Name: _____

1 Objective: One Way ANOVA Data Reading

In this lab we'll read the data and then put it into anova form for analysis.

2 Key Words

`proc glm`, ANOVA, multiple comparison.

3 Part I

These data are taken from D.J. Hand, et. al., Small Data Sets. They give the number of fatalities in coal mining disasters in Britain which claimed ten or more lives. Column 1 gives the number of deaths at various accidents in the decade 1860-69; the next 3 columns have data for the decades beginning in 1870, 80, and 90; the last 3 columns have data for the double decades beginning in 1900, 20, and 40. Data file is `mining.dat`.

Problem: Are the means the same over the various time periods?

[A] Open the data in some word processor and look at its structure: 7 columns of varying lengths. How can we read the data and then put it into anova form: two long columns, one giving the time category and the other the number of deaths?

We have two ways to restructure the data so that SAS can handle it.

[1] Put data into XL and insert columns which specify the decade for the next column (see the data set `mining_aov.dat`).

[2] Put data into XL and fill all missing values with 0, so that all columns are the same length (see the data set `mining_0.dat`).

[3] Read the file `mining_aov.dat` with SAS in such a way that it has anova structure. HINT: `input t d @ @;`

[4] Read the file `mining_0.dat` into SAS as 7 columns, `c1-c7`.
Use another SAS data step to put it into anova form. HINT: if `c[k] ~= 0` then do `d=c[k]; output; end;`

[5] Once you have the data in the proper form, sort by time period.

[6] Find means and std dev's of various time periods.
TURN IN mean and std dev for first and last time periods.

[7] Get box plots.

[8] Test if means are the same across time. TURN IN p-value:

[9] ANOVA procedures in [8] assume normal errors with the same variance. Do you think those conditions are true for these data? Why? Explain in a short paragraph.

4 Part II

The data are also taken from D.J. Hand : "the data are from a foster feeding experiment with rat mothers and litters of four different genotypes: A, B, I, and J. The measurement is litter weight (in g.) after a trial feeding period."

The first column is weight (`wt`), the second genotype of litter (`gl`) and the last genotype of mother (`qm`).

[B] Read in the data file 'C:\...\fosterall.dat'

[1] Find cell means, `"by gl gm"`.

[2] Plot cell `wtmean` vs `gl` with symbol given by `gm`.
NOTE: the following code turns the figures into a manageable size:

```
PROC plot data=wtmean;
        plot wtmean*gl = gm / hpos = 60 vpos = 30 vaxis = 45 to 65 by 5;
run;
```

[3] Plot `wtmean*gl = gm` by connecting like symbols (optional: find, mark, label the value of the omitted point) TURN IN.

[4] From your plot do you think there is interaction between `gl` and `gm`? Explain.

[5] Run the `glm` test for [4]. What is the p-value? TURN IN.

[6] Estimate and TURN IN the mean litter weight when the litter type is `a`.
HINT: `estimate 'glt = a' intercept 1 glt 1 0 0 0 gm .25 .25 .25 .25 glt*gm .25 .25 .25 .25 / e;`
LOOK AT OUTPUT TO SEE WHAT `'/e'` DOES.

[7] Return [B1] getting cell means just `"by gl"`. What did you get? TUR IN. The reason it differs from [6] is because cells have different number of observations: the data are unbalanced.