An Introduction to Complex Survey Sampling

Trent D. Buskirk, Ph.D.
Vice President Statistics and Methodology
Marketing Systems Group
May 7, 2014
Warm-Up

- Three men are in a hot-air balloon. Soon, they find themselves lost in a canyon somewhere. One of the three men says, "I've got an idea. We can call for help in this canyon and the echo will carry our voices far."

- So he leans over the basket and yells out, "Hellllloooooo! Where are we?" (They hear the echo several times.)

- Fifteen minutes pass. Then they hear this echoing voice: "Hellllloooooo! You're lost!!"

- One of the men says, "That must have been a statistician."

- Puzzled, one of the other men asks, "Why do you say that?"

- The reply: "For three reasons. (1) he took a long time to answer, (2) he was absolutely correct, and (3) his answer was absolutely useless."
So What is a Survey Sample?

• **Sample Surveys**
  – A sample survey is a study that obtains data from a subset of a population, in order to estimate population attributes.
  

• **Probability Samples (Sample Surveys)**
  – A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.
  

• **Non-Probability Samples (Convenience/Opt-In Surveys)**
  – A type of unit sampling where it is not known which of the units will be picked to be sampled, and where some of the units have a zero probability of being chosen.
  
  Source: [http://www.businessdictionary.com/definition/non-probability-sample.html#ixzz2jVT5f57f](http://www.businessdictionary.com/definition/non-probability-sample.html#ixzz2jVT5f57f)
Types of Samples - Not All Sample Surveys are the Same!

- **Probability-Based Samples**
  - Address-based samples (National Health Interview Survey)
  - Random Digit Dial (RDD) Dual Frame Samples (Behavioral Risk Factor Surveillance Survey)
  - Generally used to form basis of “official statistics” because they yield projectable estimates along with standard errors

- **Non-probability Based Samples**
  - Quota Samples (Mall Intercept Studies)
  - Convenience Samples (Opt-in Internet Panels)
  - More difficult to use these samples for projectability in that no estimates of precision can typically be given
So what makes a Sample Survey Complex?

• If selection of final units of observation is accomplished through a series of stages in which larger units are first selected then the sample survey is said to be complex (or multistage).

• Generally, complex sample surveys employ simple random sampling at each stage in a series of stages to culminate in the final sample of observation units at the desired level.
  
  — An estimate of the number of attempts required to visit the Healthcare.gov Website before successful completion of an application is desired.

  — Because internet volume varies regionally, selections are to be made from Western and Eastern states separately.

  — Moreover, because multiple users with varying levels of health information might live within a given household, we will seek to randomly query two users per household.

In this case households are first organized into Western and Eastern states and users within households. Once a household is randomly selected, all eligible household members are enumerate and then two are randomly selected.
Steps in Conducting/Analyzing Complex Survey Samples

1. Define Problem/State Objectives
2. Understand the Sample Design
3. Select the sample (or obtain the data)
4. Process and Analyze Data
5. Interpret and evaluate the results of the analysis
6. Report results from survey data
Step 1. Define Problem/State Objectives

- Specify the problem to be addressed
- Objectives support/seek to answer problems stated
- Problems can be highly variable, even within one survey –
  - for example, in the NHANES survey, you could estimate the proportion of people who are obese, nationally and then estimate the serum cortisol levels among menopausal women.
Step 2: Understand the Sample Design

• Understand how the survey was conducted, the intended population of interest and the levels at which data are available...

• If a problem seeks to understand trends over time and the survey is cross-sectional, then the utility of the survey data in meeting estimation objectives will be limited.

• If county-level estimates of health literacy are of interest from the state-wide health literacy survey that did not plan for data coverage in every county, then some estimates will be unavailable, others highly unstable and others useful.
Step 2: Understand the Sample Design, cont.

• When designing a survey, problems and objectives are often balanced by the various levels of estimates that are required/desired.

• Often, one survey is not enough to answer all key objectives for a particular problem. Knowing how the sample was designed/redesigned is also an important component of combining sources...

  ➔ For example- NHANES was redesigned in 1999 from periodic administration (every 4-6 years) to continuous/yearly administration. Now point estimates should be based on two-year data rather than on single periodic release data.
Sampling Elements Defined…
A Closer Look.

• **Observation Unit**: the object of measurement in a survey. The person, machine, office building, housing unit or account or otherwise for which some variable of interest is measured by using either a precise scientific or psychological instrument.

• **Target Population**: Collection of observation units about which we want questions answered. Defining this target population is a crucial step in executing a survey - it is the population we want to make inference to.

• **Sample**: A subset of OBSERVATION UNITS from the target population.
More concepts, defined...

- **Sampled Population**: All possible observation units that could have theoretically been included in the sample. This collection of units represents the population from which the sample was taken. Inference drawn from samples can be made to this population.

- **Sampling Unit**: Actual unit that we include in our sample. The primary sampling units are those selected within the first stage of a design (PSU); the secondary sampling units are those selected in the second stage (SSUs) etc.

- **Sampling Frame**: A listing of all possible sampling units. In many cases this frame coincides with the sampled population.
Target versus Sampling Populations

- Target Population
- Sampling Population: (Sampling Frame)
- Overlap Between the Two Populations
- Sample
Sampling Frame Details...

- The sampling frame will generally coincide with the sampled population, but sometimes the frame will be larger or smaller

  - Think about exclusion criteria (frame is all phone numbers, but some businesses will be excluded)

  - Multiple frames may be needed to cover the population (i.e. dual frame phone surveys that have both a landline and cell phone frame).
Sampling Frame Example

- An RDD of cell phone numbers selected from a master frame of all available cell phone banks is the Sampled Population for a study desiring feedback from the U.S. Adult population about opinions on current politicians.

  - The target population will in large part be “covered” by the sampled population

  - however the sampled population contains many numbers that are
    - not working
    - are associated with children.

  - Adults living with only a landline telephone or no telephone at all will be under-covered by this sampling frame. If these adults are different in some way to those who own a cell phone on our survey outcomes, then selection bias may result – in particular, coverage bias.
Step 3: Select a Probability Sample

- There are many types of random sampling designs that can be used at various stages of a complex, multistage sampling design.
  - Simple Random Sampling (with/without replacement)
  - Systematic Sampling
  - Stratified Sampling
  - Cluster Sampling
Sampling Designs: Simple Random Sampling

- SIMPLE RANDOM SAMPLING designs are those in which every subset of a fixed size of the population has the same probability of being the selected sample. Most typical in Human Subjects Research involves WITHOUT REPLACEMENT Designs...

- To implement this type of sampling design in practice, an enumerated list containing [all/most] members of the target population should be available. Every person on such a list will have an equal chance of being selected for the sample!

- Analysis of resulting data requires no additional statistical adjustment – if design is without replacement, estimates of standard errors will be conservative using standard software analyses (i.e. larger than actual)
First, we define the frame.

Then we select our sample.

Under SRS, every subset of size 2 from our population has the same likelihood of being selected as the sample.
Sampling Designs Continued: Stratification

- **STRATIFIED RANDOM SAMPLING** designs identify subpopulations or STRATA within the population based on auxiliary information that is known from every member.
  - Strata are exhaustive and non-overlapping

- Independent random samples are then selected from each of the identified STRATA.
  - Selection approach can differ across strata

- Precision of estimates is expected to be greater when members are more homogeneous within strata as compared to between strata.

- Statistical Analyses are straightforward within strata, but cumulative estimates require additional weighting.
Before Sampling, we could stratify our population by age.

One clinic could be for kids... While the other could be for parents...
Why Stratify?

- Stratification can be used to ensure that the sample taken will be “representative” of the population based on certain known demographic facts.
  - If information is available on an auxiliary variable *apriori* sampling, then it would be possible to **stratify** the population and to then take a stratified sample.

- For example: You wanted to know the mean post-lunch weight gain of students attending a heavily federally subsidized school in the Bronx. It is estimated that 90% of the students receive free school lunch while the remaining 10% bring their own food for lunch.

- A simple random sample of 30 of the 300 students enrolled may produce a sample containing either a single student or no student who brings their own lunch. The probability of obtaining such a sample would be about 17%. These samples might produce estimates that may appear too large (although unbiased over repeated sampling...)
  - May be better to stratify first by lunch cost- free or fee.
Why Stratify? Continued

• We can obtain individual estimates to a specified level of precision for demographic subgroups of the population. The demographic variables can be used as the stratification variables.
  – Need to cover certain parts of the population with certainty

• Stratified sampling may lower cost and increase data quality by allowing for different sampling schemes in each stratum.

• Prior information and cost information can be used to form strata and to decide upon respective sampling schemes within strata and/or to allocate overall sample sizes across strata.
Cluster Sampling (in one or two stages)

• Cluster sampling involves selection of collections of final sampling units that have been grouped into population subgroups or clusters
  – each population member can belong to only one cluster

• Sample selection can occur in one or two stages (or more)
  – One stage cluster sampling selects all units within each selected cluster

  – Two-stage cluster sampling first selects clusters then selects a subset of population units in each chosen cluster
Or we could cluster people by living quarter...

Information on individual members of a population can be made available by first selecting clusters and then selecting all or some of the members within those clusters.
One-Stage Cluster Sampling

• Step One: Select Primary Sampling Units (i.e. Clusters)

Step Two: Select ALL Secondary Sampling Units within selected PSU’s (i.e. Clusters)
Two-Stage Cluster Sampling

• **Stage One**: Select Primary Sampling Units
Two-Stage Cluster Sampling

• **Stage Two: Select Secondary Sampling Units**
# Two-Stage Cluster Sampling Examples

<table>
<thead>
<tr>
<th>First Stage</th>
<th>Second Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSU</strong></td>
<td><strong>SSU</strong></td>
</tr>
<tr>
<td>City Block</td>
<td>Household</td>
</tr>
<tr>
<td>Household</td>
<td>Family Member</td>
</tr>
<tr>
<td>Clinic</td>
<td>Patient</td>
</tr>
<tr>
<td>Classroom</td>
<td>Student</td>
</tr>
<tr>
<td>Geographic Area</td>
<td>Small Plot</td>
</tr>
<tr>
<td>Stream</td>
<td>3-meter Section</td>
</tr>
<tr>
<td>Hospital</td>
<td>Department/Record</td>
</tr>
<tr>
<td>Community Center</td>
<td>Member</td>
</tr>
</tbody>
</table>
Why Cluster Sampling, Cont.

• Cluster sampling is appropriate in these situations:
  – when a sampling frame of the final elements of interest is unavailable
  – when data collection is expensive because the elements of interest are widely dispersed
  – when units are naturally nested together in larger, more accessible units (i.e. students within schools, patients within hospitals)

• This design is more preferred to the other designs whenever members in a cluster are highly dissimilar to one another. Generally, cluster sampling is less expensive but statistical estimates are more variable.
Step 4: Analysis Issues – Sources of Error in Surveys

There are two types of errors associated with sample surveys: Sampling and Nonsampling Error.

Total Survey error accounts for both of these in planning for the allocation of resources. Sampling error is often easier to quantify and estimate, but may be more negligible when compared to nonsampling error.

Types of Nonsampling Error:  
- Selection Bias  
- Measurement Error  
- Survey Nonresponse

Sampling Error:  
Error that is associated with taking a sample instead of examining the entire population.
SAMPLING ERRORS

• Sampling errors are generally associated with taking a single sample as opposed to interviewing the entire population (i.e. taking a CENSUS).

  – We would expect that estimates based on a particular sample (no matter how well chosen) would differ slightly from those obtained using a different sample. Such variation is usually called sampling error.

  – In the media, the term “margin of error” typically refers to this type of variability in the estimate.

  – Impact of multistage design: complex samples are not likely to not be as efficient as a simple random sample.

  • The impact of the overall design on the precision of an estimate can be summarized by the “design effect.”
(Possible) Effect of Stratification for the Primary Survey Item

Accounting for Stratification

Ignoring Stratification

\( \hat{\theta} \)
Effect of Cluster Sampling Designs on Estimation

Ignoring Cluster Sampling

Accounting for Cluster Sampling
Software Packages for the Analysis of Survey Data

- **SAS- Proc Survey Select**: Helps you select a random, stratified or cluster sample (creates appropriate first stage weights).

- **SAS Proc Survey Means/Survey Reg**: Descriptive statistics and regression models derived from weighted survey data.

- **SUDAAN/WestVar PC**: Can derive descriptive and inferential models including regression and logistic regression using a weighted survey data file.

- **SPSS Version 21/22 Complex Samples Component**: Offers a sampling wizard for the design and analysis of basic sampling designs.

- **R**: Has several packages dedicated to the analysis of survey data – most notably the **Survey Package** from T. Lumley.
  - [http://faculty.washington.edu/tlumley/survey/](http://faculty.washington.edu/tlumley/survey/)
Resources for Survey Data Analysis

• American Association of Public Opinion Research has resources for response rate calculations and periodically offers webinars on topics like questionnaire design, sample design and weighting.
  – www.aapor.org

• American Statistical Association’s Survey Research Methods Section

• Institute for Digital Research and Education at UCLA
  – http://www.ats.ucla.edu/stat/seminars/svy_intro/
THANK YOU – THE END!

tdbuskirk@gmail.com

OR

TBuskirk@m-s-g.com

314-695-1378

www.m-s-g.com
APPENDIX
## Cluster Sampling versus Stratified Sampling

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Cluster Sampling</th>
<th>Stratified Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost reduction</td>
<td>Subpopulation estimates</td>
</tr>
<tr>
<td></td>
<td>Sampling frame units are clusters of elements</td>
<td>Variance reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control costs across subpopulations</td>
</tr>
<tr>
<td>Selection</td>
<td>Elements selected only from selected clusters</td>
<td>Elements selected from every stratum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>Precision primarily depends on variance <strong>between</strong> clusters</td>
<td>Precision primarily depends on variance <strong>within</strong> strata</td>
</tr>
</tbody>
</table>
Step 2: Example – Putting it all together

• Public health officials are interested in estimating the mental illness rate among homeless people.

• Wright (1988) estimates that 33% of all homeless people are mentally ill, by sampling homeless persons who received medical attention from one of the clinics in the Health Care for the Homeless (HCH) Project.

• He argues that selection bias is not a serious problem because the clinics were easily accessible to the homeless and because the demographic profiles of HCH clients were close to those of the general homeless population in each city in the sample.

a. What is the Target Population?
b. Sampling Frame?
c. Sampling Unit?
d. Observation Unit?
e. Your thoughts on Wright’s Claim?

Step 2 Example - Continued.

a. What is the Target Population?
   All homeless persons in the study area (nation/region?)

b. Sampling Frame?
   Clinics participating in the Health Care for Homeless Project

c. Sampling Unit?
   Unclear – but this could have been the clinic or a homeless person within a clinic (or both- multiple stage)

d. Observation Unit?
   Person (homeless)

e. Your thoughts on Wright’s Claim?
   Information about mental illness was gathered from those homeless who were seeking medical attention in these clinics (for physical or mental health reasons). These homeless folks sought needed medical attention, so they may have more/less mental illness than homeless people who are not patients. Differences in illness might extend beyond simple demographics...