Applied Statistics Lab

quantitative initiative for policy and social research

Applied Survey Data Analysis

Module 1: Sampling Methods

March 30, 2013

Why The Sample Design Matters in the Analysis of Survey Data

Survey Sampling

"Methods for selecting a sample of the population in order to make inferences about the whole population." (Kish 1965)

Method for selecting a sample will determine how to estimate the sampling variability or standard error of a parameter of interest.

$$se(\bar{y}) = \sqrt{V(\bar{y})}$$

Our focus: Probability samples

- In probability sampling, every element in the **population** has a known non-zero chance of selection.

- This is different from “convenience sampling” (e.g., snowball sampling)

Survey Sampling Methods

- When every element in the population does have the same probability of selection, this is known as an “equal probability of selection method” (EPSM) design.

- Example of EPSM: Simple Random Sampling (SRS)

Simple Random Sampling

- SRS is EPSEM
- A probability sample in which each element has an independent and equal *chance* of being selected equal to

$$f = \frac{n}{N}$$

- Applicable when population is small, homogeneous & readily available.
- Problems
  - Expensive and not very practical
  - May not be representative
  - Minority subgroups of interest in population may not be present in sample in sufficient numbers for the study

The Alternative: Complex Samples

- What is a “complex sample”?

A probability sample developed using sampling procedures such as **stratification, clustering and weighting** designed to:

  - improve statistical efficiency
  - reduce costs
  - improve precision for subgroup analyses relative to SRS

- Equal probabilities of selection may no longer hold

- Observations might not be independent

- Unbiased estimates with measurable sampling variability (i.e. standard errors) are still possible

Why do we need special statistical techniques?

- Most regular statistical software analyzes survey data as if the data were collected using simple random sampling (SRS), which is rarely the case.

- Analyzing complex sample data assuming a simple random sample can lead to underestimated standard errors since the standard errors of complex sample designs tend to be smaller or larger, but usually larger than those of a simple random sample.

- When surveys are stratified, clustered or weighted, special statistical techniques are needed to take into account the design features of complex samples.

Examples from two surveys

General Social Survey (GSS)

American National Election Study (ANES)

General Social Survey (GSS)

"Conducts basic scientific research on the structure and development of American society with a data collection program designed to both monitor societal change within the United States and to compare the United States to other countries"

http://www3.norc.org/gss+website/

Statements from GSS's Technical Note

"The General Social Survey (GSS) is a …multi-stage cluster sample of housing units for the entire United States. Since the sample for the GSS is a cluster sample, standard errors are larger for the GSS than simple random sample calculations" (page 1)

"To correctly calculate standard errors, design variables must be used in statistical software... Without these design variables, statistical software will assume a simple random sample and underestimate standard errors " (page 1)

Source: Calculating Design-Corrected Standard Errors for the General Social Survey, 1988-2010.
http://publicdata.norc.org:41000/gss/documents//OTHR/GSS%20design%20variables.pdf

Example: Generalized Trust Model based on GSS Data

Variable Name	Description	recoding
Trust	Generally speaking, would you say that most people can be trusted or that you can't be too careful in life.	Most people can be trusted=1 Can't be too careful=0
Age	Age	Years
Sex	Gender	(Female =1)
Educ	Education	Years
Marital	Marital status	(married=1)
wrkstat	Employment status	(employed=1)
Goodlife	The way things are in America, people like me and my family have a good chance of improving our standard of living -- do you agree or disagree?	Strongly agree, agree vs. Neither agree or disagree, disagree, strongly disagree
Race	Race	Non-Hispanic White, Hispanic, Black, Others
Income06	(1) Less than $19,999 (2) $20,000 -39,999 (3) $40,000-74,999 (4) $75,000 or more	Dummy variables for each quartile

Logistic Regression Model of Generalized Trust

	Model 1		
	SRS SEs		
	Coef.	s.e.	p
Age	0.03***	0.00	0.000
Female	**-0.38****	**0.14**	**0.007**
Years of Schooling	0.23***	0.03	0.000
Married	0.10	0.15	0.517
Employed	0.17	0.15	0.266
Optimistic views of life	**0.40****	**0.14**	**0.005**
Household income 1st quartile	-0.17	0.24	0.470
Household income 2nd quartile	-0.37	0.22	0.093
Household income 3rd quartile	-0.15	0.19	0.433
Hispanic	**-0.59***	**0.26**	**0.026**
Black	-1.23***	0.24	0.000
Others	-0.24	0.36	0.507
Constant	-4.82***	0.57	0.000
N	1135		
* p<0.05, ** p<0.01, *** p<0.001			

Logistic Regression Models of Generalized Trust

	Model 1			Model 2		
	SRS SEs			Design-based SEs		
	Coef.	s.e.	p	Coef.	s.e.	p
Age	0.03***	0.00	0.000	0.03***	0.01	0.000
Female	**-0.38****	**0.14**	**0.007**	**-0.40***	**0.18**	**0.026**
School completed	0.23***	0.03	0.000	0.27***	0.03	0.000
Married	0.10	0.15	0.517	0.30	0.17	0.084
Employed	0.17	0.15	0.266	0.18	0.19	0.349
Optimistic views of life	**0.40****	**0.14**	**0.005**	**0.48***	**0.19**	**0.012**
Household income 1st quartile	-0.17	0.24	0.470	-0.02	0.30	0.937
Household income 2nd quartile	-0.37	0.22	0.093	-0.23	0.24	0.350
Household income 3rd quartile	-0.15	0.19	0.433	-0.17	0.19	0.380
Hispanic	**-0.59***	**0.26**	**0.026**	**-0.52**	**0.31**	**0.101**
Black	-1.23***	0.24	0.000	-1.30***	0.28	0.000
Others	-0.24	0.36	0.507	-0.31	0.47	0.509
Constant	-4.82***	0.57	0.000	-5.91***	0.72	0.000
N	1135			1135		
* p<0.05, ** p<0.01, *** p<0.001						

Applied Statistics Lab

qipsr quantitative initiative for policy and social research

ANES surveys

"The mission of the ANES is to inform explanations of election outcomes by providing data that support rich hypothesis testing, maximize methodological excellence, measure many variables, and promote comparisons across people, context, and time"

http://www.electionstudies.org/

Statements from ANES's Technical Note

"ANES data require special analysis techniques because the respondents are not selected using a simple random sample…" (page 14)

"Proper analysis of ANES data requires the use of software that can weight the data and produce design-consistent estimates." (page 14)

Source: How to Analyze ANES Survey Data
http://www.electionstudies.org/resources/papers/nes012492.pdf

Example based on ANES data: Vote for Obama in the 2008 Election

- **Dependent Variable: Voting for Obama in the 2008 election**
- Coded 1 if the respondent voted for Obama and 0 if the respondent voted for another candidate.
- Nonvoters excluded

Independent Variables

- Party ID *(1 strong Democrat to 7 strong Republican)*
- Feeling thermometers for Obama and McCain *(Scale=0-100)*
- Sex and educational attainment
- Respondent is black
- Belief about the Bible being the word of God
- Belief whether the Iraq war was worth the cost *(worth/not worth it)*
- Belief whether homosexuals should be allowed to serve in the armed forces *(5 strongly should not be allowed to 1 strongly should be allowed)*
- Interviewer's assessment of whether the respondent seemed well informed. *(1 very low to 5 very high)*

Logistic regression analyses with the 2008 ANES Time Series data

	Model 1			Model 2		
	SRS SEs			Design-based SEs		
	Coef.	s.e.	p	Coef.	s.e.	p
Party ID	-3.4***	0.47	0.000	-3.7***	0.39	0.000
Obama FT	7.8***	0.72	0.000	8.2***	1.31	0.000
MC Cain FT	-5.4***	0.73	0.000	-5.4***	0.77	0.000
Bible word of God	-1.1**	0.40	0.006	-1.2*	0.55	0.032
Gays in Military	0.6	0.40	0.127	0.9*	0.37	0.021
Iraq war worth costs	-0.8*	0.31	0.012	-0.7	0.36	0.056
Appeared informed	1.1*	0.52	0.036	1.0	0.66	0.119
Education	-3.7***	1.05	0.000	-4.2***	1.19	0.001
Female	0.2	0.27	0.382	0.0	0.25	0.955
Black	2.5***	0.63	0.000	2.6***	0.68	0.000
Constant	2.7**	1.05	0.010	2.9**	1.01	0.005
* p<0.05, ** p<0.01, *** p<0.001						

Variance under SRS and Complex Samples

Variance under SRS

- Estimate the mean from the sample as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Then the variance of this estimate is

$$V(\bar{y}) = (1 - f)\frac{s^2}{n}$$

$$\text{where } f = \frac{n}{N} \text{ and } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Variance under Complex Samples: Design Effects

- Design effect (DEFF) is a measure of the relative effectiveness of the sample design, compared to a SRS.

$$d^2 = v(\bar{y}) / v_{srs}(\bar{y})$$

Where,

$$v_{srs}(\bar{y}) = s^2 / n$$

- The square root of DEFF is known as the “root design effect,” or DEFT, and serves as an “inflation factor” for the standard errors obtained using the complex sample.

$$DEFT = \sqrt{d^2}$$

Interpretation of DEFT

- We will consider how to incorporate complex design and calculate design effects

Interpretation of DEFT

  - Deft = 1: No effect of sample design on standard error.
  - Deft>1: Sample design inflates the standard error of the estimate.
  - Deft<1: Sample design increases efficiency (reduces s.e.) of estimate.

Design Effect for "Age": Examples from Three Surveys

Survey	Mean	Std. Err. under Complex Samples	DEFF	DEFT
ANES	46.09	0.551	1.92	1.39
GSS	44.96	0.543	1.88	1.37
AmericasBarometer-Colombia 2010	37.15	0.195	0.246	0.50

"Typical" Consequences of Complex Sample Designs

Impact on precision of estimates (size of standard error):

- Stratification: $-\Delta$
- Clustering: $+\Delta$
- Weighting: $+\Delta$

Stratification

Stratification

- Stratification is the process by which the population is divided into subgroups.

  Example: males & females; age groups; regions…

- Through stratification, we can enforce the sample to be representative based on the characteristics we use to stratify the sample

- Sampling is then conducted separately in each subgroup or stratum

- Stratification helps us increase the precision of the sample. It reduces the standard error.

Example: Six Regions in Colombia

- Source: AmericasBarometer 2012 survey

Strata in Colombia

Strata	Freq.	Percent	Cum.
Atlantic	326	21.65	21.65
Bogota	231	15.34	36.99
Central	349	23.17	60.16
Eastern	277	18.39	78.55
Pacific	269	17.86	96.41
Old National Territories	54	3.59	100.00
Total	1,506	100.00	

Proportionate vs. Disproportionate Stratification.

- The sampling fraction is the size of the sample (n) divided by the size of the population (N).
- If the same sampling fraction is used per stratum this is referred to as **proportionate** stratification.
- If the sampling fraction is not the same in each stratum, this is referred to as **disproportionate** stratification.

Example: Proportionate Stratification

- The sampling fraction is 10% for each stratum
- Notice: In this case, the sample has the same percent distribution as the population (self-weighted sample).

Gender	Population	**Percent**	Sample	**Percent**
Men	800	**80%**	80	**80%**
Women	200	**20%**	20	**20%**
Total	1,000		100	

Example: Disproportionate Stratification

- The sampling fraction is 6.25% for Men and 25% for Women.
- Notice: In this case, the sample has a different percent distribution than the population.
- Weight is the inverse of the sampling fraction (16 and 4).

Gender	Population	**Percent**	Sample	**Percent**
Men	800	**80%**	50	**50%**
Women	200	**20%**	50	**50%**
Total	1,000		100	

Design Effect due to Stratification

- Proportionate stratification leads to an increase in survey precision (smaller standard errors), when compared to a design with no stratification.

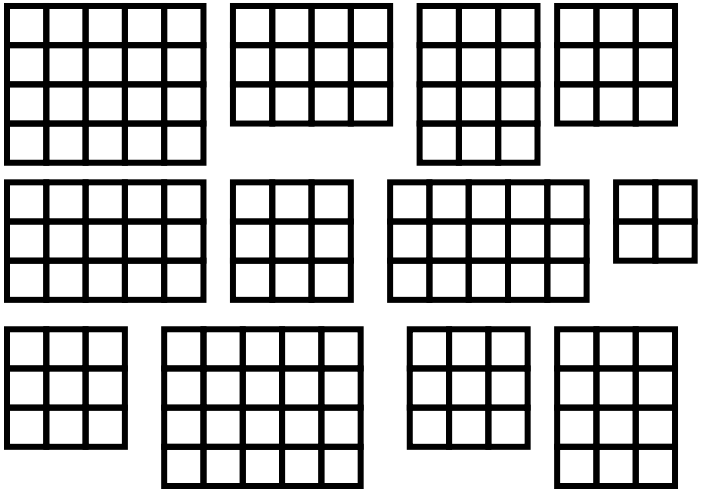
  Ex: Self-weighted samples

- Disproportionate stratification in contrast can have varying effects, increasing or decreasing precision, depending on the level of variance for a given characteristic within the over-sampled stratum.

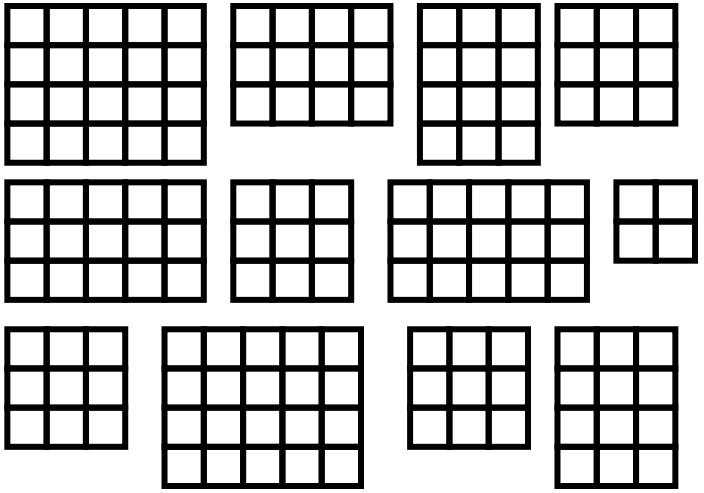
- Disproportionate stratification also requires weights to give unbiased cross-strata estimates. Otherwise, if weights are not used, the over-sampled strata will have an influence on overall population estimates disproportionate to their actual population size.

Clustering

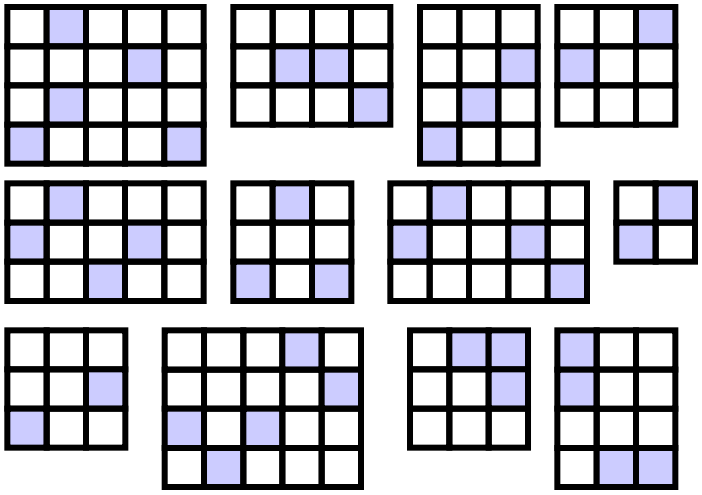
Example: Difference Between Clustering and Stratification



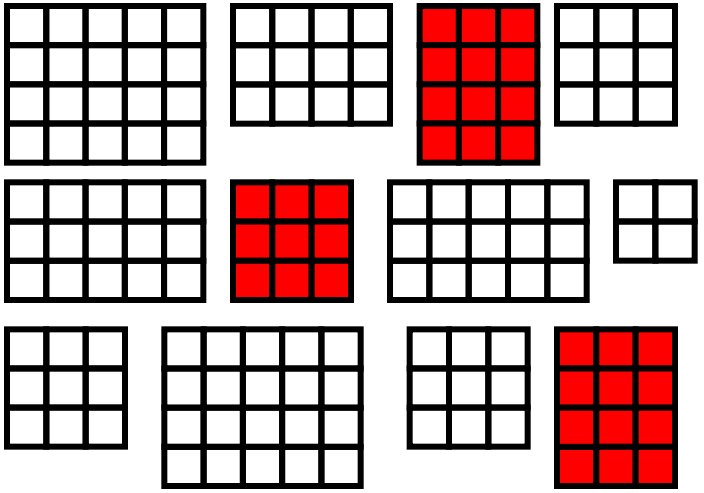
Population of H strata, stratum h contains n_l units



Population of C clusters



Take simple random sample in *every* stratum



Take SRS of clusters, sample every unit in chosen clusters

Clustering

- Clusters are sampling units containing several elements
  - **One stage sample:** sample households and individuals within households
  - **Multi-stage sample:** sample blocks, then households within blocks and finally, individuals within households

- **Higher-level cluster=Primary Sampling Units (PSU).**

- More cost effective than simple random sampling, cutting down fieldwork costs through making interviews more geographically concentrated

- Usually, clusters are homogeneous. That causes:

  -Reduced number of “effective” observations

  -Confidence intervals widen (increases variance in the sample estimations)

Clusters in 2010 AmericasBarometer survey for Colombia

Stratum	#PSU (clusters)	#Obs	Min	Mean	Max
Atlantic	13	326	12	25.1	42
Bogota	4	231	8	57.8	175
Central	15	349	12	23.3	70
Eastern	12	277	12	23.1	27
Pacific	9	269	15	29.9	76
Old National Territories	3	54	14	18.0	22
Total	56	1506	8	26.9	175

Clusters in 2008 ANES Survey

Stratum	#PSU (clusters)	#Obs	Min	Mean	Max
1	2	21	7	10.5	14
2	3	29	5	9.7	16
3	6	59	7	9.8	16
4	62	1959	13	31.6	74
5	3	52	9	17.3	26
6	9	85	3	9.4	18
7	3	44	14	14.7	15
8	2	31	11	15.5	20
9	4	42	5	10.5	14
Total	94	2322	3	24.7	74

Design Effect due to Clustering

- Clustering tends to increase the standard error of survey estimates relative to SRS of the same size because observations within a cluster are similar and there add less information than independently selected observations

- The design effect for a cluster sample depends on the average size of the clusters (*B*) and the homogeneity of the elements within the clusters, measured by the intra-class correlation (ρ):

$$D^2(\bar{y}) = 1 + (B-1)\rho.$$

- The intra-class correlation may be estimated by

$$\hat{\rho} = \frac{d^2(\bar{y}) - 1}{B-1}.$$

Intra-Class Correlation ρ

- Since elements in a cluster tend to be similar to one another, ρ is virtually always positive. For human populations, a positive ρ may be due to:
  - *Self selection*: wealthy households tend to reside in wealthy neighborhoods and poor households in poor neighborhoods.
  - *Interaction*: shared attitudes among neighbors.

- The magnitude of ρ depends on:
  - The variable under study (e.g., political leaning, newspaper read, age).
  - The nature of the clusters (e.g., households, city blocks, counties).
  - The size of the clusters.

Variance under Clustering

- Variances are inflated under clustering by a factor depending on
  - Cluster size (denoted *B*)
  - Intra-class correlation (denoted ρ)

$$V\left(\bar{y}\right) = \left(1 + \rho\left(B - 1\right)\right) V\left(\bar{y}_{\text{SRS}}\right)$$

$$d^2 = v(\bar{y}) / v_{srs}(\bar{y})$$

Weighting

What is a “Weight” ?

- A weight is used to indicate the relative strength of an observation.
- With unweighted data, each case is counted equally.
- With weighted data, each case is counted relative to its representation in the population.
- Weights allow analyses that represent the target population.

Weighting

Weighting is used to compensate for…

- Unequal probabilities of selection -- Over-sampling of specific cases or disproportionate stratification
- Nonresponse (typically, a unit that fails to respond)--Propensity to respond may depend on age, race/ethnicity, gender, place of residence
- In post-stratification to adjust weighted sample distributions for certain variables to make them conform to the known population distribution
- **Summary: weights are used to improve the accuracy (minimize bias) of sample estimates and to compensate for non-coverage and nonresponse**.

Design Effect due to Weighting

- Weights almost always increase the standard errors of your estimates
- When the variance of the weight variable is large, this results in standard errors that are larger than they would be for un-weighted estimates
- Weights introduce instability into your data. Some researchers like to “trim” or normalize to reduce the variance of the weights
- Trade off between less instability or more accurate representativeness.

Applied Statistics Lab

quantitative initiative for
policy and social research