

Data Preparation and Exploration

25 Sep 2009
Dr. Sean Ho
CPSY501

cpsy501.seanho.com

Please download from
“Example Datasets”:
SPSSExam.sav

Outline for today

- Hermeneutics of data preparation
- Creating secondary “derived” variables
- Descriptives of a dataset, error checking
- Missing data and outliers
- Assumptions of Parametricity
- What to report:
“Informationally adequate” statistics
- Appendix: Project data sources



```
graph TD; A[Research Question] --> B[Research Design & Planning]; B --> C[Data Gathering]; C --> D[Data Preparation & Exploration]; D --> E[Analysis & Interpretation]; E --> F[ ]; D <--> E;
```

Research Question

Research Design & Planning

Data Preparation & Exploration

Data Gathering

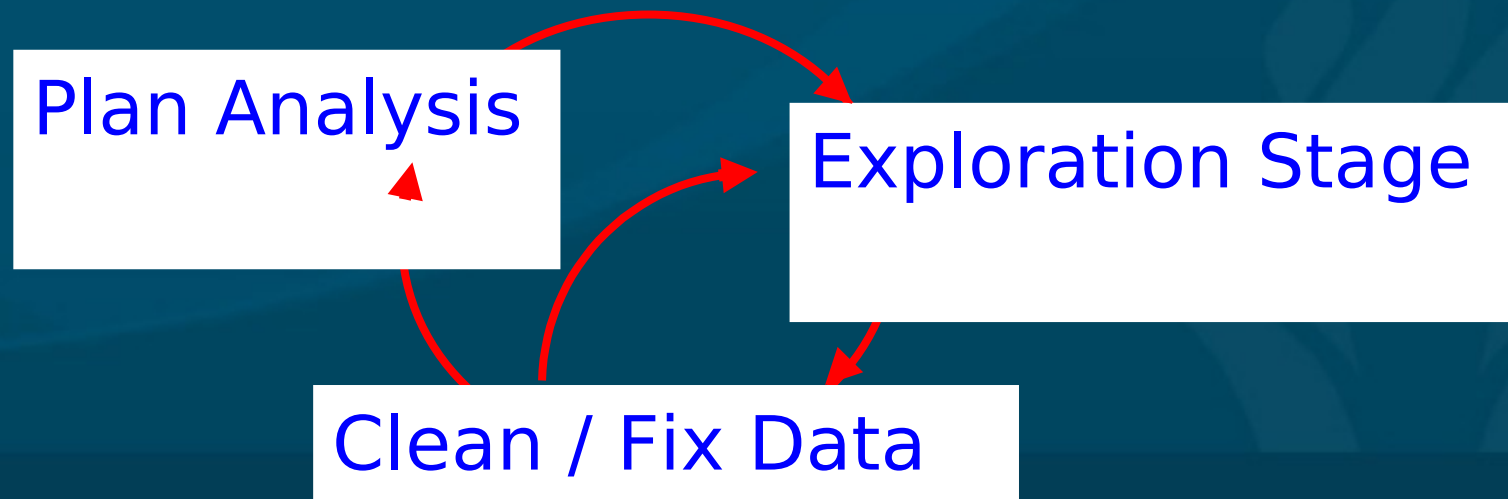
Analysis & Interpretation

Overview of Data Exploration and Preparation

- With real data, we often spend far more time **examining** and **preparing** the data than actually **running** the statistical tests!
- Failure to identify and correct problems will prevent the tests from **working** properly
 - Or worse: lead to **misleading** results that do not reflect real relationships / effects.
- Stages:
 - **Create** / **fix** variables
 - Deal with **missing** data & **outliers**
 - Assess **assumptions** of tests

Hermeneutics of Data Exploration and Preparation

- The relationship between **exploring**, **preparing** and **analyzing** data is reciprocal and hermeneutic.
- Each stage of exploration can precipitate changes to the data set, and possible re-evaluation of plans for analysis, resulting in the following **circular** process:



Outline for today

- Hermeneutics of data preparation
- Creating secondary “derived” variables
- Descriptives of a dataset, error checking
- Missing data and outliers
- Assumptions of Parametricity
- What to report:
“Informationally adequate” statistics
- Appendix: Project data sources

Derived Variables

- Not new data, but **derived** from existing vars
- **Why** generate secondary, “derived” variables?
 - Build **total** or “subscale” scores from individual test items
 - Create **grouping** vars or modify categories
 - ◆ e.g., age groups by year of birth
 - Correct **problems** in the data
 - ◆ e.g., **transform** data to make it more normally distributed
 - ◆ e.g., **missing** data

SPSS: Compute Variable

- Transform → Compute Variable
- Many kinds of **transformations** / computations from existing variables into new variables
- Can **combine** information from **multiple** old variables into a single new variable
 - e.g., **sum**, average
 - Must address **missing** scores first
- Can compute for specific **subsets**/groups within a data set using the “if” option

Application Examples

- Example dataset: SPSSExam.sav
- Combining marks:
- Create a new var: “Average assignment mark”
 - Transform → Compute Variable → Mean
- Mean vs. Sum:
 - Generally, means are easier to interpret:
does not depend on number of assignments
 - However, many standard subscales do
use sum of individual questions

Recode Options

- **Recoding** can be done into the **same** variable, or into a **new** variable
 - I generally recommend recoding into **new**:
 - Transform → Recode into different vars
 - ◆ Specify what **variable** you want to recode
 - ◆ Specify **old** and **new** values
- We can recode specific **parts** of a variable using the “if” option

Example: Coding for Ethnicity

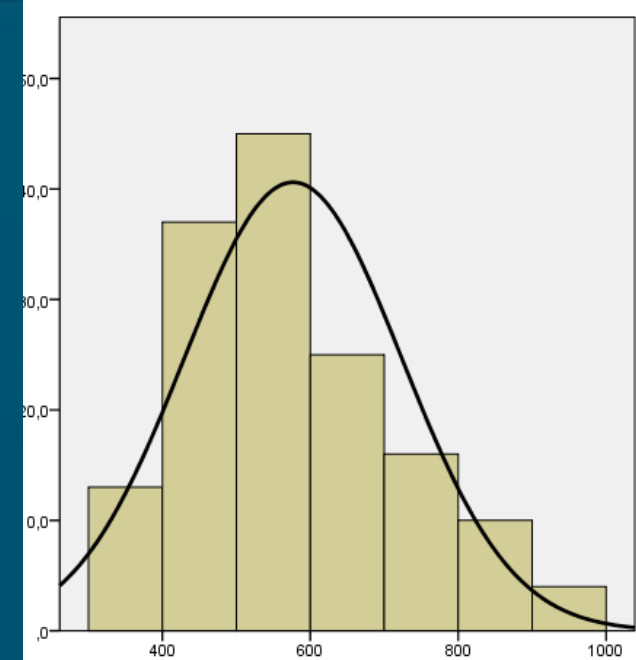
- Generally, researchers in Canada draw upon categories used by StatsCan and information on generational status
- Once data gathering has been completed and demographic description is conducted, then
- That sample can be recoded for ethnic, cultural, linguistic, and generational status
- e.g., recode from country of heritage
 - e.g., China, Philippines, etc.
- To ethnicity:
 - e.g., Asian, Pacific Islander, etc.

Outline for today

- Hermeneutics of data preparation
- Creating secondary “derived” variables
- Descriptives of a dataset, error checking
- Missing data and outliers
- Assumptions of Parametricity
- What to report:
“Informationally adequate” statistics
- Appendix: Project data sources

Data Exploration

- Get **overall** feel of the dataset and identify any **problems** with the data (for fixing)
- In final rounds: **report** & describe basic **characteristics** of the data (use APA style)
- **SPSS**: Analyze → Descriptive Statistics
 - → Descriptives, Frequencies, P-P plots, ...
 - → Explore has more stats, plots
- **Graphs**: Legacy Dialogs → Boxplot, Error Bar, Histogram, Scatterplot, ...



Signs of Data Entry Errors

- Scores that are **outside** of what is **possible**
- Unexplained **gaps** in the frequency output / **histogram**, or odd patterns in **boxplots**
- Cases on your boxplots or histograms that do **not fit** with the rest of the sample (outliers?)
- **Standard deviations** much larger than expected
- **Means** much different from expected
- For possible errors, compare the SPSS file against your **original data source**

Missing Data

- Reasons for missing data, in real research?
 - Systematic gaps indicate a real problem
 - Random gaps may be okay, as long as less than ~5% of the data is missing
- To check for missing data:
 - (1) code it as a separate variable
 - ◆ Dichotomous: “missing” vs. “not missing”
 - (2) compare responses on other variables for “missing” vs. “not missing” groups

Random Missing: Strategies

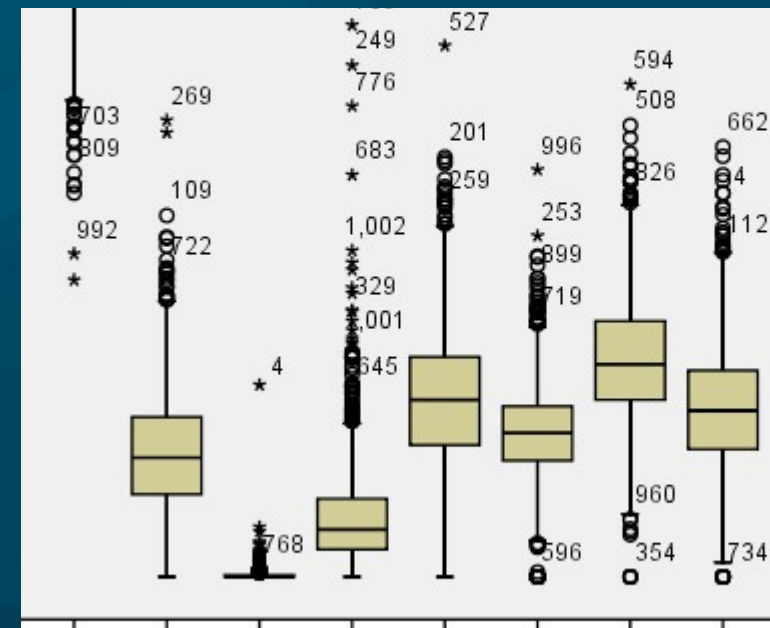
- Go back and **obtain** the bits that were missing
 - Not always **practical**, ethical, or effective
- **Eliminate variables** with the most problems
 - Unless they are **central** to the study!
- **Drop cases**/people with too much missing data
 - May result in insufficient **sample size**
- Estimate and **replace** missing values
 - Simple **imputation** methods like mean substitution have problems

Systematic Missing: Strategies

- Code “missing” var and include it in analysis
 - Might not help clarify import of omissions
- Return to participants & collect missing data
 - No guarantee of improved response rates
- Exclude the kind of participants that choose not to respond from your sample
 - Reduces generalizability of the results
 - May end up altering research question
- Start over, change design and data collection procedures

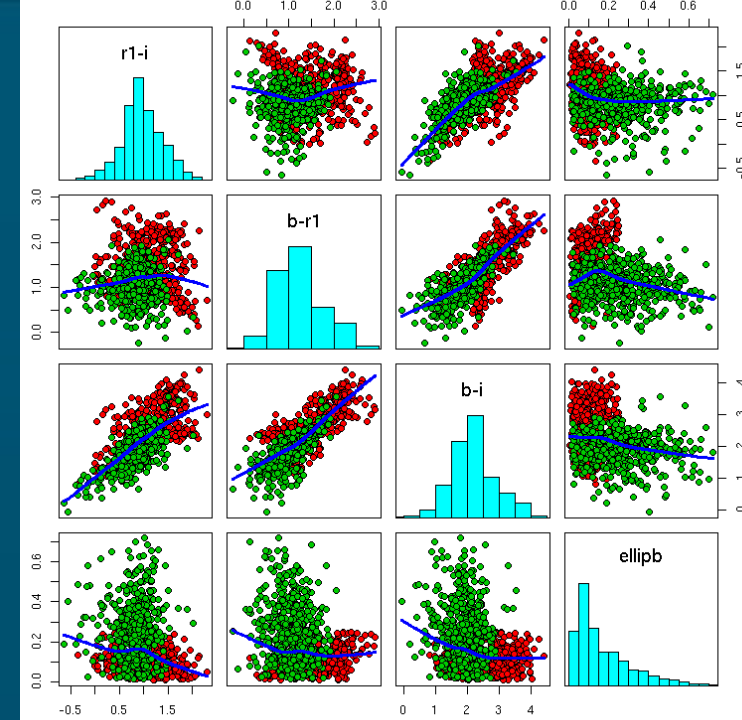
Univariate Outliers

- Definition: Patterns that reflect **persons** from a **different population** from the rest of the sample
- Use **boxplots** to spot them (outliers: *)
- From (a) knowledge of the **literature** and (b) **extremity** of an outlying score, decide
 - (1) whether to **keep** or **exclude** the outliers
 - (2) whether to use **robust** strategies for analysis
- **Multivariate** outliers can also exist



Multivariate Outliers

- Cases that might not appear extreme in any **one** variable, but that don't fit the pattern across **several** variables
- Try using **scatter plots** to detect outliers:
 - Graphs → Legacy Dialogs → Scatter/Dot
 - Choose “**matrix scatter**” to examine several quantitative variables at once.
- Use the “**simple scatter**” option to “zoom in” on a combination of variables of interest.



Richard Holbrey, Uni. Leeds

Handling Outliers

- If the **person** is a valid case,
keep it in the dataset and **check** for impact.
 - Compare results with and without it
- Otherwise, remove the **entire case** from dataset
 - Other strategies exist, but more complex
- Check for **univariate** outliers ONCE, then
- Check for **multivariate** outliers ONCE
 - Do not repeat!
 - Excessive elimination of outliers may result in a very **distorted** dataset!

Outline for today

- Hermeneutics of data preparation
- Creating secondary “derived” variables
- Descriptives of a dataset, error checking
- Missing data and outliers
- Assumptions of Parametricity
- What to report:
“Informationally adequate” statistics
- Appendix: Project data sources

Checking Test Assumptions

- Most statistical tests make assumptions of data.
 - Must always check these assumptions!
- If any assumptions are violated:
 - Use a different procedure, and/or
 - “Clean up” the data to meet assumptions
- Procedures based on the General Linear Model all assume that the data are parametric.
 - e.g., parametric correlation, *t*-tests, multiple regression, the ANOVA family, factor analysis, multi-level modeling, ...!

Assumptions of Parametricity

- *(These generally apply to the DV)*
- **Scale** (interval/ratio) level of measurement
- **Independence**:
 - No **systematic** links between cases
- **Homogeneity of variance**:
 - Variability of scores should be roughly similar across all **variables** and all participant **groups**
- **Normally** distributed:
 - Bell-shape, not **skewed** or **kurtotic**

Parametricity: Interval/ratio

- **Check:** level of measurement for each variable
- **Solutions:** try alternative, **non-parametric** tests:

Parametric	Non-parametric
Pearson's r	Spearman's r_s , Kendall's τ
Multiple Regression	Log-linear or Logistic Regression
Between-group t -test, Within-group t -test	Mann-Whitney U , Wilcoxon Signed-Rank
ANOVA family	Friedman's ANOVA, Chi-square, ...

Parametricity: Independence

- Check: sampling process, meaning of vars
 - e.g., participants from same family?
- Solutions:
 - Eliminate the offending cases or vars
 - Methods designed for “dependent” data
 - ◆ Repeated-measures ANOVA, multi-level (hierarchical) modelling, ...
 - Split data into independent subgroups
 - ◆ e.g., analyze males and females separately

Homogeneity of Variance

- Check: examine **variance** of each subgroup:
 - ◆ Analyze → Descriptive Statistics → Explore
 - ◆ Plots → Spread vs. Level → Untransformed
 - ◆ Choose **group** (factor) vars and **DVs**
- If ratio of **largest** variance to **smallest** is more than **4:1**, we have a problem
- Or: use **Levene's** test:
 - If $p < 0.05$, then assumption may be violated. (**non-significance** = good)
- **Solutions**: may be okay with more **conservative** α , or try **non-parametric** methods

Parametricity: Normality

- Check: Kolmogorov-Smirnov, Shapiro-Wilk tests
 - Skewness / kurtosis (w/confid. interval)
 - ◆ If $p < 0.05$, then we may have a problem
 - Normality plots, P-P plots
 - ◆ Analyze → Descriptive Statistics → Explore
→ Plots → check “Normality plots w/tests”
- Solutions:
 - Transform data as appropriate, or
 - Change specific problematic scores
 - ◆ (not usually recommended!)

Transform Data for Normality

(cf. Behrens, 1997)

- Minor deviations: try square-root ($\sqrt{\text{score}}$)
- Medium deviations: try logarithmic ($\log_{10}(\text{score} + 1)$)
- Severe deviations: try reciprocal ($1/\text{score}$)
 - If tail is to the right, reflect data first:
 - ◆ $\text{maximum_score} - \text{variable}$
- SPSS: Transform → Compute Variable
 - Apply appropriate transformation
 - Use menus (Arithmetic, etc.) to find operations

Practise: Check Parametricity

- Datasets: SPSSExam.sav, AttnDefDis.sav
- Let's study the “effect” of **gender** (IV) on number of **lectures** (DV) or **ADD** symptoms (DV)
- Are the data **parametric**? Try it yourself!
 - **Level** of measurement
 - **Independence**
 - Homogeneity of **variance**
 - **Normality**

Order of Data Preparation

- Examine overall **patterns** of data; identify & correct **data entry** errors; use lots of **graphs**!
- Explore & deal with **missing** data; create **derived** variables like subscales, demographics
- Identify and choose how to deal with univariate and multivariate **outliers**
- Evaluate **assumptions** of your chosen procedure (e.g., parametricity) and deal with any **violations** of assumptions
- Obtain **descriptive** information for the final dataset, and proceed with the analysis

Outline for today

- Hermeneutics of data preparation
- Creating secondary “derived” variables
- Descriptives of a dataset, error checking
- Missing data and outliers
- Assumptions of Parametricity
- What to report:
“Informationally adequate” statistics
- Appendix: Project data sources

“Informationally Adequate”

- **APA** has standards on what you must report, in order to be “informationally adequate”:
- **Sample size** (total N and n_i for each subgroup)
- **Mean** and **SD** for each outcome variable
 - Globally and for each sub-group
- Statistical **significance** (“exact” p -values)
- Measures of **effect size**
- Evidence of sufficient statistical **power**
- **Other** test-specific info (APA, §2.07, pp. 32-35)

Further Reading -- Articles

- Wintre et al. (2000). Generational status and ethnicity in Canada
- Tabachnick et al. (2007). On cleaning data
- Schafer et al. (2002). Missing data primer

APPENDICES

- **Valuable background for research in Canada**
- **Valuable data archive sources available on-line**

StatsCan Downloads

- Selected Demographic and Cultural Characteristics (102), Visible Minority Groups (15), Age Groups (6) and Sex (3) for Population, for Canada, Provinces, Territories and Census Metropolitan Areas, 2001 Census - 20% Sample Data
 - <http://www.statcan.ca/bsolc/english/bsolc?catno=97F0010X2001044>
- Selected Demographic and Cultural Characteristics (105), Selected Ethnic Groups (100), Age Groups (6), Sex (3) and Single and Multiple Ethnic Origin Responses (3) for Population, for Canada, Provinces, Territories and Census Metropolitan Areas, 2001 Census - 20% Sample Data
 - <http://www.statcan.ca/bsolc/english/bsolc?catno=97F0010X2001040>
- Place of Birth of Father (35), Place of Birth of Mother (35) and Generation Status (4) for the Population 15 Years and Over of Canada, Provinces, Territories, Census Metropolitan Areas and Census Agglomerations, 2006 Census - 20% Sample Data
 - <http://www.statcan.ca/bsolc/english/bsolc?catno=97-557-X2006009>

StatsCan Definitions

Source: *Ethnic Diversity Survey - Methodology and Data Quality*

- “questions on the birthplace of respondents and their parents were used to establish the respondent’s generational status. The first generation includes respondents born outside Canada. The second generation includes respondents born in Canada with at least one parent born outside Canada. The third-plus generation includes respondents born in Canada to two Canadian-born parents.”
- “Responses to the ethnic origin question were divided up to form the two main categories of interest: **CBFA+** (Canadian or British or French or Americans or Australians and/or New Zealanders) and Non-CBFA+ (all other responses containing at least one origin other than CBFA+). The **non-CBFA+** category was divided into European origins (for example, German, Italian, Dutch, Portuguese) and non-European (for example, Chinese, Jamaican, Lebanese, Iranian).”

StatsCan Definitions (cont.)

- “CBFA+
- Canadian only – Generation 1 and 2
- Canadian only – Generation 3 and more
- Canadian with BFA+ – Generations 1 and 2
- Canadian with BFA+ – Generations 3 and more
- BFA+ – Generation 1 and 2
- BFA+ – Generation 3 and more
- Other Europeans with Canadian – Generation 1 and 2
- Other Europeans with Canadian – Generation 3 and more
- Other Europeans – Generation 1
- Other Europeans – Generation 2
- Other Europeans – Generation 3 and more
- Other non-Europeans with Canadian – All generations
- Other non-Europeans – Generation 1
- Other non-Europeans – Generation 2 and more

■ Non-CBFA+

ICPSR

- “Established in 1962, ICPSR is the world's largest archive of digital social science data.”
 - <http://www.icpsr.umich.edu/ICPSR/>
- e.g., National Institute of Mental Health Collaborative Psychiatric Epidemiology Surveys (CPES)