



Applied Econometrics for Practitioners: Using Household Surveys to Inform Policy in Nepal

Dean Jolliffe, Hiroki Uematsu, Ganesh Thapa
Kathmandu, Nepal
01/24/2018




1



Announcements / reminders


- ◆ Reminder: P Set 1 due Jan 26, beginning of class
- ◆ For next Monday, read the Iraq Lancet papers and commentary.
- ◆ Background materials for this course, see Deaton text and two pdf chapters from Levy & Lemeshow
- ◆ NPR report on prevalence of smoking in the gay community in the US. Interpret in the context of correlation and causation.

2



Sample weight comments (HIV story, zero weights and finish example from last class)

1. Why is HIV article a story about sample weights (and sampling more generally)?
2. 2011 Nepal Living Standard Survey (NLSS)
Remaining comment: Under/oversampling

weights_example.do  `bw_example.do [Command line]`

3

Comment

(Stata weights example)

- ◆ Variance and standard deviation of the distribution of y

$$\text{Var}(y) = \frac{1}{(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{Std. Dev.} = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- ◆ (Variance &) standard error of the mean of y, **Assuming SRS**

$$\text{Var}(\bar{y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{SE}(\bar{y}) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2}$$

4

Outline for Sample Design issues

- ◆ Sample Frame & **Simple Random Sample**
- ◆ Describe **Stratification**, purpose 1: reduce variance, algebraic result & graphical illustration.
- ◆ Stratification, purpose 2: ensure sample coverage. Eg. of effect on **weights & means**.
- ◆ Describe **Multi-stage** design, purpose & derive result of increased variance. Examples showing effect and how to program.

5

Review

Simple Random Sample (SRS)

SRS: A single-stage draw without stratification from the frame.

Characteristic of SRS: All observations from the frame have equal probability of selection.

Advantage of SRS: Conceptual and statistical simplicity. Most all formulas you've seen in stats and econometrics are based on SRS.

Disadvantages of SRS, primary motivation for stratification:

- ◆ Can't guarantee coverage of particular subpopulations. Comparison of regions (see stratification)
- ◆ Low-probability outcomes hard to measure (see stratification)

Stratification can also improve precision (reduce std. errors).

6

Stratification - review

Because each sample is an independent, random draw within each stratum, the variance of \bar{y} for the full sample can be expressed as ...

$$\hat{\sigma}_{\bar{y}, SS}^2 = \frac{1}{n} \sum_{h=1}^H \omega_h \hat{\sigma}_{y,h}^2,$$

where SS abbreviates 'stratified sample', n is the sample size, H is the number of strata, and

ω_h is the proportion of the sample in h (n_h/n).

Or, in words, the sample variance of \bar{y} is the weighted average of the variance in each stratum.

7

Stratification – review (cont.)

$$\hat{\sigma}_{SS}^2 \cong \hat{\sigma}_{SRS}^2 - \frac{1}{n} \sum_{h=1}^H \omega_h (\bar{y}_h - \bar{y})^2$$

=> Precision is increased as

... heterogeneity across strata \uparrow .

Or in other words the more \bar{y}_h differs from \bar{y} .

... homogeneity within strata \uparrow .

Decreasing the variance within each stratum reduces the sum of the strata variances, or reduces $\hat{\sigma}_h^2$.

Caveats: ω_h and relative magnitudes in practice.

8

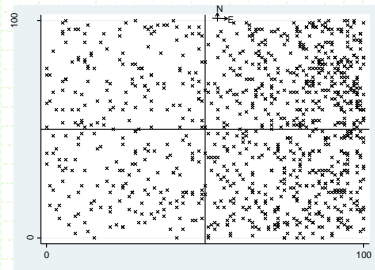
Stratification - Illustration of increased precision

Overhead Slides to provide some intuition

$$\hat{\sigma}_{SS}^2 < \hat{\sigma}_{SRS}^2$$

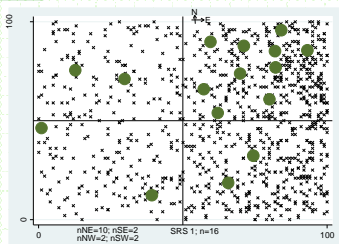
9

Population



10

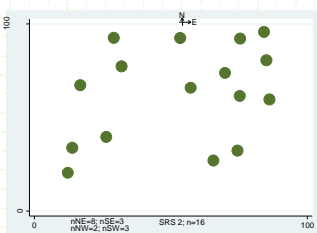
SRS1



$$\bar{y}_{SRS1} \approx \bar{Y}_{POP}$$

11

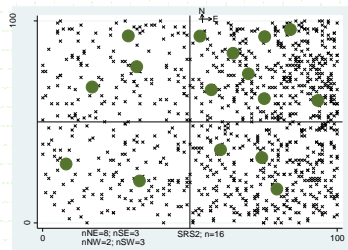
SRS2



$$\bar{y}_{SRS2} < \bar{Y}_{POP}$$

12

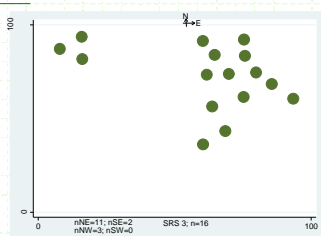
SRS2



$$\bar{y}_{SRS2} < \bar{y}_{POP}$$

13

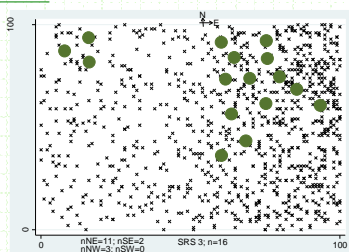
SRS3



$$\bar{y}_{SRS3} > \bar{y}_{POP}$$

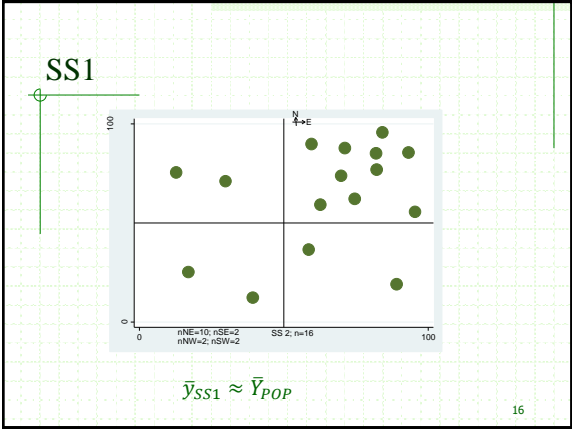
14

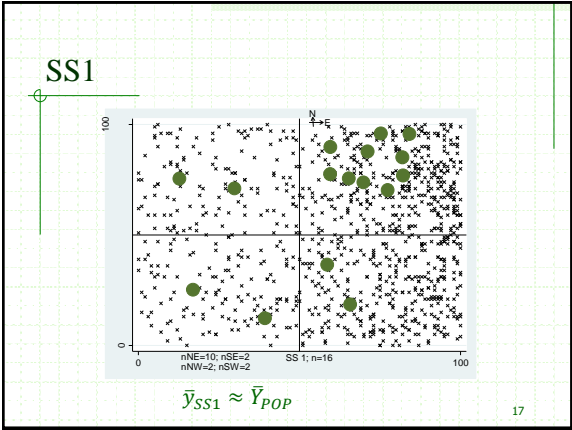
SRS3

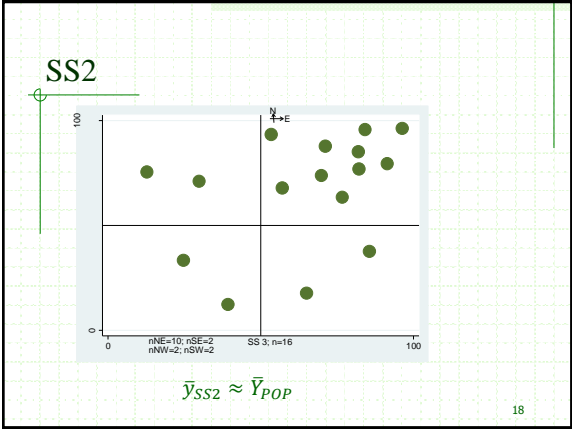


$$\bar{y}_{SRS3} > \bar{y}_{POP}$$

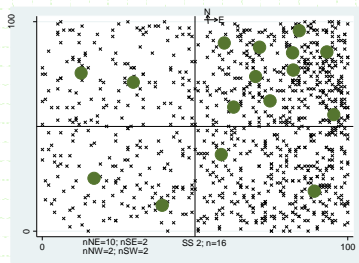
15







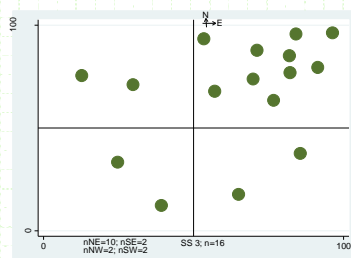
SS2



$$\bar{y}_{SS2} \approx \bar{y}_{POP}$$

19

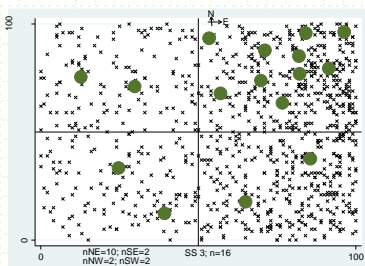
SS3



$$\bar{y}_{SS3} \approx \bar{y}_{POP}$$

20

SS3



$$\bar{y}_{SS3} \approx \bar{y}_{POP}$$

21

Repeat SRS and SS 100 times

Estimate $\text{Var}(\bar{y}_{\text{SRS}})$ and $\text{Var}(\bar{y}_{\text{SS}})$

$\text{Var}(\bar{y}_{\text{SRS}})$ and $\text{Var}(\bar{y}_{\text{SS}})$

$$\text{Var}(\bar{y}_{\text{SRS}}) = \frac{1}{(n-1)} \sum_{i=1}^{100} (\bar{y}_{\text{SRS}_i} - \bar{Y})^2$$

$$\text{Var}(\bar{y}_{\text{SS}}) = \frac{1}{(n-1)} \sum_{i=1}^{100} (\bar{y}_{\text{SS}_i} - \bar{Y})^2$$

?

$\text{Var}(\bar{y}_{\text{SS}}) > = < \text{Var}(\bar{y}_{\text{SRS}})$

22

Sample Design Continued

Multi-stage or clustered designs

- ◆ Describe **Multi-stage** design & purpose
- ◆ Derive result of increased variance
- ◆ Example showing magnitude of effect, and how to program.

23

Second aspect of sample design:

Multi-stage, explanation

- ◆ Basic idea: After stratifying the sample frame. Carry out random draw within each stratum. Instead of one-draw, SRS; select observations in stages.
- ◆ Stage 1: sort and group observations in each stratum. For example, create grid of blocks covering all obs in each stratum. Randomly select some blocks. (Blocks, Enumeration Areas - EAs, Primary Sampling Units - PSUs, Clusters)
- ◆ Stage 2: From the selected blocks, randomly draw ultimate sampling units (USUs), such as households. (or individuals, families, firms, etc.)

24

Multi-stage or clustering – purpose

- ◆ What's the purpose of clustering?
- ◆ Almost entirely an issue of cost. (2 points)
- ◆ 1. Frame needs to contain full list of PSUs with pop estimate of USUs, not necessary to have full list of USUs. (village example)
- ◆ 2. Consider drawing a sample of a few thousand people from some country with population in the millions. With SRS, you wouldn't expect any of these observations to be close to each other. Odds very low of this. Travel costs are high. By clustering, you can group together observations.
- ◆ Disadvantage: LOSS of PRECISION

25

Clustering - an algebraic statement regarding loss of precision.

Consider some observation i from cluster c :

$$y_{i,c} = \mu + \alpha_c + \varepsilon_{i,c}$$

All observations can be decomposed into parts where

μ is the overall mean, α_c is the mean cluster effect, and $\varepsilon_{i,c}$ is a random variable with mean 0 and variance σ_ε^2 .

$$\text{Simple Random Sampling: } \text{Var}(\bar{y}_{\text{SRS}}) = \frac{\sigma_\varepsilon^2}{n}$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{1}{k} \sigma_\alpha^2 + \frac{1}{km} \sigma_\varepsilon^2$$

where σ_α^2 is the variance of α_c , k is the number of clusters and m is the number of observations within each cluster. ($km=n$)

26

Clustering - an algebraic statement regarding loss of precision.

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{1}{k} \sigma_\alpha^2 + \frac{1}{km} \sigma_\varepsilon^2$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{m \sigma_\alpha^2 + \sigma_\varepsilon^2}{k m}$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{\sigma_\alpha^2 + m \sigma_\alpha^2 - \sigma_\alpha^2 + \sigma_\varepsilon^2}{k m}$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2 + (m-1) \sigma_\alpha^2}{n}$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2}{n} + \frac{(\sigma_\alpha^2 + \sigma_\varepsilon^2) (m-1) \sigma_\alpha^2}{(\sigma_\alpha^2 + \sigma_\varepsilon^2) n}$$

27

Clustering - an algebraic statement regarding loss of precision.

$$\text{Var}(\bar{y}_{CS}) = \frac{\sigma_a^2 + \sigma_e^2}{n} + \frac{(\sigma_a^2 + \sigma_e^2)(m-1)\sigma_a^2}{(\sigma_a^2 + \sigma_e^2)n}$$

Note: $\text{Var}(\bar{y}_{SRS}) = \frac{\sigma_a^2 + \sigma_e^2}{n}$ and let $\rho = \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_e^2)}$

$$\text{Var}(\bar{y}_{CS}) = \text{Var}(\bar{y}_{SRS}) + \frac{\text{Var}(\bar{y}_{SRS})(m-1)\sigma_a^2}{(\sigma_a^2 + \sigma_e^2)}$$

$$\text{Var}(\bar{y}_{CS}) = \text{Var}(\bar{y}_{SRS})[1 + (m-1)\rho]$$

$$\text{Design Effect, or Deff} = [1 + (m-1)\rho] = \frac{\text{Var}(\bar{y}_{CS})}{\text{Var}(\bar{y}_{SRS})}$$

Clustering - loss of precision

$$\text{Var}(\bar{y}_{CS}) = \text{Var}(\bar{y}_{SRS})[1 + (m-1)\rho]$$

$$\Rightarrow \text{Var}(\bar{y}_{CS}) \geq \text{Var}(\bar{y}_{SRS})$$

Precision is decreased as

... ρ increases, or in words, as the correlation of the y_{ic} obs increases within each cluster.

The more similar the observations are within clusters.

... m increases. For a fixed sample size, as the number of observations within a cluster increases.

Why? Less coverage of the variation in the frame.

As m shrinks to 1, CS moves to SRS.

29

Clustering - magnitude of loss in precision

Poverty. Howes and Lanjouw (1998), using household data from a stratified, two-stage sample, show that correcting for design effects increases standard errors for the poverty indices between 26 and 64 percent.

Means. Deaton (1997), using household survey data from a complex design, shows that correcting for design effects increases the standard error for the national average household expenditure increases by 72 percent.

Regression. Scott, Andrew and Holt, Tim. "The Effect of two-stage Sampling on Ordinary Least Squares Methods." *Journal of American Statistical Association*, December 1982, 77 (380): 848-854.

30

90% Confidence Intervals for 1999 Poverty Estimates by Race
Correcting for Complex Sample Design (stratification & clustering)

90% Conf. Intervals from 2000 P60

	Percent Poor	Corrected for Complex Sample Design	Simple Random Sample
<i>Individuals by Race & Ethnicity</i>			
White	9.8	0.3	0.16
Non-Hispanic White	7.7	0.3	0.16
Black	23.6	1.2	0.66
Asian	10.7	1.6	0.83
Hispanic	22.8	1.2	0.50

Loss in precision from the complex sample design is very large. Confidence intervals doubled in size. We'll consider another example where the change is much smaller.

31

Clustering -
Illustration of decreased precision

Overhead Slides to provide some intuition

$$\hat{\sigma}_{SS}^2 < \hat{\sigma}_{SRS}^2 < \hat{\sigma}_{CS}^2$$

32

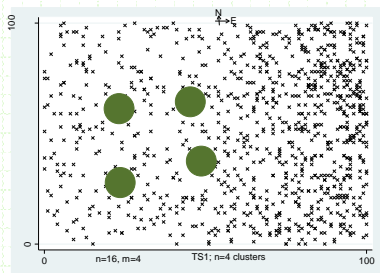
Two-stage



$$\bar{y}_{CS1} < \bar{y}_{POP}$$

33

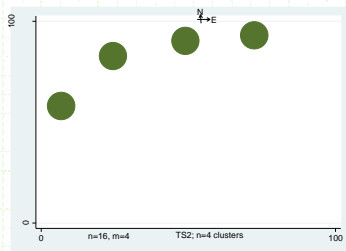
Two-stage TS1



$$\bar{y}_{CS1} < \bar{y}_{POP}$$

34

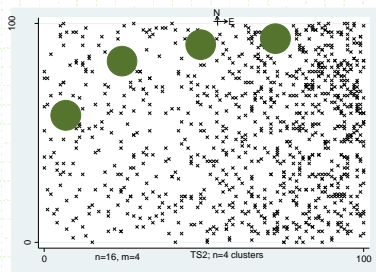
Two-stage TS2



$$\bar{y}_{CS2} > \bar{y}_{POP}$$

35

Two-stage TS2



$$\bar{y}_{CS2} > \bar{y}_{POP}$$

36

Repeat SRS and CS 100 times

Estimate $\text{Var}(\bar{y}_{\text{SRS}})$ and $\text{Var}(\bar{y}_{\text{CS}})$

$\text{Var}(\bar{y}_{\text{SRS}})$ and $\text{Var}(\bar{y}_{\text{CS}})$

$$\text{Var}(\bar{y}_{\text{SRS}}) = \frac{1}{(n-1)} \sum_{i=1}^{100} (\bar{y}_{\text{SRS}_i} - \bar{Y})^2$$

$$\text{Var}(\bar{y}_{\text{CS}}) = \frac{1}{(n-1)} \sum_{i=1}^{100} (\bar{y}_{\text{CS}_i} - \bar{Y})^2$$

?

$$\text{Var}(\bar{y}_{\text{CS}}) \geq \text{Var}(\bar{y}_{\text{SRS}})$$

37

Clustering - loss of precision (m)

$\text{Var}(\bar{y}_{\text{CS}}) = \text{Var}(\bar{y}_{\text{SRS}})[1 + (m-1)\rho]$
 $\Rightarrow \text{Var}(\bar{y}_{\text{CS}}) \geq \text{Var}(\bar{y}_{\text{SRS}})$

Precision is decreased as

... ρ increases, or in words, as the correlation of the $y_{i,c}$ obs increases within each cluster.

The more similar the observations are within clusters.

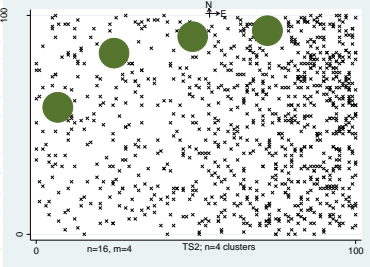
... m increases. For a fixed sample size, as the number of observations within a cluster increases.

Why? Less coverage of the variation in the frame.

As m shrinks to 1, CS moves to SRS.

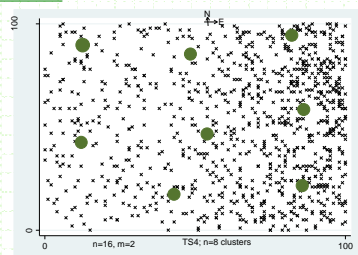
38

Two-stage TS2



39

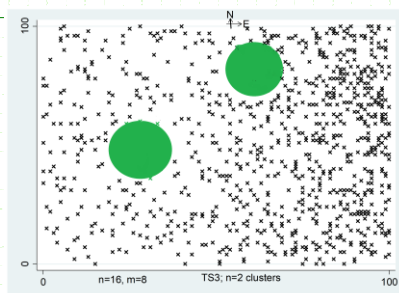
Two-stage TS4



$$\bar{y}_{CS3} \approx \bar{y}_{POP}$$

40

Two-stage TS3



$$\bar{y}_{CS3} > \bar{y}_{POP}$$

41

Complex sample design - some 'hands on' examples

Lots of software packages now handle complex designs:
CENVAR, SUDAAN, WesVar, Stata, PCCARP, Epi Info.
Stata has a suite of commands for complex design: svy*
In particular, svyset, svy: regress and svy: mean

help svy:

Help File Help File

42

Complex sample design -
some ‘hands on’ examples

Data: 2011 Nepal Living Standard Survey
5988 households. 14 Strata. 499 PSUs.

- 1. Unweighted sample estimate
- 2. Estimates corrected for stratification.
- 3. Estimates corrected for stratification and clustering.
- 4. Estimates corrected for stratification, clustering and weights.
- 5. Compare poverty in urban and rural Nepal.

Complex_example.do

Help File
