Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2016-68

Vivid illustration of the idea of the Horvitz-Thompson and other estimators applying the concept of pseudo-populations

Andreas Quatember

August 2016

Vivid illustration of the idea of the Horvitz-Thompson and other estimators applying the concept of pseudo-populations

Andreas Quatember
Department of Applied Statistics, Johannes Kepler University Linz, Austria (Europe)
andreas.quatember@jku.at

Abstract The estimation of population total from sample data is an important task in empirical research. The basic features of the estimation process can be vividly illustrated by the generation of an artificial population called pseudo-population. It substitutes the original population in the estimation process with regard to the variable involved. Experience in teaching the sampling theory shows that this teaching concept has the potential of substantially improving the students' comprehension of the basic concepts of the sampling theory.

Keywords Statistical education, teaching statistics, Horvitz-Thompson estimator; ratio estimator; regression estimator; post-stratification; iterative proportional fitting; pseudo-populations

1 Introduction

The preface of Volume 29A of the "Handbook of Statistics" starts as follows: "Thirty five years ago, the Central Bureau of Statistics in Israel held a big farewell party for the then retiring Prime Minister of Israel, Mrs Golda Meir. In her short thank you speech, the prime minister told the audience: "you are real magicians, you ask 1,000 people what they think, and you know what the whole country thinks". Magicians or not, this is what sample surveys are all about: to learn about the population from (often small) sample, dealing with issues such as how to select the sample, how to process and analyse the data, how to compute the estimates, and face it, since we are not magicians, also how to access the margin of error of the estimates" (Pfeffermann, Rao 2009, p.v).

Classical sampling theory addresses the effect of different sampling strategies consisting of a sample design applied to select the sample units from the finite population, and estimation method, on the efficiency of the estimation of a parameter under study. In this field, the estimation of a total

$$t = \sum_U y_k \,, \tag{1}$$

of a study variable y in a finite population U of size N plays a leading role in the estimation of parameters (Σ_U is abbreviated notation of the sum over all N units of population U). The reasons for its importance are as follows: Firstly, there are many cases where the total of a certain variable y is really the parameter of interest (for instance, the monthly total of private household consumption in all single-person households of a region or the number of unemployed people in the country). Secondly, statistical measures such as variance, covariance, or other moments, are also totals. Thirdly, a function of several totals, such as the harvest yields per hectare or the employment rate, may be the parameter under study.

Experience in teaching the statistical sampling theory shows that students, particularly students with only little or even no knowledge of the probability theory, do not easily understand the idea behind certain estimators of the population total as presented in classical

textbooks such as Cochran (1977), Särndal et al. (1992), and Lohr (2010). Hence, there is a need to illustrate these ideas in a vivid way. Section 2 of this paper suggests such a presentation by the pseudo-population concept. In Section 3, this concept is applied to the well-known Horvitz-Thompson estimator of the population total. Various other estimators of a total are presented under the unified roof of this teaching concept in Section 4.

2 The pseudo-population concept

For the purpose of estimating t by a probability sample s of size n (with $s \subset U$), the observed sample values of the study variable y have to be weighted simply because for the sample total of y, $\sum_s y_k < t$ applies. Hence a point estimator $t.$ of t is given by

$$t. = \sum_s y_k \cdot \rho_k \ . \tag{2}$$

For this purpose, the weights ρ_k determine how many units of U a single sample value y_k has to represent ($k \in s$). In this respect, we speak of the weights as the respondents' burden.

For the purpose of increasing the comprehension of the estimation procedure, the rationale behind (2) can vividly be described as the generation of an artificial or "pseudo-population" $U.^*$ as a set-valued estimator of the original population U with respect to the parameter t (cf. here and in the following: Quatember 2015, p.9ff). To create $U.^*$, each sample value y_k of s delivers exactly ρ_k copies. From this point of view, the weights can be interpreted as the replication factors ρ_k of the process generating a pseudo-population $U.^*$ of size $N.^* = \sum_s \rho_k$. Hence, for the actual generation of $U.^*$, the variable value y_1 of the first element of the sample s is replicated ρ_1 times delivering ρ_1 "clones" of y_1 for the replications $y.^*$ in $U.^*$, value y_2 of the second sample element is replicated ρ_2 times delivering ρ_2 copies of y_2, and so forth.

As a population the pseudo-population $U.^*$ is extraordinary only in the fact that it contains not only whole but also parts of the whole units because, of course, the replication factors ρ_k are non-integers as a rule. Ignoring this fact in the notation, with the replicated variable $y.^*$ consisting of the replications of y, estimator $t.$ of parameter t can be re-written as

$$t. = \sum_{U.^*} y.^*_k \ . \tag{3}$$

Hence, the general estimator $t.$ of the total t of a variable y in the population U is nothing else but the total of the replicated variable $y.^*$ in the pseudo-population $U.^*$. (see Figure 1).

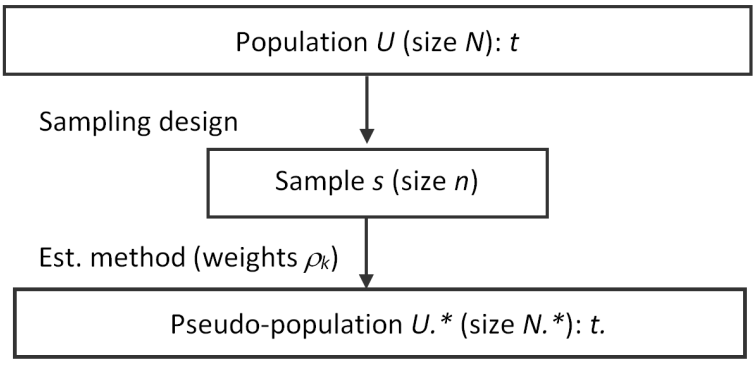

Figure 1: Generating a pseudo-population for the estimator $t.$

Obviously, for given distribution of study variable y, the quality of $t.$ with regard to the estimation of t depends solely on the replication factors used in (2).

3 The Horvitz-Thompson estimator

In the well-known Horvitz-Thompson (HT) approach to the general estimator *t.*, each sample unit has to represent exactly $\rho_k = d_k$ population units:

$$t_{HT} = \sum_s y_k \cdot d_k \qquad (4)$$

Therein, the weights d_k are defined as the reciprocals of the sample inclusion probabilities π_k of the population units (cf. Horvitz and Thompson 1952), which are completely determined by the sample design used to select the sample units. Therefore, the weights d_k are called the design weights of the survey units. Replicating each y_k-value of the sample $\rho_k = d_k = 1/\pi_k$ times (see Figure 1) guarantees that the HT estimator is unbiased for *t*.

Within the pseudo-population concept presented in Section 2 describing the estimation process, this means that the sample value y_1 is replicated d_1 times, value y_2 is replicated d_2 times, and so forth. This yields a specific composition of the pseudo-population U_{HT}^* with respect to variable *y*, which depends solely on the sample design applied. Therefore, the HT estimator is called a design-based estimator and, for a given variable *y*, its accuracy depends solely on the sample design chosen to select the sample *s*. With the replications y_{HT}^* consisting of the $N_{HT}^* = \sum_s d_k$ copies of *y*, estimator (4) can be re-written as

$$t_{HT} = \sum_{U_{HT}^*} y_{HT,k}^* \,. \qquad (5)$$

From the viewpoint of the pseudo-population concept, the HT estimator t_{HT} of the total *t* of a variable *y* in *U* is nothing else but the total of the replicated variable y_{HT}^* in the pseudo-population U_{HT}^*.

The pseudo-population U_{HT}^* with N_{HT}^* elements has an expected size of $E(N_{HT}^*) = N$, the size of the original population *U*. The actual size of U_{HT}^* depends on the concrete design weights of the randomly selected sample units.

For samples selected by the sample design of simple random sampling without replacement (SI), the HT approach (4) to estimate *t* is given by

$$t_{HT(SI)} = \sum_s y_k \cdot \frac{N}{n} \qquad (6)$$

because for the SI sample design $\pi_k = n/N$ applies $\forall\, k \in U$. This means that the HT pseudo-population $U_{HT(SI)}^*$ is created by simply replicating each *y*-value of the SI sample exactly $\rho_k = N/n$ times, which results in a pseudo-population of fixed size $N_{HT(SI)}^* = N$. For this particular sampling scheme, the HT pseudo-population has always the same size as the original population *U*.

The optimum efficiency of the estimator t_{HT} would be ensured by choosing the first-order sample inclusion probabilities π_k proportional to y_k, the "size" of population unit *k* with respect to study variable *y*, which, of course in practically impossible. Assuming $y_k > 0$ and $\pi_k \leq 1\ \forall\, k \in U$ with $\pi_k = \frac{y_k}{t} \cdot n$ for the probability proportional to size (PPS) sample design, the HT estimator is given by

$$t_{HT(PPS)} = \sum_s y_k \cdot \frac{t}{y_k \cdot n} = t\,,$$

which means that $t_{HT(PPS)}$ would perfectly estimate *t*.

One can see that for the PPS sample design, the HT approach generates a pseudo-population $U_{HT(P)}^*$ from the y-values in s with replication factors $\rho_k = \frac{t}{y_k \cdot n}$. The fact that the size $N_{HT(PPS)}^* = \sum_s \frac{t}{y_k \cdot n}$ of $U_{HT(PSP)}^*$ might not be equal to the true size N of U, is irrelevant with regard to the accuracy of $t_{HT(PPS)}$ because each generated pseudo-population $U_{HT(PPS)}^*$ reproduces exactly the interesting total t of y. If $U_{HT(PPS)}^*$ it actually contains less (more) than N elements, then this is compensated perfectly by reciprocally larger (smaller) values of the replications y^* in $U_{HT(PPS)}^*$ compared to the values of y in U. Of course, to be applicable in practice, a positive auxiliary size variable x known for all population units has to serve as a substitute of study variable y in the calculation of the sample inclusion probabilities π_k. Then,

$$t_{HT(P)}^{(x)} = \sum_s x_k \cdot \frac{t}{x_k \cdot n} = \sum_{U_{HT(PPS)}^*} x_{HT(PPS)}^* = t^{(x)}$$

applies with the population total $t^{(x)}$ of variable x and its HT estimator $t_{HT(P)}^{(x)}$. Hence, for PPS sampling with variable x approximately proportional to y, the HT estimator

$$t_{HT(PPS)} = \sum_s y_k \cdot \frac{t}{x_k \cdot n} \tag{7}$$

with replication factors $\rho_k = \frac{t}{x_k \cdot n}$ will always be close to parameter t.

4 Further estimators of a total

After the idea to illustrate the estimation process by the pseudo-population concept was introduced in Section 2 and applied to the HT estimator in Section 3, it may as well be used for the illustration of other estimators of t. For example, the ratio estimator (R) of a total t is defined as

$$t_R = t_{HT} \cdot \frac{t^{(x)}}{t_{HT}^{(x)}}. \tag{8}$$

Therein, the HT estimator t_{HT} is corrected by the ratio of the known population total $t^{(x)}$ of an auxiliary variable x in U and its HT estimator $t_{HT}^{(x)}$ calculated in s (cf., for instance, Särndal et al. 1992, p.181). Hence, also for the ratio estimator of $t^{(x)}$,

$$t_R^{(x)} = t_{HT}^{(x)} \cdot \frac{t^{(x)}}{t_{HT}^{(x)}} = t^{(x)}$$

applies.

Looking at (8), the estimation process can be described again by generating a pseudo-population, in this case denoted by U_R^*. However, compared to the HT estimator, the situation changes in a way that can easily be seen by re-writing (8) in the following way:

$$t_R = \underbrace{\sum_s y_k \cdot d_k}_{t_{HT}} \cdot c \tag{9}$$

with the "correction factor" $c = t^{(x)}/t_{HT}^{(x)}$. Obviously, to generate U_R^*, each sample value y_k has to be replicated not only d_k times but $\rho_k = d_k \cdot c$ times ($k \in s$) (see Figure 1). This means that compared to the HT estimator for the R estimator, the replication factor ρ_k is the product of the replication factor of the HT approach and the constant correction term c. The variable value y_1 of the first sample element is replicated $d_1 \cdot c$ times delivering $d_1 \cdot c$ "clones" of y_1

for the replications y_R^*, value y_2 of the second sample element is replicated $d_2 \cdot c$ times delivering this number of copies of y_2 for y_R^*, and so on. The size of pseudo-population U_R^* is given by $N_R^* = \sum_s d_k \cdot c = N_{HT}^* \cdot c$.

For c being larger than one, the number N_R^* of pseudo-population units will be larger than N_{HT}^*, and vice versa. This underlines in a vivid way the idea of the R estimator. It differs from HT estimation at the replication stage of the generation of the pseudo-population, as shown in Figure 1. If for the auxiliary variable x $t_{HT}^{(x)} > t^{(x)}$ applies, it corrects the HT estimator t_{HT} downwards, and vice versa, by reducing the number of copies ρ_k of each element k of the sample s compared to U_{HT}^* in the generation of the R pseudo-population U_R^*. Obviously, the efficiency of the estimation of t will then be increased if y and x are strongly positively correlated.

Ignoring again in the notation the fact that U_R^* might contain also parts of whole units, in the pseudo-population context, (9) can be re-written by

$$t_R = \sum_{U_R^*} y_{R,k}^* . \tag{10}$$

Hence, the ratio estimator t_R of the total t of variable y in the population U is nothing else but the total of the replicated variable y_R^* in the population U_R^*. With regard to auxiliary variable x,

$$t_R^{(x)} = \sum_{U_R^*} x_{R,k}^* = t^{(x)}$$

applies in U_R^* with the replications x_R^* consisting of the $\rho_k = d_k \cdot c$ replications of value x_k for all sample units k.

For $N_{HT}^* = \sum_s d_k \neq N$, a special ratio estimator $t_{R(N)}$ can be applied by using the true size N of U as the auxiliary information. This yields

$$t_{R(N)} = t_{HT} \cdot \frac{N}{N_{HT}^*} = \underbrace{\sum_s y_k \cdot d_k}_{t_{HT}} \cdot c_N \tag{11}$$

By the correction factor $c_N = N/N_{HT}^*$, the estimator $t_{R(N)}$ adapts the HT estimator t_{HT} to the size N of the original population U. This can be interpreted as the generation a pseudo-population $U_{R(N)}^*$, for which each sample value y_k has to be replicated $\rho_k = d_k \cdot c_N$ times ($k \in s$). Variable value y_1 of the sample selected from the population U is replicated $d_1 \cdot c_N$ times delivering $d_1 \cdot c_N$ "clones" of y_1 for the replications $y_{R(N)}^*$, value y_2 of the second sample element is replicated $d_2 \cdot c_N$ times delivering $d_2 \cdot c_N$ copies of y_2 for $y_{R(N)}^*$, and so forth. This creates a pseudo-population $U_{R(N)}^*$ of correct size $N_{R(N)}^* = N$. For this reason, estimator $t_{R(N)}$, which can also be expressed by

$$t_{R(N)} = \sum_{U_{R(N)}^*} y_{R(N),k}^* ,$$

often performs better than t_{HT} (cf., for instance, Särndal et al. 1992, Sect. 5.7).

Another type of ratio estimation is based on known sizes N_h of H population strata and can be used to increase the efficiency of an estimator of t after the data collection ($h = 1,\ldots,H$). Post-stratification (P) corrects the HT estimators of the y-totals in the H different post-strata simply by applying (11) to each stratum. This results in the P estimator

$$t_P = \sum_{h=1}^{H} t_{R(N),h} = \sum_{h=1}^{H} t_{HT,h} \cdot \frac{N_h}{N_{HT,h}^*} = \cdot \sum_{h=1}^{H} \underbrace{\sum_{s_h} y_k \cdot d_k}_{t_{HT,h}} \cdot c_h \tag{12}$$

with the correction term $c_h = N_h/N_{HT,h}^*$ within the h-th stratum, $t_{HT,h}$, the HT estimator of the total t_h of y in stratum h, and s_h, the part of the sample s that consists of population units belonging to the h-th post-stratum. Furthermore, $\sum_{s_h} d_k = N_{HT,h}^*$ applies.

From the pseudo-population point of view, the estimator t_{PS} corrects the replication factors $\rho_k = d_k$ greating the HT pseudo-population U_{HT}^* in such a way that the resulting estimated population U_{PS}^* will be correctly distributed over the post-strata variable. As a consequence, the size N_{PS}^* of the pseudo-population U_{PS}^* equals the size N of the original population U. This shall correct for incorrect stratum sizes in the HT pseudo-population U_{HT}^* and, for large n, pay off in terms of accuracy when the study variable is related to the stratum variable.

The same idea builds the basis for the iterative proportional fitting approach (IPF) to determine efficiently the replication factors ρ_k in (2) (cf. Deming, Stephan 1940). This estimation method can be applied when there is a more-dimensional variable used for post-stratification of sample s, for which only (one- or more-dimensional) marginal distributions are known. The iterative adjustment of the original HT design weights d_k of the sample units starts with the first post-stratification or "fitting" variable by adapting these weights in the same way as for the P estimator. Consequently, the sum of the adapted design weights of the sample units belonging to each post-stratum h of the first variable will equal the true size of this stratum in population U and, hence, the sum of the corrected weights of the whole sample will equal N ($h = 1,\ldots, H$). In the next step, these newly calculated weights are corrected again with respect to the true stratum sizes of the second post-stratification variable, which, in turn, will destroy the true representation of the first variable. This process is repeated for all fitting variables again and again until the stratum sizes of all these variables deviate from the true sizes in U by no more than a prescribed maximum. These final weights of elements k in s are the IPF weights $d_{IPF,k}$ and the total t of y is estimated by the IPF estimator

$$t_{IPF} = \sum_s y_k \cdot d_{IPF,k} \,. \tag{13}$$

Within the pseudo-population concept, the idea of the IPF estimator can be described in the following way: the process starts with the generation of a HT pseudo-population U_{HT}^* of size $N_{HT}^* = \sum_s d_k$ applying the HT replication factors $\rho_k = d_k$ to all sample units. Then, in the first iteration step, by an adjustment of these replication factors, the composition of the HT pseudo-population is changed in the same way as in the P strategy to equal the true category sizes of the first marginal post-stratification variable used and, consequently, the size of this adapted pseudo-population already equals the true size N of the original population. In the second iteration step, this adapted pseudo-population is again adjusted by a correction of the replication factors, calculated for each $k \in s$ in the first step, with respect to the distribution of the second fitting variable. This, in turn, destroys the correct distribution of the pseudo-population over the first variable. In the next step, the adjustment is done according to the third marginal variable, which again destroys the distribution according to the second variable, and so on. The process is repeated as long as the composition of the adjusted pseudo-population with respect to the marginal distribution of at least one post-stratification variable exceeds a prescribed limit of deviation from the true distribution. If the marginal deviations for all categories of all post-stratification variables fall below this limit, the iterative process is stopped and the current replication weights ρ_k are denoted by $d_{IPF,k}$. When each sample unit k is replicated $d_{IPF,k}$ times, a pseudo-population U_{IPF}^* of size $N_{IPF}^* = N$ with the replication variable y_{IPF}^* is generated (see Figure 1), in which the IPF estimator is calculated by

$$t_{IPF} = \sum_{U^*_{IPF}} y^*_{IPF,k} . \quad (14)$$

In the IPF approach, the estimator t_{IPF} of the total t of variable y in the population U is nothing else but the total of the replicated variable y_{IPF}^* in the population U_{IPF}^*. The pseudo-population U_{IPF}^* corresponds closely to the original population U with respect to the marginal distributions of all post-stratification variables used, which shall have a positive effect on the performance of the estimation of t if y is related to the fitting variables applied.

A last example for the application of the pseudo-population concept to facilitate students' understanding of estimation procedures is the regression (REG) estimator t_{REG}. For a single auxiliary variable x, t_{REG} is given by

$$t_{REG} = t_{HT} + b_1 \cdot (t^{(x)} - t^{(x)}_{HT}) \quad (15)$$

with b_1, the estimated regression coefficient β_1 of the linear regression equation $y = \beta_1 \cdot x + b_2 + \varepsilon$ with the residuals ε (cf., for instance, Särndal et al. 1992, p.230ff). Setting $g \equiv 1 + \frac{b_1}{t_{HT}} \cdot (t^{(x)} - t^{(x)}_{HT})$, expression (15) yields

$$t_{REG} = \sum_s y_k \cdot d_k \cdot g . \quad (16)$$

For the REG estimator, $t_{REG}^{(x)} = t^{(x)}$ applies again.

Trying to describe the estimation method by the picture of generating a pseudo-population, the REG procedure creates a pseudo-population by $\rho_k = d_k \cdot g$ times replicating each sample value y_k (see Figure 1). This yields a pseudo-population U_{REG}^* of size $N^*_{REG} = \sum_s d_k \cdot g = N^*_{HT} \cdot g$. Re-writing (16) from the pseudo-population point of view yields

$$t_{REG} = \sum_{U^*_{REG}} y^*_{REG,k} \quad (17)$$

Again, this means that the estimator, in this case t_{REG}, of the total t of variable y in the population U is nothing else but the total of the replicated variable y_{REG}^* in the population U_{REG}^*. At the same time, for auxiliary variable x with the variable x_{REG}^* of the replicated x-values of s in U_{REG}^*,

$$t^{(x)}_{REG} = \sum_{U^*_{REG}} x^*_{REG,k} = t^{(x)}$$

applies. This shall lead to a more efficient estimation of the total t of y when x and y are related (cf. Särndal et al. 1992, p. ch.6).

5 Conclusion

Experience with the pseudo-population approach in teaching the sampling theory shows that this representation of different sample strategies, consisting of the sample design and estimation method, has the potential of improving students' (or other users') basic understanding of these concepts. One can describe, for instance, the Horvitz-Thompson estimator of the total of a study variable by the generation of a pseudo-population estimating the original finite population with respect to this parameter. For this purpose, the variable values observed in the sample are assigned to the units of the pseudo-population by replicating each of these values by a factor ρ_k that reflects the sample strategy. The Horvitz-Thompson estimator of the total in the original population then is nothing else but the total of the same variable in the pseudo-population consisting of not only whole units, but also parts of whole units because the replications factors are non-integer as a rule. Further concepts of estimators of the total, such as the ratio or regression estimator, can be illustrated in the same way (see Table 1). With a fundamental methodological understanding, students and other users should be able to focus

on questions concerning the difference between methods and their practical implementation is statistical software, for instance.

Estimator	Repl. factor ρ_k
t_{HT}	d_k
$t_{HT(SI)}$	N/n
$t_{HT(P)}$	$t/(x_k \cdot n)$
t_R	$d_k \cdot t^{(x)}/t_{HT}^{(x)}$
$t_{R(N)}$	$d_k \cdot N/N_{HT}^*$
t_{PS}	$d_k \cdot N_h/N_{HT,h}^*$
t_{IPF}	$d_{IPF,k}$
t_{REG}	$d_k \cdot [1+b_1/t_{HT} \cdot (t^{(x)}-t_{HT}^{(x)})]$

Table 1: The replication factors ρ_k (Figure 1) of different estimation procedures with regard of the total t in the pseudo-population approach to the sampling theory

The idea of describing different procedures by the generation of a pseudo-population can also be successfully applied to many other aspects of the sampling theory and survey methodology. For example, the statistical compensation techniques for occurred nonresponse, weighting adjustment and data imputation, can also be presented under the single roof of the pseudo-population concept. But, for this paper, the emphasis was upon an understanding of the basics of the sampling theory.

References

Cochran, W. G. (1977). Sampling techniques. 3rd edition. New York: Wiley and Sons.

Deming, W. E., Stephan, F. F. (1940). On a least squares adjustment of a sampled frequence table when the expected marginal totals are known. *The Annals of Mathematical Statistics*. Vol. 11, p. 427–444.

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, p. 663–685.

Lohr, S. (2010). *Sampling: Design and Analysis*. Boston: Brooks/Cole.

Pfeffermann, D., Rao, C. R. (2009). Preface to Handbook 29A. In: Pfeffermann, D, Rao, C. R. *Handbook of Statistics, Volume 29A. Sample Surveys: Design, Methods and Applications*. Amsterdam: Elsevier.

Quatember, A. (2015). *Pseudo-Populations. A Basic Concept in Statistical Surveys*. Cham: Springer.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.