2022 SISCER Module 2 Small Area Estimation: Lecture 1: Motivation and Survey Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Overview

Motivation

- Small area estimation (SAE) entails estimating characteristics of interest for domains, often geographical areas, in which there may be few or no samples available – “small” refers to the number of samples in, and not to the geographical size of, the areas.

- SAE has a long history and a wide variety of methods have been suggested, from a bewildering range of philosophical standpoints.

- Application areas include: epidemiology, public and global health, agriculture, economics, education,...

- The classic text is Rao (2003) which was updated to Rao and Molina (2015).

Data

Examples we discuss include:

- Subnational variation in the under 5 mortality rate (U5MR) in low- and middle-income countries (LMIC).
- Diabetes prevalence in King County health reporting areas.
- Corn and Soy crop yield in Iowa counties.
- Poverty mapping in Spanish regions.

Health and Demographic Indicators

My own research focusses on health/demographic indicators in LMICs:

- Charactering and understanding subnational variation is an important public health endeavor.
- For example, in the Sustainable Development Goals (SDGs), Goal 3.2 states, “By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births”.
- Many other indicators have SDG targets, including poverty.

Prevalence Mapping and Geostatistics

- Examination of proportions across space, is known as prevalence mapping – we may map continuously in space, or across discrete administrative areas – we focus on the latter.

- SAE methods provide one approach to performing prevalence mapping, for administrative areas.

- "The term geostatistics is a short-hand for the collection of statistical methods relevant to the analysis of geolocated data, in which the aim is to study geographical variation throughout a region of interest, but the available data are limited to observations from a finite number of sampled locations." (Diggle and Giorgi, 2019).

- Model-based geostatistics (MBG) provide another approach to performing prevalence mapping, over continuous space, though these continuous surfaces can be averaged for area-level inference.

Characterization of Methods and Approaches

Some important distinctions:

Area-level	versus	Unit-level	Modeling
Direct	versus	Indirect	Estimation
Linear	versus	Non-Linear	Modeling
No Auxiliary	versus	Auxiliary	Data
Non-spatial	versus	Spatial	Mixed Modeling
Design-based	versus	Model-based	Inference
Frequentist	versus	Bayesian	Inference

In general, the lack of information in small samples is compensated for by:

- The use of covariates (auxiliary variables) in a regression model.
- Employing smoothing via mixed effects models, perhaps including spatial smoothing.

Overview of Short Course

- Data: We consider the common situation in which the available data arise from surveys with a complex design.
- A Problem: If small sample sizes in some areas/time periods, there is high instability. In the limit, there may be no data...
- Supplementary Data: On covariates to aid in modeling.
- Survey Sampling Methodology: Required for design and analysis.
- Shrinkage and Spatial Smoothing: To reduce instability, use the totality of data to smooth both locally and globally over space.
- Different Approaches to SAE: Both traditional and Bayesian methods that use spatial smoothing.
- Implementation: In `R` programming environment, using the `SUMMER` package.
- Visualization: Maps of uncertainty, accompanied with uncertainty, produced using the GIS capabilities of `R`.

Overview of Short Course

Lectures:

- Lecture 1: Motivation and approaches to analyzing complex survey data.
- Lecture 2: Mixed effects area-level models.
- Lecture 3: Mixed effects unit-level models.

Methods illustrated in `R`, in particular using the `sae` and `SUMMER` packages.

Course website:

`http://faculty.washington.edu/jonno/SISCER-SAE.html`

My SAE Background

My own interest in SAE:

- Started with work on BRFSS, with local government.
- Moved to estimating subnational estimation of U5MR, neonatal mortality, vaccination, HIV prevalence,... in LMIC.

Details on my research is here:

`http://faculty.washington.edu/jonno/space-station.html`

Demographic Health Surveys

- Motivation: In many developing world countries, vital registration is not carried out, so that births and deaths go unreported.
- Objective: To provide reliable estimates of demographic/health indicators at the (say) Admin1 or Admin2 level[1], at which policy interventions are often carried out.
- We will illustrate using 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 Admin2 level in particular.
- SAE: Leverage space-time similarity to construct a Bayesian smoothing model.

[1] Admin0 = country level boundaries, Admin1 = first level administrative boundaries (states in US), Admin 2 = second level administrative boundaries (counties in US)

2014 Kenyan DHS

- The 3 most recent Kenya DHS were carried out in 2003, 2008 and 2014.
- The DHS use stratified two-stage cluster sampling. The strata consist of urban/rural crossed with geographic administrative strata.
- In each strata, enumeration areas (EAs) are selected with probability proportional to size using a sampling frame developed from the most recent census.
- In each of the clusters, households are selected. Within each household, women between the ages of 15 and 49 are interviewed.

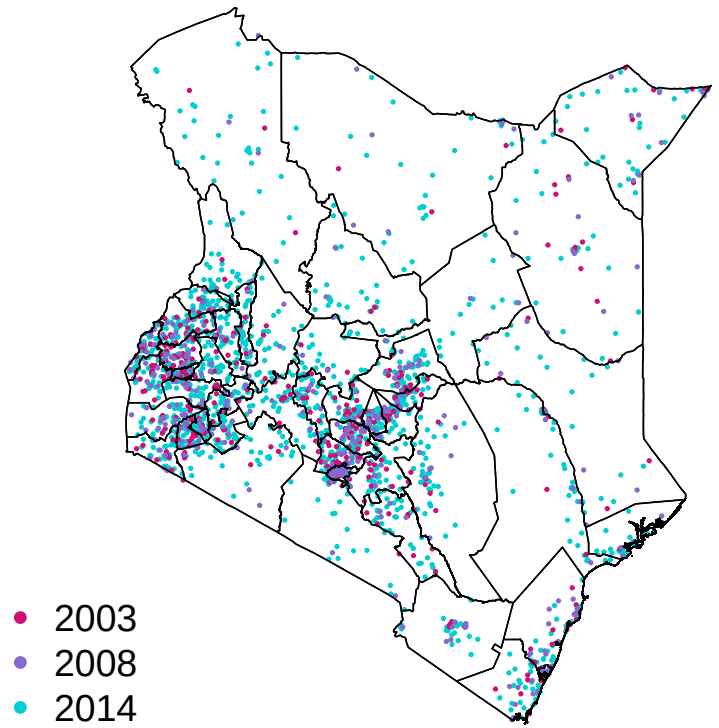


Figure 1: Cluster locations in three Kenya DHS, with county boundaries.

2014 Kenya DHS

- In the 2014 Kenya DHS, the stratification was county (47) and urban/rural (2).
- Nairobi and Mombasa are entirely urban, so there are 92 strata in total.
- We have data from a total of 1584 EAs across the 92 strata. In the second stage, 40,300 households are sampled.
- DHS provides sampling (design) weights, assigned to each individual in the dataset.

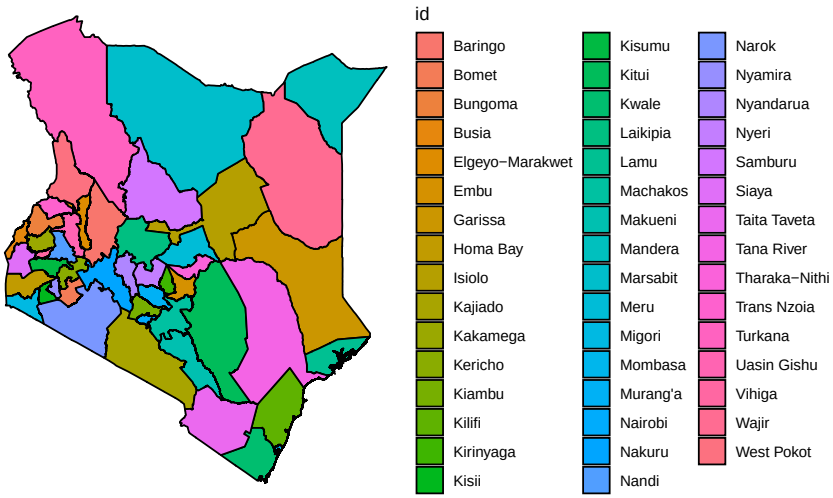


Figure 2: Counties of Kenya.

Aim: Inference for U5MR over Counties and Years

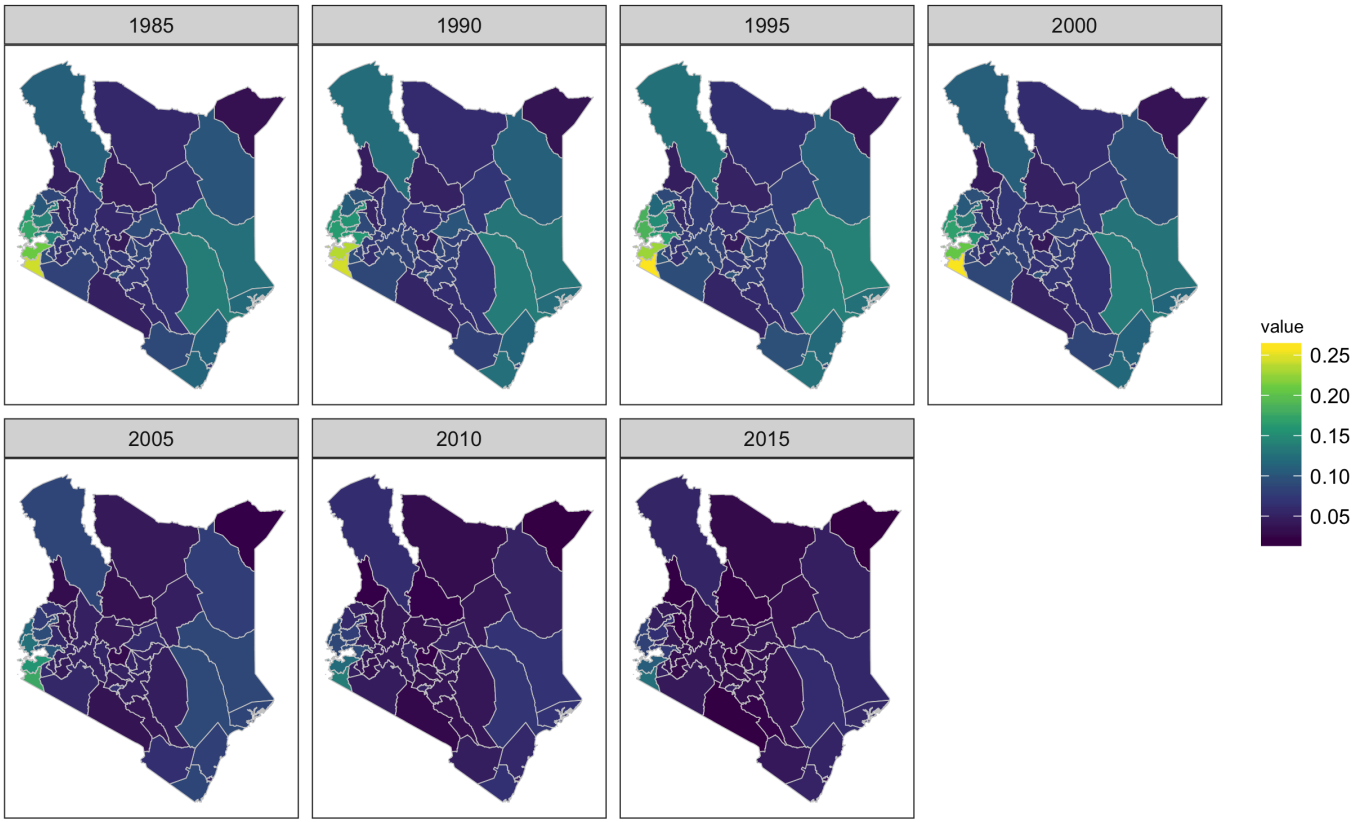


Figure 3: SAE estimates of under-5 mortality risk, across time, and Kenyan counties. These estimates were obtained using the `SUMMER` package.

2013 Nigeria DHS

- As a second DHS example, we consider measles vaccination rates in Nigeria, from the 2013 Nigerian DHS.
- Across African countries, there is great variability in the number of Admin2 areas.
- In Nigeria, the Admin2 areas correspond to Local Government Areas (LGAs) and there are 774 in total – with such a large number there are many LGAs with little/no data.
- There are no clusters in 255 LGAs.

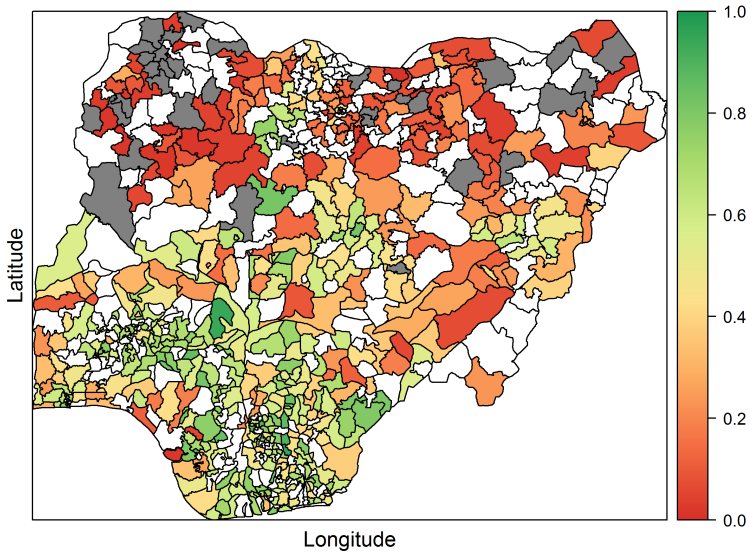


Figure 4: Vaccination prevalence for LGAs in Nigeria. LGAs with no data are in white.

Motivating Example: Diabetes in King County

Arises out of a joint project between me and Laina Mercer and Seattle and King County Public Health, which lead to the work reported in Song *et al.* (2016).

Aim we will concentrate on here is to estimate the number of 18 years or older individuals with diabetes, by health reporting areas (HRAs) in King County in 2011.

HRAs are city-based sub-county areas with a total of 48 HRAs in King County. Some of these are as 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.

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

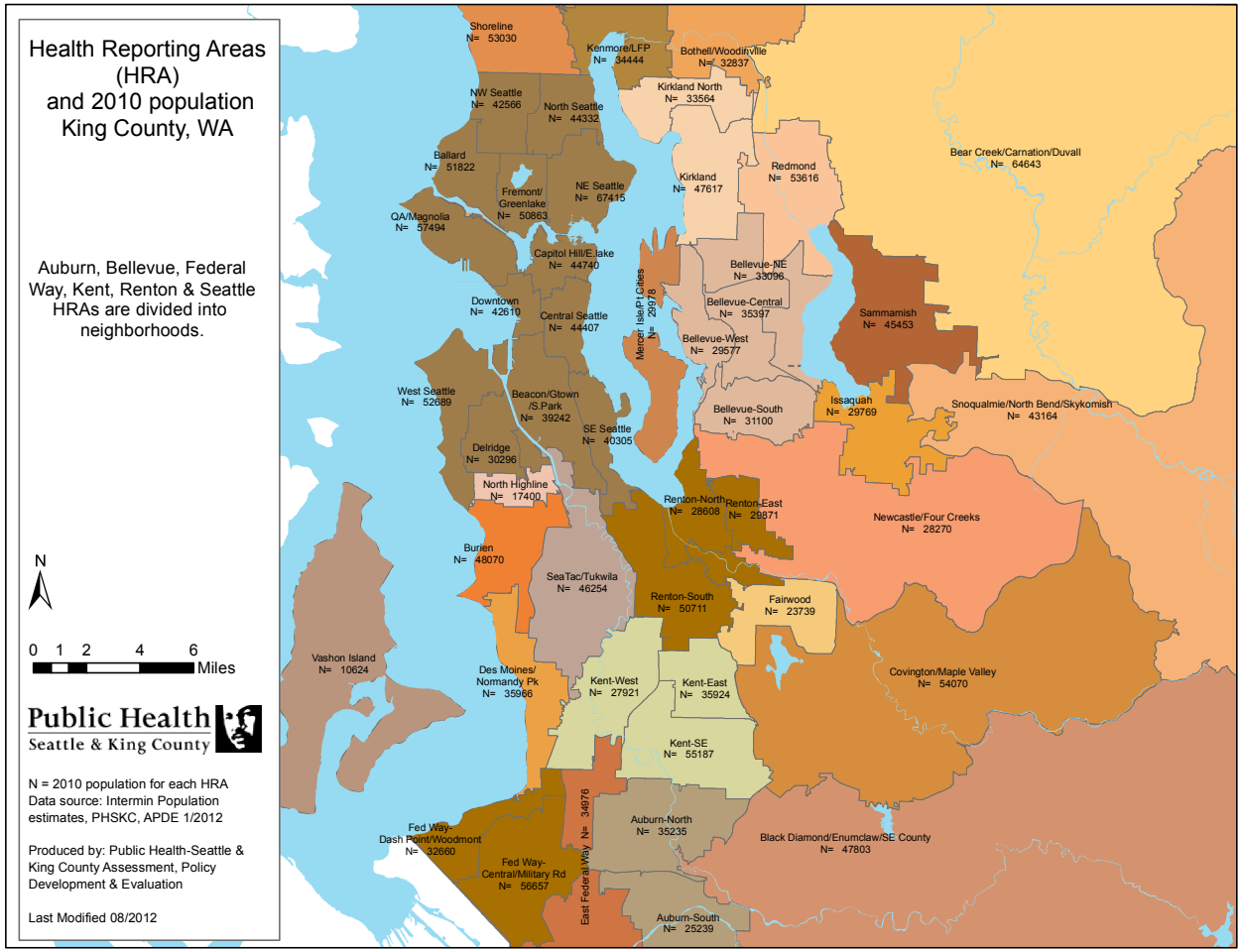


Figure 5: Health reporting areas (HRAs) in King County.

Motivating BRFSS Example

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.

Estimation is based on Behavioral Risk Factor Surveillance System (BRFSS) data.

The BRFSS is an annual telephone health survey conducted by the Centers for Disease Control and Prevention (CDC) that tracks health conditions and risk behaviors in the United States and its territories since 1984.

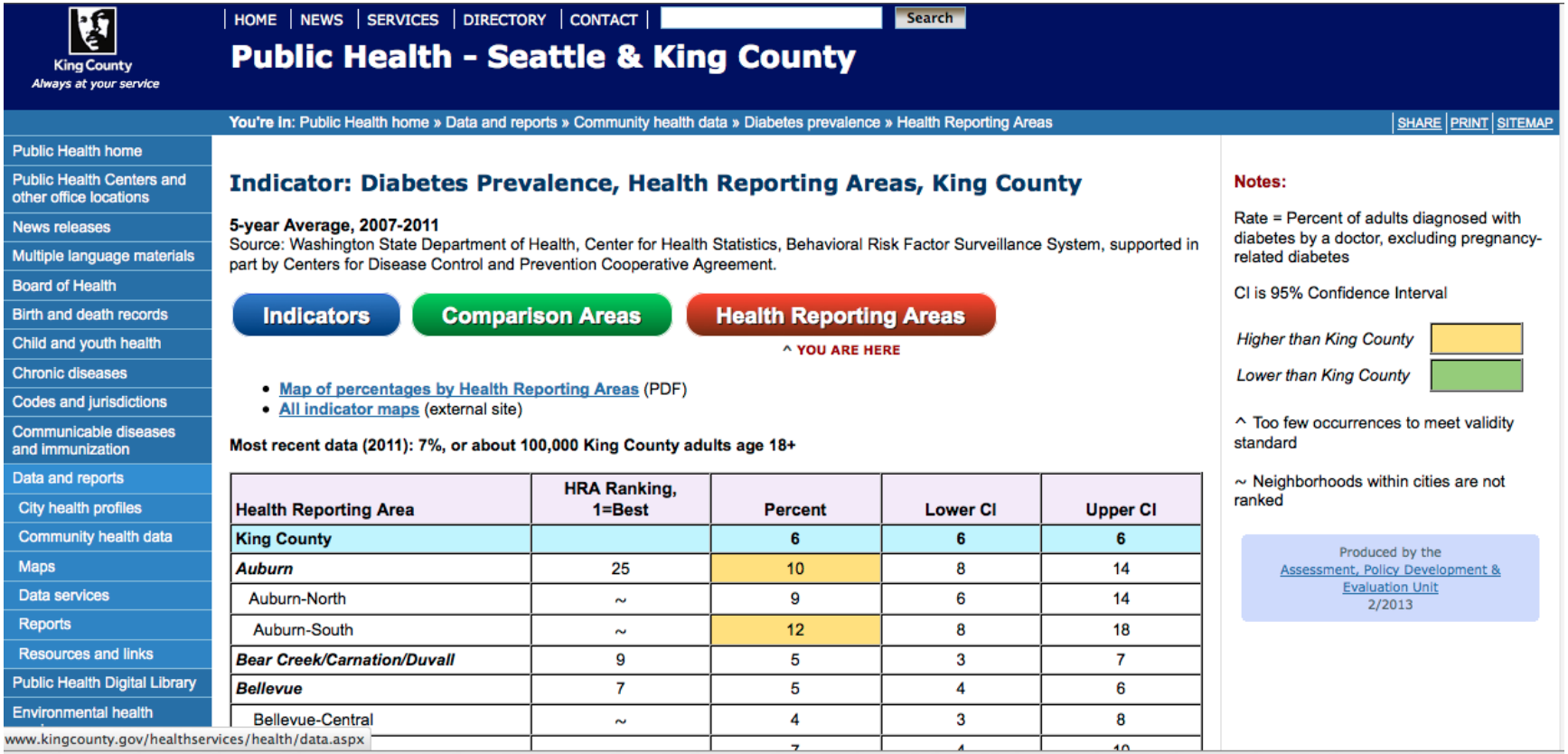


Figure 6: Public Health: Seattle and King County website.

Motivating BRFSS Example

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

The `Design-Wt`, is calculated as the product of four terms

$$\texttt{Design-Wt} = \texttt{Strat-Wt} \times \frac{1}{\texttt{No-Telephones}} \times \texttt{No-Adults}$$

where `Strat-Wt` is the inverse probability of a "likely" or "unlikely" stratum being selected (stratification based on county and "phone likelihood").

Then a raking adjustment. From the documentation, "BRFSS rakes the design weight to 8 margins (age group by gender, race/ethnicity, education, marital status, tenure, gender by race/ethnicity, age group by race/ethnicity, phone ownership). If BRFSS includes geographic regions, four additional margins (region, region by age group, region by gender, region by race/ethnicity) are included."

Motivating BRFSS Example

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

About 35% of the areas have sample sizes less than 50 (CDC recommended cut-off), so that the diabetes prevalence estimates are unstable in these areas.

We would like to use the totality of the data to aid in estimation in the data sparse areas.

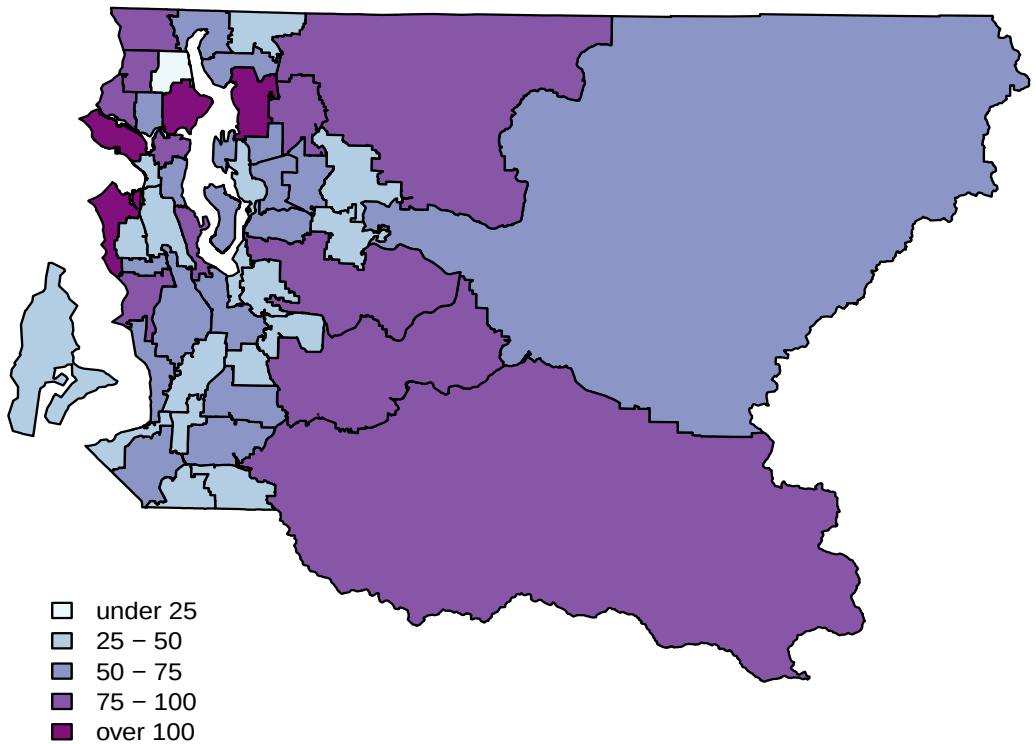


Figure 7: Sample sizes across 48 HRAs in 2011.

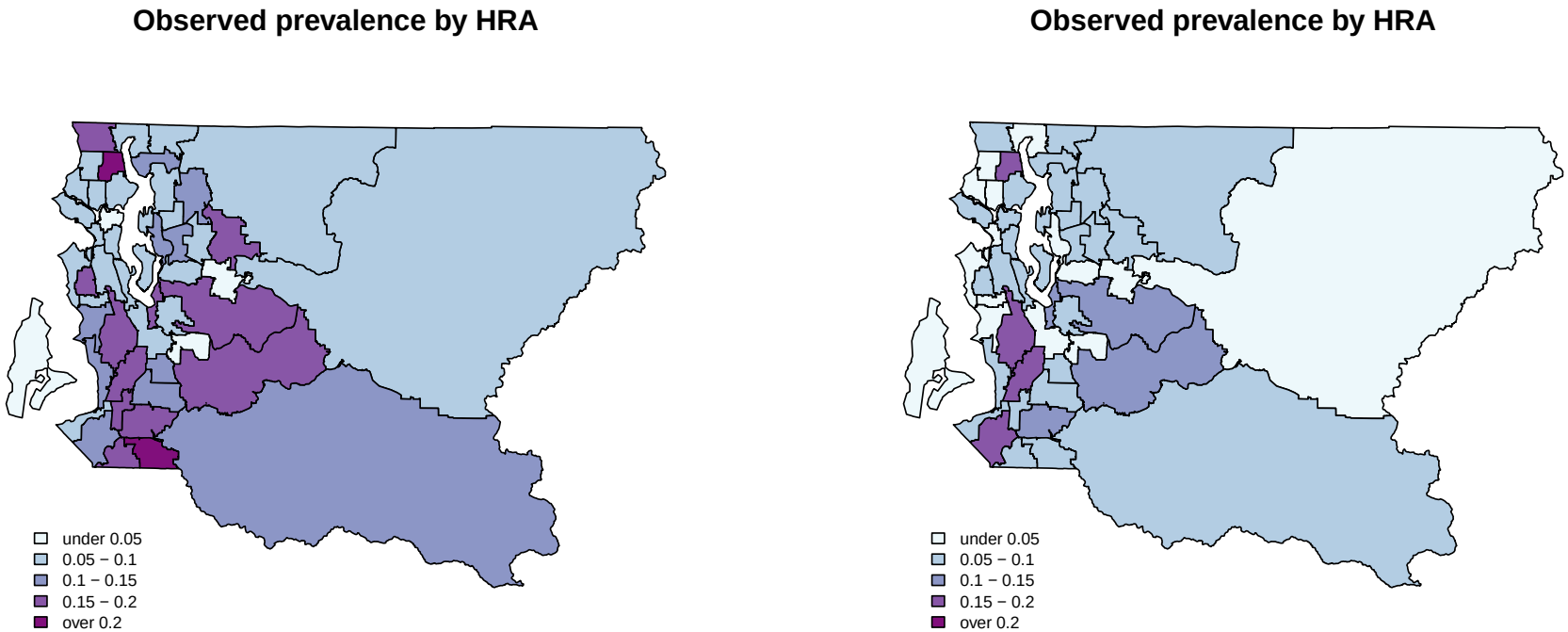
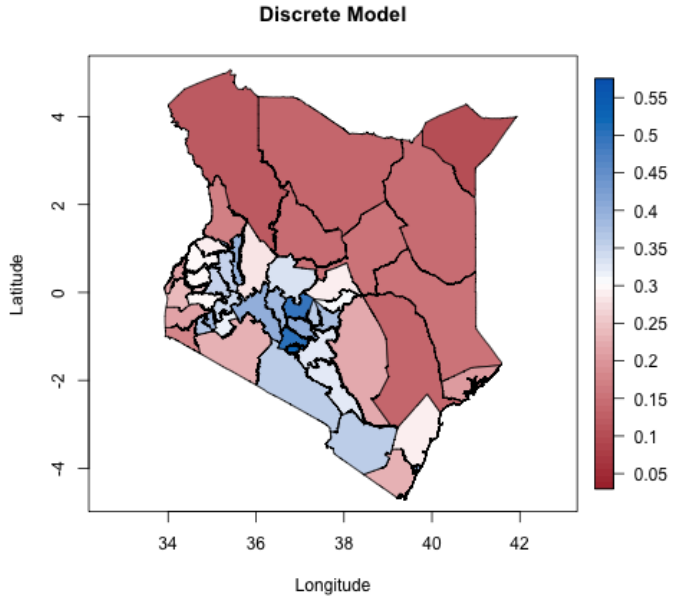


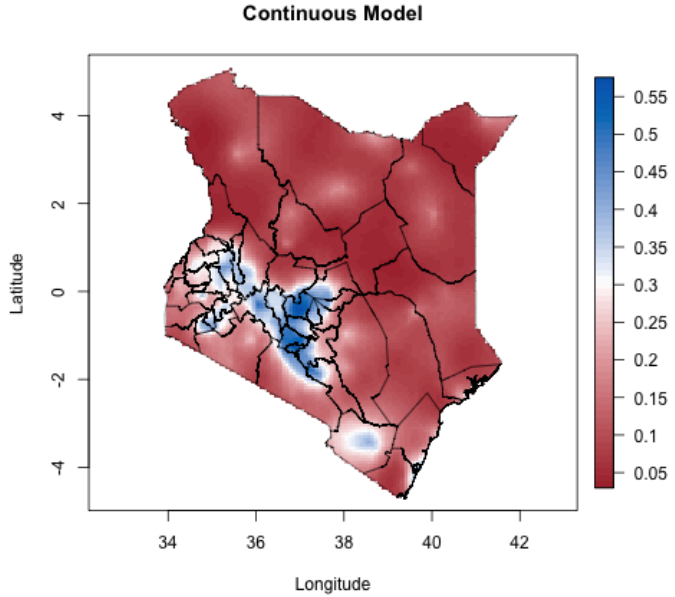
Figure 8: Diabetes prevalence by HRAs in 2011. Left: Crude proportions. Right: Horvitz-Thompson weighted estimator.

Two Approaches to Spatial Smoothing

- Model at the area level using a discrete spatial model. These are the SAE models that are implemented in the SUMMER package.



- Model at the point level using a continuous spatial model. Model-based geostatistics is a popular approach.



2013 Nigeria DHS

- Recall that almost a third of the LGAs in Nigeria have no data (left plot below).
- We fit a discrete spatial model in which the rates in neighboring areas (as defined by sharing a boundary) are "encouraged" to be similar (right plot below).

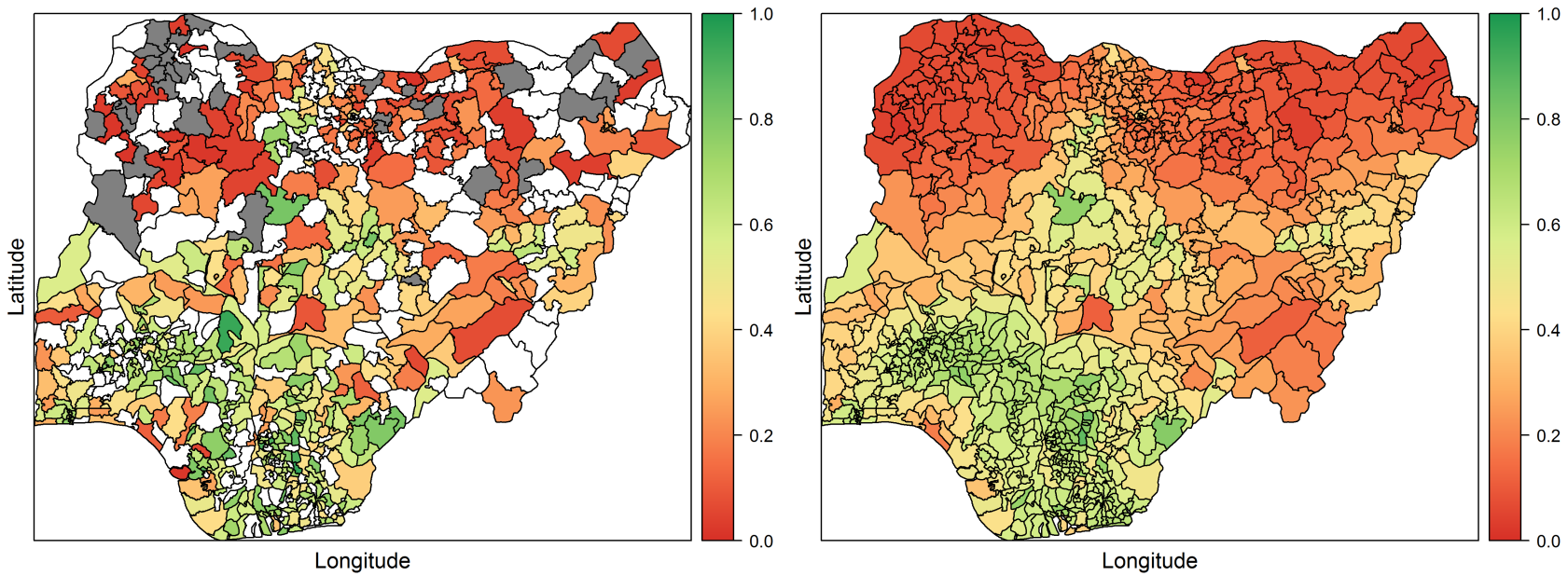


Figure 9: Vaccination prevalences in Nigeria in 2013. Left: Weighted estimates. Right: Estimates from a discrete spatial smoothing model.

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.

We will discuss:

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

First, we briefly explain why taking account of the survey design (data collection process) is important.

Acknowledging the Design: Stratification

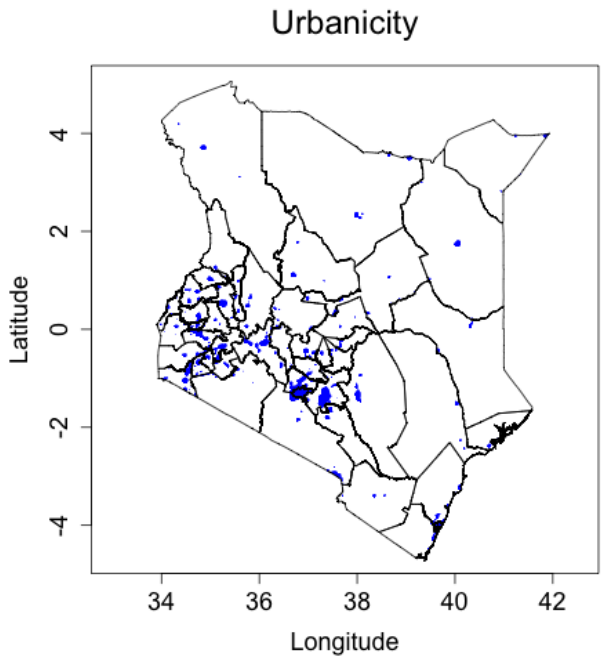


Figure 10: In the DHS, stratification is based on counties (the solid lines) and on a binary urban/rural variable (urban indicated in blue, the white is rural).

- Suppose we are interested in the proportion of women aged 20–29 who complete secondary education – this is much higher in urban areas
- If we oversample urban areas but ignore this when we analyze the data we will overestimate the fraction of women who complete secondary education, i.e., we will introduce bias.
- Taking into account of the stratification also reduces the variance of the estimator.
- In the design-based approach to inference, the stratification is accounted for via design weights.
- In the model-based approach to inference, the stratification is accounted for in the mean model.

Acknowledging the Design: Cluster Sampling

- The DHS also employs cluster sampling, in which multiple units (individuals) within the same cluster are interviewed.
- Units within the same cluster tend to be more similar than units in different clusters, which reduces the information content of the clustered sample, relative to independently sampled units.
- The dependence can be measured via the intraclass correlation coefficient.
- In the design-based approach to inference, the clustering is accounted for in the variance calculation that is carried out.
- In the model-based approach to inference, the clustering is accounted for by including a cluster-specific random effect in the model.

Modes of Inference

- Surveys can be analyzed using **design- and model-based inference**. In this lecture, the former will be focused upon.
- The target of inference are the set of means for areas indexed by i (e.g., Admin2 regions).
- Let y_{ik} be the binary indicator on the k-th unit sampled in area i, for $k \in S_i$ (the set of selected individuals) and $i = 1, \ldots, n$.

Design-Based Inference

- Labels S_i of sampled units are random.
- Responses y_{ik} are fixed.
- Asymptotic inference, perhaps using resampling.

Model-Based Inference

- Condition on units that are actually sampled.
- Responses Y_{ik} are random.
- Exact inference, conditional on model.

Model-Based Inference

Suppose we carry out stratified cluster sampling, with one-stage of clusters, and the outcome is continuous.

Let y_{ck} be the outcome from sampling unit k in sampled cluster c, and $\boldsymbol{s}_c$ the location of cluster c,

Suppose the data were collected within two strata, urban and rural.

A model-based approach to inference might begin with

$$Y_{ck} = \alpha + \gamma \underbrace{I(\boldsymbol{s}_c \in \text{ rural })}_{\text{indicator for rural}} + \epsilon_c + \upsilon_{ck},$$

where

- α is the mean for urban and $\alpha + \gamma$ is the mean for rural.
- within-cluster dependence is modeled via the random effect $\epsilon_c \sim_{iid} \text{N}(0, \sigma^2_\epsilon)$.
- Measurement error is $\upsilon_{ck} \sim_{iid} \text{N}(0, \sigma^2_\upsilon)$.

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, while the indices of the individuals who are sampled are random.
- Imagine a population of size $N = 4$ and we sample $n = 2$
- There are 6 possible samples, with sampled unit indices in red and non-sampled in blue:

$$y_1, y_2, y_3, y_4$$
$$y_1, y_2, y_3, y_4$$
$$y_1, y_2, y_3, y_4$$
$$y_1, y_2, y_3, y_4$$
$$y_1, y_2, y_3, y_4$$
$$y_1, y_2, y_3, y_4$$

- Different designs are possible, and the probabilities we assign to each sample depend on which is used.

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 simple random sample (SRS) in the sense of obtaining the same precision at lower cost, e.g., stratified sampling.
- May be worse in the sense of precision, but be required logistically, e.g., cluster sampling.

Probability Samples

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

- N is population size.
- n is sample size.
- π_k is the 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.

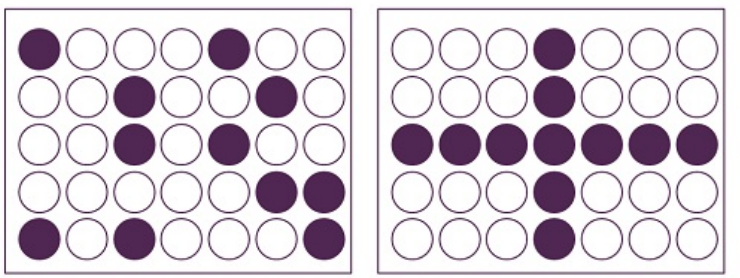
For example, in stratified random sampling, the probabilities of selection differ, in different strata.

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). Clusters are then sampled from this the set of PSUs, with units within these clusters being subsequently sampled.
- Multistage sampling: Stratified cluster sampling, with multiple levels of clustering.

Probability Samples

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

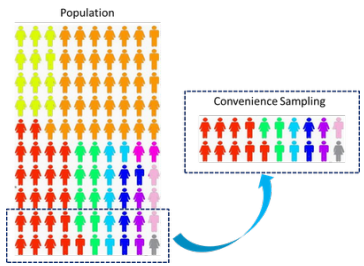


Probability Sampling Vs Non-Probability Sampling

- 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 uses 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!).

Probability Samples: Point Estimation

For design-based inference:

- To obtain an unbiased estimator, every individual k in the population needs to have a non-zero probability π_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 units, k and l, in the sample, there must a non-zero probability of being sampled together, call this probability, π_{kl} for units k and l, $k = 1, \ldots, N$, $l = 1, \ldots, N$, $k \neq l$.
- 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, with mean μ.
- In the model-based view, $Y_1, \ldots, Y_N$ are random variables and properties are defined with respect to $p(\cdot)$; often we say Y_k are independent and identically distributed (iid) from $p(\cdot)$.
- As an estimator of μ, we may take the sample mean:

$$\widehat{\mu} = \frac{1}{n} \sum_{k=1}^{n} Y_k.$$

- $\widehat{\mu}$ is a random variable because $Y_1, \ldots, Y_n$ 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:

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

- Variance:

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

Model-Based Inference

- The variance σ^2 is unknown so we estimate by the unbiased estimator

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

- A 95% asymptotic confidence interval is,

$$\widehat{\mu} \pm 1.96 \times \frac{s}{\sqrt{n}}.$$

- In practice, "asymptotic" means that n is sufficiently large that the sampling distribution of $\widehat{\mu}$ (i.e., it's distribution in hypothetical repeated samples) is close to normal.

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 mean (Horvitz and Thompson, 1952; Hájek, 1971)

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

- This is an estimator of the finite population mean $\overline{y}_U$.
- So long as the weights are correctly calculated, and the sample size is not small, this estimator is appealing, though it may have high variance, if n is small.

The weighted mean is the basic **direct estimator** that is the first choice for SAE.

Simple Random Sample (SRS)

- The simplest probability sampling technique is simple random sampling without replacement.
- 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.

Simple Random Sample (SRS)

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

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

  $k = 1, \ldots, N$.

- Since sampling without replacement 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, of the estimator can be easily estimated.
  - Unless n is quite close to N, the uncertainty does not depend on N, only on n.

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

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

  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 $w_k = 1/\pi_k$ 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

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

  the sample mean, which is reassuring under SRS!

- This is an unbiased estimator, i.e.,

$$\mathrm{E}\left[\widehat{\overline{y}}_U\right] = \overline{y}_U,$$

  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 examine 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 $I_k \sim \text{Bernoulli}(\pi_k)$:

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

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

- It can be shown that the variance is

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

  where,

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

- Contrast (1) with the model-based variance which is σ^2/n.
- The factor

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

  is the famous finite population correction (fpc) factor.
- 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, is unknown, and we estimate using the unbiased estimator:

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

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

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

- The standard error is

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

- 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"[2], a 95% asymptotic confidence interval for $\overline{y}_U$ is

$$\overline{y} \pm 1.96 \times \sqrt{1 - \frac{n}{N}} \frac{s}{\sqrt{n}}. \tag{3}$$

- The interval given by (3) is random (across samples) because $\overline{y}$ 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).

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

Stratified 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 carried out independently.
- An important practical consideration of whether stratified sampling can be carried out is whether stratum membership is known for every individual in the population, i.e., we need a sampling frame containing the strata variable.

Rationale for Stratified Sampling

Lohr (2010, Section 3.1) provides a good discussion of the benefits of stratified sampling, we summarize here.

- 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 – this is usual for the DHS.
- For example, from the Kenya DHS sampling manual (Kenya National Bureau of Statistics, 2015):

  "The 2014 KDHS was designed to produce representative estimates for most of the survey indicators at the national level, for urban and rural areas separately, at the regional (former provincial) level, and for selected indicators at the county level."

Rationale for Stratified Sampling

- Flexible since sampling frames can be constructed differently in different strata.
- For example, one may carry out different sampling in urban and rural areas.
- 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.
- In a national study, the most natural form of sampling may be based on geographical regions.
- Due to the independent sampling in different stratum, variance estimation is straightforward, as long as within-stratum sampling variance estimators are available.

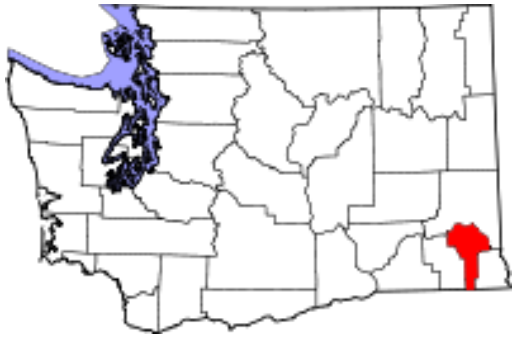
Example: Washington State

- According to the census there were 2,629,126 households in Washington State in the period 2009–2013.
- Consider a simple random sample (SRS) 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 (highlighted left) and Garfield (highlighted right) 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, which would not be a good idea given the objective here.
- Stratified sampling would allow control of the number of samples in each county.

Notation

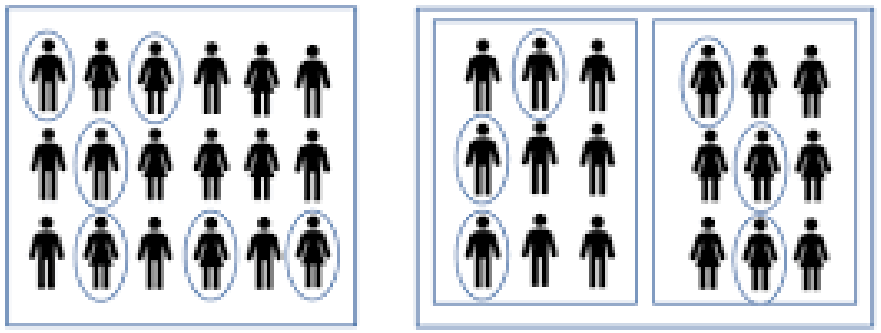
- Stratum levels are denoted $h = 1, \ldots, H$, so H in total.
- Let $N_1, \ldots, N_H$ be the known population totals in the stratum with

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

  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 probabilities 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:

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

  so that summing over the weights recovers the population size – this is consistent with the idea that each sampled individual represents a number of people (the weight) in the population.

Estimators

- Weighted estimator:

$$\widehat{\overline{y}}_U = \frac{\sum_{h=1}^{H}\sum_{k\in S_h} w_{hk}y_{hk}}{\sum_{h=1}^{H}\sum_{k\in S_h} w_{hk}} = \sum_{h=1}^{H}\frac{N_h}{N}\overline{y}_h$$

where

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

- Since we are sampling independently from each stratum using SRS, we have[3]

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

where the within stratum variances are:

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

[3] using the variance formula for SRS, (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: 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**

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

Motivation for Cluster Sampling

For logistical reasons, 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 for in-person interviews.
- It is far more cost effective (in terms of travel costs, etc.) to cluster sample.

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).
- This can clearly be extended to multistage cluster sampling.

Differences Between Cluster and Stratified sampling

Stratified Random Sampling	One-Stage Cluster Sampling
A sample 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	The cluster is the sampling unit and the more clusters sampled the smaller the variance – which depends primarily on between cluster means
For greatest precision, we want 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.
- The design effect (deff) is often to summarize the effect on the variance of the design:

$$\text{deff} = \frac{\text{Variance of estimator under design}}{\text{Variance of estimator under SRS}},$$

where in the denominator we use the same number of observations as in the complex design in the numerator.

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:

$$\pi_{ik} = \Pr(\text{ SSU } k \text{ in cluster } i \text{ is selected }) = \frac{n}{N}$$

$$w_{ik} = \text{Design weight for SSU } k \text{ in cluster } i = \frac{N}{n}.$$

Let S represent the set of sampled clusters.

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

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

- An unbiased estimator is

$$\widehat{\bar{y}}_U = \frac{\sum_{i \in s} \sum_{k \in s_i} w_{ik} y_{ik}}{M_0}.$$

- Then,

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

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.

We discuss the equal-probability two stage cluster design:

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

Two-Stage Cluster Sampling Weights

- The selection 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

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

- An unbiased estimator is

$$\widehat{\bar{y}}_U = \frac{\sum_{i \in s} \sum_{k \in s_i} w_{ik} y_{ik}}{M_0}.$$

- 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

$$\operatorname{var}(\widehat{\bar{y}}_U) = \frac{1}{M_0^2}\left[\underbrace{N^2\left(1-\frac{n}{N}\right)\frac{s_T^2}{n}}_{\text{One-stage cluster variance}} + \underbrace{\frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}}_{\text{Two-stage cluster variance}}\right] \tag{5}$$

  where
  - s_T^2 is the estimated variance of the cluster totals,
  - s_i^2 is the estimated variance within the i-th PSU.
- In most software packages, the second term in (5) is ignored, since it is small when compared to the first term, when N is large.

The Jackknife

- The jackknife is a very general technique for calculating the variance of an estimator.
- The basic idea is to delete portions of the data, and then fit the model on the remainder – if one repeats this process for different portions, one can empirically obtain the distribution of the estimator.
- The key is to select the carefully select the portion of the data so that the design is respected.
- We describe in the context of multistage cluster sampling.
- Observations within a PSU should be kept together when constructing the data portions, which preserves the dependence among observations in the same PSU.

The Jackknife for Multistage Cluster Sampling

- Assume we have H strata and n_h PSUs in strata h, and assume PSUs are chosen with replacement.
- To apply the jackknife, delete one PSU at a time.
- Let $\widehat{\mu}_{(hi)}$ be the estimator when PSU i of stratum h is omitted.
- To calculate $\widehat{\mu}_{(hi)}$ we define a new weight variable:

$$w_{k(hi)} = \begin{cases} w_{k(hi)} & \text{if observation } k \text{ is not in stratum } h \\ 0 & \text{if observation } k \text{ is in PSU } i \text{ of stratum } h \\ \frac{n_h}{n_h - 1} w_k & \text{if observation } k \text{ is not in PSU } i \text{ but in stratum } h \end{cases}$$

Then we can use the weights $w_{k(hi)}$ to calculate $\widehat{\mu}_{(hi)}$ and

$$\widehat{V}_{\text{JK}}(\widehat{\mu}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\widehat{\mu}_{(hi)} - \widehat{\mu})^2.$$

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).
- Weighted estimators are used and a common approach to variance estimation is the jackknife (Pedersen and Liu, 2012)
- In later lectures, we will show how model-based inference can be carried out for the DHS.

Use of Auxiliary Information

Synthetic Estimator

Many approaches have been suggested to obtain estimators with greater precision – we give a flavor by discussing some different approaches to modeling.

We consider estimation of a generic finite population mean, $\overline{y}_{Ui}$, in area i.

Synthetic Estimator

The **synthetic estimator**, based on $p-1$ covariates, is

$$\widehat{\overline{y}}_{Ui}^{syn} = \frac{1}{N_i}\sum_{k=1}^{N_i} \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{B} = \overline{\boldsymbol{x}}_i^{\mathsf{T}}\widehat{B},$$

where we have an appropriate weighted estimator,

$$\widehat{B} = \left[\sum_{i=1}^{n}\sum_{k\in S_i} w_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{x}_{ik}\right]^{-1} \sum_{i=1}^{n}\sum_{k\in S_i} w_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}y_{ik}.$$

This is an example of pseudo-likelihood estimation (Binder, 1983).

The area means of covariates needed across all of the population.

This approach assumes the regression model $\overline{y}_{ik} = \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{B}$ is appropriate for all areas.

In general, gives high precision estimates, but with a strong possibility of large bias as the model assumes all between-area variability arises from the differences in covariates, which is a heroic assumption.

Synthetic Estimator

If we fit a logistic regression model that includes auxiliary variables, then the procedure is more complex, and more information is needed.

- Consider the model
$$\text{logit}\, \overline{y}_{ik} = \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{B}.$$
- The weighted estimates of $\boldsymbol{B}$ are obtained through solving the weighted score, another example of pseudo-likelihood estimation.
- The estimate is:
$$\widehat{\overline{y}}_{Ui}^{syn} = \frac{1}{N_i} \sum_{k=1}^{N_i} \text{expit}\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \widehat{\boldsymbol{B}}\right),$$
so that the covariates are required for all members of the population (because of the nonlinear model, we can't get away with just the areas means).

Synthetic Estimator

Synthetic estimation is very naive but if only very few areas are sampled it can be used (since there aren't a lot of alternatives) – in this situation, one would hope there are lots of covariates with strong predictive power.

If there are sufficient samples within areas, it may be possible to estimate separate regression coefficients in each area, $\widehat{\boldsymbol{B}}_i$, which will reduce bias.

But we would like a method that allows for area-specific discrepancies from the regression model – an obvious extension is to add random effects, and we see methods of this type shortly.

Generalized Regression (GREG) Estimator

Now we consider general estimation of a mean or total when we have again have $p - 1$ variables to assist in the modeling, following Lohr (2010, Section 11.7).

In an SAE context, the GREG estimator attempts to correct for at least some of the bias in the synthetic estimator.

A detailed account of model-assisted estimators is given by Särndal *et al.* (1992).

The GREG Estimator

Working super-population model:

$$y_{ik} = \underbrace{\boldsymbol{x}_{ik}^{\mathsf{T}}}_{1\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1} + \epsilon_{ik},$$

$\boldsymbol{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})$, $\text{var}_M(\epsilon_{ik}) = \sigma_k^2$, $\text{cov}_M(\epsilon_{ik}, \epsilon_{i'j}) = 0$ $j \neq k$, $i, i' = 1, \ldots, n$.

Define

$$\boldsymbol{B} = (\boldsymbol{X}_U^{\mathsf{T}}\Sigma_U^{-1}\boldsymbol{X}_U)^{-1}\boldsymbol{X}_U^{\mathsf{T}}\Sigma_U^{-1}\boldsymbol{y}_U$$

with $\boldsymbol{X}_U = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathsf{T}}$ is the $N \times p$ matrix of population covariates, $\boldsymbol{y}_U = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$, and $\boldsymbol{\Sigma}_U = \text{diag}(\sigma_{11}^2, \ldots, \sigma_{nN_n}^2)$, we let $N = \sum_i N_i$.

$\boldsymbol{B}$ is the finite population estimator of β.

The GREG Estimator

We have

$$\underbrace{\boldsymbol{X}_U^{\mathsf{T}}}_{p\times N}\underbrace{\Sigma_U^{-1}}_{N\times N}\underbrace{\boldsymbol{X}_U}_{N\times p}=\sum_{i=1}^{n}\sum_{k=1}^{N_i}\frac{\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}}{\sigma_{ik}^2},\qquad \underbrace{\boldsymbol{X}_U^{\mathsf{T}}}_{p\times N}\underbrace{\Sigma_U^{-1}}_{N\times N}\underbrace{\boldsymbol{y}_U}_{N\times 1}=\sum_{i=1}^{b}\sum_{k=1}^{N_i}\frac{\boldsymbol{x}_{ik}\boldsymbol{y}_{ik}}{\sigma_{ik}^2}$$

The sample survey estimator of $\boldsymbol{B}$ is

$$\widehat{\boldsymbol{B}}=\left(\sum_{i=1}^{n}\sum_{k\in S_i}w_k\frac{\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}}{\sigma_{ik}^2}\right)^{-1}\sum_{i=1}^{n}\sum_{k\in S_i}w_{ik}\frac{\boldsymbol{x}_{ik}\boldsymbol{y}_{ik}}{\sigma_{ik}^2}$$

This is the pseudo-likelihood estimator.

We can obtain predicted values:

$$\widehat{y}_{ik}=\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}}.$$

The GREG Estimator

We can write the true area-level mean as,

$$\begin{aligned}\overline{y}_{Ui} = \frac{1}{N_i}\sum_{k\in U} y_{ik} &= \frac{1}{N_i}\sum_{k\in U}\widehat{y}_{ik} + \left(\frac{1}{N_i}\sum_{k\in U} y_{ik} - \frac{1}{N_i}\sum_{k\in U}\widehat{y}_{ik}\right) \\ &= \frac{1}{N_i}\sum_{k\in U}\widehat{y}_{ik} + \frac{1}{N_i}\sum_{k\in U} R_{ik}\end{aligned}$$

where R_{ik} are the residuals.

We can estimate the sum of the residuals using the usual weighted estimator:

$$\sum_{k\in U}\widehat{R}_{ik} = \sum_{k\in S} w_{ik}(y_{ik} - \widehat{y}_{ik}).$$

Substituting in the working model we obtain the GREG,

$$\widehat{\overline{y}}_{Ui}^{greg} = \frac{1}{N_i}\sum_{k\in U}\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}} + \frac{1}{N_i}\sum_{k\in S} w_{ik}(y_{ik} - \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}})$$

Note: adjusts the synthetic estimator (the first term on the RHS).

The GREG Estimator

We can rewrite the GREG estimator as

$$
\begin{aligned}
\widehat{\overline{y}}_{Ui}^{greg} &= \frac{1}{N_i}\sum_{k\in U}\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}} + \frac{1}{N_i}\sum_{k\in s} w_{ik}y_{ik} - \frac{1}{N_i}\sum_{k\in S} w_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}} \\
&= \frac{1}{N_i}\sum_{k\in S} w_{ik}y_{ik} + \left(\frac{1}{N_i}\sum_{k\in U}\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}} - \frac{1}{N_i}\sum_{k\in S} w_{ik}\boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}}\right) \\
&= \widehat{\overline{y}}_{Ui}^{ht} + \left(\overline{\boldsymbol{x}}_i^{\mathsf{T}}\widehat{\boldsymbol{B}} - (\widehat{\overline{\boldsymbol{x}}}_{Ui}^{ht})^{\mathsf{T}}\widehat{\boldsymbol{B}}\right) \\
&= \widehat{\overline{y}}_{Ui}^{ht} + \left(\overline{\boldsymbol{x}}_{Ui}^{\mathsf{T}} - (\widehat{\overline{\boldsymbol{x}}}_{Ui}^{ht})^{\mathsf{T}}\right)\widehat{\boldsymbol{B}}
\end{aligned}
$$

where $\widehat{\overline{\boldsymbol{x}}}_{Ui}^{ht}$ is the weighted (HT) estimator of the vector of means of the covariates in area i – this estimator is also known as the survey regression estimator.

We see that the HT estimator is adjusted by a term that depends on the strength of the regression association.

The GREG is design consistent, since $\widehat{\overline{\boldsymbol{x}}}_{Ui}^{ht} \to \overline{\boldsymbol{x}}_{Ui}$ and $\widehat{\overline{y}}_{Ui}^{ht} \to \overline{y}_{Ui}$.

The GREG Estimator

The generalized regression (GREG) estimator of the total is

$$\begin{aligned}\widehat{t}_y^{greg} &= \widehat{t}_y^{ht} + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{ht})^\top \widehat{\boldsymbol{B}} \\ &= \sum_{k \in S} w_k g_k y_k\end{aligned}$$

where we have dropped the subscripts i, since the properties we next derive are for the total population,

$$g_k = \left[1 + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{ht})^\top \left(\sum_{j \in S} w_j \frac{\boldsymbol{x}_j \boldsymbol{x}_j^\top}{\sigma_j^2} \right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \right].$$

For any choice of the constants σ_k^2, the GREG estimator calibrates the sample to the population total of each x variable used in the regression model.

We now show this, i.e., that the estimator calibrates to the complete population total $\boldsymbol{t}_x$:

$$\widehat{\boldsymbol{t}}_x^{greg} = \boldsymbol{t}_x$$

The GREG Estimator

$$
\begin{aligned}
\widehat{\boldsymbol{t}}_x^{greg} &= \sum_{k\in S} w_k g_k \boldsymbol{x}_k \\
&= \sum_{k\in S} w_k \left[1 + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{ht})^{\mathsf{T}} \left(\sum_{j\in S} w_j \frac{\boldsymbol{x}_j \boldsymbol{x}_j^{\mathsf{T}}}{\sigma_j^2}\right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2}\right] \boldsymbol{x}_k \\
&= \widehat{\boldsymbol{t}}_x + \sum_{k\in S} w_k \underbrace{\left[(\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{ht})^{\mathsf{T}} \left(\sum_{j\in S} w_j \frac{\boldsymbol{x}_j \boldsymbol{x}_j^{\mathsf{T}}}{\sigma_j^2}\right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2}\right]}_{\text{A scalar, so equal to its transpose}} \boldsymbol{x}_k \\
&= \widehat{\boldsymbol{t}}_x^{ht} + \sum_{k\in S} w_k \frac{\boldsymbol{x}_k}{\sigma_k^2} \left[\boldsymbol{x}_k^{\mathsf{T}} \left(\sum_{j\in S} w_j \frac{\boldsymbol{x}_j \boldsymbol{x}_j^{\mathsf{T}}}{\sigma_j^{-1}}\right)^{-1} (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x)\right] \\
&= \widehat{\boldsymbol{t}}_x^{ht} + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{ht}) = \boldsymbol{t}_x
\end{aligned}
$$

Note that this recovery is at the population and not at the area level – the latter will only occur if the working model includes area-specific coefficients.

Variance of the GREG Estimator

Using linearization:

$$\mathrm{var}(\widehat{t}_{yi}^{greg}) = V\left(\widehat{t}_{yi} + (\boldsymbol{t}_{xi} - \widehat{\boldsymbol{t}}_{xi}^{ht})^{\mathsf{T}}\widehat{\boldsymbol{B}}\right) \approx V\left(\widehat{t}_{yi}^{ht} - (\widehat{\boldsymbol{t}}^{ht})_{xi}^{\mathsf{T}}\boldsymbol{B}\right)$$

Let $e_{ik} = y_{ik} - \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{B}}$ be the k-th residual in the i-th area; then the variance can be estimated by

$$\widehat{\mathrm{var}}_1(\widehat{t}_{yi}^{greg}) = \widehat{\mathrm{var}}\left(\sum_{k\in s} w_{ik}\, e_{ik}\right).$$

Alternative estimator:

$$\widehat{\mathrm{var}}_2(\widehat{t}_{yi}^{greg}) = \widehat{\mathrm{var}}\left(\sum_{k\in s} w_{ik}\, g_{ik}\, e_{ik}\right).$$

Variance of the GREG Estimator

Now suppose we have SRS.

For the HT estimator:

$$\widehat{\text{var}}(\hat{t}_{yi}^{ht}) = \frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\frac{\sum_{k\in S_i}(y_{ik} - \bar{y}_i)^2}{n_i - 1}$$

For the GREG estimator:

$$\widehat{\text{var}}(\hat{t}_{yi}^{greg}) = \frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\frac{\sum_{k\in S_i} e_{ik}^2}{n_i - 1}.$$

So the variance of the estimator is smaller if the regression model gives smaller residuals than in the original observations (relative to their mean).

Generalized Regression (GREG) Estimator

Notes:

- If the synthetic estimator is based on a model with area specific regression parameters, it is equivalent to the GREG estimator.
- For a linear GREG estimator only the area totals or means for the $\boldsymbol{x}$'s are needed (not unit values for all of the population) – this is very desirable.
- Logistic GREG (LGREG) is a version that has a logistic working model (Kennel and Valliant, 2010).

Generalized Regression (GREG) Estimator

- In general, we have

$$\widehat{\overline{y}}_{Ui}^{greg} = \frac{1}{N_i}\sum_{k\in U}\widehat{\mu}_{ik}(\boldsymbol{x}_{ik}) + \frac{1}{N_i}\sum_{k\in S} w_{ik}(y_{ik} - \widehat{\mu}_{ik}(\boldsymbol{x}_{ik}))$$

  for a general mean prediction model $\mu_{ik}(\boldsymbol{x}_{ik})$.

- See Dagdoug *et al.* (2022) for a random forest prediction model.

- Gao and Wakefield (2022) exploit this in an SAE context, by using the LGREG estimator as the observed data (as in Fay-Herriot) and adding random effects.

Composite Estimator

In Rao and Molina (2015, Section 3.3) describe a composite estimator of the form

$$\widehat{\overline{y}}_i^{com} = d_i \widehat{\overline{y}}_{Ui}^{dir} + (1 - d_i) \widehat{\overline{y}}_{Ui}^{syn},$$

with weight $0 \leq d_i \leq 1$ – the idea is to trade off the unbiased but potential high variance of a direct estimator $\widehat{\overline{y}}_{Ui}^{dir}$ and the opposite properties of the synthetic estimator $\widehat{\overline{y}}_{Ui}^{syn}$.

The optimal weight d_i in a design-based framework is a function of the MSEs of $\widehat{\overline{y}}_{Ui}^{dir}$ and $\widehat{\overline{y}}_{Ui}^{syn}$.

Rao and Molina (2015, p. 59) report that the estimated weights can be highly unstable.

Different composite estimators are available, but we prefer to pursue a random effects representation, to have a formal model for choosing the weighting parameter.

Discussion

Discussion

- SAE is often based on samples collected under a complex design, and if this case one must account for the design in the analysis.
- Direct (weighted) estimates are the starting point for analysis, and will be suitable, if the sample size is sufficiently large.
- Variance estimation that accounts for the design has been a topic of much research.
- However, for the major designs (e.g., SRS, stratified SRS, cluster sampling, multistage sampling), weighted estimates and their variances are available within all the major statistical packages.

Discussion

- When the variance is large, because of small sample sizes, we would like to use smoothing methods, with Bayes being a convenient way to do this – this is the topic of the next lecture.
- We will also consider how covariate information can be used.
- The majority of survey sampling texts take a design-based view of inference – this is a different paradigm to model-based inference, for which most spatial statistical models were developed!
- Later we will see how spatial methods can incorporate the survey design.

References

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, pages 1–18. To appear.

Diggle, P. J. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman and Hall/CRC.

Gao, P. A. and Wakefield, J. (2022). Smoothed model-assisted small area estimation. *arXiv preprint arXiv:2201.08775*.

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.

Kennel, T. L. and Valliant, R. (2010). Logistic generalized regression (lgreg) estimator in cluster samples. In *Proceedings of the Section on Survey Research Methods*, pages 4756–4770.

Kenya National Bureau of Statistics (2015). Kenya Demographic and Health Survey 2014. Technical report, Kenya National Bureau of Statistics.

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

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

Rao, J. (2003). *Small Area Estimation*. John Wiley, New York.

Rao, J. and Molina, I. (2015). *Small Area Estimation, Second Edition*. John Wiley, New York.

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**.