Survey Methods

20 February 2009

Non-random (non-probability) sampling

- Convenience
- Purposive
- Snowball Sampling
- Quota

Data analysis

Simplified by

simple random sample

(no stratification or clustering)

everyone responds

(no unit non-response)

respondents answer all (correct) questions

(no item non-response)

→ no weights, no design effect

Why unlikely?

Unit Non-response

- Check for bias
  - In characteristics and responses
    - Multi-mode follow-up
  - In characteristics
    - External data
    - Frame
- Adjust for non-response
  - Post-stratification weights

Complications

- **Weights:** Each respondent represents a different number of population elements
- **Design effect:** Adjusts standard error to account for the additional or reduced information on distribution of parameters in the population (deff = ratio of variance of estimator under actual sampling design to variance under SRS of same sample size)

Weighting

Reasons?

Post-stratification to adjust for non-response

Increase precision of estimates and comparisons of sub-groups through disproportionate stratification

Reduce costs through multi-stage clustering

Frame units different from population elements

Overlapping sampling frames

Weighting

Reasons?

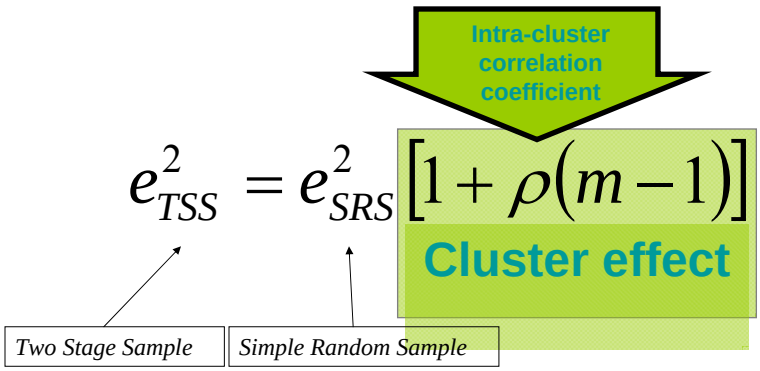
Descriptive statistics that accurately represent overall population

Complications

- **Weights:** Each respondent represents a different number of population elements
- **Design effect:** Adjusts standard error to account for the additional or reduced information on distribution of parameters in the population (deff = ratio of variance of estimator under actual sampling design to variance under SRS of same sample size)

Cluster effect

Standard error grows when the sample of size n is drawn from k PSUs, with m households in each PSU ($n=k \bullet m$)



Design effects

Motivations

- Increase precision
  - Proportionate stratification
- Increase precision of estimates/ comparisons for sub-groups
  - Disproportionate stratification
- Reduce costs
  - Multi-stage clustering
- Non-respondents
  - Post-stratification

Design effects

- Design effect specific to sample design **and** estimator

- Find a statistical consultant
- STATA
- Guidance in meta-data of public surveys

Item non-response

- MAR = Missing At Random = Probability of an item being missing depends only on other items that have been measured for that unit and no additional information as to the probability of being missing would be obtained from the unmeasured values of the missing items.
- Missing Completely At Random
- Missing Not At Random

Imputation

- Mean or median
- Mean or median by strata
- Model and sample
- Hot deck
- Nearest neighbor (propensity score match)
- …

Useful for modeling

NOT for descriptive statistics

Coding & checking

- Spreadsheets

- Full sample
  - Double entry
  - Check for invalid and extreme values
  - Codes
  - Keep
    - All data in “raw” form
    - Multiple versions of database
    - Complete meta-data

Reporting

Categorical vs. continuous/numerical

Univariate

Categorical

- Percent (or count)
- Lump categories, create “other”
- Bar and pie charts

Reporting

Categorical vs. continuous/numerical

Univariate

Numerical

- Mean and standard deviation
- Median and inner quartile
- Coefficient of variation (SD/mean)
- Histogram

Reporting

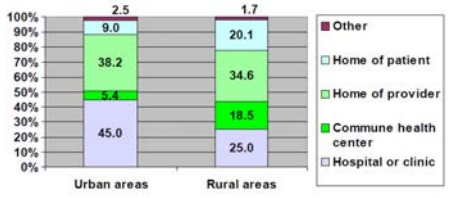
Categorical vs. continuous/numerical

Bivariate

Categorical

Cross-tabulate

Figure XVI.4. Use of health facilities among the population (all ages) that visited a health facility in the past four weeks, by urban and rural areas of Viet Nam, in 1992-1993 (Percentage)



Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 2,276.

Reporting

Categorical vs. continuous/numerical

Bivariate

Numerical

Correlation

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

Scatter plot

Categorical - numerical

Presenting results in tables & graphs

- Should contain sufficient information for interpretation independent of text
  - Sample/ sub-sample description and size
    - Master title: date, location, sampling
    - Each table/graph: sample size
- Variable description completely consistent with survey instrument
- Reasonable number of significant digits
- Provide insight on your research question