DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 763–782, http://dx.doi.org/10.1515/JOS-2015-0044

Effects of Cluster Sizes on Variance Components in Two-Stage Sampling

Richard Valliant[1], *Jill A. Dever*[2], *and Frauke Kreuter*[3]

Determining sample sizes in multistage samples requires variance components for each stage of selection. The relative sizes of the variance components in a cluster sample are dramatically affected by how much the clusters vary in size, by the type of sample design, and by the form of estimator used. Measures of the homogeneity of survey variables within clusters are related to the variance components and affect the numbers of sample units that should be selected at each stage to achieve the desired precision levels. Measures of homogeneity can be estimated using standard software for random-effects models but the model-based intracluster correlations may need to be transformed to be appropriate for use with the sample design. We illustrate these points and implications for sample size calculation for two-stage sample designs using a realistic population derived from household surveys and the decennial census in the U.S.

Key words: Anticipated variance; measure of homogeneity; sample size calculation.

1. Introduction

Samples from finite populations are often selected in two or more stages for reasons of cost or operational necessity. For example, household samples in the U.S. may be selected through geographic areas like counties or groups of counties at the first stage, smaller areas like blocks at the second, and households at the last stage. Using multiple stages concentrates the sample in a limited number of areas, which is important when data are collected by personal interview at the respective households. In a survey of students, permission to conduct a survey may have to be obtained from school districts. Selecting districts first, then schools within sample districts, and finally a sample of students within a certain grade level within the school is operationally convenient and economical. Another example is a survey of employees in one or more business sectors, such as retail trade or services. Selecting establishments and then employees within establishments is a natural way of obtaining the sample.

Designing an efficient sample depends on estimating the contribution to the variance of an estimator associated with each stage of sampling. This involves estimating variance components for each stage that depend on the type of estimator and the types of units

[1] Universities of Michigan and Maryland – Joint Program for Survey Methodology, 1218 Lefrak Hall, College Park MD 20742 20742, U.S.A. Email: rvallian@umd.edu
[2] RTI International, Washington, District of Columbia, U.S.A. Email: jdever@rti.org
[3] University of Maryland – Joint Program for Survey Methodology, 1218 Lefrak Hall, College Park MD 20742 20742, U.S.A. Email: fkreuter@umd.edu
Acknowledgments: The authors are grateful to the editor and referees whose comments led to some important clarifications.

selected for the stage in question. This topic is covered in many standard texts on theoretical and applied sampling (Cochran 1977; Lohr 2010; Särndal et al. 1992). In the textbooks, formulae are available for the variance components for general sample designs; these formulae are usually specialized and simplified to obtain versions that facilitate sample size calculations. The relative sizes of the variance components are quite sensitive to how large the sampling units are at the different stages, how much variation there is among the sizes of the units, and the type of estimator used. Although this is implicit in the general variance-component formulae, this sensitivity is given little emphasis in most texts but can have a critical effect on calculated sample sizes and the achieved precision of estimators.

In certain applications, a survey designer has some control over the relative size of the sampling units. For example, in a household survey, extremely large metropolitan areas in the U.S., like New York or Chicago, are treated as strata and not as clusters of units. The first-stage units within such strata are groups of blocks defined by the U.S. Census Bureau for census taking and other survey data collections. Attempts are usually made to create groups by combining individual blocks so that the groups have about the same total population. In other applications, the survey designer has very little control over the units' sizes. In a school or establishment survey, the number of students or employees in each school or establishment is given. The survey must work with the existing sizes and combining these clusters further would not be meaningful.

In this article, we illustrate the effect of varying cluster sizes on design effects and measures of homogeneity within clusters for two-stage sampling. Section 2 discusses the variance-component formulae for two-stage sampling when the first-stage units are selected by either simple random sampling or probability proportional to size sampling. The effects of variation in cluster size are illustrated using an artificial, but realistic, population created using decennial census data from one county in the state of Maryland (Section 3). In Section 4, we describe how variance components from random-effects models can be used to calculate the measures of homogeneity needed for a two-stage sample. We summarize our results in the last section.

2. Background: Two-Stage Sampling

In this section, we present some background material for two-stage sampling and estimators of totals used for such designs. The units in the first stage of selection will be called primary sampling units (PSUs) or clusters. Units within PSUs are called elements and are the units for which data are collected. We use the following notation in the subsequent formulae:

$U =$ universe of PSUs
$M =$ number of PSUs in the universe
$U_i =$ universe of elements in PSU i
$N_i =$ number of elements in the population for PSU i
$N = \sum_{i \in U} N_i$, the total number of elements in the population
$\bar{N} = N/M$, the average number of elements per PSU
$m =$ number of sample PSUs
$n_i =$ number of sample elements in PSU i

s = set of sample PSUs
s_i = set of sample elements in PSU i
π_i = selection probability of PSU i
$\pi_{k|i}$ = selection probability of element k given PSU i was selected
y_k = value of a variable Y observed for element k
$\bar{y}_U = \sum_{i\in U}\sum_{k\in U_i} y_k/N$, the mean per element in the population
$\bar{y}_{Ui} = \sum_{k\in U_i} y_k/N_i$, the mean per element in the population in PSU i
$S_U^2 = \sum_{i\in U}\sum_{k\in U_i}(y_k - \bar{y}_U)^2/(N-1)$, the population variance of Y
$t_i = \sum_{k\in U_i} y_k$, the universe total of Y in PSU i
$t_U = \sum_{i\in U} t_i$, the universe total
$\bar{t}_U = t_U/M$, the average PSU total.

The π-estimator of a population total weights the value for element k inversely by its selection probability, π_k. Särndal et al. (1992, Result 4.3.1) give a formula for the variance of the π-estimator for a very general two-stage sample design. However, the general formula is not useful for designing samples because it involves joint selection probabilities of units at each stage that do not explicitly involve sample sizes. In this section, we present the variance formulae for different two-stage sample designs where the variance of the estimated total is simple enough for use in sample size calculation. In the first, PSUs are selected by simple random sampling; in the second, PSUs are selected with varying probabilities. For both designs, we assume that elements within PSUs are selected by simple random sampling. We follow the discussions of the π-estimator and probability with replacement (*pwr*) estimator of a total in Subsections 2.1 and 2.2 with the ratio estimator of a total in Subsection 2.3.

2.1. *Equal-Probability Sampling at Both Stages*

Suppose the first stage is a simple random sample selected without replacement (*srswor*) of m PSUs from a population of M PSUs, and the second stage is a sample of n_i elements selected by *srswor* from the population of N_i. As a shorthand, denote this design by *srs/srs.* The selection probability of element k in PSU i is $\pi_k = \pi_i\pi_{k|i} = (m/M)(n_i/N_i)$. The π-estimator of a population total is

$$\hat{t}_\pi = \frac{M}{m}\sum_{i\in s}\frac{N_i}{n_i}\sum_{k\in s_i} y_k = \frac{M}{m}\sum_{i\in s}\hat{t}_i \tag{1}$$

where $\hat{t}_i = (N_i/n_i)\sum_{k\in s_i} y_k$, the estimate of the total for PSU i with a simple random sample. The design variance, that is, the variance computed with respect to repeated sampling, of the π-estimator is

$$V(\hat{t}_\pi) = \frac{M^2}{m}\frac{M-m}{M}S_{U1}^2 + \frac{M}{m}\sum_{i\in U}\frac{N_i^2}{n_i}\frac{N_i - n_i}{N_i}S_{U2i}^2 \tag{2}$$

where $S_{U1}^2 = \frac{\sum_{i\in U}(t_i - \bar{t}_U)^2}{M-1}$, and $S_{U2i}^2 = \frac{\sum_{k\in U_i}(y_k - \bar{y}_{Ui})^2}{N_i - 1}$, the unit variance of Y among the elements in PSU i.

The first component of (2), the "between" term, can also be written as a function of the variance among means per element within the PSUs. However, expressing the between

term as a function of PSU totals as shown above allows a more intuitive explanation to be given for some subsequent results.

The relative variance (relvariance) of $\hat{t}_\pi$ is its variance divided by the square of the population total, $V(\hat{t}_\pi)/t_U^2$, and is especially useful for sample size calculation since the relvariance is unaffected by the scale of y. If the same number of sample elements, $n_i = \bar{n}$, is selected from each PSU, and the first-stage sampling fraction, m/M, and the second-stage sampling fraction, $\bar{n}/N_i$, are both small, the relvariance can be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{B^2}{m} + \frac{W^2}{m\bar{n}} \tag{3}$$

where $B^2 = S_{U1}^2/\bar{t}_U^2$ is the unit relvariance among PSU totals and $W^2 = M^{-1}\sum_{i\in U}\left(\frac{N_i}{N}\right)^2 \frac{S_{U2i}^2}{\bar{y}_U^2}$. A common simplification used in Cochran (1977) and Hansen et al. (1953a) is to further assume that all PSUs contain the same number of elements, that is, $N_i \equiv \bar{N}$, so that $W^2 = M^{-1}\sum_{i\in U} S_{U2i}^2/\bar{y}_U^2$. Roughly speaking, W^2 is an average relvariance per PSU with the per-PSU relvariance expressed as $S_{U2i}^2/\bar{y}_U^2$, that is, with the overall mean in the denominator. Expression (3) can be rearranged to give

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{\tilde{V}}{m\bar{n}} k[1 + \delta(\bar{n} - 1)] \tag{4}$$

where $\tilde{V} = S_U^2/\bar{y}_U^2$, $k = (B^2 + W^2)/\tilde{V}$, and $\delta = B^2/(B^2 + W^2)$, often referred to as a *measure of homogeneity*. With single-stage *srs* sampling of clusters from a population in which all clusters have the same size $\bar{N}$, δ is an *intraclass correlation* (see Cochran 1977, ch. 9; Lohr 2010 sec. 5.2.2) that can be computed as a type of Pearson correlation. With two-stage sampling, however, δ is not a correlation but still is related to the degree of homogeneity of elements within clusters. Note that an *fpc*, $1 - m\bar{n}/M\bar{N}$, is sometimes inserted into Expression (4) if the sampling fractions are not small, but this is an *ad hoc* addition that does not follow directly from rewriting (3).

The formula found in most textbooks is Expression (4) with $k = 1$, which comes from first writing the population variance of y as

$$(M\bar{N} - 1)S_U^2 = \sum_{i\in U} N_i\left(\frac{t_i}{Ni} - \frac{t_U}{M\bar{N}}\right)^2 + \sum_{i\in U}(N_i - 1)S_{U2i}^2.$$

Then, with some algebra (see Hansen et al. 1953a, sec. 6.6; Hansen et al. 1953b, sec. 6.5), it can be shown that when all clusters have the same size, $\bar{N}$, and both M and $\bar{N}$ are large,

$$\frac{S_U^2}{\bar{y}_U^2} = \frac{1 - M^{-1}}{1 - (M\bar{N})^{-1}}B^2 + \frac{1 - \bar{N}^{-1}}{1 - (M\bar{N})^{-1}}W^2 \quad \doteq B^2 + W^2 \tag{5}$$

that is, $k = 1$. In that case, (4) reduces to the relvariance of the estimated total in *srs*, $\tilde{V}/m\bar{n}$, times a design effect, $1 + \delta\,(\bar{n} - 1)$. The design-effect concept has been extended to more complex situations by Gabler et al. (1999), Lynn and Gabler (2005), and Park and Lee (2004).

The assumptions to obtain (5) that the number of population clusters and number of population elements per cluster are large is often reasonable, but assuming that the clusters all have the same size ($N_i = \bar{N}$) may not be. Although this special case is emphasized in texts like Kish (1965) and Lohr (2010), it can be misleading when clusters vary in size.

An alternative design for the second stage is to select elements at a fixed rate r within each cluster. The expected sample size in cluster i then is $n_i = rN_i$. This design might be preferred to *srs*/*srs* with a fixed-size sample at the second stage because all sample elements will have the same weight, $M/(mr)$. There are different ways of selecting such a sample. Bernoulli sampling is one; systematic sampling from a randomly ordered list is another. In the latter design, which we use here, the achieved sample size is either the integer floor or ceiling of rN_i. This type of systematic sample can reasonably be treated as *srswor* when the list is randomly ordered. Substituting $n_i = rN_i$ in (2), dividing by t_U^2, and using the equivalent expressions for the population total, $t_U = M\bar{t}_U = M\bar{N}\bar{y}_U$, gives the approximate relvariance as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{B^2}{m} + \frac{\tilde{W}^2}{m\bar{n}^*} \tag{6}$$

where B^2 is the same quantity as in (3), $\bar{n}^* = r\bar{N}$, and $\tilde{W}^2 = M^{-1}\sum_{i\in U}\frac{N_i}{N}\frac{S_{U2i}^2}{\bar{y}_U^2}$. There is some randomness in the achieved second-stage sample size when rN_i is not an integer. Note that $\bar{n}^*$ is an average cluster sample size in the sense that the average sample size over all clusters in the universe is $\sum_{i\in U} n_i/M = r\bar{N}$. The corresponding value of the measure of homogeneity is $\tilde{\delta} = B^2/(B^2 + \tilde{W}^2)$. The relvariance in (6) can also be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{\tilde{V}}{m\bar{n}^*}\tilde{k}[1 + \tilde{\delta}(\bar{n}^* - 1)] \tag{7}$$

where $\tilde{k} = (B^2 + \tilde{W}^2)/\tilde{V}$. Note that (7) reduces to the usual textbook formula if $\tilde{k} = 1$, which requires that $S_U^2/\bar{y}_U^2 \doteq B^2 + \tilde{W}^2$. Since the design with a fixed sampling rate at the second stage may be more common in practice than one with a common $\bar{n}$ when the design is *srs*/*srs*, we concentrate on it in the numerical illustrations.

Expressions (4) or (7) are useful for sample size calculation since the number of sample PSUs, m, sample elements per PSU, $\bar{n}$, or the within-PSU rate, $r = \bar{n}^*/\bar{N}$, are explicit in the formula. Expressions like (4) and (7) often seem to be treated as if they apply regardless of how the samples of PSUs and elements within PSUs are selected. If, for example, a probability proportional to size (*pps*) sample of PSUs is selected, (4) and (7) do not reflect that feature. In Subsection 2.2 we therefore give a relvariance that is similar in form to (4) and (7) but is appropriate for *pps* sampling of PSUs.

When designing samples, practitioners sometimes use rough rules of thumb for values of $\tilde{\delta}$ (or δ), say $\tilde{\delta} \leq 0.10$, based on how "alike" elements within PSUs are thought to be. However, the form of S_{U1}^2 and, therefore, B^2 implies that the size of $\tilde{\delta}$ (or δ) can also be determined by the relative variability of the cluster totals, t_i. As we will illustrate, one way in which $\tilde{\delta}$ can be large is by having clusters that vary in size.

2.2. *Varying Probabilities at the First Stage*

Variances of estimators in designs more complicated than simple random sampling at each stage can also be written as a sum of components. However, the most general of these have limited value in determining sample sizes (e.g., see Särndal et al. 1992, result 4.3.1).

A more useful formulation is the case where PSUs are selected with varying probabilities but with replacement (*ppswr*), and the sample within each PSU is selected by

srswor. We refer to this design as *ppswr/srs*. With-replacement designs may not often be used in practice but have simple variance formulae, which makes them useful for sample size calculation. The probability with-replacement (*pwr*) estimator of a total is

$$\hat{t}_{pwr} = \frac{1}{m}\sum_{i\in s}\frac{\hat{t}_i}{p_i}$$

where $\hat{t}_i$ was defined in Subsection 2.1 and p_i is the one-draw selection probability of PSU i. The variance of $\hat{t}_{pwr}$ is (Cochran 1977, 308-310)

$$V(\hat{t}_{pwr}) = \frac{1}{m}\sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i\in U}\frac{{N_i}^2}{mp_in_i}\left(1 - \frac{n_i}{N_i}\right)S^2_{U2i}. \tag{8}$$

Making the assumption that $\bar{n}$ elements are selected in each PSU and that $\bar{n}/N_i$ is negligible, the variance reduces to

$$V(\hat{t}_{pwr}) = \frac{S^2_{U1(pwr)}}{m} + \frac{1}{m\bar{n}}\sum_{i\in U}\frac{N_i^2S^2_{U2i}}{p_i}$$

where, in this case, $S^2_{U1(pwr)} = \sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2$ and S^2_{U2i} is defined for Expression (2). Dividing this by t_U^2 and simplifying, we obtain the relvariance of $\hat{t}_{pwr}$ as, approximately,

$$\frac{V(\hat{t}_{pwr})}{t_U^2} \doteq \frac{B_*^2}{m} + \frac{W_*^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}}k_*[1 + \delta_*(\bar{n} - 1)] \tag{9}$$

with $B_*^2 = \frac{S^2_{U1(pwr)}}{t_U^2}$, $W_*^2 = \frac{1}{t_U^2}\sum_{i\in U} N_i^2\frac{S^2_{U2i}}{p_i}$, $k_* = \left(B_*^2 + W_*^2\right)/\tilde{V}$, and $\delta_* = B_*^2/\left(B_*^2 + W_*^2\right)$. If $k_* = 1$, then (9) has the interpretation of an *srs* relvariance times a design effect, $1 + \delta_*(\bar{n} - 1)$.

The approximation in (9) does depend on the sampling fraction of elements within each sample cluster being small, and more importantly on using the with-replacement variance formula for the first stage. On the other hand, it does allow the number of population elements per cluster to vary, which is an important feature to account for in some populations.

A special case of the design above would be $p_i = N_i/N$, that is, probability proportional to the size of cluster i. If the weight of cluster i in a with-replacement sample of m clusters is $N/(mN_i)$ and an equal-probability sample of $\bar{n}$ elements are selected in each cluster, the sample is "self-weighting" as the weight of each sample element in the *pwr* estimator is the same: $(N/mN_i)(N_i/\bar{n}) = N/(m\bar{n})$. This combination of design and weighting method is common in household surveys where a practical goal is often to have an equal workload in each cluster and limit variation in weights.

A more general point to note is that the measures of homogeneity in (4), (7), and (9) depend on both the sample design and the estimator being used. This is because the decomposition of the variance of an estimator depends on both. A different decomposition would be needed for, say, the general regression (GREG) estimator of a total or an estimator of a mean that uses an estimate of N in its denominator.

2.3. *Ratio Estimator of a Total*

The π-estimator of a total may be inefficient in some designs compared to alternatives like the ratio estimator or a GREG estimator. In this section, we present the variance-component formula in the *srs/srs* design for the ratio estimator defined as

$$\hat{t}_R = \hat{t}_\pi \frac{N}{\hat{N}_\pi}$$

where $\hat{N}_\pi$ is the π-estimator of the number of elements in the population, N, defined as $\hat{N}_\pi = M \sum_s N_i/m$. (Note that, in probability proportional to N_i sampling with $p_i = N_i/N$, the estimated total number of elements is $\hat{N}_{pwr} = m^{-1} \sum_{i \in s} \hat{t}_i/p_i = N$, and there is no gain from ratio estimation.) Assuming that the sample size of clusters m is large and using a first-order linear approximation,

$$\hat{t}_R - t_U = \hat{t}_\pi - \bar{y}_U \hat{N}_\pi + O_p(M/m) \doteq \frac{M}{m} \sum_s \hat{t}_{zi} \tag{10}$$

where $\hat{t}_{zi} = N_i \sum_{k \in s_i} z_k/n_i$ with $z_k = y_k - \bar{y}_U$. Expression (10) follows from assuming that $M^{-1}(\hat{t}_\pi - t_U)$ and $M^{-1}(\hat{N}_\pi - N)$ are $O_p(m^{-1/2})$ as they would be if $m^{1/2}(\hat{t}_\pi - t_U)/M$ and $m^{1/2}(\hat{N}_\pi - N)/M$ had asymptotic standard normal distributions. In that case the remainder in the first-order Taylor series approximation to $M^{-1}(\hat{t}_R - t_U)$ is $O_p([m^{-1/2}]^2) = O_p(m^{-1})$ (see Wolter 2007, Theorem 6.2.2). Under those conditions, $\hat{t}_\pi - \bar{y}_U \hat{N}_\pi = O_p(M/m^{1/2})$, that is, a higher order than the remainder term in (10). Approximation (10) has the same form as the π-estimator in (1). Consequently, a variance-component formula analogous to (2) and a relvariance formula similar to (3) can be derived. In particular,

$$V(\hat{t}_R) \doteq \frac{M^2}{m} \frac{M-m}{M} S^2_{Uz1} + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S^2_{U2zi}$$

with $S^2_{Uz1} = (M-1)^{-1} \sum_{i \in U} (t_{zi} - \bar{t}_{Uz})^2$, and $S^2_{U2zi} = (N_i - 1)^{-1} \sum_{k \in U_i} (z_k - \bar{z}_{Ui})^2$ where $t_{zi} = \sum_{k \in U_i} z_k$, $\bar{t}_{Uz} = \sum_{i \in U} t_{zi}/M$, and $\bar{z}_{Ui} = \sum_{k \in U_i} z_k/N_i$. Assuming that the *fpc*s, $(M-m)/M$ and $(N_i - n_i)/N_i$, are approximately 1 and that the sample size in PSU i is rN_i, the relvariance formula is

$$V(\hat{t}_R) \doteq \frac{B_z^2}{m} + \frac{\tilde{W}_z^2}{m\bar{n}^*} = \frac{\tilde{V}}{m\bar{n}^*} k_z [1 + \delta_z(\bar{n}^* - 1)] \tag{11}$$

where $B_z^2 = S^2_{Uz1}/\bar{t}_U^2$, $\tilde{W}_z^2 = M^{-1} \sum_{i \in U} (N_i/\bar{N}) S^2_{U2zi}/\bar{y}_U^2$, $k_z = \left(B_z^2 + \tilde{W}_z^2\right)/\tilde{V}$, and $\delta_z = B_z^2/\left(B_z^2 + \tilde{W}_z^2\right)$. Compared to the (*srs*/*srs*, π-estimator) strategy the ratio estimator can reduce the measure of homogeneity, leading to more precise estimators as illustrated in Example 4 of Section 3.

3. Examples of Variance Components and Measures of Homogeneity

We created an example population based on U.S. Census counts from the year 2000 for Anne Arundel County in the state of Maryland and refer to this data set as `MDarea.pop`. The population is also included in the R package `PracTools` (Valliant et al. 2013, 2015). The population contains three continuous and two binary variables denoted by `y1, y2, y3, ins.cov,` and `hosp.stay`, respectively. The variables are generated using models, since individual-person data for small geographic areas is suppressed in the actual census for reasons of confidentiality. The variables in `MDarea.pop` were created by fitting models for several variables in the 2001-2002 National Health and Nutrition

Table 1. Descriptive statistics for the Maryland area population

	Tract population	BG population	y1	y2	y3
Minimum	86	52	−62.7	−2.9	32.6
1st quartile	2,728	780	18.7	2.3	66.7
Median	4,132	1,240	50.8	5.4	81.4
Mean	4,253	1,316	69.7	7.7	87.5
3rd quartile	5,684	1,732	104.4	10.7	101.2
Maximum	13,579	4,744	1163.7	101.1	479.2
Population CV	0.51	0.58	1.21	1.01	0.34

CV = coefficient of variation.

Examination Survey (Center for Disease Control and Prevention 2009) and 2003 National Health Interview Survey (Center for Disease Control and Prevention 2012) data sets to obtain regression means that depended on whether a person was Hispanic and on the person's gender and age. Person-level values were created using random-effects models that had error terms for tracts, block groups, and persons. The three continuous variables (`y1, y2, y3`) are positively skewed with mean values based on models for body weight, body mass index, and systolic blood pressure (although the scales of the generated variables do not match those of these physical measurements). The binary variables, `ins.cov` and `hosp.stay`, are based on the rates of insurance coverage and overnight hospital stay in a twelve-month period.

The geographic divisions used in this data set are tracts and block groups, which are geographic areas defined by the Census Bureau (U.S. Census Bureau 2011). Tracts are constructed to have a desired population size of 4,000 people. Block groups (BGs) are smaller, with a target size of 1,500 people. However, the sizes of both tracts and BGs vary because the Census Bureau also attempts to limit the geographic area covered by a BG. Counts of persons in the data set are the same for most tracts and BGs as in the 2000 Census; exceptions are five BGs that were augmented to have at least 50 persons each.

The example population contains 403,997 persons, 95 tracts, and 307 BGs. The proportion of persons with insurance coverage is 0.793; the proportion with a hospital stay in the prior twelve months is 0.072. Descriptive statistics for other variables are given in Table 1.

Because the tracts and BGs in the Maryland population are extremely variable in size, we created two other variables called `PSU` and `SSU` to demonstrate the effect of having equal-sized units. Each artificial `PSU` has approximately the same number of persons; likewise the `SSUs` were created to have about the same number of persons. The `PSUs` and `SSUs` were formed after sorting the file by tract and BG within tract, thus retaining geographic proximity of persons grouped together. Each `PSU` has about 5,050 persons while an `SSU` has about 1,010. Although the assumption of equal PSU size made to obtain (5) or to set $\tilde{k} = 1$ may seem innocuous, it is far from that, as we will illustrate below.

We use the Maryland population to illustrate the effects of using different sizes of primary and secondary sampling units on the measures of homogeneity for two-stage sampling. In all of the examples, calculations are made assuming that the entire population is in hand. This means that the theoretical values in the preceding formulae can be evaluated rather than estimated from a sample as would be required in practice.

When examining the effects of varying unit sizes, working with a population is an advantage as the complication of sampling variability is eliminated.

Example 1. Between- and within-variance components in srs/srs design. Using the variables in the Maryland population, we computed the unit relvariance of each variable $\left(S_U^2/\bar{y}_U^2\right)$, $B^2 + \tilde{W}^2$ and $\tilde{k}$ for comparison, and $\tilde{\delta} = B^2/(B^2 + \tilde{W}^2)$ for the *srs/srs* design and the *pwr*-estimator. (Note that the π-estimator and *pwr*-estimator in *srs/srs* have the same form when the first stage is selected with replacement. In the examples, we will refer to the (*srs/srs*, *pwr*-strategy).) First, the results are shown in Table 2 using the `PSU` and `SSU` variables as clusters. Values of $\tilde{\delta}$ range from 0.001 to 0.079 when `PSUs` are clusters. Deltas are somewhat larger when `SSUs` are clusters, reflecting the common phenomenon that smaller geographic areas are somewhat more homogeneous than large ones in household populations. The third through fifth columns show that the approximation that $S_U^2/\bar{y}_U^2 \doteq B^2 + W^2$ works well in this case.

Next, to illustrate the dramatic effect that varying sizes of clusters can have, in Table 3 we present the same statistics as above using tracts and BGs within tracts as clusters. Values of δ range from 0.023 to 0.730 when tracts are clusters. When BGs are used as clusters, $\tilde{\delta}$s range from 0.032 to 0.791. The measures of homogeneity increase substantially when tracts or BGs are the first-stage clusters. For example, when `PSUs` are clusters, $\delta = 0.005$ for `y1` but is 0.152 when tracts are clusters. This is almost entirely due to the increase in the between-variance component, B^2, when units with highly variable sizes are used. For example, $B^2 = 0.0079$ for `y1` when `PSU` is a cluster, but is 0.2605 when tract is a cluster. The third through fifth columns in Table 3 show that the approximation $S_U^2/\bar{y}_U^2 \doteq B^2 + \tilde{W}^2$ does not work well when either tracts or BGs are clusters. This again is due to the clusters not all having the same size. This implies that when making advance estimates of the relvariance of an estimated total, $\tilde{k}$ cannot be safely set to 1 in (7) when PSUs vary in size.

Example 2. Effect of incorrect measures of homogeneity on achieved precision. If incorrect values of the measure of homogeneity are used to compute sample sizes, the sample can be much less efficient than anticipated. This example looks at the effect of

Table 2. Variance components and measures of homogeneity in the Maryland population using `PSUs` and `SSUs` as clusters with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage

	B^2	$\tilde{W}^2$	$S_U^2/\bar{y}_U^2$	$B^2 + \tilde{W}^2$	$\tilde{k}$	$\tilde{\delta}$
`PSUs` as clusters						
`y1`	0.0079	1.4553	1.4627	1.4631	1.0003	0.005
`y2`	0.0069	1.0097	1.0163	1.0166	1.0003	0.007
`y3`	0.0090	0.1048	0.1136	0.1137	1.0012	0.079
`ins.cov`	0.0012	0.2599	0.2611	0.2611	1.0003	0.005
`hosp.stay`	0.0175	12.8831	12.8979	12.9006	1.0002	0.001
`SSUs` as clusters						
`y1`	0.0365	1.4277	1.4627	1.4642	1.0010	0.025
`y2`	0.0169	1.0004	1.0163	1.0173	1.0010	0.017
`y3`	0.0184	0.0954	0.1136	0.1137	1.0012	0.161
`ins.cov`	0.0032	0.2581	0.2611	0.2613	1.0010	0.012
`hosp.stay`	0.0558	12.8549	12.8979	12.9107	1.0010	0.004

Table 3. Variance components and measures of homogeneity in the Maryland population using tracts and block groups as clusters with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage

	B^2	$\tilde{W}^2$	$S_U^2/\bar{y}_U^2$	$B^2+\tilde{W}^2$	$\tilde{k}$	$\tilde{\delta}$
Tracts as clusters						
y1	0.2605	1.4539	1.4627	1.7144	1.1720	0.152
y2	0.2687	1.0058	1.0163	1.2745	1.2540	0.211
y3	0.2707	0.1001	0.1136	0.3707	3.2634	0.730
ins.cov	0.2624	0.2593	0.2611	0.5217	1.9985	0.503
hosp.stay	0.3078	12.8786	12.8979	13.1864	1.0224	0.023
Block groups as clusters						
y1	0.3489	1.4478	1.4627	1.7967	1.2283	0.194
y2	0.3485	0.9994	1.0163	1.3479	1.3263	0.259
y3	0.3492	0.0926	0.1136	0.4418	3.8887	0.791
ins.cov	0.3408	0.2574	0.2611	0.5982	2.2916	0.570
hosp.stay	0.4246	12.8567	12.8979	13.2813	1.0297	0.032

using $\tilde{\delta}$s computed as if clusters all had the same size when clusters actually vary. Suppose that the costs which vary with the number of sample clusters and elements can be written as $C = C_1 m + C_2 m\bar{n}$ where C_1 is the cost per cluster and C_2 is the cost per sample element. If the budget for variable costs is fixed at C and the relvariance is given by (7), the optimal numbers of elements and clusters are (cf. Hansen et al. 1953a sec. 16.6):

$$\bar{n}_{opt} = \sqrt{\frac{C_1}{C_2}\frac{1-\tilde{\delta}}{\tilde{\delta}}} \quad \text{and} \quad m_{opt} = \frac{C}{C_1 + C_2\bar{n}_{opt}}. \tag{12}$$

(The results in (12) hold for both $\tilde{k} = 1$ and a general value of $\tilde{k}$.) In this example, the cost assumptions are $C = \$100{,}000$, $C_1 = \$1,000$, and $C_2 = \$100$. Suppose that the sample sizes in (12) are computed using the $\tilde{\delta}$s in Table 2, assuming that clusters are PSUs or SSUs (i.e., clusters with the same size). These values of $\bar{n}_{opt}$ and m_{opt} are shown in Table 4 using the $\tilde{\delta}$s and values of $\tilde{k}$ from Table 2. The estimated coefficients of variation (CVs), that is, the square root of the estimated relvariances that would be obtained with the equal-size cluster $\tilde{\delta}$s, are in the fourth column, assuming that $\tilde{V} = 1$. Suppose that the correct $\tilde{\delta}$s and $\tilde{k}$s are in reality those in Table 3, which account for varying cluster sizes. The actual CVs that would be obtained with these $\tilde{\delta}$s are also shown in the sixth column of Table 4, again assuming that $\tilde{V} = 1$. The ratio of actual CVs with $\tilde{\delta}$s from Table 3 to the estimated CVs with $\tilde{\delta}$s from Table 2 range from 1.5 to 6.3. In other words, the actual CVs range from 50% to 530% higher than estimated because varying cluster sizes increase the measures of homogeneity and values of $\tilde{k}$. This implies that if the correct $\tilde{\delta}$s and $\tilde{k}$s were used, more clusters and fewer elements per cluster should be selected than the m_{opt} and $\bar{n}_{opt}$ values in Table 4.

Example 3. ***ppswr* at first stage, *srs* at second**. This example repeats the calculations in Example 1 for the variables in the Maryland area population but with a different sample design. Assume that clusters will be selected proportional to the count of persons N_i in each cluster and that an *srs* with a small sampling fraction is selected in each sample cluster, that is, a particular case of *ppswr/srs*. Table 5 shows the values of B_*^2, W_*^2, and δ_*

Table 4. Loss of precision from using incorrect measures of homogeneity with an srs/srs design, the pwr-estimator, and a fixed sampling rate at the second stage

	$\tilde{\delta}$ with equal-size clusters (Table 2)	$\bar{n}_{opt}$	m_{opt}	Estimated CV (%)	$\tilde{\delta}$ with varying-size clusters (Table 3)	Actual CV (%)	Ratio: Actual to estimated CVs
Tracts as clusters							
y1	0.005	43	19	3.9	0.123	10.3	2.7
y2	0.007	38	21	4.0	0.173	11.8	3.0
y3	0.079	11	48	5.8	0.681	22.6	3.9
ins.cov	0.005	47	18	3.8	0.443	24.1	6.3
hosp.stay	0.001	84	11	3.5	0.018	5.8	1.6
Block groups as clusters							
y1	0.024	20	33	4.7	0.151	9.3	2.0
y2	0.016	25	29	4.4	0.206	11.5	2.6
y3	0.160	7	58	6.9	0.740	23.4	3.4
ins.cov	0.011	30	25	4.3	0.498	22.6	5.3
hosp.stay	0.003	58	15	3.8	0.023	5.6	1.5

Table 5. Variance components and measures of homogeneity in the Maryland population using PSUs and SSUs as clusters with a ppswr/srs design and the pwr-estimator

	B_*^2	W_*^2	k_*	δ_*
PSUs as clusters				
y1	0.0078	1.4553	1.0002	0.005
y2	0.0068	1.0097	1.0002	0.007
y3	0.0088	0.1048	1.0002	0.078
ins.cov	0.0012	0.2599	1.0002	0.005
hosp.stay	0.0173	12.8831	1.0002	0.001
SSUs as clusters				
y1	0.0364	1.4277	1.0010	0.025
y2	0.0169	1.0004	1.0010	0.017
y3	0.0183	0.0954	1.0008	0.161
ins.cov	0.0032	0.2581	1.0010	0.012
hosp.stay	0.0557	12.8549	1.0010	0.004

when PSUs and SSUs are clusters. Because each PSU and SSU was formed to have almost the same number of persons, the values in Table 5 are virtually the same as the *srs/srs* results in Table 2.

Table 6 shows the results when tracts and BGs are used as clusters. With the *ppswr/srs* design, the between term is much smaller than the within term compared to the results in Example 1. This is true whether PSU and SSU are used as clusters or tracts and BGs are used. For example, with y1, $\delta = 0.152$ when tracts are clusters in the *srs/srs* design (Table 3). However, $\delta_* = 0.006$ for y1 with tracts as clusters in the *ppswr/srs* design in Table 6. The measures of homogeneity for other variables are also substantially less in Table 6 than in Table 3.

When clusters are selected by *srs*, S_{U1}^2 is the variance of the cluster totals around the average cluster total. In contrast, with *pps* sampling of clusters, $S_{U1(pwr)}^2$ is the variance of the estimated population totals, t_i/p_i around the population total, t_U. When clusters are selected with probability proportional to N_i, $t_i/p_i = N_i\bar{y}_{Ui}/(N_i/N) = N\bar{y}_{Ui}$. If these

Table 6. Variance components and measures of homogeneity in the Maryland population using tracts and BGs as clusters with a ppswr/srs design and the pwr-estimator

	B_*^2	W_*^2	k_*	δ_*
Tracts as clusters				
y1	0.0092	1.4539	1.0002	0.006
y2	0.0107	1.0058	1.0002	0.011
y3	0.0136	0.1001	1.0002	0.119
ins.cov	0.0018	0.2593	1.0002	0.007
hosp.stay	0.0223	12.8786	1.0002	0.002
Block groups as clusters				
y1	0.0160	1.4478	1.0007	0.011
y2	0.0176	0.9994	1.0007	0.017
y3	0.0211	0.0926	1.0006	0.186
ins.cov	0.0039	0.2574	1.0007	0.015
hosp.stay	0.0509	12.8567	1.0008	0.004

Table 7. Variance components and measures of homogeneity in the Maryland population using tracts and block groups as clusters with an srs/srs design, a fixed rate at the second stage, and a ratio estimator of a total

	B_z^2	$\tilde{W}_z^2$	k_z	δ_z
Tracts as clusters				
y1	0.0093	1.8390	1.2636	0.005
y2	0.0114	1.2662	1.2571	0.009
y3	0.0143	0.1253	1.2285	0.102
ins.cov	0.0021	0.3260	1.2568	0.007
hosp.stay	0.0265	16.3171	1.2672	0.002
Block groups as clusters				
y1	0.0193	1.9499	1.3462	0.010
y2	0.0223	1.3338	1.3344	0.017
y3	0.0271	0.1220	1.3127	0.182
ins.cov	0.0052	0.3426	1.3324	0.015
hosp.stay	0.0681	17.2695	1.3442	0.004

1-cluster estimates of the population total are fairly accurate, as they are here, the B^2 term can be quite small. This leads to much smaller values of the measure of homogeneity in *pps* sampling of clusters, implying that the effect of clustering is less important in this population for a design that selects clusters with probabilities proportional to their population counts.

Practitioners habitually gravitate toward *pps* sampling of clusters rather than *srs*. This example makes it clear why this choice is often a good one.

Example 4. *srs/srs* design with ratio estimator of the total. Next, we consider whether use of the ratio estimator of the total in an *srs/srs* design reduces the effects of using clusters with varying sizes. Table 7 displays results for the variance components, B_z^2 and $\tilde{W}_z^2$, k_z, and δ_z defined in Subsection 2.3 when tracts or block groups are used as clusters. The values of δ_z in Table 7 are much lower than those of $\tilde{\delta}$ in Table 3, implying that use of a ratio estimator in *srs/srs* substantially reduces the effect of clustering compared to using the *pwr*-estimator. The values of δ_z are very close to those of δ_* in Table 6 for the *ppswr/srs* design combined with the *pwr*-estimator. However, the values of k_* in Table 6 are all near 1 while k_z in Table 7 ranges from about 1.23 to 1.35. Thus, for a given number of sample clusters m and elements $\bar{n}$ in the *ppswr/srs* case or $\bar{n}^*$ in the *srs/srs* fixed rate case, the (*ppswr/srs*, *pwr*-estimator) strategy will be more efficient than the (*srs/srs*, ratio estimator) strategy. For example, suppose that BGs are clusters, the total of y1 is estimated and $\bar{n} = \bar{n}^* = 50$. If *srs/srs* and the *pwr*-estimator is used, then $\tilde{k}[1 + \tilde{\delta}\,(\bar{n} - 1)] = 1.2283[1 + 0.194(50 - 1)] = 12.905$ using the figures in Table 3. For the (*ppswr/srs*, *pwr*-estimator) strategy, $k_*[1 + \delta_*(\bar{n} - 1)] = 1.007[1 + 0.011(50 - 1)] = 1.550$ using the values in Table 6. For (*srs/srs*, ratio estimator) the corresponding value is $k_z[1 + \delta_z\,(\bar{n}^* - 1)] = 1.3462[1 + 0.010(50 - 1)] = 2.006$ using the figures in Table 7. Accordingly, the relvariance for (*srs/srs*, *pwr*-estimator) is 8.33 (12.905/1.550) times as large as that of (*ppswr/srs*, *pwr*-estimator), while the relvariance of (*srs/srs*, ratio estimator) is 1.29 (2.006/1.55) times as large. Using the ratio estimator in *srs/srs* is much better than using the *pwr*-estimator, but still is considerably less efficient than the (*ppswr/srs*, *pwr*-estimator) strategy.

4. Estimating Variance Components Using Anticipated Variances

In normal circumstances, only a sample is available from a population and variance components must be estimated. Design-based estimators can be found in Särndal et al. (1992, sec. 4.3.2) and will not be covered here. As noted earlier, the general formulae for estimation of variance components are specialized, complex, and difficult to use in practice. Being able to use the software routines that are available for variance-component estimation would be a real advantage if they estimate the components properly. The best of these routines use algorithms designed to handle a variety of numerical problems that are hard to anticipate in practice. Searle et al. (1992) review the methods available, including minimum variance quadratic unbiased estimation (MIVQUE0), maximum likelihood, and restricted maximum likelihood (REML). Note that these estimates are derived through a specified model and not a particular sample design.

Model variance components can be introduced by using an anticipated variance (Isaki and Fuller 1982) defined as

$$AV(\hat{t}) = E_M[E_\pi(\hat{t} - t_U)^2] - [E_M E_\pi(\hat{t} - t_U)]^2$$

where E_M is the theoretical expectation (or average) with respect to the specified population model and E_π is the (design-based) expectation under repeated sampling. If the estimator is design-unbiased or approximately so, then the anticipated variance is $AV(\hat{t}) = E_M[\text{var}_\pi(\hat{t} - t_U)]$ since $E_\pi(\hat{t}) = t_U$. Thus the model expectation of a formula like (3) or (4) can be computed, resulting in a formula that includes model variance components that can be estimated using standard software. An additional advantage to this approach is the clarification of the key role that PSU and SSU sizes play in determining the measures of homogeneity. Expressions (4), (7), (9), and (11) contain measures of homogeneity, δ, $\tilde{\delta}$, δ_*, and δ_z, respectively, that are critical determinants of sample sizes. However, δ, $\tilde{\delta}$, δ_*, and δ_z are not equal to the model correlation of elements in the same cluster, except in some special circumstances, as we will illustrate.

Examples in the literature of using model variance-component estimates in survey design seem limited, even though practitioners often use the technique. A few examples are Chromy and Myers (2001); Hunter et al. (2005); Judkins and Van de Kerckhove (2003); and Waksberg et al. (1993). We demonstrate the basic approach using a random-effects model.

In a clustered population, the simplest model to consider is one with common mean and random effects for clusters and elements:

$$y_k = \mu + \alpha_i + \varepsilon_{ik}, \quad k \in U_i, \tag{13}$$

with $\alpha_i \sim (0, \sigma_\alpha^2)$, $\varepsilon_{ik} \sim (0, \sigma_\varepsilon^2)$, and the errors being independent. The model correlation of any two elements in the same cluster is

$$corr(y_k, y_{k'}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \equiv \rho. \tag{14}$$

The model expectation of the design variance can be computed under this model, but for sample size calculation, only the approximate expectation of the between- and within-variance components for two-stage sampling are needed. First, take the case of an *srs/srs*

design and the *pwr*-estimator where a common sampling rate r is used in all clusters. The approximate model expectations are needed for $B^2 = S_{U1}^2/\bar{t}_U^2$ and $\tilde{W}^2 = M^{-1}\sum_{i\in U}\frac{N_i}{\bar{N}}\frac{S_{U2i}^2}{\bar{y}_U^2}$ in (6). After some algebra, the model expectations of S_{U1}^2 and S_{U2i}^2 defined below (2) are:

$$E_M\left(S_{U1}^2\right) \doteq \left(\sigma_\alpha^2 + \mu^2\right)S_N^2 + \bar{N}^2\sigma_\alpha^2 + \sigma_\varepsilon^2$$

$$E_M\left(S_{U2i}^2\right) = \sigma_\varepsilon^2$$

where $\bar{N} = \sum_{i\in U} N_i/M$ is the average number of elements per cluster, and $S_N^2 = \sum_{i\in U}(N_i - \bar{N})^2/(M-1)$ is the population variance of the PSU sizes, N_i. We also assume that M is large so that $M - 1 \doteq M$. Assuming that the expectation of a ratio, like $S_{U1}^2/\bar{t}_U^2$, is approximately the ratio of the expectations, the model expectation of the measure of homogeneity $\tilde{\delta}$ in (7) is

$$E_M(\tilde{\delta}) \doteq \frac{\left(\sigma_\alpha^2 + \mu^2\right)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}^2}{\left(\sigma_\alpha^2 + \mu^2\right)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2(1 + \bar{N}^{-2})} \tag{15}$$

where $\nu_N^2 = S_N^2/\bar{N}^2$ is the relvariance of the N_is. If $N_i = \bar{N}$, that is, all the clusters are the same size, then $\nu_N^2 = 0$ and (15) reduces to

$$E_M(\tilde{\delta}) \doteq \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2(1 + \bar{N}^{-2})}. \tag{16}$$

If, in addition, $\bar{N}$ is sufficiently large for $\sigma_\varepsilon^2/\bar{N}^2$ to be negligible compared to σ_α^2, then $E_M(\tilde{\delta})$ does equal the model correlation in (14). However, when clusters vary in size, (15) will be a closer approximation to the measure of homogeneity needed for sample size calculation.

The result for δ in (4) is very similar. The model expectation of δ is equal to (15) but $1 + \bar{N}^{-2}$ in the denominator is replaced with $1 + \nu_N^2 + \bar{N}^{-2}$. Numerically, the model expectation of $\tilde{\delta}$ will be somewhat larger than that of δ. For δ_* and δ_z the calculations would have to be specialized to be appropriate to the forms of B_*^2, W_*^2, B_z^2, and $\tilde{W}_z^2$ used in the definitions of those measures of homogeneity. We consider only δ_* below.

Next, consider the *ppswr/srs* design where the one-draw probability of cluster i is proportional to its number of elements, that is, $p_i = N_i/M\bar{N}$. The model expectation of $S_{U1(pwr)}^2$ is

$$E_M\left(S_{U1(pwr)}^2\right) = (M\bar{N})^2\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + M^2\bar{N}\sigma_\varepsilon^2.$$

The model expectation of δ_* is then approximately

$$E_M(\delta_*) \doteq \frac{\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \frac{\sigma_\varepsilon^2}{\bar{N}}}{\sigma_\alpha^2\left[1 - \frac{1}{M}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \sigma_\varepsilon^2\left(1 + \frac{1}{\bar{N}}\right)} \tag{17}$$

If $N_i \equiv \bar{N}$, selecting PSUs with probability proportional to the sizes N_i is the same as equal-probability sampling. In that case, (17) reduces to approximately the same form as

Table 8. Intracluster correlations ρ from (14) under a simple random-effects model

	Values of model intracluster correlation ρ			
	Unit used for clusters			
Variable	PSUs	SSUs	Tracts	Block groups
y1	0.005	0.024	0.008	0.012
y2	0.007	0.016	0.013	0.017
y3	0.079	0.161	0.148	0.191
ins.cov	0.004	0.011	0.008	0.014
hosp.stay	0.001	0.003	0.002	0.003

(16), which is essentially equal to the model correlation in (14) when $N_i \equiv \bar{N}$ and the average cluster size is large.

Example 5. Anticipated variance components in two-stage sampling from a model. A number of software routines are available for estimating variance components – the R package lme4 (Bates et al. 2011), the SAS® procedure proc mixed, and the xtmixed routine in Stata® are examples. We used the function lmer in lme4 to estimate the variance components for the model in (13) and the corresponding intracluster correlations in (14). The type of sample design used (*srs/srs* or *ppswr/srs*) does not affect these estimates, since they are based strictly on the model in (13). The results for all variables using PSUs, SSUs, tracts, and BGs as clusters are shown in Table 8. The estimates for ρ when PSUs and SSUs are clusters are almost the same as the values of $\tilde{\delta}$ in Table 2 where *srs* is used at each stage. But when tracts and BGs of varying sizes are used as the clusters, the ρs in Table 8 are very different and much smaller than the $\tilde{\delta}$s in Table 3. As noted above, the design-based formula for $B^2/(B^2 + \tilde{W}^2)$ will estimate the same thing as the model-based calculation if the clusters have the same large size, but not otherwise.

Table 9 shows the measures of homogeneity computed from Formula (15) for an *srs/srs* design and Formula (17) for a *ppswr/srs* design, both with the *pwr*-estimator of a total. Values of $\tilde{\delta}$ in Table 9 for *srs/srs* when PSUs and SSUs are clusters are similar to those in Table 2 and Table 8. For example, $\tilde{\delta} = 0.005$ in Table 2 for y1 with PSUs as clusters and

Table 9. Measures of homogeneity $E_M(\tilde{\delta})$ and $E_M(\delta_*)$ estimated from Expression (15) for an (srs/srs, pwr-estimator) strategy and from Expression (17) for a (ppswr/srs, pwr-estimator) strategy

	PSUs	SSUs	Tracts	Block groups
$\tilde{\delta}$s for *srs/srs* design using (15)				
y1	0.005	0.024	0.159	0.198
y2	0.007	0.016	0.216	0.264
y3	0.079	0.161	0.738	0.797
ins.cov	0.004	0.011	0.503	0.569
hosp.stay	0.001	0.003	0.022	0.029
δ_*s for *ppswr/srs* design using (17)				
y1	0.005	0.025	0.008	0.012
y2	0.007	0.017	0.013	0.018
y3	0.077	0.161	0.144	0.190
ins.cov	0.005	0.012	0.008	0.015
hosp.stay	0.001	0.004	0.002	0.004

Table 10. Sample sizes of PSUs and elements computed with incorrect and correct measures of homogeneity

	Tracts		Block groups	
	m	$\bar{n}$	m	$\bar{n}$
ρs for *srs/srs* design using (14)				
y1	22	35	26	29
y2	27	28	29	24
y3	57	8	61	7
ins.cov	22	35	27	27
hosp.stay	12	71	15	58
δs for *srs/srs* design using (15)				
y1	58	7	61	6
y2	62	6	65	5
y3	84	2	86	2
ins.cov	76	3	78	3
hosp.stay	32	21	35	18

is 0.005 in both Tables 8 and 9. PSUs and SSUs have almost the same size, and therefore (15) reduces to the model formula for the correlation in (14). When tracts or BGs are clusters, the values of ρ in Table 8 and $\tilde{\delta}$ and δ_* in Table 9 are substantially different – for example, when tracts are clusters $\rho = 0.148$ for y3 but $\tilde{\delta} = 0.738$ for *srs/srs* in Table 9. However, 0.738 is close to the value of 0.730 for (tracts, *srs/srs*) in Table 3. That is, using the correlation estimated from the model in the variance formula for a total in (7) would be a mistake, as shown in Example 6 below. However, using the model correlation to calculate a measure of homogeneity in (15) works.

Example 6. Effect of using incorrect measure of homogeneity on sample size calculation. Suppose that the design is *srs/srs* with a fixed second-stage sampling rate and that tracts or BGs are used as PSUs. The cost assumptions are the same as those in Example 2. Table 10 in its upper bank lists the sample sizes computed from (12), assuming that the model correlations in (14) can be used for $\tilde{\delta}$. This would be appropriate if tracts and BGs were equal sized. The lower tier of Table 10 shows the sample sizes computed when the measures of homogeneity are the ones proper for tracts and BGs that are computed from (15). Since the correct $\tilde{\delta}$s in Table 9 are much larger than the model correlations in Table 8, the sample sizes of tracts and BGs in the lower tier are larger than in the upper tier of Table 10. The sample sizes of elements per tract or BG are correspondingly lower when the appropriate measures of homogeneity are used. All of the allocations in Table 10 respect the cost constraint of \$100,000, but the one in the lower tier will yield smaller CVs (i.e., more precise estimates), assuming that the values of $\tilde{\delta}$ from (15) are correct.

Results are different for a *ppswr/srs* design. The values of the model correlation ρ in Table 8 and the measures of homogeneity $\tilde{\delta}$ in Table 9 for tracts and BGs are almost identical. They are also very close to the design-based values in Table 6, resulting in relatively similar sample sizes for the two stages of the design using either method. As noted earlier, probability proportional to cluster-size sampling was extremely effective in reducing the between component of variance in the Maryland population. The upshot of this is that the measures of homogeneity and thus the sample sizes are quite similar to ones for a population in which clusters all have the same size.

5. Conclusion

Using variance components and measures of homogeneity are key parts of designing multistage samples. The relative sizes of the variance components are very sensitive to the sizes of the first-stage units or clusters themselves. Many textbooks present specialized variance formulae that assume that all clusters contain the same number of elements. However, varying cluster sizes can increase the measures of homogeneity that affect the precision of estimates from a two-stage sample. Having clusters that are more internally homogeneous will require more clusters and fewer elements per cluster to be sampled to achieve a desired level of precision. The effect of having variable-sized clusters also depends on the method of selecting clusters and the type of estimator that is used. Probability proportional to cluster-size sampling is more efficient than simple random sampling of clusters. Use of a ratio estimator when clusters are sampled via *srs* will temper some of the precision losses when cluster sizes vary, but still will be less efficient than *pps* sampling. As a result, recognizing the effects of varying cluster sizes is important for designing efficient samples and choosing estimators.

The variation of the tract sizes in the Maryland population used in our examples is considerably more than practitioners would prefer when defining PSUs for a household survey. For example, the range of the number of persons per tract is 86 to 13,579. Having such a large variation in PSU sizes leads to large differences in the cluster totals of analysis variables. This causes the between-cluster variance component to be large, which in turn leads to high measures of homogeneity and inefficiency if an equal-probability sample of clusters is selected. Standard practice would be to combine the small tracts or BGs so that all PSUs have some prescribed minimum number of persons. Although variation in cluster sizes can have a dramatic effect on the measures of homogeneity needed to design a sample, this seems to be rarely emphasized in sampling texts.

If the designer has some flexibility in forming the clusters, as would usually be the case in a household survey, clusters with nearly equal numbers of elements should definitely be created. In some surveys, however, the clusters are naturally occurring units, like schools, classrooms, or establishments. In those cases, one may have to live with the predefined units, but considering the variation in cluster size will be important when determining sample sizes. This will be true whether clusters are selected with equal probability or with probabilities proportional to their sizes as measured by counts of elements. Generally speaking, sampling unequal-sized clusters with probabilities proportional to their sizes will be more efficient as long as the measure of sizes (MOSs) are accurate and cluster totals of analysis variables are closely related to MOSs. If clusters are selected with equal probability, some efficiency can be recovered by using a ratio estimator of a total rather than a π-estimator; however, in the examples we presented, *pps* sampling will still be more efficient.

We have not covered several topics that are important in practice: three-stage sampling and nonlinear estimators more general than a ratio estimator. Three-stage sampling is used in many household surveys, but involves more complex variance formulae that we plan to address in a separate paper. Although we did not cover nonlinear estimators, such as the poststratification estimator or the general regression estimator, the analyses presented here will apply after forming a linear approximation to the nonlinear estimator (see, e.g., Binder

1995). The sizes of design effects for these nonlinear estimators can be quite different from those for the π-estimator, as pointed out by Park and Lee (2004).

Another important topic that we have omitted is domain estimation. The general technique of breaking the variance of an estimator into components will apply to subpopulation estimates. However, using the usual method of coding y to 0 for units not in the subpopulation will have an effect on the size of between- and within-variance components, which in turn affects the measures of homogeneity and sample size calculations. Whether a domain is spread over most clusters or present only in a subset of them will also affect the efficiency of sampling probability proportional to an MOS compared to equal-probability sampling of clusters.

Sample size calculation is an important aspect of survey design. Using formulae with assumptions that are not supported by the population at hand can result in either wasted project funding, an insufficient sample size with lower precision than desired, or inconclusive hypothesis tests. We demonstrated techniques not clearly specified in the literature to properly account for the variance components under two first-stage sample designs and the implications for assuming equal cluster sizes when in fact this is not the case. With knowledge in hand, survey statisticians are better equipped to design multistage surveys, and teachers will be better able to explain some of the nuances of sample design to students.

6. References

Bates, D., M. Maechler, and B. Bolker. 2011. *lme4: Linear Mixed-Effects Models Using S4 Classes*. Available at: http://CRAN.R-project.org/package=lme4. (accessed October 12, 2015).

Binder, D. 1995. "Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach." *Survey Methodology* 22: 17–22.

Center for Disease Control and Prevention. 2009. *National Health and Nutrition Examination Survey: 1999–2010 survey content*. Washington, DC: Department of Health and Human Services. Retrieved from www.cdc.gov/nchs/data/nhanes/survey_content_99_10.pdf.

Center for Disease Control and Prevention. 2012. *National Health Interview Survey*. Retrieved from National Center for Health Statistics: http://www.cdc.gov/nchs/nhis.htm.

Chromy, J. and L. Myers. 2001. "Variance Models Applicable to the NHSDA." In Proceedings of the Survey Research Methods Section: American Statistical Association, August 5–9, 2001. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/Proceedings/. (accessed October 12, 2015).

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

Gabler, S., S. Haeder, and P. Lahiri. 1999. "A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering." *Survey Methodology* 25: 105–106.

Hansen, M., W. Hurwitz, and M. Madow. 1953a. *Sample Survey Methods and Theory*, (Vol. I) New York: John Wiley & Sons.

Hansen, M., W. Hurwitz, and W. Madow. 1953b. *Sample Survey Methods and Theory*, (Vol. II) New York: John Wiley & Sons.

Hunter, S., K. Bowman, and J. Chromy. 2005. "Results of the Variance Component Analysis of Sample Allocation by Age in the National Survey on Drug Use and Health." In Proceedings of the Survey Research Methods Section: American Statistical Association, August 7–11, 2005 (pp. 3132–3136). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/Proceedings/. (accessed October 12, 2015).

Isaki, C. and W. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. Doi: http://dx.doi.org/10.1080/01621459.1982.10477770.

Judkins, D. and W. van de Kerckhove. 2003. *Residential Energy Consumption Survey 2005 Optimization*. Washington, DC: Department of Energy.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley.

Lynn, P. and S. Gabler. 2005. "Approximations to b^* in the Prediction of Design Effects Due to Clustering." *Survey Methodology* 31: 101–104.

Lohr, S. 2010. *Sampling: Design and Analysis*, (2nd edition). Boston, MA: Brooks/Cole CENGAGE Learning.

Park, I. and H. Lee. 2004. "Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling." *Survey Methodology* 30: 183–193.

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

Searle, S., G. Casella, and C. McCulloch. 1992. *Variance Components*. New York: John Wiley & Sons.

U.S. Census Bureau. 2011. *2010 Census Redistricting Data (Public Law 94–171) Summary File*. Washington, DC: Department of Commerce. Available at: http://www.census.gov/prod/cen2010/doc/pl94-171.pdf. (accessed October 12, 2015).

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., J.A. Dever, and F. Kreuter. 2015. `PracTools`: Tools for Designing and Weighting Survey Samples. R package version 0.3. Available at: http://CRAN.R-project.org/package=PracTools. (accessed November 25, 2015).

Waksberg, J., S. Sperry, D. Judkins, and V. Smith. 1993. "National Survey of Family Growth: Evaluation of Linked Design." *Vital and Health Statistics* 117: July 1993. 20pp. (PHS) 93-1391. PB94-103462. PC A04 MF A01. Available at: http://www.cdc.gov/nchs/products/series/series02.html (accessed October 12, 2015).

Wolter, K.M. 2007. *Introduction to Variance Estimation*. New York: Springer.

Received January 2013
Revised January 2015
Accepted January 2015