Aplicação de **Sondagens Indirectas no Turismo**

LUÍS NOBRE PEREIRA * [lmper@ualg.pt]
PEDRO SIMÕES COELHO ** [psc@isegi.unl.pt]

Resumo | Na maioria das sondagens realizadas na área do turismo não existe uma base de amostragem da população alvo. Em seu lugar, pode existir uma base de amostragem formada por unidades de amostragem de outra população, as quais estão relacionadas, de alguma forma, com as unidades estatísticas de análise que se pretendem inquirir. Esta situação provoca um problema no que concerne à dificuldade de associar uma probabilidade de inclusão a cada unidade amostral da população alvo. Uma solução para esse problema consiste em aplicar o método de sondagem indirecta e utilizar o Método de Partilha de Pesos Generalizado (MPPG) desenvolvido por Lavallée (1995, 2002), na atribuição dos pesos de estimação a cada unidade amostral da população alvo. Este artigo é dedicado à apresentação do método de sondagem indirecta e da forma de determinação dos pesos da estimação, através do MPPG, e a sua adaptação à resolução de problemas de estimação na área do turismo em ambiente aberto.

Palavras-chave | Método de Sondagem Indirecta, Método de Partilha de Pesos Generalizado, Sondagem Multi-etapas, Estimação, Sondagens no Turismo.

Abstract | In most of the surveys accomplished in the tourism area there is not a sampling frame of the target population. Instead, it may exist a sampling frame formed by elements of another population, which are in some way related with the sample elements. The problem comes from the difficulty to associate a selection probability to the surveyed elements of the target population. In order to solve this type of estimation problem, one can use the Generalized Weight Share Method (GWSM) developed by Lavallée (1995, 2002). The GWSM provides an estimation weight for every surveyed element of the target population. This paper aims at describing the indirect sampling method and how to obtain the estimation weights from the GWSM. It also intends to propose an adaptation of this methodology to the solution of open environment touristic surveys.

Keywords | Indirect Sampling Method, Generalized Weight Share Method, Multi-stage Sampling, Estimation, Tourism surveys.

* **Doutorado em Métodos Quantitativos Aplicados à Economia e à Gestão** pela Universidade do Algarve e **Professor Adjunto** na Escola Superior de Gestão, Hotelaria e Turismo da Universidade do Algarve.

** **Doutorado em Estatística** pela Universidade Nova de Lisboa e **Professor Associado com Agregação** do Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa.

1. Introdução

As necessidades de conhecimento de uma população, relativamente a uma ou mais características, são normalmente satisfeitas através de sondagens. Uma sondagem é um processo de recolha e análise de informação, que permite estudar características de uma população a partir da observação de um subconjunto dos seus elementos, denominado por amostra, e da extrapolação dos resultados amostrais para a população. A selecção de uma amostra aleatória no âmbito de uma sondagem exige a utilização de uma base de amostragem, ou seja, de uma listagem com a identificação dos elementos da população em estudo, bem como a possibilidade de contacto com esses elementos.

Infelizmente, na maioria das sondagens realizadas na área do turismo não existe uma base de amostragem dessa população, o que dificulta o estabelecimento de um contacto com esses elementos, a partir daqui denominados por turistas, com o intuito de recolha de informação. Para além disso, a inexistência de uma base de amostragem da população em estudo inviabiliza a utilização de um método de amostragem probabilístico, com todas as desvantagens daí resultantes. A utilização de um método de amostragem não probabilístico na selecção da amostra, como é frequente nos estudos da área do turismo, impossibilita, por exemplo, a avaliação objectiva da qualidade dos resultados, ou seja, a determinação de precisão associada aos resultados obtidos na estimação realizada a partir de uma amostra de determinada dimensão.

O problema da inexistência de uma base de amostragem de turistas pode ser resolvido através da amostragem aleatória de serviços utilizados pelos turistas em várias localizações, nos quais são aplicados os instrumentos de recolha de dados. Obviamente que um determinado turista pode usar um ou mais serviços da amostra, pelo menos uma vez durante o período da recolha de dados. Para ser possível estimar os parâmetros de interesse relativos à população de turistas, tem que ser possível seleccionar uma amostra aleatória desses serviços, a partir da qual se estabelece uma ligação entre as probabilidades de inclusão dos serviços pertencentes à amostra e as probabilidades de inclusão dos turistas que utilizaram esses serviços. O Método de Partilha de Pesos Generalizado (MPPG), desenvolvido por Lavallée (1995, 2002), permite efectuar essa ligação.

O objectivo deste artigo consiste em apresentar o método de sondagem indirecto, o qual constitui os fundamentos do MPPG, e uma aplicação destes métodos na resolução de problemas na área do turismo. O MPPG é um método utilizado para a obtenção dos pesos da estimação. Na secção 2 efectua-se uma breve descrição teórica do método de sondagem indirecto. Nesta secção expõe-se também o MPPG e as suas propriedades gerais. Na secção 3 apresenta-se uma aplicação do método de sondagem indirecto ao turismo, na qual se propõe uma medida de qualidade do parâmetro de interesse utilizando uma sondagem complexa. Um exemplo numérico baseado nessa aplicação é também apresentado na secção 3. Por último, na secção 4 apresentam-se as principais conclusões deste trabalho.

2. Método de sondagem indirecto

Tal como foi referido na introdução, na maioria das sondagens realizadas na área do turismo não existe uma base de amostragem da população alvo. Em seu lugar, pode existir uma base de amostragem formada por unidades de amostragem de outra população, as quais estão relacionadas, de alguma forma, com as unidades estatísticas de análise que se pretende inquirir. Nesta situação, está-se perante duas populações relacionadas, denominadas por U^A e U^B, e pretende-se produzir estimativas para um parâmetro de interesse da população em estudo, U^B, como por exemplo uma média, uma proporção ou um total. Contudo, só existe uma base de amostragem

de outra população, U^A. Admita-se que U^A e U^B são duas populações finitas constituídas por N^A e N^B unidades, indexadas por j e i, respectivamente. A correspondência unívoca entre essas duas populações pode ser representada por uma matriz de ligação, $\Theta_{AB}=\left[\theta_{ji}^{AB}\right]$, de dimensão N^A x N^B com elementos de ligação não negativos. Melhor explicitando, admite-se que se a j-ésima unidade de U^A estiver relacionada com a i-ésima unidade de U^B, então $\theta_{ji}^{AB}>0$; e se essas duas unidades não estão relacionadas, então $\theta_{ji}^{AB}=0$. Desta forma, θ_{ji}^{AB} pode representar a intensidade da associação entre as unidades j e i (e.g. frequência de visitas a um determinado local por parte de um turista), ou mais simplesmente a presença ou ausência de associação entre essas unidades.

O método de sondagem indirecto apresenta-se como uma solução para o problema da inexistência de uma base de amostragem da população em estudo. Este método consiste em seleccionar, de acordo com um determinado método de amostragem, uma amostra aleatória s^A de dimensão n^A da base de amostragem disponível de U^A, com o objectivo de produzir estimativas para os parâmetros de interesse de U^B, utilizando as relações existentes entre as duas populações. Seja π_j^A a probabilidade de inclusão de ordem 1 da j-ésima unidade, ou seja, a probabilidade de selecção dessa unidade para a amostra s^A. Assume-se que $\pi_j^A>0$ para todas as unidades de U^A. Veja-se, por exemplo, em Särndal, *et al*. (1992), os vários métodos de amostragem que existem, assim como as respectivas probabilidades de inclusão. Para cada unidade j seleccionada para a amostra s^A, identificam-se as i-ésimas unidades de U^B que apresentam uma correspondência não nula, isto é, com $\theta_{ji}^{AB}>0$. Desta forma constitui-se um conjunto (ou amostra) s^B com todas as n^B unidades de U^B identificadas pelas unidades da amostra seleccionada. Na prática, apesar da dimensão da amostra n^A poder ser calculada *a priori*, a dimensão da amostra n^B é desconhecida porque depende da composição da amostra s^A e das ligações existentes entre os elementos das duas populações.

Para cada unidade i do conjunto s^B, é observada uma característica de interesse, y_i, na população em estudo U^B. Seja $\mathbf{Y}=(y_1,...,y_{N^B})'$ o vector coluna da variável de interesse. Por exemplo, se um dos objectivos de uma determinada investigação consistir na estimação do total dessa variável de interesse na população em estudo U^B, $\tau_Y^B=\sum_{i=1}^{N^B}y_i$, então deve utilizar-se o seguinte estimador:

$$\hat{\tau}_Y^B=\sum_{i=1}^{N^B}w_i y_i\,, \qquad (1)$$

onde w_i é o peso atribuído à i-ésima unidade da amostra s^B, com $w_i=0$ para $i\notin s^B$. Como é muito difícil, ou mesmo impossível, obter-se a probabilidade de inclusão de cada uma das unidades da amostra s^B, de acordo com Lavallée (1995, 2002), então não se pode utilizar directamente o estimador de Horvitz-Thompson (Horvitz e Thompson, 1952). No sentido de resolver este problema, Lavallée (1995, 2002) propôs a utilização de um novo estimador que usa o método MPPG, no âmbito do qual o peso atribuído a cada unidade da amostra s^B é dado por:

$$w_i=\sum_{j=1}^{N^A}\frac{t_j^A\tilde{\theta}_{ji}^{AB}}{\pi_j^A}\,, \qquad (2)$$

onde t_j^A é uma variável indicatriz com $t_j^A=1$ se $j\in s^A$ e $t_j^A=0$ em caso contrário, e $\tilde{\theta}_{ji}^{AB}=\theta_{ji}^{AB}\Big/\sum_{j=1}^{N^A}\theta_{ji}^{AB}$. O estimador do total τ_Y^B tem então a forma:

$$\hat{\tau}_Y^B=\sum_{i=1}^{N^B}\sum_{j=1}^{N^A}\frac{t_j^A\tilde{\theta}_{ji}^{AB}}{\pi_j^A}y_i=\sum_{i=1}^{N^B}\sum_{j\in s^A}\frac{\tilde{\theta}_{ji}^{AB}y_i}{\pi_j^A}\,. \qquad (3)$$

Note-se, que ao contrário do estimador de Horvitz-Thompson, onde $w_i=0$ para $i\notin s^B$, todas as unidades populacionais em U^B terão agora um ponderador diferente de zero, desde que associadas a pelo menos uma unidade da população U^A.

A compreensão do estimador (3) é facilitada notando que o total populacional da variável de interesse pode ser escrito como $\tau_Y^B=\sum_{j=1}^{N^A}\sum_{i=1}^{N^B}\tilde{\theta}_{ji}^{AB}y_i$. Utilizando um peso do tipo (2), Lavallée (1995) mostrou que o estimador (3) é não enviesado e que a sua variância é dada por:

$$V\left(\hat{\tau}_Y^B\right)=\sum_{j=1}^{N^A}\sum_{j'=1}^{N^A}\frac{\left(\pi_{jj'}^A-\pi_j^A\pi_{j'}^A\right)}{\pi_j^A\pi_{j'}^A}z_j z_{j'}, \qquad (4)$$

onde $\pi_{jj'}^A$ é a probabilidade das unidades j e j' serem seleccionadas para a amostra s^A e $z_j=\sum_{i=1}^{N^B}\tilde{\theta}_{ji}^{AB}y_i$. Veja-se mais uma vez, por exemplo, em Särndal, *et al*. (1992) a forma de cálculo das probabilidades de inclusão subjacentes aos vários métodos de amostragem.

O conhecimento da matriz de ligação, Θ_{AB}, é, portanto, fundamental para a aplicação de uma sondagem indirecta. A aplicação do método parte de dois pressupostos: (*a*) que é possível identificar todas as unidades $i\in U^B$ com correspondência unívoca não nula com cada unidade $j\in s^A$; (*b*) que é possível identificar todas as unidades $j\in U^A$ com correspondência não nula com cada unidade $i\in s^B$. Esta informação pode, em muitos casos, ser obtida através de entrevista directa ou de fonte administrativa. Contudo, segundo Lavallée (2002), é dispensável o conhecimento dos valores de ligação entre todos os elementos das duas populações.

Este problema torna-se ainda mais interessante pelo facto da escolha de $\theta_{ji}^{AB}>0$ poder afectar a precisão das estimativas do parâmetro de interesse. Kalton e Brick (1995), Lavallée e Caron (2001), Lavallée (2002) e Deville e Lavallée (2006), estudaram o problema da escolha óptima dos valores que expressam as ligações entre as unidades duas populações, θ_{ji}^{AB}, de forma a minimizar a variância das estimativas dos parâmetros de interesse. Utilizando diferentes metodologias, Kalton e Brick (1995), Lavallée e Caron (2001) e Lavallée (2002) sugeriram que $\theta_{ji}^{AB,opt}=1$ quando $\theta_{ji}^{AB}>0$ e $\theta_{ji}^{AB,opt}=0$ quando $\theta_{ji}^{AB}=0$. Posteriormente, Deville e Lavallée (2006) apresentaram a mesma sugestão para os casos particulares de uma sondagem aleatória simples sem reposição e de uma sondagem de Poisson. No trabalho de Deville e Lavallée (2006) podem também ser encontrados resultados gerais para a escolha óptima de θ_{ji}^{AB} no contexto do MPPG.

3. Aplicações do método de sondagem indirecto no turismo em ambiente aberto

3.1. Introdução

Um dos indicadores que tem grande interesse medir quando se analisa o fenómeno turístico é o montante total dos gastos locais efectuados pelos turistas de sol e praia. Os gastos locais dos turistas definem-se como todos os gastos realizados pelos turistas, excluindo os gastos com o alojamento e a deslocação de e para o país de origem ou local de residência. A impossibilidade de obtenção de uma base de amostragem da população de turistas de sol e praia numa dada região, dificulta não só a estimação do montante total dos gastos locais, mas principalmente a produção de medidas de precisão associadas ao estimador desse parâmetro de interesse (total dos gastos locais). A produção de medidas de precisão é, como é do conhecimento geral, fundamental para avaliar a qualidade dos resultados obtidos em qualquer sondagem. O método de sondagem indirecto apresentado na secção 2 permite resolver esse problema.

A estimação do número total de turistas que visitam uma determinada área, o número total de turistas a passar férias numa região segundo a sua origem geográfica, o número total de dias de férias dos turistas numa determinada região, a avaliação da satisfação e necessidades dos turistas, são outros problemas que podem ser resolvidos com o apoio do método de sondagem indirecto.

3.2. População alvo

Defina-se a região de interesse por *G* (e.g. Algarve) e o período de referência por *P* (e.g. durante o mês de Junho de um determinado ano). Define-se a *i*-ésima unidade estatística de análise como sendo o grupo de turistas que passa férias de sol e praia na região *G* durante o período *P*. Um grupo de turistas é um grupo de u_i pessoas (formado por familiares,

amigos ou outros) que, em conjunto, está a fazer uma viagem, num período compreendido em *P*, e que apresenta um comportamento semelhante nas principais variáveis de interesse (gastos locais, por exemplo). A população alvo U^B é, portanto, formada por um agregado de grupos de turistas que passa férias de sol e praia em *G* durante *P*. Define-se ainda que o entrevistado é uma pessoa que conhece, representa ou gere os gastos locais diários do grupo que está a fazer a viagem em conjunto.

3.3. Desenho da amostra

A produção de estimativas do montante total dos gastos locais é efectuada através do método de sondagem indirecto, o qual se apoia numa população, U^A, formada por um vasto agregado de serviços utilizados pelos turistas na região de interesse. Alguns exemplos de serviços utilizados pelos turistas na região de interesse são os seguintes: visitas às praias existentes nessa região; visitas aos pontos de interesse (a nível cultural, histórico, arquitectónico, gastronómico, entre outros), populares e bem conhecidos, existentes nessa região; compras efectuadas nas lojas e refeições adquiridas nos restaurantes dessa região.

Por facilidade de exposição, admita-se que a base de amostragem é formada apenas pelos seguintes dois estratos: praias e pontos de interesse. De forma lata, considere-se que as praias e os pontos de interesse são estabelecimentos. Admita-se também que no primeiro estrato se utiliza uma amostra aleatória em três etapas, na qual se extrai:

- Uma amostra de praias;
- Uma amostra de dias de recolha de dados pertencente ao período de referência;
- Uma amostra de turistas frequentadores de uma determinada praia num determinado dia.

Admita-se ainda que se utiliza igualmente uma amostra aleatória em três etapas no segundo estrato, na qual se extrai:

- Uma amostra de pontos de interesse;
- Uma amostra de dias de recolha de dados pertencente ao período de referência;
- Uma amostra de turistas frequentadores de um determinado ponto de interesse num determinado dia.

Desta forma, a população em estudo é formada por todos os turistas que "consomem" pelo menos um desses serviços durante o período de referência. Admita-se que o número de turistas que não "consome" pelos menos um desses serviços é negligenciável. No que se refere ao desenho da sondagem em cada estrato (praias, pontos de interesse), considere-se que:

- As praias são seleccionadas utilizando uma amostragem aleatória simples sem reposição com probabilidades iguais;
- Os pontos de interesse são seleccionados utilizando uma amostragem aleatória simples sem reposição com probabilidades iguais ou uma amostragem aleatória estratificada proporcional. Neste último caso, os estratos podem ser definidos por nível de interesse: cultural, histórico, arquitectónico, gastronómico, outros;
- Em cada praia e em cada ponto de interesse seleccionado para a amostra de estabelecimentos, os locais de recolha de dados não são seleccionados de forma aleatória, mas sim definidos de acordo com a notoriedade e o potencial contacto com os turistas (e.g. entradas/saídas principais). Assume-se assim que são seleccionados os locais de recolha de dados que permitem uma observação exaustiva dos estabelecimentos;
- Os dias de recolha de dados são seleccionados utilizando uma amostragem aleatória simples sem reposição com probabilidades iguais;
- Em cada dia seleccionado para a amostra de cada estabelecimento, os horários de recolha de dados são definidos de acordo com o potencial contacto com os turistas em função do respectivo

estabelecimento. Note-se mais uma vez, que se assume que são seleccionados todos os horários relevantes para a obtenção de uma observação exaustiva das visitas do dia e estabelecimento seleccionados;

– Em cada subpopulação "estabelecimento × dia", os turistas são seleccionados para a amostra utilizando técnicas de selecção aleatória em chegadas. A amostragem pseudo-sistemática é uma dessas técnicas. A amostragem pseudo--sistemática pode ser utilizada em ambientes fechados (com ou sem controlo de entradas) e em ambientes abertos, mas não permite estimar directamente o número total de visitantes. Assim, assume-se que a entrada de um grupo de turistas num certo "estabelecimento × dia" corresponde formalmente à prestação de um serviço num certo dia e estabelecimento.

Adicionalmente, pode considerar-se que o desenho de sondagem descrito anteriormente é utilizado num contexto de amostragem a partir de múltiplas bases de amostragem. Quando aplicado a este problema, o MPPG considera cada base de amostragem como um estrato, desde que seja possível identificar para cada unidade amostral todas as bases de amostragem das quais faz parte.

3.4. Parâmetro de interesse

Defina-se a aplicação F, a qual estabelece a ligação entre qualquer serviço, j, da região de interesse G durante o período P e o grupo de turistas, i, que utiliza esse serviço:

$$\begin{array}{ccc} F: \text{Serviços} & \rightarrow & \text{Grupo de turistas} \\ j & \rightarrow & i \end{array}$$

A população alvo, U^B, é a imagem dada por $F(j)$ do agregado de serviços durante P. A população auxiliar U^A é a imagem dada por $F^{-1}(i)$ do agregado de grupos de turistas durante P. Defina-se então o número de serviços j utilizados por um determinado grupo de turistas i, durante o período de referência, como $R_i(B) = card\,[F^{-1}(i)]$, $\forall\, i \in U^B$. Note-se que em alguns casos práticos pode ser conveniente trabalhar exclusivamente com grupos unitários. Estabelecendo um paralelismo com o MPPG apresentado na Secção 2, $R_i(B)=\sum_{j=1}^{N^A}\theta_{ji}^{AB}$ onde $\theta_{ji}^{AB}=1\ \forall\theta_{ji}>0$ e onde N^A representa o número total de serviços disponíveis para observação no período e região de referência.

Admitindo que se está interessado em estimar um total, τ, relativo ao montante de gastos locais efectuados pelos turistas de sol e praia (variável de interesse y) definido na população U^B, então o parâmetro de interesse pode ser apresentado da seguinte forma (Deville e Maumy-Bertrand, 2006):

$$\tau_Y^B=\sum_{i\in U^B} y_i=\sum_{l=1}^{2}\sum_{a_l\in A_l}\sum_{d_l\in D_l}\sum_{j\in C_{d_l}} z_j\,, \qquad (5)$$

onde $z_j = y_i / R_i(B)$, para $j \in F^{-1}(i)$, i.e. y_i representa o valor da variável de interesse para um grupo de turistas i que consumiu o serviço j; A_1 é o conjunto de praias da população U^{A_1} identificadas pelo índice a_1; A_2 é o conjunto de pontos de interesse da população U^{A_2} identificados pelo índice a_2; D_l é o conjunto de dias da população, identificados pelo índice d_l num estabelecimento a_l do conjunto A_l, l=1,2; C_{d_l} é o conjunto de serviços de um estabelecimento a_l no dia d_l, identificado pelo índice j. Por simplicidade de notação, as respectivas amostras são representadas por s_{A_l}, s_{D_l} e s_{Cd_l}, l=1, 2. Naturalmente que $U^A=U^{A_1}\cup U^{A_2}$ e $s^A=s_{A_1}\cup s_{A_2}$.

3.5. Estimador do parâmetro de interesse

Admita-se a existência de uma amostra aleatória de "estabelecimentos × dias × serviços", s^A, para a qual são conhecidas as probabilidade de inclusão, $\pi_j^{A_l}$ (ver Särndal *et al.*, 1992 no que se refere à determinação das probabilidades de inclusão em amostragens multi-etapas). Note-se que habitualmente o serviço j corresponderá a uma entrada (utilização) na unidade ad por parte de um elemento de um grupo de turistas. Assume-se igualmente

que é conhecido o número de grupos de turistas que utilizam cada unidade *ad* (estabelecimento × dia), bem como o número de serviços utilizados por cada grupo de turistas seleccionado para a amostra durante o período de referência, $R_i(B)$. É igualmente de notar que $R_i(B)$ poderá ser difícil de obter. Em muitos casos poderá ser obtido por via declarativa, perguntando ao representante do grupo o número de estabelecimentos que visitou no período de referência, mas na maioria dos casos a entrevista será realizada antes do final do seu período de permanência na região *G*, o que faz com que a informação transmitida no questionário não seja factual, mas meramente baseada em intenções.

Um estimador centrado para τ_Y^B (Deville e Maumy-Bertrand, 2006) é:

$$\hat{\tau}_Y^B = \sum_{i \in s^B} w_i y_i \,, \qquad (6)$$

onde os pesos w_i são inspirados pelo MPPG de acordo com (2), admitindo $\theta_{ji}^{AB}=1$ quando o serviço *j* for utilizado pelo grupo *i* e $\theta_{ji}^{AB}=0$ em caso contrário:

$$w_i = \frac{\sum_{l=1}^{2} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{Cd_l}} \frac{1}{\pi_j^{A_l}}}{R_i(B)} \,. \qquad (7)$$

Apesar da semelhança com o estimador do MPPG, este estimador aparece como uma adaptação do estimador original (2), distinguindo-se deste pelo facto de já não se supor que os valores de θ_{ji}^{AB} sejam conhecidos para todas as unidade $i=1,\ldots,N^B$. Isto é conseguido através de um artifício, supondo que as unidades da população U^A não são os "estabelecimentos × dias", mas os serviços por estes oferecidos e entendendo-se um serviço como uma utilização ou visita a um estabelecimento por parte de um grupo de turistas. Assim a população U^A é uma população de utilizações de estabelecimentos e a realização de uma amostragem dessas utilizações faz com que já não seja necessário conhecer todos os grupos de turistas. De facto o estimador é agora dado por:

$$\hat{\tau}_Y^B = \sum_{i \in s^B} \sum_{l=1}^{2} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{Cd_l}} \frac{z_j}{\pi_j^{A_l}} \,, \qquad (8)$$

onde $z_j = y_i / R_i(B)$, para $j \in F^{-1}(i)$. Note-se que a probabilidade de selecção $\pi_j^{A_l}$ é dada por $\pi_j^{A_l} = \pi_{l,ad}^{A} n_{l,ad} / N_{l,ad}$, onde $\pi_{l,ad}^{A}$ é a probabilidade de selecção do dia d_l no estabelecimento a_l (dependente do desenho amostral adoptado), $n_{l,ad}$ é o número de serviços observados no mesmo "estabelecimento × dia" e $N_{l,ad}$ é o número total de serviços consumidos no mesmo "estabelecimento × dia". Contudo, só em algumas situações práticas é possível obter directamente as probabilidades de inclusão $\pi_j^{A_l}$, $l=1, 2$, porque se desconhece a dimensão das populações, ou seja, porque se desconhece o número total de turistas que utiliza um serviço de um estabelecimento num determinado dia, $N_{l,ad}$. Quando $N_{l,ad}$ for desconhecido, então tem que ser estimado de forma a se poder estimar $\pi_j^{A_l}$.

Admite-se que a dimensão das populações é conhecida quando os turistas consomem serviços mediante a sua compra (e.g. através de um bilhete), ou quando existe um sistema de controlo e contagem automática de entradas/saídas. Quando a dimensão das populações é desconhecida, admite-se apenas que os turistas estão divididos em grupos mutuamente exclusivos e que é possível conhecer o número total de grupos na população. Admite-se que o número total de grupos pode ser obtido, por exemplo, através de contagem automática efectuada com apoio de sistemas de videovigilância.

Uma abordagem para a estimação do número de serviços consumidos pode passar por perguntar a cada turista, *i*, seleccionado para a amostra, o número u_i de indivíduos que constituem o grupo ao qual pertence. Nesta situação, Deville e Maumy-Bertrand (2006) propuseram o seguinte estimador de $N_{l,ad}$:

$$\hat{N}_{l,ad} = N_{l,ad}^{g} \hat{\bar{v}}_{l,ad} = N_{l,ad}^{g} \frac{n_{l,ad}}{\sum_{i \in s_{Cd_l}^B} \frac{1}{u_i}} \,, \qquad (9)$$

onde $N_{l,ad}^{g}$ é o número total de grupos, $\hat{\bar{v}}_{l,ad}$ é um

estimador da dimensão média dos grupos e $s^B_{Cd_l}$ representa a amostra de turistas que consome os serviços prestados por um estabelecimento a_l num determinado dia d_l. Segundo Deville e Maumy-Bertrand (2006), o estimador (9) é assimptoticamente não enviesado. Neste caso, um estimador da probabilidade de selecção $\pi_j^{A_l}$ é dado por $\hat{\pi}_j^{A_l}=\frac{\pi^A_{l,ad}}{N^g_{l,ad}}\left(\sum_{i\in s^B_{C_{d_l}}}\frac{1}{u_i}\right)$.

Depois de completamente deduzido o estimador do total, é altura de se avaliar a qualidade desse estimador. O enviesamento e a precisão são os critérios normalmente utilizados para avaliar essa qualidade. Tal como foi referido anteriormente, o estimador (6) é centrado, tal como foi provado por Lavallé (1995) e posteriormente por Deville e Lavallée (2006). Nesta situação, a capacidade do estimador (6) produzir estimativas próximas do verdadeiro valor do parâmetro a estimar pode ser avaliada com base na sua variância, a qual depende do desenho amostral adoptado.

Admitindo que a sondagem é efectuada com probabilidades iguais nas duas primeiras etapas, então tem-se que $\pi^A_{l,ad}=m_{l,ad}/M_{l,ad}=m_l/M_l$, l=1, 2, onde m_l e M_l representam, respectivamente, o número de subpopulações "estabelecimento × dia" pertencentes à amostra e à população do estrato U^{A_l}. Neste caso particular, esse estimador pode ainda ser apresentado como:

$$\hat{\tau}^B_Y=\sum_{l=1}^{2}\frac{M_l}{m_l}\sum_{s\in A_l}\sum_{s\in D_l}\hat{N}_{l,ad}\overline{z}_{l,ad}\,, \qquad (10)$$

onde $\overline{z}_{l,ad}=\frac{1}{n_{l,ad}}\sum_{j\in C_{dj}} z_j$ é um estimador da média da variável Z, $\mu_{zl,ad}$, ao nível individual em cada subpopulação "estabelecimento × dia". Note-se que podem existir situações em que $N_{l,ad}$ é conhecido, sendo desnecessária a sua estimação. Admitindo igual dimensão média das subpopulações "estabelecimento × dia" na amostra e na população de cada estrato, então uma aproximação do estimador da variância do estimador (8) é dada por:

$$\hat{V}\left(\hat{\tau}^B_Y\right)=\sum_{l=1}^{2}\hat{V}\left(\hat{\tau}^B_{zl}\right), \qquad (11)$$

onde $\hat{V}\left(\hat{\tau}^B_{zl}\right)$ é um estimador do total associado à variável Z no estrato l. Na estimação desta variância devem ser considerados dois casos distintos: o caso em que $N_{l,ad}$ é conhecido e o caso em que $N_{l,ad}$ é desconhecido. No primeiro caso, pelos resultados da sondagem em duas etapas tem-se que:

$$\hat{V}\left(\hat{\tau}^B_{zl}\right)=M_l^2\left(\frac{1}{m_l}-\frac{1}{M_l}\right)\hat{s}^2_{zl,\tau}+\frac{M_l}{m_l}\sum_{s\in D_{A_l}}N^2_{l,ad}\left(\frac{N_{l,ad}-n_{l,ad}}{N_{l,ad}n_{l,ad}}\right)\hat{s}^2_{zl,ad}\,, \qquad (12)$$

onde $\hat{s}^2_{zl,\tau}=\frac{1}{m_l-1}\sum_{s\in D_{A_l}}\left(\hat{\tau}_{zl,ad}-\hat{\mu}_{zl,\tau}\right)^2$,

$\hat{s}^2_{zl,ad}=\frac{1}{n_{l,ad}-1}\sum_{j\in C_{D_l}}\left(z_j-\overline{z}_{l,ad}\right)^2$, sendo

$\hat{\tau}_{zl,ad}=N_{l,ad}\overline{z}_{l,ad}$ um estimador do total na subpopulação do dia d_l no estabelecimento a_l e $\hat{\mu}_{zl,\tau}=\frac{1}{m_l}\sum_{s\in D_{A_l}}\hat{\tau}_{zl,ad}$ um estimador do total médio por "estabelecimento × dia" no estrato l. No caso em que $N_{l,ad}$ é desconhecido, existem duas componentes de aleatoriedade em (10), pelo que:

$$\hat{V}\left(\hat{\tau}^B_{zl}\right)\approx M_l^2\left(\frac{1}{m_l}-\frac{1}{M_l}\right)\hat{s}^2_{zl,\tau}+\frac{M_l}{m_l}\sum_{s\in D_{A_l}}\hat{V}\left(\hat{N}_{l,ad}\overline{z}_{l,ad}\right). \qquad (13)$$

Como se pode observar, a primeira parcela do estimador (13) é igual à primeira parcela do estimador (12). No que se refere à segunda parcela, e admitindo que as variáveis $\hat{N}_{l,ad}$ e $\overline{z}_{l,ad}$ são independentes, obtém-se pelo teorema de Huygens que $V\left(\hat{N}_{l,ad}\overline{z}_{l,ad}\right)=N^2_{l,ad}V\left(\overline{z}_{l,ad}\right)+\overline{Z}^2_{l,ad}V\left(\hat{N}_{l,ad}\right)+V\left(\hat{N}_{l,ad}\right)V\left(\overline{z}_{l,ad}\right)$. Pelos resultados obtidos por Deville e Maumy-Bertrand (2006), tem-se que $V\left(\hat{N}_{l,ad}\right)=\left(N^g_{l,ad}\right)^2\overline{v}^4_{l,ad}\left(\frac{1}{n_{l,ad}}-\frac{1}{N_{l,ad}}\right).S^2_{\sqrt{u},l,ad}$ e

$V\left(\overline{z}_{l,ad}\right)=\frac{N_{l,ad}-n_{l,ad}}{N_{l,ad}n_{l,ad}}S^2_{zl,ad}$. Os estimadores destas últimas variâncias são dados, respectivamente, por $\hat{V}\left(\hat{N}_{l,ad}\right)=\left(N^g_{l,ad}\right)^2\hat{\bar{v}}^4_{l,ad}\frac{\hat{s}^2_{\sqrt{u},l,ad}}{n_{l,ad}}$, admitindo que $1/N_{l,ad}\to 0$, e por $\hat{V}\left(\overline{z}_{l,ad}\right)=\frac{\hat{N}_{l,ad}-n_{l,ad}}{\hat{N}_{l,ad}n_{l,ad}}\hat{s}^2_{zl,ad}$, onde $\hat{\bar{v}}_{l,ad}=\frac{n_{l,ad}}{\sum\limits_{i\in s^B_{C_{dl}}}\frac{1}{u_i}}$ é um estimador da dimensão média dos grupos e $\hat{s}^2_{\sqrt{u},l,ad}=\frac{1}{n_{l,ad}-1}\sum\limits_{i\in s^B_{C_{dl}}}\left(\frac{1}{u_i}-\frac{1}{\hat{\bar{v}}_{l,ad}}\right)^2$.

De posse destes resultados, pode-se não só estimar o total da variável de interesse, assim como avaliar a qualidade dessa estimação através das conhecidas medidas de precisão (coeficiente de variação, precisão absoluta, precisão relativa, entre outras). A medida de precisão mais utilizada no contexto das sondagens é a precisão relativa, a qual indica, com um determinado grau de confiança, o afastamento relativo máximo entre o verdadeiro parâmetro e a estimativa. Um estimador dessa medida, apresentado sob a forma de percentagem, é o seguinte:

$$\hat{PR}=\frac{z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\tau}^B_Y\right)}}{\hat{\tau}^B_Y}\times 100\%\,, \qquad (14)$$

onde $\hat{\tau}^B_Y$ é dado por (10), $\hat{V}\left(\hat{\tau}^B_Y\right)$ é dado por (11) e $z_{1-\alpha/2}$ é o quantil de ordem $\left(1-\alpha/2\right)$ de $Z\sim N(0,1)$.

3.6. Aplicação numérica

De forma a apresentar a aplicação das sondagens indirectas na estimação do montante total dos gastos locais efectuados pelos turistas de sol e praia, vai apresentar-se um exemplo numérico. Admita-se que na região de interesse, *G*, a população é formada por todas as visitas a 10 praias e 8 pontos de interesse (existindo dois estratos: $l=1$ para a população das praias e $l=2$ para a população dos pontos de interesse). Admita-se também que o período de referência, *P*, é de 30 dias e que é efectuada uma sondagem em três etapas em cada estrato, tal como descrito anteriormente ($M_1 = 10 \times 30 = 300$ e $M_2 = 8 \times 30 = 240$). Por simplicidade de cálculo, admita-se que:

- foi seleccionada uma amostra aleatória simples sem reposição de 4 praias num determinado dia (A_1 é formado pelas praias 1, 2, 3 e 4; $m_1 = 4$);
- foi seleccionada uma amostra aleatória simples sem reposição de 3 pontos de interesse num determinado dia (A_2 é formado pelos pontos de interesse 1, 2 e 3; $m_2 = 3$);
- foi seleccionada uma amostra de 10, 5, 10 e 8 turistas, respectivamente, nas praias 1, 2, 3 e 4 ($n_{1,11}=10$, $n_{1,21}=5$, $n_{1,31}=10$ e $n_{1,41}=8$); e uma amostra de 8, 6 e 10 turistas, respectivamente, nos pontos de interesse 1, 2 e 3 ($n_{2,11}=8$, $n_{2,21}=6$ e $n_{2,31}=10$);
- a recolha de dados foi efectuada entre as 9 e as 19 horas.

Admita-se que o número total de serviços consumidos nos três pontos de interesse no dia de recolha de dados ($N_{2,11}=200$, $N_{2,21}=190$ e $N_{2,31}=210$) é conhecido, por exemplo pelo facto do consumo exigir a compra de um ingresso. Pelo contrário, admita-se que só é conhecido o número total de grupos na população que visita cada uma das quatro praias ($N^g_{1,11}=80$, $N^g_{1,21}=60$, $N^g_{1,31}=75$ e $N^g_{1,41}=70$). A cada turista, *i*, seleccionado para cada amostra s^A foi questionado, entre outras coisas, o número de pessoas que constituem o grupo ao qual pertence, u_i, o montante despendido pelo grupo em gastos locais (em unidades monetárias), y_i, e o número de serviços utilizados pelo grupo, $R_i(B)$, durante o período de referência. Esses dados encontram-se no quadro 1.

No quadro 2 encontram-se todas as estimativas necessárias para o cálculo da estimativa do montante total dos gastos locais efectuados pelos turistas na região de interesse *G* durante o período *P*, assim como da sua precisão. Com base no estimador (10), obtém-se a seguinte estimativa do montante total dos gastos locais efectuados pelos turistas:

$$\hat{\tau}_Y^B = \frac{300}{4}\left(181\times 86.50 + 189\times 92.00 + 186\times 90.50 + \right.$$
$$\left. + 164\times 86.25\right) + \frac{240}{3}\left(200\times 60.00 + 190\times 68.33 + \right.$$
$$\left. + 210\times 61.50\right) = 7,833,504$$

A avaliação da qualidade da estimativa anterior é baseada na estimativa da sua variância. Utilizando os estimadores (13), (12) e (10), tem-se respectivamente:

$$\hat{V}\left(\hat{\tau}_{z1}^B\right) = 300^2\left(\frac{1}{4} - \frac{1}{300}\right)\times 2,105,798.65 +$$
$$\frac{300}{4}\left[2,072,696.28 + 3,515,221.56 + \right.$$
$$\left. 5,830,553.44 + 4,801,724.30\right] =$$
$$47,965,224,698.50$$

$$\hat{V}\left(\hat{\tau}_{z2}^B\right) = 240^2\left(\frac{1}{3} - \frac{1}{240}\right)\times 301,473.15 +$$
$$+ \frac{240}{3}\left[200^2\left(\frac{200-8}{200\times 8}\right)\times 1,107.14 + \right.$$
$$+ 190^2\left(\frac{190-6}{190\times 6}\right)\times 1,176.67 +$$
$$\left. + 210^2\left(\frac{210-10}{210\times 10}\right)\times 689.17\right] = 6,921,118,913.33$$

e $\hat{V}\left(\hat{\tau}_Y^B\right) = 47,965,244,698.50 + 6,921,118,913.33$
$= 54,886,363,612$.

Por último, obtém-se através da expressão (14) e para um grau de confiança de 95%, que a precisão relativa associada à estimativa do montante total dos gastos locais efectuados pelos turistas na região de interesse *G* durante o período *P* é de 5.9%.

Quadro 1 | Dados recolhidos em cada amostra de turistas

i		1	2	3	4	5	6	7	8	9	10
$a_{1,11}$	u_i	3	4	4	2	1	2	2	3	4	2
	y_i	500	240	160	700	130	840	180	140	360	400
	$R_i(B)$	5	2	2	7	2	12	3	2	3	5
$a_{1,21}$	u_i	4	2	4	3	4	-	-	-	-	-
	y_i	1,200	300	540	1,000	360	-	-	-	-	-
	$R_i(B)$	10	5	6	10	4	-	-	-	-	-
$a_{1,31}$	u_i	1	3	3	5	4	3	2	3	4	2
	y_i	600	300	140	800	1,260	160	130	700	440	500
	$R_i(B)$	20	3	2	5	9	2	2	7	4	10
$a_{1,41}$	u_i	2	3	4	1	4	2	3	4	-	-
	y_i	360	800	280	280	300	480	500	1,650	-	-
	$R_i(B)$	6	8	2	7	3	12	5	15	-	-
$a_{2,11}$	u_i	4	3	4	5	2	2	1	3	-	-
	y_i	500	360	480	600	100	1,000	300	180	-	-
	$R_i(B)$	10	3	8	6	4	20	10	4	-	-
$a_{2,21}$	u_i	2	3	2	4	4	2	-	-	-	-
	y_i	420	100	600	960	500	240	-	-	-	-
	$R_i(B)$	7	2	12	8	5	8	-	-	-	-
$a_{2,31}$	u_i	2	3	6	5	4	2	3	4	2	2
	y_i	360	700	850	120	450	250	70	1,600	360	210
	$R_i(B)$	6	7	10	2	5	10	2	20	8	6

Quadro 2 | Estimativas de médias, totais e variâncias

	$\bar{z}_{l,ad}$	$\hat{\tau}_{zl,ad}$	$\hat{s}^2_{zl,ad}$	$\hat{\bar{v}}_{l,ad}$	$\hat{S}^2_{\forall u,l,ad}$	$\hat{N}_{l,ad}$	$\hat{\mu}_{zl,\tau}$	$\hat{s}^2_{zl,\tau}$
$a_{1,11}$	86.50	15,667.92	489.17	2.26	0.051	181	16,016.15	2,105,798.65
$a_{1,21}$	92.00	17,431.58	470.00	3.16	0.012	189		
$a_{1,31}$	90.50	16,828.51	1,591.39	2.48	0.054	186		
$a_{1,41}$	86.25	14,136.59	1,283.93	2.34	0.064	164		
$a_{2,11}$	60.00	12,000.00	1,107.14	-	-	-	12,632.78	301,473.15
$a_{2,21}$	68.33	12,983.33	1,176.67	-	-	-		
$a_{2,31}$	61.50	12,915.00	689.17	-	-	-		

Pode-se então afirmar, com 95% de confiança, que o verdadeiro montante total dos gastos locais efectuados pelos turistas não se afasta mais do que 5,9% de 7,833,504.

4. Conclusões

Neste artigo apresentou-se o método de sondagem indirecto em conjunto com o método desenvolvido para a obtenção dos pesos de estimação: o MPPG. A utilização desta metodologia permite resolver o problema da inexistência de uma base de amostragem da população alvo, o qual é frequente nas sondagens realizadas na área do turismo em ambiente aberto. Este método baseia-se na observação de turistas quando estes estão a consumir um determinado serviço em estabelecimentos seleccionados aleatoriamente de uma população auxiliar. Com base numa matriz de ligação, é possível atribuir os pesos da estimação através da utilização do MPPG. Neste artigo foi igualmente proposta uma adaptação desta metodologia geral ao caso de amostragem multi-etápica, a qual é frequentemente utilizada em estudos na área do turismo no contexto do turismo e desenvolvido um formulário que suporta a estimação e avaliação da precisão neste contexto.

Torna-se assim possível efectuar amostragem probabilística no contexto do turismo, onde as abordagens até há pouco tempo disponíveis estavam confinadas ao domínio empírico. Note-se que um dos principais méritos desta nova abordagem consiste em tornar possível o cálculo de medidas de precisão, fornecendo medidas objectivas da validade das estimativas disponibilizadas (com abordagens empíricas essa validade não podia ser acedida senão subjectivamente).

Apresentou-se também uma aplicação numérica do método de sondagem indirecto, na estimação do montante total dos gastos locais efectuados pelos turistas de sol e praia, numa dada região de interesse e durante um certo período de referência.

A capacidade de adaptação do método de sondagem indirecto a uma grande variedade de situações torna-o bastante apelativo. Todavia, este método pode carecer de adaptações várias quando aplicado a problemas de estimação específicos, nomeadamente devido à utilização de desenhos amostrais diferentes daquele que foi apresentado neste artigo. De facto, o plano de sondagem subjacente à aplicação apresentada neste artigo, é suportado por um desenho amostral multi-etápico. Apesar de este ser provavelmente o desenho mais adaptado à generalidade dos inquéritos utilizados na área do turismo (especialmente se em ambiente aberto), não é de excluir a existência de situações específicas onde se aconselhe o recurso a outros desenhos amostrais. Por exemplo, uma cadeia de hotéis ou uma cadeia detentora de determinados equipamentos de lazer poderá recorrer a um inquérito suportado por uma amostragem estratificada por hotel/equipamento. Apesar das adaptações necessárias à metodologia proposta dependerem do conhecimento do enquadramento concreto do problema, é de sublinhar que esta metodologia apresenta a flexibilidade necessária para poder ser aplicada a problemas que ocorram na área do turismo, independentemente do plano de sondagem subjacente aos mesmos.

Agradecimentos

Luís N. Pereira foi beneficiário de uma bolsa de investigação para doutoramento da Fundação para a Ciência e a Tecnologia (SFRH/BD/36764/2007). Os autores agradecem os comentários e sugestões pertinentes dos revisores, os quais permitiram melhorar este artigo.

Bibliografia

Deville, J.-C., Lavallée, P., 2006, Indirect Sampling: The Foundations of the Generalized Weight Share Method, *Survey Methodology*, Vol. 32(2), pp. 165-176.

Deville, J.-C., Maumy-Bertrand, M., 2006, Extension of the Indirect Sampling Method and its Application to Tourism, *Survey Methodology*, Vol. 32(2), pp. 177-185.

Horvitz, D.G., Thompson, D.J., 1952, A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, Vol. 47, pp. 663-685.

Kalton, G., Brick, J.M., 1995, Weighting Schemes for Household Panel Surveys, *Survey Methodology*, Vol. 21(1), pp. 33-44.

Lavallée, P., 1995, Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method, *Survey Methodology*, Vol. 21(1), pp. 25-32.

Lavallée, P., 2002, *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*, Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P., Caron, P., 2001, Estimation Using the Generalised Weight Share Method: The Case of Record Linkage, *Survey Methodology*, Vol. 27(2), pp. 155-169.

Särndal, C.-E., Swensson, B., Wretman, J., 1992, *Model assisted survey sampling*, Springer-Verlag, New-York.

[**Submitted** 11 March 2008; **accepted** 28 October 2008]