

Estimation of a Proportion with Survey Data

Pierre Duchesne
Université de Montréal

Journal of Statistics Education Volume 11, Number 3 (2003),
<http://www.amstat.org/publications/jse/v11n3/duchesne.pdf>

Copyright © 2003 by Pierre Duchesne, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: Auxiliary information; Bernoulli sampling; Confidence interval; Logistic regression estimator; Sampling plan; Survey sampling.

Abstract

The estimation of proportions is a subject which cannot be circumvented in a first survey sampling course. Estimating the proportion of voters in favour of a political party, based on a political opinion survey, is just one concrete example of this procedure. However, another important issue in survey sampling concerns the proper use of auxiliary information, which typically comes from external sources, such as administrative records or past surveys. Very often, an efficient insertion of the auxiliary information available will improve the precision of the estimations of the mean or the total when a regression estimator is used. Conceptually, it is difficult to justify using a regression estimator for estimating proportions. A student might want to know how the estimation of proportions can be improved when auxiliary information is available. In this article, I present estimators for a proportion which use the logistic regression estimator. Based on logistic models, this estimator efficiently facilitates a good modelling of survey data. The paper's second objective is to estimate a proportion using various sampling plans (such as a Bernoulli sampling and stratified designs). In survey sampling, each sample possesses its own probability and for a given unit, the inclusion probability denotes the probability that the sample will contain that particular unit. Bernoulli sampling may have an important pedagogical value, because students often have trouble with the concept of the inclusion probability. Stratified sampling plans may provide more insight and more precision. Some empirical results derived from applying four sampling plans to a real data base show that estimators of proportions may be made more efficient by the proper use of auxiliary information and that choosing a more satisfactory model may give additional precision. The paper also contains computer code written in S-Plus and a number of exercises.

1. Introduction

In the analysis of a survey, the response variables encountered are often discrete. This would be the case for public opinion research, marketing research, and government survey research. Take, for example, estimation of the employment status: This would require the introduction of an indicator variable showing a value of one if the unit is employed and zero if not. Another example is the estimation of the proportion of voters in favour of a presidential candidate. In an introductory survey sampling course, the estimation of proportions is usually discussed from the perspective of various sampling plans (Kish 1965; Cochran 1977, amongst others). Later in the course, ratio and regression estimators are introduced. These estimators rely on auxiliary information that may come from a past census or from other administrative sources. At this point, a curious student might ask: “Why not use the auxiliary information to improve the estimation of a proportion?” In that case, ratio and regression estimators could be proposed but, since these estimators are fully justified for continuous variables, they would be rather hard to motivate. They are not a good choice for a variable which is discrete and typically consists of a sequence of ones and zeros.

In line with the paper’s first objective, we present the logistic regression estimator proposed by [Lehtonen and Veijanen \(1998a\)](#) and which they call the LREG estimator; it may be used to estimate a proportion when auxiliary information is made available. The LREG estimator is based on a logistic model which describes the joint distribution of class indicators. Logistic models are sometimes introduced at the undergraduate level (see, for example, [Moore and McCabe 1999](#)) and at the advanced undergraduate level (see [Neter, Kutner, Nachtsheim, and Wasserman 1996](#)). In survey sampling texts, they are discussed in [Lohr \(1999\)](#) and in [Särndal, Swensson, and Wretman \(1992\)](#). We shall see that the discussion of logistic models allows the teacher to focus on the modelling of survey data. The idea of introducing a modelling approach in a survey sampling course is advocated in an edited version of a panel discussion on the teaching of survey sampling (see [Fecso, Kalsbeek, Lohr, Scheaffer, Scheuren, and Stasny 1996](#)).

A variety of sampling plans such as simple random sampling, stratified sampling, and cluster sampling are generally introduced in a first survey sampling course. This article’s secondary objective is to discuss the estimation of proportions using Bernoulli (BE) sampling and stratified designs. Many students find it hard to understand the concept of inclusion probability. The BE sampling plan may help them see what inclusion probabilities are all about. It is very easy to implement, and it may cast greater light on the random part of the sampling experiment. In conjunction with the usual, simple random-sampling plan without replacement (SRS) and BE sampling, we shall also consider stratified designs. Stratified sampling plans may be useful when the analyst needs separate estimations for different groups in the population. An efficient stratification variable may also be of help in obtaining more accurate estimations, since many unrepresentative samples can be eliminated.

Monte Carlo experiments may serve as empirical illustrations of several statistical concepts, such as the bias and variance of the estimators or the coverage properties of the confidence intervals. They are particularly useful when it is voluminous to enumerate all the samples in a moderate or large-sized population. Simulations with four sampling plans were carried out. The population under consideration in the empirical study was the 2000 *Academic Performance Index* (API) Base data file. These data contain performance scores and ethnic and socio-economic information for the schools in the State of California, USA. The data file in question may be useful for academic purposes, as it is publicly available and contain many variables. In our application, a natural

stratification variable was school type (elementary, high, middle or small). We show that stratified sampling plans may give a more insightful analysis since they allow us to obtain a separate estimation for each school type. Furthermore, the stratification variable helped to reduce the variance of the logistic regression estimator. Our analysis shows that incorporating auxiliary information into a suitable model may substantially enhance the efficiency of estimating proportions, demonstrating that the appropriate modelling of survey data may result in more suitable procedures.

2. Estimators of a Proportion Under Different Sampling Plans

Let $U=\{1,2,\dots,N\}$ be a finite population. A sample $s \subset U$ is obtained using a sampling design $p(\cdot)$. We denote $\pi_k = \Pr(s \ni k)$ the first order inclusion probability of a given unit k . The symbol “ \ni ” should be read “contains”, since after the sampling plan has been executed, the random sample s may or may not contain unit k . For units k and l , we let $\pi_{kl} = \Pr(s \ni k, l)$ be the second order inclusion probability. We consider the estimation of the class frequencies of a discrete random variable Y with possible values $\{0, 1\}$, that is we want to estimate the population proportion of ones using the random sample s . We denote y_k the realization of the variable Y for the unit k and the quantity of interest is noted $P = N^{-1}T_y$, where $T_y = \sum_U y_k$ represents the total number of ones in the population (In general, if A is any set of units, $A \subseteq U$, then $\sum_A y_k$ will be our shorthand for the quantity $\sum_{k \in A} y_k$). Examples include unemployment rate ($y_k = 1$ if k employed, $y_k = 0$ if not), the proportion of voters in favour of a presidential candidate and so on.

2.1 Simple random sampling without replacement and Bernoulli sampling

Several sampling plans are possible. The more commonly used is perhaps the SRS, where each sample of a given size n_s has the same probability, giving an inclusion probability equal to $\pi_k = n_s/N$. Several statistical packages contain a macro or a function for obtaining an SRS sample. Other sampling plans are much more difficult to find. Students sometimes have trouble interpreting the inclusion probability n_s/N . The reason is the following: students in their statistics course too often encounter the common premise “*Let X_1, X_2, \dots, X_n identically and independently distributed (iid) with mean μ and variance σ^2 .*” Usually, the X_i ’s are the random variables. However, in survey sampling, each sample s possesses its own probability, given by $p(s)$. The value of the variable of interest for the sampling unit k could be given by the numerical value X_k and the random element would be whether or not unit k is included in the sample. A design which is simple to implement could help the student see what is random and what is not random in the sample experiment. The BE sampling plan serves well this purpose. That sampling plan is discussed in [Särndal, et al. \(1992\)](#). To implement the plan, it suffices to proceed in the following manner:

Step 1. Let n be the *expected* sample size.

Step 2. Generate N variables *independently* from a uniform distribution $U[0, 1]$. Denote the values obtained as u_1, u_2, \dots, u_N .

Step 3. If $u_k < n/N$, choose unit k . If not, do not include k in the sample.

Step 4. Repeat step 3 for each unit in the population.

An illustration using a real dataset of size $N = 30$ is given in the following example. Note that much larger real population could be used in class without additional complications (using conventional slides or PowerPoint software for example), adding more realism to the presentation.

Example 1

Royal Lepage is a Canadian company that provides real estate services. They produce annually a survey of Canadian housing prices. In that survey, several specific categories of housing are surveyed. For example, for Greater Montreal (in the province of Quebec), the housing values of executive, detached two-storey houses for July 2002 are described in [Table 1](#). The prices are in Canadian dollars (CAN\$).

Table 1. Values of the executive, detached two-story houses for Greater Montreal in July 2002.

k	City	Price
1	Ahuntsic	229000
2	Beaconsfield	275000
3	Beloeil	152000
4	Blainville	314000
5	Boucherville	205000
6	Chomedey	212000
7	Cote-St-Luc	475000
8	Dorval	157000
9	Duvernay	243000
10	Fabreville	169800
11	Hudson	245000
12	Kirkland	198000
13	Lachine	189000
14	Lasalle	175000
15	Lorraine	309000
16	Montreal West	360000
17	Mount Royal	370000
18	Notre-Dame-De-Grace	375000
19	Outremont	600000
20	Pierrefonds	138000
21	Pointe Claire	290000
22	Rosemere	338000
23	St-Bruno-De-Montarville	235000
24	St-Eustache	250000
25	St-Lambert	300000
26	St-Laurent	250000

27	Ste-Therese	255000
28	Terrebonne	165000
29	Vimont	259000
30	Westmount	758000

For example, an executive, detached two-storey house in Dorval would be worth 157,000 CAN\$. However, the same house located in Westmount would cost 758,000 CAN\$. This kind of database allows us to compare real estate prices according to location. Suppose that we draw a sample from that population using BE sampling. In the step 1, we set the expected sample size $n = 14$, which gives an expected sampling fraction equal to $n/N = 14/30 = 46.6\%$. For step 2 in obtaining a BE sample, we generate uniform random variables using the S-Plus function `runif()`. To illustrate our discussion, three samples are chosen from that population.

```

> set.seed(1) # Fix the seed
> round(runif(30), digits=3) # Commands for the first sample
[1] 0.163 0.425 0.317 0.646 0.084 0.083 0.203 0.978 0.439 0.272 0.968 0.788
[13] 0.021 0.908 0.904 0.559 0.373 0.798 0.385 0.818 0.525 0.857 0.492 0.348
[25] 0.117 0.216 0.572 0.807 0.859 0.955
> round(runif(30), digits=3) # Commands for the second sample
[1] 0.913 0.922 0.863 0.210 0.548 0.472 0.772 0.068 0.052 0.384 0.613 0.404
[13] 0.224 0.151 0.560 0.061 0.099 0.937 0.270 0.620 0.275 0.411 0.617 0.570
[25] 0.001 0.586 0.323 0.326 0.335 0.465
> round(runif(30), digits=3) # Commands for the third sample
[1] 0.273 0.998 0.056 0.037 0.127 0.032 0.287 0.968 0.003 0.866 0.160 0.353
[13] 0.398 0.703 0.951 0.375 0.220 0.090 0.328 0.512 0.710 0.170 0.437 0.376
[25] 0.984 0.676 0.660 0.355 0.127 0.339

```

In [Table 2](#), the columns labelled “ u_k ” give the realizations of the uniform random variables for each unit in the population and the additional columns labelled “Included?” indicate whether or not unit k is included in the sample.

Table 2. Three BE samples for the housing data.

k	City	Price	u_k	Included?	u_k	Included?	u_k	Included?
1	Ahuntsic	229000	0.163	Yes	0.913	No	0.273	Yes
2	Beaconsfield	275000	0.425	Yes	0.922	No	0.998	No
3	Beloeil	152000	0.317	Yes	0.863	No	0.056	Yes
4	Blainville	314000	0.646	No	0.210	Yes	0.037	Yes
5	Boucherville	205000	0.084	Yes	0.548	No	0.127	Yes
6	Chomedey	212000	0.083	Yes	0.472	No	0.032	Yes
7	Cote-St-Luc	475000	0.203	Yes	0.772	No	0.287	Yes
8	Dorval	157000	0.978	No	0.068	Yes	0.968	No
9	Duvernay	243000	0.439	Yes	0.052	Yes	0.003	Yes
10	Fabreville	169800	0.272	Yes	0.384	Yes	0.866	No
11	Hudson	245000	0.968	No	0.613	No	0.160	Yes
12	Kirkland	198000	0.788	No	0.404	Yes	0.353	Yes
13	Lachine	189000	0.021	Yes	0.224	Yes	0.398	Yes

14	Lasalle	175000	0.908	No	0.151	Yes	0.703	No
15	Lorraine	309000	0.904	No	0.560	No	0.951	No
16	Montreal West	360000	0.559	No	0.061	Yes	0.375	Yes
17	Mount Royal	370000	0.373	Yes	0.099	Yes	0.220	Yes
18	Notre-Dame-De-Grace	375000	0.798	No	0.937	No	0.090	Yes
19	Outremont	600000	0.385	Yes	0.270	Yes	0.328	Yes
20	Pierrefonds	138000	0.818	No	0.620	No	0.512	No
21	Pointe Claire	290000	0.525	No	0.275	Yes	0.710	No
22	Rosemere	338000	0.857	No	0.411	Yes	0.170	Yes
23	St-Bruno-De-Montarville	235000	0.492	No	0.617	No	0.437	Yes
24	St-Eustache	250000	0.348	Yes	0.570	No	0.376	Yes
25	St-Lambert	300000	0.117	Yes	0.001	Yes	0.984	No
26	St-Laurent	250000	0.216	Yes	0.586	No	0.676	No
27	Ste-Therese	255000	0.572	No	0.323	Yes	0.660	No
28	Terrebonne	165000	0.807	No	0.326	Yes	0.355	Yes
29	Vimont	259000	0.859	No	0.335	Yes	0.127	Yes
30	Westmount	758000	0.955	No	0.465	Yes	0.339	Yes

More specifically, for steps 3 and 4, each number in the column “ u_k ” is compared with 0.466 and a unit k is chosen if $u_k < 0.466$, $k = 1, \dots, 30$. The resulting samples, s_1 , s_2 , and s_3 are given by: $s_1 = \{1, 2, 3, 5, 6, 7, 9, 10, 13, 17, 19, 24, 25, 26\}$; $s_2 = \{4, 8, 9, 10, 12, 13, 14, 16, 17, 19, 21, 22, 25, 27, 28, 29, 30\}$; $s_3 = \{1, 3, 4, 5, 6, 7, 9, 11, 12, 13, 16, 17, 18, 19, 22, 23, 24, 28, 29, 30\}$.

Since the N experiments are independent and using the basic property of the uniform distribution, the inclusion probability of the sampling unit k is clearly n/N . The student may appreciate that some samples contain a given unit k while others not, and that the inclusion probability corresponds to the chances that the sample s contains the fixed unit k . The instructor may wish to stress the fact that what is random is the sample s and that the Y_k ’s are not random quantities. From one sample to the next, what is random is the inclusion of a given unit in the sample. For example, from the [Example 1](#), the unit *Ahuntsic* ($k = 1$) is included in the first and third samples but not in the second. However, the price of an house in Ahuntsic is the fixed real number $y_1 = 229,000$.

By comparison, to illustrate the inclusion probability under the SRS design, the instructor would need a more elaborate illustration, based either on a long enumeration of all the samples (or on many samples) and on the idea of the Monte Carlo simulation (which is presented later in the course). Based on BE sampling, the requirements seem minimal. Additional technical exercises on small populations are naturally useful in understanding inclusion probabilities ([Särndal, et al. 1992](#); [Lohr 1999](#)). Our illustration with BE sampling represents an intuitive complement, without the exasperation of calculations. Note that a generalization of BE sampling, called Poisson (PO) sampling, could possibly be useful in illustrating plans with unequal inclusion probabilities. To apply a PO design, the sampler needs to specify the π_k ’s. He then proceeds as in the BE design, except that he replaces step 3 with the following:

Step 3’. If $u_k < \pi_k$, choose unit k . If not, do not include k in the sample.

The π_k 's correspond to the inclusion probabilities. When $\pi_k = n/N, \forall k \in U$, we retrieve the BE sampling plan. A natural question is how to choose the π_k 's. If x is a positive auxiliary variable, available for each unit k in the population, a possible choice consists in specifying:

$$\pi_k = \frac{nx_k}{\sum_U x_k}.$$

For the estimation of the population mean or the population total, it is known that if the variable of interest y is proportional to the auxiliary variable x , then that choice of the π_k 's will give a small variance for certain estimators of the total T_y . The PO design falls somewhat beyond the scope of the present paper and we refer the reader to [Särndal, et al. \(1992\)](#) for more details on this particular sampling plan.

We should note that with BE sampling, the sample size of s , say n_s , could differ from the planned or expected size $E(n_s) = n$. Indeed, a possible drawback of BE sampling is that the sample size is a random quantity. Thus, in [Example 1](#), the expected sample size was $n = 14$ and the final sample sizes of s_1, s_2 and s_3 were 14, 17 and 20, respectively. For some samplers, this represents a serious disadvantage. For others, it is of little importance, since in practice, due to possible non-response, the *final* sample size will be probably different of the planned sample size. According to [Särndal \(1996\)](#), we should not consider BE design inferior because of the random sample size. In his paper, he mentions several successful applications of sampling plans and strategies (a strategy is a combination of an estimator and a sampling plan) with random sample size. From a pedagogical point of view, the successful illustration of the inclusion probability largely compensates for the random sample size.

We shall now discuss the point estimation of P . Under SRS sampling, the natural unbiased estimator is the sampling proportion, that is

$$P_s = \frac{1}{n_s} \sum_s y_k,$$

where n_s is the fixed planned size. For BE sampling, an unbiased estimator is

$$P_{BEs} = \frac{1}{n} \sum_s y_k = \frac{n_s}{n} \frac{1}{n_s} \sum_s y_k = \frac{n_s}{n} P_s,$$

where n_s is now the final random sample size. In the following example, we compute point estimators of P with the BE samples taken from [Example 1](#).

Example 2

Consider the housing data described in [Example 1](#). Suppose that we are interested in estimating the proportion of regions in Greater Montreal such that the price of an executive, detached two-storey house is higher than 260,000 CAN\$. Note that according to [Table 1](#) the true unknown proportion is $P = 12/30 = 40\%$. We need to introduce the following dichotomous variable y :

$$\begin{aligned} y_k &= 1, \text{ if the price of the house for region } k \text{ is higher than 260,000 CAN\$}, \\ &= 0, \text{ if not.} \end{aligned}$$

Recall that the expected sample size is $n = 14$ and the final sample sizes are given by $n_{s_1}=14$, $n_{s_2}=17$ and $n_{s_3}=20$. From [Table 2](#), we obtain that $\sum_{s_1} y_k=5$, $\sum_{s_2} y_k=8$ and $\sum_{s_3} y_k=8$. Consequently, the point estimations for the estimator P_{BES} are given in the [Table 3](#).

Table 3. Point estimators based on the three samples considered in [Example 1](#).

P_{BES}	s_1 $5/14=35.7\%$	s_2 $8/14=57.1\%$	s_3 $8/14=57.1\%$
-----------	------------------------	------------------------	------------------------

At first look, the point estimators in [Table 3](#) may seem counterintuitive for the students. They may find that the sample proportions $P_{s_1} = 5/14 = 35.7\%$, $P_{s_2} = 8/17 = 47.1\%$ and $P_{s_3} = 8/20 = 40.0\%$ are more natural estimators. Furthermore, it seems intuitively that the sample proportions are closer to the population proportion P ! However, the estimator P_{BES} is exactly an unbiased estimator of P , when the sample comes from a BE sampling plan. This illustrates that the form of the estimator may be affected by the sampling plan. The apparent large variations of the estimators in [Table 3](#) are explained in part by the fixed denominator n of the estimator P_{BES} . It can be shown that a better estimator than P_{BES} for the estimation of the proportion P is precisely the sample proportion P_s , even if the sample is obtained according to the BE sampling plan. Though slightly biased, this estimator does exhibit less variability. Replacing n by the random size n_s in the denominator of P_{BES} reduces the part of the variability related to the sample size variation. Another example and a discussion are given in [Särndal, et al. \(1992\)](#).

The estimators P_s and P_{BES} are unbiased estimators for the true proportion P , under SRS and BE sampling plans, respectively. They are special cases of the general Horvitz-Thompson (HT) estimator. That estimator is a key quantity in [Särndal, et al. \(1992\)](#). The HT estimator allows us to obtain unbiased estimators when the sample comes from a general sampling plan $p(\cdot)$ taken in a finite population U . The general formula for the HT estimator for the total $\sum_U y_k$ is

$$T_{ps} = \sum_s y_k / \pi_k .$$

The variance of T_{ps} is given by $V_p(T_{ps}) = \sum \sum_U \Delta_{kl} (y_k / \pi_k) (y_l / \pi_l)$, where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ and $\pi_{kk} = \pi_k$. An unbiased estimator of $V_p(T_{ps})$ is $\hat{V}_p(T_{ps}) = \sum \sum_s (\Delta_{kl} / \pi_{kl}) (y_k / \pi_k) (y_l / \pi_l)$ (see [Särndal, et al. 1992](#)). The HT estimator for the proportion P is noted $P_{ps} = \frac{1}{N} \sum_s y_k / \pi_k$. The associated variance estimator is given by $\hat{V}_p(P_{ps}) = N^{-2} \sum \sum_s (\Delta_{kl} / \pi_{kl}) (y_k / \pi_k) (y_l / \pi_l)$.

In the SRS sampling plan, $\hat{V}_{SRS}(T_{ps})$ reduces to the formula $N^2 \frac{(1-f)}{n_s} S_{ys}^2$, which is usually derived in a first course, with $S_{ys}^2 = \frac{1}{n_s-1} \sum_s (y_k - \bar{y}_s)^2$ and where the sampling fraction is $f=n_s/N$. Using the property that y_k is either 0 or 1, we deduce that the estimator of variance for P_s reduces to $\hat{V}_{SRS}(P_{ps}) = \frac{(1-f)}{n_s-1} P_s(1-P_s)$.

For BE sampling, the formula is even simpler, since $\pi_{kl} = \pi_k \pi_l$ for $k \neq l$. A valid variance estimator for P_{BEs} is then given after some algebraic manipulations by:

$$\hat{V}_{BE}(P_{BEs}) = N^{-2} \sum_s \frac{1}{\pi_k} \left(\frac{1}{\pi_k} - 1 \right) y_k^2 = \frac{1}{n} \left(1 - \frac{n}{N} \right) \frac{n_s}{n} P_s.$$

If the sampling distribution of P_{ps} is approximately normal, this allows us to construct a confidence interval for P having the familiar form

$$P_{ps} \pm t_{n-1; \alpha/2} \sqrt{\hat{V}_p(P_{ps})},$$

where $t_{n-1; \alpha/2}$ is the $(1-\alpha/2)$ th quantile of a Student t distribution with $n-1$ degrees of freedom. For large n , we can replace $t_{n-1; \alpha/2}$ by the $(1-\alpha/2)$ th quantile $z_{\alpha/2}$ of a normal distribution. With $\alpha=5\%$, such confidence intervals should contain the true parameter P around 95% of the time. For SRS sampling plan, the adequacy of the normal approximation for a general variable of interest y will depend on the sample size and on how closely the population U resembles a population generated from the normal distribution. See also [Lohr \(1999\)](#), who presents an interesting discussion on confidence intervals in finite population sampling problems. In estimating proportion P , the usual rule $nP \geq 5$ and $n(1-P) \geq 5$ is a useful guideline in deciding whether the sample size is large enough to use the normal approximation. [Cochran \(1977\)](#) discusses the validity of the normal approximation of the sample proportion under SRS design.

Example 3

In [Example 2](#), we computed point estimators. We can now provide the variance estimators of P_{BEs} for the three samples obtained under the BE design in the [Example 1](#). Using the results given in the [Example 2](#), the variance estimators $\hat{V}_{BE}(P_{BEs})$ for the samples s_1, s_2 and s_3 are $2/147, 16/735$ and $16/735$, respectively. It might seem tempting to use the point and variance estimators to produce confidence intervals. However, it seems that the sample size and the population size are rather small. For illustrative purposes, we set $\alpha=5\%$, giving a quantile equal to $t_{13, 2.5\%}=2.16$. Consequently, the confidence intervals for these three samples are $[0.11, 0.61]$, $[0.25, 0.89]$ and $[0.25, 0.89]$, at the 95% confidence level. These intervals are quite large, reflecting the variability of the estimator P_{BEs} .

2.2 Stratified sampling with SRS and BE sampling plans

Sometimes the population can be naturally divided into H groups, called strata. Common variables of stratification are regions, geographic areas, etc. At the stratum level, the sample s_h is obtained by drawing in stratum h , $h=1,2, \dots, H$, a sample of size n_{hs} independently in each stratum of size N_h . For example, we could consider using the SRS sampling plan in each stratum to select s_h , $h=1,2, \dots, H$; the resulting sample at the population level is $s=\bigcup_{h=1}^H s_h$. This sampling design is

called the stratified simple random sampling, noted STSRS. Another possibility is to draw in each stratum h a random sample using BE sampling. We denote the stratified Bernoulli sampling STBE. Such sampling plans are considered in [Särndal \(1996\)](#).

Under STSRS, a natural unbiased estimator is given by $P_{st,s} = \sum_{h=1}^H W_h P_{hs}$, where $W_h=N_h/N$ is the proportion of units in stratum h and $P_{hs} = \frac{1}{n_{hs}} \sum_{s_h} y_k$ is the sample proportion in stratum h .

Essentially $P_{st,s}$ consists of a weighted average of the proportions in each stratum. Since we draw samples independently in each stratum, the variance of $P_{st,s}$ is the weighted sum of the variance inside each stratum. An unbiased estimator for the variance of $P_{st,s}$ is given by

$$\sum_{h=1}^H W_h^2 \frac{(1-f_h)}{n_{hs}-1} P_{hs} (1-P_{hs}), \text{ where } f_h = n_{hs}/N_h.$$

The same reasoning holds for STBE. As an exercise, we propose finding an unbiased estimator of the variance of $P_{stBE,s} = \sum_{h=1}^H W_h P_{BEhs}$, where $P_{BEhs} = \frac{1}{n_h} \sum_{s_h} y_k$ and n_h is the expected sample size in stratum h . (The answer is $\hat{V}_{stBE}(P_{stBE,s}) = \sum_{h=1}^H W_h^2 \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) \frac{n_{hs}}{n_h} P_{hs}$).

In fact, we should note that $P_{st,s}$ and $P_{stBE,s}$ are the HT estimators under STSRS and STBE respectively. The inclusion probabilities under STSRS and STBE are given by $\pi_k = n_{hs}/N_h$ and $\pi_k = n_h/N_h$, respectively, when the sampling unit k lies in stratum h .

3. The Logistic Regression Estimator

Auxiliary information is often available in survey sampling. This information, which may come from past census or from other administrative sources, can be used to obtain more accurate estimators. When auxiliary information is made available, we might still decide to execute a SRS sampling plan, but we would want to change the estimation method. There are other choices available for making use of auxiliary information, such as the ratio estimator or the regression estimator. For example, to estimate the total T_y , we could decide to replace the strategy HT/SRS by the regression estimator with an SRS design:

$$\hat{T}_{yREG} = N \left\{ \bar{y}_s + \hat{B}(\bar{x}_U - \bar{x}_s) \right\},$$

which is approximately unbiased for the true total T_y . The underlying model is the simple regression model with an intercept and the slope estimator is given by \hat{B} . More generally, in a multiple regression model $y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k$, the general regression estimator (called GREG) is given by:

$$\hat{T}_{yGREG} = \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_s + \sum_s (y_k - \mathbf{x}_k' \hat{\mathbf{B}}_s) / \pi_k,$$

$$\text{where } \hat{\mathbf{B}}_s = \left(\sum_s \mathbf{x}_k \mathbf{x}_k' / \pi_k \right)^{-1} \sum_s \mathbf{x}_k y_k / \pi_k.$$

The usual estimators for a proportion usually cannot incorporate auxiliary information. A student might ask why not try to improve the estimation of the HT estimator for the proportion with a certain estimator function of the \mathbf{x}_k 's. However, the regression estimator is fully justified when the variable of interest is continuous. Since the variable Y is dichotomous when we estimate a proportion, it may be more natural to consider a logistic model for the population, where it is assumed that $\{\mathbf{x}_k, k \in U\}$ is known. For a given \mathbf{x}_k , the model is given by:

$$\Pr(Y_k = 1) = \frac{\exp(\mathbf{x}_k' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_k' \boldsymbol{\beta})},$$

and $\Pr(Y_k = 0) = 1 - \Pr(Y_k = 1)$. The parameter $\boldsymbol{\beta}$ is estimated by the following HT estimator of the log-likelihood:

$$L(\boldsymbol{\beta}) = \sum_s [I(Y_k = 0) \log(1 - \mu_k) + I(Y_k = 1) \log \mu_k] / \pi_k,$$

where $\mu_k = E(Y_k | \mathbf{x}_k, \boldsymbol{\beta}) = \Pr(Y_k = 1 | \mathbf{x}_k, \boldsymbol{\beta})$ and $I(A)$ is the indicator variable for set A . See also the logistic model described in [Särndal, et al. \(1992\)](#) and [Lohr \(1999\)](#). The predicted values for the μ_k 's are given by $\hat{\mu}_k = \Pr(Y_k = 1 | \mathbf{x}_k, \hat{\boldsymbol{\beta}})$, $k = 1, 2, \dots, N$. To obtain the LGREG estimator of [Lehtonen and Veijanen \(1998a\)](#), we need only replace the linear prediction $\mathbf{x}_k' \hat{\mathbf{B}}_s$ of y_k in the GREG by $\hat{\mu}_k$:

$$\hat{T}_{yLGREG} = \sum_U \hat{\mu}_k + \sum_s (y_k - \hat{\mu}_k) / \pi_k.$$

For a discrete variable Y , the LGREG estimator is more natural than the GREG estimator since in the logistic formulation μ_k lies between 0 and 1 and the predicted value $\hat{\mu}_k$ also shares that property. However, we should note that the GREG estimator might need only the population totals of the auxiliary information. By comparison, the LGREG estimator usually requires more knowledge of the \mathbf{x}_k 's in the population U . For more details on that specific aspect, see [Lehtonen and Veijanen \(1998a\)](#).

The LGREG estimator may be useful in constructing an estimator for a proportion P by considering $N^{-1}\hat{T}_{yLGREG}$. It is possible to compute the LGREG estimator under a general sampling plan $p(\cdot)$ including stratified sampling plans such as STSRS or STBE. In these cases, it is natural to consider LGREG estimators separately in each stratum, since we assume that the auxiliary information $\{\mathbf{x}_k, k \in U\}$ is totally known.

From a pedagogical point of view, a first sampling course is too often composed of the following routine: quantity of interest - estimator - variance of the estimator - estimator of the variance. Regression and ratio estimators are introduced as more accurate estimators when the auxiliary information is used efficiently. However, perhaps more emphasis should be placed on the underlying linear model, since it is the only one considered with that kind of estimator. At this point, students may not yet realize why the implicit modelling of the survey data is so important. The LGREG is an example of an estimator justified with logistic models. These models could be introduced as another type of model—providing motivation for the LGREG estimators, highlighting the underlying justification of the different estimators, and stressing the appropriate modelling of observed and available data. Logistic models may help students understand the underlying dichotomous variables. In some circumstances a linear model is adequate, but in some other cases (for example, with dichotomous variables) a logistic model may be preferable. In some sense, logistic models constitute a specialized topic. They could, however, be introduced at the end of a first course in survey sampling or at the beginning of a second course on the subject, whereas regression estimators are usually introduced earlier. The LGREG estimator seems to have good pedagogical merits and it may be useful in teaching with a modelling approach, which is suggested in [Fecso, et al. \(1996\)](#).

Variance estimation remains an important consideration for the practical applications. A possible variance estimator for \hat{T}_{yLGREG} discussed in [Lehtonen and Veijanen \(1998a\)](#) is given by the following formula:

$$\hat{V}_{LGREG,p} = \sum \sum_s (\Delta_{kl}/\pi_{kl})(e_k/\pi_k)(e_l/\pi_l),$$

where $e_k = Y_k - \hat{\mu}_k$. That formula takes the same form that the variance estimator \hat{V}_p for the HT estimator. It is also similar to the linearized variance estimator for the GREG estimator ([Särndal, et al. 1992](#)). We conclude this section with an exercise that is an application of the results obtained in [Section 2](#).

Exercise: Find the variance estimators for the LGREG estimator under the sampling plan a) SRS, b) BE, c) STSRS and d) STBE.

Answer to the exercise:

- Let us begin under the SRS sampling plan. Using a common trick, it suffices to realize that the same algebraic developments will occur with the variable e instead of a general variable y . Note that the residual e_k is not dichotomous. Recall that the variance estimator for the HT

estimator of a general variable y under SRS reduces to $N^2 \left(\frac{1-f}{n_s} \right) S_{ys}^2$ where

$S_{ys}^2 = \frac{1}{n_s - 1} \sum_s (y_k - \bar{y}_s)^2$ is the sampling variance. Then for the LGREG estimator the

formula is simply $N^2 \left(\frac{1-f}{n_s} \right) S_{es}^2$, where $S_{es}^2 = \frac{1}{n_s - 1} \sum_s (e_k - \bar{e}_s)^2$.

b) Under BE, the variance estimator of the HT estimator is the expression

$\sum_s \frac{1}{\pi_k} \left(\frac{1}{\pi_k} - 1 \right) y_k^2 = \frac{N}{n} \left(\frac{N}{n} - 1 \right) \sum_s y_k^2$. For the LGREG estimator, the formula is
 $\frac{N}{n} \left(\frac{N}{n} - 1 \right) \sum_s e_k^2$.

c) Under STSRS, the variance estimator of the HT estimator $\sum_{h=1}^H N_h \bar{y}_{s_h}$ is given by

$\sum_{h=1}^H N_h^2 \frac{(1-f_h)}{n_{hs}} S_{ys_h}^2$, where $f_h = n_{hs}/N_h$. For the LGREG, the expression becomes
 $\sum_{h=1}^H N_h^2 \frac{(1-f_h)}{n_{hs}} S_{es_h}^2$.

d) Under STBE, the variance estimator of the HT estimator $\sum_{h=1}^H \frac{N_h}{n_h} \sum_{s_h} y_k$ is

$\sum_{h=1}^H \frac{N_h}{n_h} \left(\frac{N_h}{n_h} - 1 \right) \sum_{s_h} y_k^2$. For the LGREG estimator, the formula reduces to

$\sum_{h=1}^H \frac{N_h}{n_h} \left(\frac{N_h}{n_h} - 1 \right) \sum_{s_h} e_k^2$.

3.1 Computing the LGREG estimator

In the logistic model, estimating the parameter β represents an important step. In general, the model is estimated by maximizing the weighted log-likelihood. A Newton-Raphson algorithm could be used to maximize the likelihood function numerically. See [Lehtonen and Veijanen \(1998b\)](#) for more numerical details. We developed some S-Plus codes to compute the LGREG estimator. In the more general case the inclusion probabilities might be unequal. In that situation we could use the general S-Plus function `nlminb`. With SRS and BE sampling plans, the inclusion probabilities are all equal. In that case we can use the built-in function `multinom` coming from the `nnet` library created by W.N.Venables and B. Ripley and described in their book ([Venables and Ripley 1994](#)). The library is included with the professional edition of S-Plus 2000 for Windows. For STSRS and STBE designs, we can use `multinom` in each stratum. In our simulations, `multinom` was much more faster than `nlminb`. We provide below some S-Plus codes. The first function computes the LGREG estimator and the second is useful in obtaining the log-likelihood. All the codes for reproducing the simulation results of the next section can be obtained by communicating with the author.

```

LGREG <-
function(beta, echan, y.s, x, pik)
{
# beta: vector that corresponds to the estimator of beta
# echan: corresponds to the indices for the sample
# y.s: values of variable y for s; y[echan] is y.s
# X: matrix of auxiliary information for the whole population
# We assume a general sampling design.
  N <- length(X[, 1])      # Size of the population
  n <- length(y.s) # Sample size
  z.s <- ifelse(y.s == 1, 1, 0)
  weight <- 1/pik
  xbeta <- as.vector(X %*% beta)
  den <- 1 + exp(xbeta)  # den is a vector of size N
  predict1 <- exp(xbeta)/den
  That1 <- sum(predict1) + sum(weight * (z.s - predict1[echan]))
  e.s <- z.s - predict1[echan]
  list(That1 = That1, e.s = e.s)
}

loglik.LGREG <-
function(beta, y = y, X = X, pik = pik)
{
# Computation of the loglikelihood
  weight <- 1/pik
  z0 <- ifelse(y == 0, 1, 0)
  z1 <- ifelse(y == 1, 1, 0)
  xbeta <- as.vector(X %*% beta)
  den <- 1 + exp(xbeta)
  predict0 <- 1/den
  predict1 <- exp(xbeta)/den
  term0 <- z0 * log(1 - predict1)
  term1 <- z1 * log(predict1)
  - sum(weight * (term0 + term1))
}

```

4. Application with the *Academic Performance Index* Base Datafile

Two topics that may be covered at the end of a first course in sampling theory are random number generation and the Monte Carlo simulation. This is often one of the first contacts students will have with pseudo-random numbers and random number generation via the inverse transformation of the distribution function. The Monte Carlo simulation serves to illustrate many fundamental issues, such as the probability concept of convergence and the statistical concepts of bias, variance and confidence interval. As a technique it is particularly useful when the exact description of an estimator's sampling distribution is rather difficult to obtain. Ideally, if all the samples s that are possible under a certain sampling plan were obtained, then one could determine the exact bias of an estimator or the exact confidence level of a certain procedure. However, this is often a task of enormous proportions, since the number of possible s increases rapidly as a function of N . To perform a Monte Carlo simulation in our context, we draw B samples independently from a certain population, where each sample s is obtained according to the sampling design $p(\cdot)$. The number

of replications B must be taken reasonably large, for example we could use $B = 1000$. Then, we use each sample to compute certain estimators and/or confidence intervals for the parameter of interest. For example, to appreciate empirically the bias of the estimator \hat{P} of the true proportion P , it suffices to compute the Monte Carlo mean of the B estimators $B^{-1} \sum_{i=1}^B \hat{P}_i$, where \hat{P}_i represents the estimator \hat{P} calculated at the i th replication; the Monte Carlo bias is given by $B^{-1} \sum_{i=1}^B \hat{P}_i - P$. This involves the law of large number, since according to that probability result $B^{-1} \sum_{i=1}^B \hat{P}_i$ converges to $E(\hat{P})$ in probability. By computing the Monte Carlo variance of the B estimators, we may also study the variability of the estimators. Another possible application of the Monte Carlo simulation is to help the students appreciate the meaning of a confidence interval, by examining its coverage properties empirically. It suffices to compute B confidence intervals based on the B samples and to count the number of times that the true proportion P belongs to the B confidence intervals. Furthermore, since the confidence intervals are justified with asymptotic arguments, the students will have the opportunity to observe whether the coverage rates are close to the asymptotic confidence level $1-\alpha$ in finite samples. See also [Särndal, et al. \(1992\)](#), who describe a complete Monte Carlo experiment.

More specifically, the Monte Carlo experiments in this section will help to illustrate the properties of the estimators considered in this paper, particularly the coverage properties of the confidence intervals and some efficiency considerations as well (for example, reduction of the variance stemming from auxiliary information; adequate modelling; and choice of the sampling plan). We carried out the simulations from the 2000 *Academic Performance Index* Base data file. In the State of California, the 1999 Public Schools Accountability Act requires that the Department of Education calculate the API every year: The API is a performance indicator (on a scale of 200 to 1000) for public schools. Schools are ranked, and the results are published and made available to the public. Performance scores are particularly important for schools in California, since rewards are granted if annual API growth targets are reached. For 2000, the API consists solely of results from the Stanford 9 norm-referenced assessment. Documents describing the 2000 API are available from the California Department of Education, and the API Base data file can be downloaded from the Internet at the California Department of Education web site.

The original database contains 7367 schools, but 174 of these did not receive a 2000 API score. For our simulation study, we simply removed these schools, leaving a database consisting of 7193 schools. When a school contains a sub-group with a given characteristic (ethnic or socio-economic) accounting for more than 15% of the total pupil population and consists of at least 30 pupils—or if 100 pupils with valid Stanford 9 scores have the given characteristic—then that sub-group is said to be numerically significant. Assume that, for a given socio-economic study, we need to conduct a survey to find an estimator for the proportion of schools with a significant socio-economically disadvantaged (SD) sub-group. Suppose that from an external source, we have at our disposal some auxiliary information. That information consists of the 2000 API score (variable API00) and the percentage of students tested who are also participants in the free or reduced-price lunch program (variable MEALS). It seems natural to include that kind of information in a model studying a variable related to socio-economic issues. In estimating the logistic model, we considered the variables $\log(\text{API00})$ and $\text{MEALS}/100$. The logarithm was taken to reduce the order of magnitude for the API00 variable; and the MEALS variable in the

original database is expressed as a percentage. Our example is fictitious but it serves the purposes behind our Monte Carlo experiment based on the real information in the API Base data file.

At this point, the instructor could challenge his students with some of the following questions:

1. Should we use the HT estimator in this problem?
2. What should be the impact of the auxiliary information on the point estimators, the variances and the confidence intervals?
3. Is the regression estimator appropriate in this situation?
4. What should be the impact of the sampling plan on the variances of the estimators?
5. From a modeller's perspective, why is the logistic regression estimator more satisfactory? Would we gain any benefits in choosing a better model? What kind of benefits?

To see if some benefits can be obtained, a Monte Carlo study is conducted to investigate empirically the accuracy of the confidence intervals when auxiliary information is included in the estimation of proportions. In the first experiment, one thousand samples of size $n=1800$ were independently drawn with the SRS and BE sampling plans (with n as the expected sample size for BE) to estimate the proportion P . That sample size gives a sampling fraction of about 25%.

In the second experiment, we assume that, prior to the survey, it will be possible to identify the type for each school. Stratification according to school type will allow us to obtain separate estimations for each category: Elementary, High, Middle, and Small. A small school has between 11 and 99 pupils with valid Stanford 9 test scores. Summary information of the auxiliary information is given in [Table 4](#). We observe that the API scores are quite similar among the school types. However, on average, it seems that the percentage of students tested who are also participants in the free or reduced program is lower in high schools when compared with the average at the population level of the MEALS variable. The size of the strata were $N_1 = 4779$, $N_2 = 854$, $N_3 = 1125$, and $N_4 = 435$, respectively. As in the first experiment, we considered a total sample size of 1800 which we allocate proportionally as $n_1 = 1196$, $n_2 = 214$, $n_3 = 281$ and $n_4 = 109$. The proportional allocation is often sufficient for the estimation of a proportion, since the gain of optimal allocation over proportional allocation is usually small ([Kish 1965](#); [Cochran 1977](#)).

Table 4. Summary of the auxiliary information.

	API00		MEALS	
	Mean	Variance	Mean	Variance
Population level	664.3	(16308.6)	47.6	(924.0)
Elementary	671.6	(17044.2)	51.8	(958.5)
High	634.2	(11613.6)	30.9	(565.5)
Middle	655.3	(15435.9)	44.1	(752.2)

Small	666.5	(17234.7)	43.7	(914.1)
-------	-------	-----------	------	---------

Based on these samples, we calculated confidence intervals for P using the HT estimator under SRS and BE sampling plans. The confidence intervals are given by

$$\frac{1}{n_s} \sum_s y_k \pm z_{\alpha/2} \sqrt{\frac{(1-f)}{n_s-1} P_s (1-P_s)}$$

and

$$\frac{1}{n} \sum_s y_k \pm z_{\alpha/2} \sqrt{\frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{n_s}{n} P_s},$$

respectively. For the LGREG estimators, the confidence intervals under the sampling plans SRS and BE are

$$N^{-1} \sum_U \hat{\mu}_k + \frac{1}{n_s} \sum_s (y_k - \hat{\mu}_k) \pm z_{\alpha/2} \sqrt{\left(\frac{1-f}{n_s}\right) S_{es}^2},$$

$$N^{-1} \sum_U \hat{\mu}_k + \frac{1}{n} \sum_s (y_k - \hat{\mu}_k) \pm z_{\alpha/2} \sqrt{N^{-2} \frac{N}{n} \left(1 - \frac{n}{N}\right) \sum_s e_k^2},$$

respectively, where $e_k = Y_k - \hat{\mu}_k$. The preceding formulas are appropriate for constructing confidence intervals at the population level or in a particular stratum (in the latter case it suffices to replace U by U_h , N by N_h , n by n_h and finally, n_s by n_{sh}). Under a stratified design, we can also calculate confidence intervals at the population level, by combining the estimations in each stratum. For the HT estimator, the confidence interval under the STSRS design is given by

$$\sum_{h=1}^H W_h P_{hs} \pm z_{\alpha/2} \sqrt{\sum_{h=1}^H W_h^2 \frac{(1-f_h)}{n_{hs}-1} P_{hs} (1-P_{hs})}$$

and under the STBE sampling plan the formula is

$$\sum_{h=1}^H W_h P_{BEhs} \pm z_{\alpha/2} \sqrt{\sum_{h=1}^H W_h^2 \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) \frac{n_{hs}}{n_h} P_{hs}}.$$

For the LGREG estimators, the confidence intervals under STSRS and STBE are

$$N^{-1} \sum_{h=1}^H \hat{T}_{yhLGREG} \pm z_{\alpha/2} \sqrt{\sum_{h=1}^H W_h^2 \frac{(1-f_h)}{n_{hs}} S_{es_h}^2},$$

$$N^{-1} \sum_{h=1}^H \hat{T}_{yhLGREG} \pm z_{\alpha/2} \sqrt{N^{-2} \sum_{h=1}^H \frac{N_h}{n_h} \left(\frac{N_h}{n_h} - 1 \right) \sum_{s_h} e_k^2},$$

respectively, where $\hat{T}_{yhLGREG}$ represents the LGREG estimator in stratum h . All the formulas for the estimators and the variance estimators are reproduced in the Appendix of the paper. The GREG estimator is also included in our study. The confidence intervals are similar to the confidence intervals of the LGREG estimator, if one replaces $\hat{\mu}_k$ by $\mathbf{x}'_k \hat{\mathbf{B}}_s$ and the residuals are computed as $\tilde{e}_k = Y_k - \mathbf{x}'_k \hat{\mathbf{B}}_s$. The true values for the unknown parameters of interest at the population level and at the stratum level are given in [Table 5](#). We note a large difference between small schools and the other schools. This may be partly related to the fact that, because of the 30-pupils rule, no subgroup will be numerically significant for a very small school (less than 30 pupils). That variable should therefore be interpreted with great caution. However, from a practical point of view, it may be useful to have separate estimations for each strata. In this example, the small proportion of significant SD sub-group for small schools was naturally hidden at the population level.

Table 5. True values of the parameters of interest.

	True proportion
Population level	5772/7193 = 80.24%
Elementary	3956/4779 = 82.78%
High	712/854 = 83.37%
Middle	963/1125 = 85.60%
Small	141/435 = 32.41%

For an estimator \hat{P} , we noted $E_{MC}(\hat{P})$ the Monte Carlo mean and $\text{var}_{MC}(\hat{P})$ the Monte Carlo variance. The Monte Carlo mean of the estimators of variance is given by $E_{MC}(\hat{V})$. Finally, the empirical coverage rates of the confidence interval of the form $\hat{P} \pm z_{\alpha/2} \hat{V}^{1/2}$ are given in the CR column. With 1000 samples, acceptable values are in the interval [93.65%, 96.35%].

The results of the first experiment are presented in [Table 6](#). All the estimators of P had a slight bias, and the mean of the variance estimators was reasonably close to the Monte Carlo variance. All the confidence intervals had empirical coverage rates close to the nominal level, and all the values were in the interval [93.65%, 96.35%]. The LGREG estimator was more efficient than the HT estimator. Under the SRS sampling plan, since $6.6102/2.5631 = 2.58$, in our experiment the

LGREG was much more efficient than the HT estimator. This illustrates the merits of an estimator that can take advantage of the auxiliary information available. The GREG seems unbiased for the true total, but it is less efficient than the LGREG estimator. This illustrates the robustness property of the GREG estimator, since even if the model is misspecified, the regression estimator remains unbiased and the variance estimator formula is still valid. These empirical results show that the GREG estimator is model assisted, but not model dependent (Särndal, et al. 1992). However, a substantial reduction of variance is possible with the LGREG estimator, illustrating that a linear regression model was not appropriate for these data.

The HT estimator under BE sampling was less efficient than under the SRS design. It is a well known fact that the HT estimator may suffer of a variance penalty when the sample size is random (Särndal, et al. 1992). Interestingly, in our experiment, there is no variance penalty when a variable sample size is used with LGREG or GREG estimators, since the Monte Carlo variances under SRS or BE designs were similar. The sample mean (sample variance) of the final sample size for BE sampling was 1801.66 (1318.53).

Table 6. Results of the first experiment.

	SRS sampling plan			
	$E_{MC}(\hat{P})$	$\text{var}_{MC}(\hat{P})$ (*)	$E_{MC}(\hat{V})$ (*)	CR
HT	80.27%	6.6102	6.5964	95.4%
GREG	80.24%	4.2942	4.3509	96.0%
LGREG	80.23%	2.5631	2.6249	95.5%
	BE sampling plan			
	$E_{MC}(\hat{P})$	$\text{var}_{MC}(\hat{P})$ (*)	$E_{MC}(\hat{V})$ (*)	CR
HT	80.33%	32.367	33.460	95.1%
GREG	80.27%	4.3433	4.3552	95.5%
LGREG	80.25%	2.5375	2.6256	95.4%

* column $\times 10^{-5}$

In Table 6, only results at the population level could be computed, since, in our experiments, the SRS and BE sampling plans did not incorporate the school type as an auxiliary variable. With stratified designs, estimations are calculated at the population level and also for each school type. The results of the second experiment are presented in Table 7. First we study the estimation of the proportion at the population level. We observe that including school type as a stratification variable gave a lower variance for the LGREG estimator. If we compare the SRS and STSRS sampling plans, we observe that the gain in efficiency is $2.5631/1.2492 = 2.05$. In our experiment, we see that the gain was more modest when we compare the SRS/HT and STSRS/HT strategies. The GREG estimator was still less efficient than the LGREG estimator. When the HT/BE and

HT/STBE strategies are compared, it appears that stratification did not improve the estimation. The sample means (sample variances) of the final sample size under STBE for each stratum were 1197.20 (852.54), 214.28 (184.63), 281.23 (218.67) and 108.94 (85.09). The confidence intervals had empirical coverage rates close to the nominal level and all the values were in the interval [93.65%, 96.35%].

Second, we study the estimation of the proportions at the stratum level. It seems that we obtained more accurate results for the LGREG estimator, particularly in the stratum consisting of Elementary schools, since the reduction of variance for STSRS is $8.4534/1.0179 = 8.30$. It does, however, seem that the coverage rates were slightly below the nominal coverage rate, particularly for Middle schools. This is related to the fact that the estimators of variance seem to underestimate the true variance in that stratum; indicating that the confidence interval is too narrow. A similar behaviour has been reported in [Lehtonen and Veijanen \(1998a\)](#). The LGREG estimator was more efficient than the GREG estimator in each stratum, except for the small schools where it was slightly less efficient. Under STBE sampling with LGREG estimator, spectacular variance reductions were obtained, particularly in Elementary and Middle strata. The LGREG/STBE and LGREG/STSRS strategies gave very similar results.

Table 7. Results of the second experiment.

	$E_{MC}(\hat{P})$	$STSRS$ sampling plan $var_{MC}(\hat{P}) (*)$	$E_{MC}(\hat{V}) (*)$	CR
HT, pop level	80.27%	5.8346	5.9836	95.0%
GREG, pop level	80.28%	3.6950	3.6499	94.0%
LGREG, pop level	80.25%	1.2492	1.1217	93.8%
HT, Elem school	82.82%	8.4534	8.9202	95.0%
GREG, Elem school	82.82%	4.9456	5.0862	95.1%
LGREG, Elem school	82.77%	1.0179	0.9138	93.8%
HT, High school	83.34%	45.942	48.698	94.8%
HT, High school	83.38%	34.726	35.341	94.1%
LGREG, High school	83.41%	19.257	17.185	93.6%
HT, Middle school	85.58%	33.267	32.972	95.3%
GREG, Middle school	85.61%	22.423	22.038	95.4%
LGREG, Middle school	85.59%	5.114	4.268	90.1%
HT, Small school	32.50%	144.615	151.227	95.1%
GREG, Small school	32.52%	96.588	100.472	94.2%
LGREG, Small school	32.50%	101.114	101.644	93.4%
	$E_{MC}(\hat{P})$	$STBE$ sampling plan $var_{MC}(\hat{P}) (*)$	$E_{MC}(\hat{V}) (*)$	CR
HT, pop level	80.32%	33.786	33.460	94.6
GREG, pop level	80.29%	3.2108	3.6472	96.2

LGREG, pop level	80.26%	1.1539	1.1105	93.7
HT, Elem school	82.87%	51.826	51.951	94.8
GREG, Elem school	82.82%	4.5150	5.0959	96.2
LGREG, Elem school	82.79%	0.9393	0.9043	94.5
HT, High school	83.54%	337.10	292.55	92.4
GREG, High school	83.49%	34.668	35.162	94.5
LGREG, High school	83.39%	17.775	17.050	93.9
HT, Middle school	85.55%	247.98	228.41	93.9
GREG, Middle school	85.57%	22.438	22.069	94.4
LGREG, Middle school	85.59%	5.1460	4.2775	90.3
HT, Small school	32.49%	215.498	223.352	95.3
GREG, Small school	32.57%	99.343	99.069	94.0
LGREG, Small school	32.57%	100.416	100.168	94.3

* column $\times 10^{-5}$

To summarize, our analysis of the Monte Carlo experiments allows us to answer the typical questions asked above.

1. The HT estimator remained appropriate, since it provided an unbiased estimator of the true proportion. However, since auxiliary information was made available, much lower variances were obtained with an estimator that makes efficient use of the available variables.
2. The auxiliary information helped to reduce the variance of the estimators. While all the estimators were unbiased or approximately unbiased, the GREG estimators and LGREG estimators showed less variability than the HT estimator and the confidence intervals were more precise for the same level of confidence.
3. Since the regression estimator is model assisted and not model dependent, the estimators of variance for the GREG estimators were valid and the coverage properties were rather close to the nominal confidence level. However, a linear regression model is hard to motivate with discrete data.
4. The empirical variance of the HT estimator was higher under BE sampling than under SRS sampling. Interestingly, the differences in efficiency between these two designs were smaller for the GREG and LGREG estimators. This suggests that the auxiliary information compensated for the random sample size. With stratified sampling plans, more precise estimations at the population level were generally observed, since many unrepresentative samples had been eliminated with an appropriate stratification variable. Furthermore, stratified designs offered the possibility of obtaining separate estimations for each stratum.

5. Conceptually, the logistic model seemed more satisfactory since the underlying variable is dichotomous. The LGREG estimator proved more efficient than the GREG estimator in our experiments, that is the former estimator usually generated lower variances than the latter.

5. Conclusion

The estimation of proportions is an important subject with many practical applications. In a first survey sampling course, auxiliary information is an important topic and the ordinary sample mean or estimated total can be improved with the general regression estimator, since it is capable of incorporating the auxiliary information. However, regression estimators are more suitably used for continuous variables. In estimating a proportion, we might also want to incorporate auxiliary information. In this paper, we demonstrated how this could be done with the logistic regression estimator, which is based on a logistic model. It is more natural to motivate the use of a logistic model for a discrete variable. We also discussed different sampling plans, such as Bernoulli's, which might have interesting pedagogical merits. We further considered stratified sampling plans whose usefulness resides in their capacity to compute a separate estimation for each stratum. Furthermore, in many situations, an efficient stratification variable may be of help in obtaining accurate estimations. In the simulation section, we use a real database to show that smaller variances might be obtained with the logistic regression estimator than with the HT estimator or with the classical regression estimator. In our empirical study, the LGREG estimator gave accurate estimations under various sampling plans, but the best results were observed using stratified designs. In conclusion, when estimating a proportion, the use of auxiliary information may give large gains in efficiency and the choice of an appropriate model may lead to smaller variances. When the time comes to discuss the modelling of survey data, it would seem that instructors may find logistic models and the LGREG estimator to be of great help.

6. Acknowledgements

The author is grateful to the Associate Editor and an anonymous referee for their helpful comments and suggestions that improved this paper.

7. Appendix. Summary of the different estimators

Strategy	Estimator	Estimator of variance
SRS sampling / HT estimator	$P_s = \frac{1}{n_s} \sum_s y_k$	$\frac{(1-f)}{n_s-1} P_s (1-P_s)$
BE sampling / HT estimator	$P_{BES} = \frac{1}{n} \sum_s y_k$	$\frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{n_s}{n} P_s$

STSRS sampling / HT estimator	$P_{st,s} = \sum_{h=1}^H W_h P_{hs}$	$\sum_{h=1}^H W_h^2 \frac{(1-f_h)}{n_{hs}-1} P_{hs} (1-P_{hs})$
STBE sampling / HT estimator	$P_{stBE,s} = \sum_{h=1}^H W_h P_{BEhs}$	$\hat{V}_{stBE} = \sum_{h=1}^H W_h^2 \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) \frac{n_{hs}}{n_h} P_{hs}$
SRS sampling / LGREG estimator	$N^{-1} \hat{T}_{yLGREG} = N^{-1} \sum_U \hat{\mu}_k + \frac{1}{n_s} \sum_s (y_k - \hat{\mu}_k)$	$\left(\frac{1-f}{n_s}\right) S_{es}^2$
BE sampling / LGREG estimator	$N^{-1} \hat{T}_{yLGREG} = N^{-1} \sum_U \hat{\mu}_k + \frac{1}{n} \sum_s (y_k - \hat{\mu}_k)$	$N^{-2} \frac{N}{n} \left(\frac{N}{n} - 1\right) \sum_s e_k^2$
STSRS sampling / LGREG estimator	$N^{-1} \sum_{h=1}^H \hat{T}_{yhLGREG}$	$\sum_{h=1}^H W_h^2 \frac{(1-f_h)}{n_{hs}} S_{es_h}^2$
STBE sampling / LGREG estimator	$N^{-1} \sum_{h=1}^H \hat{T}_{yhLGREG}$	$N^{-2} \sum_{h=1}^H \frac{N_h}{n_h} \left(\frac{N_h}{n_h} - 1\right) \sum_{s_h} e_k^2$

References

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

Fecso, R. S., Kalsbeek, W. D., Lohr, S. L., Scheaffer, R. L., Scheuren, F. J., and Stasny, E. A. (1996), "Teaching Survey Sampling," *The American Statistician*, 50, pp. 328-340.

Kish, L. (1965), *Survey Sampling*, New-York: John Wiley and Sons, Inc.

Lehtonen, R., and Veijanen, A. (1998a), "Logistic generalized regression estimators," *Survey Methodology*, 24, pp. 51-55.

Lehtonen, R., and Veijanen, A. (1998b), "On multinomial logistic generalized regression estimators," Preprint from the Department of Statistics, University of Jyväskylä, Number 22.

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

Moore, D. S., and McCabe, G. P. (1999), *Introduction to the Practice of Statistics* (3rd ed.), New York: W. H. Freeman and Co., Inc.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models* (4th ed.), Chicago: Irwin.

Särndal, C. E. (1996), "Efficient estimators with simple variance in unequal probability sampling," *Journal of the American Statistical Association*, 91, pp. 1289-1300.

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

Venables, W. N., and Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag.

Pierre Duchesne
Département de Mathématiques et Statistique
Université de Montréal
Montréal QC H3C 3J7
Canada
duchesne@dms.umontreal.ca

[Volume 11 \(2003\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Information Service](#) | [Editorial Board](#) | [Guidelines for Authors](#) | [Guidelines for Data Contributors](#) | [Home Page](#) | [Contact JSE](#) | [ASA Publications](#)