**STAT 6740**

**Final Exam**

**Autumn 2015**

Lake Hartwell is one of the Southeast’s largest and most popular public recreation lakes. Built by the U.S. Army Corps of Engineers between 1955 and 1963, the authorized purposes are flood risk management, water quality, water supply, downstream navigation, hydropower production, fish and wildlife protection, and recreation. Between 1955 and 1977 the Sangamo-Weston capacitor manufacturer released an estimated 400,000 pounds of polychlorinated biphenyls (PCBs) into Town Creek, a tributary of Twelve Mile River, which is a major arm of Lake Hartwell. PCBs are toxins with a broad range of effects, including physiological, immunological, developmental, and reproductive. PCBs degrade very slowly and remain attached to clay and organic sediments for many decades. Lake sediment was sampled in mid-1980’s to determine the location and levels of PCBs.

There are 209 individual PCBs. Based on their chemical structure, the PCBs can be partitioned into 10 groups, known as homologs, based on the number of chlorine molecules attached to the biphenyl structure. While individual PCB concentrations are usually measured, PCB data are often reported as the total PCB concentration within each of the 10 homolog groups.

PCB data from the Lake Hartwell sediment sampling have been placed into two data files in CSV format. The file “Hartwell Concentrations.csv” contain concentration data for the 10 homologs from samples taken at varying depths at several sampling locations (with different numbers of depths per location). The file “Hartwell Depth.csv” contains data showing the upper and lower depth limits of each sediment sample. Both data files contain an ID number for the sample, although the format of the ID number differs between the files. In the concentration data file, the sample ID consists of the label ‘SLN’ followed by a hyphen and the sample location; in the depth data file, the sample ID is the same as the concentration file, but it includes an additional integer that corresponds to the depth category after a second hyphen. This depth integer corresponds to the last character in the column labels for the concentration in the concentration data file. These two data files have been placed on the Stat 6740 Carmen site in the Final Exam folder.

For your final exam, write a SAS® program to perform the following tasks:

1. Read the two datasets into SAS®. For each data file, define a macro variable using %LET statements to identify the location (filename and directory) of the input data files and reference that macro variable during data importation. (This will allow me to replace your %LET statements with my own to identify the location of the datasets when I run your program on my computer.) Use macro variables “CONC” for concentration and “DEP” for depth.
2. Transpose the concentration data so that each depth is in a separate row and each homolog is in a separate column. Merge the transposed concentration data with the depth data, and add labels to each variable in this dataset. (Hint: The FIND function may be useful to identify the location of the second hyphen in the sample ID for the depth data.)
3. Produce a table that contains the following summary information for each combination of homolog and sampling location:
   1. the number of sampling depths,
   2. the minimum and maximum concentrations across all depths,
   3. the mean concentration across all depths; and
   4. upper and lower 95% confidence limits for the mean.

Label all the statistics appropriately.

1. Define shallow samples as all those where the upper depth is less than 25, medium samples as all those where the upper depth is between 25 and 75, and deep samples as all those where the upper depth is greater than or equal 75. Produce two graphic-quality charts, one for Homolog 3 and the other for Homolog 6, that show the mean concentrations for each depth category within each sampling location. Use AXIS and PATTERN statements to define labels and color patterns, and make the patterns change with each depth category. Use a logarithmic (base 10) scale for the mean. Define a format for the depth categories to create labels for the output.
2. Produce graphics-quality scatterplots of homolog concentration for each pair of homologs numbered 3 through 7. Use different plotting symbols or colors for each depth category. Also, for each homolog (3 – 7), plot a histogram of the distribution of the concentration and overlay a normal distribution.
3. Create a univariate summary of the ratio of the concentrations of homologs 4 and 5 (homolog 4 / homolog 5) that includes a test for normality and a histogram with fitted (overlaid) normal and lognormal distributions. Include the results of the normality test in the plot of the histogram. Define the SAS output so that it includes ONLY the histogram (with inset).
4. Create a dataset that contains the maximum concentration across all depths for each homolog and sampling location with a separate observation for each combination of sampling location and homolog. Save this dataset as a permanent SAS dataset, using a %LET statement to identify the location of the data library where the dataset is to be stored (with macro variable OUTDAT).

Your SAS® program should be written using the following guidelines:

1. Name the program “final\_exam\_*yourlastname*.sas”.
2. Place your name, student ID number, and the submission date in a comment at the beginning of the program.
3. Create separate sections of code for each of the tasks using comments to separate and describe the objective of each task/section of code.
4. Add appropriate titles to all tables and figures.
5. Add a right-justified footnote to all outputs that indicates that the data come from the U.S. EPA.

Submit your program in Carmen by 4 pm on Thursday, December 17.