What is secondary data?

- Data collected by a person or organization other than the users of the data

Advantages of Secondary Data

- Unobtrusive
- Fast & inexpensive
- Avoid data collection problems
- Provide bases for comparison
Disadvantages of Secondary Data

- Data availability
- Level of observation
- Quality of documentation
- Data quality control
- Outdated data

Data Sources

- Inter-university Consortium for Political & Social Research (ICPSR)
  http://www.icpsr.umich.edu/
  (data archivist in ATLAS, room 212 Lincoln Hall)
- National Center for Health Statistics (NCHS)
  http://www.cdc.gov/nchs/default.htm
- Centers for Medicare & Medicaid Services (CMS)
  Formerly HCFA
  http://www.cms.hhs.gov
- US Census Bureau
  http://www.census.gov

Data Sources (cont’d)

Examples of Directly Downloadable Data from NCHS:
- National Health and Nutrition Examination Survey (NHANES)
- National Ambulatory Medical Care Survey (NAMCS)
- National Hospital Ambulatory Medical Care Survey (NHAMCS)
- National Hospital Discharge Survey (NHDS)
- National Home and Hospice Care Survey (NHHCS)
- National Nursing Home Survey (NNHS)
- National Survey of Ambulatory Surgery (NSAS)
- National Employer Health Insurance Survey (NEHIS)
- National Vital Statistics System (NVSS)
- National Health Interview Survey (NHIS)
Survey Documentation & Analysis (SDA)

Web-based analysis and documentation

- http://sda.berkeley.edu/
- http://www.icpsr.umich.edu/access/sda.html
- http://www.icpsr.umich.edu/NACJD/das.html
- http://www.icpsr.umich.edu/SAMHDA/

Data Available for Use with SDA

Aging Data
- Longitudinal Study of Aging, 70 Years & Older, 1984-1990
- National Survey of Self-Care & Aging: Follow-Up, 1994
- National Health & Nutrition Examination Survey II: Mortality Study, 1992
- National Hospital Discharge Survey, 1994-1997
- National Health Interview Survey, 1994, 2nd Supplement on Aging

Criminal Justice Data
- International Crime Data
- Homicide Data
- National Crime Victimization Survey Data
- Corrections Data

Data Available for Use with SDA (continued)

Substance Abuse Data
- Drug Abuse Warning Network
- Monitoring the Future
- National Household Survey on Drug Abuse
- National Pregnancy & Health Survey
- National Treatment Improvement Evaluation Study
- Treatment Episode Data Set
- Uniform Facility Data Set
- Washington, DC Metropolitan Area Drug Study (DC*MADS)
Evaluation of Data Sources

- Purpose of the study
- Sponsor/collector of the data
- Mode of data collection
- Sampling procedures
- Consistency of data with other sources

Evaluation of Data Sources (cont'd)

- Documentation
- Number of observations
- Number of variables
- Coding scheme
- Summary statistics

Types of Survey Sample Design

- Simple random sampling
- Systematic sampling
- Complex sample designs
  - Stratified designs
  - Cluster designs
  - Mixed-mode designs
Why complex survey design?

• Increased efficiency
• Decreased costs

Complex Survey Design

• Complex designs with clustering & unequal selection probabilities generally increase the sampling variance.
• Not accounting for the impact of complex sample design can lead to Type I error (rejecting $H_0$ when it is true).

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

Selection Weights
• inverse of probability of selection
• in simple random sample, selection weights are the same for all respondents

Types of Sample Weights (cont’d)

Post-stratification weights
• Typically used to adjust for minor differences in nonresponse by demographic subgroup
• Bring sample proportions in these subgroups into agreement with the population proportion in the subgroups.
• Requires auxiliary data set to use as a comparison.

Post-Stratification Weights Example

<table>
<thead>
<tr>
<th></th>
<th>Sample Percent</th>
<th>Population Percent</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>42%</td>
<td>49%</td>
<td>1.16</td>
</tr>
<tr>
<td>Female</td>
<td>58%</td>
<td>51%</td>
<td>0.879</td>
</tr>
</tbody>
</table>
Types of Sample Weights (cont'd)

Nonresponse weights:
• designed to inflate the weights of survey respondents to compensate for nonrespondents with similar characteristics.
• Only useful if nonresponse varies by stratum (unless inflating sample size to population size).

Types of Sample Weights (cont'd)

• “Blow-up” (expansion) weights:
  ▪ weights sum to population totals
  ▪ provide estimates for the total population of interest

Summary of Weights

• Weight for probability of selection
• Adjust for non-response
• Post-stratify
• Expand or contract to population/sample totals
Analysis of Complex Sample Designs

• Weighting data necessary but not sufficient
• Weighting allows for accurate point estimates (means and totals)
• Weighting by itself is insufficient to adjust for effect of complex sample design on sampling variances (or its square root, the standard error)

Impact of Design on Standard Errors

What is a standard error?
• Standard deviation of sampling distribution
• Variation in parameter estimates resulting from repeated samples being drawn from same population.
• Because we generally draw only one sample, it has to be estimated from that sample.
• In its most basic form is equal to the sample standard deviation divided by the square root of the sample size.

Why do we care?
• As designs become more complex, calculating standard errors become more complex.
• Standard error for most complex designs is larger than that of a simple random sample.
• The ratio of the two is called the design effect.
• Confidence intervals and significance tests are based on standard errors.
• Underestimating standard errors results in over-stating significance.
Impact of Design on Standard Errors (cont'd)

Why do we care (cont’d)?

- Standard statistical software packages often assume simple random sampling.
- Significance testing is therefore incorrect.
- Need to analyze data with a software that can adjust for complex design.

Syntax Examples of Design-Based Analysis in STATA & SUDAAN

**STATA**

```
svyset strata
svyset 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;
```

Syntax Example of Design-Based Analysis in SAS

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

• Several software packages can analyze data from complex designs
• These packages vary with respect to:
  ▪ procedures available
  ▪ how they estimate variances
• For a good summary see:
  http://www.hcp.med.harvard.edu/statistics/survey-soft/