STAT 8200 — Design of Experiments for Research Workers
Lab – Due Friday, Sept. 6, 2013

Example:

Zelazo *et al.* (1972) describe an experiment in which the age at which children first walked is studied. Newborn children were randomly placed into one of four treatment groups: active exercise, passive exercise, no exercise, or an 8-week control group. Infants in the active-exercise group received walking and placing stimulation four times a day for 8 weeks, infants in the passive-exercise group received an equal amount of gross motor stimulation, infants in the no-exercise group were tested along with the first two groups at weekly intervals, and the 8-week control group consisted of infants observed only at 8 weeks to control for possible effects of repeated examination. The resulting ages at first walking (in months) appear below.

Active Group	Passive Group	No-exercise Group	8-week Control Group
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	

These data also are contained in the file walking.sas which is a SAS program to analyze the data. Copy this file from the course website to your flash drive and open it in SAS. Uncomment the line that reads

```
*ods pdf file="mypath/myfile.pdf";
```

and change the path so that it points to a folder on your flash drive. Also change the filename to walking.pdf (instead of myfile.pdf) and uncomment the last line in the program, which reads

```
*odf pdf close;
```

Then run the program and examine its output.

This experiment is an example of a one-way layout. There is a single treatment factor with four levels (the four groups) in a completely randomized design. An appropriate statistical model for these data is

$$y_{ij} = \mu_i + e_{ij}$$

where y_{ij} is the age at first walking for the j^{th} child in the i^{th} treatment group. Here, $i = 1, 2, 3, 4$ and $j = 1, \ldots, n_i$ where $n_1 = n_2 = n_3 = 6$ and $n_4 = 5$. This model allows for there to be four separate population means μ_i, $i = 1, \ldots, 4$, and the model is an example of a cell-means model.

Of interest here is whether the various treatments affect age at first walking. To address this question, it is appropriate to test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ in terms of the parameters in the above cell-means model.

An equivalent model for the data which uses a different parameterization is

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \text{where } \textstyle\sum_{i=1}^{4} \alpha_i = 0,$$

where now μ can be interpreted as the grand mean across all treatment groups and $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are treatment effects which allow the treatment means to be greater than or less than the grand mean μ. In terms of this effects model, the mean for the i^{th} treatment group is equal to $\mu + \alpha_i$ or $\mu_i = \mu + \alpha_i$. The null hypothesis above from the means model translates to the equivalent hypothesis $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ in terms of the fixed effects model.

In walking.sas PROC GLM is used to analyze these data. The data are first read in to data set walking. Notice the input statement defines two variables, age and group. the variable group is followed by a '\$' to identify it as a character variable rather than a numeric variable. In PROC GLM, the CLASS statement defines group to be a factor (or CLASSification variable). In the model statement age appears on the left-hand side of the equals sign to identify age as the response. On the right hand side of the equals sign appears any variables for which effects are desired. In this case, the variable group appears to define the α_i's. The grand mean μ from the effects model, and the error term e_{ij} are implied automatically. The CONTRAST statements request contrasts to compare the control mean with the other three group means, the no-exercise group mean with the active and passive means, and a contrast comparing the active and passive means. Notice that the ORDER=DATA option on the PROC GLM statement requests that the levels of the factors (in this case group) are ordered in the same way as they appear in the data set. If this option were not included, SAS would simply alphabetize the levels of group so that, for example, the third contrast would actually compare active and control (the first two levels of group in an alphabetical sense). The ESTIMATE statment request an estimate of the contrast between the active and passive treatment means. Finally, the LSMEANS statement requests means for each level of the variable group. The option CL (options appear after a slash), requests 95% confidence intervals for the means requested on the LSMEANS statement.

From the output we can obtain several results easily:

1) The null hypothesis $H_0 : \alpha_1 = \cdots \alpha_4 = 0$ is tested with the F test on the group variable. $F = 2.14$ has a $p-$value of .1285, so using the conventional significance level of 0.05 we do not reject H_0.

Conclusion: There is insufficient evidence here to conclude that the mean age at first walking differs across these four treatments.

2) Not surprisingly, none of the more specific comparisons performed by the CONTRAST statements are significant either.

3) We can estimate the group means $\mu_1, \ldots, \mu_4$ and the grand mean μ. We estimate these population quantities with the corresponding sample quantities: $\bar{y}_{1\cdot} = 10.125$, $\bar{y}_{2\cdot} = 12.35$, $\bar{y}_{3\cdot} = 11.708$, $\bar{y}_{4\cdot} = 11.375$, and $\bar{y}_{\cdot\cdot} = 11.348$, respectively. These group means are on p.3 of the output and the grand mean is listed on p.2.

4) The difference between the active and passive treatment means is estimated on the bottom of p.2 as -1.25. This value can easily be confirmed by examining the means listed on p.3.

This page intentionally blank.

STAT 8200 — Lab 2

Name:____

Exercise:

A study of river contamination by polychlorinated biphenyls (PCBs), a hazardous chemical used in the manufacture of large electrical transformers and capacitors, resulted in the following PCB concentrations (parts per million) in samples of fish from five rivers:

River 1	River 2	River 3	River 4	River 5
2	4	12	7	13
3	6	9	5	9
1	3	11	5	15
5	5	8	9	10
	7			11
				7

These data are contained in the file pcb.dat on the course website. Copy this file to your flash drive and write and run a SAS (or R) program to answer the following questions.

1. What is the F statistic for testing the null hypothesis of equal means across the five rivers? Report the corresponding $p-$value and state the appropriate conclusion.

2. Estimate the five treatment (river) means.

3. Use the ESTIMATE statement to estimate the difference between rivers 2 and 5. What is the standard error of this estimate?