

The Advantage and Disadvantage of Implicitly Stratified Sampling

Peter Lynn

Institute for Social and Economic Research, University of Essex

Abstract

Explicitly stratified sampling (ESS) and implicitly stratified sampling (ISS) are well-established alternative methods for controlling the distribution of a survey sample in terms of variables that define the strata. If these variables are correlated with survey estimates, the estimates will benefit from improved precision. With ESS, unbiased estimation of the standard errors of survey estimates is possible, provided that sampling strata membership is identified on the survey dataset. With ISS this is not possible and usual practice is to invoke an approximation that tends to result in systematic over-estimation of standard errors. This can be perceived as a disadvantage of ISS. However, this article demonstrates, both theoretically and through a simulation study, that true standard errors can be smaller with ISS and argues that this advantage may be more important than the ability to obtain unbiased estimates of the standard errors. The simulation findings also suggest that the extent of over-estimation with the usual approximate variance estimator may be modest.

Keywords: standard error, stratified sampling, survey sampling



© The Author(s) 2018. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Introduction

Most surveys use stratified sampling designs. This is done in order to benefit from the precision gains that such designs can bring. For a modest effort in designing the sample, the precision gains can often be equivalent to those that would accrue from carrying out tens or even hundreds of extra interviews. Stratified sampling is therefore highly cost-effective. However, there are many different ways that it can be done. The researcher must choose which variables to use, and how to combine them to define the strata. She must also decide whether all strata should be sampled at the same rate (proportionate stratified sampling) or whether some should be over-sampled, perhaps in order to increase the representation in the sample of certain subgroups (disproportionate stratified sampling). Though the researcher is typically constrained to define strata in terms of information that is either available on the sampling frame or can be linked to the frame, this still usually leaves a lot of options regarding exactly how the information should be used. The better the decisions, the more cost-effective the survey design will be.

This article focuses on one specific decision that the researcher must make: whether to use explicitly stratified sampling (ESS) or implicitly stratified sampling (ISS). For simplicity, the arguments are illustrated in the context of proportionate stratified sampling, but the arguments apply equally when sampling is disproportionate, as a similar decision must be made within each top-level sampling domain. The arguments also apply when a decision is being made about how to stratify at a secondary level, i.e. within primary explicit strata.

ESS involves sorting the population elements into explicit groups (strata) and then selecting a sample independently from each stratum. ISS involves ranking the elements following some ordering principle and then applying systematic sampling, i.e. selecting every n^{th} element. For example, if the sampling frame were a list of people containing a single auxiliary variable, date of birth, proportionate ESS would involve creating strata corresponding to a number of discrete age groups and then selecting, using simple random sampling (SRS), a number of people from each group such that the proportion of the sample in each group equals the proportion of the population in the group. ISS, on the other hand, would involve sorting the people from youngest to oldest (or oldest to youngest; this is equivalent) and then selecting every n^{th} person on the list (after generating a random start point).

The advantage of ESS is that unbiased estimation of the standard errors of survey estimates is possible, provided that the sampling stratum membership is

Acknowledgments

The author is grateful to Seppo Laaksonen for provoking him to write this article.

Direct correspondence to

Peter Lynn, ISER, University of Essex, Wivenhoe Park, Colchester,
Essex CO4 3SQ, UK
E-mail: plynn@essex.ac.uk

identified on the survey dataset and provided that at least two sample elements are selected from each stratum. With ISS this is not possible and usual practice is to invoke an approximation that tends to result in systematic over-estimation of standard errors. This can be perceived as a disadvantage of ISS. However, this begs the question of whether it is better to know the precision of one's estimates or to have more precise estimates without knowing exactly how much more precise they are.

On the other hand, there are several disadvantages of ESS relative to ISS. One of these relates to the focus of this article: a greater precision gain due to stratified sampling can be achieved with ISS than with ESS (Madow and Madow, 1944; Cochran, 1946). Another disadvantage of ESS is that it is not possible to obtain an equal-probability sample unless each stratum size is an exact multiple of the sampling interval. Consequently, unequal design weights must be applied, with an associated further loss in precision. Furthermore, it is not possible to stratify deeply on a combination of many variables, due to restricting limitations on the number of strata and an associated risk of greater variation in the design weights the larger the number of explicit strata relative to the sample size. Deeper stratification is possible with ISS.

ESS is often used in order for different sampling fractions to be applied to different sub-domains of the population (disproportionate stratified sampling), by creating the strata to reflect the sub-domains. However, this should not be perceived as an advantage of ESS as the same can be achieved with ISS by assigning a size measure to each element proportional to the desired sampling fraction and making selections with probability proportional to this size measure. Variance estimation for variable probability systematic sampling is considered by Stehman and Overton (1994).

The potential of ISS to provide a greater precision gain than ESS is recognised in the statistical literature (e.g. Madow & Madow, 1944; Kish, 1965) but is not given attention in the sample design sections of generalist survey research handbooks or textbooks. For example, Groves et al. (2009) explain ESS and how, with proportionate allocation to strata, it can improve precision compared to simple random sampling (pp. 113-120). They then introduce systematic selection as "a simpler way to implement stratified sampling" (p. 122), but make no mention of the implications for precision, other than a rather general statement that "Systematic sampling from an ordered list is sometimes termed "implicitly stratified sampling" because it gives approximately the equivalent of a stratified proportionately allocated sample" (p. 124). Even texts that are devoted specifically to sampling, when written for non-statisticians, do not mention explicitly how ESS and ISS compare in terms of precision. For example, Henry (1990) states that "Systematic sampling has statistical properties that are similar to simple random sampling" (p. 98), and subsequently, "Another advantage is that systematic sampling can be used for de facto stratification to insure proportional representation of the population for some characteristic" (p. 98), but with no further mention of precision. In a subsequent section on ESS, however, Henry states that "stratification reduces standard errors" (p. 101) and demonstrates how this works with formulas and a worked example. Kalton (1983) too explains the variance properties of ESS at some length (pp. 20-24), while

the shorter section devoted to ISS focusses instead on the practicality of implementation: “systematic sampling provides a mean of substantially reducing the effort required for sample selection” (p. 16).

Even the most recent specialist texts on survey sampling provide very little detail on the statistical properties of ISS. Bethlehem (2009) merely points out that, “the sample variance [...] need not necessarily be a good indicator of the variance of the estimator” (p. 79) and then suggests that the only way to obtain an unbiased estimator for the variance is to select multiple samples and combine the observed sample means. Approximations are not mentioned. Valliant et al (2013) state that, “Systematic sampling is often used in practice because it is fairly easy to implement and it can be used to control the distribution of a sample across a combination of auxiliary variables” (p. 63) and “Regardless of the reasons for its use, statisticians usually collapse the selection intervals into one or more analytic strata and pretend the method of selection was something else, like *stsrswor*, *stsrswr*, or *ppswr*, in order to estimate a variance.” (p. 64). It is therefore unsurprising if survey researchers may have the impression that ESS is the (only) way to improve precision compared to SRS.

Furthermore, empirical demonstrations of the relative performance of ESS and ISS are surprisingly hard to find. This article provides an exposition of the distinction between ESS and ISS and attempts, via a simulation study using real survey data, to quantify the extent of the improvement in precision with ISS and the extent of the uncertainty about the improvement in precision if the usual approximation is used to estimate standard errors. In the next section, the relevant aspects of sampling theory are presented and are used to derive an expression for the difference in sampling variance between ESS and ISS. The subsequent sections describe how a simulation study will be used to quantify the true difference in sampling variance between the two designs and the extent to which sampling variance will tend to be over-estimated if the usual approximation is used in the case of ISS. The results from the study are then presented and the implications are discussed in the final section.

Sample Designs and Variance Estimators

For simplicity of exposition, it will be assumed that survey estimates are means or proportions. Under ESS, the sampling variance of the sample mean can be expressed (Kish, 1965, p. 81; Cochran, 1977, p. 69) as:

$$Var(\bar{y}) = \sum_{i=1}^I \frac{N_i S_i^2 (N_i - n_i)}{(N^2 n_i)} \quad (1)$$

where $S_i^2 = Var_i(y_{ik})$ is the variance of y within stratum i (y_{ik} is the value of y for individual k in stratum i);

n_i is the number of sample elements in stratum i ;

N_i is the number of population elements in stratum i ;

and $N = \sum_{i=1}^I N_i$ is the total number of elements in the population.

In this article we will assume the context of proportionate sampling, in which case $\frac{n_i}{N_i} = \frac{n}{N}, i = 1, \dots, I$. With this assumption, expression (1) simplifies to:

$$Var(\bar{y}) = \frac{\sum_{i=1}^I S_i^2 (N_i - n_i)}{nN} \quad (2)$$

From this expression it can be seen that differences between strata in terms of y do not contribute to the sampling variance. The sampling variance depends only on the variance of y within the strata. This demonstrates how stratified sampling improves the precision of estimates; by eliminating any influence on the sample of one part of the variance of y , namely the part that is between-strata. Once a survey has been carried out, assuming equal probabilities of selection, $Var(\hat{y})$ can be estimated in a straight-forward manner from the survey data, by substituting the observed within-stratum sample variances (s_i^2) for the corresponding population variances (S_i^2), thus:

$$\widehat{Var}(\bar{y}) = \frac{\sum_{i=1}^I s_i^2 (N_i - n_i)}{nN} \quad (3)$$

For ISS designs there is of course no concept of explicit strata, so the $\{i\}$ in expression (2) are not defined. The design-based variance of a sample mean is equivalent to that under cluster sampling with a sample size of one cluster (Madow & Madow, 1944). Unbiased sample-based estimators of this variance do not exist. While a number of estimators have been proposed, all of them are biased and all will over-estimate the variance whenever the stratification effect is anything more than negligible (Wolter, 1984; Wolter, 1985, pp. 258-262). A commonly-used variance estimation method is to treat the ordered list of selected elements as if each consecutive pair had been selected from the same stratum, a method referred to by Kish (1965, p. 119) as the “paired selections model”, and by Wolter (1985, pp. 250-251) as the “estimator based on nonoverlapping differences”. Thus, a systematic sample of n elements from an implicitly-stratified list is treated as if it consisted of simple random samples of size 2 from each of $n/2$ explicit strata. Analogous methods, in which elements selected from more than one stratum are treated as if they had been selected from the same stratum, are also sometimes used in the context of ESS, particularly when there exists one or more strata in which only one element is selected or observed (Cochran, 1977; Seth, 1966; Rust & Kalton, 1987). In order to compare the sampling variance of ISS and ESS, we can consider the situation in

which the ISS pseudo-strata are subsets of the ESS strata. This is a realistic reflection of the example mentioned in the previous section of stratifying either explicitly or implicitly using date of birth. We will denote the ISS substrata by $j = 1, \dots, J_j$. Then, the approximation usually invoked to estimate the sampling variance associated with ISS is:

$$\widehat{Var}(\bar{y}) = \frac{\sum_{i=1}^I \sum_{j=1}^{J_i} s_j^2 (N_j - n_j)}{nN} \quad (4)$$

whereas the true ISS sampling variance is:

$$Var(\bar{y}) = \frac{\sum_{h=1}^{N/n} (\bar{y}_h - \bar{\bar{y}})^2}{(N/n - 1)} \quad (5)$$

where

there are N/n possible samples that could be selected, corresponding to the N/n possible random start points;

\bar{y}_h is the sample mean of y for sample h ;

$\bar{\bar{y}} = \frac{n}{N} \sum_{h=1}^{N/n} \bar{y}_h$ is the mean of the N/n sample means.

This true variance can be thought of as the sampling variance of a mean under cluster sampling, with a sample size of one cluster, where the population is divided into N/n clusters, \bar{y}_h are the cluster means, and $\bar{\bar{y}}$ is the population mean.

Expression (4) is also used as an estimator for 1-per-stratum designs. In this case, the estimator is known to be upwardly-biased (Fuller, 2009, p. 202; Breidt et al., 2016). ISS is similar to 1-per-stratum sampling, so the bias in using expression (4) as an estimator for (5) might be assumed to be similar, but the designs are not exactly equivalent. In particular, with ISS stratum boundaries are arbitrary and are constrained only conditionally on the random start, and ordering within strata is not random. It should be clear from expression (4) that both ISS and 1-per-stratum ESS should provide greater precision than the most precise form of ESS that enables unbiased estimation of standard errors, namely 2-per-stratum ESS ($\sum_{i=1}^I J_i = n/2$). If the ordering of elements within each stratum in a 2-per-stratum design is completely random, then further sub-dividing each stratum j into two substrata ($k_j = 1, 2$) to create a 1-per-stratum design will have no effect on the sampling variance as $s_{k_j}^2 = s_j^2 \forall k_j$. But any meaningful ordering of elements within at least some of the strata will result in $s_{k_j}^2 < s_j^2$ for at least some j, k , and hence reduced sampling variance. Wolter (1985) presents a series of simulations in which the estimator based on nonoverlapping differences is shown to sometimes be upwardly-biased and sometimes downwardly-biased as an estimator of the ISS variance, depending on the nature of the population ordering.

Simulation Methodology

Data from wave 1 of *Understanding Society, the UK Household Longitudinal Study*, are treated as population data. These data are used to calculate the sampling variance of means and proportions under simple random sampling, ESS and ISS, in ways that will be described in this section. *Understanding Society* is a large nationally-representative multi-topic general population survey. A stratified, multi-stage sample of addresses was selected (Lynn, 2009) and all persons aged 16 or over resident at a sample address were eligible for an individual interview at wave 1. Members of ethnic minority groups and residents of Northern Ireland were sampled at higher rates than the remainder of the population. Data collection took place face-to-face in respondent's homes using computer-assisted personal interviewing (CAPI) between January 2009 and March 2011. At wave 1, 50,295 individual interviews were completed with sample members. For the illustrative purposes of this article, these individuals are treated as a population from which survey samples are to be selected.

A set of eleven target parameters were selected for study. Of these, five are means of continuous variables and six are proportions based on binary variables. For each, we are interested in comparing the sampling variance of the sample statistic under alternative sampling designs and the estimate of the ISS sampling variance using the successive pairing approach. For ease of exposition and calculation, for each parameter we first amend the population such that N is a multiple of 100. This allows the subsequent creation of equal-sized explicit strata (each containing $N_i = 100$ elements) and the application of implicitly stratified systematic sampling designs in which the sampling interval takes the integer value of 50, the convenience of which will be explained below. From the 50,295 elements, we first drop any with item missing values. This is done separately for each of the eleven target variables, so the dropped elements will differ between the eleven simulated populations. Then, a further set of m elements are dropped (m between 0 and 99) in order to round the population size down to a multiple of 100. The m elements with the smallest analysis weights (largest inclusion probabilities) are chosen. Descriptive statistics regarding this process are presented in Table 1.

For each estimate, the variance and estimated variance for samples of size $N/50$ will be compared under different designs. These designs are simpler than those that tend to be used for real social surveys. Specifically they are all equal-probability single-stage designs, without clustering, and with stratification based on a single auxiliary variable, whereas real designs often involve variable probabilities, multi-stage selection, clustering and multiple stratification variables. The simplifications are introduced in order to provide a simple illustration in which differences between the designs are strictly limited to the aspects of design that are the focus of this article. The following sub-sections describe the sampling variance metrics that were calculated for each of the eleven parameters to be estimated. All but one of the metrics rely on knowledge of the population size, N , and the popula-

Table 1 Simulated Populations for 11 Parameter Estimates

	Understanding Society sample size	Item missing	Also dropped (smallest weights)	Simulated population size, N	Sample size, n
<i>Continuous variables</i>					
Total monthly income	50,295	78	17	50,200	1,004
Monthly benefit income	50,295	3,236	59	47,000	940
Number of children	50,295	50	45	50,200	1,004
Hours of sleep	50,295	12,420	75	37,800	756
Body mass index	50,295	6,432	63	43,800	876
<i>Binary variables</i>					
Limiting long-term illness (%)	50,295	0	95	50,200	1,004
Arthritis (%)	50,295	3,234	61	47,000	940
In paid employment (%)	50,295	90	5	50,200	1,004
Has degree (%)	50,295	86	9	50,200	1,004
Lives with spouse/partner (%)	50,295	0	95	50,200	1,004
Religion makes a great difference (%)	50,295	3,234	61	47,000	940

Note: Hours of sleep was asked in a supplemental self-completion questionnaire that was returned by only 85.9% of interview respondents, whereas all other items were administered in the face-to-face interview. The items on body mass index, arthritis and religion were not included in the proxy version of the face-to-face interview, which was administered for 6.4% of respondents.

tion variance of y , S^2 , each of which were derived in the usual way from the population simulated as described above.

Simple Random Sampling

The variance of \bar{y} under simple random sampling is computed as a benchmark and will be used later in the calculation of design effects for the various sample designs under consideration, to help with interpretation of the findings. It is calculated in the usual way:

$$Var_{SRS}(\bar{y}) = \frac{S^2(N-n)}{nN} \quad (6)$$

Explicit Stratified Sampling with 11 Strata

The first stratified design considered is one with eleven explicit strata, defined by the person's age. The first stratum consists of persons aged 16 to 19; the following nine strata consist of five-year age bands from 20-24 to 60-64; the final stratum consists of person 65 years old or older. Proportionate stratified sampling with a sampling fraction of 1 in 50 is used. The sampling variance of a mean is therefore calculated as in expression (2) above, with $n_i = \frac{N_i}{50}$ and $I = 11$.

Explicit Stratified Sampling with $N/100$ Strata

The second stratified design considered is one with $N/100$ equal-sized explicit strata, again defined by the person's age. It can be seen from Table 1 that this corresponds to between 378 and 502 strata. The strata are created by first sorting the population in increasing order of age and then treating the first 100 in sorted order as the first stratum, and so on. A simple random sample of $n = 2$ is selected from each stratum. The sampling variance of a mean is therefore calculated as in expression (2) above, with $n_i = 2$ and $I = N/100$.

Implicit Stratified Sampling with $n = N/50$

The third design considered involves sorting the population in increasing order of age and then selecting a systematic random sample of $N/50$ cases using a random start between 1 and $N/50$. There are therefore $N/50$ possible samples that could be selected and the sampling variance of a mean is calculated as the variance of the $N/50$ corresponding sample means, as in expression (5), with $n = 50$.

In addition to calculating the true sampling variance for this design, the expected value of the estimated sampling variance was calculated using the consecutive pairs method outlined in section 2 above. This was done by calculating the estimate produced by expression (5) for each of the $N/50$ possible samples and then taking the mean of these $N/50$ values.

Results

For each of the eleven variables, Table 2 presents the true standard error of the sample mean under each of the four sample designs under consideration, as well as the expected value of the estimate of the standard error for the ISS design under the consecutive pairs method. The true value of the population mean is also presented for reference (first column). It is worth noting firstly that the relative standard errors vary greatly between the eleven estimates. Under SRS, they range from 0.01 to

0.08, with the exception of body mass index, which has a relative standard error of 0.65 (driven by a number of influential outliers). This provides a range of circumstances in which to compare the effects of alternative stratified sample designs.

As expected, standard errors are in all cases smaller under stratified sampling than under simple random sampling. In fact the rank order of the four designs in terms of standard error is the same for all eleven estimates: ESS with eleven strata provides an improvement in precision over SRS, ESS with around 500 strata ($N/100$) provides a further improvement, and ISS improves precision further still. The relative extent of the standard error reduction varies between the estimates, however. For example, for estimating mean number of children or the proportion of people in paid employment most of the gains to be had from stratifying by age accrue with the use of just eleven explicit strata: extensions to 500 strata or ISS provide only very modest marginal gains. For body mass index and for the proportion suffering from arthritis, on the other hand, the gains in moving from eleven to 500 explicit strata are similar or greater in magnitude to those in moving from no strata (SRS) to eleven. These differences evidently reflect the differing nature of the associations of the variables with age and are illustrated in Figure 1, which presents the design effect for each of the three stratified designs (ratio of sampling variance under ESS or ISS to that under SRS). The proportion suffering from arthritis stands out as the estimate that gains most in terms of precision from each of the successive enhancements to stratification. The precision gain in moving from the ESS11 to the ESS($N/100$) design demonstrates that tendency to suffer from arthritis is quite strongly associated with age, even within the eleven strata of the ESS11 design. However, the further gain in moving to the ISS design shows that even within (at least some of) the 470 strata in the ESS($N/100$) design there remains an association of arthritis with age. This may seem surprising considering that each of the 470 strata covers an age range of only around 2.5 months, on average, but is explained by the strata towards the upper end of the age range – where arthritis is most prevalent – covering larger age ranges, reflecting the smaller population sizes. The design effect of around 0.65 for this estimate with ISS – the smallest of all the design effects in this study – represents a very considerable precision gain. Without stratification, this improvement in precision would require an increase in the sample size with SRS from 940 to 1,443 – an increase that would have considerable cost.

The other variable that stands out in Figure 1 is the only attitudinal variable in the study, the proportion of people agreeing with the statement that religion makes a big difference in life. This variable stands out because the precision gains from stratification are much more modest than for all other variables. Beliefs about the importance of religion are only very weakly associated with age.

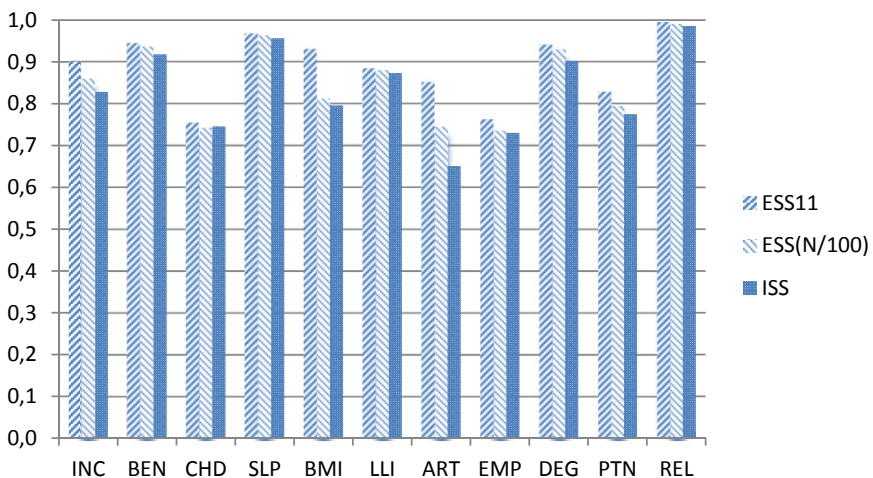
Turning now to the final column of Table 2, it can be seen that the consecutive pairs method of variance estimation for ISS results in a modest over-estimation of standard errors, i.e. an under-estimation of the precision gain from stratification. The expected value of the estimated standard error is typically similar to, or just slightly smaller than, the true standard error with the ESS($N/100$) design. This is of

course the design that is assumed by expression (4) with $n_j = 2$, but the estimated standard errors differ from the true standard errors under this design due to the data having been generated by a different mechanism.

Table 2 Standard errors of means and proportions under four sample designs, and mean estimated standard errors for implicit stratified sampling

	Mean	s.e.				Est.(s.e.)
		SRS	ESS(11)	ESS(N/100)	ISS	ISS
<i>Continuous variables</i>						
Total monthly income	1479.0	49.82	47.28	46.22	45.33	46.24
Monthly benefit income	466.0	37.28	36.23	36.08	35.72	36.15
Number of children	1.600	0.0467	0.0406	0.0403	0.0403	0.0403
Hours of sleep	6.97	0.0587	0.0577	0.0576	0.0574	0.0576
Body mass index	26.06	17.03	16.42	15.35	15.19	15.28
<i>Binary variables</i>						
Limiting long-term illness (%)	34.93	1.489	1.400	1.397	1.392	1.396
Arthritis (%)	14.29	1.130	1.042	0.976	0.912	0.968
In paid employment (%)	52.29	1.560	1.362	1.339	1.333	1.339
Has degree (%)	21.37	1.281	1.242	1.236	1.216	1.226
Lives with spouse/partner (%)	61.51	1.520	1.384	1.356	1.338	1.344
Religion makes a great difference (%)	22.13	1.340	1.337	1.334	1.331	1.333

Figure 1 Design effects for three sample designs



Discussion

The simulation study has shown, using real survey data, that ISS provides useful precision gains relative to ESS. This is true even when comparing to the most detailed form of ESS possible, namely that which involves creating strata such that just two selections are made from each stratum (i.e. the minimum number that permits variance estimation.) This result should lead researchers to question why, whenever useful auxiliary data are available for sample stratification, one would ever choose not to use implicit stratification, given that estimates will be less precise as a result. In practice, ESS typically involves a rather smaller number of strata, such that the average number of sample elements selected from each stratum is very considerably greater than two, perhaps more akin to the ESS11 design presented here, in which around 90 elements are selected per stratum. In this study, the ISS design produced substantially smaller standard errors than the ESS11 design. Gains are apparent, though more modest, even relative to the ESS(N/100) design. There consequently seems to be a strong case for ISS designs rather than ESS designs of this kind.

Furthermore, the approximation commonly used to estimate standard errors with ISS results in only a modest over-estimation. This would make statistical tests slightly conservative, which is probably more desirable than the false precision that would be provided by the opposite. In any case, the extent of the over-estimation (systematic error) is most likely small compared to the extent of sampling variance in the standard error estimate (random error). This conclusion is consistent with that of Wolter (1985, p. 283) who compared eight different possible variance estimators for systematic sampling and concluded that the consecutive pairs estimator “performed, on average, as well as any of the estimators” and “in very small samples ... might be the preferred estimator”.

The choice between ESS and ISS would therefore seem to come down to a choice between improved precision of the survey estimate or unbiased estimation of the precision of the survey estimate. To take the estimation of the proportion of people suffering from arthritis as a concrete example, would researchers prefer to have a standard error of 0.976 associated with their estimate (expected value) of 14.29 (the smallest standard error that would be possible with ESS) and to have an estimate of the standard error with an expected value of 0.976, or to have a standard error of 0.912 (with ISS) and an estimate of the standard error with an expected value of 0.968? For descriptive estimation, it is hard to imagine why the less precise estimate might be preferred. The choice could be less clear, however, when the objective is statistical inference. Analysts could justifiably prefer unbiased hypothesis tests, including those that are implicit in the fitting of statistical models. This distinction between different kinds of analysis objectives is particularly problematic for surveys that are used for both types of analysis, as only one sample design can be used. The ideal solution might be to develop ways of adjusting in inferential analysis for the bias in the variance estimator.

It should be noted that results could be different if a combination of multiple stratification variables were used rather than a single variable, as in the simulations presented here. With a single stratification variable, it is likely that any relationship of the implicitly stratified ordering with the target parameters will be monotonic, or at most quadratic in nature, whereas when combining variables large discontinuities in the distribution can occur at the boundaries of categories of a variable. However, there is no suggestion in Wolter (1985, p.268) that the bias in the consecutive pairs estimator is strongly dependent on whether one, two or three stratification variables are used.

A limitation of the empirical results presented here is that they are restricted to full-sample means and proportions. Some additional simulations (results not shown) for subclass means and proportions based on the same variables suggest that ISS less frequently provides a noticeable improvement in precision over the ESS(N/100) design. This could be because relatively few of the strata in the ESS(N/100) design provide more than one element in the subclass, in which case there is little scope for further precision gains. However, to explore this limitation further, analysis should be extended to a range of subclasses, with different distributions over strata, and to other types of ratio estimates. Such investigation is beyond the scope of this article.

A final point to note is that the situation considered here is that of single-stage sample selection. In practice, stratification is also sometimes used at one or more stages of a multi-stage design. For example, many address-based surveys use stratification at the first stage but not at the final stage (e.g. Lynn, 2009; Lynn & Lievesley, 1991). The precision gains due to stratification are generally likely to be more modest in such designs than in single-stage designs, and consequently the differences between ISS and ESS may also be more modest. A different situation is where stratification is used at the final stage of a multi-stage design. An example might be the selection of pupils within schools after first selecting a sample of schools. In this situation, precision gains can be considerable and it seems likely that the effects described in this article should apply. Indeed, the likely small sample size within each primary sampling unit is likely to result in ISS having even greater advantages, for the design weight reasons discussed in section 1 above.

References

- Breidt, F.J., Opsomer, J.D., & Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *Journal of the American Statistical Association*, 111:514, 822-833.
- Cochran, W. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, 17(2), 164-177.
- Cochran, W. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley.
- Fuller, Wayne A. (2009). *Sampling Statistics*. Hoboken, NJ: Wiley.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd edition). Hoboken, NJ: Wiley.

- Henry, G.T. (1990). *Practical Sampling*. Newbury Park, California: Sage.
- Kalton, G. (1983). *Introduction to Survey Sampling*. Sage Quantitative Applications in the Social Sciences Series, paper 35, Beverly Hills, California: Sage.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley.
- Lynn, P. (2009). Sample design for Understanding Society. *Understanding Society Working Paper 2009-01*, Colchester: University of Essex.
- Lynn, P., & Lievesley, L. (1991). *Drawing General Population Samples in Great Britain*. London: SCPR
- Madow, W. G., & Madow, L. G. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*. 15, 1–24.
- Rust, K., & Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*. 3, 69-81.
- Seth, G. R. (1966). On collapsing of strata. *Journal of the Indian Society of Agricultural Statistics*. 18, 1-3.
- Stehman, S.V. & Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomised variable probability systematic sampling. *Journal of the American Statistical Association*, 89(425), 30-43.
- Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79(388), 781-790.
- Wolter, K. M. (1985). *Introduction to Variance Estimation*. Berlin: Springer-Verlag.