STATISTICS 511 – Spring 2013

Homework 3 - Due Friday, 1 Feb 2013

1. **Orthogonal polynomials** Imagine an experiment where the treatment is a quantitative variable, e.g. length of time or amount of toxicant, or number of children in a group. Four treatments were randomly assigned to subjects in a completely randomized design. There are 3 subjects per treatment. The treatments have the values: 0, 2, 5, 10. The data are in op.txt on the class web site.

   (a) NOT counting the intercept column, how many orthogonal polynomials are needed to recreate the SS for testing the hypothesis that all treatments have the same mean? Briefly explain.

   (b) Calculate and report the coefficients for all the orthogonal polynomials. Please express the linear coefficient in "nice" form. Some of the others do not have "nice" forms, at least that I can find.
   R hint: dividing the coefficients by min(abs(coeff vector)) often gets to a "nice" form or sufficiently close that you can see what else to do.

   (c) Calculate and report the SS associated with each orthogonal polynomial.

   (d) Fit a simple linear regression. Is the SS associated with the linear orthogonal polynomial the same as the SS for x in the regression. (I'm talking about model SS here).

   (e) Explain why the F statistic for the linear trend in the anova with orthogonal polynomials differs from that in the regression.

   (f) Use an ANOVA to test the hypothesis that the three groups have the same mean. You do not need to report anything from this anova. Instead, explain why the conclusion about effect(s) of the treatment in the ANOVA differs substantially from the conclusion about effect(s) of the treatment in the orthogonal polynomial analysis.
   Hint: It may help to plot the data.

   (g) If you anticipated that the treatment would have a linear effect, because of background knowledge, which p-value should you report in a paper? Explain why.

2. **Two-factor ANOVA** — This data set is from an experiment on childrens's memory. A random sample of 36 fourth-graders from one city were used in the experiment. Two factors are varied: level of reinforcement (none or verbal) and time of isolation (20, 40, or 60 minutes). Students were told to memorize a paragraph and given positive verbal reinforcement or no reinforcement while learning it according to their treatment assignment. Then students were isolated for the specified amount of time. There were 6 students randomly assigned to each of the six treatment groups. The response is a score measuring the student's memory for the learned paragraph. The data are contained in the file "paragraph.txt" with the first column indicating the level of reinforcement (none or verbal), the second column indicating the isolation time (20, 40, 60), and the third column giving the observed memory score.

   (a) Plot the observations, putting response on the Y axis and isolation time on the X axis. On this plot indicate the location of the mean response for each treatment combination. Connect the means for all treatments with no reinforcement. Repeat for all treatments with verbal reinforcement. You do not need to include this graph in your answers. Based on the graph what significant effects do you expect to find? Explain your answer.

   (b) Obtain the analysis of variance table for the two-way factorial model. Report the p-values.

   (c) Consider the analysis as a one-way ANOVA using 6 treatments. This is equivalent to fitting the cell means model. Construct the two-way factorial ANOVA table using contrasts among the 6 treatments. Report the SS associated with each contrast, then use these SS to calculate the SS for each effect in part 2b

   (d) Check for constant variance and outliers using the residuals. Do you see any concerns? Is the assumption of independence reasonable? Explain why or why not for both questions.

(e) Since the interaction is significant, the researchers ask you to test the effects of reinforcement (none or verbal) at each isolation time. A test is sufficient. You don't need to compute estimates and standard errors. Report the p-values for each comparison.

(f) Summarize your results. How do reinforcement and isolation time effect memory? Also, describe the population for which you think these results are relevant.

3. **Two-factor ANOVA again** The data in alpine1.txt come from some of the research I did for my PhD in ecology. I was interested in the distribution of plants in alpine tundra (grassland above treeline (ca 12,000 ft, 3600 m) in the Colorado Rockies). In one part of my work, I excavated small individuals of two species (Potentilla nivea, which is a small plant that naturally occurs only in dry ridgetops, and Potentilla gracilis, which is a larger plant that naturally occurs only in moist sites) and transplanted them to new sites (Dry, Mid, Wet)

   If you want pictures, Cumberland Pass is shown here:
   http://salidastagestart.com/wp-content/uploads/2011/02/Jeep-at-Cumberland-Pass.jpg The dry site is right by the jeep, the mid site is the very right of the picture and the wet site is the at the bottom of this hill. to the left of the ATV's. My ecological questions were:

   (a) Did the transplant site have any effect on the growth of Potentilla, averaged over the two species? If so, what is the nature of those differences.

   (b) Do the two species respond differently to site, i.e. did P. nivea grow better than P. gracilis at one site but about the same (or worse) at other sites. If so, what is the nature of those differences.

   The two species are naturally different sizes, so the comparison between species averaged over sites is not very interesting.

   I started out with 10 reps per species per site, but some seedlings died, so the number that I could harvest three years later varied between 5 and 10. The response I measured was the dry mass of this year's growth (in mg). Because the within-group variance varied considerably between groups, I log transformed the biomass. This has already been done for you. The variables in the data file are: Experiment, Species, Site, and log biomass. You should ignore Experiment for this problem, since it has only one value in the alpine1 data set.

   (a) Report the F statistics and p-values for the appropriate F tests that answer the biological questions:

      i. Averaged over the two species, did site affect the growth?

      ii. Did site have similar effects on the growth of each species?

   (b) Plot the cell means (for each combination of species and site). Use this plot and the test results to write sentences answering my two biological questions. Your answers should be sentences suitable for the results section of a research paper Hint: by 'nature of the pattern', I am interested in the direction, magnitude, and pattern of the effects.

   (c) There are two ways to analyse each species separately. One is to separate the data into two groups, one for each species, and do separate analyses for each species. The other is to analyze all the data together and use contrasts to examine differences among sites within each species. Will these two approaches give the same differences in means? The same p-values? Explain.

   (d) How should you decide which approach in part 3c is more appropriate?

   (e) Analyze the effect of site separately in each species using the approach you believe most appropriate.

   (f) Treating dead plants as missing assumes that they died for reasons unrelated to the treatments. An alternative is to treat dead plants as plants with a biomass of 0. This is especially appropriate if you were interested in total biomass produced in an area. Zero is a problem with log transformations (log(0) is a big problem). Replace logbiom of the dead plants with the value 2 and reanalyze the data. Are the conclusions similar?

R hint: The alpine1 data set contains 10 lines for each treatment. The missing plants are explicitly indicated by a missing logbiom value. These missing values can be replaced by any other value by y[is.na(y)] <- 2