SONDAGENS A TURISTAS: COMO MEDIR O ERRO AMOSTRAL?

Luís N. Pereira

Mestre em Estatística e Gestão de Informação, Professor Adjunto, ESGHT - Universidade do Algarve
Lmper@ualg.pt

Lara N. Ferreira

Doutora em Economia Matemática e Modelos Econométricos, Professora Adjunta, ESGHT - Universidade do Algarve
Investigadora do Centro de Estudos e Investigação em Saúde da Universidade de Coimbra (CEISUC)
Lnferrei@ualg.pt

Resumo:
Na maioria das sondagens realizadas na área do turismo não tem existido a preocupação na medição do erro amostral das estimativas produzidas para os parâmetros de interesse. Desta forma, uma vez realizada uma sondagem a turistas e publicados os seus resultados, desconhece-se geralmente quão próximas se encontram essas estimativas dos respectivos verdadeiros valores dos parâmetros populacionais desconhecidos. Este artigo pretende fazer uma revisão das principais metodologias existentes para medir o erro amostral em sondagens a turistas.

Palavras-Chave: Dimensão amostral, erro amostral, estimação, método de sondagem indirecta, sondagens no turismo.

Abstract:
In the majority of the tourism surveys there have not been a concern on measuring the sampling error of the parameters of interest estimates computed. Once a tourism survey has been conducted and its results have been published, we usually don't know how close are the estimates and the respective true values of the unknown population parameters. This paper aims to review the main known methodologies that can be used to measure the sampling error in tourism surveys.

Keywords: Sample size, survey error, estimation, indirect sampling method, tourism surveys.

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 inferência dos resultados amostrais para a população. No contexto do turismo são frequentemente utilizadas sondagens para estudar características de interesse, mas raros são os estudos publicados nos quais são divulgadas as margens de erro amostral associadas às estimativas produzidas. Na verdade, esta situação resulta do facto de muitos investigadores não disporem de uma lista de turistas em condições de constituir uma base de sondagem com qualidade ou do facto da simples identificação dos turistas poder tornar-se numa tarefa muito complexa, ou mesmo impossível. Por sua vez, estas dificuldades conduzem à utilização de um método de amostragem não probabilístico na selecção da amostra, impossibilitando, por esta via, a avaliação objectiva da qualidade das estimativas produzidas em estudos a turistas. Desta forma, a medição do erro amostral, ou, por outras palavras, a determinação da precisão associada aos resultados obtidos na estimação realizada a partir de uma amostra de determinada dimensão, tem-se tornado o propósito da investigação de alguns autores, tendo inclusive sido publicados na literatura recente alguns artigos sobre esta e outras matérias relacionadas (veja-se, por exemplo, Beaman *et al.*, 2004). O objectivo deste trabalho consiste em revisitar duas das principais metodologias existentes para medir, se bem que de forma aproximada, o erro amostral em sondagens a turistas.

Na secção 2 é apresentada uma metodologia baseada no pressuposto que a sondagem utilizada, que se admite ser não probabilística devido à inexistência de uma base de sondagem da população alvo de turistas, segue um padrão semelhante à de uma sondagem aleatória simples. Em oposição a esta metodologia simplista, é apresentada na secção 3 uma metodologia baseada no método de sondagem indirecta, a qual permite resolver o problema da inexistência de uma base de sondagem da população alvo de turistas. O artigo termina com uma conclusão na secção 4.

2. METODOLOGIA SIMPLISTA DE MEDIÇÃO DO ERRO AMOSTRAL

Na maioria das sondagens realizadas na área do turismo não existe uma base de sondagem da população alvo de turistas, o que dificulta o estabelecimento de um contacto com os turistas, com o intuito de recolha de informação. Por esta razão, muitos investigadores são levados a utilizar uma sondagem não probabilística, a qual não permite medir directamente e sem aproximações a precisão das estimativas produzidas em função da dimensão amostral utilizada e do grau de confiança pretendido nos resultados. Uma forma simplista de resolver este problema consiste em admitir que a amostra empírica de turistas segue um padrão[23] semelhante ao que se obteria numa amostra extraída segundo um plano de sondagem aleatório

[23] Note-se que podem ser considerados métodos de aleatorização numa amostra empírica, fazendo-se uma escolha aleatória dos locais e dos períodos de recolha de dados, definindo-se itinerários aleatórios a serem percorridos pelos entrevistadores, definindo-se formas sistemáticas de extracção dos turistas para a amostra, etc.

simples com reposição, utilizando-se para tal o formulário disponível para se calcularem estimativas grosseiras da precisão associadas às estimativas dos parâmetros de interesse. Como resultado da utilização desse formulário simples, disponível em qualquer livro de Estatística, parece existir a ideia comum que um determinado nível de precisão é aproximadamente atingido para todas as estimativas produzidas (quer sejam percentagens, médias ou totais) a partir de uma amostra de determinada dimensão pré-definida. Uma dimensão amostral muitas vezes utilizada como permitindo obter-se um erro amostral de 5% para uma confiança de 95% é de 385 turistas. Esta dimensão pode ser obtida a partir da seguinte fórmula de cálculo no âmbito da estimação de percentagens (i.e. de proporções) em sondagens aleatórias simples com reposição:

$$n = \frac{z_{1-\alpha/2}^2 p(1-p)}{d^2}, \quad (1)$$

onde n representa a dimensão amostral, $z_{1-\alpha/2}$ representa o quantil da distribuição normal estandardizada correspondente ao grau de confiança escolhido (para um grau de confiança de 95%, um dos mais utilizados nas aplicações práticas, obtém-se $z_{1-\alpha/2} = 1{,}96$), p é uma estimativa da proporção amostral de um parâmetro de interesse (quando este valor é desconhecido considera-se o caso mais pessimista, i.e., p=0,5) e d é a precisão absoluta, mais vulgarmente conhecida como erro amostral, que corresponde à diferença máxima entre a estimativa e o verdadeiro parâmetro desconhecido da população. A precisão relativa, muitas vezes utilizada na estimação de médias e totais por ter uma interpretação mais intuitiva pelo facto de ser normalmente multiplicada por 100 e expressa em percentagem, é obtida a partir da precisão absoluta através da sua divisão pela estimativa pontual do parâmetro de interesse.

A ideia falaciosa de que uma dimensão amostral de 385 turistas produz, em geral, estimativas aproximadamente centradas dos parâmetros de interesse (com erro não superior a 5%) pode estar totalmente errada, pois é necessário ter-se em atenção os pressupostos subjacentes à fórmula (1).

Em primeiro lugar, é necessário considerar-se que a fórmula (1) admite que a sondagem foi efectuada com reposição, o que não é a situação mais realista em sondagens efectuadas nos dias de hoje. No caso de uma sondagem aleatória simples sem reposição, a fórmula de cálculo da dimensão da amostra no âmbito da estimação de percentagens é a seguinte:

$$n = \frac{\dfrac{z_{1-\alpha/2}^2 p(1-p)}{d^2}}{1 + \dfrac{z_{1-\alpha/2}^2 p(1-p)}{d^2 N}}, \quad (2)$$

onde N representa a dimensão populacional (no caso desta dimensão ser totalmente desconhecida, é-se forçado a utilizar a fórmula (1)).

Em segundo lugar, é necessário ter em atenção que as fórmulas (1) e (2) só são válidas quando se pretende estimar proporções relativas às características de interesse (por exemplo, a proporção de turistas satisfeitos com um determinado serviço). Contudo, podem existir outros parâmetros de interesse, tais como médias (por exemplo, o número médio de dias da estadia dos turistas num determinado destino) ou totais (por exemplo, o montante total dos gastos locais efectuados pelos turistas). Nestas situações, as fórmulas de cálculo das dimensões amostrais sem reposição são as seguintes, respectivamente:

$$n = \frac{\dfrac{z_{1-\alpha/2}^2 s^2}{d^2}}{1 + \dfrac{z_{1-\alpha/2}^2 s^2}{d^2 N}} \quad (3)$$

e

$$n = \frac{\dfrac{z_{1-\alpha/2}^2 N^2 s^2}{d^2}}{1 + \dfrac{z_{1-\alpha/2}^2 N s^2}{d^2}}, \quad (4)$$

onde s^2 representa uma estimativa da variância amostral da característica de interesse.

Em terceiro lugar, é necessário considerar que as fórmulas usadas no dimensionamento das amostras dizem respeito à estimação de um único parâmetro associado a uma característica de interesse[24], mas, na prática, num estudo a turistas são normalmente estimados diferentes parâmetros relativos a diferentes características de interesse. Surge então a questão de saber qual deve ser a variável utilizada para efectuar esse dimensionamento. Este problema pode ser ultrapassado através de uma de

[24] Na verdade, esta situação pode também ser considerada como o caso mais pessimista, em que se está a dimensionar uma amostra respeitante à estimação de um qualquer parâmetro, havendo a garantia que a precisão planeada é garantida para qualquer variável do estudo. O caso mais típico consiste em dimensionar uma amostra para o caso da estimação de uma proporção e considerar a situação de variância máxima. Naturalmente que o preço a pagar nesta abordagem será o sobredimensionamento da amostra, resultando num custo de recolha de dados, monetário e temporal, superior ao que seria necessário para a precisão desejada/planeada.

duas formas: (*a*) dimensionar a amostra para a estimação da variável mais importante ou "central" do questionário, como por exemplo a satisfação global média dos turistas com um determinado destino; ou (*b*) dimensionar a amostra para a estimação de um conjunto de variáveis chave do questionário. Neste último caso poderão ser planeados níveis de precisão diferenciados para cada variável desse conjunto, tendo em conta a natureza e a importância das várias variáveis em estudo, resultando naturalmente em dimensões amostrais diferentes. A dimensão amostral final deverá ser a dimensão amostral máxima, de forma a assegurar a precisão planeada para todas as variáveis.

Em quarto lugar, é importante notar que todo o formulário apresentado pressupõe que se pretende estimar as características de interesse para toda a população. No caso de se pretender estimar essas características para grupos específicos da população de turistas (por exemplo, por nacionalidade, pelo motivo da viagem, etc.), deve-se dimensionar uma amostra de turistas para cada um dos grupos, de forma a que em cada grupo se obtenham resultados com a precisão desejada inicialmente. A dimensão da amostra global resultará da soma das dimensões das subamostras associadas a cada segmento, determinadas através do formulário apresentado acima.

Por último, note-se que a abordagem apresentada nesta secção pressupõe um adequado planeamento da sondagem, i. e., que são definidas *a priori* a precisão e a confiança desejada nos resultados, em função das quais se determina o número de turistas a inquirir. Contudo, por vezes a determinação da dimensão da amostra não é regida puramente por métodos estatísticos, mas sim por critérios económicos ou temporais (por exemplo, o tempo e/ou o orçamento disponível[25] para a recolha de dados num estudo a turistas), sendo necessário avaliar-se *a posteriori* a precisão dos resultados em função da dimensão da amostra de turistas que foi possível observar, *n*. No âmbito da estimação de proporções, de médias ou de totais, as aproximações para a precisão absoluta dos resultados podem ser obtidas, respectivamente, através das seguintes expressões:

$$d = z_{1-\alpha/2}\sqrt{\frac{(N-n)}{nN}p(1-p)}, \tag{5}$$

$$d = z_{1-\alpha/2}\sqrt{\frac{(N-n)}{nN}s^2} \tag{6}$$

e

$$d = z_{1-\alpha/2}\sqrt{\frac{(N-n)N}{n}s^2}\,. \tag{7}$$

3. MÉTODO DE SONDAGEM INDIRECTA

Tal como foi referido anteriormente, na maioria das sondagens realizadas na área do turismo não existe uma base de sondagem da população alvo de turistas, o que dificulta o estabelecimento de um contacto com os turistas, com o intuito de recolha de informação. Para além disso, a inexistência de uma base de sondagem da população em estudo inviabiliza a utilização de um método de amostragem probabilístico, com todas as desvantagens daí resultantes. Na secção anterior apresentou-se uma metodologia simplista para a medição do erro amostral, mesmo utilizando um método de amostragem não probabilístico e, portanto, abdicando da obtenção de uma base de sondagem da população de turistas. Nesta secção, apresenta-se outra metodologia que permite medir esse erro amostral, a qual é suportada na existência de uma base de amostragem de uma população relacionada com a população de turistas.

O problema da inexistência de uma base de sondagem 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.

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, $\mathbf{\Theta}_{AB} = \left[\theta_{ji}^{AB}\right]$, de dimensão $N^A \times N^B$ 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.

---

[25] Quando o orçamento de um estudo a turistas é definido à partida e admitindo que a parte referente à recolha de dados é a única componente variável desse orçamento, então o número de turistas a inquirir é determinada a partir da seguinte expressão: n = [(orçamento total – custos fixos)/custo de observação de cada turista].

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 sondagem disponível de U^A, com o objectivo de produzir estimativas para os parâmetro 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, 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 $\mathbf{Y} = \left(y_1, \ldots, y_{N^B}\right)'$ 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:

$$\hat{\tau}_Y^B = \sum_{i=1}^{N^B} w_i y_i , \tag{8}$$

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 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). Foi então proposto por Lavallée (1995, 2002) 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:

$$w_i = \sum_{j=1}^{N^A} \frac{t_j^A \tilde{\theta}_{ji}^{AB}}{\pi_j^A}, \tag{9}$$

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:

$$\hat{\tau}_Y^B = \sum_{i=1}^{N^B} \sum_{j=1}^{N^A} \frac{t_j^A \tilde{\theta}_{ji}^{AB}}{\pi_j^A} y_i = \sum_{i=1}^{N^B} \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB} y_i}{\pi_j^A}. \tag{10}$$

Utilizando um peso do tipo (2), Lavallée (1995) mostrou que o estimador (3) é não enviesado. Lavallée (1995) mostrou também que a variância do estimador (3) é dada por:

$$V\left(\hat{\tau}_Y^B\right) = \sum_{j=1}^{N^A} \sum_{j'=1}^{N^A} \frac{\left(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A\right)}{\pi_j^A \pi_{j'}^A} z_j z_{j'}, \tag{11}$$

onde $\pi_{jj'}^A$ é a probabilidade das unidades *j* e *j'* serem seleccionadas para a amostra s^A e $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$. Utilizando diferentes metodologias, Kalton e Brick (1995), Lavallée e Caron (2001) e Lavallée (2002) sugeriram que as ligações entre as duas populações sejam definidas da seguinte forma: $\theta_{ji}^{AB,opt} = 1$ quando $\theta_{ji}^{AB} > 0$ e $\theta_{ji}^{AB,opt} = 0$ quando $\theta_{ji}^{AB} = 0$.

De posse destes resultados, torna-se então possível avaliar a precisão das estimativas produzidas através de sondagens realizadas a turistas, qualquer que seja o método de amostragem probabilístico utilizado. Para tal, basta apenas adaptar o formulário desta metodologia geral ao caso particular de uma qualquer amostragem probabilística, seja ela aleatória simples, estratificada, por conglomerados, multi-etápica ou qualquer outra sondagem complexa. Após esta adaptação, a precisão absoluta é dada pela seguinte expressão geral:

$$d = z_{1-\alpha/2} \sqrt{\hat{V}\left(\hat{\tau}_Y^B\right)}. \tag{12}$$

Uma aplicação desta metodologia para o caso de uma sondagem multi-etápica pode ser encontrada em Pereira e Coelho (2009).

4. CONCLUSÃO

No contexto do turismo são frequentemente utilizadas sondagens para estudar características de interesse de uma população, embora a maioria dos estudos efectuados não apresente o erro amostral das estimativas produzidas. No entanto, à medida que nos trabalhos realizados no turismo têm vindo a ser cada vez mais incluídas análises de dados, tem também vindo a crescer a necessidade de incluir a temática do erro amostral das estimativas produzidas, como forma de avaliar a qualidade dos resultados obtidos. Neste sentido, disponibilizar uma revisão dos métodos de medição do erro amostral tem implicações práticas importantes.

Neste artigo foram revisitadas duas metodologias de cálculo do erro amostral em sondagens a turistas, designadamente a simplista e a baseada no método de sondagem indirecta. A revisão destas metodologias tem como intuito esclarecer os utilizadores sobre algo fundamental em qualquer tipo de sondagens e que, naturalmente, deve ser incluído na apresentação dos resultados de sondagens a turistas.

Em conclusão, parece ter ficado claro que a precisão desejada nos resultados e o orçamento disponível determinam as dimensões amostrais de turistas utilizadas nas aplicações práticas. Quando a determinação da dimensão da amostra é efectuada com base em critérios exclusivamente orçamentais há o risco de se virem a produzir resultados com um nível de precisão desadequado aos objectivos do estudo, o que em alguns casos poderá inviabilizar a sua utilização. Por outro lado, quando a determinação da dimensão da amostra é efectuada com base na precisão desejada existe a possibilidade de se determinarem dimensões amostrais de custo financeiro e temporal incomportável. Assim, será prudente a realização de simulações que permitam ao investigador escolher de entre um conjunto de cenários a dimensão amostral a utilizar, tendo simultaneamente em conta a precisão e os custos associados.

AGRADECIMENTOS

Luís N. Pereira é beneficiário de uma bolsa de investigação para doutoramento (SFRH/BD/36764/2007) da Fundação para a Ciência e a Tecnologia.

REFERÊNCIAS

BEAMAN, J.G., HUAN, T.C., e BEAMAN, J.P. (2004), "Tourism Surveys: Sample Size, Accuracy, Reliability, and Acceptable Error", *Journal of Travel Research*, 43, 1, 67-74.

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, 260, 663-685.

KALTON, G. e BRICK, J.M. (1995), "Weighting Schemes for Household Panel Surveys", *Survey Methodology*, 21, 1, 33-44.

LAVALLÉE, P. (1995), "Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method", *Survey Methodology*, 21, 1, 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., e CARON, P. (2001), "Estimation Using the Generalised Weight Share Method: The Case of Record Linkage", *Survey Methodology*, 27, 2, 155-169.

PEREIRA, L. e COELHO, P. (2009), "Aplicação de Sondagens Indirecta no Turismo", *Revista Turismo e Desenvolvimento*, 12. (aceite).

SÄRNDAL, C. E., SWENSSON, B. e WRETMAN, J. (1992), *Model assisted survey sampling*, Springer-Verlag, New-York.