Introduction to Survey Data Analysis

Young Ik Cho, PhD
Research Assistant Professor
Survey Research Laboratory
University of Illinois
Focus of the Seminar

- Data Cleaning/ Missing Data
- Sampling Bias Reduction
When analyzing survey data...

1. Understand & evaluate survey design
2. Screen the data
3. Adjust for sampling design
1. **Understand & Evaluate Survey**

- Conductor of survey
- Sponsor of survey
- Measured variables
- Unit of analysis
- Mode of data collection
- Dates of data collection
1. Understand & Evaluate Survey

- Geographic coverage
- Respondent eligibility criteria
- Sample design
- Sample size & response rate
Levels of Measurement

- Nominal
- Ordinal
- Interval
- Ratio
2. **Data Screening**

Examine raw frequency distributions for...

(a) out-of-range values (outliers)
(b) missing values
2. Data Screening

Out-of-range values:
- Delete data
- Recode values
Missing Data:

- Can reduce effective sample size
- May introduce bias
Reasons for Missing Data

- Refusals (question sensitivity)
- Don’t know responses (cognitive problems, memory problems)
- Not applicable
- Data processing errors
- Questionnaire programming errors
- Design factors
- Attrition in panel studies
Effects of Ignoring Missing Data

- Reduced sample size—loss of statistical power
- Data may no longer be representative—introduces bias
- Difficult to identify effects
Assumptions on Missing Data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Ignorable
- Nonignorable
**Missing Completely at Random**

- Being missing is independent from any variables.
- Cases with complete data are indistinguishable from cases with missing data.
- Missing cases are a random sub-sample of original sample.
**Missing at Random (MAR)**

- 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.
Ignorable Missing Data

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

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

- **Listwise (casewise) deletion**: uses only complete cases
- **Pairwise deletion**: uses all available cases
- **Dummy variable adjustment**: missing value indicator method
Methods of Handling Missing Data

- **Mean substitution**: substitute mean value computed from available cases (cf. unconditional or conditional)

- **Regression methods**: predict value based on regression equation with other variables as predictors

- **Hot deck**: identify the most similar case to the case with a missing and impute the value
Methods of Handling Missing Data

- **Maximum likelihood methods**: use all available data to generate maximum likelihood-based statistics.

- **Multiple imputation**: combines the methods of ML to produce multiple data sets with imputed values for missing cases.
Types of Survey Sample Designs

- Simple random sampling
- Systematic sampling
- Complex sample designs
  - stratified designs
  - cluster designs
  - mixed mode designs
Why complex sample designs?

- Increased efficiency
- Decreased costs
Why complex sample designs?

- 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 overestimate 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 nonresponse.


**Types of Sample Weights**

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

- **Nonresponse weights**: designed to inflate the weights of survey respondents to compensate for nonrespondents with similar characteristics.

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

**STATA**

- `svyset strata strata`
- `svyset psu psu`
- `svyset pweight finalwt`
- `svyreg fatitk age male black hispanic`

**SUDAAN**

- `proc regress data="c:\nhanes.sav" filetype=spss desgn=wr;`
- `nest strata psu;`
- `weight finalwt`
- `subpgroup sex race;`
- `levels 2 3;`
- `model fatintk = age sex race;`
Syntax Examples of Design-Based Analysis in STATA, SUDAAN, & SAS

**SAS**

```sas
proc surveyreg data=nhanes;
strata strata;
cluster psu;
class sex race;
model fatintk = age sex race;
weight finalwt
```
In Summary

When analyzing survey data...

- Understand & evaluate survey design
  Screen the data—deal with missing data & outliers.

- If necessary, adjust for study design using weights & appropriate computer software.
Thank You!

www.srl.uic.edu