Data Management and Descriptive Statistics

Onoja M. Akpa

Data Analysis and Manuscript Writing Workshop

University of Ibadan Medical Education Partnership Initiative Junior Faculty Research Training Programme
(UI-MEPI-J)

24-28th April, 2017

What is data management?

- All activities relating to preparation of data for analysis.
- It is worth spending sometime planning data management - this may save considerable effort at later stages
  - My experiences with postgraduate students…

Data Management Activities

- Data collection
- Data entry template development
- Data entry
- Data storage

Data Management Activities

- Data processing:
  - cleaning,
  - editing,
  - manipulation

- Data analysis
- Data presentation

Data entry template

- Develop a data entry template using your data collection tool.
- A well coded template should at least replicate your data collection tool.
- This ensures uniformity of database across study sites,
- Controls may be included in the template.

Needs For Coding Guide/Data Dictionary

- Prepare data in format to allow use of computers for statistical analysis.
- Prepare code book or data dictionary for the questionnaire.
- Specify range of values expected.

Needs For Coding Guide/Data Dictionary

- Numerical data:
  - Should be entered with the same precision as they are measured.
- Unit of measurement should be consistent for all observations on a variable.
  - e.g weight should be recorded in kg or in pounds , but not both interchangeably

Handling missing data

- You should consider what to do with missing values before you enter the data.
  - use some symbols to represent a missing value
- Statistical packages deal with missing values in different ways

Handling missing data

- Some use special characters (e.g. a full stop or asterisk) to indicate missing values,
  - whereas others require you to define your own code for a missing value (commonly used values are 9, 99 or 999)
- The value that is chosen should be one that is not possible for that variable

Handling missing data

- When entering a categorical variable with four categories (coded 1,2,3, and 4), you may choose the value 9 to represent missing values.

- However, if the variable is age of a child, then a different code should be chosen.

Handling missing data

- If a large proportion of data is missing, then the results are likely to be unreliable
- Investigate reasons for missing: how much is missing and why?

Handling missing data

- It may be that the data is simply sitting on a piece of paper in someone's drawer and are yet to be entered!

Handling missing data

- What type of missingness:
- **Missing completely at random (MCAR)**
  - The missing data are just a random subset of the data
  - There's no relationship between whether a data point is missing and any value in the data set, missing or observed

Handling missing data

- What type of missingness:
- **Missing at Random (MAR)**
  - The propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

Data Quality Control

- Record verification (double entry)
  - Does not rule out the possibility that the same error has been incorrectly entered on the two occasions
  - A disadvantage of this approach is that it takes twice as long to enter the data, which may have major cost or time implications

Data Quality Control

- Random checking can be done at random but should represent all forms being entered
  - Selection should be systematic

Checking for errors

- In categorical data:
  - relatively easy, values not allowable must be errors
  - Check frequency distribution of each variable e.g Sex: 1=male, 2=female, 3?

Checking for errors

- In numerical data:
  - Produce descriptive statistics for all variables.
  - Standard deviation higher than mean check for an outlier observation
  - range checks, upper and lower limits can be specified for each variable

Checking for errors

- Dates:
  - For example 30th feb. must be incorrect,
  - Any day of the month greater than 31, may be an error
  - Any month greater than 12 may be an error etc

Checking for errors

- Apply logical checks:
  - date of birth should correspond to patient's age,
  - patients who have died should not appear on subsequent follow up visits,
  - there should be no pregnant men

Checking for errors

- With all error checks, a value should only be corrected if there is evidence that a mistake has been made
- You should not change values simply because they look unusual; investigate

How to deal with outliers

- Outliers are observations that are distinct from the main body of data, and are incompatible with the rest of the data
- They may be genuine value, data collection errors or entry/typing errors

How to deal with outliers

- Any suspicious values should be checked
- Value should only be changed if there is evidence that it is incorrect

What to do with genuine outliers

- Repeat analysis with and without the value
- Results could be similar or change drastically
- If similar, then outlier does not have great influence on the result
- If it changes drastically, use appropriate methods not affected by outliers to analyze data.
  - These include use of transformations and non parametric tests

Data management software

- Use of computer software allows:
  - Storage of large quantities of data
  - Ease of checking and correcting errors
  - Ease of tabulation and presentation of results
  - Quick statistical analysis

Data management software

- Choice may be influenced by
  - Organisational preference
  - User-friendliness
  - Portability
  - Ease of data conversion
  - Availability
  - System requirement etc.

Data management software

- Common examples
  - *Epi Info*
  - *Epi Data*
  - *SPSS*
  - *Stata**
  - *SAS**
  - *Ms Excel; Access etc*

NOTE:

- Some statistical packages have problems dealing with non numerical data
- You may need to assign numerical codes to categorical data before entering data

NOTE:

- For example, you may choose to assign codes of 1, 2, 3 and 4 to categories of
  - "no pain",
  - 'mild pain",
  - "moderate pain" and
  - "severe pain" respectively

NOTE:

- These codes can be put in the questionnaire when collecting the data.
- For binary data e.g yes/no answers, it is often convenient to assign codes 1 (e.g for yes) and 0 or 2 (for no).
- Need to prepare a coding guide/data dictionary