Homework 1

STAT448 - Advanced Data Analysis

Due: Thursday, Sep 14, at 11:00 pm

To complete this assignment, you must use the Iris data set which is an existing SAS data set. We include a brief table of the variables below. Since the data set is part of the SAShelp data sets, you can access is directly in using `data=sashelp.iris` as in

```
proc print data=sashelp.iris ;
run;
```

Variable Name	Description
sepallength	sepal length, in millimeters
sepalwidth	sepal width, in millimeters
petallength	petal length, in millimeters
petalwidth	petal width, in millimeters
species	species of iris

1. **Descriptive Statistics** (5 parts)

   (a) Obtain a single box plot for **sepallength** and comment on its appearance.

   (b) Obtain box plots for **sepallength** by **species** and comment on any differences you notice between the different species sepal lengths.

   (c) Obtain basic descriptive statistics for the sepal lengths for all species together. Comment on any general features of the data (e.g., typical sepal lengths, range of values, etc.). Use visual and quantitative methods to comment on whether an assumption of normality would be reasonable for the general population of iris plants.

   (d) Obtain basic descriptive statistics for **sepallength** by **species**. Comment on any general features for this stratification. Comment on the assumption of normality for the species-wise population's sepal length. Be sure to use quantitative and visual methods.

(e) Comment on how the species-wise statistics and graphics differ from those for all species combined.

2. **Hypothesis Testing** (3 parts)

(a) Test the null hypothesis that the true mean or median sepal length is **60** against the alternative that it is not **60**. Based on the normality tests from Exercise 1, which location test should we use and what do we conclude about the true mean or median sepal length of the population?

(b) Of the three species, **virginica** has the highest mean and median sepal length, but is it significantly greater than the general population? Use the sample median for sepal length of all species as the null value and perform a t test, sign test, or signed rank test to decide whether **virginica** has significantly larger sepal length. (Hint: the median value in results and tests from Exercise 1 may help to inform you about your decision.)

(c) Consider the **setosa** and **versicolor** sepal lengths. Test for differences of the two populations (e.g., a test for a difference of mean or median *or* a test for one population being stochastically greater than the other). State your conclusion which should include why you chose that test for the differences of the two populations.

3. **Correlation** (3 parts)

(a) Obtain the Pearson correlation matrix for the entire data set. Comment on what the results tell us about significant relationships between the four length and width measurements.

(b) Obtain the Pearson correlation matrix by **species**. Comment on what the results tell us about significant relationships between the four length and width measurements for each species.

(c) Compare and contrast the relationships in parts (a) and (b).