Bayesian SAE using Complex Survey Data
Lecture 5A: Survey Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Overview

Design-Based Inference

Simple Random Sampling

Stratified Simple Random Sampling

Cluster Sampling

Multistage Sampling

Discussion

Technical Appendix: Simple Random Sampling

Technical Appendix: Stratified SRS

Technical Appendix: Cluster Sampling

Technical Appendix: Lonely PSUs

Overview

Outline

Many national surveys employ stratified cluster sampling, also known as multistage sampling, so that's where we'd like to get to.

In this lecture we will discuss:

- Simple Random Sampling (SRS).
- Stratified SRS.
- Cluster sampling.
- Multistage sampling.

Main texts

- Lohr, S.L. (2010). *Sampling Design and Analysis, Second Edition*. Brooks/Cole Cengage Learning. Very well-written and clear mix of theory and practice.
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*, Wiley. Written around the `R survey` package. Great if you already know a lot about survey sampling.
- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. Wiley. Well written but not as comprehensive as Lohr.
- Särndal, Swensson and Wretman (1992). *Model Assisted Survey Sampling*. Springer. Excellent on the theory though steep learning curve and hard to dip into if not familiar with the notation. Also, anti- model-based approaches.

The problem

We have a question concerning variables in a well-defined finite population (e.g., 18+ population in Washington State).

What is required of a sample plan?

We want:

- Accurate answer to the question (estimate).
- Good estimate of the uncertainty of the estimate (e.g., variance).
- Reasonable cost for the study (logistics).

We may be interested in this particular finite population only, or in generalizing to other populations/situations, i.e., the process.

If the former, then if we sample the complete population, we are done! No statistics needed...

The problem

A simple random sample (SRS) is almost always better than a non-random sample, because the former allows more allows an assessment of uncertainty.

We will focus on design-based inference: in this approach the population values of the variable of interest, $y_1, \ldots, y_N$ are viewed as fixed, what is random is the indices of the individuals who are sampled.

This approach is frequentist, so that properties are based on hypothetical replications of the data collection process; hence, we require a formal description of the replication process.

A complex random sample may be:

- better than a SRS in the sense of obtaining the same precision at lower cost,
- but may be worse in the sense of precision, but be required logistically.

Domains

Often estimation is required for sub-populations of interests, these are known as domains.

Often the decision to study the domain occurs after the design, and so the sample sizes in the domains are random, and may be small.

If the domains are defined geographically, then inference for these domains is known as small area estimation (SAE).

Domains may also be defined as socio-demographic groups and trying to obtain sufficient samples in some domains (e.g., based on race), may lead to small sample sizes in others (e.g., states).

Probability samples

Notation for random sampling, in a single population (and not distinguishing areas):

- N, population size.
- n sample size.
- π_k, sampling probability for a unit (which will often correspond to a person) k, $k = 1, \ldots, N$.

Random does not mean “equal chance”, but means that the choice does not depend on variables/characteristics (either measured or unmeasured), except as explicitly stated[1] via known sampling probabilities.

In a simple random sample, the sampling probabilities are all equal,

$$\pi_k = \frac{n}{N}.$$

[1] For example, in stratified random sampling, certain groups may have fixed numbers sampled.

Probability samples

For design-based inference, which we shall discuss in detail:

- To obtain an unbiased estimator every individual k in the population to have a non-zero probability of being sampled; this probability will be defined as π_k, for $k = 1, \ldots, N$.
- To carry out inference, this probability π_k must be known for every individual in the sample (so not needed for the unsampled individuals).
- To obtain a form for the variance of an estimator: for every pair of individuals, k and l, in the sample, there must a non-zero probability of being sampled together, call this probability, π_{kl}, for $k = 1, \ldots, N$, $l = 1, \ldots, N$, $k \neq l$.
- The probability π_{kl} must be known for every pair in the sample[2].

Lower case values will denote population values, $y_1, \ldots, y_N$.

[2] in practice, these are often approximated, or the variance is calculated via a resampling technique such as the jackknife

Probability samples

The label probability sample is often used instead of random sample.

Non-probability sampling approaches include

- convenience (accidental, haphazard) sampling (e.g., asking for volunteers);
- purposive (also known as judgmental) sampling in which a researcher users their subject knowledge to select participants (e.g, selecting an “average” looking individual).
- Quota sampling in which quotas in different groups are satisfied (but unlike stratified sampling, probability sampling is not carried out).

Non-probability samples cannot be analyzed with design-based approaches, because there is no π_k or π_{kl}.

Representative samples

Surveys are broadly of two types: questionnaire and interview.

A fundamental concept in sampling is whether the sample is representative.

There is no perfect, “scaled down” version of the population for whom we would like to make inference is available. Lohr (2010, p. 3) says,

“...a good sample will be **representative** in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy”.

Post-stratification and raking are techniques for making a sample more representative.

When does having a representative sample matter?

Inference for population quantities such as means, totals, medians, etc., are not reliable except with random samples.

Estimation of relationships between variables, for example, whether a diet high in salt increases blood pressure, can often be estimated from non-random samples, with careful modeling — called model-based analysis.

Definitions: Based on Lohr (2010, Section 1.2).

Definitions to allow the idea of a good sample to be make precise:

- Observation unit: An object on which a measurement is taken, sometimes called an element. In human populations, observation units are individuals.
- Target population: The complete collection of observations we want to study.
- Sample: A subset of a population.
- Sampled population: The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was chosen. The sampled population will often not correspond to the target population; it may be a more accessible version for example.

Definitions

- Sampling unit: A unit that can be selected for a sample. Although we might want to study individuals, we may not have a list of individuals in the target population. For example, households may serve as the sampling units, with the individuals in the household being the observation units.
- Sampling frame: A list, map or other specification of sampling units in the population from which a sample may be selected. i.e.. it allows access to the sampling units. For a multistage survey, a sampling frame should exist for each stage of sampling.

  Examples:
  - In BRFSS, the sampling frame is a list of telephone numbers (actually 2 lists, landline and cell).
  - for the DHS, the sampling frame is often derived from the census and corresponds to a list of enumeration areas (EAs); within each EA, there should be a list of households;
  - In NHANES the sampling frame is counties.

Selection bias

- Selection bias occurs when some population units are sampled at a different rate than intended by the investigator.
- Undercoverage can lead to selection bias, e.g., BRFSS is a telephone survey; started in 1984 at which time many households did not have landline telephones, and so such people are not a random sample of the target population (over 18 years of age).
- Overcoverage includes population units in the sampling frame that are not in the target population, e.g., desire over-18 year olds by phone, but younger individuals are included.
- Multiplicity in lists, e.g., households with more than one phone have a greater probability of being selected.
- Non-response frequently leads to selection bias since non-responders often differ from responders. It is better to have a small survey with a high response rate, than a large survey with a low response rate.
- Surveys in which the participants volunteer (e.g., internet polls) are fraught with selection bias.

Common sampling designs

- Simple random sampling: Select each individual with probability $\pi_k = n/N$.
- Stratified random sampling: Use information on each individual in the population to define strata h, and then sample n_h units independently within each stratum.
- Probability-proportional-to-size sampling: Given a variable related to the size of the sampling unit, Z_k, on each unit in the population, sample with probabilities $\pi_k \propto Z_k$.
- Cluster sampling: All units in the population are aggregated into larger units called clusters, known as primary sampling units (PSUs), and clusters are sampled initially, with units within clusters then being sampled.
- Multistage sampling: Stratified cluster sampling, with multiple levels of clustering.

Measurement error

Measurement error reflects inaccurate responses.

Multitude of reasons; people:

- lie,
- do not understand the question,
- forget,
- respond how they think the interviewer would like them to respond.

Interviewers may "cheat".

Nonsampling and sampling errors

Selection bias and measurement error are examples of nonsampling errors.

Sampling ‘errors’ occur because we take a sample and not the complete population of individuals; each potential sample we can take will give a particular answer, and the sample to sample variability can be expressed in probabilistic terms.

Design-Based Inference

Overview of approaches to inference

In general, data from survey samples may be analyzed using:

1. Design-based inference.
2. Model-based inference.
3. Model-assisted inference.

We focus on 1. and 2.

Inference

Suppose we are interested in a variable denoted y, with the population values being $y_1, \dots, y_N$.

Random variables will be represented by upper case letters, and constants by lower case letters.

Finite population view: We have a population of size N and we are interested in characteristics of this population, for example, the mean

$$\overline{y}_U = \frac{1}{N} \sum_{k=1}^{N} y_k.$$

Infinite population view: The population variables are drawn from a hypothetical distribution (the model) $f(\cdot)$ with mean μ.

In the latter (model-based) view, $Y_1, \dots, Y_N$ are random variables and properties are defined with respect to $f(\cdot)$; often we say Y_k are independent and identically distributed (iid) from $f(\cdot)$.

Model-based inference

As an example, we take the sample mean:

$$\overline{Y} = \frac{1}{n} \sum_{k=1}^{n} Y_k$$

is a random variable because $Y_1, \ldots, Y_n$ are each random variables.

Assume Y_k are iid observations from a distribution (f) with mean μ and variance σ^2.

The sample mean is an ubiased estimator, and has variance σ^2/n.

Model-based inference

Unbiased estimator:

$$\begin{aligned} \mathrm{E}[\overline{Y}] &= \mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{n} Y_k\right] = \frac{1}{n}\sum_{k=1}^{n} \underbrace{\mathrm{E}\,[Y_k]}_{=\mu} \\ &= \frac{1}{n}\sum_{k=1}^{n} \mu = \mu \end{aligned}$$

Variance:

$$\begin{aligned} \mathrm{var}(\overline{Y}) &= \mathrm{var}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) \underbrace{=}_{\text{iid}} \frac{1}{n^2}\sum_{k=1}^{n} \underbrace{\mathrm{var}\,(Y_k)}_{=\sigma^2} \\ &= \frac{1}{n^2}\sum_{k=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Design-based inference

In the design-based approach to inference the y values are treated as unknown but fixed[3] (so we write $y_1, \dots, y_N$), and the randomness, with respect to which all procedures are assessed, is associated with the particular sample of individuals that is selected, call the random set of indices S.

Minimal reliance on distributional assumptions.

Sometimes referred to as inference under the randomization distribution.

In general, the procedure for selecting the sample is under the control of the researcher.

The basic estimator is the weighted form (Horvitz and Thompson, 1952; Hájek, 1971)

$$\widehat{\overline{Y}}_U = \frac{\sum_{k \in S} w_k y_k}{\sum_{k \in S} w_k}.$$

[3] To emphasize: the y's are not viewed as random variables

Simple Random Sampling

Simple random sample (SRS)

The simplest probability sampling technique is simple random s without replacement, or srswor.

Suppose we wish to estimate the population mean in a particular population of size *N*.

In everyday language: consider a population of size *N*; a random sample of size $n \leq N$ means that any subset of *n* people from the total number *N* is equally likely to be selected.

This is known as simple random sampling.

Simple random sample (SRS)

We sample n people from N, choosing each person independently at random and with the same probability of being chosen:

$$\pi_k = \frac{n}{N},$$

$k = 1, \ldots, N$.

Note: sampling without replacement and the joint sampling probabilities are

$$\pi_{kl} = \frac{n}{N} \times \frac{n-1}{N-1}$$

for $k, l = 1, \ldots, N$, $k \neq l$.

In this situation:

- The sample mean is an unbiased estimator.
- The uncertainty, i.e. the variance in the estimator can be easily estimated.
- Unless n is quite close to N, the uncertainty does not depend on N, only on n (see later for numerical examples).

Design-based inference

Example: $N = 4, n = 2$ with SRS. There are 6 possibilities:

$$\{y_1, y_2\}, \quad \{y_1, y_3\}, \quad \{y_1, y_4\}, \quad \{y_2, y_3\}, \quad \{y_2, y_4\}, \quad \{y_3, y_4\}.$$

The random variable describing this design is S, the set of indices of those selected.

The sample space of S is

$$\{(1, 2), \quad (1, 3), \quad (1, 4), \quad , (2, 3), \quad (2, 4), \quad (3, 4)\},$$

and under SRS, the probability of sampling one of these possibilities is 1/6.

The selection probabilities are
$\pi_k = \Pr($ individual i in sample $) = \frac{3}{6} = \frac{1}{2}$, which of course is $\frac{n}{N}$.

In general, we can work out the selection probabilities without enumerating all the possibilities!

Design-based inference

Fundamental idea behind design-based inference: An individual with a sampling probability of π_k can be thought of as representing $1/\pi_k$ individuals in the population.

Example: in SRS each person selected represents $\frac{N}{n}$ people.

The value $w_k = 1/\pi_k$ is called the sampling (or design) weight.

Estimator of $\overline{y}_U$ and properties under SRS

The weighted estimator is

$$\begin{aligned}
\widehat{\overline{Y}}_U &= \frac{\sum_{k \in S} w_k y_k}{\sum_{k \in S} w_k} \\
&= \frac{\sum_{k \in S} \frac{N}{n} y_k}{\sum_{k \in S} \frac{N}{n}} \\
&= \frac{\sum_{k \in S} y_k}{n} = \overline{y}_S,
\end{aligned}$$

the sample mean.

This is an unbiased estimator (i.e., $\mathrm{E}[\widehat{\overline{Y}}_U] = \overline{Y}_U$), where we average over all possible samples we could have drawn, i.e., S.

Estimator of $\overline{y}_U$ and properties under SRS

Variance is

$$\mathrm{var}(\overline{y}_S) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, \tag{1}$$

where,

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \overline{y}_U)^2.$$

Contrast this with the model-based variance which is σ^2/n.

The factor

$$1 - \frac{n}{N}$$

is the finite population correction (fpc).

Because we are estimating a finite population mean, the greater the sample size relative to the population size, the more information we have (relatively speaking), and so the smaller the variance.

In the limit, if $n = N$ we have no uncertainty, because we know the population mean!

Estimator of $\overline{y}_U$ and properties under SRS

The variance of the estimator depends on the population variance S^2, which is usually unknown, so instead we estimate the variance using the unbiased estimator:

$$s^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \overline{y}_S)^2.$$

Substitution into (1) gives an unbiased estimator of the variance:

$$\widehat{\text{var}}(\overline{y}_S) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}. \tag{2}$$

The standard error is

$$\text{SE}(\overline{y}_S) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}.$$

Note: S^2 is not a random variable but s^2 is.

Estimator of $\overline{y}_U$ and properties under SRS

If n, N and $N - n$ are “sufficiently large”[4], a $100(1 - \alpha)\%$ confidence interval for $\overline{y}_U$ is

$$\left[\overline{y}_S - z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{s}{\sqrt{n}}, \quad \overline{y}_S + z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{s}{\sqrt{n}}\right], \tag{3}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th percentile of a standard normal random variable.

The interval in (3) is random (across samples) because $\overline{y}_S$ and s^2 (the estimate of the variance) are random.

In practice therefore, if $n \ll N$, we obtain the same confidence interval whether we take a design- or a model-based approach to inference (though the interpretation is different).

[4] so that the normal distribution provides a good approximation to the sampling distribution of the estimator

Weighted estimator of $\overline{y}_U$

Recall that the sampling weights are $w_k = 1/\pi_k$ where π_k is the inclusion probability, which for SRS is $\pi_k = n/N$.

Hence, we can think of each sampled individual as representing N/n individuals.

Sometimes the population size may be unknown and the sum of the weights provides an unbiased estimator.

In general, examination of the sum of the weights can be useful as if it far from the population size (if known) then it can be indicative of a problem with the calculation of the weights.

Weighted estimator of $\overline{y}_U$

The weighted sum of the sampled *y*'s is the estimator of the total:

$$\sum_{k \in S} w_k y_k \quad = \quad \sum_{k \in S} \frac{N}{n} y_k = \widehat{t}$$

A weighted estimator of this form as known as a Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

For SRS:

$$\sum_{k \in S} w_k \quad = \quad \sum_{k \in S} \frac{N}{n} = N$$

so the sum of the weights is exactly the population total.

This is true for more general sampling schemes, and is useful if the population total is unknown (and is also often used if the population total is known).

Weighted estimator of $\overline{y}_U$

Hence, the mean estimator can be written as

$$\frac{\sum_{k \in S} w_k y_k}{\sum_{k \in S} w_k} = \frac{\widehat{t}}{N} = \overline{y}_S.$$

This form will reappear many times, for more general weighting schemes.

Dividing by the estimated population total is known as the Hájek estimator (Hájek, 1971).

Stratified Simple Random Sampling

Stratified simple random sampling

Simple random samples are rarely taken in surveys because they are logistically difficult and there are more efficient designs for gaining the same precision at lower cost.

Stratified random sampling is one way of increasing precision and involves dividing the population into groups called strata and drawing probability samples from within each one, with sampling from different strata being independent.

The stratified simple random sampling without replacement design is sufficiently popular to merit a ridiculous acronym, stsrswor.

An important practical consideration of whether stratified sampling can be carried out is whether stratum membership is known (for whatever variable is defining the strata) for every individual in the population.

Reasons for stratified random sampling

- Protection from the possibility of a “really bad sample”, i.e., very few or zero samples in certain stratum giving an unrepresentative sample.
- Obtain known precision required for subgroups (domains) of the population.
- Convenience of administration since sampling frames can be constructed differently in different strata. The different stratum may contain units that differ greatly in practical aspects of response, measurement, and auxiliary information, and so being able to treat each stratum individually in terms of design and estimation, may be beneficial.

Reasons for stratified random sampling

- More precise estimates can be obtained if stratum can be found that are associated with the response of interest, for example, age and gender in studies of human disease.
- The most natural form of sampling may be based on geographical regions, and treating each region as a separate stratum is then suggested.
- Due to the independent sampling in different stratum, variance estimation straightforward (so long as within-stratum sampling variance estimators are available).

See Lohr (2010, Section 3.1) for further discussion.

Example: NMIHS

Korn and Graubard (1999) discuss the National Maternal and Infant Health Survey (NMIHS) which collected information on live births, fetal deaths and infant deaths that occurred in 1998 in the United States (excluding Montana and South Dakota).

Six strata were used, as the cross of race (black/non-black) and birthweight of the baby as reported on the birth certificate ($<$1500, 1500–2499, $\geq$2500 grams).

These strata include groups at risk for adverse pregnancy outcomes and so they were oversampled in the NMIHS to increase the reliability of estimates for these subdomains.

Example: Washington State

According to

`http://quickfacts.census.gov/qfd/states/53000.html`

there were 2,629,126 households in WA in 2009–2013.

Consider a simple random sample of 2000 households, so that each household has a

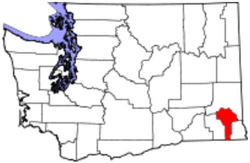
$$\frac{2000}{2629126} = 0.00076,$$

chance of selection.

Suppose we wish to estimate characteristics of household in all 39 counties of WA.

Example: Washington State

King and Garfield counties had 802,606 and 970 households so that under SRS we will have, on average, about 610 households sampled from King County and about 0.74 from Garfield county.



The probability of having no-one from Garfield County is about 22%, (binomial experiment) and the probability of having more than one is about 45%.

If we took exactly 610 from King and 1 (rounding up) from Garfield we have an example of proportional allocation (but see later for problems with samples of size 1).

Stratified sampling allows control of the number of samples in each county.

Notation

Stratum levels are denoted $h = 1, \ldots, H$, so H is total.

Let $N_1, \ldots, N_H$ be the known population totals in the stratum with

$$N_1 + \cdots + N_H = N,$$

where N is the total size of the population.

In stratified random sampling, the simplest from of stratified sampling, we take a SRS from each stratum with n_h samples being randomly taken from stratum h, so that the total sample size is

$$n_1 + \cdots + n_H = n.$$

Notation

Population quantities:

- y_{hk} value of kth unit in stratum h, $h = 1, \ldots, H$, $k = 1, \ldots, N_h$.
- $t_h = \sum_{k=1}^{N_h} y_{hk} =$ population total in stratum h.
- $t = \sum_{h=1}^{H} t_h =$ population total.
- $\overline{y}_{hU} = \frac{1}{N_h} \sum_{k=1}^{N_h} y_{hk} =$ population mean in stratum h.
- $\overline{y}_U = \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{N_h} y_{hk} = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_{hU} =$ population mean.
- $S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (y_{hk} - \overline{y}_{hU})^2 =$ population variance in stratum h.

Estimators

We can view stratified random sampling as carrying out SRS in each of the H stratum; we let S_h represent the probability sample in stratum h.

We also let S refer to the overall probability sample.

Confusing notation: S_h is both the standard deviation and the random probability sample, in strata h but hopefully clear which we are referring to by the context.

Estimators

Sample estimators:

- Stratum h mean (the S in the subscript emphasizes that this is a random variable with respect to the random sample):

$$\overline{y}_{hS} = \frac{\sum_{k \in S_h} y_{hk}}{n_h}.$$

- Stratum h total:

$$\widehat{t}_h = N_h \overline{y}_{hS} = \frac{N_h}{n_h} \sum_{k \in S_h} y_{hk}.$$

Note that the n_h are not random because the survey is defined with a fixed n_h in mind.

Estimators

Sample estimators:

- Population total:

$$\widehat{t}_{\text{strat}} = \sum_{h=1}^{H} \widehat{t}_h = \sum_{h=1}^{H} N_h \overline{y}_{hS}. \tag{4}$$

- Population mean:

$$\overline{y}_{\text{strat}} = \frac{\widehat{t}_{\text{strat}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_{hS}. \tag{5}$$

- Stratum variance:

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_{hk} - \overline{y}_{hS})^2.$$

Estimators

It is straigtforward to show that (4) and (5) are unbiased estimators, since we have linear combinations of SRS estimators.

Since we are sampling independently from each stratum using SRS, the variance of the mean and total estimators is simply the sum of the variances within each stratum:

$$\begin{aligned}
\mathrm{var}(\widehat{t}_{\mathrm{strat}}) &= \sum_{h=1}^{H} \mathrm{var}(\widehat{t}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h} && (6) \\
\mathrm{var}(\overline{y}_{\mathrm{strat}}) &= \frac{\mathrm{var}(\widehat{t}_{\mathrm{strat}})}{N^2} = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} && (7)
\end{aligned}$$

Example: 1988 NMIHS

Table 1: Mother's age, as reported on birth certificate, and other statistics, by stratum (race and birthweight, in grams), from 1988 NMIHS. Data reproduced from Korn and Graubard (1999, Table 2.2-1).

Stratum h	Estimated Population Size (N_h)	Sample Size (n_h)	Sampling Fraction (n_h/N_h)	Mean Age ($\overline{y}_{hs}$)	Standard Deviation Age (s_h)
1. Black, $<$1500	18,130	1295	1/14	24.64	5.84
2. Black, 1500–2499	65,670	1194	1/55	24.42	5.76
3. Black, $\geq$2500	559,124	4948	1/113	24.41	5.68
4. Non-Black, $<$1500	27,550	950	1/29	26.44	5.88
5. Non-Black, 1500–2499	150,080	938	1/160	26.11	5.85
6. Non-Black, $\geq$2500	2,944,800	4090	1/720	26.70	5.45

Example: 1988 NMIHS

The target population is live births in the United States in 1988 from mothers who were 15 years or older.

Using (4) we can estimate the mean as

$$\begin{aligned}\overline{y}_{\text{strat}} &= \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_{hS} \\ &= \frac{1}{3765354}(18130 \times 24.64 + \cdots + 2944800 \times 26.70) \\ &= 26.28 \text{ years.}\end{aligned}$$

Notice that the mean is far closer to the non-black summaries, since the oversampling of black mothers is accounted for.

Example: 1988 NMIHS

The variance is estimated, from (7), as

$$\begin{aligned}\widehat{\text{var}}(\overline{y}_{\text{strat}}) &= \frac{1}{(3765354)^2}\left[(18130)^2\left(1-\frac{1}{14}\right)\frac{(5.84)^2}{1295}+\cdots\right.\\ &+ \left.(2944800)^2\left(1-\frac{1}{720}\right)\frac{(5.45)^2}{4090}\right] = 0.004647.\end{aligned}$$

A 95% confidence interval for the average age (in years) of mothers (15 years or older) of live births in the United States is

$$26.28 \pm 1.96 \times \sqrt{0.004647} = (26.15, 26.41).$$

Defining strata

Since we almost always gain in precision over SRS, why not always use stratification?

A very good reason is that we need the stratification variable to be available on all of the population.

Taking a stratified sample adds to complexity.

Stratification is best when the stratum means differ greatly; ideally we would stratify on the basis of y, but of course these are unknown in the population (that's the point of the survey!).

Stratification should aim to produce strata within which the outcomes of interest have low variance.

Cluster Sampling

References on cluster sampling

Lumley (2010, Chapter 3): not very extensive but describes the use of the `survey` package.

Lohr (2010, Chapters 5 and 6): very good description.

Särndal et al. (1992, Chapter 4): concentrates on the estimation side.

Korn and Graubard (1999, Section 2.3): a brief overview.

Motivation

Cluster sampling is an extremely common design that is often used for government surveys.

Two main reasons for the use of cluster sampling:

- A sampling frame for the population of interest does not exist, i.e., no list of population units.
- The population units have a large geographical spread and so direct sampling is not logistically feasible to implement. It is far more cost effective (in terms of travel costs, etc.) to cluster sample.

Motivation

The clusters can be:

- Genuine features of the populations, e.g., households, schools, or workplaces.
- Subsets chosen for convenience, e.g., counties, zipcodes, telephone number blocks.

Terminology

In single-stage cluster sampling or one-stage cluster sampling, the population is grouped into subpopulations (as with stratified sampling) and a probability sample of these clusters is taken, and every unit within the selected clusters is surveyed.

In one-stage cluster sampling either all or none of the elements that compose a cluster (PSU) are in the sample.

The subpopulations are known as clusters or primary sampling units (PSUs).

In two-stage cluster sampling, rather than sample all units within a PSU, a further cluster sample is taken; the possible groups to select within clusters are known as secondary sampling units (SSUs).

For example, if we take a SRS within each PSU sampled, we have a two-stage cluster sampling design.

This can clearly be extended to multistage cluster sampling.

Motivation: NHANES

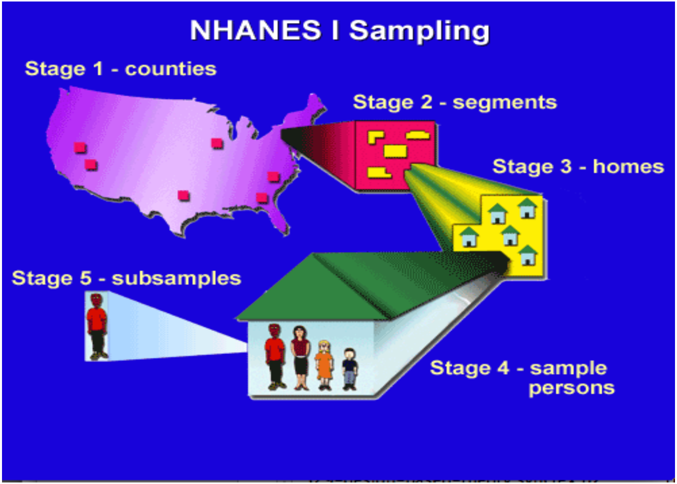


Figure 1: Cartoon of sample design in NHANES I; a multistage stratified clustered sample of civilian, non-institutionalized population.

Motivation: NHANES

In NHANES, participants had an interview, clinical examination and blood samples were taken and needed to be stored, and this carried out mobile examination trailers.

27,000 individuals were sampled over 4 years and not practical to move the trailers to thousands of locations.

Figure 2 shows what a SRS of 10,000 looks like; the sampled individuals live in 1184 counties.

In NHANES III the design used involved sampling 81 PSUs locations (clusters) with a plan to recruit multiple participants in each cluster.

Motvation: NHANES

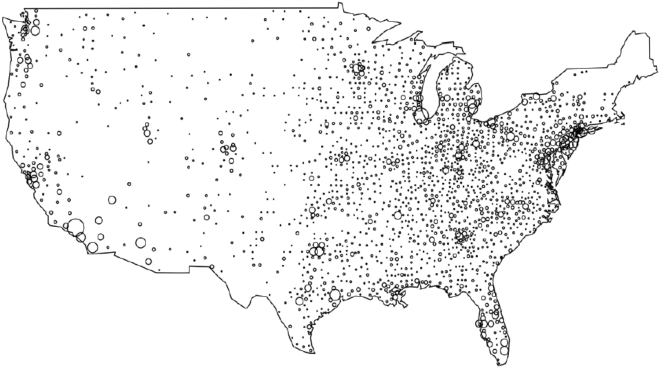


Figure 2: A SRS of 10,000 voter locations from the USA with circles at the county centroids and areas proportional to the number sampled. Los Angeles County contains the largest sample of 257.

Differences between cluster sampling and stratified random sampling

Stratified Random Sampling	One-Stage Cluster Sampling
SRS is taken from every stratum	Observe all elements only within the sampled clusters
Variance of estimate of $\overline{y}_U$ depends on within strata variability	Cluster is sampling unit and the more clusters sampled the smaller the variance. The variance depends primarily on between cluster means
For greatest precision, low within-strata variability but large between-strata variability	For greatest precision, high within-cluster variability and similar cluster means.
Precision generally better than SRS	Precision generally worse than SRS

Heterogeneity

The reason that cluster sampling loses efficiency over SRS is that within clusters we only gain partial information from additional sampling within the same cluster, since within clusters two individuals tend to be more similar than two individuals within different clusters.

The similarity of elements within clusters is due to unobserved (or unmodeled) variables.

Estimation: Unbiased estimation for one-stage cluster sampling

There are two ways we might estimate totals and means: via an unbiased estimator, or using ratio estimation; we briefly describe the former.

We suppose that a SRS of n PSUs is taken.

The key idea is to realize that since all SSUs in the selected clusters are observed, we can use results directly from SRS.

Reminder, for SRS:

$$\begin{aligned}
\widehat{t} &= \frac{N}{n}\sum_{k\in S} y_k \\
\text{var}(\widehat{t}) &= N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} \\
\widehat{\text{var}}(\widehat{t}) &= N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}.
\end{aligned}$$

Estimation: Unbiased estimation for one-stage cluster sampling

To use the above results in the one-stage cluster sampling context, replace y_k by t_i, the total in cluster i.

Then using the results for an SRS of n from N we have, for one-stage sampling:

$$
\begin{aligned}
\widehat{t}_{\text{unb}} &= \frac{N}{n}\sum_{i\in S} t_i & (8)\\
\text{var}(\widehat{t}_{\text{unb}}) &= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} & (9)\\
\widehat{\text{var}}(\widehat{t}_{\text{unb}}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}, & (10)
\end{aligned}
$$

where S_t^2 is the variance of the PSU totals and s_t^2 is the estimate of this variance (see Technical Appendix for more on notation).

Estimation: Unbiased estimation for one-stage cluster sampling

The probability of sampling a PSU is n/N, and since all the SSUs are sampled in each selected PSU we have selection probabilities and design weights

$$\begin{aligned}\pi_{ik} &= \text{Pr(SSU } k \text{ in cluster } i \text{ is selected)} = \frac{n}{N}\\ w_{ik} &= \text{ Design weight for SSU } k \text{ in cluster } i \ = \frac{N}{n}.\end{aligned}$$

Hence, we can write (8)

$$\widehat{t}_{\text{unb}} = \sum_{i \in S} \sum_{k \in S_i} w_{ik} y_{ik}$$

since $t_i = \sum_{k \in S_i} y_{ik}$.

Estimation: Unbiased estimation for one-stage cluster sampling

We now turn our attention to estimation of the population mean $\overline{y}_U$.

Let $M_0 = \sum_{i=1}^{N} M_i$ be the total number of secondary sampling units (SSUs) (i.e., elements in the population) so that

$$\begin{aligned}\overline{y}_U &= \frac{1}{M_0}\sum_{i=1}^{N}\sum_{k=1}^{M_i} y_{ik} \\ &= \frac{1}{M_0}\sum_{i=1}^{N} t_i = \frac{t}{M_0}\end{aligned}$$

Then,

$$\begin{aligned}\widehat{\overline{y}}_{\text{unb}} &= \frac{\widehat{t}_{\text{unb}}}{M_0} \\ \widehat{\text{var}}(\widehat{\overline{y}}_{\text{unb}}) &= \frac{1}{M_0^2}\widehat{\text{var}}(\widehat{t}_{\text{unb}}) = \frac{N^2}{M_0^2}\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n}\end{aligned}$$

Two-stage cluster sampling with equal-probability sampling

It may be wasteful to measure all SSUs in the selected PSUs, since the units may be very similar and so there are diminishing returns on the amount of information we obtain.

Here, we discuss the equal-probability two stage cluster design:

1. Select an SRS S of n PSUs from the population of N PSUs.
2. Select an SRS of m_i SSUs from each selected PSU, the probability sample collected will be denoted S_i.

Estimation for two-stage cluster sampling

Since we do not observe all the SSUs in the sampled PSUs we estimate

$$\widehat{t}_i = \sum_{k \in S_i} \frac{M_i}{m_i} y_{ik} = M_i \overline{y}_i,$$

to give the unbiased estimator of the population total:

$$\widehat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in S} \widehat{t}_i = \frac{N}{n} \sum_{i \in S} M_i \overline{y}_i = \sum_{i \in S} \sum_{k \in S_i} \frac{N}{n} \frac{M_i}{m_i} y_{ik}. \tag{11}$$

Estimation for two-stage cluster sampling using weights

The inclusion probabilities are:

$$\begin{aligned} \Pr(\ k\text{th SSU in } i\text{th PSU selected }) &= \Pr(\ i\text{th PSU selected }) \\ &\times \Pr(\ k\text{th SSU} \mid i\text{th PSU selected }) \\ &= \frac{n}{N} \times \frac{m_i}{M_i} \end{aligned}$$

Hence, the weights are

$$w_{ik} = \pi_{ik}^{-1} = \frac{N}{n} \times \frac{M_i}{m_i}.$$

The unbiased estimator (11) may then be written as

$$\widehat{t}_{\text{unb}} = \sum_{i \in S} \sum_{k \in S_i} w_{ik} y_{ik}.$$

Variance calculation is not trivial, and requires more than knowledge of the weights.

Variance estimation for two-stage cluster sampling

With respect to (11), in contrast to one-stage cluster sampling we have to acknowledge the uncertainty in both stages of sampling; in one-stage cluster sampling the t_i are known in the sampled PSUs, whereas in two stage sampling we have estimates $\widehat{t}_i$.

In Lohr (2010, Chapter 6) it is shown that

$$\text{var}(\widehat{t}_{\text{unb}}) = \underbrace{N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}}_{\text{one-stage cluster variance}} + \underbrace{\frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}}_{\text{two-stage cluster variance}} \tag{12}$$

where

- S_t^2 is the population variance of the cluster totals,
- S_i^2 is the population variance within the *i*th PSU.

If all SSUs are included in the sampled PSU, i.e. $m_i = M_i$, we return to one-stage cluster sampling as the second term in (12) is zero.

Variance estimation for two-stage cluster sampling

Again from Lohr (2010, Chapter 6), an unbiased variance estimate is

$$\widehat{\text{var}}(\widehat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i} \qquad (13)$$

where

- $s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\widehat{t}_i - \frac{\widehat{t}_{\text{unb}}}{N}\right)^2$ is the sample variance of the estimated PSU totals,
- $s_i^2 = \frac{1}{m_i - 1} \sum_{k \in S_i} (y_{ik} - \overline{y}_i)^2$ is the sample variance of the sampled SSUs within the ith PSU.

Variance estimation for two-stage cluster sampling

If N is large, the first term in (13) dominates, and often software uses this term only, even omitting the fpc to give the with replacement estimator

$$\widehat{\text{var}}_{\text{wr}}(\widehat{t}_{\text{unb}}) = N^2 \frac{s_t^2}{n}.$$

As in one-stage cluster sampling with unequal cluster sizes,

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\widehat{t}_i - \frac{\widehat{t}_{\text{unb}}}{N} \right)^2 = \frac{1}{n-1} \sum_{i \in S} M_i^2 \left(t\overline{y}_i - \frac{\widehat{t}_{\text{unb}}}{M_i N} \right)^2$$

can be very large since it is affected by both variation in the unit sizes (the M_i) and by variations in the $\overline{y}_i$.

If the cluster sizes are variable the variance can be large even if the cluster means $\overline{y}_i$ are relatively constant.

Estimation of the mean for two-stage cluster sampling

If total number of units, M_0, is known we can estimate the population mean by

$$\widehat{\overline{y}}_{\text{unb}} = \frac{\widehat{t}_{\text{unb}}}{M_0},$$

with variance

$$\widehat{\text{var}}(\widehat{\overline{y}}_{\text{unb}}) = \frac{\widehat{\text{var}}(\widehat{t}_{\text{unb}})}{M_0^2}$$

Multistage Sampling

Multistage Sampling in the DHS

A common design in national surveys is multistage sampling, in which cluster sampling is carried out within strata.

We will not go into inference for this design, but basically weighted estimates are readily available, and accompanying variance estimates can be calculated.

DHS Program: Typically, 2-stage stratified cluster sampling:

- Strata are urban/rural and region.
- Enumeration Areas (EAs) sampled within strata (PSUs).
- Households within EAs (SSUs).

Information is collected on population, health, HIV and nutrition; more than 300 surveys carried out in over 90 countries, beginning in 1984.

Discussion

Discussion

The majority of survey sampling texts are based on design-based inference, which is a different paradigm to model-based inference!

However, for the major designs (SRS, stratified SRS, cluster sampling, multistage sampling), weighted estimates and their variances are available within all the major statistical packages.

What is required in the data are the weights, and the design information for each individuals, for example, the strata and cluster membership.

We will exclusively use the `survey` package in `R`.

When the variance is large, we would like to use Bayesian methods to smooth, but where's the likelihood?

References

Hájek, J. (1971). Discussion of, “An essay on the logical foundations of survey sampling, part I”, by D. Basu. In V. Godambe and D. Sprott, editors, *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Kish, L. (1965). *Survey sampling*. John Wiley and Sons, Chichester.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys*. John Wiley and Sons, New York.

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

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. John Wiley and Sons, Hoboken, Jersey.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–625.

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

Technical Appendix: Simple Random Sampling

Statistical details on SRS

Let $y_1, \ldots, y_N$ be the population values of a variable of interest and suppose we carry out SRS without replacement (in practice N may not always be known). Example: $N = 1000, n = 50$.

Approach:

1 Select with probability $\frac{1}{N}$ the first element from the N units. Example: probability is $\frac{1}{1000}$.

2 Select with probability $\frac{1}{N-1}$ the second element from the remaining $N - 1$ units. Example: probability is $\frac{1}{999}$

...

n Select with probability $\frac{1}{N-n+1}$ the n-th element from the remaining $N - n + 1$ units. Example: probability is $\frac{1}{951}$

Statistical details on SRS

Let $U = \{1, \ldots, N\}$ be the index set of the finite population and s be the index set of the sampled units, with S being the random variable representing the sample selected.

Suppose S can take the values $s_1, \ldots, s_M$, i.e. these are the possible sample set of indices that could be selected; let $p(s)$ be the probability distribution over the possible sets that can be selected.

For SRS without replacement there are $M = \binom{N}{n}$ possible sets of n elements that can be selected and

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } s \text{ has } n \text{ elements} \\ 0 & \text{otherwise} \end{cases}$$

Example: $N = 4$ and $n = 2$ so that $M = \binom{4}{2} = 6$. Write down the possible samples and the probabilities of these samples.

Statistical details

In general, the size of the sample is denoted n_S, where we subscript by S because the size of the sample may depend on the sample that is (randomly) chosen.

When this is not the case we write n.

In Bernoulli sampling, each unit is selected independently with probability $0 < q < 1$.

The number of units selected, n_S, is binomial with parameters N and q.

Poisson sampling is a generalizaion of Bernoulli sampling in which the probabilities associated with each unit can vary (see Särndal et al. 1992, p. 85). Each sample has probability

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U - s} (1 - \pi_k).$$

Estimator of $\overline{y}_U$ and properties under SRS

For estimating the population mean $\overline{y}_U$ we use the sample mean $\overline{y}_S$; note the dependence on S (the random variable representing the probability sample).

$$\overline{y}_S = \frac{1}{n} \sum_{k \in S} y_k \tag{14}$$

The notation is a bit cumbsersome (and is not consistent in the literature).

We write $\overline{y}_s$ to be the estimate (i.e. a number, not a random variable).

Estimator of $\overline{y}_U$ and properties under SRS

The variance is

$$\text{var}(\overline{y}_S) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \tag{15}$$

where S^2 is the variance of the population values about the mean:

$$S^2 = \frac{1}{N-1} \sum_{k=1}^{N} (y_k - \overline{y}_U)^2.$$

The variance (15) measures the variability of estimates of $\overline{y}_U$ over different samples.

Estimator of a proportion under SRS

Estimating a proportion is a special case of estimating a mean, with

$$p_U = \frac{1}{N}\sum_{k=1}^{N} y_k = \overline{y}_U$$

and estimator

$$\widehat{p}_S = \overline{y}_S = \frac{1}{n}\sum_{k \in S} y_k,$$

the proportion of 1's in the sample.

Can show that

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - p_U)^2 = \frac{N}{N-1}p_U(1-p_U).$$

Hence, from (1), we have the estimator,

$$\mathrm{var}(\widehat{p}_S) = \left(\frac{N-n}{N-1}\right)\frac{p_U(1-p_U)}{n}.$$

Estimator of a proportion under SRS

We have the unbiased estimator,

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{N}(y_k - \widehat{p}_S)^2 = \frac{n}{n-1}\widehat{p}_S(1-\widehat{p}_S).$$

From (2), we therefore have the estimator,

$$\widehat{\text{var}}(\widehat{p}_S) = \left(1 - \frac{n}{N}\right)\frac{\widehat{p}_S(1-\widehat{p}_S)}{n-1}.$$

Estimator of population total and properties under SRS

We can simply extend these results to the population total:

$$t = \sum_{k=1}^{N} y_k = N\overline{y}_U.$$

We have the unbiased estimator

$$\widehat{t} = N\overline{y}_S = \frac{N}{n}\sum_{k\in S} y_k. \tag{16}$$

From (1) we have variance

$$\text{var}(\,\widehat{t}\,) = N^2\text{var}(\overline{y}_S) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n} \tag{17}$$

and estimated variance

$$\widehat{\text{var}}(\,\widehat{t}\,) = N^2\text{var}(\overline{y}_S) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}. \tag{18}$$

Estimator of $\overline{y}_U$ and properties under SRS

Lohr (2010, Section 2.5) provides an accessible discussion of the use of confidence intervals in survey sampling, and notes that sample sizes needed for (3) to be accurate are often larger than we are used to in non-survey situations. A formula for a recommended “minimum *n* for accuracy of CI” is provided.

The percentiles of a Student’s *t* distribution with $n - 1$ degrees of freedom may replace the normal percentile points in (3).

Technical Appendix: Stratified SRS

Proportions

Using (4)–(7) with $\overline{y}_{hS} = \widehat{p}_{hS}$ and $s_h^2 = \frac{n_h}{n_h - 1}\widehat{p}_{hS}(1 - \widehat{p}_{hS})$.

Then,

$$\widehat{p}_{\text{strat}} = \sum_{h=1}^{H} \frac{N_h}{N} \widehat{p}_{hS}$$

and

$$\widehat{\text{var}}(\widehat{p}_{\text{strat}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\widehat{p}_{hS}(1 - \widehat{p}_{hS})}{n_h - 1}.$$

The total number possessing the attribute of interest is estimated as

$$\widehat{t}_{\text{strat}} = \sum_{h=1}^{H} N_h \widehat{p}_{hS}.$$

and $\widehat{\text{var}}(\widehat{t}_{\text{strat}}) = N^2 \widehat{\text{var}}(\widehat{p}_{\text{strat}})$.

Selection probabilities

π_{hk} is the probability of selecting unit k in stratum h, $k = 1, \ldots, N_h$, $h = 1, \ldots, H$.

$\pi_{hkh'l}$ is the joint probability of selecting unit k in stratum h, and unit l in stratum h', $k, l = 1, \ldots, N_h$, $k \neq l$, $h, h' = 1, \ldots, H$.

For SRS within each sample,

$$\pi_{hk} = \frac{n_h}{N_h}$$

to give sampling weights, $w_{hk} = \pi_{hk}^{-1} = N_h/n_h$.

Joint probabilities, for $k \neq l$:

$$\begin{aligned} \pi_{hkh'l} &= \pi_{hk} \times \pi_{h'l} \\ &= \frac{n_h}{N_h} \times \frac{n_{h'}}{N_{h'}} && \text{for } h \neq h' \\ &= \frac{n_h}{N_h} \times \frac{n_{h'} - 1}{N_{h'} - 1} && \text{for } h = h' \end{aligned}$$

Small samples

If a stratum has $n_h = 0$ the sampling weight is infinite, so if you don't sample from every stratum you can't get an unbiased estimate of the population total or mean.

Recall we need $\pi_{hk} > 0$ for everyone in the population.

If $n_h = 1$ it is not possible to get an unbiased estimate of the standard error (recall we need $\pi_{hkh'l} > 0$ for all k and l): to estimate the variability within a stratum takes at least two observations in the stratum.

If $n_h = 2$ we are OK with respect to unbiasedness and variance estimation; designs with $n_h = 2$ are susceptible to non-response, however, since this can be easily reduced to $n_h = 0$ or $n_h = 1$.

If $N_h = n_h = 1$ then we are OK because we know the answer and so the variance is zero!

Small samples

Tricks for handling $n_h = 1$:

- Collapse two strata (but NOT based on the observed values).
- Use the estimated population standard deviation instead of the sample standard deviation in the standard error formula.
- Replace the standard error for that stratum by the average standard error for all strata with $n_h > 1$.

Collapsing the strata is popular.

If we don't sample from all stratum (and we don't collapse), we are probably doing cluster sampling.

Design issues

We need to consider what our objective is, to we wish to estimate a total or a mean across the whole population, or do we want estimates of the mean or total in each stratum (i.e., as in domain estimation).

In the latter case we can take n_h sufficiently large to gain the required precision in each stratum.

For design in the case of a population total or mean, we can think about what the strata should be, and how many samples to pick in each strata; we first consider the latter problem.

Proportional allocation

We first consider proportional allocation which assigns $n_h \propto N_h$ and so makes the sample a small version of the population.

The inclusion probabilities are $\pi_{hk} = n_h/N_h$ and are the same $(= n/N)$ for all strata h (self-weighting).

In a population of 2400 men and 1600 women a 10% proportional allocation sample would sample 240 men and 160 women; sample weights are $1/\pi_{1k} = 1/\pi_{2k} = 1/10 = 10$.

It can be shown (Lohr, 2010, Section 3.4.1) that when the strata are large enough, the variance of the mean (or the total) under proportional allocation stratified sampling will almost always be less than under SRS.

Design: optimal allocation

If the variances S_h^2 are similar across strata, proportional allocation is a good option for increasing precision (as we will see mathematically below).

If S_h^2 vary greatly, optimal allocation can increase precision over proportional allocation: we should sample a greater fraction of larger units.

Let C represent total cost, c_0 baseline costs, and c_h the (known) cost of sampling an observation in stratum h so that total cost is

$$C = c_0 + \sum_{h=1}^{H} c_h n_h.$$

A natural criteria is to minimize $\text{var}(\overline{y}_{\text{strat}})$ for a given C and it can be shown (using Lagrange multipliers) that the optimal allocation is to take n_h proportional to

$$\frac{N_h S_h}{\sqrt{c_h}}.$$

Design: optimal allocation

The optimal size is

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}} \right) n.$$

We sample more in a stratum if:

1. The stratum accounts for a large part of the population.
2. The variance within the stratum is large; larger samples acknowledge the heterogeneity. If there were no heterogeneity then we would only require a single sample from that stratum.
3. Sampling in the stratum is inexpensive.

Neyman allocation, so called because of the derivation in Neyman (1934), is a special case in which n_h is proportional to $N_h S_h$.

Design: optimal allocation

The variances S_h^2 are not known, but in repeated survey there may be estimates from previous surveys.

If the variances and costs are equal across stratum we obtain proportional allocation.

The optimal allocations will vary if different variables (i.e. different y's) are examined.

Given the potential non-response, will need to increase the numbers beyond the optimal (see the NMIHS document).

A selected unit k in stratum h represents $w_{hk} = \frac{N_h}{n_h}$ units in the population, and for non-proportional allocation this will lead to different sample weights in different strata (though constant within a strata).

Technical Appendix: Cluster Sampling

Notation

PSU level, population quantities:

- $N =$ number of PSUs in the population.
- $M_i =$ number of SSUs in PSU i.
- $M_0 = \sum_{i=1}^{N} M_i =$ total number of SSUs in the population.
- $t_i = \sum_{k=1}^{M_i} y_{ik} =$ total in PSU i.
- $t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{k=1}^{M_i} y_{ik} =$population total.
- $S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N}\right)^2 =$ population variance of the PSU totals.

Notation

SSU level, population quantities:

- $\overline{y}_U = \frac{1}{M_0} \sum_{i=1}^{N} \sum_{k=1}^{M_i} y_{ik} =$ population mean.
- $\overline{y}_{iU} = \frac{1}{M_i} \sum_{k=1}^{M_i} y_{ik} = \frac{t_i}{M_i}$ population mean in PSU i.
- $S^2 = \frac{1}{M_0 - 1} \sum_{i=1}^{N} \sum_{k=1}^{M_i} (y_{ik} - \overline{y}_U)^2 =$ population variance (per SSU).
- $S_i^2 = \frac{1}{M_i - 1} \sum_{k=1}^{M_i} (y_{ik} - \overline{y}_{iU})^2 =$ population variance within PSU i.

Notation

Sample quantities:

- $n =$ number of PSUs in the sample.
- $m_i =$ number of SSUs in the sample.
- $S =$ the random variable describing the sampled PSUs.
- $S_i =$ the random variable describing the sample of SSUs in cluster i.
- $\overline{y}_{iS} = \sum_{k \in S_i} \frac{y_{ik}}{m_i} =$ sample mean (per SSU) for PSU i.
- $\widehat{t}_i = \sum_{k \in S_i} \frac{M_i}{m_i} y_{ik} =$ estimated total for PSU i.
- $\widehat{t}_{\text{unb}} = \sum_{i \in S} \frac{N}{n} \widehat{t}_i =$ unbiased estimator of population total.
- $s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\widehat{t}_i - \frac{\widehat{t}_{\text{unb}}}{N} \right)^2 =$ population variance of the PSU totals.
- $s_i^2 = \frac{1}{m_i - 1} \sum_{k \in S_i} (y_{ik} - \overline{y}_{iS})^2 =$ sample variance within PSU i.
- $w_{ik} =$ sampling weight for SSU j in PSU i.

Notation

Notes:

- This notation is close to that introduced by Lohr (2010, Section 5.1).
- N denotes the number of clusters from which we may sample n (by analogy with SRS).
- In one-stage cluster sampling the number of sampled is equal to the number in the SSU, i.e. $m_i = M_i$, if cluster i is selected.
- This notation will be used for one-stage, two-stage and multistage sampling.

Estimation: Unbiased estimation for one-stage cluster sampling

There are two problems with the unbiased estimators:

1. We may only know M_i for the sampled clusters and M_0 may not be known.
2. The variance may be large because it depends on the variance of the cluster totals

   $$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\widehat{t}_{\text{unb}}}{N} \right)^2 ,$$

   which may be large, particular if the M_i vary a lot (greater totals will often be associated with greater numbers of units).

An alternative approach is provided by ratio estimation.

A tangent on design-based (randomization) inference

Design-based inference is fundamentally frequentist and in its most extreme form, model free.

The frequentist slant suggests that estimators will be judged in terms of their frequentist properties, over repeat samples being taken.

Estimators are judged via mean-squared error and its two components, bias and variance.

There is no universal prescription for deriving estimators, which may be seen as a disadvantage, or as an advantage, depending on your statistical convictions.

In the Bayesian approach, inference is completely prescriptive.

This is illustrated in this section where we derive two different estimators, one unbiased and the other intended to be efficient (low variance).

Efficiency of sampling compared to SRS

We now compare the efficiency of stratified random sampling and cluster sampling as compared to SRS.

We introduce the design effect (abbreviated to deff), which is defined (Kish, 1965) as

$$\text{deff} = \frac{\text{var(estimator from sampling plan)}}{\text{var(estimator from SRS)}} \tag{19}$$

where both designs use the same number of observations.

This is a very important summary that is often used to compare designs.

The denominator in (19) is $\left(1 - \frac{n}{N}\right) \frac{S^2}{n}$.

Efficiency of stratified sampling compared to SRS

Table 2: ANOVA table for stratified sampling; SSB and SSW are the sums of squares between and within strata and SSTO is the total sum of squares.

Source	df	Sum of Squares
Between strata	$H - 1$	SSB = $\sum_{h=1}^{H} \sum_{k=1}^{N_h} (\overline{y}_{hU} - \overline{y}_U)^2$ $= \sum_{h=1}^{H} N_h(\overline{y}_{hU} - \overline{y}_U)^2$
Within strata	$N - H$	SSW = $\sum_{h=1}^{H} \sum_{k=1}^{N_h} (y_{hk} - \overline{y}_{hU})^2$ $= \sum_{h=1}^{H} (N_h - 1)S_h^2$
Total	$N - 1$	SSTO $= \sum_{h=1}^{H} \sum_{k=1}^{N_h} (y_{hk} - \overline{y}_U)^2 = (N - 1)S^2$

Efficiency of stratified sampling compared to SRS

In a stratified sample with proportional allocation ($n_h/N_h = n/N$):

$$\mathrm{var}_{\text{strat}}(\widehat{t}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h} = \left(1 - \frac{n}{N}\right) \frac{N}{n} \left(\text{SSW } + \sum_{h=1}^{H} S_h^2 \right),$$

so that it is the within-strata variability that is key.
For SRS:

$$\begin{aligned} \mathrm{var}_{\text{SRS}}(\widehat{t}) &= \left(1 - \frac{n}{N}\right) N^2 \frac{S^2}{n} \\ &= v_{\text{strat}}(\widehat{t}) + \left(1 - \frac{n}{N}\right) \frac{N}{n(N-1)} \left[N \times \text{SSB} - \sum_{h=1}^{H} (N - N_h) S_h^2 \right] \end{aligned}$$

Proportional allocation stratified sampling always gives smaller variance than SRS unless

$$\text{SSB } < \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_h^2.$$

The more unequal the stratum means, the more efficiency is gained.

Inefficiency of cluster sampling compared to SRS

We saw that for stratified sampling the variance is small if SSW is small relative to SSTO.

For simplicity consider one-stage cluster sampling with equal numbers in each SSU ($= M$).

Table 3: ANOVA table for cluster sampling; SSB and SSW are the sums of squares between and within PSUs and SSTO is the total sum of squares.

Source	df	Sum of Squares
Between PSUs	$N - 1$	SSB = $\sum_{i=1}^{N} \sum_{k=1}^{M} (\overline{y}_{iU} - \overline{y}_U)^2$
Within PSUs	$N(M - 1)$	SSW = $\sum_{i=1}^{N} \sum_{k=1}^{M} (y_{ik} - \overline{y}_{iU})^2$
Total	$NM - 1$	$\text{SSTO} = \sum_{i=1}^{N} \sum_{k=1}^{N_h} (y_{ik} - \overline{y}_U)^2 = (NM - 1)S^2$

Inefficiency of cluster sampling compared to SRS

Note that

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i - \overline{t}_U)^2 = \frac{1}{N-1}\sum_{i=1}^{N} M^2(y_{iU} - \overline{y}_U)^2 = M \times \frac{\text{SSB}}{N-1}$$

Hence, for cluster sampling

$$\text{var}_{\text{clust}}(\widehat{t}) = N^2\left(1 - \frac{n}{N}\right)\frac{M}{n} \times \frac{\text{SSB}}{N-1}$$

so that it is the between-cluster variability that is key.
For SRS with *nM* observations:

$$\begin{aligned}\text{var}_{\text{SRS}}(\widehat{t}) &= (NM)^2\left(1 - \frac{nM}{NM}\right)\frac{S^2}{nM} \\ &= N^2\left(1 - \frac{n}{N}\right)\frac{M}{n} \times S^2\end{aligned}$$

So cluster sampling is less efficient than SRS if

$$\frac{\text{SSB}}{N-1} > S^2.$$

Intraclass correlation coefficient

We have seen that the efficiency of cluster sampling depends on the between-cluster variability, in contrast to stratified sampling.

The intraclass (or intracluster) correlation coefficient (ICC) measures the homogeneity within the clusters, i.e. how similar observations in the same cluster are.

It is the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ik}, y_{il}) for $i = 1, \ldots, N$ and $k \neq l$ and can be written

$$\text{ICC} \;= 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}. \tag{20}$$

If the clusters are perfectly homogenous SSW$=0$ and ICC$=1$. From (20),

$$\frac{\text{SSB}}{N-1} = \frac{NM-1}{M(N-1)} S^2[1 + (M-1)\text{ICC}].$$

Intraclass correlation coefficient

Hence,

$$\frac{\text{var}_{\text{clust}}(\widehat{t})}{\text{var}_{\text{SRS}}(\widehat{t})} = \frac{\text{SSB}}{N-1} S^2 = \frac{NM-1}{M(N-1)}[1+(M-1)\text{ICC}].$$

If the number of PSUs in the population N is large so that

$$NM-1 \approx M(N-1)$$

then the ratio of the variances is approximately

$$1+(M-1)\text{ICC}.$$

So $1+(M-1)$ICC SSUs, taken in a one-stage cluster sample, gives us approximately the same amount of information as one SSU from an SRS.

Intraclass correlation coefficient

Example: If ICC $= 0.5$ and $M = 5$ then

$$1 + (M - 1)\text{ICC} = 3$$

and we would need to take 300 elements using a cluster sample to get the same precision as 100 elements from an SRS.

But often it is much cheaper to logistically carry out cluster sampling.

Ratio estimation for two-stage cluster sampling

The ratio estimator for the population mean is

$$\widehat{\overline{y}}_{\mathrm{r}} = \frac{\sum_{i \in S} \widehat{t}_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \overline{y}_i}{\sum_{i \in S} M_i}$$

and an approximation to the variance is available (Lohr, 2010, p. 186).

Technical Appendix: Lonely PSUs

Strata with only one PSU (Lumley 2010, Section 3.2.1)

We have discussed multistage sampling scheme in which cluster sampling is carried out within strata.

Consider a stratum with only one PSU (cluster); the sampling probability for this stratum must be 1 under stratified sampling.

This stratum will not contribute to the first stage of the variance calculation, but may contribute to later stages: in this case the first stage of sampling may be ignored.

If the stratum contains more than one potential PSU, but only one is sampled then this violates our requirement of probability, namely that all pairs of units can be sampled (units in different PSUs in the stratum cannot both be selected).

This situation occurs because of non-response, or because fine stratification is carried out.

Strata with only one PSU

The best way of dealing with a stratum with a ‘lonely’ PSU is to combine with another stratum, based on population data (which is available before sampling. Basing on sample data will lead to bias.

Some NHANES studies use this approach, taking one PSU per stratum and then creating ‘pseudo-strata’, with two PSUs, for analysis.

In the `survey` package, if a strata with a single PSU is detected the default behavior is to report an error. Two solutions are also provided, however.

Strata with only one PSU

Setting `options(survey.lonely.psu="adjust")` gives a conservative variance estimate that uses residuals from the population mean rather than from the stratum mean.

Setting `options(survey.lonely.psu="average")` sets the variance contribution to the average for all strata with more than one PSU, and this is also conservative.

When there is only a single population PSU in a stratum it should be clear from the fpc information that the sampling fraction is 100%; if the population size information is not supplied, single-PSU strata can be dropped from the variance calculation with `options(survey.lonely.psu="remove")`.

The same adjustments are applied to each level of sampling, e.g., to second-stage strata with only a single SSU.