Getting Started with Data and Stats

For N710 Project – by Melinda Higgins, PhD

Dated 09/13/2018

Suggestions to get started with your data and stats:

1. Codebook
   1. Variable name – description
   2. Categories? Exact? Units?
   3. Missing data coding?
      1. 97, 98, 99 – don’t know, refused, missing
2. Instruments, protocol
   1. Version?
   2. Subscales? (BDI – beck depression index)
   3. Know the valid range for the scores
   4. Clinical points or ranges mean – interpretation
   5. Population norms? For example population norm 50 +/- 10. <50 poorer QOL, >50 better QOL (e.g. PROMIS, NeuroQOL, SF36/12…)
   6. Which way is positive? Higher scores better or worse? “goldilocks”
3. Things to check data
   1. Descriptive stats
      1. Mean, standard deviation
      2. Median – 50th percentile
      3. 25th, 75th percentile (Q1, Q3) – IQR interquartile range
      4. Range – min, max
         1. Check outliers
         2. Unusual or just wrong – typos, units
         3. Incorrect scoring – 102? CESD depression 20 scored 0-3 = total 0 to 60. If have CESD score 65, something went wrong
   2. Distributions
      1. Histograms – look for symmetry and single “mode”
      2. normal probability plots – smaller datasets
      3. skewness and kurtosis
         1. if you have skewness – longer tail to right (higher number) – number of times xxx, length of stay, biomarkers
            1. math transformation – sqrt, ln (natural log)
         2. less likely to have left skewness (sparse values on the smaller/lower end) – knowledge tests, adherence
            1. sqrt(max (100) – x)
         3. suggestion – split – dichotomize – high, low, clinical point (not depressed, depressed)
         4. zero-inflated – 0’s, non-zero (>0) – have no comorbidities, people with 1 or more – dichotomize
   3. categorical data/ordinal data
      1. frequencies, counts, percentages
      2. small numbers <5% (outcome data)
      3. collapse categories together – be careful with “other”
      4. how you want to interpret categories
      5. dummy coding – which reference?
      6. “general linear model” – regression and ANOVA (pairwise comparisons)