Introduction to Survey Data Analysis

Young Cho

Survey Research Laboratory
University of Illinois at Chicago
The Circle of Research Process

Real World

- Theory Evaluation
- Hypotheses Test
- Data Collection
- Sample
- Operationalization/Measurement
- Hypotheses
- Theory

Survey Research Laboratory
University of Illinois

info@srl.uic.edu
http://www.srl.uic.edu
Focus of the Seminar

- Data Cleaning/ Missing Data
- Sampling Bias Reduction
When Analyzing Survey Data . . .

With clear goals of analysis:

(1) Understand & evaluate survey design

(2) Screen the data

(3) Adjust for sampling design
1. Understand & Evaluate Survey

(1) Conductor of survey
(2) Sponsor of survey
(3) Measured variables
(4) Unit of analysis
(5) Mode of data collection
(6) Dates of data collection
(7) Geographic coverage
(8) Respondent eligibility criteria
(9) Sample design
(10) Sample size & response rate
Levels of Measurement

(1) Nominal
(2) Ordinal
(3) Interval
(4) Ratio
2. Data Screening

Examine raw frequency distributions for …

(a) out-of-range values (outliers)
(b) missing values
Out-of-Range Values

(1) Delete data

(2) Recode values
Missing Data

(a) Can reduce effective sample size

(b) May introduce bias
Reasons for Missing Data

1. Refusals (question sensitivity)
2. “Don’t know” responses (cognitive problems, memory problems)
3. Not applicable
4. Data processing errors
5. Questionnaire programming errors
6. Design factors
7. Attrition in panel studies
Effects of Ignoring Missing Data

(1) Reduced sample size - loss of statistical power

(2) Data may no longer be representative - introduces bias

(3) Difficult to identify effects
Assumptions on Missing Data

(1) Missing completely at random (MCAR)

(2) Missing at random (MAR)

(3) Ignorable

(4) Nonignorable
(1) **Missing completely at random**

- Being missing is independent of any variables.
- Cases with complete data are indistinguishable from cases with missing data.
- Missing cases are a random subsample of original sample.
(2) Missing at random (MR)

- The probability of a variable being observed is independent of the true value of that variable controlling for one or more variables.

- **Example:** Probability of missing income is unrelated to income within levels of education.
(3) **Ignorable missing data**

- The data are MAR.
- The missing data mechanism is unrelated to the parameters we want to estimate.

(4) **Nonignorable missing data**

- The pattern of data “missingness” is non-MAR.
Methods of Handling Missing Data

(1) **Listwise (casewise) Deletion**: uses only complete cases

(2) **Pairwise Deletion**: uses all available cases

(3) **Dummy Variable Adjustment**: missing value indicator method

(4) **Mean Substitution**: substitute mean value computed from available cases (cf. unconditional or conditional)
Methods of Handling Missing Data (cont’d)

(5) **Regression methods**: predict value based on regression equation with other variables as predictors

(6) **Hot deck**: identify the most similar case to the case with a missing value and impute the value
Methods of Handling Missing Data
(cont’d)

(7) **Maximum likelihood methods:** uses all available data to generate maximum likelihood-based statistics

- **Expectation Maximization (EM):** based on iterations of E (expected value) and M (maximization) to impute missing values that allow stochastic error terms to be added to imputations.

(8) **Multiple Imputation:** combines the methods of EM and ML to produce multiple data sets with imputed values for missing cases
Types of Survey Sample Designs

1. Simple Random Sampling
2. Systematic Sampling
3. Complex sample designs
   - stratified designs
   - cluster designs
   - mixed mode designs
Why complex sample designs?

(1) Increased efficiency

(2) Decreased costs

• Complex designs with clustering and unequal selection probabilities generally increase the sampling variance.

• Not accounting for the impact of complex sample design can lead to an underestimate of the sampling variance.
Sample Weights

• Used to adjust for differing probabilities of selection

• In theory, simple random samples are self-weighted.

• In practice, simple random samples are likely to also require adjustments for non-response.
Types of Sample Weights

(1) **Poststratification weights**: designed to bring the sample proportions in demographic subgroups into agreement with the population proportion in the subgroups.

(2) **Non-response weights**: designed to inflate the weights of survey respondents to compensate for nonrespondents with similar characteristics.

(3) **“Blow-up” (expansion) weights**: provide estimates for the total population of interest.
Syntax Examples of Design-Based Analysis in STATA, SUDAAN and SAS

**STATA**

```stata
svyset strata strata
svyset psu psu
svyset pweight finalwt
svyreg fatinlk age male black hispanic
```

**SUDAAN**

```sudaan
proc regress data="c:\nhanes.sav" filetype=spss desgn=wr;
nest strata psu;
weight finalwt
subpgroup sex race;
levels 2 3;
model fatinlk = age sex race;
```

**SAS**

```sas
proc surveyreg data=nhanes;
strata strata;
cluster psu;
class sex race;
model fatinlk = age sex race;
weight finalwt
```
In Summary, When Analyzing Survey Data . . .

(1) Understand & evaluate survey design

(2) Screen the data – deal with missing data & outliers

(3) If necessary, adjust for study design using weights and appropriate computer software