HW 1

STAT 448 - Advanced Data Analysis

Due: September 8, 2018 12:00:00 AM

Submitting your work to Compass

You are to submit two files for your homework submission.

1. Your SAS program file which should be saved as `HW#_YourNetID.sas`. For example, my file for the HW1 assignment would be `HW1_kinson2.sas`. All program statements and code should be included in one program file.

2. Your Report including all relevant code and output to address the exercises which should be saved as `HW#_YourNetID.pdf`. For example, my file for HW1 would be `HW1_kinson2.pdf`.

You have an unlimited number of submissions, but only the last submission (which contains those two files) will be viewed and graded. To submit, click on the title of the assignment in Compass. The homework questions begin below the line. Be sure to attach the relevant files as dictated in that week's assignment.

Starting SAS program for this assignment

To complete this assignment, you will need to analyze the data sets in the `Program_HW1_Data_Fall2018.sas` file on Compass.

---

1. The `fishy` dataset is a subset of the `sashelp.fish` dataset. In this subset, there are 60 observations and 5 variables recording fish measurements. All fishes were caught from Lake Laengelmavesi. The data are from a paper from 1917 by Juha Puranen, Department of Statistics, University of Helsinki, Finland. Fish weights is the response variable. (10 parts)

Variable Name	**Description**
species	species of fish
weight	weight of fish in grams
length1	length from the nose to the beginning of the tail in centimenters (cm)
length3	length from the nose to the end of the tail in cm
height	maximal height as a percentage of length3

   (a) Describe the relationships that exist among the numeric variables using a scatter plot matrix as your main visualization.

   (b) Describe the relationships that exist among the variables using the correlation analyses (correlation values and correlation tests) appropriate for linear and nonlinear relationships, respectively.

   (c) Describe the relationships that exist among the variables, when grouped by the categorical variable, using the correlation analyses appropriate for linear and nonlinear relationships, respectively.

   (d) Describe and interpret a box plot of the **weight** variable.

   (e) Describe and interpret box plots of the **weight** by **species**.

   (f) Using basic descriptive statistics, histograms, and QQ plots (or probability plots), describe any notable features of the distributions for **length1** and **weight**. Also, determine whether normality would be reasonable for the **length1** and **weight** variables, respectively.

   (g) Using basic descriptive statistics, histograms, and QQ plots (or probability plots), describe any notable features of the distributions for **length3** by **species** and **height** by **species**. Also, determine whether normality would be reasonable for the **length3** by **species** and **height** by **species**, respectively.

   (h) Using statistical evidence, i.e., hypothesis tests and confidence intervals, determine whether it is reasonable to say that the true central **length1** among all fishes is 30 cm.

   (i) Perform a hypothesis test of whether Bream have significantly greater **length3** than Perch, and state your conclusions. Remember that goodness of fit tests can help determine which kind of hypothesis test is appropriate.

   (j) Does using a log transformation of **weight** allow for a two sample t-test of the mean weight to be performed by **species**? Discuss why or why not.