Survey Methodology, December 2006
Vol. 32, No. 2, pp. 177-185
Statistics Canada, Catalogue No. 12-001

Extension of the Indirect Sampling Method and its Application to Tourism

Jean-Claude Deville and Myriam Maumy-Bertrand [1]

Abstract

A survey of tourist visits originating intra and extra-region in Brittany was needed. For concrete material reasons, “border surveys” could no longer be used. The major problem is the lack of a sampling frame that allows for direct contact with tourists. This problem was addressed by applying the *indirect sampling method*, the weighting for which is obtained using the *generalized weight share method* developed recently by Lavallée (1995), Lavallée (2002), Deville (1999) and also presented recently in Lavallée and Caron (2001). This article shows how to adapt the method to the survey. A number of extensions are required. One of the extensions, designed to estimate the total of a population from which a Bernouilli sample has been taken, will be developed.

Key Words: Generalized weight share method; Incomplete frame and multiple frames.

1. Introduction

A “border survey” of extra-region tourist visits in Brittany (those not by residents of Brittany) was conducted over the period from April to September 1997. The Observatoire Régional du Tourisme de Bretagne and the Comités Départementaux de Tourisme were interested in doing another one. Unfortunately, they no longer had the means to gather a certain mass of data at the regional or intra-regional borders because the police forces were no longer interested in collaborating on roadside surveys.

For this reason, the Observatoire Régional du Tourisme de Bretagne, with the assistance of a technical committee comprised of methodologists and field operators, decided to introduce a new survey methodology to replace the “border survey” methodology. In addition, evaluation of intra-regional tourism (of residents of Brittany vacationing in Brittany, for example) is vital to identifying development factors.

One of the major problems is the lack of a sampling frame that allows direct communication with tourists. This problem was addressed by using an approach previously used in the Asturias in Spain (Valdés, De La Ballina, Aza, Loredo, Torres, Estébanez, Domínguez and Del Valle (2001) and Torres Manzanera, Sustacha Melijosa, Menéndez Estébanez and Valdés Pelaáez (2002)), which involves sampling services intended mainly for tourists and asking them questions at the various locations of these many tourist service sites. Obviously, a tourist may use one or more of the services in the sampling frame once or several times during the survey period in question. To be able to estimate the parameters of interest with respect to tourists, it must be possible to conduct a rigorous sample of certain services and then link the set of weights of the sampled services to the set of weights of the tourists the tourists who used these services. The purpose of this article is to present a method that makes this calculation possible. This method relies mainly on the generalized weight share method (GWSM) developed by Lavallée (1995), Lavallée (2002) and Deville (1999).

2. Generalized Weight Share Method

We will briefly review the principle of the *generalized weight share method* (GWSM). For more information, see Lavallée (1995), Lavallée (2002) and Deville (1999).

We will let U^A be a finite population containing N^A units, where each unit is denoted by j and U^B is a finite population containing N^B units, where each unit is denoted by i. The correspondence between U^A and U^B can be represented by a matrix of links $\Theta_{AB} = [\theta_{ji}^{AB}]$, of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. In other words, the unit j of U^A is linked to unit i of U^B provided that $\theta_{ji}^{AB} > 0$; otherwise, there is no link between these two units.

In the case of the indirect survey, we select the sample s^A of n^A units from U^A based on a given sampling design. Let $\pi_j^A > 0$, be the probability of selection of the unit j. For each unit j selected in s^A, we identify the units i of U^B for which $\theta_{ji}^{AB} > 0$. Then we let s^B, be all of the n^B units of U^B identified using the units $j \in s^A$, that is,

$$s^B = \{i \in U^B\, ;\, \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}.$$

For each unit i of s^B, a variable of interest y_i is measured.

It is assumed that, for any unit j of s^A, it is possible to obtain the values of θ_{ji}^{AB} for $i = 1, ..., N^B$ by a direct interview or from an administrative source. For any unit i

[1] Jean-Claude Deville, Laboratoire de Statistique d’Enquêtes, ENSAI/CREST, Campus de Ker-Lann, 35170 BRUZ (France). E-mail: deville@ensai.fr; Myriam Maumy-Bertrand, Laboratoire de Statistique, Université Louis Pasteur, 7, rue René Descartes 67084 STRASBOURG Cedex (France). E-mail: mmaumy@math.u-strasbg.fr.

identified of U^B (or only of s^B), it is assumed that we can obtain the values of θ_{ji}^{AB} for $j = 1, ..., N^A$. For this reason, it is not necessary to know the values of θ_{ji}^{AB} for all of the matrix of links $\mathbf{\Theta}_{AB}$. Indeed, we only need to know the values of θ_{ji}^{AB} for lines j of $\mathbf{\Theta}_{AB}$, where $j \in s^A$, and for columns i of $\mathbf{\Theta}_{AB}$ where $i \in s^B$.

For example, if the purpose is to estimate a variable of interest Y^B of target population U^B, where

$$Y^B = \sum_{i=1}^{N^B} y_i, \tag{2.1}$$

with y_i measured according to the aggregate U^B. We then use an estimator in the form

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i\, y_i, \tag{2.1}$$

where w_i is the estimated weight of unit i of s^B, with $w_i = 0$ for $i \notin s^B$. To obtain an unbiased estimate of a variable of interest Y^B, we must use as weight w_i the inverse of the probability of selection π_i^B of unit i. As mentioned in Lavallée (1995) and Lavallée (2002), it is generally difficult, if not impossible, to obtain these probabilities. Consequently, we turn to the GWSM, where the weights are given by

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A},$$

where $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. Using this construction, the estimator $\hat{Y}^B$ is unbiased. Similarly, it is possible to calculate and estimate the variance of this estimator because it is the same as that of

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A},$$

with $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB}\, y_i$.

3. Tourism Survey in an Open Environment

3.1 Survey Objectives

The principle of the survey is as follows:

> *"reach tourists (foreigners or French citizens whether or not they live in Brittany) through services aimed at meeting the basic or specific needs"*

such as accommodation, food, leisure activities and transportation.

3.2 Population of Interest

We will let G be a *geographic field* (the four provinces of Brittany) and P be a *reference period* (in this case, it is from February 2005 to December 2005).

A tourist is defined as a person who spent at least one night in G outside his principle residence (tourist-night).

For a tourist, a *trip* is an period sej of P, the length of the cardinal of sej noted as $|sej|$, during which the tourist spends all his nights in G outside his principle residence, the nights immediately before or after the trip sej having been spent outside G (or at the principle residence.).

A tour is a group of tourists (tourist household) sharing the same trip and with the same accommodation during the trip. The term tourist household will also be used through a slight misuse of the terminology (the same tourist household can have several tours over a period, but we have no way to distinguish them).

The statistical unit i of the survey is the tour.

The sub-units of the survey are the trips, tourists and tourist-nights. A tour i consists of n_i tourists during a trip of duration $|sej|$ and thus $n_i \times |sej|$ tourist-nights. Here population U^B is therefore the aggregate of the tours in G during P. ($sej \cap P \neq \emptyset$).

3.3 Survey Sampling Design

To use the GWSM, the theoretical population U^A is formed by a "services" aggregate. In this survey, these services consist of:

- Purchases in bakeries, being the first stratum of U^A.
- Visits to a set of well known cultural, recreational or family sites. In practice, for each of them, a "mandatory pass point" has been defined. It consists of the total number of people passing by this point, which is the second stratum of U^A.
- The number of people exiting Brittany by way of the La Gravelle highway toll, which accounts for 80% of the exits by tourists from Brittany by car. This method of transport itself accounts for 80% of the trips by non-resident of Brittany. People passing this point constitute the third stratum of U^A.

In other words, the *sampling frame* is formally constructed of three strata:

1. purchases in bakeries;
2. visits to a set of sites typical of Brittany;
3. people at the La Gravelle highway toll.

In the *first stratum*, we use a three-stage sample:

- a sample of bakeries;
- a sample of survey days;
- a sample of clients in the bakery on a given day.

In the *second stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through one of the 16 chosen sites on a given day.

Lastly, in the *third stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through the La Gravelle highway toll on a given day.

It is acknowledged that any tourist household consumes at least one of the "services" (bakery purchases, visits to typical Brittany sites, the La Gravelle highway toll)), or at least, that very few households do not consume any of them.

Each sampling (bakery, days, "service") requires specific techniques and it would take considerable time to provide details on each of them. Nevertheless, we will provide the following key technical elements:

- bakeries are sampled using a traditional design stratified geographically (five strata: coastal area of four Brittany departments, the interior of Brittany). In each stratum, the bakeries are sampled with probabilities proportional to their "tourist potential" constructed from their business revenue, the tourist accommodation capacity, and the number of prince-pal residences in the commune to which they belong. This was the theoretical approach, but in practice, the sample was somewhat "forced" by unforeseen circumstances (refusal of bakers, closures during certain period, for example).
- The sites are not sampled, but rather selected for their notoriety and the technical possibility of identifying a "mandatory pass point" (sometimes approximate).
- For each bakery, each site and the La Gravelle highway toll, we defined completely homogeneous "clusters of days" in each period P. A cluster was assigned randomly to each bakery, site and the La Gravelle highway toll. In practice, this means that a full-time enumerator is mobilized for several clusters.
- For each "service", tourists are sampled using the normal techniques of random selection of arrivals: pseudo-systematic sample because, while the enumerator is handing out one questionnaire, other people are going by without being counted. This means that the total number of visitors cannot be estimated directly. If a site is accessible through a ticket booth (museum or chateau, for example), the sampling relies on this means. Ultimately, the sample of users of a "service" on a given day is considered a Bernouilli sample, that is, a simple random sample if we know the size of the population (the number of visitors on a given day).

Comments 3.1. The definition of *tourist* itself is linked to accommodation and it seems natural to use a frame directly related to this service. Practice shows that this is difficult to achieve.

To begin with, there is no correct sampling frame for non-commercial accommodation (relatives, friends, secondary residence) or for seasonal furnished rentals.

In the case of hotels, campgrounds and family holiday homes, the trials runs in summer 2004 revealed the existence of catastrophic bias due to the intervention of hotel owners in the survey selection process. The hoteliers did not respect the random sample instructions and "essentially" distributed the questionnaires to their best clients. This part of the survey had to be set aside and replaced by the count through the La Gravelle highway toll, which is regularly subject to honest quality surveys by various organizations.

The questionnaires collected at the bakeries and at the Brittany tourism sites during summer 2004 apparently produced good qualitative and quantitative results regarding the various modes of accommodation.

Food consumption would undoubtedly have been captured better by questionnaires at the exit of supermarkets, but the problem there lies in the heterogeneity of these establishments and in the cutthroat competition between them; group C … agrees to the surveys in its establishments only if group I … is excluded! In contrast, the collaboration of local bakers in the survey was excellent.

Comments 3.2. By the very definition of the method used, we operate formally within the context of sampling from multiple frames. The problem has given rise to considerable literature (Hartley (1962), Lund (1968) and Hartley (1974) for a start). The GWSM applies to this problem by simply considering each sampling frame as a stratum provided that it is possible to identify for each unit sampled all of frames of which it is a part. This approach provides a rigorous and unique design-based solution to this problem. This comment is worthy of its own article, but the authors know that it is not worth the trouble: an idea that can be explained in ten lines does not need an article or a book for it to survive.

4. Parameters of Interest

Application F, which links to any service j during the reference period P in the three types of establishments of the survey coverage tour i that used this service, is defined as:

$$\begin{array}{lcl} F:\text{services} & \rightarrow & \text{tour} \\ j & \rightarrow & F(j)=i. \end{array}$$

We will let U^B, be the population of tours i of reference period P. This population of interest U^B is the image by F of the aggregate of services during reference period P

in the three types of establishments of the survey coverage. Population U^A is the image by F^{-1} of the aggregate of tours during reference period P. For all $i \in U^B$, we define $R_i(B) = \text{card}(F^{-1}(i))$, the number of antecedents of i during the survey period, that is, the number of services j used by the given tourist household i.

The *parameters of interest* can be totals, sizes or ratios. Let us assume, for example, that we are interested in the estimate of a total relative to a variable y defined on population U^B,

$$Y^B = \sum_{i \in U^B} y_i. \tag{4.1}$$

A specific example of these totals is the size of U^B, written N^B and defined by

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

For example, Y^B can be the number of people who practiced this activity, the total budget spent by the tourist household in Brittany, the geographic origin of the tourist households, or the number of days that the tourist household spends in Brittany. It should be noted that for many variables, the total Y^B depends on the size of the tourist household, that is, the number of people who make up this group and on the length of the trip (only those days spent in Brittany).

Now, we can write

$$Y^B = \sum_{i \in U^B} y_i = \sum_{l=1}^{3} \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j, \tag{4.2}$$

where

$$z_j = \frac{y_i}{R_i(B)}, \quad \text{for} \quad j \in F^{-1}(i),$$

where

- A_1: the aggregate of bakeries in the survey coverage identified by index a_1
- A_2: the 16 visit locations in the survey coverage identified by index a_2
- A_3: the La Gravelle highway toll identified by index a_3
- D_l: the aggregate of survey days, identified by index d_l in an establishment a_l of A_l, for the variant of 1 to 3
- C_{d_l}: the aggregate of services in an establishment a_l of A_l of day d_l of D_l identified by index j.

5. Unbiased Estimates of a Total

In the previous paragraph, we showed that the total of interest is written as a total over the aggregate of the services in the coverage. Let us assume that we have a sample of respondent services j, to which we can link sampling weight δ_j. These weights are assumed to be unbiased because the sample of services follows the canons of a multi-stage sample, each component sample being unbiased.

To make the notations easier to read, we will not show below all stages of the sample draw based on establishment a_l. Let:

- s^B: be the aggregate of tourist household i corresponding to the aggregate of services sampled during the survey period
- s_{A_l}: be the aggregate of sampled establishments
- s_{D_l}: be the aggregate of days sampled in establishment a_l
- s_{d_l}: be the sub-sample of services j corresponding to establishment day a_l.

Since we have a set of sampling weights δ_j for the respondent services, and if we know $R_i(B)$, we can estimate the unbiased total Y^B by

$$\hat{Y}^B = \sum_{i \in s^B} w_i \; y_i \tag{5.1}$$

where

$$w_i = \frac{\sum_{l=1}^{3} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

This gives us an estimate of the population of tourist households. This formula is none other than that given by the GWSM mentioned in section 2. Note that $U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} = \bigcup_{l=1}^{3} U^{A_l}$, $\theta_{ji}^{AB} = 1$ of service j was used by tour i and then $\delta_j = 1/\pi_j^A$.

The variance can be estimated using the same principles (see Lavallée (2002)). We will not go into the details here because it is simply an application of general principles that requires somewhat onerous calculations.

Furthermore, using auxiliary information in the form of totals, whether in populations U^{A_l} or in population U^B, does not pose any particular problems for the point estimation or the estimation of the variance (see Lavallée (2002)).

Comments 5.1. The procedure we have just described for sharing weights may be considered naïve. In fact, we know how to optimize the links matrix $\boldsymbol{\Theta}_{AB}$ as shown in Deville and Lavallée (2006). The application of the Brittany survey is described in Deville, Lavallée and Maumy (2005).

6. An example of a Specific Problem: Visit Points in Open Country

As has already been mentioned, developing the survey of tourism in Brittany required many complementary studies.

We have already mentioned the optimization of weight sharing. Using auxiliary data related to the various frames and to the various stages of the sampling is another task. In this section, we want to focus on estimating some of these auxiliary data, in particular for visits to tourism sites in open country.

In certain cases, we unfortunately do not know the total number of people, denoted as $T_P^{A_2}$, coming to the site on a given day. In effect, in aggregate A_2, we do not know all the services (here the number of visits) of the population. It is therefore not possible to obtain $\pi_j^{A_2}$ directly and therefore δ_j for $j \in A_2$. To overcome this problem, we estimate the number of daily visitors in order to deduct $\tilde{\pi}_j^{A_2} = n_{A_2} / \hat{T}_P^{A_2}$.

Our next step was to develop two approaches to estimating the number of daily visitors for sites accessible by vehicles only (or almost!). The first approach is based on a vehicle sampling system intended to estimate the number of visitors to the site. The second approach uses a sampling of visitors and is aimed at estimating the same quantity by interviewing individuals who give the number of people who travelled with him or her in the vehicle. These two approaches are developed in sections 7 and 8 below.

7. Constructing an Estimator of the Number of Visitors Using a Vehicle Sample

In this paragraph, we examine the approach where an enumerator counts the number of occupants in vehicles that break the line of an electronic eye, or an equivalent system has been set up to count vehicles for which the total number, written as T_V, is known with a virtually negligible measurement error.

7.1 Definition and Variance of $\hat{T}_P^{A_2}$

The total number of vehicles equals

$$T_V = \sum_{\kappa=1,\,\ldots} t_\kappa = \sum_{l \in U_v} 1, \qquad (7.1)$$

where t_κ represents the number of vehicles carrying κ persons and U_V the vehicle universe.

Comments 7.1. To make the notations easier to read, we will use here and until the end this article T_P to denote $T_P^{A_2}$.

The total number of people visiting the site equals

$$T_P = \sum_{\kappa=1,\,\ldots} \kappa t_\kappa = \sum_{k \in U_P} 1, \qquad (7.2)$$

where U_P denotes the universe of people. We also have the equation

$$T_P = \sum_{l \in U_v} v_l, \qquad (7.3)$$

where v_l is the number of people in vehicle l.

As mentioned in the previous section, the total number of people T_P is unknown. Consequently, we must construct an estimator of T_P. If we let $\hat{T}_P$ be $\pi -$estimator based on s_V, a simple random sample of vehicles of size n and with a probability of inclusion n / T_V

$$\hat{T}_P = \frac{T_V}{n} \sum_{l \in s_V} v_l = T_V \, \bar{v}, \qquad (7.4)$$

assuming

$$\bar{v} = \frac{1}{n} \left(\sum_{l \in s_V} v_l \right).$$

It is clear that $\hat{T}_P$ is an unbiased estimator of the total number of people T_P and that $\bar{v}$ is an unbiased estimate of the average number $\bar{V}$ of people in a vehicle.

The variance of $\hat{T}_P$ is therefore equal to

$$\begin{aligned} \operatorname{Var}[\hat{T}_P] &= T_V^2 \left(\frac{1}{n} - \frac{1}{T_V} \right) S_V^2 \\ &= \frac{1}{n} T_V^2 \, S_V^2 - T_V \, S_V^2, \qquad (7.5) \end{aligned}$$

where S_V^2 denotes the corrected variance of population U_V.

7.2 Constructing an Estimator of a Variable of Interest in the Case of a Vehicle Sample

We want to estimate a variable of interest Y of population U_P written as

$$Y = \sum_{k \in U_P} y_k, \qquad (7.6)$$

where y_k is the variable of interest measured in the final questionnaire. Let $\hat{Y}$ be $\pi -$estimator defined by

$$\hat{Y} = \sum_{k \in s_P} w_k^P \, y_k, \qquad (7.7)$$

where weight w_k^P is equal to $\hat{T}_P / m$. Consequently, estimator $\hat{Y}$ can be written

$$\hat{Y} = \frac{\hat{T}_P}{m} \sum_{k \in s_P} y_k = \hat{T}_P \, \bar{y} \qquad (7.8)$$

assuming

$$\bar{y} = \frac{1}{m} \left(\sum_{k \in s_P} y_k \right).$$

Subsequently, variables $\hat{T}_P$ and $\bar{y}$ will be assumed to be independent. The assumption is realistic, because we use two independent enumerators in the field.

7.2.1 Calculation of the Variance of the Estimator $\hat{Y}$

According to Huygens' theorem (1673), conditioning on sample s_V, we get

$$\begin{aligned} V_Y &= \text{Var}[\hat{Y}] \\ &= \bar{Y}^2\,\text{Var}[\hat{T}_P] + T_P^2\,\text{Var}[\bar{y}] \\ &\quad + \text{Var}[\hat{T}_P]\,\text{Var}[\bar{y}]. \end{aligned} \tag{7.9}$$

In the present case, we liken the sample to a simple random sampling without replacement. Equation (7.9) thus becomes

$$\begin{aligned} V_Y &= \bar{Y}^2\left(\frac{1}{n}\,T_V^2\,S_V^2 - T_V\,S_V^2\right) \\ &\quad + T_P^2\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right) \\ &\quad + \left(\frac{1}{n}T_V^2\,S_V^2 - T_V\,S_V^2\right)\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right), \end{aligned}$$

with $S_Y^2 = 1/(T_P - 1)\,\sum_{k \in U_P}(y_k - \bar{Y})^2$. Reorganizing the terms gives

$$\begin{aligned} V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2. \end{aligned}$$

The next step is to determine the allocation of the sample sizes s_P and s_V that minimizes the variance of estimator $\hat{Y}$ for fixed population sizes T_P and T_V.

We must therefore minimize equation (7.10) in n, m subject to

$$C_V\,n + C_P\,m = C,$$

where C_V denotes the cost (in time for example) of the questionnaires related to vehicles, C_P the cost (in time) of the questionnaires related to people, and C the total cost.

The Lagrangian equation can be written as

$$\begin{aligned} L(n,\,m,\,\lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2 \\ &\quad + \lambda(C_V\,n + C_P\,m - C). \end{aligned} \tag{7.11}$$

Taking the partial derivatives with respect to variables n, m, λ and setting them equal to zero gives

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\,m,\,\lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\left(-\frac{1}{n^2}\right) \\ &\quad + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{mn^2}\right) \\ &\quad + \lambda\,C_V = 0, \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial m}(n,\,m,\,\lambda) &= (T_P^2 - T_V\,S_V^2)\,S_Y^2\left(-\frac{1}{m^2}\right) \\ &\quad + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{nm^2}\right) \\ &\quad + \lambda\,C_P = 0, \end{aligned}$$

$$\frac{\partial L}{\partial \lambda}(n,\,m,\,\lambda) = C_V\,n + C_P\,m - C = 0.$$

After calculations, we get a third-degree equation in n that is written

$$\begin{aligned} &\lambda C_V^2\,n^3 - \lambda C_V C n^2 \\ &\quad - C_V T_V^2\,S_V^2\left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) n \\ &\quad + T_V^2\,S_V^2\left(C\left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) + C_P\,S_Y^2\right) = 0. \end{aligned}$$

This third-degree equation in n allows a real solution that can be determined using numeric methods.

Using the same reasoning, we get a third-degree equation in m

$$\begin{aligned} &\lambda C_P^2\,m^3 - \lambda C_P C m^2 \\ &\quad - C_P\,S_Y^2\,(T_P^2 - T_V\,S_V^2)\,m \\ &\quad + S_Y^2\,(C(T_P^2 + T_V\,S_V^2) + C_V\,T_V^2\,S_V^2) = 0. \end{aligned}$$

7.2.2 Simplified Case

To simplify the variance calculation of estimator $\hat{Y}$, we can make an approximation in equation (7.10). In effect, we can assume that term $1/nm$ is negligible before terms $1/n$ and $1/m$.

This then gives us the following transformation of equation (7.10)

$$\begin{aligned} V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + \frac{T_V}{T_P}S_V^2\,S_Y^2 - \bar{Y}^2\,T_V\,S_V^2 \\ &\quad - T_P\,S_Y^2. \end{aligned} \tag{7.12}$$

The next step is determining the allocation of the sample sizes s_P and s_V that minimize the variance of estimator $\hat{Y}$ for fixed population sizes T_P and T_V.

We must therefore minimize equation (7.12) in n, m subject to

$$C_V\, n + C_P\, m = C.$$

The Lagrangian equation can be written as

$$\begin{aligned} L(n,\ m,\ \lambda) = {} & \left(\bar{Y}^2 - \frac{S_Y^2}{T_P} \right) T_V^2\, S_V^2 \frac{1}{n} \\ & + \ (T_P^2 - T_V\, S_V^2)\, S_Y^2 \frac{1}{m} \\ & + \frac{T_V}{T_P} S_V^2\, S_Y^2 - \bar{Y}^2\, T_V\, S_V^2 \\ & - T_P\, S_Y^2 \\ & + \lambda\, (C_V\, n + C_P\, m - C). \end{aligned} \qquad (7.13)$$

Taking the partial derivatives with respect to variables n, m, λ and setting them equal to zero gives

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\ m,\ \lambda) = {} & \left(\bar{Y}^2 - \frac{S_Y^2}{T_P} \right) T_V^2\, S_V^2 \left(-\frac{1}{n^2} \right) \\ & + \lambda\, C_V = 0, \\ \frac{\partial L}{\partial m}(n,\ m,\ \lambda) = {} & (T_P^2 - T_V\, S_V^2)\, S_Y^2 \left(-\frac{1}{m^2} \right) \\ & + \lambda\, C_P = 0, \\ \frac{\partial L}{\partial \lambda}(n,\ m,\ \lambda) = {} & C_V\, n + C_P\, m - C = 0. \end{aligned}$$

After the calculations, we get

$$n_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V \dfrac{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}{T_V^2\, S_V^2\, (T_P\, \bar{Y}^2 - S_Y^2)}} \right)},$$

$$m_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V \dfrac{T_V^2\, S_V^2\, (T_P\, \bar{Y}^2 - S_Y^2)}{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}} \right)}.$$

8. Constructing an Estimator of the Number of Visitors Using a Sampling of Visitors

The previous method can be complicated and costly to use at certain sites. A simpler data collection method involves asking person k the number u_k of passengers in vehicle i that transported him or her. This number u_k is equal here to v_l for vehicle l that transported person k. This method has the further advantage of accurately capturing the number of passengers within the meaning of the survey (are babies counted?).

8.1 Definition of $\hat{\hat{T}}_P$

Let us go back to the following equation

$$T_P = \sum_{l \in U_V} v_l,$$

where v_l denotes the number of passengers in vehicle l. Let us also recall

$$T_P = \sum_{l \in U_P} 1.$$

The average number of passengers in a vehicle $\bar{V}$ can be expressed as

$$\bar{V} = \frac{\sum\limits_{l \in U_V} v_l}{\sum\limits_{l \in U_V} 1} = \frac{\sum\limits_{\kappa = 1,\, \ldots} \kappa t_\kappa}{\sum\limits_{\kappa = 1,\, \ldots} t_\kappa} = \frac{\sum\limits_{\kappa = 1,\, \ldots} m_\kappa}{\sum\limits_{\kappa = 1,\, \ldots} M_\kappa / \kappa}, \qquad (8.1)$$

where t_κ is the number of κ – passenger vehicles and M_κ is the number of people who came in a κ – passenger vehicle.

We can use this last relation to obtain a new version of T_P

$$T_P = T_V\, \bar{V}. \qquad (8.2)$$

Consequently, an estimator of T_P can be written as

$$\hat{\hat{T}}_P = T_V\, \hat{\bar{V}}, \qquad (8.3)$$

where the total number of vehicles T_V is perfectly known. Observing this expression, we see that, in order to know estimator $\hat{\hat{T}}_P$, all that is required is to determine the quantity $\bar{V}$. Let us therefore introduce the following estimator of $\bar{V}$

$$\hat{\bar{V}} = \frac{\sum\limits_{\kappa \in s_P} m_\kappa}{\sum\limits_{\kappa \in s_P} m\kappa / \kappa},$$

where m_κ is the number of people in the sample travelling in a κ passenger vehicle. Estimator $\hat{\bar{V}}$ can also be written as follows:

$$\hat{\bar{V}} = \frac{\sum\limits_{k \in s_P} 1}{\sum\limits_{k \in s_P} 1/u_k}$$

or as

$$\hat{\bar{V}} = \frac{m}{\sum\limits_{k \in s_P} 1/u_k}. \qquad (8.4)$$

The last equation makes it possible to write the following equation

$$\frac{1}{\hat{\bar{V}}} = \frac{1}{m} \sum_{k \in s_P} \frac{1}{u_k}. \qquad (8.5)$$

This new quantity represents the empirical average of $1/u_k$ and $\hat{\bar{V}}$ is the harmonic average of u_k. It is also possible to calculate its variance, which is equal to

$$\operatorname{Var}\left[\frac{1}{\hat{\bar{V}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.6)$$

8.2 Calculating the Variance of Estimator $\hat{\hat{T}}_P$ Without a Vehicle Sample

Now we have to calculate the variance of estimator $\hat{\bar{V}}$ knowing (8.6). To this end, note that we can write

$$\frac{1}{\hat{\bar{V}}} = \frac{1}{\bar{V}\left(\frac{\hat{\bar{V}}}{\bar{V}} - 1 + 1\right)}$$

$$= \frac{1}{\bar{V}} \times \frac{1}{1 + \frac{\hat{\bar{V}} - \bar{V}}{\bar{V}}}$$

$$= \frac{1}{\bar{V}}\left(1 - \frac{\hat{\bar{V}} - \bar{V}}{\bar{V}} + o\left(\frac{\hat{\bar{V}} - \bar{V}}{\bar{V}}\right)\right).$$

Accordingly, this gives

$$\operatorname{Var}\left[\frac{1}{\hat{\bar{V}}}\right] \simeq \left(\frac{1}{\bar{V}}\right)^2 \times \frac{\operatorname{Var}[\hat{\bar{V}}]}{\bar{V}^2}.$$

Lastly, we have

$$\operatorname{Var}[\hat{\bar{V}}] \simeq \bar{V}^4 \times \operatorname{Var}\left[\frac{1}{\hat{\bar{V}}}\right],$$

or, with (8.6)

$$\operatorname{Var}[\hat{\bar{V}}] \simeq \bar{V}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.7)$$

By definition, variance $S^2_{1/u}$ is equal to

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{k \in U_P} \left(\frac{1}{u_k} - \frac{1}{\bar{V}}\right)^2. \qquad (8.8)$$

Since quantity T_P is unknown, this relation can be estimated by

$$\frac{1}{m - 1} \sum_{k \in s_P} \left(\frac{1}{u_k} - \frac{1}{\bar{V}}\right)^2. \qquad (8.9)$$

Given (8.7) and (8.9), we can easily determine the variance of estimator $\hat{\bar{V}}$ and consequently, that of estimator $\hat{\hat{T}}_P$ and lastly, that of the variable of interest $\hat{Y}$.

Comments 8.1. Estimator $\hat{\hat{T}}_P$ is biased and asymptomatically unbiased.

Comment 8.2. If variables $\hat{\hat{T}}_P$ and $\bar{y}$ are not independent then we would have

$$\operatorname{Var}\left[\hat{\hat{T}}_P \, \bar{y}\right] = \bar{Y}^2 \operatorname{Var}\left[\hat{\hat{T}}_P\right] + T_P^2 \operatorname{Var}[\bar{y}]$$

$$+ \operatorname{Var}\left[\hat{\hat{T}}_P \, \bar{y}\right] \operatorname{Var}[\bar{y}]$$

$$+ \text{ terms not linked to the eventual non-independance of the variables } \hat{\hat{T}}_P \text{ and } \bar{y}.$$

9. Numeric Illustration

A mechanical counter at a site in open country gives $T_V = 100$ vehicles. We assume that 20% of the vehicles have one person, 20% have two people, 20% have three people, 20% have four people and 20% have five people. This means there are 300 visitors to the site. The variance S_V^2 is equal to two disregarding finite population corrections. The average number of passengers $\bar{V}$ is three. In effect, we have:

$$\frac{1}{\bar{V}} = \frac{1}{1} \times \frac{20}{300} + \frac{1}{2} \times \frac{40}{300} + \frac{1}{3} \times \frac{60}{300}$$

$$+ \frac{1}{4} \times \frac{80}{300} + \frac{1}{5} \times \frac{100}{300} = \frac{1}{3}.$$

which gives $\bar{V} = 3$.

Let us now calculate an estimate of $S^2_{1/u}$. After simplifications of (8.8) and assuming that T_P is large enough compared to one, we have

$$S^2_{1/u} \simeq \frac{1}{T_P} \sum_{k \in U_P} \frac{1}{u_k^2} - \left(\frac{1}{\bar{V}}\right)^2.$$

Thus, we get

$$S^2_{1/u} = \frac{1}{30}\left(2 + 1 + \frac{2}{3} + \frac{1}{2} + \frac{2}{5}\right) - \frac{1}{3^2}$$

$$= \frac{1}{30}\left(\frac{60 + 30 + 20 + 15 + 12}{30}\right) - \frac{1}{3^2}$$

$$= \frac{137}{30^2} - \frac{1}{3^2} = \frac{37}{30^2}.$$

Since we know $S^2_{1/u}$, we can calculate the variance of estimator $\bar{V}$. This gives

$$\operatorname{Var}[\hat{\bar{V}}] \simeq 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.$$

Lastly, we can calculate the variance of estimator $\hat{\hat{T}}_P$

$$\operatorname{Var}\left[\hat{\hat{T}}_P\right] = T_V^2 \operatorname{Var}\left[\hat{\hat{V}}\right]$$

$$\simeq 10^4 \times 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.$$

The first approach gives a variance of estimator $\hat{T}_P$ equal to

$$\operatorname{Var}\left[\hat{T}_P\right] = 10^4 \times 2 \times \frac{1}{n}.$$

Thus, for estimator $\hat{\hat{T}}_P$ to have the same variance as estimator $\hat{T}_P$, size m of sample s_P must be equal to

$$m \simeq 1.66n.$$

Our initial conclusion is that the second approach makes field operations simpler and less costly in terms of personnel because it only requires one enumerator. It is more accurate than a count that does not involve direct contact to obtain the composition of the tourist household. It requires only one sample about one and a half times larger than the first approach to produce the same accuracy, which is tolerable given the resulting simplification of collection. In practice, at all sites, the second approach will be the preferred application.

Conclusion

This article presented a broad description of a new method applicable to tourism statistics. It involves capturing tourists based on the consumption of certain services on which probabilistic samples are constructed. The weight share method makes it possible to shift from statistical accuracy of the services to the accuracy of the relevant tourism statistical units: the tour, the trip, the tourist household, the tourist or the tourist-night. However, the method requires numerous adaptations and complements to the weight share. We described one of these in detail, which is the estimate of the number of visitors to a site in open country. Two methods were tested. One, which was more accurate in terms of sample size, requires a relatively extensive organization and runs the risk of unacceptable errors in measurement. At the price of collecting slightly more data, the second method is preferred.

Other studies of this nature were conducted before and during the time of the survey so that it is difficult to present the full methodology in a single article.

Acknowledgements

The authors sincerely thank the two reviewers and associate editor who all made a significant contribution to improving the readability of this paper.

References

Deville, J.-C. (1999). Les enquêtes par panel : En quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*, 84-85-86, 63-82.

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

Deville, J.-C., Lavallée, P. and Maumy, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected apllications. *Sankhyā*, Series C, 36, 99-118.

Huygens, C. (1673). Horologium Oscillatorium sive de motu pendulorum.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, 21, 25-32.

Lavallée, P. (2002). Le sondage indirect, ou la méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles, éditions Ellipses, Bruxelles.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 155-169.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 282-288.

Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J.M. and Valdés Pelaáez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Éd. Ákos Probáld). *Proceedings Of The Sixth International Forum On Tourisme Statistics. Hungarian Central Statistical Office, Budapest.*

Valdés, L., De La Ballina, J., Aza, R., Loredo, E., Torres, E., Estébanez, J.M., Domínguez, J.S. and Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. In *Tourism Statistics*: *International Perspectives and Current Issues*, (Ed. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.