Two-stage cluster survey sampling and analysis: a complex survey in Zambézia Province, Mozambique.

Meridith Blevins, MS

Vanderbilt University
Department of Biostatistics
Outline

1. Motivation
2. Sample Design
   - Sample Size
   - Sample Selection
3. Design-based Analysis (*light*)
Sampling of Populations: Methods and Applications

Sampling Techniques

Complex Surveys: A Guide to Analysis Using R
Motivation
Motivating Example #1

To give an example, let’s say that you want to interview health workers across the US to learn about their job satisfaction and their intention to stay in their jobs.

You could get a list of all health workers in the country and then take a random sample of them.

Can you think of some challenges with this approach?
Some challenges:

- sample would represent all health workers, but may consider stratification by gender, position, geography, or other factors
- tracking down individual health workers by name and last known location of work may exceed the amount of resources available for your study
- if the survey is administered in person, travel costs could be exorbitant
- a good list may not exist!
Motivating Example #2

A baseline survey will estimate population characteristics in Zambézia Province (≈ size of TN, nearly 1 million households, >85% rural).
Motivating Example #2, cont’d

$20 million worth of USAID-sponsored interventions meant to benefit the nutrition, health, education, sanitation, and social agency of women and children will take place during the next five years.

We want to interview households across the province to learn about various indicators that should measure the impact our initiatives.
Suppose we could get a census listing of all households from 2007 and plan to take a random sample of households.

Can you think of some challenges with this approach?
Some challenges:

- Sample would represent all households, but may consider stratification by planned intervention, urban/rural, district, or access to health care.

- Consider the setting, very poor transportation infrastructure and extremely rural areas which could require a day’s travel for one survey.

- Census data is not current (2007); new structures have been built and old homes demolished.

- Furthermore, we do not actually have the census data!
Motivation

What makes a survey complex?
Use of the following elements in the sample design:
- Stratification
- Clustering
- Sampling fractions
- Sampling weights
- Poststratification

Why a complex survey?
Allows for some control over distribution, precision, and cost:
- Ensure a representative sample.
- Limit the geographic coverage.
- Use of sampling frame.
Sample Design
The sampling frame for the baseline survey relied on the 2007 Mozambique Population and Housing Census. A list of *Enumeration Areas* (EA) covering the province were made available (*by proxy*) with basic housing and population information and cartographic materials.

- The *population* for Zambézia is estimated at 3,794,489.
  - This is 918,025 households.
  - Divided into 9,073 EA.
  - There are 155,202 urban households (1,458 EA).
  - There are 762,823 rural households (7,615 EA).
Two-Stage Stratified Cluster Sample

The enumeration areas are stratified according to type of residence (urban-rural). Implicit stratification may be used as well (more on this later).

**Stage 1**
A number of primary sampling units, or clusters, are selected from a sampling frame independently in each stratum. In Zambézia, these are EA.

**Stage 2**
A fixed number of secondary sampling units are selected with equal probability. In Zambézia, these are households.
This two-stage cluster sampling procedure has several advantages:

- provides good coverage,
- simple to implement, and
- allows for control of field-work quality.

In order to achieve both economy and good precision, sample sizes at both stages of the survey must be determined in such a way that they minimize the sampling error under a given sampling cost.
Determining optimal sample size is critical because it requires a trade-off between the budget and the desired survey precision.

We will determine

- cost ratio
- intracluster correlation (ICC)
- cluster size
- desired precision
- number of clusters to be selected
Cost Ratio, $\frac{C_1}{C_2}$

- Cost of interviewing a cluster, $C_1$
  - Household listing cost
  - Travel between clusters (village-to-village)

- Cost of interviewing an individual, $C_2$
  - Travel within cluster (house-to-house)

The cost ratio varies depending on population density and infrastructure.
Intracluster Correlation

ICC measures the similarity of the individuals on one survey characteristic within a cluster, $\delta$.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Total</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical care</td>
<td>0.13</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>Knowledge of contraception</td>
<td>0.11</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>Background or lifetime variables</td>
<td>0.08</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Current use of contraception</td>
<td>0.03</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>Child health</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Fertility</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Current fertility intentions</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Infant mortality</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Total average</td>
<td>0.055</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Note: The intracluster correlation coefficient represents the ratio of between-cluster variability to total variability, $\sigma_b^2/(\sigma_b^2 + \sigma_w^2)$. 
Cluster Size

The number of individuals interviewed is a function of the cost ratio and intracluster correlation (ICC).

The optimal number of households to sample:

\[
\bar{n} = \left[ \frac{C_1}{C_2} \cdot \frac{1 - \delta}{\delta} \right]^{1/2}
\]
The **design effect** quantifies the increase in the standard error of the estimate due to the sampling procedure used.

\[ D = 1 + \delta(\bar{n} - 1). \]

Using optimal cluster size \( \bar{n} \), the number of clusters \( \bar{m} \) to be selected within strata may be determined:

\[ \bar{m} = \frac{p(1 - p)D}{s^2 \bar{n}} \]

- For a confidence interval of \( \pm 5\% \) we shall need \( s = 0.025 \).
- If we have some idea of the proportion \( p \) in advance, then it may be used.
The population for Zambézia is estimated at 3,794,489.

- This is 918,025 households.
- Divided into 9,073 EA.
- There are 155,202 urban households (1,458 EA).
- There are 762,823 rural households (7,615 EA).
<table>
<thead>
<tr>
<th>$C_1/C_2$</th>
<th>ICC</th>
<th>Precision (%)</th>
<th>$\bar{n}$</th>
<th>$\bar{m}$</th>
<th>Per Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.01</td>
<td>5</td>
<td>45</td>
<td>13</td>
<td>585</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>10</td>
<td>45</td>
<td>4</td>
<td>180</td>
</tr>
<tr>
<td>20</td>
<td>0.02</td>
<td>5</td>
<td>32</td>
<td>21</td>
<td>672</td>
</tr>
<tr>
<td>20</td>
<td>0.02</td>
<td>10</td>
<td>32</td>
<td>6</td>
<td>192</td>
</tr>
<tr>
<td>20</td>
<td>0.05</td>
<td>5</td>
<td>20</td>
<td>39</td>
<td>780</td>
</tr>
<tr>
<td>20</td>
<td>0.05</td>
<td>10</td>
<td>20</td>
<td>10</td>
<td>200</td>
</tr>
<tr>
<td>20</td>
<td>0.1</td>
<td>5</td>
<td>14</td>
<td>66</td>
<td>924</td>
</tr>
<tr>
<td>20</td>
<td>0.1</td>
<td>10</td>
<td>14</td>
<td>17</td>
<td>238</td>
</tr>
<tr>
<td>20</td>
<td>0.15</td>
<td>5</td>
<td>11</td>
<td>91</td>
<td>1001</td>
</tr>
<tr>
<td>20</td>
<td>0.15</td>
<td>10</td>
<td>11</td>
<td>23</td>
<td>253</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>5</td>
<td>9</td>
<td>116</td>
<td>1044</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>10</td>
<td>9</td>
<td>29</td>
<td>261</td>
</tr>
</tbody>
</table>
Estimating change from Baseline to 5 years

- Stratify on regions where particular interventions were performed (would be ideal to randomize region to intervention)
If both baseline and 5 year surveys are designed for given precision:

<table>
<thead>
<tr>
<th>Precision at Baseline and 5 Year</th>
<th>Detectable difference independent samples</th>
<th>Detectable difference same clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>0.035</td>
<td>0.030</td>
</tr>
<tr>
<td>4.5%</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>5%</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>7.5%</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>10%</td>
<td>0.14</td>
<td>0.12</td>
</tr>
</tbody>
</table>

NOTE: UN Guidelines suggest that when the same clusters are used but different households surveyed, a variance estimate on baseline to 5 year change is 40 percent \textit{higher} than the baseline level. Independent samples produce an estimated variance that is \textit{double} that of baseline.
What is the detectable difference for intervention effect assuming half of the clusters get intervention? If both baseline and 5 year surveys are designed for given precision:

<table>
<thead>
<tr>
<th>Precision at Baseline and 5 Year</th>
<th>Detectable difference independent samples</th>
<th>Detectable difference same clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>4.5%</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>5%</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>7.5%</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>10%</td>
<td>0.28</td>
<td>0.24</td>
</tr>
</tbody>
</table>
## Actual Sample Size

<table>
<thead>
<tr>
<th>ICC</th>
<th>Precision (%)</th>
<th>$\bar{n}$</th>
<th>$\bar{m}$</th>
<th>Per Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>7.5</td>
<td>15</td>
<td>68</td>
<td>1020</td>
</tr>
<tr>
<td>0.15</td>
<td>4.5</td>
<td>15</td>
<td>103</td>
<td>1545</td>
</tr>
</tbody>
</table>

![Map 1](image1.png)

- Households per EA
  - under 50
  - 50 - 100
  - 100 - 150
  - over 150

![Map 2](image2.png)

- Households per EA
  - under 50
  - 50 - 100
  - 100 - 150
  - over 150

Blevins (Vandy Biostats)  Zambézia Survey  26 / 39
Sample Selection (by Proxy)

Carlos Singano Creva, PhD
Probability Proportional to Size (PPS)

Because EA instead of individuals are now our unit of analysis, it does not make statistical sense to give two districts with very different populations an equal probability of being chosen for the sample. The end goal is to have households (the final sampling unit) to be representative of the population.

This means that an EA with a small population, should not have an equal chance of being selected as an EA, with a very large population. If the probability was equal, your final sample would not be generalizeable to the population.
Divide the total cumulative population (6,628,637) by PSUs (20).
This gives a sampling interval (331,432).
Select a random start between 1 and sampling interval (126,599).
These counties have implicit stratification by median household income.
Selecting Households (Plan A)
Selecting Households (Plan B)
Design-Based Analysis
Purpose of Design-based Inference

The researcher specifies a population, whose data values are unknown but are regarded as fixed, not random.

The observed sample is random because it depends on the random selection of individuals from this fixed population.

The random selection procedure of individuals (the sample design) is under the control of the researcher, so all the probabilities involved can, in principle, be known precisely.

The goal of the analysis is to estimate features of the fixed population, and design-based inference does not support generalizing the findings to other populations.

— T. Lumley, 2010
Horvitz-Thompson Estimator

Proposed in 1952 as an estimator of a total that can be constructed for any sampling design. For a single-stage cluster sampling design, its form is given by

\[ y'_{hte} = \sum_{i=1}^{m} \frac{Y_i}{\pi_i} \]

where

- \( Y_i = \) the total for the \( i \)th sample cluster.
- \( \pi_i = \) the probability of the \( i \)th cluster being selected in the sample.
- \( m = \) the number of clusters sampled (with replacement).

\( y'_{hte} \) is an unbiased estimator of \( Y \) with standard error given by

\[ SE(y'_{hte}) = \sqrt{\sum_{i=1}^{m} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{m} \sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j} \]
Survey Package for R

Maintained by Thomas Lumley.

First specify sample design, then request appropriate summaries...

DEMONSTRATION
Additional Considerations

- Data Collection Setbacks
- Nonresponse
- Imputation
- Graphics/Plots
- Regression
- Replicate Results (SUDAAN/SAS/STATA)


**References**

- **Sampling of Populations: Methods and Applications**  

- **Sampling Techniques**  

- **Complex Surveys: A Guide to Analysis Using R**  
Obrigada! (Thanks!)

Questões? (Questions?)