

Homework 12, Handed out: 28 November 2007, Corrected version for off campus students
 Due: on campus, Wednesday, 5 December 2007, by 4 pm to Norma Elwick in 115 Snedecor.
 off campus, Monday, 8 December 2007, by 4 pm to Ying Shi

1. **One-factor random effects:** In a pig breeding study two offspring from each of ten litters were measured for average daily weight gain (ADWG). The individual pig measurements can be thought of as having two pieces,
 pig gain = litter effect + individual pig effect.

Litter	ADWGs	Mean	Litter	ADWGs	Mean
1	2.76 2.38	2.57	6	2.72 2.74	2.73
2	2.58 2.94	2.76	7	2.87 2.47	2.67
3	2.28 2.22	2.25	8	2.31 2.23	2.27
4	3.01 2.61	2.81	9	2.74 2.56	2.65
5	2.36 2.72	2.54	10	2.50 2.48	2.49

- (a) Complete the d.f. in the analysis of variance table. You should also be able to compute the SS from the data, but I've saved you the trouble here.

Source	d.f.	SS
litters	?	0.6705
pigs within litters	?	0.3834
total	?	1.0539

- (b) The problem gives an informal model statement (pig gain = litter effect + indiv effect). Write this as a formal model statement, specifying the assumptions made about each term.
- (c) Obtain point estimates of the individual variation variance component (the error variance) and the variance of litter effects. HINT: use the lecture notes to figure out the expected mean squares, then find the method of moments estimates of the variances.
- (d) Suppose I intend to select a pig at random from the population and measure its average daily weight gain. According to the model what is the variance associate with this quantity. What fraction of this variance is due to genetic factors (i.e., litter)?
- (e) Suppose I intend to design an experiment where the treatments are assigned to mothers (i.e. litters), but the response is observed on pigs. Many teratology (effects of drugs or other chemicals while critters are in utero) studies are designed this way. The response of interest is a treatment mean, averaged over r litters and n pigs. The investigators can use one of three designs:
- r=2 litters, 4 pigs per litter
 - r=4 litters, 2 pigs per litter
 - r=8 litters, 1 pig per litter

Assume that the variance components estimated in part c are appropriate for this new study. Calculate the s.e. of the treatment mean for each of the three designs. Which design gives them the most precise estimate of the treatment mean?

2. **Nested designs** (based loosely from a problem in the text): A production engineer is studying the effects of machine model (factor A) and machine operator (factor B) on the output in a bottling plant. Three bottling machines were used, each a different model. Twelve operators were employed with four operators assigned to a machine for a six-hour shift. The number of cases produced (avg per hour) is the response. There are five responses for each operator (one per day for one week). Consider days as a nested effect, nested in operator and machine. This is reasonable when days correspond to measurements and there is no consistent effect of Monday (day 1) on all operators. If there is, we're going to ignore it here, so we'll treat days as nested.

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 machine and operator 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). You can use code similar to that in `ratweight.sas` to reformat the data. Other languages (e.g. SPSS) provide menu options to do this. If all else fails, you can edit the data file to create 60 lines.

Machine i :	1				2				3			
Operator 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

- Is the 'operator' factor crossed or nested with the 'machine' factor? Explain.
- If conclusions are to be made about the average productivity of these three specific machines and these twelve specific operators, would you consider machines as a fixed factor or a random factor? Would you consider operators as a fixed factor or a random factor?
- If these are three of many bottling machines and 12 of many operators, and conclusions are to be made about the variability between all machines and the variability between all operators, what factors are random and what are fixed?
- Assume that machines and operators are random. Write down the model for the response Y_{ijk} , the production of operator j on machine i on day k . Explain what each symbol represents. Your model should include 3 variance components.
- Estimate the variance components for machines ($\sigma^2_{machines}$), for operators within machines ($\sigma^2_{operator(machine)}$) and for days within operators and machines (σ^2_{days}).
- Test the hypothesis that $\sigma^2_{machines} = 0$; test the hypothesis that $\sigma^2_{operator(machine)} = 0$. Report your test statistic and a p-value.
- Do you have any concerns about estimating $\sigma^2_{machines}$ in this study? What about $\sigma^2_{operator(machine)}$? Hint: think about the sample sizes used to estimate each variance component.

3. **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 inoculated 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 running 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 3c 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.

4. Analysis of the barley data.

The barley experiment was actually done using design 3c. 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.

5. Cracks in concrete.

Problem 3 on last week's HW (HW 11) described a study that compared connecting rods for joining sections of freeways. To repeat from the problem description from last week:

“They randomly assigned the three connecting rods to 1 mile segments, with each connecting rod type being used in 3 segments. The road contractors then built the freeway using the assigned connecting rods in the appropriate 1 mile road segments. After 5 years, the average crack width is measured in 6 places per freeway segment. The 6 locations are the outside, center, and inside wheel track in both the slow and the fast lane. ”

Last week I told you to analyze this as a 3 way factorial. This week, we reconsider that choice.

- (a) Describe a way of randomizing treatments (rod types) for which last weeks analysis (3 way factorial, completely randomized design) is appropriate.
- (b) For simplicity in thinking about the study, consider the 6 places (outside, center and inside wheel track in both the slow and fast lane) to be randomly assigned within each 1 mile freeway segment. If you make this simplifying assumption, is there one or two sizes of experimental units in the study? Identify each of the experimental units and the treatment factor(s) assigned to it.
- (c) Is the analysis used last week appropriate? If not, write out an appropriate skeleton ANOVA table indicating sources of variation and d.f.
- (d) Analyze the log crack width data as a split plot. Report the ANOVA table with F statistics and p-values.