Biost/Stat 578 B: Data Analysis
Emerson, September 29, 2003
Handout #1

Organizing Your Approach to a Data Analysis

The general theme should be to maximize thinking about the data analysis and to minimize the time spent trying to interpret randomly selected data analyses. In most cases, the statistical methods used to answer the primary question can be selected prior to looking at the data. (Data driven selection of analysis methods weakens confidence in the statistical inference.) Exploratory analyses (i.e., those analyses based on models selected after looking at the data) should be clearly labeled as such.

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 would be the subjects

         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 and sample size
   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 scientific meaning
   a. Demographic (age, sex, etc.)
   b. Baseline physiology (SBP, performance status)
   c. Baseline disease risk factors, prognosis
   d. Measures of treatment intervention
   e. Measures of ancillary clinical course during treatment (e.g., ancillary treatments, environmental conditions)
   f. Measures of treatment outcome
   g. (Others specific to the scientific setting)
4. 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
      - Causally associated with response variable (in truth), but not in causal pathway of interest
      - Association with predictor of interest (in the sample)
   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 determine 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.
    a. Quick numerical methods of detecting ‘outliers’ by detecting asymmetry
        - Mean and median markedly different
        - Mean, median not midway between minimum and maximum
        - Mean, median not midway between 25th and 75th percentiles
        - For positive variables: standard deviation larger than two-thirds of the mean
        - Minimum or maximum too many standard deviations away from the mean (‘too many’ depends on sample size)
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. Exploratory: 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 unanticipated 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)

Reporting Results and Interpretation

The basic principles of reporting the results of a statistical analysis are the same ones we learned about in elementary school science. The elements of a proper scientific lab report are:

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

General Requirements for Ph.D. Applied Exam

In the report of your analysis, you should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naive client (i.e., the researcher who brought you the data and/or involved you in the analysis) or an interested lay person.

Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science. To wit:

1. **Summary**: Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Give the most pertinent estimates, confidence intervals, and P values. **Note that estimates and confidence intervals regarding the main question of interest are also important even when there is no statistically significant effect.** Don't give too much detail here, but do note any significant problems that were encountered. The basic goal is to have all the key information in your summary, and the rest of your report is the supporting detail.

2. **Background**: Provide a description of the scientific motivation for the analysis. Use your own words rather than copying the description provided by the client. By providing your understanding of the problem, the client may be able to correct any misconceptions that you had about the science. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis.

3. **Questions of Interest**: List the specific questions that your client posed as well as the questions that you answered. Highlight discrepancies between the two categories of questions. (It is not at all uncommon that the question posed by a researcher could not be answered statistically with the data they provide.

4. **Source of the Data**: Describe the source and sampling methods for the data, if known. Note that the source and sampling methods can not generally be divined from the data (e.g., you cannot tell from the data whether the subjects were randomized to intervention or not). Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under Results. But if there are aspects of the descriptive statistics that are of interest solely for a technical description of the sampling plan, they can go here.

5. **Statistical Methods**: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for the client to use in the manuscript. Include references for non-standard techniques. You may want to describe the software used, and certainly want to describe the methods used for ensuring the appropriateness of your models. Explain how you handled common problems like missing data, multiple comparisons, etc. 2) Explain the basic philosophy behind the analysis techniques in layman's terms. Provide interpretations for all parameter estimates. Motivate transformations. Describe the use of P values and confidence intervals if they play an important role in your analysis. Explain why you didn't use more common techniques if necessary.

6. **Results**: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the client up to the analyses gradually.

   a. Start off with descriptive statistics. This is an area often given short shrift in previous years. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

b. Then go to the major models used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, CI, P values). Make sure you provide scientific units for the estimates. Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here.

c. Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses. Present the results of your analyses in tables and publishing quality figures.

DO NOT INCLUDE OUTPUT FROM STATISTICAL PROGRAMS. (Such means little to me and nothing to a client). When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naive researcher. Thus, if you log transform your response, present geometric mean ratios rather than linear regression parameters. Present confidence intervals rather than the values of Z, t, F, or χ^2 statistics.

7. **Discussion**: Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses.
8. **Appendix**: Anything of an overly technical nature should be put in an appendix. You may want to include extensive tables in an appendix instead of the main results section.

The major theme of the above is to write to the client and the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence ("We found no evidence to suggest that we could not rely on the results from our analysis.") You are reporting your major results and impressions of the data. If the client wanted to see every detail, he/she would have to do the analysis himself/herself.

Grading

Written report

Your papers will be graded with respect to three major areas:

1. Scientific approach
   a. Did you investigate problems in the sampling that might materially affect the results?
   b. In addressing each of the questions, did you choose appropriate models to answer the scientific questions?
2. Statistical approach
   a. Were the methods chosen appropriate for the data at hand? Were any key assumptions violate?
   b. Were the methods chosen reasonably efficient?
3. Written report
   a. Were your findings well documented in a succinct manner?
   b. Was the report written at an appropriately low level?