


Introduction to Survey Data Analysis

Linda K. Owens, PhD
Assistant Director for Sampling & Analysis
Survey Research Laboratory
University of Illinois at Chicago

1

Focus of the seminar


- Data cleaning/missing data
- Sampling bias reduction

Survey Research Laboratory 

2

When analyzing survey data...

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

Survey Research Laboratory 

3

1. Understand & evaluate survey

- Conductor of survey
- Sponsor of survey
- Measured variables
- Unit of analysis
- Mode of data collection
- Dates of data collection

Survey Research Laboratory 

4

1. Understand & evaluate survey

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

Survey Research Laboratory 

5

Levels of measurement

- Nominal
- Ordinal
- Interval
- Ratio


Survey Research Laboratory 

6

2. Data screening

Examine raw frequency distributions for...

- (a) out-of-range values (outliers)
- (b) missing values


Survey Research Laboratory 

7

2. Data screening

Out-of-range values:


- Delete data
- Recode values

Survey Research Laboratory 

8

Missing data:


- can reduce effective sample size
- may introduce bias

Survey Research Laboratory 

9

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

Survey Research Laboratory 

10

Effects of ignoring missing data


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

Survey Research Laboratory 

11

Assumptions on missing data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Ignorable
- Nonignorable

Survey Research Laboratory 

12

Missing completely at random (MCAR)

- 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.



13

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.



14

Ignorable missing data


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



15

Nonignorable missing data


- The pattern of data missingness is non-MAR.

Survey Research Laboratory 

16

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

Survey Research Laboratory 

17

Methods of handling missing data

- *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

Survey Research Laboratory 

18

Methods of handling missing data

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

Methods of handling missing data

- *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?

- Statistical software packages with an assumption of SRS underestimate the sampling variance.
- Not accounting for the impact of complex sample design can lead to a biased estimate of the sampling variance (Type I error).

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.

Survey Research Laboratory 

25

Syntax examples of design-based analysis in STATA, SUDAAN, & SAS

STATA

```
svyset strata strata  
svyset psu psu  
svyset pweight finalwt  
svyreg fatintk age male black hispanic
```

SUDAAN

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

Survey Research Laboratory 

26

Syntax examples of design-based analysis in STATA, SUDAAN, & SAS

SAS

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

Survey Research Laboratory 

27

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 and appropriate computer software.



28

Thank You!

www.srl.uic.edu



29
