GEOG 455 Exercise Two
DATA SUMMARIZATION
15 points

Name:_____
Student#:_____

Purpose: *Familiarity with a few methods for summarizing, characterizing, and comparing data sets*

This exercise requires summarization, characterization, and some inferences about several groups of data. Class discussion and notes should provide all of the background needed to complete the tasks in each part of the exercise, but if you need additional help, please see me. The final product in each case should include a short explanation of your findings suitable for presentation to the person who requested your assistance. Your theoretical employer (see text of the problem) will be interested in receiving something readable and applicable to the problem at hand, and will be uninterested in actual scratch calculations. You will be using SPSS to assist you with the calculations. Data for the exercise (Table 1 below) are available on a computer file, readable by SPSS. This file can be downloaded from the course web page, or accessed on the course CD.

One of the tasks facing the climatologist working in government or industry is the comparison of present conditions with those of some climatological normal. For example, many Kansas municipalities depend upon surface-recharged aquifers for their water supply. Water from such aquifers varies in quality and quantity as drier or wetter than normal conditions prevail at a particular site. Surprisingly, **both wet and dry situations** increase the dissolved salt content in the water, in some cases to a value that exceeds safety standards. This exercise is abstracted from a study on water supply and drought relationships in Kansas; results of the study were used as a planning and emergency preparedness device by a number of communities.

Table 1, Sample Periods and Population

Period One: 4.8,4.5,4.6,3.6,3.9,3.0,4.1,4.3,4.6,5.1,4.0,3.9,4.1

Period Two: -1.2,-1.8,-2.1,-1.7,-1.3,-1.2,-1.4,-2.3,-3.0,-2.4,-2.5,-3.2,-3.2,-3.5,-3.1,-3.1

Period Three: -1.0,-0.9,0.6,0.7,0.6,1.3,1.6,-1.2,-0.9

Five Year Population at Havensville: 1.50,2.33,2.40,2.07,1.30,1.08,1.05,1.67,1.07,2.00, 1.48,1.79,1.31,1.85,2.94,2.77,3.66,3.03,3.26,2.87,4.71,4.83,5.71,5.25,5.22,5.03, 4.71,-0.79, -1.39,-1.32,0.08,0.20,1.48,-0.21,-0.05,-0.16,-0.11,-0.39, -0.36, -1.38,-1.84,-1.88,-2.05,-2.79, -3.10,-3.19,-3.39,-3.85,-4.00,-4.53,-4.56,0.62,0.02,0.50,0.25,0.94,0.76,0.14,1.41,1.44

Situation

You have been charged with determining whether a particular period at Havensville, Kansas is wetter than normal, normal, or drier than normal, and to show the characteristics of variables under these three circumstances. The user, a water municipality, wants to be able to readily identify the three situations and is untrained in climatology (i.e., techniques and displays should be as simple as possible but representative of good professional practice).

Data

You have five years of data collected from a meteorological tower located at the municipal well site in Havensville. The data have been used to calculate Palmer Drought Index Values (PDI). The PDI whether a certain period was drier or wetter than normal at a particular location. Values typically range from plus to minus 4, although plus 7 to minus 7 have been recorded in Kansas (Table 2). The index is structured in such a way that values over a long time period follow a Normal Distributional curve. PDI is the most accepted drought index and is reported regularly by the National Weather Service. Calculation of the individual values is a lengthy affair, but commonly done. The procedure is outlined in Whittemore, D., G. Marotz, and K. McGregor, (1982) *Variations in Water Quality with Drought.* Washington, D.C.: US Department of Agriculture. PDI values for 3 different periods at Havensville and for the entire sixth Kansas climatic district have been calculated for you and are in Table 1 and the SPSS file; the three periods represent the *samples*, and the five year period represents the *population*.

Procedure

The questions that you need to address are:

l. What are the characteristics of the sample and population?
   Answering this question requires the use of simple descriptive statistical procedures. Among the things you should determine are the **Mean**, **Median**, **Mode**, **Upper Quartile** (75^{th} percentile), and **Lower Quartile** (25^{th} percentile), of each sample and population.
2. Are values in the sample and population clustered about a single number, or are they widely dispersed? Answering this question involves calculation of the **Standard Deviation**, **Standard Error**, **Maximum**, **Minimum**, and **Range** of each sample and population.
3. Is the population heterogeneous enough to be divided further, say, into normal, wet and dry periods?
   This question is normally settled by using somewhat advanced statistical techniques, such as cluster analysis, which would divide the data into groups with as little internal variation as possible. You do not want to make your answer unintelligible to the user, so you decide to use characteristics of the normal distributional curve to divide the population into three groups. 68 percent of the observations on such a curve fall within one standard deviation of the mean, 95 percent fall within 2, and 99.7 percent fall within 3 standard deviations. Thus you can use +0.6 and -0.6 standard deviations from the mean as the points to divide the population into three groups, each comprising roughly 33% of

the population (from Z-score tables found in the back of any introductory statistics text). After division, calculate the statistics from questions 1&2 for each of the three sub-groups of the population.

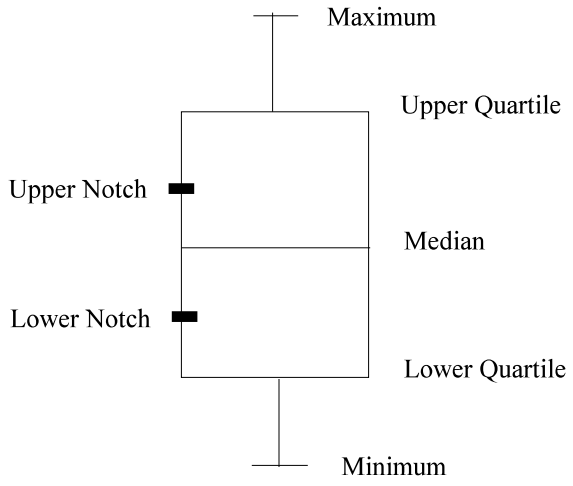
4. Are each of the three sample groups members of the same population?

Again, under most circumstances, you would probably employ a test that compares the means and variance of each group in a statistically meaningful way, but which provides an answer couched in probabilistic terms -- something that may not be useful to our intended audience. Simple graphical methods are available for the same purpose; these include the Box Plot technique, a method from a field called Exploratory Data Analysis (Velleman, P. And Hoaglin, D. (1982) *Applications, Basics, and Computing of Exploratory Data Analysis* (Boston: Duxbury Press). Box Plots (see example below) allow you to immediately see where a particular value lies within a distribution, and whether the distribution itself is similar to that of another variable.

You should already have calculated most of the values needed to complete the six necessary box plots, except for the notches. The latter are obtained from:

$$\text{Median} \pm 1.58 * \frac{(\text{Upper Quartile Value} - \text{Lower Quartile Value})}{\sqrt{\text{Number of Data Values}}}$$

Two groups whose notches do not overlap are statistically different ninety-five percent of the time. Compare each of the sample groups box plots with the "wet", "normal", and "dry" population box plot and determine if the city manager should worry about water quality in any of the three periods.



Summary

So, in summary you should provide the following when you turn in your report:

1. The statistics in the table provided below.
2. Box plots for the three periods, plus your derived "dry", "normal" and "wet" sub-populations on the graph paper provided (6 box plots in total).
3. Your assessment as to which of the three categories (if any) each period should be classified as, and what this means in terms of water quality problems for Havensville.

Statistic	Period 1	Period 2	Period 3	Pop. 5yr.	Dry Pop.	Nor. Pop	Wet Pop.
Mean							
Median							
Mode							
Standard Deviation							
Standard Error							
Max.							
Min.							
Range							
Upper Quartile							
Lower Quartile							
Upper Notch				XXXXX XXXXX			
Lower Notch				XXXXX XXXXX			

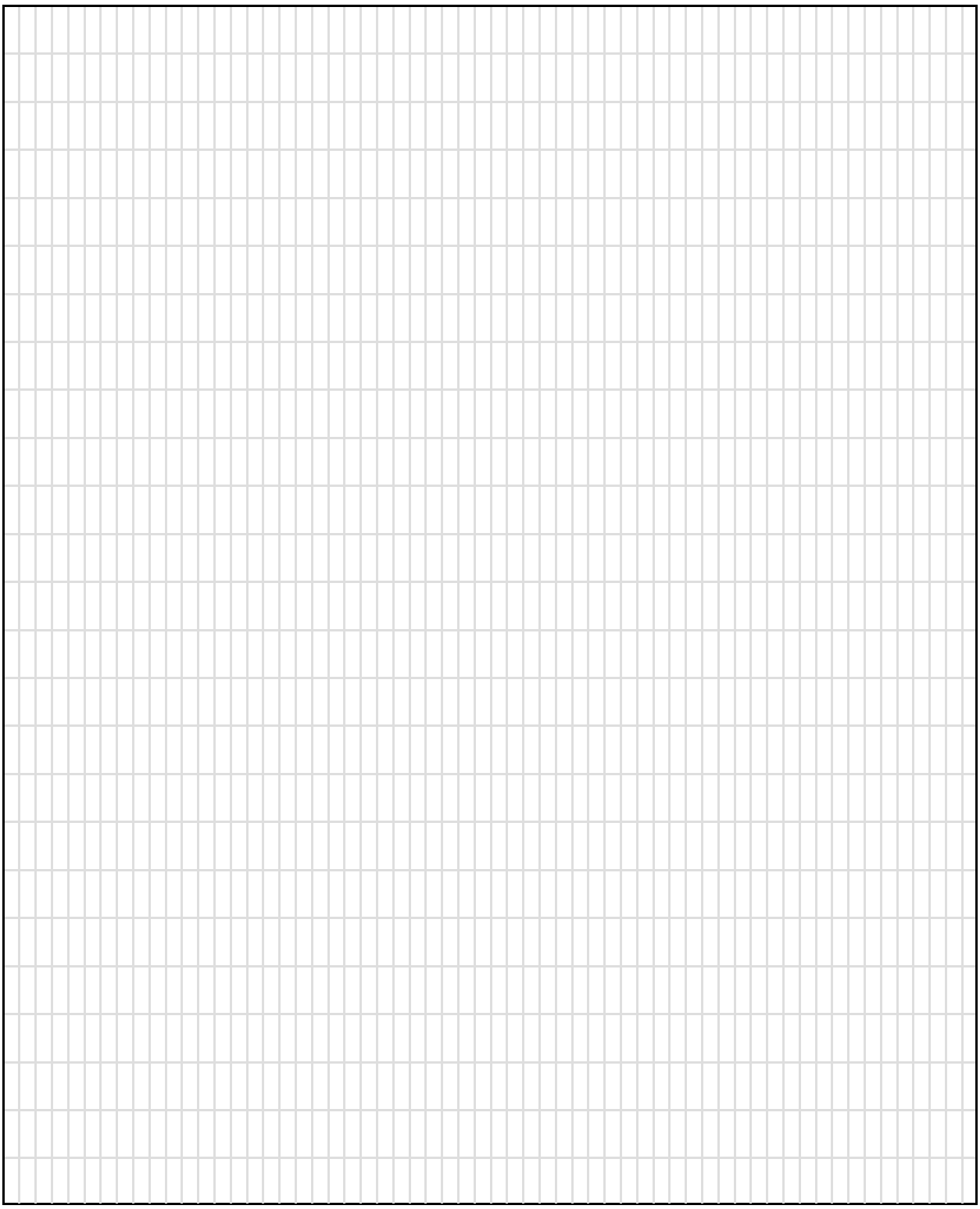


Table 2

Drought Classification by

Palmer Drought Severity Index (PDI)

Palmer Index	Characteristic
PDI $\leq$ -4.0	Extremely dry
-4.0 < PDI $\leq$ -3.0	Severely dry
-3.0 < PDI $\leq$ -2.0	Moderately dry
-2.0 < PDI $\leq$ -1.0	Mildly dry
-1.0 < PDI < +1.0	Near normal
+l.0 $\leq$ PDI < +2.0	Mildly wet
+2.0$\leq$ PDI < +3.0	Moderately wet
+3.0$\leq$ PDI < +4.0	Severely wet
+4.0$\leq$ PDI	Extremely wet