Organizing Your Approach to a Data Analysis

Written by Professor Daniel Gillen, UC Irvine
Modified from a document by Scott Emerson, U of Washington

I. Before looking at the data
   A. Identify overall goal of the study
   B. Identify specific aims and how they relate to overall goal
      1. Identify the current state of scientific knowledge
      2. Identify the competing hypotheses that the study is designed to discriminate between
      3. (Often dictated by available data)
   C. Refine scientific hypotheses into statistical hypotheses
      1. Identify type of question
         a. Prediction, estimation, or testing
         b. Identifying groups, quantifying distributions, or comparing distributions
      2. Where appropriate, specify statistical hypotheses in terms of a summary measure for the distribution of measurements
         a. e.g., mean, median, proportion above a threshold, event rate
   D. Consider design of ideal experiment
      1. Ignore practical, ethical limitations in order to be able to later compare how close the actual situation is to the ideal
         a. Who/what would be the sampling units
         b. What would be the intervention
         c. How would subjects be assigned to the intervention
         d. What would be the variables measured
   E. Available data
      1. Sampling scheme
         a. Retrospective vs prospective
         b. Observational vs intervention
         c. Inclusion, exclusion criteria
      2. Variables in the data set
         a. Names
         b. Relationship to real world quantities
         c. Conditions under which they were measured
         d. Units of measurement (limitations)
            – e.g., qualitative vs quantitative, continuous vs discrete, patterns of missing data
      3. Categorization of variables according to use in analysis
         a. Response (outcome) variables
         b. Predictor variable of interest (variable identifying groups)

c. Variables identifying subgroups to explore effect modification

d. Potential confounders

- Association with response variable (in truth)
- Association with predictor of interest (in the sample)
- Not in causal pathway of interest

e. Variables which allow increased precision

- Variables predictive of response, but not associated with predictor of interest
- Questions about effects within such groups can be answered with more precision than questions about effects in the larger population (e.g., adjusting for age)

f. Surrogates for response

- Variables in the causal pathway of interest
- Variables measuring a later effect of the response

g. Irrelevant

II. Univariate descriptive statistics

A. Goals

1. Identify errors in the data
   a. Particularly unusual measurements (out of range)
   b. Unusual combinations of measurements
2. Verify your understanding of the measurements
3. Identify patterns of missing data
4. Identify exact population used in study (Materials and Methods)
5. Identify aspects of the data that may present technical statistical issues
   a. Ideal: allows easiest, most precise statistical inference with smaller sample sizes
      - equal information about all groups being investigated (? equal sample sizes)
      - measurements of response within each group distributed symmetrically with no 'long tails' (outliers)
      - no missing data
   b. Potential problems suggesting possibility of problematic scientific interpretation (problems which can not necessarily be solved with the available data)
      - missing data patterns
   c. Potential problems suggesting less generalizable statistical analysis (problems not necessarily indicated by the measures of statistical confidence)
      - 'Outliers' in distribution of grouping variables (predictors): i.e., low sample sizes in some groups that are far away from the rest of the data (e.g., trying to determine an age effect in a sample in which most are between 10 and 20 years old, but one subject is 80)
   d. Potential technical problems suggesting possibility of less precise inference (problems that will tend to lower our reported level of statistical precision)
      - 'Outliers' in distribution of response
      - Too little variation in the distribution of the grouping variables (e.g, trying to de-

termine an age effect from a sample in which everyone is between 20 and 21 years old)

- Too much association among the different grouping variables (e.g., trying to determine an age effect when all the young subjects are male and all the old subjects are female)

e. Potential technical problems which suggest we might need to use more complicated statistical methods

- Repeated measurements on the same sampling unit (correlated response)
- When comparing means: unequal variability across groups being compared
- When comparing time to events: lack of proportional hazards
- When adjusting for covariates: nonlinear effects; interactions

C. Order of investigation

1. Potential confounders
2. Predictor of interest
3. Response

D. Tools

1. Frequency tables
2. Mean, median, standard deviation, etc.
3. Box plots, histograms

III. Bivariate and trivariate descriptive statistics

A. Goals

1. Identify confounding relationships
   a. Associations between other variables and predictor of interest
   b. Associations between other variables and response
2. Identify important predictors of response
   a. Univariate effects
   b. Effect modification (interactions)
3. Identify surrogates of response
4. Characterize form of functional relationships (linear, etc.)

B. Ideal

1. Predictor of interest has no association with any other predictors
2. Only a few variables are markedly associated with response
3. All associations look like a straight line relationship
4. No interactions (effect modification)

C. Order of investigation

1. Relationships among other predictors
2. Relationships between predictor of interest and other predictors
3. Relationships between response and other predictors
4. Relationships between predictor of interest and response overall
5. Relationships between predictor of interest and response within subgroups

D. Tools
    1. Contingency tables
    2. Stratified means, medians, standard deviations, etc.
    3. Stratified box plots, histograms, etc.
    4. Scatterplots
    5. Stratified scatterplots
    6. Correlations

IV. Defining a suitable context for modeling
A. Goals
    1. Choosing appropriate form for response variables
        a. Selection of measure of response
            – Transformations of available data
        b. Summary measure to use as basis for statistical model
    2. Selection of groups to be investigated / compared
            – Form for predictor of interest
            – Identification and form of interactions (effect modification)
            – Identification and form of potential confounders to be modeled
            – Identification and form of precision variables to be modeled
    3. Choosing analysis method (type of regression)
B. Methods
    1. Ideal: Statistical model dictated entirely by scientific question (before looking at the data)
    2. Practical: Model building
        a. Educated guess for first models
        b. Fit models
        c. Evaluate validity of necessary assumptions

V. Model Building to Address Primary Question
A. Goals (in order of importance)
    1. Selection of variables to address scientific questions (main effects and interactions)
    2. Selection of variables to minimize bias (address confounding)
    3. Selection of variables to maximize precision
    4. Selection of models which are easiest to implement (usually: have the least technical requirements on the distribution of response)
B. Methods
    1. Addressing scientific question: Thinking about the problem
    2. Addressing confounding: Adding or removing variables and observing effect on other regression parameters relative to findings in bivariate description of data
    3. Addressing precision: Determining which variables tend to predict response (many difficult issues here)
    4. Evaluate extent to which data meets technical requirements of statistical procedures

VI. Exploratory Analyses for Hypothesis Generation

A. Modeling of exact form of predictor-response relationship (e.g., dose-response)

B. Identification of other predictors of response

C. Subgroup analyses: Compare effect of predictor of interest on response within subgroups (effect modification)

VI. Reporting Results and Interpretation

A. Scientific Background and Hypotheses

B. Materials and Methods

1. Sampling scheme
2. Most basic descriptive statistics

C. Results (more objective first)

1. Descriptive statistics
2. Results of analyses about primary question
    a. Estimates of effect
        – Point estimates (single best estimate)
        - Interval estimates (range of estimates indicating precision)
    b. Decisions about hypotheses
        – Binary decision (yes or no)
        – Measure of statistical confidence in precision
3. Results of analyses about prespecified secondary questions or questions which demonstrate consistency (or lack of same) across alternative approaches
4. Results of analyses about questions that arose during analysis and that the vast majority of readers would agree could and should be answered by the data

D. Discussion (subjective, including particularly data-driven analyses)

1. Elaboration on ways that these analyses address the overall goal of the study
2. Results of the most speculative analyses of the data