CRCF Workshop Series 2017: Data Organization and Management

Biru Zhou Ph.D.

September 25, 2017
1 Unofficial Title

2 Disclaimers

3 New Data

4 A Plan
   - Two Scenarios
   - Current Data
   - Future Data
   - Special Cases
   - Organization Softwares
   - Data Management Checklist

5 Basic Data Cleaning
Things I Wish I Did Not Do To My Data
Disclaimers

- Nothing is set in stone.
- There is more than one way to manage research data.
- I have my own biases in terms of data management.
Calm Down Please

OK! JUST CALM DOWN
Data are means to an end. Data contain numbers but they have meanings in social sciences. We need a plan! Seriously!
Data are means to an end
Data are means to an end

Data contain numbers but they have meanings in social sciences
- Data are means to an end
- Data contain numbers but they have meanings in social sciences
- We need a plan! Seriously!
Current Data – Data already collected
A Plan

Two Scenarios

1. Current Data – Data already collected
2. Future Data – Data not yet collected
First Thing First:

Always Ask: Is this the only copy of the dataset?

1. **YES** → Please make a backup copy and save it elsewhere.

2. **NO** → Who is keeping a copy? Is the copy identical to the version I have?
First Thing First:

Always Ask: Is this the only copy of the dataset?
First Thing First:

Always Ask: Is this the only copy of the dataset?

YES → Please make a backup copy and save it elsewhere.
First Thing First:

Always Ask: Is this the only copy of the dataset?

1. YES → Please make a backup copy and save it elsewhere.
2. NO → Who is keeping a copy? Is the copy identical to the version I have?
Second Thing To Ask:

Who is the data manager?
Second Thing To Ask:

Who is the data manager?
Who is the Go–To person for this project?
General Data Management

Do

1. Keep one original copy
2. Create a working file
3. Modify the working file only
4. Always save your data
5. Keep logs of modifications
6. Keep dates of modifications
7. Transfer data file with encryption
8. Use end-to-end encryption whenever possible

Don't

1. Never calculate scores using the original copy
2. Never leave data file open unattended
3. Never use public computer for data modification
4. Never use public free Wi-Fi to transfer data file
5. Never use unknown/unsecured server for data file transfer
General Data Management

**Do**
1. Keep one original copy
2. Create a working file
3. Modify the working file only
4. Always save your data
5. Keep logs of modifications
6. Keep dates of modifications
7. Transfer data file with encryption
8. Use end-to-end encryption whenever possible

**Don’t**
1. Never calculate scores using the original copy
2. Never leave data file open unattended
3. Never use public computer for data modification
4. Never use public free Wi-Fi to transfer data file
5. Never use unknown/uns cured server for data file transfer
General Data File Management

- Know how to name your data file – **Avoid space, dash, symbols or dots in your file name**
  1. Be concise and consistent
  2. Put date in file names – YYYYMMDD
  3. Preferably less than 32 characters
  4. Avoid using generic naming system, e.g., “BiruData.sav” or “MyProject.sav”
  5. Can be understood by yourself and **others** 6 months, 1 year, 2 years later
  6. Example: CIS2008v2_20101105.sav
General Data File Management

- Know where to save your data – **Not on your desktop, Not in your personal folder**
- Set up a structure for folder organization
- Know where to backup your data – **Not in the Cloud, personal OneDrive or Dropbox**
General Data File Management

- Know how often you need to backup – Make a schedule and stick to it.

- Once data collection or project is concluded, archive all your backup data files.

- Only use one version of data as your working file for future analysis.
Be proactive and assume leadership if you know something can be done to improve data management procedures
User Friendly Codebook

- Codebook is a reference for researchers
User Friendly Codebook

- Codebook is a reference for researchers
- It can help newcomers to understand the project
User Friendly Codebook

- Codebook is a reference for researchers
- It can help newcomers to understand the project
- It can be used by statisticians for analyses
User Friendly Codebook

- Codebook is a reference for researchers
- It can help newcomers to understand the project
- It can be used by statisticians for analyses
- It can help for data sharing and dissemination
User Friendly Codebook

- Codebook is a reference for researchers
- It can help newcomers to understand the project
- It can be used by statisticians for analyses
- It can help for data sharing and dissemination
- Codebook format should be the same as the data format (Excel, CSV, SPSS, SAS, Word), preferably
User Friendly Codebook

- Codebook is a reference for researchers
- It can help newcomers to understand the project
- It can be used by statisticians for analyses
- It can help for data sharing and dissemination
- Codebook format should be the same as the data format (Excel, CSV, SPSS, SAS, Word), preferably
- Codebook should be customized to meet the needs of the project
User Friendly Codebook

Basic Structure:

- Brief description of the project
  1. What is this project about
  2. Who did the project – PIs and RAs
  3. When and how was the data collected – Time span for data collection
User Friendly Codebook

Basic Structure:

- Brief description of the project
  1. What is this project about
  2. Who did the project – PIs and RAs
  3. When and how was the data collected – Time span for data collection

- Sampling information
  1. Population studied
  2. What were the sampling/recruitment strategies
  3. Response rate if available
User Friendly Codebook

Basic Structure:

- Brief description of the project
  1. What is this project about
  2. Who did the project – PIs and RAs
  3. When and how was the data collected – Time span for data collection

- Sampling information
  1. Population studied
  2. What were the sampling/recruitment strategies
  3. Response rate if available

- Ns – Number of participants, number of observations, number of observations per participant, etc.
User Friendly Codebook

Basic Structure con’t:

- Details about the data
  - Structure of the data in the file – Long vs. Wide
User Friendly Codebook

Basic Structure con’t:

- Details about the data
  - Structure of the data in the file – Long vs. Wide
  - Variable names and labels – Keep them understandable, consistent, and simple
User Friendly Codebook

Basic Structure con’t:

- Details about the data

  1. Structure of the data in the file – Long vs. Wide
  2. Variable names and labels – Keep them understandable, consistent, and simple
  3. Consistent Variable codes (1 = Yes, 0 = No) – Include recoded variables, missing value codes
User Friendly Codebook

Basic Structure con’t:

- Details about the data
  1. Structure of the data in the file – Long vs. Wide
  2. Variable names and labels – Keep them understandable, consistent, and simple
  3. Consistent Variable codes (1 = Yes, 0 = No) – Include recoded variables, missing value codes
  4. Reliability for scales – Cronbach’s $\alpha$ (based on groups if applicable)
User Friendly Codebook

Basic Structure con’t:

- Missing data
  - What types of missing – refuse to respond, conditional non-response, not valid, etc.
User Friendly Codebook

Basic Structure con’t:

- **Missing data**
  - What types of missing – refuse to respond, conditional non-response, not valid, etc.
  - How are missing data coded – 9999, 8888, 7777
User Friendly Codebook

Basic Structure con’t:

- Missing data
  1. What types of missing – refuse to respond, conditional non-response, not valid, etc.
  2. How are missing data coded – 9999, 8888, 7777
  3. How many participants have missing values

- Percentage of total missing = Total number of missing values / (Total number of items * Total number of observations)
- Percentage of participants with missing values = Total number of participants with missing values / Total number of participants
User Friendly Codebook

Basic Structure con’t:

- Missing data
  - What types of missing – refuse to respond, conditional non-response, not valid, etc.
  - How are missing data coded – 9999, 8888, 7777
  - How many participants have missing values
  - How many participants have missing values on each item or on each scale
User Friendly Codebook

Basic Structure con’t:

- **Missing data**
  1. What types of missing – refuse to respond, conditional non-response, not valid, etc.
  2. How are missing data coded – 9999, 8888, 7777
  3. How many participants have missing values
  4. How many participants have missing values on each item or on each scale
  5. Percentage of total missing = Total number of missing values / (Total number of items * Total number of observations)
User Friendly Codebook

Basic Structure con’t:

- **Missing data**
  1. What types of missing – refuse to respond, conditional non-response, not valid, etc.
  2. How are missing data coded – 9999, 8888, 7777
  3. How many participants have missing values
  4. How many participants have missing values on each item or on each scale
  5. Percentage of total missing = Total number of missing values / (Total number of items * Total number of observations)
  6. Percentage of participants with missing values = Total number of participants with missing values / Total number of participants
ONE DOES NOT SIMPLY START A PROJECT

WITHOUT A DATA MANAGEMENT PLAN
First Thing First:

First Question to Ask: Who is the data manager?

Second Question: If data manager is not assigned yet, do we need one for this project? Or Who will be the data related Go – To person?

P.S. Project manager = Data manager
Second thing:

What does a data manager do actually?

- who drafts up the data management plan
Second thing:

What does a data manager do actually?

- who drafts up the data management plan
- who enforces data management plans and rules
Second thing:

What does a data manager do actually?

- who drafts up the data management plan
- who enforces data management plans and rules
- who knows about plans for future analysis and can facilitate data organization to fit the needs of future analyses
Second thing:

What does a data manager do actually?

- who drafts up the data management plan
- who enforces data management plans and rules
- who knows about plans for future analysis and can facilitate data organization to fit the needs of future analyses
- who oversees and monitors all data related issues on recruitment, data collection, data cleaning, data verification, data backup, etc.
Second thing:

What does a data manager do actually?

- who drafts up the data management plan
- who enforces data management plans and rules
- who knows about plans for future analysis and can facilitate data organization to fit the needs of future analyses
- who oversees and monitors all data related issues on recruitment, data collection, data cleaning, data verification, data backup, etc.
- who is in charge of the codebook
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
- Use PDF whenever possible
- If study is conducted in multiple languages, avoid using abbreviations
- Be on the good side of the RAs or graduate students of your collaborators
- Expect delays, so be liberal when planning your timeline
Multiple Collaborators

- Minimize the number of files to be transferred
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
- Use PDF whenever possible
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
- Use PDF whenever possible
- If study is conducted in multiple languages, avoid using abbreviations
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
- Use PDF whenever possible
- If study is conducted in multiple languages, avoid using abbreviations
- Be on the good side of the RAs or graduate students of your collaborators
Multiple Collaborators

- Minimize the number of files to be transferred
- Merge everything (multiple files) into one file
- Use PDF whenever possible
- If study is conducted in multiple languages, avoid using abbreviations
- Be on the good side of the RAs or graduate students of your collaborators
- Expect delays, so be liberal when planning your timeline
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
- Have multiple identifiers: person ID, location ID, institution ID or unit ID, etc.
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
- Have multiple identifiers: person ID, location ID, institution ID or unit ID, etc.
- Multiple datasets v.s. One single dataset
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
- Have multiple identifiers: person ID, location ID, institution ID or unit ID, etc.
- Multiple datasets v.s. One single dataset
- Project manager and/or data manager is a must-have in Longitudinal projects
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
- Have multiple identifiers: person ID, location ID, institution ID or unit ID, etc.
- Multiple datasets v.s. One single dataset
- Project manager and/or data manager is a **must-have** in Longitudinal projects
- Hire your project manager and/or data manager wisely, *not* just for the duration of data collection, but a bit longer, at least until the first manuscript or report is ready
Longitudinal Data Management

- Have a clear data collection schedule and keep track of the discrepancies
- Have multiple identifiers: person ID, location ID, institution ID or unit ID, etc.
- Multiple datasets v.s. One single dataset
- Project manager and/or data manager is a must-have in Longitudinal projects
- Hire your project manager and/or data manager wisely, *not* just for the duration of data collection, but a bit longer, at least until the first manuscript or report is ready
- Two data structures for data entry: Person-Level v.s. Person-Period
### Longitudinal Data Management cont’

- Person-Level data structure aka **Wide Data Structure**

<table>
<thead>
<tr>
<th>ID</th>
<th>Location</th>
<th>Variable1T1</th>
<th>Variable2T1</th>
<th>Variable1T2</th>
<th>Variable2T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam</td>
<td>18</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Joe</td>
<td>20</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

Biru Zhou Ph.D.

CRCF 2017 Workshop

September 25, 2017
Longitudinal Data Management cont’

- Person-Level data structure aka **Wide Data Structure**
- One person = One case entry = One row of data
Longitudinal Data Management cont’

- Person-Level data structure aka **Wide Data Structure**
- One person = One case entry = One row of data
- When entering data for multiple time points, you are adding new variables horizontally

<table>
<thead>
<tr>
<th>ID</th>
<th>Location</th>
<th>Variable1T1</th>
<th>Variable2T1</th>
<th>Variable1T2</th>
<th>Variable2T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam</td>
<td>18</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Joe</td>
<td>20</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
Longitudinal Data Management cont’

- Person-period data structure aka **Long Data Structure**
Longitudinal Data Management cont’

- Person-period data structure aka **Long Data Structure**
- One person = Multiple cases entry = Multiple rows of data
Longitudinal Data Management cont’

- Person-period data structure aka **Long Data Structure**
- One person = Multiple cases entry = Multiple rows of data
- When entering data for multiple time points, you are adding new cases/rows vertically

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>Location</th>
<th>Eye Color</th>
<th>Variable1</th>
<th>Variable2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam</td>
<td>1</td>
<td>18</td>
<td>Brown</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Sam</td>
<td>2</td>
<td>18</td>
<td>Brown</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Joe</td>
<td>1</td>
<td>20</td>
<td>Blue</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Joe</td>
<td>2</td>
<td>20</td>
<td>Blue</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
References Organization and Management

zotero

- References organization and management add-on
- Works within Firefox or as a stand alone program
- FREE! Open Source!
- Create citations and bibliographies into text or programs
- Integrate with Microsoft Word to do automatic bibliographies
- Collect, organize, cite and share your references – Just one click away
Systematic Review Software

Rayyan

- Web application for systematic reviews
- FREE! Open Source!
- Can directly import reference libraries to the web application for reviews
- Add keywords for inclusion or exclusion criteria, keep track of your review decisions
- Can collaborate with others on the review process
- “Blind On” function to prevent your collaborators from seeing each other’s reviews
- McGill Library Guide for Rayyan
- Get to know your data
- Think and plan ahead what you and your team will do with the data
- Very basic cleaning for everyone
- Remove identification variable and keep it strictly confidential