DATA PREPARATION

Supplemental Slides for Multilevel Modeling Workshop

1. PREPARE DATA

1. 1. Inspect data for plausibility and normality

1.1.1. Transform as necessary

1. 2. Dummy coding

1. 3. Centering

1.4. Organize data

- Without Step 1.2: You can't perform the analysis
- Without Step 1.1: You could find non-significance when your experimental hypothesis is true!!!

1.1. IS THE DATA PLAUSIBLE?

- Data plausibility
  - *The data represent possible values on the measured dimension*
- Causes of *im*plausibility
  - Data entry errors
  - Error in measurement reliability or validity

TYPES OF PLAUSIBILITY

- Entered data
- Missing data

ENTERED DATA PLAUSIBILITY

- Data plausibility check:
  - Calculate min and max observations for variable
  - Compare to known min and max for each scale

DATA PLAUSIBILITY IN SPSS & R

- **SPSS:**
  - FREQUENCIES VARIABLES=x y

    /STATISTICS= MINIMUM MAXIMUM.
- **R:**
  - min(x), max(x)

MISSING DATA PLAUSIBILITY

- *Missing data* plausibility check:
  - Are missing data missing or just misplaced?
  - Are they missing at random or systematically?

MISSING DATA PLAUSIBILITY

- Code
  - SPSS: Look through your data in the data viewer
  - R: cbind(x1,x2,x3,w,y,…)

DATA STRUCTURE

Level 2	Level 1	Level 2	Level 1	Level 1
Participant *(id)*	Clinic Visit *(visit)*	Education *(education)*	Arterial BP *(bp)*	Body Mass Index *(BMI)*
2448	1	4	82.0	26.97
2448	3	4	84.3	.
6238	1	2	94.3	28.73
6238	2	2	81.3	29.43
6238	3	2	80.0	28.5
…	…	…	…	…

1.1. IS THE DATA NORMAL?

- Print a histogram of each variable
- Determine the mean, median, variance, and skew of the variable
  - If skew > +1.5 or < -1.5, then transform the data
- If your data aren't even supposed to be normal, use a generalized linear model (*end of today*)

DATA NORMALITY

- **SPSS:**
  - GRAPH HISTOGRAM=x.
  - FREQUENCIES VARIABLES=x1 x2 y

    /STATISTICS= MEAN MEDIAN MODE STDDEV SKEWNESS

    /ORDER=ANALYSIS.
- **R:**
  - hist(x)
  - mean(x)
  - median(x)
  - sqrt(var(x))
  - skewness(x)

NON-NORMALITY

- What to do?
  - Transform that variable!
    - Skewness transformations:
      - Positive skew: log(x), ln(x), or x^{-k}
      - Negative skew: x^2, x^3, x^k
- Choose another comparison distribution

1.2. DUMMY CODING

- Correctly dummy code categorical variable:
  - 2-level variables: -1 and 1
  - 3-level variables or more:
    - For *k* variables, make *k* - 1 dummy variables
    - Choose one group to be the “default” (always coded -1)
    - Name each dummy variable with names of the non-default levels
    - Code each dummy variable with “-1” and “1”

1.3. CENTERING

- Subtract the mean from all your variables
- Allows for proper estimation of lower-order (main) effects

CENTERING IN SPSS & R

- **SPSS:**
  - COMPUTE x=x-[mean].
  - EXECUTE.
- **R:**
  - x <- x - mean(x)

1.4. ORGANIZE YOUR DATA

- Every value of a Level 1 variable should have its own row
  - E.g., you studied two partners from 100 married couples
    - You would have 200 rows, with each individual participant having 1 row
- E.g., you measured reaction times to 50 stimuli in 12 task blocks?
  - You would have 600 rows, with reaction times to each stimulus having 1 row

CORRECTLY ORGANIZED DATA

Participant	Day	Veggie Servings
1	1	3
1	2	4
1	3	2
2	1	3
2	2	4
2	3	5
3	1	1
3	2	2
3	3	2

INCORRECTLY ORGANIZED DATA

Participant	Veggies Day 1	Veggies Day 2	Veggies Day 3
1	3	4	2
2	3	4	5
3	1	2	2
4	3	3	3
5	6	7	7
6	3	4	3
7	4	4	2
8	2	5	4
9	4	3	5

1.4. ORGANIZE YOUR DATA

- Every value of a Level 1 variable should have its own row
  - E.g., you studied two partners from 100 married couples
    - You would have 200 rows, with each individual participant having 1 row
- E.g., you measured reaction times to 50 stimuli in 12 task blocks?
  - You would have 600 rows, with reaction times to each stimulus having 1 row

CORRECTLY ORGANIZED DATA

Participant	Day	Veggie Servings
1	1	3
1	2	4
1	3	2
2	1	3
2	2	4
2	3	5
3	1	1
3	2	2
3	3	2

INCORRECTLY ORGANIZED DATA

Participant	Veggies Day 1	Veggies Day 2	Veggies Day 3
1	3	4	2
2	3	4	5
3	1	2	2
4	3	3	3
5	6	7	7
6	3	4	3
7	4	4	2
8	2	5	4
9	4	3	5