Data Management

Data entry and data cleaning
Building codebooks
Data documentation, data archiving, and replication

Data Management

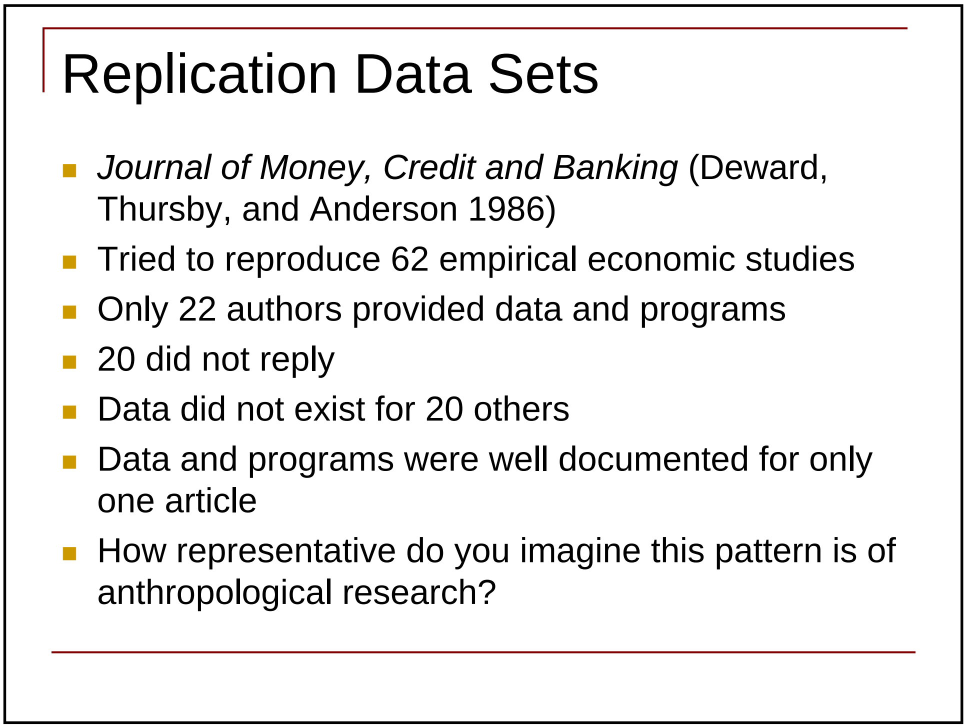
- First step in data analysis is getting raw data into usable form
- Initial tasks
  - Creating dataset
  - Editing to correct errors
  - Adding internal documentation (e.g., variable and value labels)
- Ongoing tasks
  - Adding new observations or variables
  - Reorganizing datasets
  - Separating, combining, collapsing datasets
  - Converting variable types
  - Creating new variables

Data Management in Stata

- Input data: `input, edit, insheet, infile, infix`
- View or summarize data: `describe, list, summarize, by, edit, browse`
- Specify subsets of data: `in, if`
- Create and change variables: `generate, replace, recode, egen`
- Internal documentation: `label variable, label define, label values, label data, codebook`
- Convert between string and numeric: `encode, decode, destring`
- Create new nominal and ordinal variables: `tabulate` *`catvar`*`, gen(`*`catvar`*`); autocode`
- Combining Stata files: `append, merge`
- Reorganizing datasets: `xpose, reshape`

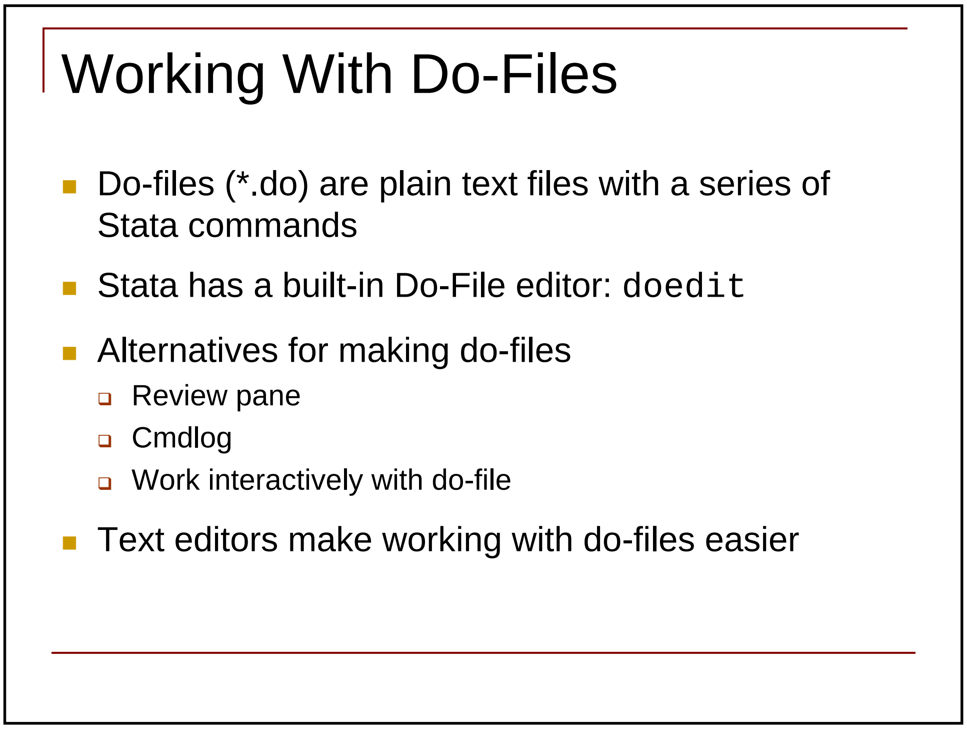
Data Management, Data Documentation, and Replication

- Key criterion for science's claim to objectivity is intersubjective traceability (Popper)
- Other researchers should be able to confirm results or object to them, based on details of data collection and analysis
- Requires diligent documentation of decisions in every step of research
- Statistical data analysis is perhaps easiest aspect of research to document, but it seldom happens



Replication Data Sets

- *Journal of Money, Credit and Banking* (Deward, Thursby, and Anderson 1986)
- Tried to reproduce 62 empirical economic studies
- Only 22 authors provided data and programs
- 20 did not reply
- Data did not exist for 20 others
- Data and programs were well documented for only one article
- How representative do you imagine this pattern is of anthropological research?



Working With Do-Files

- Do-files (*.do) are plain text files with a series of Stata commands
- Stata has a built-in Do-File editor: `doedit`
- Alternatives for making do-files
  - Review pane
  - Cmdlog
  - Work interactively with do-file
- Text editors make working with do-files easier

Designing Do-Files

- Comments make it easier to understand do-files

  -
    ```
    * Data: Puerto Rico survey, 2000-2001
    use prsurvey, clear
    ```
- Two ways to continue commands over more than one line

  -
    ```
    #delimit ;
    label define empl
       1  “Full time”
       2  “Part time”
       3  “Unemployed
    ```

  -
    ```
    label define empl 1 “Full time” 2 “Part time”
    ///
       3 “Unemployed”
    ```
- Use `exit` at end of do-file to include description or notes below

Organizing Your Work

- Organize all files by project, not document type
- Develop a consistent naming system
  - Juul suggests `gen.dataset.do, an.result.do`
  - Others suggest using only 8-character names, beginning with `cr` or `an` to distinguish
- Maintain a log to keep track of links between datasets, do-files, and results