Bayesian SAE using Complex Survey Data: Methods and Applications
Lecture 1: Complex Survey Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

CSSS Short Course: Thursday 23rd May, 2019

Outline

Motivation

Overview of Survey Sampling

Design-Based Inference

Simple Random Sampling

Stratified Simple Random Sampling

Cluster Sampling

Multistage Sampling

Discussion

Motivation

Motivating Example 1: BRFSS

- Arises out of a joint project between Laina Mercer/Jon Wakefield and Seattle and King County Public Health, which lead to the work reported in Song *et al.* (2016).
- We aim to estimate the number of 18+ individuals with diabetes, by health reporting areas (HRAs) in King County in 2011.
- HRAs are city-based sub-county areas with 48 in King County.
- Some are a single city, some are a group of smaller cities, and some are unincorporated areas. Larger cities such as Seattle and Bellevue include more than one HRA.

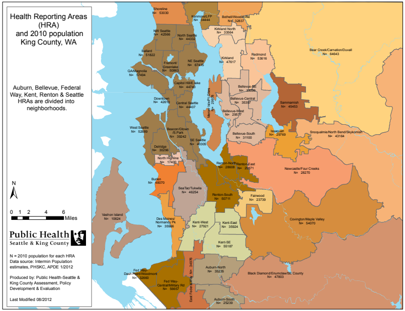


Figure 1: Health reporting areas (HRAs) in King County.

- Data are based on the question, “Has a doctor, nurse, or other health professional ever told you that you had diabetes?”, in 2011.

Motivating Example 1: BRFSS

- Estimates are used for a variety of purposes including summarization for the local communities and assessment of health needs.
- Analysis and dissemination of place-based disparities is of great importance to allow efficient targeting of place-based interventions.
- Because of its demographics, King County looks good compared to other areas in the U.S., but some of its disparities are among the largest of major metro areas.

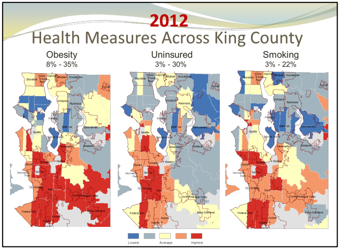
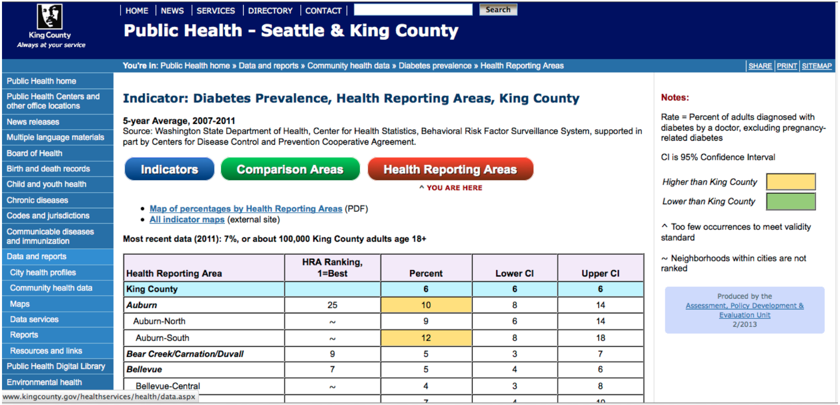


Figure 2: Summaries from Public Health: Seattle King County.

BRFSS

- Estimation is based on Behavioral Risk Factor Surveillance System (BRFSS) data.
- The BRFSS is an annual telephone health survey conducted by the CDC that tracks health conditions and risk behaviors in the United States and its territories since 1984.



Public Health - Seattle & King County

You're In: Public Health home » Data and reports » Community health data » Diabetes prevalence » Health Reporting Areas

Indicator: Diabetes Prevalence, Health Reporting Areas, King County

5-year Average, 2007-2011
Source: Washington State Department of Health, Center for Health Statistics, Behavioral Risk Factor Surveillance System, supported in part by Centers for Disease Control and Prevention Cooperative Agreement.

- Map of percentages by Health Reporting Areas (PDF)
- All indicator maps (external site)

Most recent data (2011): 7%, or about 100,000 King County adults age 18+

Health Reporting Area	HRA Ranking, 1=Best	Percent	Lower CI	Upper CI
King County		**6**	**6**	**6**
Auburn	25	10	8	14
Auburn-North	~	9	6	14
Auburn-South	~	12	8	18
Bear Creek/Carnation/Duvall	9	5	3	7
Bellevue	7	5	4	6
Bellevue-Central	~	4	3	8
		7	4	10

Notes:

Rate = Percent of adults diagnosed with diabetes by a doctor, excluding pregnancy-related diabetes

CI is 95% Confidence Interval

Higher than King County	🟨
Lower than King County	🟩

^ Too few occurrences to meet validity standard

~ Neighborhoods within cities are not ranked

Produced by the
Assessment, Policy Development & Evaluation Unit
2/2013

Those Mysterious Weights

The BRFSS sampling scheme is complex: it uses a disproportionate stratified sampling scheme.

The SampleWt, is calculated as the product of four terms

$$\texttt{Sample Wt} = \texttt{StratWt} \times \frac{1}{\texttt{NoTelephones}} \times \texttt{NoAdults} \times \texttt{PostStratWt}$$

where StratWt is the inverse probability of a “likely” or “unlikely” stratum being selected (stratification based on county and “phone likelihood”).

Table 1: Summary statistics for population data, and 2011 King County BRFSS diabetes data, across health reporting areas.

	Mean	*Std. Dev.*	*Median*	*Min*	*Max*	*Total*
Population ($>$18)	31,619	10,107	30,579	8,556	56,755	1,517,712
Sample Sizes	62.9	24.3	56.5	20	124	3,020
Diabetes Cases	6.3	3.1	6.3	1	15	302
Sample Weights	494.3	626.7	280.4	48.0	5,461	1,491,880

BRFSS: Sample Sizes

- A total of $3,020$ individuals answered the diabetes question.
- About 35% of the areas have sample sizes less than 50 (CDC recommended cut-off), so that the diabetes prevalence estimates are relatively unstable in these areas.
- We would like to use the totality of the data to aid in estimation in the data sparse areas.

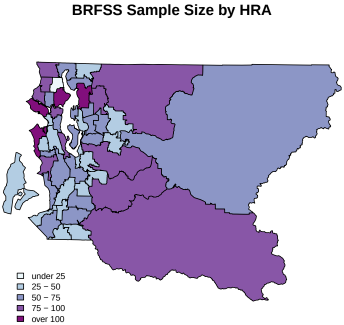


Figure 3: Sample sizes across 48 HRAs.

BRFSS: Comparison of Estimates

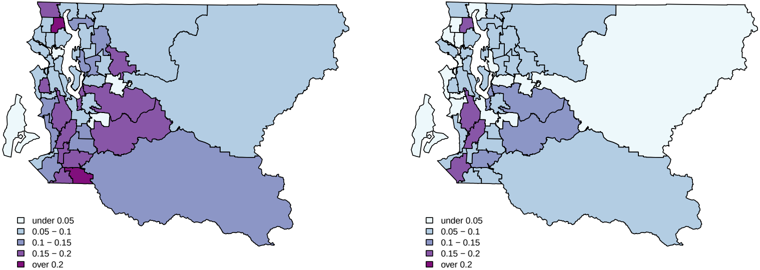


Figure 4: Diabetes prevalence by HRAs in 2011: crude proportions (left) Horvitz-Thompson weighted estimates (right).

Motivating Example 2: NHANES

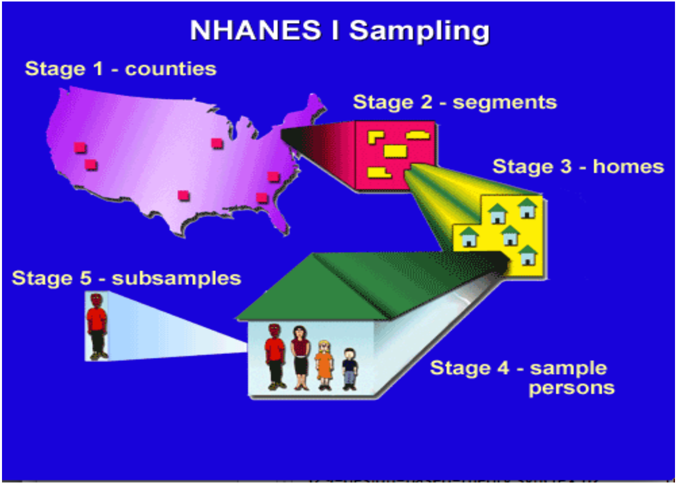


Figure 5: Cartoon of sample design in NHANES I; a multistage stratified clustered sample of civilian, non-institutionalized population.

NHANES: Study Design

- In NHANES, participants had an interview, clinical examination and blood samples were taken and needed to be stored, and this required mobile examination trailers.
- 27,000 individuals were sampled over 4 years and not practical to move the trailers to thousands of locations.
- Figure 6 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.

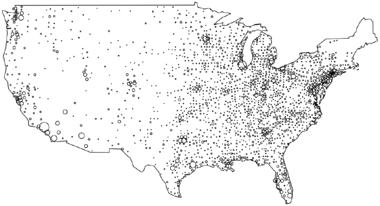


Figure 6: A SRS of 10,000 voter locations from the USA with circles at the county centroids and areas proportional to the number sampled. Los Angeles County contains the largest sample of 257.

Motivating Example 3: DHS

- Motivation: In many developing world countries, vital registration is not carried out, so that births and deaths go unreported. We aim to provide reliable U5MR estimates at the Admin 1 level, at which policy interventions are often carried out. We use data from Demographic Health Surveys (DHS).
- DHS Program: Typically stratified cluster sampling to collect information on population, health, HIV and nutrition; more than 300 surveys carried out in over 90 countries, beginning in 1984.
- The Problem: Data are sparse at the Admin 1 level.
- SAE: Leverage space-time similarity to construct a Bayesian smoothing model.

Kenyan DHS

- The 3 most recent Kenya DHS were carried out in 2003, 2008 and 2014.
- These DHS use stratified (urban/rural, 8 regions), two-stage cluster sampling (enumeration areas, and then households).
- All women age 15 to 49 who slept in the household the night before were interviewed in each selected household.
- DHS provides sampling (design) weights, assigned to each individual in the dataset. From each full birth history information was obtained.

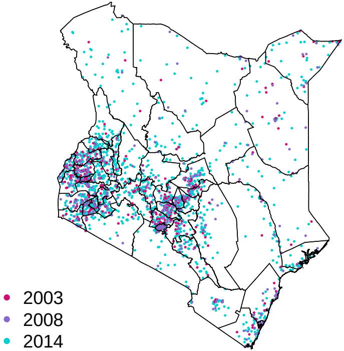


Figure 7: Cluster locations in three Kenya DHS, with county boundaries.

Small Area Estimation

Small Area Estimation concerns making areal estimates of a quantity of interest, with the data in some areas being possibly sparse.

Spatial Statistics covers many endeavors:

- Disease Mapping: Spatial dependence is a virtue.
- Spatial Regression: Spatial dependence is a nuisance – confounding by location.
- Cluster Detection: Spatial pattern of data is of primary interest.
- Assessment of Clustering: Spatial pattern of data is of primary interest.
- Small Area Estimation: Spatial dependence is a virtue.

Course Overview

Lectures:

- Complex Survey Data.
- Bayesian Smoothing Models.
- Small-Area Estimation.
- Examples with the `SUMMER` package.

Website:

http://faculty.washington.edu/jonno/CSSS20-Spatial.html

Overview of Survey Sampling

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.

Some Reference Books

- 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.

Requirements

- 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 (i.e., we have a census), we are done! No statistics needed...
- A random sample is almost always better than a non-random sample, because the former allows a more straightforward assessment of uncertainty.

Design-Based Inference

- 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.
- Imagine a population of size $N = 4$ and we sample $n = 2$
- Possible samples, with sampled unit indices in red and non-sampled in blue:

$$y_{\color{red}1}, y_{\color{red}2}, y_{\color{blue}3}, y_{\color{blue}4}$$

$$y_{\color{red}1}, y_{\color{blue}2}, y_{\color{red}3}, y_{\color{blue}4}$$

$$y_{\color{red}1}, y_{\color{blue}2}, y_{\color{blue}3}, y_{\color{red}4}$$

$$y_{\color{blue}1}, y_{\color{red}2}, y_{\color{red}3}, y_{\color{blue}4}$$

$$y_{\color{blue}1}, y_{\color{red}2}, y_{\color{blue}3}, y_{\color{red}4}$$

$$y_{\color{blue}1}, y_{\color{blue}2}, y_{\color{red}3}, y_{\color{red}4}$$

- Different designs follow from which probabilities we assign to each of these possibilities.

Design-Based Inference

Design-based inference 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.
- May be worse in the sense of precision, but be required logistically.

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 via known sampling probabilities.

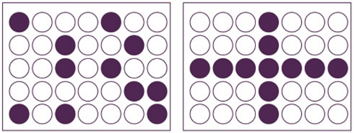
For example, in stratified random sampling, certain groups may have fixed numbers sampled.

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.

Probability Samples

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

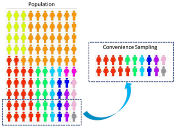


Probability Sampling Vs Non-Probability Sampling

- Non-probability samples cannot be analyzed with design-based approaches, because there are no π_k.

Non-probability sampling approaches include:

- Convenience sampling (e.g., asking for volunteers). Also known as accidental or haphazard sampling.



- 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, for example, the interviewer may choose friendly looking people!).

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.

Probability Samples: Point Estimation

For design-based inference:

- To obtain an unbiased estimator, every individual k in the population to have a non-zero probability

$$\pi_k$$

  of being sampled, $k = 1, \ldots, N$.
- To carry out inference, this probability π_k must be known only for every individual in the sample.
- So not needed for the unsampled individuals, which is key to implementation, since we will usually not know the sampling probabilities for those not sampled.

Probability Samples: Variance Estimation

For design-based inference:

- 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,

  $$\pi_{kl}$$

  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.
- in practice, these are often approximated, or the variance is calculated via a resampling technique such as the jackknife.

Inference

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

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

Model-Based Inference

- Infinite population view: The population variables are drawn from a hypothetical distribution, the model, $p(\cdot)$ 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 $p(\cdot)$.
- As an example, we take the sample mean:

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

  is a random variable because $Y_1, \dots, Y_n$ are each random variables.
- Assume Y_k are iid observations from a distribution, $p(\cdot)$, with mean μ and variance σ^2.
- The sample mean is an unbiased estimator, and has variance σ^2/n.

Model-Based Inference

Unbiased estimator:

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

Variance:

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

In general, can write down a sampling model and then proceed with likelihood or Bayesian inference.

Design-Based Inference

- In the design-based approach to inference the y values are treated as unknown but fixed.
- To emphasize: the y's are not viewed as random variables, so we write

$$y_1, \ldots, y_N,$$

  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.

Design-Based Inference

- Define design weights as

$$w_k = \frac{1}{\pi_k}.$$

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

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

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).

The Indices are Random!

- **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

$$\pi_k = \Pr(\text{ individual } i \text{ in sample }) = \frac{3}{6} = \frac{1}{2},$$

  which is of course $\frac{n}{N}$.
- 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 sum of the design weights,

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

  is the total population.
- 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.

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., E$[\widehat{\overline{Y}}_U] = \overline{Y}_U$, where we average over all possible samples we could have drawn, i.e., over *S*.

Unbiasedness

- For many designs: $\sum_{k\in S} w_k = N$ so we demonstrate with the estimator
$$\widehat{\overline{Y}}_U = \frac{1}{N}\sum_{k\in S} w_k y_k .$$
- There's a neat trick in here, we introduce an indicator random variable of selection $I_k \sim \text{Bernoulli}(\pi_k)$:

$$\begin{aligned}
\mathrm{E}[\widehat{\overline{Y}}_U] &= \underbrace{\mathrm{E}\left[\frac{1}{N}\sum_{k\in S} w_k y_k\right]}_{S\text{ is random in here}} = \underbrace{\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} I_k w_k y_k\right]}_{I_k\text{ are random in here}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[I_k\right] w_k y_k = \frac{1}{N}\sum_{i=1}^{N}\pi_k w_k y_k = \frac{1}{N}\sum_{i=1}^{N} y_k = \overline{Y}_U
\end{aligned}$$

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"[1], a $100(1-\alpha)\%$ confidence interval for $\overline{y}_U$ is

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

  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).

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

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 Simple 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: Washington State

- According to

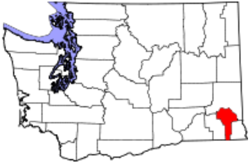
  `http://quickfacts.census.gov/qfd/states/53000.html`

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

Notation

- Stratum levels are denoted $h = 1, \ldots, H$, so H in total.
- Let $N_1, \ldots, N_H$ be the known population totals in the stratum with

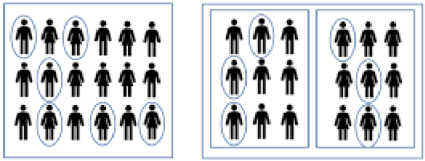
$$N_1 + \cdots + N_H = N,$$

so that N is the total size of the population.

- In stratified simple 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

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

Figure 1: Comparison of Simple Random Sampling to Stratified Random Sampling



Visual of Simple Random Sampling: Selection of 6 out of 18 People

Visual of Stratified Random Sampling: Selection of 3 out of 9 Men and 3 out of 9 Women

- We can view stratified SRS 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.

Estimators

- The sampling probaiblities for unit k in strata h are

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

  which do not depend on k.

- Therefore the design weights are

$$w_{hk} = \frac{N_h}{n_h}.$$

- Note that:

$$\sum_{h=1}^{H} \sum_{k \in S_h} w_{hk} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{N_h}{n_h} = \sum_{h=1}^{H} n_h \frac{N_h}{n_h} = N$$

Estimators

- Weighted estimator:

$$\begin{aligned}\widehat{\overline{y}}_U &= \frac{\sum_{h=1}^H \sum_{k \in S_h} w_{hk} y_{hk}}{\sum_{h=1}^H \sum_{k \in S_h} w_{hk}} \\ &= \sum_{h=1}^H \frac{N_h}{N} \overline{y}_{hS}\end{aligned}$$

  where

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

- Since we are sampling independently from each stratum using SRS, we have

$$\text{var}(\widehat{\overline{y}}_U) = \sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} \tag{4}$$

  where the within stratum variances are:

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

Weighted Estimation

Recall: The weight w_k can be thought of as the number of people in the population represented by sampled person k.

Example 1: Simple Random Sampling

Suppose an area contains 1000 people:

- Using simple random sampling (SRS), 100 people are sampled.
- Sampled individuals have weight $w_k = 1/\pi_k = 1000/100 = 10$.

Example 2: Stratified Simple Random Sampling

Suppose an area contains 1000 people, 200 urban and 800 rural.

- Using stratified SRS, 50 urban and 50 rural individuals are sampled.
- Urban sampled individuals have weight $w_k = 1/\pi_k = 200/50 = 4$.
- Rural sampled individuals have weight $w_k = 1/\pi_k = 800/50 = 16$.

Weighted Estimation

Example 2 Revisited: Stratified Simple Random Sampling

Suppose an area contains 1000 people, 200 urban and 800 rural.

- Urban risk = 0.1.
- Rural risk = 0.2.
- **True risk = 0.18.**

Take a stratified SRS, 50 urban and 50 rural individuals sampled:

- Urban sampled individuals have weight 4; 5 cases out of 50.
- Rural sampled individuals have weight 16; 10 cases out of 50.
- **Simple mean is** $15/100 = 0.15 \neq 0.18$**.**
- **Weighted mean is**

$$\frac{4 \times 5 + 16 \times 10}{4 \times 50 + 16 \times 50} = \frac{180}{1000} = 0.18.$$

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: 1988 NMIHS

Table 2: 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.
- We estimate the mean as

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

- 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 (4), as

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

- 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 for Cluster Sampling

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.

Cluster Sampling

The clusters can be:

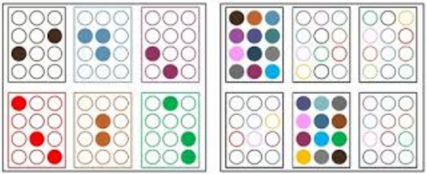
- Genuine features of the populations, e.g., households, schools, or workplaces.
- Subsets chosen for convenience, e.g., counties, zipcodes, telephone number blocks.

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.

Differences between cluster and stratified 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 $\bar{y}_U$ 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



Stratified Sampling Vs Cluster Sampling

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

- We suppose that a SRS of n PSUs is taken.
- 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

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

Estimation: Unbiased estimation for one-stage cluster sampling

- 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 the population mean is

$$\overline{y}_U \quad = \quad \frac{1}{M_0} \sum_{i=1}^{N} \sum_{k=1}^{M_i} y_{ik}$$

- An unbiased estimator is

$$\widehat{\overline{y}}_U = \frac{\sum_{i \in S} \sum_{k \in S_i} w_{ik} y_{ik}}{M_0}.$$

- Then,

$$\widehat{\text{var}}(\widehat{\overline{y}}) \quad = \quad = \frac{N^2}{M_0^2} \left(1 - \frac{n}{N}\right) \frac{s_T^2}{n}$$

where s_T^2 is the estimated variance of the PSU totals.

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.

Two-Stage Cluster Sampling Weights

- The inclusion probabilities are:

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

- Hence, the weights are

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

- An unbiased estimator is

$$\widehat{\overline{y}}_U = \frac{\sum_{i \in S} \sum_{k \in S_i} w_{ik} y_{ik}}{M_0}.$$

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

Variance Estimation for Two-Stage Cluster Sampling

- In contrast to one-stage cluster sampling we have to acknowledge the uncertainty in both stages of sampling; in one-stage cluster sampling the totals 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

$$M_0^2 \text{var}(\widehat{\overline{y}}_U) = \underbrace{N^2\left(1-\frac{n}{N}\right)\frac{s_T^2}{n}}_{\text{one-stage cluster variance}} + \underbrace{\frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}}_{\text{two-stage cluster variance}} \tag{5}$$

  where

  - s_T^2 are the estimated variance of the cluster totals,
  - s_i^2 is the estimated 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 (5) is zero.

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.
- 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.
- We will not go into inference for this design, but basically weighted estimates are readily available, and accompanying variance estimates can be calculated.
- Weighted estimators are used and a common approach to variance estimation is the jackknife (Pedersen and Liu, 2012)

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.

Pedersen, J. and Liu, J. (2012). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Medicine*, **9**, e1001289.

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

Song, L., Mercer, L., Wakefield, J., Laurent, A., and Solet, D. (2016). Peer reviewed: Using small-area estimation to calculate the prevalence of smoking by subcounty geographic areas in King County, Washington, behavioral risk factor surveillance system, 2009–2013. *Preventing Chronic Disease*, **13**.