STA 303 H1F / 1002 HF – Winter 2012 – Assignment 1

What factors increase mercury levels in fish?

Due: Monday, February 6 at 14:00

No late assignments will be accepted without a valid reason.

Presentation of solutions is important. It particular, it is inappropriate to hand in pages of SAS output without explanation or interpretation. The only SAS output you need to submit with assignments is relevant plots. Quote relevant numbers from your SAS output as part of your solutions. You don't need to hand in your SAS code.

The Data:

The source of these data is *Statistical Case Studies* by Peck, Haugh, and Goodman. The data are in a text file at

`www.utstat.utoronto.ca/alisong/Teaching/1112/Sta303/assignments.html`.

Human consumption of mercury is known to lead to neurological and physical disorders. A study was carried out in Maine to investigate which characteristics of lakes are associated with higher levels of mercury in the fish in the lakes. Data were collected on 120 lakes. Many other questions were investigated as part of this study, but we will focus on the following two questions:

1. Does the presence of dams affect mercury levels?
2. Do different types of lakes have different mercury levels?

The variables in the dataset are:

- Name of the lake.
- The mercury level in the fish in parts per million.
- An indicator variable which is 0 if there is no functional dam present (so all all water flow is natural) and is 1 if there is a man-made dam in the drainage area of the lake.
- Lake type. Lakes are classified as 1. *oligotrophic* (sustains fish based on its vegetation and oxygen), 2. *eutrophic* (has few fish), or 3. *mesotrophic* (in between oligotrophic and eutrophic).

The column of 1's at the end of the data file can be ignored.

Note that the data file is space-delimited (the default in SAS), so there is no need to specify the delimiter (`dlm`) in your program.

Use SAS to do the analysis for the following questions. Treat your work as an exploratory analysis. In exploratory analyses, Type II errors are typically of more concern than Type I errors so consider results statistically significant if p-values are < 0.10.

1. Create a new variable that combines lake type and whether or not a functional dam is present. Assuming that the indicator variable for presence of a dam is `Dam` and `Type` is a variable for lake type, you can add code like the following in your data step in your SAS program:

```
if (Dam=1 and Type=1) than DamType='OligoDam';
if (Dam=0 and Type=1) than DamType='OligoNoDam';
if (Dam=1 and Type=2) than DamType='EuDam';
if (Dam=0 and Type=2) than DamType='EuNoDam';
if (Dam=1 and Type=3) than DamType='MesoDam';
if (Dam=0 and Type=3) than DamType='MesoNoDam';
```

Construct three sets of side-by-side boxplots: 1. to compare the mercury levels between lakes which have and do not have a functional dam, 2. to compare the mercury levels among the three types of lakes, and 3. to compare mercury levels among the 6 categories of lakes grouped by the combination of their lake type and dam status. Do there appear to be differences?

2. Using the SAS `ttest` procedure, investigate whether or not there is a difference in the mean mercury level between lakes with and without dams. In question 1, you probably noticed a large outlier in the mercury level in the fish from one of the lakes. In order to assess the influence of the outlier, compare the results (means, test statistics, p-values) with the outlier in the data and with the outlier removed. Your answer should be given in practical terms. Some SAS code for producing a SAS dataset with the outlier removed is given at the end of this assignment.

3. Investigate whether or not there is a difference in mean mercury level between the three types of lakes using one-way analysis of variance. In order to assess the influence of the outlier, compare the results (means, test statistics, p-values) with the outlier in the data and with the outlier removed. If there is evidence of differences among the types of lakes, carry out an appropriate analysis to see which types of lakes differ. Your answer should be given in practical terms.

4. Use one-way analysis of variance to investigate whether or not there is a difference in mean mercury level between the six categories of lakes categorized by the combination of their type and dam status. In order to assess the influence of the outlier, compare the results (means, test statistics, p-values) with the outlier in the data and with the outlier removed. If there is evidence of differences among the six categories of lakes, carry out an appropriate analysis to see which differ. Your answer should be given in practical terms.

5. Do you trust the results of the statistical tests carried out in question 4? Assess whether the necessary assumptions of the model hold.

6. Instead of the one-way classification model used in question 4, a two-way analysis of variance model could have been used with dam status, lake type, and their interaction. Do **NOT** use SAS to fit this model. But answer the following questions about it.

   (a) Would the number of predictor variables be the same as in the model used in question 4? Why or why not?

   (b) Would the F-test for the presence of the interaction between dam status and lake type be statistically significant? How do you know from your results of question 4?

7. In each lake, the mercury level was found by combining flesh from a few fish caught in the lake and then analyzing the mixed sample. The number of fish per lake ranged from 2 to 5. Should we be concerned that different lakes had a different number of fish? Why or why not?

Here is some potentially useful SAS code to create a dataset without the outlier. Assume that the data were read in and called `originaldata` and the name of the lake was stored in a variable called `lake`. This code creates a new SAS dataset called `datawithoutoutlier` with Hodgdon Pond, which is the outlier, removed.

```
data datawithoutoutlier;
   set originaldata;
   if lake ne 'HODGDON.';
```

Marking scheme:
Each of the questions is worth 3 marks (for a total of 21). 3 marks will be awarded for complete, correct answers, or answers with only very minor problems. Good answers that are unclear or have some mistakes or are missing some aspects of the solution will be awarded 2 marks. Poor answers that have some value will be awarded one mark. Note that sometimes an answer awarded 3 marks will not be perfect. You should always look at the solutions.