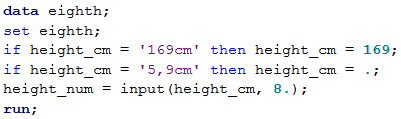
Lab 3

*Census at School* (<https://ww2.amstat.org/censusatschool/>) is a project that engages students in grades 4-12 in statistical problem solving. Students complete a brief online survey, providing them with a census of their classroom that they can compare with random samples of students throughout the country. The list of questions (the questionnaire) is given on Blackboard as the file “Lab 3 C@SQuestionnaire.pdf”. The ordering of these questions corresponds with the order of the variables in the Excel file posted on Blackboard (after the variables Country, Region, DataYear, and ClassGrade). This data was obtained from the Random Sampler Tool on the website (<https://ww2.amstat.org/censusatschool/RandomSampleForm.cfm>), which allows you to take random samples of sizes 10, 25, 50, 100, 250, or 500 from all respondents from specific states, grades, genders, or years that you select.

We will compare the heights of male and female 8th graders. The data is in the file “lab 3 - eighth100.xlsx”, which was produced from a random sample of size 100 from all 8th graders nationwide completing the survey.

1. Height is asked on number 4 of the questionnaire.
   1. State how the question is asked in the questionnaire.
   2. After importing the data file into SAS, identify the name of the SAS variable corresponding to height.
   3. Show the code used to make side-by-side boxplots of the height of students by gender. (Note: This will produce an error.)
   4. Copy and paste the following message from the Log window. “ERROR: The VBOX variable must be numeric”. (Note: it would say HBOX instead of VBOX if you used that instead; this is fine.)
2. This data is “messy data”. In other words, it’s real data containing exactly what was typed in by the 8th graders when they did the survey online.
   1. While the height\_cm variable should be numeric, it is read in as a character variable by SAS because two of its observations contain letters. Look at the data and identify these two observations. (Write them as your answer.)
   2. Use the following code to modify these two observations and then make a new variable “height\_num” which is numeric.  
         
      Also make a label for this new variable. Now make the side-by-side boxplots and copy and paste them.
   3. There are 4 outliers in the male boxplot. Identify the height values for each of these outliers by looking at the dataset.
   4. For each of these 4 outliers: Do you think it represents the heights in cm of the student who typed it in? Why or why not?
   5. Choose one of the outliers and explain what you think the student was thinking when they typed them in.
   6. Copy and paste a table that contains the sample size (N), mean, and standard deviation of the heights of the male and female students. Use two decimal places.
   7. This table should have columns including both N Obs and N. “N Obs” represents how many students are in each of the categories Female and Male.
      1. Why isn’t the sum of N Obs equal to 100, the total number of individuals in the sample? (Hint: it may help to look at the dataset.)
      2. What do you think N represents and how is this different from N Obs?
3. We will modify the dataset to exclude values of height we think don’t represent the students who entered them. We will do this in SAS, and not by modifying the Excel file. (Though it might be easier to modify them in Excel, it is good statistical practice to keep a copy of the raw data file and make changes to it in the SAS program so we have a record of the changes made for reproducibility’s sake.)
   1. Create a new dataset where the height values of 14, 65, and 72 are changed to missing values (a period). We will keep the height value of 203. Show the code used to do this.
   2. Copy and paste a table like in 4(a) but for this new dataset.
   3. Compare the values of the summary statistics for the males in the tables before and after removing these three height values (that is, in the tables for 4a and 5b). Describe the difference and explain its cause for
      1. N
      2. Mean
4. The standard deviation shrinks by more than a factor of three from removing the three observations. To see why, let’s look at the contribution of these three points to the sum of squared deviations (that goes into calculating the standard deviation).
   1. Starting the dataset before removing the three points, create a new dataset that made up only of the males. Show the code that does this.
   2. Add to this dataset variables “dev” that give the deviations of the height observations from the mean and “dev2” that give the squared deviations. Show the code that does this.
   3. Show SAS output that gives the sum of the squared deviations.
   4. Copy and paste SAS output that shows the values of the squared deviations for the three heights.
   5. What percent of the sum of squared deviations is contributed by these three observations?
5. Copy and paste the side-by-side boxplot after removing the three height values.
6. Using PROC TTEST, evaluate the model assumption of normality, including (and using) the plots provided by PROC TTEST).   
   1. Which of the following populations would be more appropriate for inference from the sample: (A) the population of all 8th grade students in the United States or (B) the population of all 8th grade students providing answers to the survey.
   2. Explain you answer to (a).
7. Test whether 8th grade males are on average taller than 8th grade females. Show the five steps of a hypothesis test:
   1. Write null and alternative hypothesis in terms of mu1 and mu2, and define mu1 and mu2.
   2. Find the test statistic and p-value from PROC TTEST. Copy and paste the table from the output that gives these values.
   3. Write the decision about Ho at the alpha=.01 level.
   4. Write the conclusion.
   5. Find a 99% confidence interval for the difference in mean heights. Copy and paste the table from PROC TTEST that gives it.
   6. Interpret this confidence interval within the context of the problem.