Stat 500 – Homework 11 answers – 2009

My notes are marked by •

1. Power of ANOVA tests

   (a) The standard errors of each contrast are:
   1) $\sqrt{2/3}\sigma/\sqrt{n}$, 2) $\sqrt{2}\sigma/\sqrt{n}$, 3) $\sigma/\sqrt{n}$, 4) $2\sigma/\sqrt{n}$,
   The s.e. for contrast 4 (the interaction) is the largest.

   (b) When all else is equal, the estimate with the largest s.e. leads to the test with the smallest power and vice-versa. The test of interaction is the least powerful; the test of the main effect of reinforcement (contrast 1) is the most powerful.

   (c) The general equation is $\delta = (t_{1-\alpha/2} + t_\beta)s.e. = (t_{1-\alpha/2} + t_\beta)2\sigma/\sqrt{n}$. Solving for n gives you $n \geq 2^2(t_{1-\alpha/2} + t_\beta)^2(\sigma/\delta)^2$. Starting with $t_{1-\alpha/2} = 2$ and $t_\beta = 0.85$ gives $n \geq 4 * 2.85^2 * (16/25) = 20.7$, i.e. use $n = 21$ replicates per treatment.

   The above design has 6*(21-1) = 120 error d.f. so better quantiles are $t_{1-\alpha/2} = 1.98$ and $t_\beta = 0.8446$. Using these gives $n \geq 4 * 2.825^2 * (16/25) = 20.4$. Use $n = 21$ replicates per treatment.

2. **Three-factor factorial**
   My SAS code:

```
data freeway;
infile 'C:\500\Homework 11\freeway.txt' firstobs=2;
input rod segment lane location width;
logwidth=log(width);
run;

proc glm data=freeway;
class rod lane location;
model width=rod lane location rod*lane rod*location
   lane*location rod*lane*location;
run;

proc glm data=freeway;
class rod lane location;
model logwidth=rod lane location rod*lane rod*location
```

```
   lane*location rod*lane*location;
estimate ’rod1-rod2’ rod  1 -1 0;
estimate ’rod1-rod3’ rod 1 0 -1;
estimate ’rod2-rod3’ rod 0 1 -1;
run;
```

(a)

```
Source             DF Type III SS  Mean Squ.  F Value    Pr > F
rod                 2    539.036   269.518     53.12    <.0001
lane                1    118.163   118.163     23.29    <.0001
location            2    209.201   104.600     20.62    <.0001
rod*lane            2     39.795    19.897      3.92    0.0288
rod*location        4     71.964    17.991      3.55    0.0154
lane*location       2     23.118    11.559      2.28    0.1170
rod*lane*location  4      8.969     2.242      0.44    0.7774
```

The interaction effects of rod*lane and rod*location are significant at the 0.05 level. Hence the differences between rod types are not the same in all locations and lanes.

(b)

```
Source              DF Type III SS  Mean Squ. F Value    Pr > F
rod                  2  21.6953      10.8476    85.84    <.0001
lane                 1   4.7811       4.7811    37.83    <.0001
location             2   8.6820       4.3410    34.35    <.0001
rod*lane             2   0.0023       0.0011     0.01    0.9907
rod*location         4   0.0260       0.0065     0.05    0.9948
lane*location        2   0.0191       0.0095     0.08    0.9272
rod*lane*location   4   0.0180       0.0045     0.04    0.9975
```

From the ANOVA table, none of the two-way interactions is significant. And, the three-way interaction is not significant. These results show no evidence that the log-scale differences between rod types change across locations and lanes. Another way of wording this is: These results are consistent with the differences between rod types on the log scale being the same in all locations and lanes.

These conclusions are carefully worded to avoid saying that the differences are the same. They are very likely to be not the same, even on the log scale. It is just that we don’t have enough data to know if there is a very small difference. However, the interactions are much smaller than the other effects so I will go ahead and report the results as if there was no interaction, i.e. the differences between rod types are the same in all locations and lanes.

(c) To construct estimates, we use the log transformed data. Because the two and three way interactions are not significant, which implies that the differences between rod types do not depend on lanes or locations. Consequently, we estimate the following differences between rod types averaged across all lanes and locations.

```
                                   Standard
Parameter        Estimate            Error      t Value      Pr > |t|
rod1-rod2           -1.55             0.12       -13.10        <.0001
rod1-rod3           -0.76             0.12        -6.42        <.0001
rod2-rod3            0.79             0.12         6.68        <.0001
```

3. Average Daily Weight Gain in pigs

(a) The d.f. are litters = (10-1) = 9, pigs(litters) = (2-1)*10 = 10. Total = 20 - 1 = 19

(b) If Y_{ij} is the ADWG from the j'th pig in litter i, then one way to write the model is:

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2) \\ \mu_i &\sim N(0, \sigma^2_\mu). \end{aligned}$$

You could also write an effects model:

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2) \\ \tau_i &\sim N(0, \sigma^2_\mu). \end{aligned}$$

The only difference is that the mean of the pig effects is separated out as a distinct parameter.

(c) The ANOVA estimates of variance components are obtained by equating expected mean squares ang observed mean squares:
Litters: $\sigma^2 + 2\sigma^2_{mu} = 0.6705/9 = 0.0745$
Error: $\sigma^2 = 0.3834/10 = 0.03834$
The ANOVA estimate of σ^2 is the MSE = .03834. The ANOVA estimate of σ^2_μ is $\sigma^2_\mu = (MS_{\text{litters}} - MS_{\text{pigs in litters}})/(\#\text{pigsinlitter}) = (.0745 - .0383)/2 = .0181$

(d) Under the model in b, the variance of Y_{ij} is the sum of the two variance components, i.e. 0.0564. The fraction of this that is due to genetic factors is .0181 /(.0181+.0383)=.32.

• You may notice that this fraction is also the intraclass correlation, also known as the heritability when the classes are genetic groups.

(e) The variance of the mean for r litters and n pigs per litter is: $\sigma^2_\mu/r + \sigma^2/(rn)$. For the sample sizes given this is:

r	n	Var mean
2	4	0.01385
4	2	0.00932
8	1	0.00706

r=8, n=1 is the most precise estimate.

• While r=8, n=1 gives the most precise estimate of the mean, it does not allow you to estimate the variance components. If you want both, you need to compromise, e.g. use r=4, n=2.

(f) The ANOVA (type3) estimates are litters: $\hat{\sigma}^2_\mu = -0.0799$ and pigs w/i litters: $\hat{\sigma}^2 = 1.21$. The REML estimates are litters: $\hat{\sigma}^2_\mu = 0$ and pigs w/i litters: $\hat{\sigma}^2 = 1.08$.

BONUS: The REML estimates are not independent of each other. Changing one changes the other. In particular, the sum of the two is close to (but rarely exactly equal) to the sample variance of the observations. Increasing one estimate (the litter variance component from -0.0799 to 0) forces the other to change (from 1.21 to 1.08). (Aside to the bonus: my explanation for why the sum and sample variance are not equal is that the sample variance of the observations is a biased estimator when observations are correlated).

(g) Inspection of the data (or of a residual plot) identifies a very unusual value (pig 5, value of 7.22).

• In fact that value was an error in the data file. When corrected (to 2.72), you get the SS reported at the beginning of the problem.