Last Long Homework, Due Thursday, Dec 11th

For all problems, turn in your SAS code as one or more documents (you can have one per problem) and a separate single document in Word or PDF that explains your answers and includes figures. As usual, results in figures should be discussed in the text and complete sentences in English should be used for all answers.

1. For this problem, you will have to reverse engineer the SAS output on slide 16 of the week 15 lectures. The goal for (a) and (b) is to determine what SAS did to generate the plot.

(a) Determine the age ranges for each bar. If 58 is the midpoint, does this mean that the age range for the left most bar is all ages \( x \) in the interval \( 56 < x < 60 \), or \( 56 < x \leq 60 \) or \( 56 \leq x < 60 \) or \( 56 \leq x \leq 60 \). It might be easier to determine whether these intervals are open, closed, half-open on the left or right, and so forth for other ages that more data.

(b) Determine whether the error bars being plotted are confidence intervals or the mean plus and minus one estimated standard error. Recall that a confidence interval would be based on

\[ \bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}} \]

and that \( s/\sqrt{n} \) is the estimated standard error. You can take a particular interval and calculate by hand, in a calculator, using R, SAS, or Excel, what the confidence interval and estimated standard error are and figure out what SAS is really plotting. You might be able to find this information online from SAS manuals and so forth, but be sure to also justify your answer by calculating at least one standard error and confidence interval on your own (or using other SAS procedures) and giving the numerical values in your report (you never know, an online source might be wrong, so it is good to check for yourself). Note that to do this, you'll need to determine how many observations are in an age category based on problem (a).

(c) Make a box plot (using either SGPLOT or some other procedure such as PROC BOXPLOT using the same age and sex categories as were generated by the bar plot on slide 14 of week 15. You might have to create the age categories yourself, since otherwise you might end up with one box plot for each distinct age. Describe the results of the box plot (does there appear to be similar variation in different age categories, does the data seem roughly normal for different age categories?)

2. Use PROC SURVEYSELECT to generate a 95% bootstrap percentile interval for the difference in medians, \( \eta_1 - \eta_2 \) between normal body temperature for people aged 75 and under versus those aged over 75 using the temperature data set. Ignore the sex of the individuals in the study. Use 1000 bootstrap replicates. Use the percentile interval to test the hypothesis \( H_0 : \eta_1 = \eta_2 \) that the median temperature for the two age groups is the same, and state the conclusion of the test in the context of the problem (i.e., the conclusion is not just “there is (or is not) sufficient evidence to reject the null hypothesis”). Also make a histogram of the 1000 bootstrapped differences in medians.

3. Find out about some SAS procedure that we haven’t used or mentioned in class (not PROC BOXPLOT). Write a summary of what the procedure can do and illustrate its use on one data set used in class and one data set that you find for yourself (could be a table from Wikipedia, an atlas/almanac that you find in the library, data mining your own computer, etc.). Discuss any limitations or difficulties with using the procedure (anything you wish it could do? or wish it could do better?) This problem is obviously pretty open-ended, so it is up to you to get something out of it. The point is that by the end of the SAS class, hopefully you can start looking up how to do things in SAS on your own, and not be limited to what is taught in a class.

4 (extra-credit, 2 points extra). Find or create a data set with two variables, \( x \) and \( y \), where \( x \) has a smaller interquartile range than \( y \), but the sample standard deviation for \( x \) is larger than the sample standard deviation for \( y \). Use SAS to compute standard deviations and interquartile ranges and include the datasets in the main text for your homework solutions.