Lecture Notes in Quantitative Biology

Course Summary

Last lecture (from 29 November 1993)

Chapter 24 Question during review session Fall 2000 (on tape)

Questions during review sessions Fall 2001)

Quantitative methods are used in a variety of ways. Some people become practicing scientists, and use quantititave methods directly to analyze data and discover how biological systems function. Many more people rely upon and use the results of quantitative analyses in business and government. The practice of quantitative analysis affect the lives of people indirectly, through use in medical research, the setting of insurance rates, estimating environmental risks etc. Consequently, it is important that we be able to evaluate quantitative analyses, and to understand the ideas and principles of sound quantitative analysis. Such as:

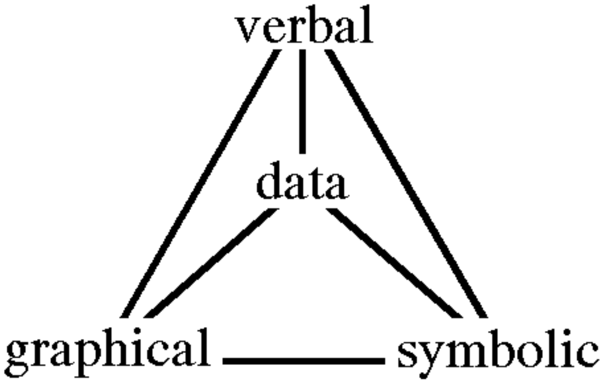
1. Reducing complex situations to useful simplifications (models).
2. Identifying patterns in complex phenomena.
3. Making decisions in the face of uncertainty.
4. Evaluating models (simplifications) relative to data.
5. Isolating causal mechanisms through efficient design and rigourous testing of formal hypotheses.

The key concepts in this course were:

QUANTITIES
FREQUENCY DISTRIBUTIONS
HYPOTHESIS TESTING
THE GENERAL LINEAR MODEL
EXAMINING RESIDUALS,
THE GENERAL*IZED* LINEAR MODEL
RANDOMIZATION AND BOOTSTRAP TESTS

For final exam:

Be able to proceed along lines shown:



For final exam:

Able to define quantities
To separate response from explanatory variables.
To assign symbols, units, and dimensions to variable quantities.
Check dimensional consistency of equations.

Frequency distributions
  Construction of:
    absolute frequency distribution F(Y=k)
    relative frequency distribution RF(Y=k)   = pdf(Y)
    absolute cumulative frequency distribution $F(Y \leq k)$
    cumulative relative frequency distribution $RF(Y \leq k)$ = cdf(Y)
  Theoretical models of frequency distributions:
      Normal, Poisson, Binomial, chisquare, F, t
      and when to apply each.
  Using theoretical frequency distributions.
    obtain outcome (statistic) for given p-value using invcdf
    obtain p-value for outcome (statistic) using cdf

Hypothesis testing
  Declare decision and interpret result.
  Type I and Type II errors

The general linear model (ANOVA, regression, ANCOVA, etc)
  write equation for specific model
      either from name ("t-test") or from data situation
  nested designs, crossed designs, interaction terms
  state H_o/H_a pair
  Calculate a mean, a slope, a variance, a covariance, a correlation, an odds ratio
            (find formula in book and apply formula)
  Partition variance (make calculations within ANOVA table)
  Calculate test statistics F t X^2 G   (find and apply formula)
  Select appropriate theoretical frequency distribution for an analysis:

Response variable:	Ratio/Interval	Binomial	Poisson
test statistic	F t z	X^2 G	X^2 G
cdf	F t Normal randomization	Chisq	Chisq

Residuals

1. Structural model acceptable ? (no bowls/arches)

Seber, G.A.F. 1966. The Linear Hypothesis: A General Theory. London, Griffin.

2. Assumption for F-distribution met ?
   $E(e) = 0$
   $Var(e) = \sigma^2$ = constant . . . . (no cones)
   $Cov(e_i\ e_j) = 0$ Independent of each other? . . . . (no pattern in lag plot)
   e_i normal ? . . . . (plot histogram, rootogram, nscores).

If assumptions not met, state appropriate action in each case.

Multivariate Analysis (no calculations). Recognize and explain
   Correlation. Autocorrelation.
   Correlation matrix (dropped 1995 onward)
   Manova (dropped 1996 onward). Mancova (dropped 1996 onward).
   Canonical correlation, factor analysis, discriminant analysis.
   Clustering algorithms (dropped 1995 onward)

For final exam. Open Book.
   Bring all material--text, tables, labs, notes, calculator
   Organize material for quick access (where to go and how to use)
      labels for quick access
      indices for paginated notes (text is already indexed)
      lists of definitions (or lists + source of def)
      lists of formula (+ sources for application)
      re-write some notes, if necessary.
      make lists of Minitab commands (n.b. these are at end of lab manual)
   Summarize material in your own words.
   Can you explain it to someone else ?

Practice problem-set-up.

Questions, during review sessions in Fall 2001

1. Odds calculated, but not used in computing G-statistic. Why not ?
   Proportions used to show how to compute G-statistic, following method in book. The frequency is the response variable for these methods.
   However, statistical packages use iterative fitting rather than direct computation by formula in book. For these calculations, the response variable is an odds.

2. How many H_A/H_o pairs should be examined in a GLM.
   Always examine interaction terms.
      If these are significant, then break analysis according to one of the factors.
      If interaction terms not significant, proceed to main effects.
   Main effects not all tested. Some are secondary, present in the model for statistical control, to eliminate effects of that factor, and reduce the error SS.

3. What is rule for 1-tail versus 2-tail testing.
   It depends on the knowledge of the investigator.
   If little is known, then 2-tailed test usually performed. The p-values reported by a statistical package will usually be 2-tailed.
   If more is known, then a 1 tailed test can be performed. The advantage of this is that the test is more sensitive, it can detect smaller differences at the 5% significance level.
   Tables report both 1-tail and 2-tail probabilities. One has to check and make sure which is being used. Minitab cdf reports 1 tail only, for t and normal.
   Note that F-and chisquare distributions produce p-value from one tail, but the p-value applies to the 2-tailed test.

4. Where does s_r come from ? (for correlation coefficient).
   This is the standard deviation for the statistic r.
      Each statistic has its own stdev.
   The formula for a standard deviation will be found in texts.

5. How do you locate the differences after performing an ANOVA ?

A priori. 1 test for each df. Preferable because they use knowledge of the investigator. Students usually have enough information about their own data to set these up.

A posteriori. These are carried out by software packages, according to any of several algorithm. They introduce a penalty for multiple testing. For ANOVA with 8 categories there are 7 legitimate comparisons because 7 df, but there are 7*6/2 = 21 possible pairwise comparisons. There is no clear agreement on the best algorithm, or penalty for multiple testing. They are a substitute for thought about the analytic situation.

5a. What if you run out of tests, using *a priori* approach?

Because there are few tests, the must be used judiciously, with thought.

5b. Can *a posteriori* tests be performed in a crossed design ?

This is possible in a paired comparisons, but difficult when factors have more than two categories.

6. What if assumptions for p-value (F-distributions, etc) not met ?

1. Ignore the problem. This unfortunately is the prevailing practice.
   Worse, most attempts to correct the problem are introduced before showing the problem exists.
2. Does it matter ?
   If n large, decision won't change because p-value via randomization will be close to that from statistical distribution.
   If p-value far from criterion, then decision won't change because p-values via randomization rarely change by more than factor of 5 (usually by factor of 2 or less).
3. Use generalized linear model to remedy the problem.
   This usually works, but requires some knowledge of how to use link and distribution functions, not covered in this course (or indeed in any undergraduate course for non-statistics majors).
4. Randomize. This gives the most defensible p-value, but it require work.

7. When do you use correlation ? Regression ?

Regression when the variables can be ordered by cause: Y is function of X.
Correlation where there is no obvious ordering.
Graphical test.
Draw graph axes, using the convention that Y (vertical) is a function of X (horizontal). If the graph cannot be switched (swap Y for X) then it is regression. If Y-variable can be put on horizontal axis (swapped for X) then correlation is appropriate.