Preparing for Data Analysis

Prof. Andrew Stokes

March 26, 2019

Managing your data

- Entering the data into a database
- Reading the data into a statistical computing package
- Checking the data for errors and inconsistencies
- Data cleaning
- Preparing the data for analysis

Cautionary Note

- Never alter the raw data directly!
- All manipulations should be performed in the statistical computing software

Data checks

- Range checks to identify out-of-range values (e.g. age)
- Cross-checks to identify inconsistencies between values (e.g. males that are pregnant)
- Cross-tabulations

Data checks continued

- Scatter plots and box-plots to compare groups and identify outliers
- Proportion of responses missing, “Other” and “Don’t Know”
- Consistency across questions that elicit similar information

Note: These preliminary checks can be run before you have finished collecting your data

Data cleaning

- The goal of data cleaning is to resolve problems that were identified during data checking process
- The aim is make the data as high quality as possible for analysis
- If a problem cannot be resolved, the incorrect data can be assigned a missing code.
- *Remember: never alter the raw data file. All changes should be made in the R script*

Variable naming and coding conventions

- Variable names should be descriptive (e.g. 'birthwt' for birth weight)
- Questions that have categories of answers should be assigned numeric values in a systematic fashion
- For example, yes/no questions should be coded assigned values 0/1

Routinely backup your data and code

- Use systematic approach to naming backup files (e.g. with the date of backup included in the filename)
- In addition to backing up your data, you should routinely create backups of your R scripts

Preparing data for analysis

- The goal of this step is to create a final analytic dataset
- The raw data as entered into the database is usually not sufficient
- You'll both be creating new variables and recoding existing ones
- In some cases, you may also need to merge multiple sources of data

Organizing your code

- You can have one script both for data pre-processing and analysis
- Or you can have two or more scripts (recommended)
- The first script can be used to generate the final analytic data set
- The second script can then be used to implement analyses

Data dictionary

The data dictionary provides a map between the questionnaire and the data files. It is a record of how the data are structurued and will be useful resource as you prepare for analysis. It should contain the following information:

- Name and description of each variable
- Data type (e.g. numeric or text; if numeric, continuous, binary or categorical)
- Coding (e.g. 0=No, 1=Yes)
- Question number to which the variable relates

Creating new variables

- Calculated variables may combine information from two or more individual variables
- Body mass index is calculated using weight in kilograms over height in meters squared.
- When you compute such a variable you may have to first translate the raw variables into the correct units
- Some variables may use external data

Checking your composite variables

- After calculating composite variables, check for validity of the responses (e.g. plot the data)
- For example, seemingly reliable weight and height data may produce unrealistic BMI values
- Sometimes data errors may only appear upon checking the range of calculated variables

Coding & Re-coding

Reasons to consider categorizing data:

- Grouping values is a form of simplication or "dimension reduction"
- Can help for identifying non-linear associations
- Will be necessary when some categories include too few observations to be analyzed separately

Re-coding continued

- Pooling like groups for variables that are already categorical (important principle here is that risk of outcome should be similar in each of the combined groups)
- Divide continuous data into quartiles (four groups) or quintiles (five groups) with equal numbers of observations

Re-coding continued

- Other times cutpoints will be based on established rules or guidelines (NHLBI/wHO BMI categories for normal weight, overweight and obese)
- If no standard cutpoints are available, another approach is to study a histogram of the data and choose cutpoints based on natural break

Planning your analysis

- What are you measuring?
- What comparisons do you want to make?
- What is your outcome and how should your outcome variable be constructed?
- What are your predictor variables?
- How should they be constructed?

Steps in the analysis

Two primary levels of analysis

- Descriptive analysis (summary of population)
- Multivariate analysis

Note: often simple bivariate analyses will be presented prior to multivariate ones to show associations without statistical adjustment for other variables.

Descriptive analysis (quantitative)

- Description of quantitative data is Table 1 in most papers
- Describes characteristics of population
- Why is this important?

Example from Stokes 2014

Table 1 Characteristics of US never-smoking adults ages 50-84

	No.	% or mean
Age at survey, years		64.09
Education		
Less than high school	2,461	28.35
High school or equiv.	1,395	29.02
More than high school	1,684	42.63
Race/ethnicity		
Hispanic	1,371	8.54
Non-Hispanic white	2,944	77.81
Non-Hispanic black	1,079	9.25
Non-Hispanic other	146	4.41
Obesity status at survey		
Normal	1,542	29.70
Overweight	2,171	38.13
Obese class I	1,152	20.00
Obese class II	675	12.18
Obesity status at maximum		
Normal	768	17.09
Overweight	1,991	36.36
Obese class I	1,649	27.44
Obese class II	1,132	19.11
Obesity status: maximum-survey		
Normal - normal	768	17.09
Over - normal	633	10.44
Obese 1 - normal	116	1.80
Obese 2 - normal	25	0.36
Over - over	1,358	25.92
Obese 1 - over	702	10.62
Obese 2 - over	111	1.59
Obese 1 - obese 1	831	15.02
Obese 2 - obese 1	321	4.98
Obese 2 - obese 2	675	12.18
Deceased	903	11.92
Total	5,540	

Categories of BMI are normal weight (18.5-25.0 kg/m^2); overweight (25.0-29.9 kg/m^2); obese class 1 (30.0-34.9 kg/m^2); and obese class 2 (35.0 kg/m^2 or greater). Percentages and means are calculated using sample weights. Entry years are 1988-2004 with mortality follow-up through 2006. Source: National Health and Nutrition Examination Survey.

Bivariate analysis

Comparing outcomes between groups

- t-tests for continuous outcomes
- chi-squared tests for categorical outcomes

Multivariate analysis

- Used to adjust for confounding
- Can be used to investigate effect modification and mediators
- Linear regression for continuous outcomes
- Logistic regression for dichotomous outcomes
- Other common models: ordinal logit, poisson, negative binomial, Cox proportional hazards

Upcoming deadlines (due Sun 5 PM)

- Methods Section
- Data Dictionary
- Table Shells