STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 1, pp. 185–193, DOI 10.21307/stattrans-2020-051
Received – 15.06.2016; accepted – 06.05.2020

Through a Random Route to the Goal: Theoretical Background and Application of the Method in Tourism Surveying in Poland

Sebastian Wójcik [1]

ABSTRACT

Classic survey methods are ineffective when surveying a small or rare population. Several methods have been developed to address this issue, but often without providing a full mathematical justification. In this paper we propose estimators of parameters relating to Random Route Sampling and explore their basic properties. A formula for the Horvitz-Thompson estimator weights is presented. Finally, a case of a tourism-related survey conducted in Poland is discussed.

Key words: random route, Horvitz-Thompson estimator.

1. Introduction

Nowadays official statistics is looking for cost-effective and time-effective survey methods. It is particularly noticeable when we deal with surveys of small populations such as unemployed, foreigners, homeless, etc. Usually a frame for such a subpopulation is not available. Some methods for solving these problems have been developed, but often without a full theoretical background. The representativeness and unbiasedness of the sample surveyed in that way is a question of concern.

2. Random Route Sampling

We shall present some details on the Random Route Sampling method. Assume that we want to survey a subpopulation S of a population P. The frame of members of P is available, but the frame of members of S is unknown. In this paper we focus on household (or dwelling) population. In the Random Route procedure, interviewers walk from house to house and survey households on a prescribed route that ensures randomness. At the first stage, a group of n households (list of starting points) is sampled. At the second stage, an interviewer sets out from the starting point. If a household is not a member of S, then the interviewer continues walking and surveying. There are two alternative models. In the first one the interviewer is surveying until he/she finds a member of S. In the second model the interviewer continues until he/she finds a member of S or he/she reaches the limit of K steps. Each visited dwelling is called *a step*. The interviewer follows some rules that ensure randomness such as: always on the right, always clockwise or always downstairs.

In this paper we unify two aforementioned models by assuming that:

[1] Institute of Mathematics, University of Rzeszow, Division of Mathematical Statistics, Statistical Office in Rzeszów, Poland. E-mail: s.wojcik@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2425-9626.

- the interviewer makes up to $K = 1, 2, ..., \infty$ steps until he/she finds a member of S. The starting point is the first step. The sequence of up to K steps will be called *a route*;
- the interviewer walks only within his/her district (Primary Statistical Unit);
- there is only one interviewer per district;
- the interviewer does not survey the household that has been already surveyed;
- if in a route interviewer surveys another starting point from the list then he/she counts it as a step. The next household to be surveyed after the route is completed, becomes a new (replaced) starting point.

Clearly, the better survey completeness and the larger size of the subpopulation S in a relation to the size of the population P, the lower number of steps made in a route. Obviously, if $K = 1$ then the Random Route (RR) becomes the Simple Random Sampling (SRS). Thus, SRS can be treated as a special case of RR and all of the results for RR for $K = 1$ should be consistent with results for SRS.

Several papers assess the quality of random route samples. Biasedness and sample representativeness are studied based on case studies (Hoffmeyer-Zlotnik (2003), de Rada, Martin (2014)) or simulations (Bauer (2014), Bauer (2016)).
Hoffmeyer-Zlotnik compared three different models of Random Route Sampling:

1) uncontrolled Random Route Sampling with Kish tables and net number of interviews defined used in German General Social Survey (ALLBUS) in 1992,

2) controlled Random Route Sampling with Kish tables and gross number of addresses defined used in German General Social Survey (ALLBUS) in 1998,

3) Random Route plus quota design with net number of interviews defined used in a national survey of the German Youth Institute.

Hoffmeyer-Zlotnik, based on data analysis, found out that the uncontrolled Random Route Sampling saves 30% of expenses in comparison to the controlled version and to the version with quota. In the extreme, case the version with quota caused the walk to be very long and the interviewer had to contact about 100 households for carrying out 10 interviews. Moreover, with modification of the sampling process in the controlled version or the version with quota, the Random Route Sampling becomes non-probability sampling and the sampling error cannot be calculated.

In this paper we refer to the uncontrolled Random Route Sampling. We propose an estimator of the fraction of a subpopulation S in a population P and prove that this estimator is asymptotically unbiased and consistent. Further, we derive sample weights.

3. Parameter estimation under the Random Route Sampling

Since the Random Route method is focused on surveying members of S the question that arises naturally is how to estimate the size of a population S. We will derive some estimators and check their consistency and unbiasedness. Let us denote by M the size of population - the only known parameter. We introduce two further parameters r and p - the unknown parameters of the level of completeness and the fraction of a subpopulation S in a population P, respectively. We assume that r and p are such that prM and pM are integers.

Now, we divide our analysis into two cases.

3.1. Unlimited number of steps in a single route

Let X be a number of visited $P \backslash S$ members (members of P but not S) until an S member is surveyed. Clearly X is a random variable. Furthermore, for every $k \in \{0, 1, ..., (1-pr)M\}$, X takes the value k provided the following two conditions are satisfied:

- in each of the first k steps of the route the interviewer either visited a member of $P \backslash S$ or he/she did not get an answer due to the incompleteness;
- in the $(k+1)^{th}$ he/she surveyed a member of S.

Therefore,

$$P(X=k) = \frac{\binom{M-k-1}{(1-pr)M-k}}{\binom{M}{(1-pr)M}} \text{ for } k = 0, 1, ..., (1-pr)M. \tag{1}$$

That is X follows the negative hypergeometric distribution $HYP^{-}(M, (1-pr)M, 1)$. In order to survey exactly n members of S the interviewer will make $n + \sum_{i=1}^{n} X_i$ steps, where X_i for $i = 1, .., n$ are i.i.d. random variables with probability distribution described by (1). Guenther (1975) proposed the following maximum-likelihood estimator of pr

$$Y^{\infty}_{HYP^{-}}(n) = \frac{n}{n + \sum_{i=1}^{n} X_i}. \tag{2}$$

The estimator $Y^{\infty}_{HYP^{-}}$ being the maximum-likelihood estimator has a number of attractive limiting properties such as consistency and efficiency (Pfanzagl (1994)). Nevertheless, the estimator given by (2) is biased (Zhang, Johnson (2011)). It is still an open question if this estimator is asymptotically unbiased. Therefore, we are going to modify the model in such a way that the estimator defined by (2) becomes asymptotically unbiased.

Assume that M is relatively large compared to n. According to the result of Johnson and Kotz (1969), if $M \to \infty$ with p and r being fixed, then

$$Y^{\infty}_{HYP^{-}}(n) \to_D Y$$

where Y is a random variable following the negative binomial distribution $BIN^{-}(1-pr, 1)$. Therefore, for relatively large M, we can treat the Random Route Sampling as a sampling with replacement. Note, however, that

$$BIN^{-}(1-pr, 1) = GEO(1-pr)$$

where $GEO(1-pr)$ denotes the geometric distribution with parameter $1-pr$. So, it is reasonable to replace the underlying negative hypergeometric distribution by the geometric one. Then, we get

$$P\left(\sum_i^n X_i = k\right) = \binom{k+n-1}{k}(pr)^n(1-pr)^k \text{ for } k = 0,1,2,\ldots.$$

Furthermore, $\sum_i^n X_i$ being a sum of i.i.d. random variables having the geometric distribution with parameter $1-pr$ (cf. Bandyopadhyay P.S., Forster M.R. (2011)), follows the negative binomial distribution $BIN^-(1-pr,n)$. Hence,

$$E\left(\sum_i^n X_i\right) = \frac{n(1-pr)}{pr} \quad \text{and} \quad D^2\left(\sum_i^n X_i\right) = \frac{n(1-pr)}{(pr)^2}. \tag{3}$$

Define an estimator $Y^{\infty}_{BIN^-}$ as follows:

$$Y^{\infty}_{BIN^-}(n) = \frac{n}{n+\sum_{i=1}^n X_i}. \tag{4}$$

For every $n \in \mathbb{N}$, $Y^{\infty}_{BIN^-}(n)$ expresses the ratio of the number of surveyed S-members in a relation to the number of surveyed P-members in an n-route survey. Note that $Y^{\infty}_{BIN^-}$ is the maximum-likelihood estimator of pr (cf. Hilbe (2011)). Moreover, taking into account (3) and applying the result by Stuart (Stuart (1998), p. 351) we obtain

$$E\left(Y^{\infty}_{BIN^-}(n)\right) = E\left(\frac{n}{\sum_{i=1}^n X_i + n}\right) = \frac{n}{E\left(\sum_{i=1}^n X_i\right)+n} + O\left(\frac{1}{n}\right) =$$

$$\frac{n}{\frac{n(1-pr)}{pr}+n} + O\left(\frac{1}{n}\right) = pr + O\left(\frac{1}{n}\right).$$

Thus,

$$\lim_{n\to\infty} E\left(Y^{\infty}_{BIN^-}(n)\right) = pr, \tag{5}$$

which shows that $Y^{\infty}_{BIN^-}$ is asymptotically unbiased.

3.2. Limited number of steps in a single route

In practice, the most important case of the Random Route is when the number of steps is finite. In this approach, on the one hand surveying is less sensitive to clustering of $P\backslash S$-members, but on the other, we need more starting points to survey the same number of S-members. Moreover, in this setting it is possible that the interviewer will not survey a S-member in his/her route. Thus, the number of surveyed S-members is a random variable, taking two values: 0 and 1. Let us denote it by L^K. Then, assuming that M is relatively large comparing to n and treating the Random Route Sampling as the sampling with replacement, we obtain the following probability distribution of L^K

$$P\left(L^K = l\right) = \begin{cases} 1-(1-pr)^K & \text{for } l = 1, \\ (1-pr)^K & \text{for } l = 0. \end{cases}$$

We shall determine the maximum-likelihood estimator of pr. To this end, assume that $l = (l_1, ..., l_n)$ is a vector of observed data in a K-step route. Then, the log-likelihood function $\mathscr{L}(pr, l, K)$ is given by

$$\mathscr{L}(pr, l, K) = \sum_{i=1}^{n} \ln[(1-(1-pr)^K)l_i + (1-pr)^K(1-l_i)].$$

Derivating $\mathscr{L}(pr, l, K)$ with respect to pr, we obtain

$$\frac{\partial \mathscr{L}(pr, l, K)}{\partial pr} = \sum_{i=1}^{n} \frac{2l_i - 1}{(1-pr)^{1-K} l_i + (1-pr)(1-2l_i)}. \tag{6}$$

Hence, the maximum-likelihood estimator of pr is of the form

$$Z^K(n) = 1 - \left(1 - \frac{\sum_{i=1}^{n} L_i}{n}\right)^{\frac{1}{K}}. \tag{7}$$

Obviously, Z^K is consistent and efficient. It is an open question if Z^K is asymptotically unbiased.

Let X^K be the number of visited households until an S member is surveyed in a K-step route. Then X^K is a random variable taking the values $1, ..., K$. Furthermore, assuming as previously that M is relatively large compared to n and treating the Random Route Sampling as the sampling with replacement, we conclude that X^K has the following probability distribution:

$$P\left(X^K = k\right) = \begin{cases} pr(1-pr)^k & \text{for } k < K, \\ (1-pr)^K & \text{for } k = K. \end{cases} \tag{8}$$

Thus, we have

$$E\left(X^K\right) = \frac{(1-pr)(1-(1-pr)^K)}{pr} \tag{9}$$

and

$$\begin{aligned} D^2\left(X^K\right) = {} & \tfrac{1}{(pr)^2}[(1-pr) - (2K-1)pr(1-pr)^K - (1-pr)^K(1-(1-pr)^K)p^2] + \\ & \tfrac{1}{(pr)^2}[2pr(1-pr)(1-(1-pr)^K)^2 - (1-pr)^{2K}] - \\ & \tfrac{1}{(pr)^2}[2pr(1-pr)(1-(1-pr)^K - K(1-pr)^{K-1} + K(1-pr)^K)]. \end{aligned} \tag{10}$$

Note that

$$\lim_{K\to\infty} E\left(X^K\right) = \frac{1-pr}{pr} \quad \text{and} \quad \lim_{K\to\infty} E\left(L^K\right) = 1.$$

Hence, (9)-(10) are consistent with the case of $K = \infty$.

Now, we are going to determine the maximum-likelihood estimator of pr. Assume that

$x = (x_1, ..., x_n)$ is a vector of observed data in a K-step route. Let $l = (l_1, ..., l_n)$ where

$$l_i = \begin{cases} 0 & \text{whenever } x_i = K, \\ 1 & \text{whenever } x_i < K. \end{cases}$$

The log-likelihood function $\mathscr{L}(pr, x, K)$ is given by

$$\mathscr{L}(pr, x, K) = \sum_{i=1}^{n} \ln[pr(1-pr)^{k_i-1} l_i + (1-pr)^{K-1}(1-l_i)]. \tag{11}$$

Derivating $\mathscr{L}(pr, x, K)$ with respect to pr, we get

$$\frac{\partial \mathscr{L}(pr, x, K)}{\partial pr} =$$

$$\sum_{i=1}^{n} \frac{(l_i-1)(K-1)(1-pr)^{K-2} + l_i((1-pr)^{x_i-1} - pr(x_i-1)(1-pr)^{k_i-2})}{l_i pr(1-pr)^{x_i-1} + (1-l_i)(1-pr)^{K-1}}.$$

Hence, the maximum-likelihood estimator of pr is of the form

$$Y^K(n) = \frac{\sum_{i=1}^{n} L_i^K}{\sum_{i=1}^{n} L_i^K X_i^K + K(n - \sum_{i=1}^{n} L_i^K) + \sum_{i=1}^{n} L_i^K}. \tag{12}$$

Note that for $i = 1, .., n$ we have

$$L_i^K \to_D 1 \text{ with } K \to \infty.$$

Moreover, if X_i follows $GEO(1-pr)$ for $i = 1, .., n$, then taking into account (8), we get

$$X_i^K \to_D X_i \text{ with } K \to \infty.$$

Thus, in view of (12), for every n, we obtain

$$Y^K(n) \to_D Y_{BIN^-}^{\infty}(n) \text{ with } K \to \infty.$$

The estimator Y^K is asymptotically unbiased. In fact, applying the result of Stuart (Stuart (1998), p. 351) and making use of (9)-(10), we obtain

$$E\left(Y^K(n)\right) = E\left(\frac{\sum_{i=1}^{n} L_i^K}{\sum_{i=1}^{n} X_i^K}\right) = \frac{E\left(\sum_{i=1}^{n} L_i^K\right)}{E\left(\sum_{i=1}^{n} X_i^K\right)} + O\left(\frac{1}{n}\right) =$$

$$\frac{n(1-(1-pr)^K)}{n\frac{1-(1-pr)^K}{pr}} + O\left(\frac{1}{n}\right) = pr + O\left(\frac{1}{n}\right).$$

Hence,

$$\lim_{n\to\infty} E\left(Y^K(n)\right) = pr.$$

4. Horvitz-Thompson estimators

In order to use direct estimators we need to know probabilities of inclusion of a household in the sample. In the Simple Random Sampling, these probabilities can be calculated before survey is conducted. In fact, in the Random Route we can derive them after we data from the survey are collected.

By $\pi_{i|S}$ and $\pi_{i|P\setminus S}$ we denote probability of inclusion of i^{th} household into the sample for S-member and $P\setminus S$-member, respectively. Assume that M is the size of population, p is the share of S-members in population, r is the level of completeness and $q = 1 - pr$.

If $K = \infty$ then

$$1-\pi_{i|S} = \frac{prM-1}{M} + \frac{qM}{M}\frac{prM-1}{M-1} + \frac{qM}{M}\frac{qM-1}{M-1}\frac{prM-1}{M-2} + \ldots = \frac{prM-1}{prM},$$

$$1-\pi_{i|P\setminus S} = \frac{prM}{M} + \frac{qM-1}{M}\frac{prM}{M-1} + \frac{qM-1}{M}\frac{qM-2}{M-1}\frac{prM}{M-2} + \ldots = \frac{prM}{prM+1}.$$

Hence,

$$\pi_{i|P\setminus S} = \frac{1}{prM+1} < \frac{1}{prM} = \pi_{i|S}$$

so $\pi_{i|P\setminus S}$ and $\pi_{i|S}$ are not equal and depend on size of S only. However, the difference between these probabilities becomes negligible for relevant size of S, e.g. if $prM \geq 1000$ then we have $\pi_{i|S} - \pi_{i|P\setminus S} \leq 10^{-6}$. In the case $K = \infty$, the parameter pr can be estimated from (4).

Consider the case where K is finite. Then

$$1-\pi_{i|S} = \frac{prM-1}{M} + \frac{qM}{M}\frac{prM-1}{M-1} + \frac{qM}{M}\frac{qM-1}{M-1}\frac{prM-1}{M-2} + \ldots + \frac{qM}{M}\frac{qM-1}{M-1}\times\ldots\times\frac{qM-K+2}{M-K+2}\frac{prM-1}{M-K+1},$$

$$1-\pi_{i|P\setminus S} = \frac{prM}{M} + \frac{qM-1}{M}\frac{prM}{M-1} + \frac{qM-1}{M}\frac{qM-2}{M-1}\frac{prM}{M-2} + \ldots + \frac{qM-1}{M}\frac{qM-2}{M-1}\times\ldots\times\frac{qM-K+1}{M-K+2}\frac{prM}{M-K+1}.$$

In the case of a finite K, the parameter pr can be estimated from (12).

5. The Random Route in practice. Case of tourism survey in Poland.

The survey "Participation of Polish residents in tourism" has been carried out by the Statistical Office in Rzeszów since the first quarter of 2014. The target population are Polish people who travelled abroad (for one day or more) and people who made a domestic trip for at least one night. Taking into account possibly low completeness rate and the fact that the population of travellers is not very big, the Random Route was applied with 8 steps and 18,750 starting points in a population over 13 million of households.

The data base from the third quarter of 2017 was analysed to assess the aforementioned methods of estimating *pr*. We collected information about the step on which in a route the household was surveyed. All of the households on the first step of the route could be treated as data collected in the Simple Random Sampling. Thus, it allowed us to estimate *pr* like a simple fraction. We expected that the level of completeness *r* should be higher in the Simple Random Sampling than in the Random Route Sampling because all of the households in the starting points are informed about the survey from a letter of President of Statistics Poland. Therefore, an estimate of *pr* should be also higher.

The table below presents the precision of formulas described above.

	Fraction formula	Formula (12)	Formula (7)	Formula (4)
Estimate of *pr*	0.104	0.091	0.083	0.081

Clearly, (12) estimate is the closest to *pr* obtained from the fraction formula. It may stem from the observation that (4) is derived under the assumption of infinite number of steps while (7) is not utilizing information on the total number of steps.

Further investigations based on simulations may create a better picture of properties of the estimators given by (4), (7) and (12).

6. Conclusions

As a cost-effective and time-effective survey method, the Random Route may be preferred to the Simple Random Sampling, especially when we deal with small populations. Under some natural assumptions the weights for the Horvitz-Thompson estimator are easy to compute. The Random Route proved its usefulness also in practice.

REFERENCES

HOFFMEYER-ZLOTNIK, J. H. P., (2003). New Sampling Designs and the Quality of Data. *Methodoloski zvezki - Advances in Methodology and Statistics*, 19. Ljubljana: FDV, pp. 205–217.

DE RADA, V. D., MARTIN, V. M., (2014). Random Route and Quota Sampling: Do They Offer Any Advantage over Probably Sampling Methods?, *Open Journal of Statistics*, 4 (5). DOI: 10.4236/ojs.2014.45038.

BAUER, J. J., (2014). Selection Errors of Random Route Samples, *Sociological Methods & Research*, 43 (3), pp. 519–544. DOI: 10.1177/0049124114521150.

BAUER, J. J., (2016). Biases in Random Route Surveys, *Journal of Survey Statistics and Methodology*, 4 (2), pp. 263–287. DOI: 10.1093/jssam/smw012.

PFANZAGL, J., (1994). *Parametric Statistical Theory*, Berlin: Walter de Gruyter.

GUENTHER, W. C., (1975). *The Inverse Hypergeometric - A Useful Model.* Statistica Neerlandica, 29, pp. 129–144.

ZHANG, L., JOHNSON, W. D., (2011). *Approximate Confidence Intervals for a Parameter of the Negative Hypergeometric Distribution* Proceedings of the Survey Research Methods Section, American Statistical Association.

JOHNSON, N. L., KOTZ, S., (1969). *Distributions in statistics, discrete distributions*, Wiley.

BANDYOPADHYAY, P. S., FORSTER, M. R., (2011). *Philosophy of Statistics*, North Holland.

HILBE, J. M., (2011). *Negative binomial regression*, Cambridge University Press.

STUART, A., (1998). *Kendall's Advanced Theory of Statistics*, Wiley.