**Lesson 7 Homework**

1. Refer to the CLINTON data. Write a SAS program which reads the data. Create a new variable which indicates the number of days elapsed between successive polls. Set the elapsed time for January 24, 1993 (the day of the first poll) to be a missing value. Then, use PROC UNIVARIATE to examine the distribution of those elapsed times. Identify the 5 shortest and 5 longest elapsed times with dates, using an appropriate format for the date. What was happening to President Clinton when the opinion polls were spaced very close together? (Stick to what you can tell from the data and not political events)
2. Refer to the USEDCARS data. Suppose that an econometrician wanted to use results applicable to the normal distribution to describe the prices of used cars. Used PROC UNIVARIATE to decide whether the prices or the logarithms of the prices more closely follow a normal distribution. Write down at least two findings from PROC UNIVARIATE to support your claim.
3. The midrange statistic is sometimes used to report a central value of a distribution. The midrange is defined as (minimum value + maximum value)/2. Refer to the LIMES data. Use PROC UNIVARIATE to calculate the midrange of the juice liquid volumes of the limes. For this problem, you must use SAS to perform all of the calculations; for example, you may not find the minimum and the maximum with SAS then calculate the midrange by hand. Use PROC IMPORT to import the data:
4. Refer to the IRIS data. Suppose that you want to present summary statistics to someone who is not familiar with SAS PROC UNIVARIATE output. Prepare a printout which lists only the sample size, mean, and standard deviation of the sepal widths separately for each of the three iris species. Use SAS commands to explicitly label these numbers on your printout as “Sample Size”, ”Mean”, and “Standard Deviation”.
5. Refer to CATS2 data. Suppose that the veterinarian wants to see if the treatment had altered kidney function within the first week after surgery. One way to do this is to perform a paired t-test. Calculate a new variable representing ((GFR of the untreated kidney in Week0) minus (GFR of the surgically-treated kidney in Week1)) for each of the 8 cats. Then, apply PROC UNIVARIATE to those differences. The p-value of the 2-sided t-test is the number marked Pr>|T|. Based on this value, would you decide that the surgery had an effect after one week?
6. In some situations, a trimmed mean is used as a measure of the central value of a distribution. To calculate a trimmed mean, equal numbers of the lowest and highest observations are removed from the data, and the remaining observations are averaged. Trimmed means are used in the Olympics. In subjectively-scored events such as diving, the lowest and highest ratings given by judges are thrown out, and the remaining scores are averaged.

Refer to the GRADES dataset. Use a SAS program to find the total number of points earned by each student. You should find that the lowest point total is 63 and the largest is 103. Then calculate the trimmed mean of the point totals after removing the single lowest point total and the single highest point total, but do not explicitly write 63 and 103 in the program. This needs to be automated so the code can run on any dataset and produce the desired result.

1. Using SAS data set College, compute the mean, median, minimum and maximum and the number of both missing and non-missing values for the variables ClassRank and GPA. Report the statistics to two decimal places.
2. Using the SAS data set College, report the mean and median GPA and the ClassRank broken down by school size (SchoolSize). Do this twice, once using a BY statement, and once using a CLASS statement.
3. Using the SAS data set College, report the mean GPA for the following categories of ClassRank: 0-50=bottom half, 51-74=3rd quartile, and 75 to 100 = top quarter. Do this by creating an appropriate format. Do not use a DATA step.
4. Using the SAS data set College, create four summary data sets containing the number of non-missing and missing values and the mean, minimum, and maximum for ClassRank and GPA, broken down by Gender and SchoolSize. The first data set (Grand) should contain the statistics for all subjects, the second data set (ByGender) should contain the statistics broken down by SchoolSize, and the fourth data set (Cell) should contain the statistics broken down by Gender and SchoolSize. Do this using PROC MEANS (with a CLASS statement) and one DATA step.

Hint: Use the CHARTYPE procedure option.

1. Using the data set PATIENTS, write the necessary SAS statements to create a data set in which the first visit for each patient is omitted. Then, using that data set, compute the mean HR, SBP, and DBP for each patient. (Patient 9 with only one visit will be eliminated.)

**DATA** PATIENTS;

INPUT @**1** ID $3.

@**4** DATE MMDDYY8.

@**12** HR **3.**

@**15** SBP **3.**

@**18** DBP **3.**

@**21** DX **3.**

@**24** DOCFEE **4.**

@**28** LABFEE **4.**;

FORMAT DATE MMDDYY10.;

DATALINES;

0071021198307012008001400400150

0071201198307213009002000500200

0090903198306611007013700300000

0050705198307414008201300900000

0050115198208018009601402001500

0050618198207017008401400800400

0050703198306414008401400800200

;

1. Write a program similar to the one in problem 1 except that we want to include all the data for each patient, excluding any patient who has had only one visit. Instead of having PROC MEANS create printed output, use it to create a SAS data set (PAT\_MEAN) containing the mean for each patient (use the AUTONAME output option to name these variables).
2. We have the data set called BLOOD that contains from one to five observations per subject. Each observation contains the variables ID, GROUP, TIME, WBC (white blood cells), and RBC (red blood cells). Run the following program to create the data set.

**DATA** BLOOD;

LENGTH GROUP $ **1**;

INPUT ID GROUP $ TIME WBC RBC @@;

DATALINES;

1 A 1 8000 4.5 1 A 2 8200 4.8 1 A 3 8400 5.2

1 A 4 8300 5.3 1 A 5 8400 5.5

2 A 1 7800 4.9 2 A 2 7900 5.0

3 B 1 8200 5.4 3 B 2 8300 5.4 3 B 3 8300 5.2

3 B 4 8200 4.9 3 B 5 8300 5.0

4 B 1 8600 5.5

5 A 1 7900 5.2 5 A 2 8000 5.2 5 A 3 8200 5.4

5 A 4 8400 5.5

;

We want to create a data set that contains the mean WBC and RBC for each subject. This new data set should contain the variables ID, GROUP, M\_WBC, and M\_RBC where M\_RBC and M\_WBC are the mean values for the subject. Finally we want to exclude any subjects from this data set who have two or fewer observations in the original data set (assume there are no missing values).

HINT: Use PROC MEANS with a CLASS statement. Since we want both ID and GROUP in the new data set, then you can make them both CLASS variables or include ID statement (ID GROUP;) to cause variable GROUP to be present in the output data set. Also remember the \_FREQ\_ variable that PROC MEANS creates. I will be useful for creating a data set that meets the last condition of excluding subjects with two or fewer observations.

1. Modify the program above to include the standard deviation of WBC and RBC for each subject.
2. We have the following data from our liver transplant database:

**DATA** LIVER;

INPUT SUBJ DOSE REACT LIVER\_WT SPLEEN;

DATALINES;

1 1 5.4 10.2 8.9

2 1 5.9 9.8 7.3

3 1 4.8 12.2 9.1

4 1 6.9 11.8 8.8

5 1 15.8 10.9 9.0

6 2 4.9 13.8 6.6

7 2 5.0 12.0 7.9

8 2 6.7 10.5 8.0

9 2 18.2 11.9 6.9

10 2 5.5 9.9 9.1

;

Use PROC UNIVARIATE to produce the histograms, normal probability plots and boxplots, and test the distributions for normality. Do this for the variables REACT LIVER\_WT SPLEEN, first for all subjects and then separately for each of the two doses. I am a moody transplant surgeon running on no sleep so explain it in a way that I won’t start yelling at you because I don’t know statistics!