STATISTICS 500 – Fall 2009 – Homework 12, corrected

on campus Friday, 11 Dec 2009, in lecture (11 am)
or by e-mailto Chuanlong, dclong@iastate.edu, no later than noon.

off campus Monday, 14 Dec 2009, by 4 pm to Nicole Rembert, email: rembeall@iastate.edu or FAX: 515-294-4040 (please include cover page with Stat 500 / Nicole Rembert).

1. **Nested designs** A production engineer is studying the effects of machine model (factor A) and machine (factor B) on the output in a pop (soda, outside the Midwest) bottling plant. The study compared three models of bottling machine, using four machines of each model. We can imagine that model is randomly "assigned" to the machine. The production of each machine ((# cases produced per day) was recorded on five successive days. There are five responses for each machine.

   The data are given below and in bottle.txt on the class web site. Bottle.tex contains the data in a condensed format with one line per model and machine with 5 values (one per day) on the line. To analyze the data, you need to create one line per day (a total of 60 observations).

Model i:	1				2				3			
Machine j:	1	2	3	4	1	2	3	4	1	2	3	4
Day $k=1$:	65	68	56	45	74	69	52	73	69	63	81	67
$k=2$:	58	62	65	56	81	76	56	78	83	70	72	79
$k=3$:	63	75	58	54	76	80	62	83	74	72	73	73
$k=4$:	57	64	70	48	80	78	58	75	78	68	76	77
$k=5$:	66	70	64	60	68	73	51	76	80	75	70	71

   (a) You are most interested in comparing the mean performance of the three models of bottling machine. A student analyzes the data and shows you the following ANOVA table and F test of Ho: $\mu_1 = \mu_2 = \mu_3$:

Source	d.f.	MS	F	p-value
model	2	847.82	14.2	< 0.0001
error	57	59.74		
c.total	59			

   You ask the student if this is the correct analysis and he tells you, "Yes, because there are no blocks in the design". Do you agree? If this isn't the correct analysis, what would be a more appropriate analysis?

   (b) Using an appropriate model for the analysis, estimate the difference between models 1 and 2, calculate the s.e. of that difference and test whether that difference equals 0.

2. **Design of a microarray experiment** The following problem is motivated by a recent microarray experiment. Fungal infections of crop plants are a common problem in the midwest where the summers are warm and humid. There is often some genetic control of which fungi infect which crop plants. This is a study of 2 genetic isolates of one fungus species (A and B) and 3 genetic isolates of barley (1,2, and 3). Barley genotypes 1 and 3 are suspected to be sensitive to fungus B and resistant to fungus A while Barley genotype 2 is sensitive to fungus A and resistant to fungus B.

   This experiment considered all 6 combinations of fungus genotype and barley genotype. A tray was randomly assigned to one of the six treatments. The appropriate barley genotype was planted in each flat then innoculated with the appropriate fungus. These trays were then grown in a growth chamber. The response is the log-transformed expression level of a particular gene thought to be involved in resistance. There are three replicates of each treatment. There is one response for each flat of plants (here I'm simplifying the problem

greatly). Because fungi spread easily, you can't have two different fungal genotypes in the same growth chamber. Here we consider four possible designs and their analysis.

You spend a long time discussing potential sources of variation with the experimenters. You decide that:
there is considerable random variability between growth chambers,
variation between the growth chambers is not consistent over time. That is two repetitions of the experiment in the same growth chamber is just as variable as runing the experiment in different growth chambers.
and, repetitions of the study differ in many uncontrollable ways.
As a consequence of these discussions, you decide to block on repetitions on the study (if a study is repeated more than one), but not block on growth chambers.

For each of the following ways of conducting the experiment, write out the skeleton ANOVA and indicate the appropriate error term for the F test of each effect (barley, fungus and the interaction).

(a) If you had 18 growth chambers, you could randomly assign each of the six treatments to growth chambers (3 chambers per treatment).

(b) If you had 6 growth chambers, you could randomly assign each of the six treatments to a growth chamber (1 chamber per treatment). After you collect the data, you clean the growth chambers and repeat the study with a new randomization of treatment to chamber. You repeat once again, for a total of 3 repetitions.
Hint: think about repetitions as blocks.

(c) If you only had 2 growth chambers (the case for the actual study), you could assign one chamber to fungus A and one to fungus B. Three flats (one of each barley genotype) are grown in chamber 1 and three more are grown in chamber 2. The entire study is repeated a total of 3 times, as in the previous design.

(d) One disadvantage of the design in part 2c is the time it requires. Your professor proposes that you make 18 flats of plants (6 of each barley genotype). One growth chamber is randomly assigned to fungus A; the other chamber is assigned to fungus B. 9 flats (3 of each barley genotype) are placed in growth chamber 1 (Fungus A); the other 9 flats are placed in growth chamber 2.

3. **Analysis of the barley data.**

The barley experiment was actually done using design 2c. I don't have the real data so I have manufactured some. It is in barley.txt on the web site. The response is logexpr, the log transformed expression level. You do not need to worry about evaluating assumptions.

(a) Test the hypotheses of:
1) no difference between fungi, averaged over barley genotypes
2) no difference between barley genotypes, averaged over fungi
3) no interaction between fungi and barley genotypes

(b) Estimate the following differences and their standard errors:
1) between fungus A and fungus B in barley genotype 1
2) between barley genotypes 1 and 2 for fungus A
3) between barley genotype 1 fungus A and barley genotype 2 and fungus B.

(c) All three estimates in the previous part are differences between cell means. Explain why estimate 2) has a different s.e.

4. **Multi-location studies** The data in stream.txt are made-up data based on the LINX II experiment. The context is Nitrogen transport from agricultural fields in the upper Midwest down the Mississippi River to the Gulf of Mexico, where excess N is having severe environmental effects. This experiment studied whether streams serve simply as pipes transporting Nitrogen downstream, or whether biological processes in streams retain some (perhaps all) of the Nitrogen instead of moving it downstream. The LINX experiment is a multi-location comparison of three types of streams (undisturbed, agricultural, and urban). At one location, researchers studied 9 streams; 3 were undisturbed, 3 were agricultural, and 3 were urban. This basic experiment was repeated at 8 different locations scattered around the US and Puerto Rico. There are a total of 72 observations (9 streams at each of 8 locations). The response is Ammonium (NH_4 uptake distance; this was measured on each stream. There are a lot of details of measurement and methodology that are irrelevant for this problem. For interpretation, it may help to know that a large uptake distance means that the stream is acting like a pipe and just moving N downstream. A short uptake distance means that a stream is retaining N.

   (a) The investigators are interested in comparing the three types of streams (treatments), while accounting for possible differences between the eight locations and the possible inconsistency of treatment effects across the 8 locations. Write out an appropriate skeleton ANOVA table for this study, showing sources of variation and d.f.

   (b) If you are interested in broad sense inference about the differences between streams relative to their consistency across locations, which terms in the ANOVA table should be considered fixed and which should be considered random?

   (c) Test whether the differences between stream types are large relative to the consistency of treatments across locations. Report the appropriate F statistic and p-value.

   (d) The previous analysis was considered discouraging, so the investigators want to consider inferences for only these 8 locations. Test the differences between stream types are large relative to the variability between streams at a location. Report the appropriate F statistic and p-value.

   (e) If you examine residuals (you don't need to), you realize that a log transformation is needed to equalize error variances. Inspection of the location*treatment means indicates that, on the log scale, the differences among treatments are more consistent across locations (again, you don't need to draw the plot, unless you want to check what I say). Log transform the responses and test whether the differences between stream types are large relative to the consistency of treatments across locations. Report the appropriate F statistic and p-value.

   (f) The investigators are interested in inferences about land use effects on NH_4 uptake distance for streams throughout the US. (Note: they know that this is an observational study, so will word their conclusions very carefully to avoid implying a cause/effect relationship). Is broad sense or narrow sense inference more appropriate? Is reporting log-transformed responses or raw data more appropriate? Explain both answers.