MHPE 494: Data Analysis

Alan Schwartz, PhD
Department of Medical Education
Memoona Hasnain, MD, PhD, MHPE
Department of Family Medicine

College of Medicine
University of Illinois at Chicago

Welcome!

- Your name, specialty, institution, position
- Experience in data analysis
- Why this class?
  What are your expectations and goals?

The Analytic Process

- Formulate research questions
- Design study
- Collect data
- Record data
- Check data for problems
- Explore data for patterns
- Test hypotheses with the data
- Interpret and report results

Covered in Research Design/Grant Writing

Covered in Writing for Scientific Publication

Monday AM

- Introduction
- Syllabus
- Data Entry
- Data Checking
- Exploratory Data Analysis

Data entry

or,

"Garbage in, garbage out"

Data Entry

- Data entry is the process of recording the behavior of research subjects (or other data) in a format that is efficient for:
  - Understanding the coded responses
  - Exploring patterns in the data
  - Conducting statistical analyses
  - Distributing your data set to others
- Data entry is often given low regard, but a little time spent now can save a lot of time later!

Methods of data entry

- Direct entry by participants
- Direct entry from observations
- Entry via coding sheets

- Entry to statistical software
- Entry to spreadsheet software
- Entry to database software

Data file layout

- Most data files in most statistical software use “standard data layout”:
  - Each row represents one subject
  - Each column represents one variable measurement
- Special formats are sometimes used for particular analyses/software
  - Doubly multivariate data (each row is a subject at a given time)
  - Matrix data

“Standard data layout”

Id	Female	YrsOld	GPA
1	1	19	3.5
2	0	21	3.4
3	1	20	3.4

Missing data

- Data can be missing for many reasons:
  - Random missing responses
  - Drop out in longitudinal studies (censoring)
  - Systematic failure to respond
  - Structure of research design
- Knowing why data is missing is often the key to deciding how to handle missing data

Missing data

- Approaches to dealing with missing data:
  - Leave data missing, and exclude that cell or subject from analyses
  - Impute values for missing data (requires a model of how data is missing)
  - Use an analytic technique that incorporates missing data as part of data structure

Naming Variables

- Variables should have both a short name (for the software) and a descriptive name (for reporting)
- Name for what is measured, not inferred
- Short names should capture something useful about the variable (its scale, its coding)
- Better names:
  - Q1-Q20, IQ, MALE, IN_TALL, IN_TALLZ
- Worse names:
  - INTEL, SEX, SIZE

Coding Variables

- Depends on *measurement scale*
  - Nominal, two categories: Name variable for one category and code 1 or 0
  - Nominal, many categories: Use a string coding or meaningful numbers
  - Ordinal: Code ranks as numbers, decide if lower or higher ranks are better
  - Interval/Ratio: Code exact value

Labeling Variable Values

- For nominal and ordinal variables, *values* should also be labeled unless using string coding.
- Value labels should precise indicate the response to which the value refers.
  - Example: Educational level ordinal variable:

1 = grade school not completed
2 = grade school completed
3 = middle school completed
4 = high school completed
5 = some college
6 = college degree

Error Checking

- Goal: Identify errors made due to:
  - Faulty data entry
  - Faulty measurement
  - Faulty responses
- Prior to analyses. Not hypothesis-based

Range checking

- The first basic check that should be performed on all variables
- Print out the range (lowest and highest value) of every variable
- Quickly catches common typos involving extra keystrokes

Distribution checking

- Examining the distribution of variables to insure that they'll be amenable to analysis.
- Problems to detect include:
  - Floor and ceiling effects
  - Lack of variance
  - Non-normality (including skew and kurtosis)
  - Heteroscedascity (in joint distributions)

Eccentric subjects

- Patterns of data can suggest that particular subjects are eccentric
  - Subjects may have misunderstood instructions
  - Subjects may understand instructions but use response scale incorrectly
  - Subjects may intentionally misreport (to protect themselves or to subvert the study as they see it)
  - Subjects may actually have different, but coherent views!

Verbal protocols

- Verbal protocols (written or otherwise recorded) can help to distinguish subjects who don't understand from subjects who understand, but feel differently than most others.
  - "What was going through your head while you were doing this?"
  - "How did you decide to response that way?"
  - "Do you have any comments about this study?"
- Debriefing interviews can be used similarly

Holding subjects out

- If a subject is indeed eccentric, you must decide whether or not to hold the subject out of the analysis. Document these choices.
  - Pros: Data will be cleaner (sample will be more homogenous, less noisy)
  - Cons: Ability to generalize is reduced, bias may be introduced
- If a group of subjects are eccentric in the same way, it's probably better to analyze them as a subgroup, or use individual-level techniques.

Cleaning data

- When only a few data points are eccentric, a case can sometimes be made for *cleaning* the data.
  - Example: Subjects were asked to respond on a computer keyboard to money won or lost in a game on a scale from -50 (very unhappy) to 50 (very happy). One subject's ratings were:
  - +$5 = "10", -$5 = "-3", -$20 = "-40", -$10 = "20"
  - Should the "20" response be changed to "-20"?
- Document these choices.

Four great SPSS commands

- *Transform…Compute*: create a new variable computed from other variables.
- *Transform…Recode*: create a new variable by recoding the values of an existing variable.
- *Data…Select cases*: choose cases on which to perform analyses, setting others aside.
- *Data…Split file*: choose variables that define groups of cases, and run following analyses individually for each group.

Exploratory Data Analysis

- The goal of EDA is to apprehend patterns in data
- The better you understand your data set, the easier later analyses will be.
- EDA is *not:*
  - "data-mining" (an atheoretical look for any significant findings in the data, capitalizing on chance)
  - hypothesis testing (though it may help with this)
  - data presentation (though it **does** help with this)

EDA Tools: Stem-and-leaf plots

```
Starting Salary Stem-and-Leaf Plot
 Frequency     Stem &  Leaf
      1.00       0 .  &
      9.00       0 .  889
      9.00       1 .  001
     22.00       1 .  2222333
     20.00       1 .  4555555
     39.00       1 .  66667777777777
     57.00       1 .  888888888999999999
    139.00       2 .  00000000000000000000000000000011111111111111111
    118.00       2 .  22222222222222222233333333333333333333
    126.00       2 .  44444444444444444455555555555555555555
    132.00       2 .  66666666666666666666667777777777777777777
     98.00       2 .  88888888888888888899999999999
    113.00       3 .  00000000000000000000011111111111111
     94.00       3 .  22222222222222223333333333333
     55.00       3 .  444444444555555555
     21.00       3 .  6666677
     15.00       3 .  88889
     11.00       4 .  0001
      6.00       4 .  23
      2.00       4 .  4
     13.00 Extremes    (>=45000)
 Stem width:  10000 Each leaf: 3 case(s) & denotes fractional leaves.
```

EDA Tools: Central Tendency

- Measures of central tendency: what one number best summarizes this distribution?
- Most common are mean, median, and mode
- Others include trimmed means, etc.
- Example:

Starting salary (N=1100)	
Mean	26064.20
Median	26000.00
Mode	20000

EDA Tools: Variability

- Measures of variability: how much and in what way do the data vary around their center?
- Most common: standard deviation, variance (sd squared), skew, kurtosis

Starting salary (N=1100)	
Mean	26064.20
Std. Deviation	6967.98
Variance	48552771.77
Skewness	.488
Std. Error of Skewness	.074
Kurtosis	1.778
Std. Error of Kurtosis	.147

EDA Tools: Norms and percentiles

- Percentiles are pieces of the frequency distribution: for what score are x% of the scores below that score. They can be used to set norms.

Starting salary (N=1100)	
Percentiles	
5	15000.00
25	21000.00
50	26000.00
75	30375.00
95	36595.00

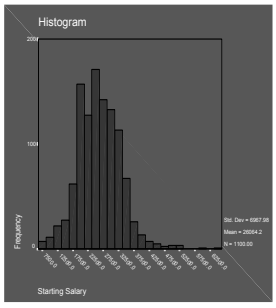
EDA Tools: Graphing

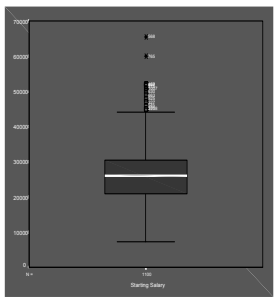
- Graphing puts the inherent power of visual perception to work in finding patterns in data
- Choice of graph depends on:
  - Number of dependent and independent variables
  - Measurement scale of variables
  - Goal of visualization (compare groups? seek relationships? identify outliers?)

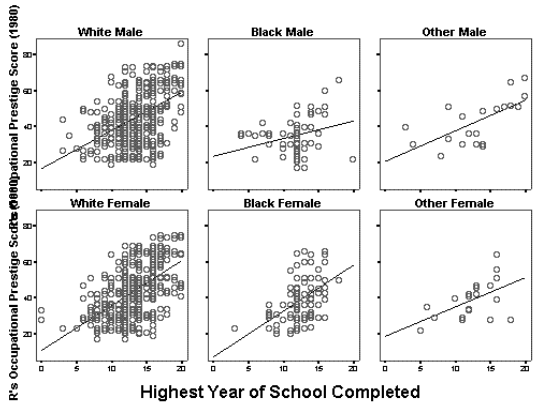
Types of graphs

- One variable: Frequency histogram, stem and leaf
- Two variables (independent x dependent):
  - nominal x interval: bar chart
  - interval x nominal: histogram
  - interval x interval: scatter plot
- Three variables (ind x ind x dep):
  - nominal x nominal x interval: 3d or clustered bar chart
  - nominal x interval x interval: line chart
  - interval x interval x interval: 3d scatter plot
- Four variables (ind x ind x ind x dep): matrix

Examples

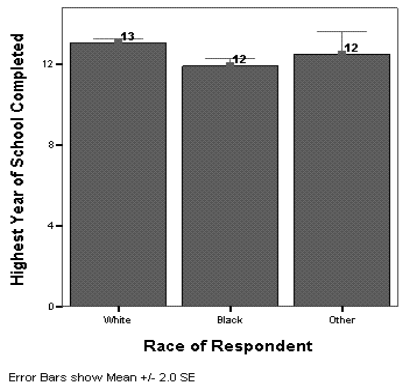






Error bars

- Most graphs provide measures of central tendency or aggregate response
- Error bars are a natural way to indicate variability as well. Some common choices to show:
  - 1 standard deviation (when describing populations)
  - 1 standard error of the mean
  - 95% confidence interval
  - 2 standard errors of the mean



Assignment

- Explore the hyp data, and describe the distribution of each of the variables.