22S:30/105, Statistical Methods and Computing

Homework 2, Spring 2014 Instructor: Cowles

Due: Fri., 02/07 in class

Please put your name at the top of your homework, and list the names of any classmates with whom you collaborated.

1. Go to the Datasets section of the course web page. Read the description "hot-dogs.info." Answer the following questions, using SAS to produce the the required descriptive plots, tables, and/or summary numbers. **Turn in all of your SAS code, but only those parts of the SAS output that are needed to answer the questions.**

   Note that hotdog type is a character variable. An example data step that would work to read in this file is:

   ```
   data wieners ;
   input type $ calories sodium ;
   datalines ;
   Beef    186     495
   Beef    181     477
   Beef    176     425
   Beef    149     322
   Beef    184     482
   .
   .
   .
   Poultry 152     588
   Poultry 146     522
   Poultry 144     545
   ;
   run ;
   ```

   (a) What type of variable is hotdog type (by a statistician's definition, not just what SAS calls a character variable)?

   (b) What would be an appropriate table or set of summary measures to describe this variable? Produce it and turn in a printout.

   (c) What proportion of the hotdog brands in the study were poultry?

   (d) Name the two kinds of plots that we have studied that would be appropriate to display the distribution of the hotdog type variable.

   (e) Produce either one of the graphs that you listed in the previous question. Turn in a printout.

   (f) What type of variable is *sodium*?

   (g) Use "proc univariate" to produce a stemplot and a boxplot of *sodium*. (For now, we will consider all the hotdog types together. In future homework, we'll consider

them separately.) Turn in a printout of the parts of your proc univariate output used to answer the following questions (include the two plots).

  i. Is the distribution of *sodium* roughly symmetric?
  ii. Are there any outliers (in the general sense of points lying far away from the main group of points)?

(h) What is the 5-number summary of *sodium*? You may either print it from SAS or just copy if off the screen.

(i) What is the interquartile range (IQR)?

(j) Refer to the paragraph after textbook problem 2.25 for the $1.5 \times IQR$ criterion for outliers. Based on that criterion, what is the smallest value of sodium that would qualify as an outlier in the high direction? Are there any high outliers?

(k) What is the largest value of sodium that would qualify as an outlier in the low direction? Are there any low outliers?

(l) Is *sodium* a variable for which it would also be reasonable to report the mean and standard deviation? If so, what are they?

(m) If sodium is measured in units called milliequivalents, what are the units for the mean and the standard deviation?

2. Without using a computer, find first and third quartiles of these values:

   27 56 82 108 29 37 56 81 94

3. The data from Table 2.2 in the textbook are on the course web page as "athsal2011.dat." With this dataset, use SAS to produce a graph and a numeric summary. Then write a brief (2 or 3 sentence) description of the important features of the distribution.

4. Textbook problems: 2.4, 2.33, 2.38, 2.46