UNDERSTANDING SAMPLE SURVEY THEORY WITH THE "REPLICATES-DUPLICATES" APPROACH

Pierre Lavallée
Statistics Canada, Canada
Pierre.Lavallee@statcan.gc.ca

In sampling, a sample is selected from a finite population in order to produce some estimates for this population. We want these estimations to be unbiased and the most precise. A good precision corresponds to the situation where different samples produce about the same estimation. In other words, we want the replicates (i.e., the results of the sample selection process) to be as duplicates. The use of auxiliary information (e.g. through a linear regression estimator) also helps in making replicates to be as duplicates, and the concept of superpopulations allows alleviating some emerging conceptual problems. Based on the "replicate-duplicate" approach, we can develop a complete philosophy of teaching sampling theory where, at the start, formulas are left behind to concentrate in the development of the intuitive aspect of sampling theory.

INTRODUCTION

Statistics is and must be of practical significance. Indeed, statistics is seldom practised for its own purposes; instead, it seeks either to summarize (or describe) a given population or to convey something about that population (inference). By a *population*, we mean a set of units (individuals, households, businesses, farms...) that we want to study (e.g., the individuals of a given city). We believe that the teaching of statistics must include concrete situations designed to anchor the theory so that it can be put to a practical use.

From all the possible ways of dividing the realm of statistics, some people choose to set up an opposition between inferential (or classical) statistics and sample survey theory. In many faculties, these two branches of statistics are generally taught separately, without bringing them together in any way. This means that students, after much frustration, resign themselves to seeing a gulf between the two branches.

What we are proposing here is a way to understand sample survey theory using the "replicates-duplicates" approach. We see this as an approach whereby—at least at the outset—sample survey theory can be taught without having to use mathematical formulas. The formulas are introduced at a later stage, after a visualisation of the situation that helps to approach the survey. This serves to get to the formulas with fewer difficulties and better understanding of students.

THE "REPLICATES-DUPLICATES" APPROACH

The *replicates-duplicates* approach described in this paper is drawn from an undergraduate course on sample survey theory that the author took at Carleton University (Ottawa, Canada) in 1985. The professor who taught the course was Dr. Dale, a retired Statistics Canada survey methodologist. Having worked on a number of Statistics Canada surveys, Dr. Dale had great hands-on experience and a pragmatic way of thinking. The author of this paper cannot say whether the replicates-duplicates approach used was developed by Dr. Dale himself, but he believes that the professor had an excellent way of using the approach to teach sample survey theory.

By *sample survey*, we mean the selection of a portion (or part) of the population in order to produce estimates of that population. We generally assume that we have a finite population of size N consisting of variables of interest y_k, x_k, z_k,..., for $k = 1,...,N$, that are unknown constants. For example, we might want to measure the production (y), the revenue (x) and the expenses (z) of the N businesses of a given country. We seek to obtain a "representative" sample of that population.

Most people have an idea, albeit a vague one, of what representativeness is. In formal terms, it can be defined in different ways, and there seems to be no consensus on this matter. Kruskal and Mosteller (1979) cite at least nine different definitions of a representative sample or a representative sample design. In general, we can say that if different samples are drawn and they are all "representative" of the population, the estimates produced from each sample should be similar. The estimates produced then have good *precision*. We therefore want each of the sample's

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

replicates to produce identical estimates, i.e. *duplicates*. In other words, "we want the replicates to be duplicates", a point that Dr. Dale drilled into his students. From this perspective, variability in estimates from different samples is seen as "the failure to obtain replicates that are duplicates."

Relationship between "duplicate replicates" and variability of estimates

Let $\$$ be the set of all possible samples *s* drawn from the population, and let $\hat{Y}_s$ be an estimate of the total $Y = \sum_{k=1}^{N} y_k$ of a variable of interest *y* for that population. We define the expectation of $\hat{Y}_s$ by $\mu = E(\hat{Y}_s) = \sum_{s \in \$} p(s)\hat{Y}_s$, where $p(s)$ is the probability of selecting sample *s*. We have that μ is nothing other than the weighted mean over all possible samples. The variance of $\hat{Y}_s$ is given by $V(\hat{Y}_s) = \sum_{s \in \$} p(s)(\hat{Y}_s - \mu)^2$, i.e., the weighted average of the squared difference from the expectation, over all possible samples.

When replicates are duplicates, each sample *s* yields the same estimate $\hat{Y}_s$, meaning that $\hat{Y}_s = \hat{Y}$ for any *s*. If all samples *s* yield the same estimate $\hat{Y}$, the expectation comes down to $\mu = E(\hat{Y}_s) = \sum_{s \in \$} p(s)\hat{Y} = \hat{Y}$, and the variance gives us $V(\hat{Y}_s) = \sum_{s \in \$} p(s)(\hat{Y} - \hat{Y})^2 = 0$! Thus, there is no variability between the different estimates that we would produce from different samples. We conclude that if the replicates are duplicates, $V(\hat{Y}_s) = 0$ and we obtain maximum precision for the estimate of the total *Y*. Note that there might still be a bias in the estimates (i.e., $E(\hat{Y}_s) - Y \neq 0$), but it will be considered as nil for the present paper. Therefore, assuming no bias, the more replicates are duplicates, the greater the precision of the estimates (variance approaching zero).

Choosing a sample design

The easiest way to draw a sample of size *n* from a population of size *N* is to select units "perfectly randomly", that is, in such a way that all possible samples have the same probability of being selected. This is what is meant by simple random sampling (SRS), which is usually illustrated by selecting a number of marbles from a bag. SRS generally yields estimates with great variability. Fortunately, some techniques exists to reduce this variability, one being stratification, which consists in dividing the population into homogeneous sub-populations.

To discuss variability among the different estimates obtained from different samples, it is important to choose a simple measure of this variability. Assume that we have two samples drawn independently from the same population. Let $\hat{Y}_1$ be the estimate obtained from sample 1, and $\hat{Y}_2$ be the estimate obtained from sample 2. A possible and simple measure of variability is $(\hat{Y}_1 - \hat{Y}_2)^2$.

If the samples are big (i.e., the sample size *n* approaches the population size *N*), we should expect the estimates $\hat{Y}_1$ and $\hat{Y}_2$ to be close to the true total *Y*, which means small variability since $(\hat{Y}_1 - \hat{Y}_2)^2$ then tends toward zero. We then have replicates— $\hat{Y}_1$ and $\hat{Y}_2$ —that tend to be duplicates. On the other hand, the smaller is the sample, the greater the risk of having a bad sample, in the sense that the estimates $\hat{Y}_1$ and $\hat{Y}_2$ obtained may prove to be very different from *Y*, which means greater variability. Note that if the population is very homogeneous, even small samples should produce estimates that are relatively close, and we should then have replicates that tend to be duplicates. Large samples are required when the population is very heterogeneous.

If the replicates are not duplicates, there will be variability among the estimates. Therefore, we want to select the sample in order to reduce variability as much as possible. The way to select the sample is connected to the *sample design* (where SRS is the simplest one). Four main factors influence the choice of a sample design. The first factor is the desired precision of estimates, which is related to the risk of making a mistake and the concept of variability. The greater the need to produce precise estimates, the greater the need to assign special importance to the sample design, which notably includes the sample size. The second factor has to do with the available resources (money, time, human and physical resources), which are important insofar as they make it

necessary, for example, to limit the size of the survey sample. The third factor is related to the characteristics of the population being surveyed. For example, in the case of a very heterogeneous population that would require a very large sample to get reliable estimates, considerable effort must be expended on developing a sample design that stays within collection budgets. Lastly, the fourth factor influencing the choice of a sample design is the data collection method chosen. For example, in a survey involving face-to-face interviews, collection costs are likely to influence the sample size.

In determining a sample design, there are three steps: (i) visualize (assess the situation) and determine the different options; (ii) compare theoretically the different options selected—for example, by determining theoretically the parameters according to which a given sample design will yield the most precise estimates; (iii) use experimental results to choose the best sample design. For this purpose, we generally use results from previous surveys or simulations (Monte-Carlo methods).

Visualization (or assessment of the situation) is a major and essential part the survey methodologist's work. This is the step where an idea begins to take shape on how to approach the survey, considering the four factors listed above. This step does not include, per say, mathematical aspects or formulas; instead, it consists of outlining the population to be surveyed and thinking of possible sample designs. This is the step where the survey methodologist assesses which of the retained sample designs will likely produce replicates that are duplicates.

Several books provide a classical approach to the different sample designs currently used in practice. Among these are the books of Cochran (1977), Särndal, Swensson and Wretman (1992) and Lohr (1999), all in English, and those of Morin (1993), Tillé (2001) and Ardilly (2006) in French. The Statistics Canada publication "Survey Methods and Practices" examines the choice of a sample design based on visualization (Statistics Canada, 2003).

A CLASSICAL EXAMPLE: BOOKS IN A PUBLIC LIBRARY

Dr. Dale's classical example dealt with a survey of books in a public library. In this example, we want to estimate the number of book borrowings during a year. There is therefore a population corresponding to the set of N books in the library. The measured variable of interest is y_k, the number of times book k is borrowed during the year. Typically, a library has two-sided racks containing books. Each side of a rack forms a row, and each rack has shelves. Books are grouped by subject and are arranged conventionally on the shelves.

These days, with computer technology, this example seems out of date, because to estimate the total number of borrowings during the year, all that needs to be done now is to consult the database on borrowings and the result is available in a few seconds. In 1985, not all libraries had been computerized, and each book contained a card on which the date of each borrowing was noted. To determine whether a given book had been borrowed during the year, it was necessary to open the book and look at its card.

To illustrate the replicates-duplicates approach, it is natural to consider SRS. With this sample design, we randomly select a sample of books and look at the number of times each selected book was borrowed during the year. Since the library contains books on different subjects and these subjects vary in popularity, it seems likely that different samples of books will yield different estimates of the total number of borrowing during the year. For example, if by chance the sample contains only heavily borrowed books, there will be an overestimate of the annual number of borrowings, and vice versa. Therefore, here it seems likely that the replicates will not be duplicates. Consequently, there will be variability in the estimates, and to reduce this variability it will be necessary to have a large sample size, which will generate high costs. It is therefore important to consider ways to reduce this variability. One solution is to divide the population into more homogeneous sub-groups, or strata.

INTRODUCING FORMULAS INTO THE TEACHING APPROACH

After several weeks of discussing sample designs and assessing the situation (visualization), still applying the reasoning of the replicates-duplicates approach with the only help

of the very simple formula $(\hat{Y}_1 - \hat{Y}_2)^2$, we can begin to introduce more complicated mathematical formulas into the teaching of sample survey theory.

Consider the example of SRS. We randomly draw (without replacement) a sample *s* of *n* books from the *N* books of the library. Let $\pi_k = n/N$ be the probability that *s* contains book *k*. To estimate *Y*, the total number of borrowings from the library, we propose to use the estimator of Horvitz-Thompson (1952): $\hat{Y} = \sum_{k=1}^{n} y_k / \pi_k = N \sum_{k=1}^{n} y_k / n = N\bar{y}$. Generally, students tend naturally to accept the choice of this estimator. In teaching sample survey theory, we then show that the variance of this estimator is given by $V(\hat{Y}) = \sum_{s \in \$} p(s)(\hat{Y}_s - E(\hat{Y}))^2 = N^2(1-n/N)S^2/n$, where $S^2 = \sum_{k=1}^{N}(y_k - \bar{Y})^2/(N-1)$ and $\bar{Y} = \sum_{k=1}^{N} y_k / N$. To prove this, we can consult, among others, Särndal, Swensson and Wretman (1992). The quantity S^2 is in fact a measure of the heterogeneity of the population.

The more heterogeneous is the population, the larger is S^2. Thus, for a large S^2, we must also have a large sample size *n* in order to maintain a given level of precision in the estimates. Finally, the closer *n* gets to *N*, the more the variance $Var(\hat{Y})$ approaches zero. These are the same observations as those made with the replicates-duplicates approach, but they are now made using formulas. The formulas therefore take on a meaning, which makes it easier for the students to assimilate them.

INCORPORATING AUXILIARY INFORMATION

As before, we want to estimate the number of book borrowings over the course of the year. For simplicity, assume for now that an SRS is performed.

We now assume that we have an auxiliary variable x_k, which is the year of publication of book *k*, and we assume that this variable is available for all the *N* books in the library. Accordingly, we assume that we have a record of all books in the library and their year of publication. It seems plausible that the older a book, the less chance it has of being borrowed. According to this hypothesis, the variable of interest y_k (number of times book *k* is borrowed during the year) should be positively correlated with the variable x_k (year of publication of book *k*).

Regression estimator

Linear regression is well known and is much used in classical statistics. A regression estimator, based on linear regression, can be used to produce estimates in a sample survey context. On this subject, we can consult Cochran (1977) or Särndal, Swensson and Wretman (1992).

To develop the regression estimator, we assume that the number of times book *k* is borrowed during the year is linearly related to the year of publication of book *k*. That is, we assume $y_k \approx A + B x_k$ for $k = 1,...,N$. Thus, we can write

$$\frac{Y}{n\bar{y}} = \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{n} y_k} \approx \frac{NA + B\sum_{k=1}^{N} x_k}{nA + B\sum_{k=1}^{n} x_k} = \frac{NA + BX}{nA + Bn\bar{x}} \qquad (1)$$

Through cross-multiplication, we obtain the regression estimator:

$$\hat{Y}_{rég} = \frac{NA + BX}{nA + Bn\bar{x}} n\bar{y}$$

Estimates of *A* and *B* may be obtained by using the least squares method, which is to minimize $\phi_s = \sum_{k=1}^{n}(y_k - A - Bx_k)^2$. We get $\hat{B} = \sum_{i=1}^{n}(y_k - \bar{y})(x_k - \bar{x}) / \sum_{i=1}^{n}(x_k - \bar{x})^2$ and $\hat{A} = \bar{y} - \hat{B}\bar{x}$. Replacing these estimates in (1), we then obtain the estimator $\hat{Y}_{rég} = N\bar{y} + \hat{B}(X - N\bar{x})$. Note that if the relationship $y_k = A + B x_k$ holds perfectly, we have $\hat{Y}_{rég} = Y$. Thus, the more the variable *y* is linearly related to *x*, the closer the estimate $\hat{Y}_{rég}$ produced from the drawn sample will be to the true total *Y*. Thus, in this case, the replicates will be almost duplicates.

The development of the regression estimator generally raises several legitimate questions from the students: How is this related to the classical theory of linear regression where $y_k = \alpha + \beta x_k + \varepsilon_k$ and $\varepsilon_1,...,\varepsilon_N \sim i.i.d.\, N(0,\sigma^2)$? Why not minimize $\phi = \sum_{k=1}^{N}(y_k - A - Bx_k)^2$ instead of $\phi_s = \sum_{k=1}^{n}(y_k - A - Bx_k)^2$? In other words, why not minimize the sum of the squares over the entire population, instead of restricting ourselves to the sum over the sample? If we need to minimize ϕ_s instead of ϕ, why not assume that we have exactly $Y = A + B\,X$? In other words, why not assume that the total *Y* is exactly linearly related to the total *X*? How can one obtain the regression estimator when the design is more complex than SRS? If the selection probabilities π_k are not the same for all units *k*, how can we consider this? To answer these questions so that the students will have a good understanding, we believe that it is necessary to avoid developing the regression estimator using relationship (1). A way to answer these questions is to use the concept of *superpopulations*.

Superpopulations

The superpopulation is the stochastic process that generates the population. It is often considered as unknown or poorly understood. For example, borrowing books from the library can be seen as a random process. The book borrowing process is seen as a complex mechanism affected by factors, such as the choice of books purchased and the tastes of readers. This complex process constitutes the superpopulation.

In sample survey theory, a sample *s* is drawn from a finite population consisting of unknown constants $y_1,...,y_N$ in order to make an inference on this population. For example, we are interested in obtaining an *estimate* of the *total* number of borrowings of books from the library. The superpopulation, on the other hand, generally corresponds to an infinite and uncountable statistical distribution found in classical statistics. For example, it can be assumed that the superpopulation has the form of the linear model $y = A + Bx + \varepsilon$ where $\varepsilon \sim N(0,\sigma^2)$. The population is then seen as a realisation—that is, a *sample*—of size *N* from the superpopulation. Thus, in this case, we have a population of *N* units where each unit *k* has the form $y_k = A + Bx_k + \varepsilon_k$ where $\varepsilon_1,...,\varepsilon_N \sim i.i.d.\, N(0,\sigma^2)$. Note that this type of sample corresponds to what is generally found in classical statistics.

If we have no measure of *y* for all the *N* units of the population, we must then use the sample to make an inference on the population, which we can then use to make an inference on the superpopulation. The superpopulation concept helps us to understand this inference process, which consists of making an inference firstly from the sample to the population, and then from the population to the superpopulation. To learn more about this subject, see Särndal, Swensson and Wretman (1992).

Using superpopulations in developing the regression estimator

To develop the regression estimator, we assume the superpopulation model $y = A + Bx + \varepsilon$ where $\varepsilon \sim N(0,\sigma^2)$. In our public library example, it corresponds to the "probabilistic process of borrowing books over the course of the year" where the auxiliary variable *x* is, once again, the year of publication of the book. From the superpopulation, we "draw" a population of *N* books where $y_k \approx A + B\,x_k$, $k = 1,...,N$.

To construct the regression estimator, we now start with the identity

$$Y = \sum_{k=1}^{N}\hat{y}_k + \sum_{k=1}^{N}(y_k - \hat{y}_k) \qquad (2)$$

where $\hat{y}_k$ is the predicted value of the y_k obtained using the adjusted superpopulation model. In our case, we have $\hat{y}_k = \hat{A} + \hat{B}x_k$. Since we have the auxiliary variable *x* for the entire population, we can calculate $\hat{y}_k$ for all the *N* units of the population. However, we have y_k only for the *n* units of the sample. We must therefore settle for estimating the second term of (2), which can be done using the Horvitz-Thompson estimator. The resulting regression estimator is

$$\tilde{Y}_{rég} = \sum_{k=1}^{N} \hat{y}_k + \sum_{k=1}^{n} \frac{(y_k - \hat{y}_k)}{\pi_k} \tag{3}$$

To obtain the predicted values $\hat{y}_k$, we can estimate the parameters *A* and *B* using the least squares method. According to classical statistics, to estimate *A* and *B*, we must minimize the sum $\phi = \sum_{k=1}^{N} (y_k - A - Bx_k)^2$. Since we have only the sample *s* and not the entire population, we cannot calculate this minimization in practice. Instead, we minimize a Horvitz-Thompson estimate of that quantity, namely $\hat{\phi} = \sum_{k=1}^{n} (y_k - A - Bx_k)^2 / \pi_k$, which correspond to estimating *A* and *B* using the weighted least squares method (Fuller, 1975). The solution of this minimization problem gives us

$$\hat{B} = \frac{\sum_{i=1}^{n} (y_k - \hat{\bar{Y}})(x_k - \hat{\bar{X}}) / \pi_k}{\sum_{i=1}^{n} (x_k - \hat{\bar{X}})^2 / \pi_k} \quad \text{and} \quad \hat{A} = \hat{\bar{Y}} - \hat{B}\hat{\bar{X}} \tag{4}$$

where $\hat{\bar{Y}} = \hat{Y} / \hat{N}$, $\hat{\bar{X}} = \hat{X} / \hat{N}$, $\hat{Y} = \sum_{k=1}^{n} y_k / \pi_k$, $\hat{X} = \sum_{k=1}^{n} x_k / \pi_k$ and $\hat{N} = \sum_{k=1}^{n} 1 / \pi_k$. Using (3) and (4), we then obtain the regression estimator:

$$\tilde{Y}_{rég} = \sum_{k=1}^{N} \hat{y}_k + \sum_{k=1}^{n} \frac{(y_k - \hat{y}_k)}{\pi_k} = \sum_{k=1}^{N} (\hat{A} + \hat{B}x_k) + 0 = N\hat{\bar{Y}} + \hat{B}(X - N\hat{\bar{X}})$$

The estimator $\tilde{Y}_{rég}$ is a generalization of $\hat{Y}_{rég}$ described in Section 5.1. If the relationship $y_k = A + B\,x_k$ holds exactly, we have $\tilde{Y}_{rég} = Y$. Thus, the better the model, the closer the estimate $\tilde{Y}_{rég,s}$ produced from the given sample *s* will be close to the true total *Y*. Therefore, the replicates — that is, the different estimates $\hat{Y}_{rég,s}$ obtained for different samples s — will be duplicates.

As well, it can be shown that in the context of an SRS, we have $V(\tilde{Y}_{rég,s}) \approx N^2(1 - n/N)S_e^2 / n$ where $S_e^2 = \sum_{k=1}^{N} (y_k - \hat{y}_k)^2 / (N-1)$. The more the predicted value $\hat{y}_k$ is close to y_k (which is the case when the relationship $y_k = A + B\,x_k$ holds exactly), the smaller is S_e^2, and better is the precision. The variance $V(\tilde{Y}_{rég,s})$ is based on the sum over all the possible samples. While it contributes to the determination of $\hat{Y}_{rég}$, the superpopulation model as such does not enter into the variance calculation.

REFERENCES

Ardilly, P. (2006). *Les techniques de sondage*. 2e édition. Paris: Éditions Technip.

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

Fuller, W. A. (1975). Regression Analysis for Sample Survey. *Sankhya: The Indian Journal of Statistics Series C, 37,* pp. 117-132.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association, 47,* 663-685.

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

Kruskal, W., & Mosteller, F. (1979). Representative Sampling III : the Current Statistical Literature. *International Statistical Review, Vol. 47,* pp. 245-265.

Lohr, S. (1999). *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove, CA.

Morin, H. (1993). *Théorie de l'échantillonnage*, Presses de l'Université Laval, Sainte-Foy, Québec.

Statistics Canada (2003). *Survey Methods and Practices*. Publication 12-587-XPE. Ottawa: Minister of Industry.

Tillé, Y. (2001). *Théorie des sondages – Échantillonnage et estimation en populations finies*. Paris: Dunod.