Lab 2: Normality Diagnostics and Regression-Correlation Descriptive Statistics

OBJECTIVES: This lab is designed to make the student familiar with the following techniques for assessing whether a given sample of observations is normal/non-normal:

boxplots

normal probability plots

It is important to observe that none of these procedures are automatic, and in many real situations deciding one way or another is very much a subjective decision. This lab is also designed to familiarize the student with the basic descriptive statistics associated with the correlation between two variables. Both summary statistics and graphical methods will be explored. The methods shown here can be used not only in the two variable applications, but also in other types of problems.

DATA: The first data we will use for this lab will be simulated (constructed) by Systat. Once we have worked with the simulated data for a while, we will then move on to two different real data sets, which will be described when we get to those points.

DIRECTIONS: Follow the instructions below, answering all questions. Write out the answers neatly and turn them in to the lab instructor.

1. Open a new Systat File and name the columns NORMAL, HEAVY, LIGHT, and ASYMETR. The abbreviations are for heavy-tailed, light-tailed and asymmetric. Use the **Fill Worksheet** option under the **Data** menu to fill the worksheet to 100 rows. Go to the command window and type:

   LET NORMAL=ZRN

   LET HEAVY=TRN(2)

   LET LIGHT=URN

   LET ASYMETR=XRN(1)

   IF CASE<21 THEN LET SAMPLE=1

   IF CASE>20 AND CASE<41 THEN LET SAMPLE=2

   IF CASE>40 AND CASE<61 THEN LET SAMPLE=3

   IF CASE>60 AND CASE<81 THEN LET SAMPLE=4

   IF CASE>80 AND CASE<101 THEN LET SAMPLE=5

   These commands fill the NORMAL column with generated data from a N(0,1) distribution, the HEAVY column with data from a t-distribution with 2 df, the LIGHT column with data from a U(0,1) distribution, and the ASYMETR column with data from a $\chi^2(1)$ distribution. You do not need to worry about what these different distributions are for now; you will learn about them later in the semester.

For this exercise, we want to have 5 groups each with sample sizes of 20. We generated samples of size 100 in the 1st 4 commands above. The second set of commands (involving CASE above) creates a variable called SAMPLE which has values corresponding to the 5 groups. Thus, we have created 5 samples of 20 observations each for each of the 4 situations: NORMAL, HEAVY, LIGHT, and ASYMETR.

a. First, we want to look at boxplots of these generated data sets to see what these plots can tell us about assessing normality. Turn on the Append/Graph option by going to **Windows**, then **Graph Placement** and then selecting **Append Graph**. There are 2 ways to make the boxplots. The slow way is to go to Graph/Box and choose NORMAL (for example) from the left window and SAMPLE from the right window. This should plot 5 boxplots side by side. The 5 groups are broken up according to the grouping variable SAMPLE.

   The fast way to make the boxplots is to go to the command window and type:

   ```
   GRAPH
   BOX NORMAL*SAMPLE/NSORT
   BOX HEAVY*SAMPLE/NSORT
   BOX LIGHT*SAMPLE/NSORT
   BOX ASYMETR*SAMPLE/NSORT
   ```

b. What characteristic do the heavy tail box plots seem to have which the other box plots do not? What seems to characterize the boxplots for LIGHT? For ASYMETR?

c. Turn on the **ByGroups** option under the **Data** menu, selecting SAMPLE as the grouping variable. Go to Graph/Prob'y/Normal and choose NORMAL. Look through the 5 normal probability plots that SYSTAT creates. If the data are normal, then the observations should fall roughly along a straight line. In this simulation set-up, we have generated the NORMAL data from a normal distribution, so we expect the normal probability plots to fall on a straight line. Does this seem to be the case? Discuss.

d. Now do normal probability plots for HEAVY. What do you observe (pay attention to the scale of the horizontal axis)? How would these plots change if we removed some of the extreme outliers? How are the extreme outliers related to the "heavy tails" of the density?

e. Do normal probability plots for LIGHT. Do the plots for light tails seem more curvy, less curvy, or about the same as the plots for NORMAL?

f. Do normal probability plots for ASYMETR. What do you observe about these plots relative to the normal?

2. Now we want to try out our normality diagnostics on some real data. The first data set we will consider is the Cavendish Data.

   DATA: In 1798, H. Cavendish set out to find out the density of the earth relative to that of water, using a torsion balance. He conducted 29 very precise experiments assessing this relative density. A working assumption which we begin with is that the values of his measurements were taken from a Normal distribution centered on the true relative density. The data from all of his 29 experiments are given below (taken from Stigler, S., "Do Robust Estimators Work With Real Data?", *Annals of Statistics*, **5**, 1977).

5.50	5.55	5.57	5.34	5.42	5.30
5.61	5.36	5.53	5.79	5.47	5.75
4.88	5.29	5.62	5.10	5.63	5.68
5.07	5.58	5.29	5.27	5.34	5.85
5.26	5.65	5.44	5.39	5.46	

a. Open a new data file, title a column CAVEND and enter the above data in that column.

b. Calculate and record the basic summary statistics, such as the sample mean, standard deviation, minimum,and maximum

c. Construct a boxplot of this data. Does it seem plausible that this data comes from a normal population (use your own judgement, there's no right or wrong answer here)?

d. We now want to use some more sophisticated tools to come up with a better opinion as to the normality of the Cavendish data. In the command window, type:

   LET STDCAV=(CAVEND-5.4479)/0.2209

   This creates a standardized version of CAVEND in a new column called STDCAV.

   Also, we want to generate several data sets from a standard normal distribution with the same sample size as the Cavendish data (i.e. 29). In the command window type:

   LET Z1=ZRN

   LET Z2=ZRN

   .

   .

   .

   LET Z10=ZRN

   This creates 10 new columns each filled with 29 generated N(0,1)'s.

e. Now make sure the Append/Graph option is on. Make normal probability plots of all the variables (don't select any variables, just click OK and Systat does it for all of them); also choose the **Axes** option in the **normal probability plot** menu and set Xmin=-3 and Xmax=3. This will keep the horizontal scales the same for all of our plots. Toggle through Z1, ..., Z10 and STDCAV, ignoring the other plots. Does it seem plausible that STDCAV could be one of the Z's? Critique the following statement: Based on the normal probability plot of the Cavendish data, we conclude that the Cavendish data is indeed a sample from a normal distribution.

3. The goal of descriptive statistics in a regression-correlation setting is to determine if and to what extent a linear relationship exists between two variables.

   DATA: The third data set in this lab is found in the Systat data file ROTATE. These data measure the angle of rotation in degrees, ANGLE, and the reaction time in seconds, RT, for a perception study. The experiment measured the time it took subjects to make the same judgements when comparing a picture of a three-dimensional object to a picture of possible rotations of the object.

   a. Look at a scatterplot of the ROTATE dat. To do this, select the **Plot** option from the **graph** menu. Then select the **Plot** option from the **Plot** menu. Now select ANGLE as the independent variable (**x**) and RT as the dependent variable (**y**). Does there seem to be a linear relationship between the two? Try the scatter plot again with the linear smooth option. Click on the **smooth** box in the Graph/Plot/Plot menu. Then select the linear option and hit **ok**. This will show the best linear least squares approximation to the data. Does this line seem to 'fit' the data?

   b. Use the **Pearson** option under the **Corr** option of the **Stats** menu to calculate the Pearson correlation coefficient for the data (this is more commonly referred to as the sample correlation coefficient). What is its value? What does this value tell you about the degree of linear relationship between the angle of rotation and the reaction time? If there were a "perfect" positive linear relationship in the data, what would be the value of the Pearson correlation coefficient? What does a correlation coefficient of -1 imply?

   c. Based on these descriptive statistics, is there a linear relationship between the angle of rotation and reaction time? Is it a strong or weak relationship? Is it a positive or negative relationship? How good do you think a prediction based on the regression line is?

   d. Suppose you had to relate these results to the psychologists that performed the experiment. What would you tell them? (*i.e.* What does your answer to the previous question mean in terms of the experiment itself?)