Data Management

Department of Political Science and Government
Aarhus University

November 24, 2014

Data Management

- Weighting
- Handling missing data
  - Categorizing missing data types
  - Imputation
- Summary measures
  - Scale construction
  - Combining question branches
- Coding and editing
  - Open-ended questions
  - Marking problematic data
- Data preparation
  - Codebook creation
  - File formats
  - Archiving, access, and rights

Goal of Survey Research

- The goal of survey research is to estimate population-level quantities (e.g., means, proportions, totals)
- Samples estimate those quantities with uncertainty (sampling error)
- Sample estimates are unbiased if they match population quantities

Realities of Survey Research

- Sample may not match population for a variety of reasons:
  - Due to constraints on design
  - Due to sampling frame coverage
  - Due to intentional over/under-sampling
  - Due to nonresponse
  - Due to sampling error

Realities of Survey Research

- Sample may not match population for a variety of reasons:
  - Due to constraints on design
  - Due to sampling frame coverage
  - Due to intentional over/under-sampling
  - Due to nonresponse
  - Due to sampling error
- Weights can be used to “correct” a sample

Realities of Survey Research

- Sample may not match population for a variety of reasons:
  - Due to constraints on design
  - Due to sampling frame coverage
  - Due to intentional over/under-sampling
  - Due to nonresponse
  - Due to sampling error

- Weights can be used to “correct” a sample

- Weighting is never perfect
  - Limited to work with observed variables
  - Rarely have good knowledge of coverage, nonresponse, or sampling error
  - Weighting can increase sampling variance

Three Kinds of Weights

- Design Weights
- Nonresponse Weights
- Post-Stratification Weights

Design Weights

- Address design-related unequal probability of selection into a sample
- Applied to *complex survey designs*:
  - Disproportionate allocation stratified sampling
  - Oversampling of subpopulations
  - Cluster sampling
  - Combinations thereof

Design Weights: Simple Random Sampling

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: Simple Random Sampling

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?
- $p = \frac{1000}{100{,}000} = .01$

Design Weights: Simple Random Sampling

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?
- $p = \frac{1000}{100{,}000} = .01$
- Design weight for all units is $w = 1/p = 100$
- SRS is *self-weighting*

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Danes & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Danes & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Danish} = \frac{900}{90{,}000} = .01$
  - $p_{Imm} = \frac{100}{10{,}000} = .01$

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Danes & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Danish} = \frac{900}{90{,}000} = .01$
  - $p_{Imm} = \frac{100}{10{,}000} = .01$
- Design weight for all units is $w = 1/p = 100$
- Proportionate allocation is *self-weighting*

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Danes & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Danes & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Danish} = \frac{500}{90,000} = .0056$
  - $p_{Imm} = \frac{500}{10,000} = .05$

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Danes & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Danes & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Danish} = \frac{500}{90{,}000} = .0056$
  - $p_{Imm} = \frac{500}{10{,}000} = .05$
- Design weights differ across units:
  - $w_{Danish} = 1/p_{Danish} = 178.57$
  - $w_{Imm} = 1/p_{Imm} = 20$
- Disproportionate allocation is not *self-weighting*

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$
- Design weights differ across units:
  - Clusters are equally likely to be sampled
  - Probability of selection within cluster varies with cluster size
- Cluster sampling is rarely *self-weighting*

Nonresponse Weights

- Correct for nonresponse
- Require knowledge of nonrespondents on variables that have been measured for respondents
- Requires data are *missing at random*
- Two common methods
  - Weighting classes
  - Propensity score subclassification

Nonresponse Weights: Example

- Imagine immigrants end up being less likely to respond[1]
  - $RR_{Danish} = 1.0$
  - $RR_{Imm} = 0.8$

[1] This refers to a lower RR in this particular survey sample, not in general.

Nonresponse Weights: Example

- Imagine immigrants end up being less likely to respond[1]
  - $RR_{Danish} = 1.0$
  - $RR_{Imm} = 0.8$

- Using weighting classes:
  - $w_{rr,Danish} = 1/1 = 1$
  - $w_{rr,Imm} = 1/0.8 = 1.25$

- Can generalize to multiple variables and strata

[1] This refers to a lower RR in this particular survey sample, not in general.

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
  - Divide sample and population into strata
  - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
  - Divide sample and population into strata
  - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does
- There are numerous other related techniques

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5		
Danish, Male	.45	.4		
Immigrant, Female	.05	.07		
Immigrant, Male	.05	.03		

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5	Over	
Danish, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5	Over	0.900
Danish, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5	Over	0.900
Danish, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5	Over	0.900
Danish, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Danish, Female	.45	.5	Over	0.900
Danish, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	1.667

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification

- Should only be done after correcting for sampling design
- Strata must be large ($n > 15$)
- Need accurate population-level stratum sizes
- Only useful if stratifying variables are related to key constructs of interest

Post-Stratification

- Should only be done after correcting for sampling design
- Strata must be large ($n > 15$)
- Need accurate population-level stratum sizes
- Only useful if stratifying variables are related to key constructs of interest
- This is the basis for inference in non-probability samples
  - Probability samples make design-based inferences
  - Non-probability samples post-stratify to obtain

Questions about weighting?

Sources of Missing Data

- Unit or item nonresponse
- Attrition or break-off
- Data loss

Effects of Missing Data

- Sampling variance and effective sample size

Effects of Missing Data

- Sampling variance and effective sample size
- Scale construction and multi-variate analysis

Effects of Missing Data

- Sampling variance and effective sample size
- Scale construction and multi-variate analysis
- Bias in estimates

Imputation

- Definition

Imputation

- Definition: Systematic replacement of missing values

Imputation

- Definition: Systematic replacement of missing values

- Why?
  - Casewise deletion creates loss of information
  - Preserve sampling variances (i.e., no loss of precision)

Imputation

- Definition: Systematic replacement of missing values
- Why?
  - Casewise deletion creates loss of information
  - Preserve sampling variances (i.e., no loss of precision)
- Considerations
  - Why are data missing?
  - How do we impute?
  - What are the consequences of imputation?

Missing Data Assumptions

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Imputation Methods

- Single Imputation

Imputation Methods

- Single Imputation
  - Mean imputation
  - Top/bottom category imputation
  - Random imputation
  - Hot deck imputation
  - Regression imputation

Imputation Methods

- Single Imputation
  - Mean imputation
  - Top/bottom category imputation
  - Random imputation
  - Hot deck imputation
  - Regression imputation

- Multiple Imputation

Imputation Methods

- Single Imputation
  - Mean imputation
  - Top/bottom category imputation
  - Random imputation
  - Hot deck imputation
  - Regression imputation
- Multiple Imputation
  - Single imputation multiple times, combining results across data sets
  - Can apply numerous imputation methods
  - Accounts for uncertainty due to missingness

Coding

- What is coding?
    - Categorizing responses
    - Assigning numeric values to categories

Coding

- What is coding?
  - Categorizing responses
  - Assigning numeric values to categories

- When in the data collection process do we code?
  - In the field
  - After data collection

Coding

- What is coding?
  - Categorizing responses
  - Assigning numeric values to categories

- When in the data collection process do we code?
  - In the field
  - After data collection

- How do we code?
  - Create set of *exhaustive*, *mutually exclusive* categories
  - Assign responses to categories
  - Add new categories, as needed

Practice Coding

1. Code the Gordon Brown responses as:
   - Correct
   - Incorrect
   - “Don’t know”

2. Code the MIP responses into issue categories

Data Editing

- Leftover of manual data recording
- Software handles most data editing now
  - Online survey tools (e.g., Qualtrics)
  - CATI systems
- May still have problematic data points that need to be marked or changed
  - If still in field, may clarify answers with respondents

Anonymizing Data

- Why do data need to be anonymous?

Anonymizing Data

- Why do data need to be anonymous?
  - Guarantees of anonymity
  - Sensitive data

Anonymizing Data

- Why do data need to be anonymous?
  - Guarantees of anonymity
  - Sensitive data

- When are data non-anonymous?

Anonymizing Data

- Why do data need to be anonymous?
  - Guarantees of anonymity
  - Sensitive data
- When are data non-anonymous?
  - Identifying information
  - Statistical identifiability

Anonymizing Data

- Why do data need to be anonymous?
  - Guarantees of anonymity
  - Sensitive data
- When are data non-anonymous?
  - Identifying information
  - Statistical identifiability
- How do we anonymize?

Anonymizing Data

- Why do data need to be anonymous?
  - Guarantees of anonymity
  - Sensitive data
- When are data non-anonymous?
  - Identifying information
  - Statistical identifiability
- How do we anonymize?
  - Restrict data access
  - Remove identifying variables

Data Storage, Archiving, and Sharing

- In what formats can we store survey data?

Data Storage, Archiving, and Sharing

- In what formats can we store survey data?
  - Paper
  - Punchcards
  - Digitally

Data Storage, Archiving, and Sharing

- In what formats can we store survey data?
  - Paper
  - Punchcards
  - Digitally
- Considerations in digital formats
  - *Open* versus *proprietary*
  - *Human-readable* versus *machine-readable*
  - File sizes
  - Study-level metadata
  - Question-level metadata

Study-level Metadata

- Title
- Creator/Author
- Sponsor
- Description
- Date of publication
- Dates of data collection
- Population, sampling frame, etc.
- Sampling design
- Sample size
- Recruitment details
- Mode
- Rights

Question-level Metadata

- Response codes
- Response labels
- Variable labels
- Variable names
- Missing data categories
- Variable types
- Mode

Question-level Metadata II

- Details of randomization or question order
- Exclusion criteria
- Source of data (if not from R)
- Frequencies or summary statistics
- Interviewer instructions
- Constraints on responses

Example Codebook[2]

```
Question B 10 DK
Which party did you vote for in that election? (Denmark)
Variable name and label: prtvtcdk Party voted for in last national
election, Denmark
Values and categories
01 Socialdemokraterne - the Danish social democtrats
02 Det Radikale Venstre - Danish Social-Liberal Party
03 Det Konservative Folkeparti - Conservative
04 SF Socialistisk Folkeparti - Socialist People's Party
05 Dansk Folkeparti - Danish peoples party
06 Kristendemokraterne - Christian democtrats
07 Venstre, Danmarks Liberale Parti - Venstre
08 Liberal Alliance - Liberal Alliance
09 Enhedslisten - Unity List - The Red-Green Alliance
10 Andet - other
66 Not applicable
77 Refusal
88 Don't know
99 No answer
Filter: If code 1 at B9
```

[2] From: http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_appendix_a7_e02_0.pdf

File Formats

- Go to the course website
- Open data files under Week 11
- All contain the same data
- What do you notice about the different files?

Metadata Standards

- Most survey data are stored in proprietary formats using codebooks constructed in arbitrary formats
  - This makes it hard to work with survey data
- There are common standards for metadata
  - Dublin Core (DC)
  - Data Documentation Initiative (DDI)

For Your Project

- Discuss appropriate file format for data storage/sharing
- Discuss how data can be used after collection (i.e., rights)
- Discuss codebook creation
  - When do you create a codebook
  - What goes in your codebook
  - Where do you record study-level metadata

Questions about handling survey data?

Total Survey Error

- Design-Related Errors
  - Coverage Error
  - Sampling Error
  - Nonresponse Error
  - Adjustment Error

Total Survey Error

- Design-Related Errors
  - Coverage Error
  - Sampling Error
  - Nonresponse Error
  - Adjustment Error
- Measurement Errors
  - Construct Validity
  - Measurement Error and Response Biases
  - Processing Error

Total Survey Error

- Design-Related Errors
  - Coverage Error
  - Sampling Error
  - Nonresponse Error
  - Adjustment Error

- Measurement Errors
  - Construct Validity
  - Measurement Error and Response Biases
  - Processing Error

- Our goal: Minimize *total* error (thus maximizing data quality), within the constraints of time, cost, and other resources

Agenda for next two classes

- Presentations
- Prepare questions to get help with
- Email me if you want to meet (after Dec. 4)