STAT 705: Exam I

***Whammo!* data analysis** (25 points each part)

Consider data on the sales of the Energy Drink *Whammo!* The makers of *Whammo!* set up a marketing experiment to determine the best way to sell their product. Fifteen stores that carry *Whammo!* spread across the state of South Carolina were randomly picked to participate. For one week, the baseline sales in cases were recorded as `time1` using the existing sales strategy, which is a lifesize cardboard cutout of a Tazmanian Devil drinking *Whammo!* next to the refridgerated section in the back of the store where *Whammo!* is kept. After the first week, one of three treatments were implemented at east store. Treatment 1 is the existing display next to the refridgerated section. Treatment 2 is a new coardboard display of a medical doctor prescribing *Whammo!* to treat lethargy, in the usual refridgerated section. Treatment 3 is the same cardboard cutout of the medical doctor, but located at the cashier's next to a small refridgerator filled only with *Whammo!* The sales were recorded for the second week as `time2`. You will analyze these data using three increasingly sophisticated models. Use `proc glimmix` for all model fits if using SAS, except where noted below. Perform all significance tests including FER at the 95% (confidence intervals) or 5% (significance testing) levels.

(a) Using the data `sales1` in the SAS code, fit a simple one-way ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ using the difference in `time2` and `time1` as the outcome Y_{ij} with different levels of `treat`: α_1, α_2, and α_3. Are there significant differences according to the F-test? Use Tukey to report a lines plot showing significant pairwise differences and comment.

(b) Now use the data `sales1` to fit the ANCOVA model $Y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \epsilon_{ij}$ to `time2` (Y_{ij}) only, keeping `treat` (α_i) but adding `time1` as a continuous concommitant variable x_{ij}, i.e. adjusting for baseline `time1`. Again compare treatment effects as in the previous ANOVA. How does adjusting for baseline sales affect the analysis? Is precision for treatment inference improved in any way? How is the ANOVA model in part (a) a special case of the ANCOVA model in part (b)? Can this be tested?

(c) Finally use the `sales2` data to examine a profile plot (SAS code included) and fit a mixed model with random store effects ρ_i, e.g. $Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$ where now $i = 1, \ldots, 15$, $j = 1, 2, 3$, and $k = 1, 2$. The model is repeated measures over time but not treatment. First obtain the averaged interaction plot for `treat` vs. `time` (i.e. time is x-axis), perhaps using `proc glm` with `plots=all`. Comment on the plot: Is there a treatment by time interaction here? Does this make intuitive sense? Now fit the appropriate random (store) effects model. Examine all pairwise differences using Tukey *sliced at the two time points*; use the `slice` command to perform this. Are there significant differences at baseline? How about for the 2nd week after the treatments were in effect? How has the precision for examining the paired difference between treatments 2 and 3 changed from ANOVA to ANCOVA to the repeated measures random effects model? Formally test $H_0 : \sigma_\rho = 0$ using a `covtest` statement.

Be sure to comment on the appropriateness of each model using standard diagnostics. What would you recommend to the makers of *Whammo!* ?

Short answer (5 points each)

1. In a one-way ANOVA with $r = 5$ treatments, the following contrasts are to be estimated with simultaneous confidence intervals: $L_1 = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{3}(\mu_3 + \mu_4 + \mu_5)$ and $L_2 = \mu_2 - \mu_1$. Which method is *likely* to be the most appropriate: Bonferroni, Scheffe, or Tukey?

2. In a one-way ANOVA boxplots of each group show extreme right-skew and many outliers for larger observations. What are two remedial actions for such data?

3. When is it appropriate to treat block effects, e.g. ρ_i, as random? Give an example.

4. Give an example of a completely randomized design. Now discuss how the completely randomized design can be modified to a randomized complete block design. Why does one introduce blocking variables?

5. In a three-factor ANOVA with factors named A, B, and C, the following model is fit:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijkl}.$$

   For these data $a = 3$, $b = 2$, and $c = 2$. What does a `lsmeans A;` statement estimate *in terms of model parameters*? How about `lsmestimate A*B -1 1 0 0 0 0, 0 0 -1 1 0 0, 0 0 0 0 -1 1;`?

```
* Whammo! data for parts (a) and (b);
data sales1;
 input treat store time1 time2; sales=time2-time1;
 datalines;
 1     1  69.7  72.2
 1     2  74.3  72.2
 1     3  74.4  71.7
 1     4  77.7  73.7
 1     5  72.7  74.4
 2     1  76.3  83.7
 2     2  71.9  83.8
 2     3  71.7  87.2
 2     4  72.7  89.6
 2     5  66.4  78.3
 3     1  74.5  92.5
 3     2  71.4  89.7
 3     3  69.4  86.4
 3     4  73.9  86.2
 3     5  77.1  90.4
;

* data for repeated measures over time, part (c);
data sales2; set sales1;
sales=time1; time=1; store=_n_; output;
sales=time2; time=2; store=_n_; output;
drop time1 time2;
proc print; run;

proc sgpanel;
panelby treat / rows=1 columns=3;
series x=time y=sales / group=store;
run;
```