau}_i / \hat{N}_i$, sendo $\hat{N}_i$ dado por (2) e

$$\hat{\tau}_i = \sum_{k \in s_i} \frac{y_k}{\pi_k} \qquad (11)$$

As propriedades deste estimador são, também, difíceis de investigar. No caso mais simples do plano de sondagem aleatória simples sem reposição o estimador (8) 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 (8) numa abordagem *model-assisted* adequada a planos de sondagem com uma etapa. Assim, é de esperar que para planos de sondagem complexos este estimador tenha também boas propriedades condicionais.

No caso da sondagem aleatória estratificada, para amostras grandes tais que os pós-estratos têm uma dimensão razoável em cada estrato inicial, espera-se que o estimador (8) seja um bom estimador de $V(\hat{\tau}_{ps,str})$, na abordagem condicional.

Quando 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.

Alternativamente ao estimador proposto por Rao (1985), podem-se considerar métodos de re-amostragem (*resampling*), como o Bootstrap, para se estimar a variância de $\hat{\tau}_{ps,str}$.

2.2. 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. 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. Uma discussão detalhada sobre os efeitos estatísticos da não resposta pode ser obtida em Lessler e Kalsbeek (1992).

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. 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. Referências bibliográficas relevantes sobre outros métodos 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.

2.2.1. INTRODUÇÃO AOS MÉTODOS DE AJUSTAMENTO DAS NÃO RESPOSTAS

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. 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

$$p_k = P(\mathbb{I}_{k\in U_1} = 1) \text{ , } k = 1, 2, \ldots, N \qquad (12)$$

onde, $\mathbb{I}_{k\in U_1}$ é a variável indicatriz que toma o valor 1 se o elemento *k* pertence à sub-população U_1 e toma o valor zero no caso contrário (i.e., se pertence a U_0).

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 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, 182]. Obtém-se, desta forma, o estimador centrado

$$\hat{\tau}^{*}_{HT} = \sum_{k=1}^{n_1} w^{*}_k y_k \qquad (13)$$

onde, $w^*_k = 1/(\pi_k p_k)$.

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.

2.2.2. 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) 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:

$$\hat{p}_{hk} = \frac{\sum_{k=1}^{n_{1h}} w_{hk}}{\sum_{k=1}^{n_h} w_{hk}} \quad , k \in s_h,\ h = 1, \ldots, H \tag{14}$$

O ponderador ajustado, a utilizar nos estimadores por ponderação em classes, é então

$$w_{hk}^{(pc)} = \frac{\sum_{k=1}^{n_h} w_{hk}}{\sum_{k=1}^{n_{1h}} w_{hk}} w_{hk} \,, \ k \in s_h,\ h = 1, \ldots, H \tag{15}$$

Este ponderador ajustado pode ser escrito como

$$w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} \,, \ k \in s_h,\ h = 1, \ldots, H \tag{16}$$

onde,

$$\hat{N}_h = \sum_{k=1}^{n_h} w_{hk} \tag{17}$$

$$\hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk} \tag{18}$$

Um **estimador do total da população por ponderação em classes** é

$$\hat{\tau}_{pc} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} w_{hk}^{(pc)} y_{hk} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} y_{hk} \qquad (19)$$

onde, $\hat{N}_h$ e $\hat{N}_{1h}$ são dados, respectivamente, por (17) e (18) e y_{hk} é o valor da variável de interesse para o elemento *k* da *h-ésima* célula de ajustamento.

2.2.3. 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 2.1 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. 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

$$w_{ik}^{(ps)} = \frac{N_i}{\hat{N}_i} w_{ik}^{(pc)} = \frac{N_i}{\hat{N}_i}\frac{\hat{N}_i}{\hat{N}_{1i}} w_{ik} = \frac{N_i}{\hat{N}_{1i}} w_{ik},\ k \in s_i,\ i = 1, \ldots, L \qquad (20)$$

com, $\hat{N}_{1i}$ dado por (18) ou seja,

$$\hat{N}_{1i} = \sum_{k=1}^{n_{1i}} w_{ik} \qquad (21)$$

Utilizando-se estes pesos ajustados, obtém-se um **estimador de pós-estratificação do total da população**, na presença de não respostas, dado por

$$\hat{\tau}_{ps} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} w_{ik}^{(ps)} y_{ik} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} \frac{N_i}{\hat{N}_{1i}} w_{ik} y_{ik} \qquad (22)$$

onde, $\hat{N}_{1i}$ é dado por (21) e y_{ik} é o valor da variável de interesse para o elemento *k* do *i-ésimo* pós-estrato (célula de ajustamento).

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 (15):

$$w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}, \; k \in s_h, \; h = 1, \ldots, H \qquad (23)$$

com, $\hat{N}_h$ e $\hat{N}_{1h}$ definidos, respectivamente, por (17) e (18); 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:

$$w_{ik}^{(pc,ps)} = \frac{N_i}{\hat{N}_{1i}^*} w_{ik}^{(pc)}, \; k \in s_i, \; i = 1, \ldots, L \qquad (24)$$

onde, $w_{ik}^{(pc)}$ é o peso ajustado por ponderação em classes, (23), do elemento *k* pertencente ao pós-estrato *i*; e $\hat{N}_{1i}^*$ é agora dado por

$$\hat{N}_{1i}^* = \sum_{k=1}^{n_{1i}} w_{ik}^{(pc)} \qquad (25)$$

Um **estimador de pós-estratificação do total da população**, com **ajustamento da não resposta por ponderação em classes**, é dado por

$$\hat{\tau}_{pc,ps} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} w_{ik}^{(pc,ps)} y_{ik} \qquad (26)$$

onde, $w_{ik}^{(pc,ps)}$ é o ponderador ajustado definido por (24).

Para planos de sondagem complexos, os estimadores dos métodos de ajustamento por ponderação em classes e por pós-estratificação são difíceis de analisar. O enviesamento e a variância dos estimadores da média da população, por ponderação em classes e por pós-estratificação, para um plano de sondagem aleatória simples sem reposição foram considerados por Thomsen (1973, 1978, citado por Little 1986), Kalton (1983, citado por Lessler e Kalsbeek 1992) e por Oh e Scheuren (1983, citados por Little 1986). É de salientar que estes autores verificaram que o estimador de pós-estratificação deverá ter erro quadrático médio inferior ao do estimador por ponderação em classes para esse plano de sondagem (mesmo considerando diferentes abordagens).

2.3. ESTIMAÇÃO DA VARIÂNCIA PELO MÉTODO *BOOTSTRAP WITHOUT REPLACEMENT* (*BWO*)

Os métodos de re-amostragem (*resampling*) 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 extensões destes métodos a dados de sondagens podem ser encontradas em Shao e Tu (1995).

A aplicação da metodologia Bootstrap a dados de sondagens requer algumas alterações ao algoritmo "puro" (também designado na literatura por bootstrap *naïf* ou *naive*) proposto por Efron (1979). Os primeiros trabalhos nesta área devem-se a Gross (1980) e a Chao e Lo (1985). O método proposto por estes autores para a sondagem aleatória simples sem reposição designa-se, geralmente, na literatura por *Bootstrap Without Replacement* ou *Without Replacement Bootstrap* (*BWO*). Sitter (1992) propôs uma extensão deste método à sondagem aleatória estratificada.

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

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

$$k_h = \frac{N_h}{n_h}\left(1-\frac{1-f_h}{n_h}\right);\ f_h = n_h/N_h,\ h = 1, \ldots, H \qquad (27)$$

2º – Seleccionar n'_h unidades de cada estrato *h* através de tiragens aleatórias sem reposição, sendo $n'_h = n_h - (1 - f_h)$, h = 1, …, 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 $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*b}, \ldots, \hat{\theta}^{*B}$ réplicas bootstrap do estimador.

4º – Estimar $V(\hat{\theta})$ por

$$\hat{V}^{*}_{BWO} = E^{*}[\hat{\theta}^{*} - E^{*}(\hat{\theta}^{*})]^2 \qquad (28)$$

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

$$\hat{V}^{*}_{BWO} \approx \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^{*}(\cdot)\right)^2 \qquad (29)$$

onde,

$$\hat{\theta}^{*}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b} \qquad (30)$$

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 (1992) 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 (1992) através de estudos por simulação.

3. APLICAÇÕES AO INQUÉRITO ÀS EMPRESAS / HARMONIZADO DE 1997

Foram realizadas aplicações práticas[5], com os dados do IEH de 1997 (IEH97), dos métodos de ajustamento das não respostas (secção 2.2) e do método de Bootstrap BWO proposto por Sitter (1992), apresentado na secção 2.3.

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.

Para efeitos de apuramento, o universo é estratificado pelos escalões definidos pelas variáveis: *Escalões de NUTS II* – Nomenclatura das Unidades Territoriais para Fins Estatísticos; *Escalões de Classificação Portuguesa das Actividades Económicas*

[5] Resultados detalhados deste estudo e aplicações aos dados de 1996 podem ser encontrados em Machado e Costa (1998, 2001).

(CAE - REV. 2); *Escalões de número de pessoas ao serviço* e *Escalões de forma jurídica*.

O inquérito é realizado por amostragem para as unidades estatísticas (empresas) com menos de 100 pessoas ao serviço e de forma exaustiva para as unidades estatísticas com 100 e mais pessoas ao serviço[6].

Designem-se por *Pequenas e médias empresas* as empresas consideradas para inquirição com recurso à teoria de amostragem e por *Grandes empresas* as consideradas para inquirição exaustiva. No presente estudo consideram-se apenas as empresas do Continente inquiridas por amostragem, ou seja, as *Pequenas e médias empresas* do Continente.

A título ilustrativo, foram seleccionadas três variáveis: *Número médio de pessoas ao serviço* (total – remunerado e não remunerado), *Vendas* e *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).

3.1. ASPECTOS METODOLÓGICOS

Os estimadores considerados foram: o estimador de ponderação em classes, o estimador de pós-estratificação e o estimador de pós-estratificação com ajustamento da não resposta por ponderação em classes. Obtiveram-se também estimativas através do estimador de Horvitz-Thompson que, apesar de neste caso ser enviesado, não deixa de ser uma referência.

No método de ajustamento por ponderação em classes consideram-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.

Nas aplicações do método de ajustamento por pós-estratificação, obtiveram-se as estimativas correspondentes à pós-estratificação da amostra através de três técnicas: segundo a variável *Escalões de número de pessoas ao serviço* (Quadro 1); segundo a variável *Escalões de volume de vendas* (Quadro 2); segundo o cruzamento dos escalões das variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas* (Quadro 3). Em qualquer dos casos, pressupõe-se que a(s) variável(s) de pós-estratificação está estreitamente relacionada com as variáveis de interesse. Neste método considera-se que os pós-estratos correspondem às células de ajustamento da não resposta e, portanto, assume-se que as probabilidades de resposta são iguais, em cada pós-estrato.

No método de pós-estratificação com ajustamento da não resposta por ponderação em classes consideram-se os estratos iniciais como sendo as classes de ajustamento da não resposta. Supõe-se, portanto, que os elementos têm valores

[6] Uma descrição mais detalhada da metodologia desta sondagem pode ser obtida em Instituto Nacional de Estatística (1997).

semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais, em cada estrato inicial.

Neste caso, a amostra foi pós-estratificada através de duas técnicas: segundo a variável *Escalões de número de pessoas ao serviço* (Quadro 1) e segundo o cruzamento dos escalões das variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas* (Quadro 3).

Quadro 1. Dimensões dos pós-estratos na população segundo a variável Escalões de número de pessoas ao serviço

Escalão	*Pessoas ao serviço*	Dimensão do pós-estrato na população
0	0	116581
1	1 a 9	610786
2	10 a 19	17965
3	20 a 49	9846
4	50 a 99	2141

Quadro 2. Dimensões dos pós-estratos na população segundo a variável *Escalões de volume de vendas*

Escalão	*Volume de vendas* (milhares de escudos)	Dimensão do pós-estrato na população
1	≤ 30 000	16678
2	> 30 000	740641

Quadro 3. Dimensões dos pós-estratos na população segundo as variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas*

Pós-estrato	*Escalões de número de pessoas ao serviço*	*Escalões de volume de vendas*	Dimensão do pós-estrato na população
1	0	1	3455
2	0	2	113126
3	1	1	13213
4	1	2	597573
5	2	1	10
6	2	2	17955
7	3	1	-
8	3	2	9846
9	4	1	-
10	4	2	2141

Para estimar a variância do estimador de pós-estratificação utilizou-se o estimador proposto por Rao (1985), dado por (8) para o caso da ausência de não respostas num plano de sondagem aleatória estratificada. 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 (1992).

As estimativas da variância foram obtidas através das aproximações de Monte Carlo. Para tal, foram retiradas 1000 amostras bootstrap, da população bootstrap construída segundo o referido algoritmo (*c.f.* secção 2.3).

3.2. RESULTADOS E CONCLUSÕES

As estimativas da média obtidas através do estimador de pós-estratificação por *Escalões de volume de vendas* (*PS por EVVN*) parecem indicar que este é extremamente enviesado. Este resultado não é de estranhar uma vez que, neste caso, se consideram apenas dois pós-estratos e, portanto, não se deve verificar o pressuposto de que os pós-estratos são homogéneos. Por esta razão, optou-se por não se aplicar o método Bootstrap BWO a este estimador.

As estimativas bootstrap da média obtidas para os estimadores considerados parecem indicar que o estimador de ponderação em classes é também bastante enviesado. Assim, as estimativas dos desvio-padrão deste estimador subestimam o verdadeiro valor do erro, uma vez que não contêm a contribuição do enviesamento. Aliás, como foi referido na secção 2.2, há evidências teóricas de que o estimador de pós-estratificação tem um erro quadrático médio inferior ao desse estimador e portanto deverá ser preferível.

Quanto aos restantes estimadores por pós-estratificação, as diferenças observadas parecem prender-se com o tipo de variáveis utilizadas para pós-estratificar a amostra. Dado que não é possível estimar o respectivo enviesamento, e consequentemente o seu erro quadrático médio, é mais difícil precisar qual das formas de pós-estratificação é mais adequada (por *Escalões de número de pessoas ao serviço* ou pelo cruzamento destes com os *Escalões de volume de vendas*). No entanto, pela análise dos desvios padrão e dos coeficientes de variação estimados por bootstrap verifica-se que, para as variáveis em estudo, há evidências de que o estimador de pós-estratificação por *Escalões de número de pessoas ao serviço* deverá ser mais adequado.

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. Este era o resultado esperado dado que, no caso do IEH97, a dimensão dos pós-estratos em cada estrato inicial é bastante razoável. 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.

Como foi referido anteriormente, os pós-estratos devem ser o mais homogéneos possível e, portanto, a variável que define os pós-estratos deve estar fortemente correlacionada com as variáveis de interesse. Sob este pressuposto e caso as dimensões dos pós-estratos sejam conhecidas, poder-se-ão considerar outras variáveis de pós-estratificação. Neste caso, seria também interessante analisar a utilização do estimador de pós-estratificação com ajustamento da não resposta por ponderação em classes.

A utilização dos estimadores de pós-estratificação tem por principal objectivo lidar com os problemas na base de sondagem. Dado que a taxa de não resposta total no IEH é muito elevada (cerca de 27%), fica também como sugestão para futuras investigações a utilização destas técnicas em simultâneo com outros métodos de tratamento das não respostas, nomeadamente os métodos de imputação.

BIBLIOGRAFIA

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.

CHAO, M. T. e LO, S. H. (1985). A Bootstrap method for finite populations. *Sankhyä* A 47, 399-405.

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. 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.

EFRON, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Mathematical Statistics* 7, 1-26.

GROSS, S. (1980). "Median estimation in sample surveys." Proceedings of the Section *on* Survey Research Methods, American Statistical Association, 181-184.

HOLT, D. e SMITH, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society* A 142, 33-46.

INSTITUTO NACIONAL DE ESTATÍSTICA (1997). *Inquérito ás Empresas / Harmonizado – Dossier Global do Projecto*. Departamento de Estatísticas das Empresas, INE/DEE, Junho 1997.

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.

MACHADO, J. F. e COSTA, A. C. (1998). *Mudanças de Estrato nos Inquéritos às Empresas / Harmonizados - Relatório Intercalar*. Contrato Programa INE/ISEGI, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, Outubro de 1998.

MACHADO, J. F. e COSTA, A. C. (2001). *Mudanças de Estrato nos Inquéritos às Empresas / Harmonizados - Relatório Final*. Contrato Programa INE/ISEGI, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, Junho de 2001.

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.

SHAO, J. e TU, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SITTER, R. R. (1992). Comparing three Bootstrap methods for survey data. *The Canadian Journal of Statistics* 20, No. 2, 135-154.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* 88, nº 421, 89-96.

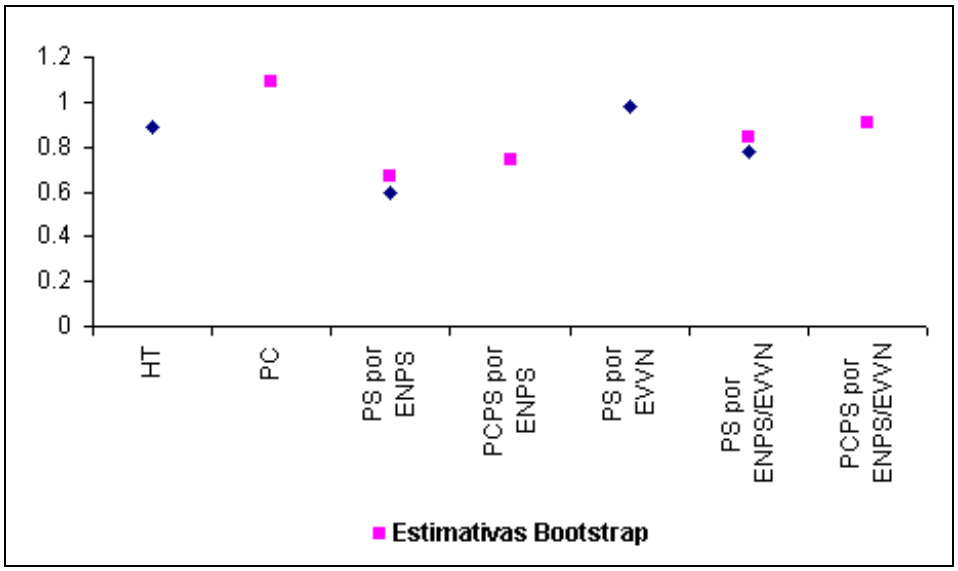
WILLIAMS, W. H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association* 57, 622-627.

ANEXO

Quadro A 1. Estimativas obtidas para a variável *Nº médio de pessoas ao serviço*

Estimador	Desvio padrão do estimador da média	Coeficiente de variação da média (%)	*Estimativas bootstrap* Desvio padrão do estimador da média	*Estimativas bootstrap* Coeficiente de variação da média (%)
Horvitz-Thompson (**HT**)	0.0173	0.89	-	-
Ponderação em classes (**PC**)	-	-	0.0173	1.09
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	0.0173	0.60	0.0200	0.67
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	0.0224	0.74
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	0.0872	0.98	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	0.0346	0.78	0.0400	0.84
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	0.0436	0.91

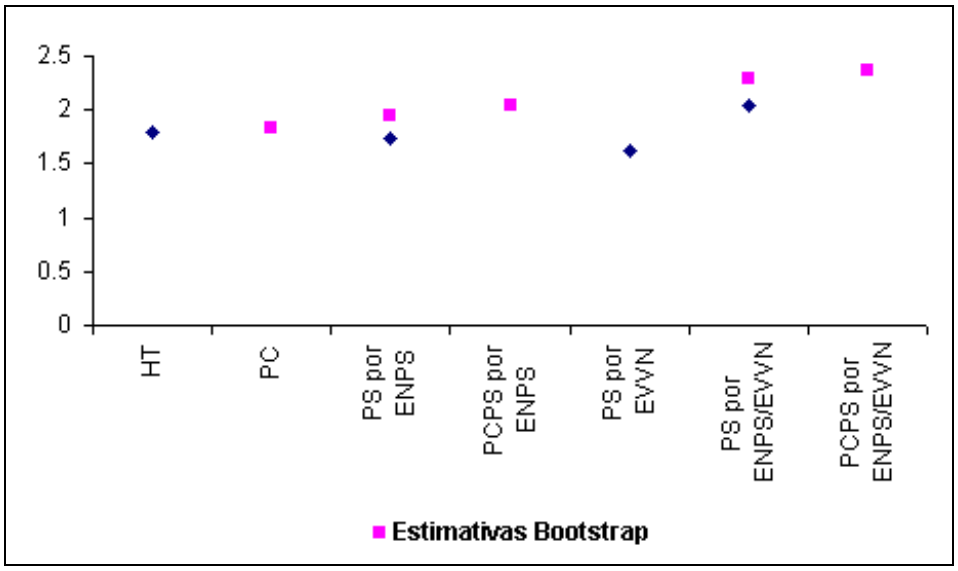
Gráfico A 1: Coeficientes de variação estimados do estimador da média (%), para a variável *Nº médio de pessoas ao serviço*



Quadro A 2. Estimativas obtidas para a variável *Vendas*

Estimador			*Estimativas bootstrap*	
	Desvio padrão do estimador da média	**Coeficiente de variação da média (%)**	**Desvio padrão do estimador da média**	**Coeficiente de variação da média (%)**
Horvitz-Thompson (**HT**)	377.65	1.79	-	-
Ponderação em classes (**PC**)	-	-	283.65	1.84
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	528.50	1.74	626.12	1.94
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	680.90	2.04
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	2018.08	1.62	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	1630.53	2.04	1928.42	2.29
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	1918.28	2.36

Gráfico A 2. Coeficientes de variação estimados do estimador da média (%), para a variável *Vendas*



Quadro A 3. Estimativas obtidas para a variável *Prestações de serviços*

Estimador			*Estimativas bootstrap*	
	Desvio padrão do estimador da média	**Coeficiente de variação da média (%)**	**Desvio padrão do estimador da média**	**Coeficiente de variação da média (%)**
Horvitz-Thompson (**HT**)	165.17	2.88	-	-
Ponderação em classes (**PC**)	-	-	104.10	2.56
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	238.51	2.97	243.30	2.87
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	221.17	2.59
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	906.20	3.00	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	828.23	4.94	850.26	4.88
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	661.02	4.09

Gráfico A 3. Coeficientes de variação estimados do estimador da média (%), para a variável *Prestações de serviços*