Introduction to Complex, Applied Survey Data Analysis Using Stata

Chelsea Goforth
Statistical Consulting Associate

StatLab: Statistics and Data Analysis
University of Virginia Library

Fall 2015

Survey Data Analysis Steps

1. Definition of the problem and statement of the objectives.
   - e.g., to describe characteristics of a target population
   - e.g., to explore and extend the understanding of multivariate relationships among variables in the target population
2. Understanding the sample design.
   - without understanding key properties of the survey sample design, e.g., clustering, stratification, and weighting, the analysis may be inefficient, biased, or otherwise lead to incorrect inferences
3. Understanding the design variables, underlying constructs, and missing data.
   - What are the empirical distributions of these design variables, and do they conform to the design characteristics outlined in the study documentation? Does the original survey question that generated a variable capture the underlying construct of interest? What is the distribution of missing data?

Survey Data Analysis Steps

4. Analyzing the data.
   - specific analytic techniques must be carefully chosen to conform to the analysis objectives adn properties of the survey data
5. Interpreting and evaluating the results of the analysis.
   - interpretation of the results from an analysis of survey data requires a consideration of the error properties of the data and is often considered in relation to simple random sampling (SRS)
6. Reporting of estimates and inferences from the survey data.

Clustering

- multi-stage selection of sample units
- a frequently used design feature, sometimes at one or more stages of the sample selection, as non-clustered designs are often impractical/inefficient for both logistical and financial reasons
  - e.g., when travel costs and related expenditures are high
  - e.g., when sample elements may not be individually identified on the available sampling frames but can be linked to aggregate cluster units (voters at precinct polling stations)
- can provide more precise population estimates, but are often subject to larger standard errors because single-stage or multi-stage clustered sampling causes correlations of observations within sample clusters
  - non-independence: when such group similarity is present, the amount of "statistical information" contained in a clustered sample of n persons is less than in an independently selected SRS of the same size

Stratification

- divides the sample up into separate sub-groups (often by demographic variables), selects random samples from within each group, then combines sub-samples to form the complete sample
- strata are nonoverlapping, homogeneous groupsings of population elements or clusters of elements that are formed by the sample designer prior to the selection of the probability sample
- in multi-stage designs, a different stratification of units can be employed at each separate stage of the sample selection
- strata can be proportionate or disproportionate to population totals
- goal of stratified designed to increase sample precision is to form strata that are "homogeneous within" and "heterogeneous between"
    - i.e., units assigned to a stratum are like one another and different from those in other strata in terms of key survey variables
    - often results in decreased standard errors

(Sampling) Weights

- generally applied to correct for unequal selection probabilities and nonresponse in order to "map" the sample back to an unbiased representation of the survey population
  - more specifically, the final (i.e., probability) weights are the product of the sample selection weight, a nonresponse adjustment factor, and the poststratification factor
- main purpose is to reduce bias in population estimates by up-weighting population sub-groups that are under-represented and down-weighting those that are over-represented in the sample
- the sample selection weight is equal to the inverse of the probability of being included in the sample due to the sampling design; under many sampling plans, the sum of the probability weights will equal the population total
- like clustering, weighting can often result in larger standard errors, especially when the variance of the weights is large

Identifying Relevant Design Features

Key 1 Is the sample selected in a single stage or multiple stages?

Key 2 Is clustering of elements used at one or more sample stages?

Key 3 Is stratification employed at one or more sample stages?

Key 4 Are elements selected with equal probabilities?

The combination of all possible answers to these questions implies that there are at least 16 possible choices of complex sample designs; however, one complex design—multistage, stratified, cluster sampling with unequal probabilities of selection for elements—is used in most in-person surveys of household populations.

Simple Example Data Set

ID	Stratum	Cluster	EconRating	Weight
1	1	1	52.8	1
2	1	1	32.5	2
3	1	2	37.3	1
4	1	2	57.0	1
5	2	3	27.7	1
6	2	3	42.3	2
7	2	4	48.8	1
8	2	4	66.8	1

So What?

Why do concepts like clustering, stratification, and weighting matter?

- Survey analysts routinely ignore complex design factors such as clustering, stratification, and weighting and proceed as if data were collected via simple random sampling.
  - → This results in biased estimates of standard errors, which in turn has implications for the inferences we're able to draw.

Instead, we ought to consider the efficiency of complex survey designs relative to SRS.

1. Design Effect Ratio
2. Effective Sample Size

Design Effect Ratio

Relative to an SRS of equal size, the complex effects of stratificaiton, clustering, and weighting on the standard errors of the estimates are together called the **design effect** $[D^2(\hat{\theta})]$ and are measured via a ratio.

Generally speaking:

$$D^2(\hat{\theta}) \approx 1 + f(G_{strat}, L_{cluster}, L_{weighting})$$

where G_{strat} refers to the relative gain in precision from stratified sampling compared to SRS; $L_{cluster}$ refers to the relative loss of precision due to clustered slection of sample elements; and, $L_{weighting}$ refers to the relative loss of precision due to unequal weighting for sample elements.

Design Effect Ratio

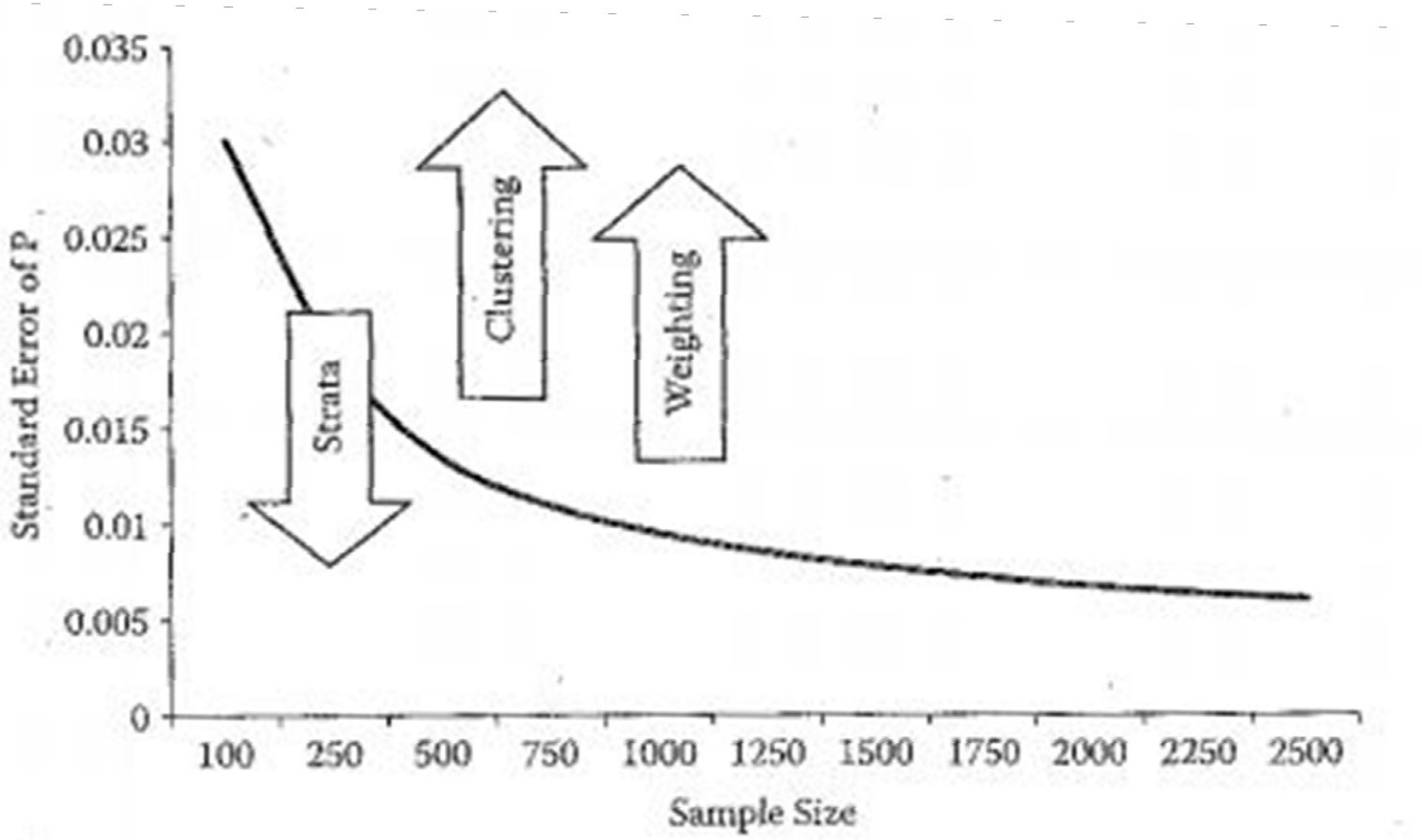
More specifically, we look at an *estimated* design effect due to the complex nature of the ways in which all of these designs features interact:

$$d^2(\hat{\theta}) = \frac{se(\hat{\theta})^2_{complex}}{se(\hat{\theta})^2_{srs}} = \frac{var(\hat{\theta})^2_{complex}}{var(\hat{\theta})^2_{srs}}$$

When this ratio is greater than 1, the design is less efficient than an SRS of equal size; when it is less than 1, it is more efficient.

In practice, this is often more useful to the survey designer than the survey analyst because the applied analysis in Stata will frequently bypass this step; however, knowledge of estimated design effects, including the component factors, helps you gauge the extent to which the sampling plan for your data has produced efficiently losses relative to an SRS design.

Complex Sample Design Effects on Standard Errors



Effective Sample Size

A related measure, the **effective sample size**, reflects the number of SRS cases required to achieve the same sample precision as the actual copmlex sample design:

$$n_{eff} = \frac{n_{complex}}{d^2(\hat{\theta})}$$

where n_{eff} refers to the effective sample size and $n_{complex}$ refers to the actual or "nominal" sample size selected under the complex sample design.

Combined, these two measures are two means of expressing the precision of a complex sample design relative to an SRS of equal size.

- e.g., for a fixed sample size, the statements "the design effect for the complex sample is 1.5" and "the complex sample of size $n = 1,000$ has an effective sample size of $n_{eff} = 667$ are equivalent statements of the expected precision loss due to the complex sample design

Finite Population Correction

A final important concept to understand is the **finite population correction** (FPC):

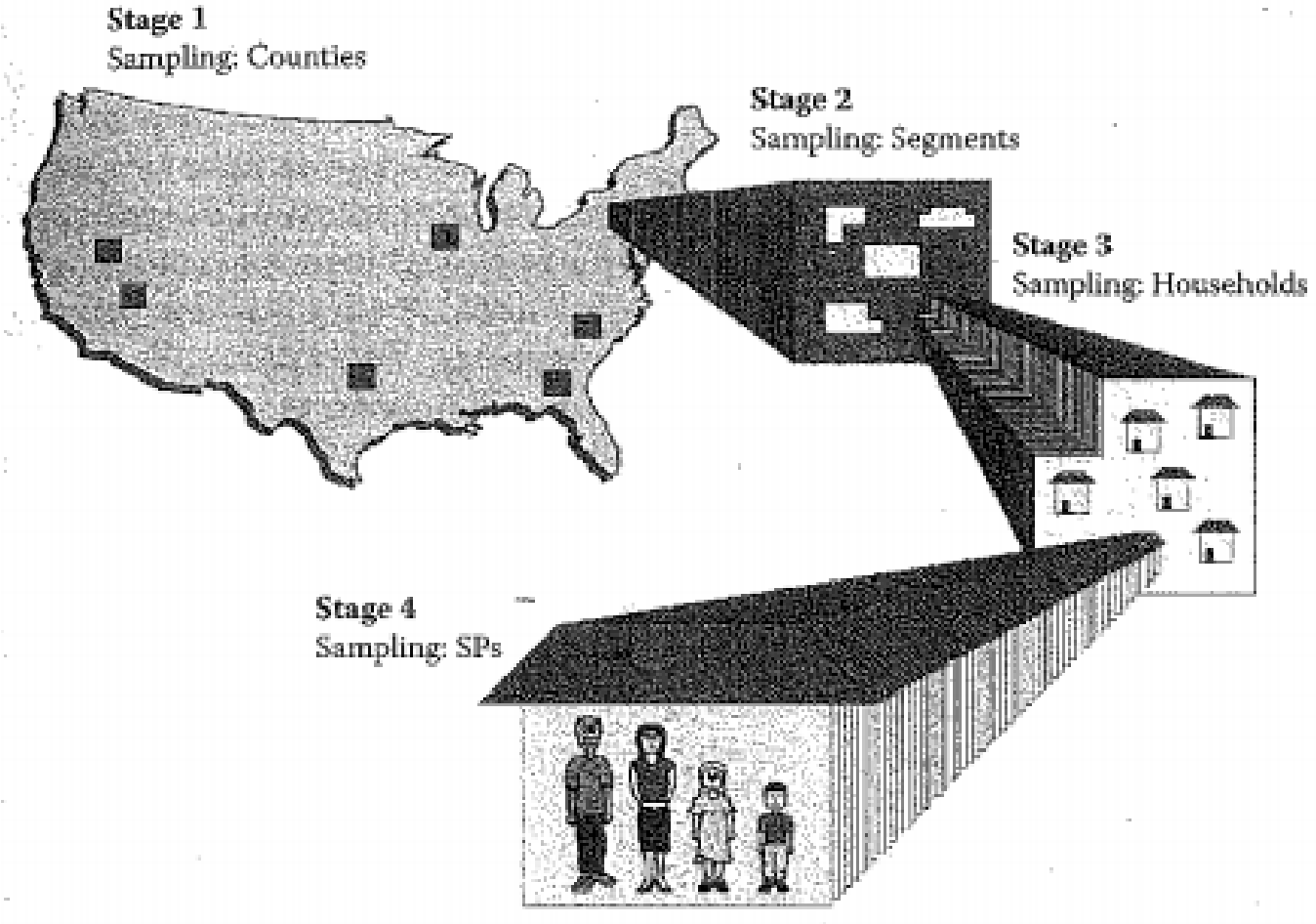
- applies only where selection of population elements is without replacement and is generally assumed to be equal to 1 in practice if the relative size of the SRS is less than 5% of the population size; in such cases, it will have little impact and can safely be ignored
- used in the calculation of the standard error of the estimate when the sampling fraction (the number of elements or respondents sampled relative to the population) becomes large

Multistage Area Probability Sampling

1. Primary Stage Sampling: the **primary sampling unit** (PSU), generally speaking, is the first unit that is sampled in the design, or the highest-level groupings or "clusters" of sample observations
   - in many multistage designs, the PSUs are often single counties or groupings of geographically contiguous counties
   - ideally, the populations within PSUs are reasonably heterogeneous to minimize intraclass correlation for survey variables but small enough in size to facilitate cost-efficient travel
2. Secondary Stage Sampling: also known as **area segments**
3. Third and Fourth Stage Sampling: often of housing units and eligible respondents

Note that you do not need to use the same sampling method at each of these sampling levels.

Illustration of Multistage Area Probability Sampling



Why Use Stata?

- includes the convenient method of declaring the survey or complex design variables one time prior to analysis using the `svyset` command
  - this allows users to specify different forms of sampling weights (probability or replicate weights), indicating the flexibility and range of this software
- the `svyset` command syntax also includes options to declare variables for finite population corrections and explicit poststratification adjustments (when available)
- every analytic command includes design-based significance testing options if statistically appropriate and tenable; each also has additional analytic options available, and most support a variety of postestimation commands
- consistency of syntax between the survey and SRS commands

Getting Started in Stata: `svyset`

Using the `svyset` command allows you to tell Stata about your specific survey design characteristics. Stata will then automatically adjust the parameter estimates to account for the sampling design.

For a single-stage design...

```
svyset [psu] [weight] [, design_options options]
```

And for a multi-stage design...

```
svyset psu [weight] [, design_options]
[ || ssu, design_options] ... [options]
```

Getting Started in Stata: `svyset`

...where everything in square brackets is optional:

- `psu` : may be _n (observation) or a variable name; in the single-stage syntax, psu is optional and defaults to _n, which indicates that individuals were randomly sampled if the design does not involve clustered sampling. If you do have a clustered sampling design, specify a variable name instead that contains identifiers for the clusters.
- `weight` : to specify a sampling or importance weight, e.g., [`pweight=`*varname*]
- Useful design options
  - `strata(`*varname*`)` : variable identifying strata
  - `fpc(`*varname*`)` : finite population correction
  - `weight(`*varname*`)` : stage-level sampling weight

See `help svyset` for more options. Now to the example do-file!

References

- Heeringa, Steven G., Brady T. West, and Patricia A. Berglund. 2010. *Applied Survey Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Korn, Edward L. and Barry I. Graubard. 1999. *Analysis of Health Surveys*. New York, NY: John Wiley & Sons, Inc.
- Levy, Paul S. and Stanley Lemeshow. 2008. *Sampling of Populations: Methods and Applications*, 4th edition. New York, NY: John Wiley & Sons, Inc.

Thank you!

Thanks for attending today!

For help and advice with your data analysis, contact the StatLab to set up an appointment: statlab@virginia.edu

Sign up for more workshops or download past workshop materials: http://data.library.virginia.edu/statlab/

Register for the Research Data Services newsletter to stay up-to-date on StatLab events and resources: http://data.library.virginia.edu/newsletters/