

**Biostat 510**  
**Homework 1**  
**Due Tuesday, January 17, 2006**

1. Download the eight Excel files for the groups from the first day of class. Save them in a folder that you will be able to access later. They are in the Biostat 510 section of my web page at <http://www.umich.edu/~kwelch>.
2. Check the data in all of the Excel files, fix any problems, and resave the fixed files.
  - a) Open each file in Excel and make sure the **first row** of the Excel file contains the variable names.
  - b) Make sure there are no empty rows between row 1 containing the variable names, and the rows containing the data.
  - c) Check to be sure that all of the variable names are *spelled the same way*, and that there are no blanks in the variable names.

You should have the following variables in each of the Excel files:

Group	
Age	(Missing value code: 999)
Sex	
Height	(Missing value code: 999)
Weight	(Missing value code: 999)
Num_sibs	
Distance	(Missing value code: -1)
Ran	
Pulse1	(Missing value code: 999)
Pulse2	(Missing value code: 999)

- d) Resave each of the Excel files after you have made any corrections. Close all of the Excel files.
3. Use the Import Wizard to import the Excel files into SAS.
  - a) Import group1.xls into SAS, and save it as a temporary data set called WORK.group1.
  - b) Save the commands that you used to import the data in a command file called homework1.sas.
  - c) Open homework1.sas in the SAS Enhanced Program Editor window, by going to File...Open program... and browsing to where you saved the commands. The commands that were used to import the group1.xls file should look something like those shown below:

```
PROC IMPORT OUT= WORK.group1
DATAFILE= "C:\510\2006\group1.xls"
DBMS=EXCEL REPLACE;
SHEET="Sheet1$";
GETNAMES=YES;
MIXED=NO;
SCANTEXT=YES;
USEDATE=YES;
SCANTIME=YES;
RUN;
```

Note: you do not have to type the commands as shown above when using the Import Wizard in SAS. You will import the Excel file by filling in the options in the Import Wizard, and SAS will create the commands automatically and save them in your command file. Be sure to save the homework1.sas commands in a folder where you can access them later. See the Power Point handout for instructions on how to use the Import Wizard to import an Excel file.

Be sure to specify the correct location for where you saved the Excel file when you use the Import Wizard to import the data for group1.

- d) Make sure the group1 data set has been imported correctly by browsing it in the VIEWTABLE Window in SAS. Close the view of the data set in SAS, being careful not to close all of SAS in the process, and return to the Enhanced Program Editor window.
  - e) In the SAS Enhanced Program Editor window, copy and paste the SAS commands for group1, and modify them to create the remaining 7 data sets, which will be called WORK.group2 through WORK.group8. Make sure you edit the names of OUT= and the Excel file names in DATAFILE= in the commands to correspond to the data for each group.
  - f) Submit the PROC IMPORT commands for all of the data sets. Check the log to see how many observations are in each Excel file.
4. Combine the data sets for all eight groups into one temporary data set called WORK.allgroups using a DATA Step in SAS and set up missing values.
- a) You will use a SET command in SAS to create the new data set.
  - b) Use if ...then ... statements to set up the missing value codes as appropriate.
  - c) Make sure all data step commands are included before the RUN; statement.
  - d) Your SAS commands will look something like those shown below:

```
data allgroups;
  set group1 group2 group3 group4
      group5 group6 group7 group8;

  if height = 999 then height = .;

  *more missing values codes will go here before "run";
run;
```

Submit these commands, by highlighting them and clicking on the running person icon at the top of the Enhanced Program Editor window.

5. Get a Proc Contents for the allgroups data set.
6. Get descriptive statistics for the numeric variables in the allgroups data set using Proc Means.
  - a) Get the n, min, max, mean, standard deviation and median for all numeric variables.
  - b) Make sure that correctly fixed the missing value codes for all variables, by checking the minimum and maximum values.

- c) If you did not specify the correct missing value code for a variable, go back and fix your SAS commands, and rerun them to re-create the allgroups data set, and rerun the commands for Proc Means.
- 7. Get histograms for the variables: Age, Height, Weight, Distance, Pulse1 and Pulse2, using Proc Univariate.
- 8. Get oneway frequencies for the variables: Group, Sex, Num\_sibs and Ran using Proc Freq.
- 9. Get horizontal barcharts for the same variables as in #8 above, using Proc Chart.
- 10. Save your SAS commands in homework1.sas. Re-run all of the commands in this command file and make sure there are no errors in your log.

Answer the following questions, based on the results of your descriptive analysis.

Please write out your answers in complete sentences; be sure it is easy to see what questions you are answering. Hand in no more than three pages for this assignment. Each question should take one or more sentences to answer it properly. In some cases a table may work better to answer some questions; feel free to use a table to answer a question where appropriate.

- a) How many observations were included in the data sets for each group?
- b) How many observations and how many variables were included in the allgroups data set? What was the type of each variable?
- c) Include the SAS output from Proc Means for the allgroups data set in your write-up. What was the sample size for each variable?
- d) What was the mean distance for the entire group? What was the median distance? What was the minimum distance? The maximum distance? Given that the circumference of the earth is about 24,900 miles, does the maximum distance make sense? Go back and set the incorrect distance to missing in the Data Step, and rerun the Data Step commands to re-create the allgroups data set. Answer all of the questions in d) again based on the new values.
- e) Include the histogram for Age in your write-up. How would you describe the distribution of Age for this class? How would you describe the distribution of Pulse1? Pulse2?
- f) How many observations are there in each Group? Make sure this matches the size of the data set for each Group in question a) above. What percent of the students in this class are females and what percent are males? How many and what percent of the class have zero siblings? What is the largest number of siblings and what percent of the class have this number of sibs?
- g) There appears to be a problem in the Ran variable, in that one person has a code of 2 for Ran. Go back and fix this value of Ran in the Data Step commands so it is set to missing. Re-run the Data Step commands to make this recoding of Ran take effect. How many and what percent of the class ran? did not run? have a missing value for the variable Ran?
- h) Hand in a separate sheet containing your SAS commands for this homework assignment.