STA 138 - Discussion 1
Fall 2020

For our discussion this week, we will ease into categorical data analysis by illustrating the use of categorical data in the kinds of data analysis that we may be familiar with for numeric data. In doing so we'll explore computational tools that will be useful moving forward.

1. On Canvas, under Files, Discussions, you will find the file `patients101.csv`. This file has the following columns:

   Column 1: `age`: The age of the patient.

   Column 2: `totalchol`: A measure of the patients total cholesterol - the higher the number, the more cholesterol. In units of mg/dL.

   Column 3: `sysBP`: The patients systolic blood pressure. In units of mm Hg.

   Column 4: `weight`: The patients weight in units of kg.

   Column 5: `height`: The patients height in units of cm.

   Column 6: `sedmins`: The patients number of sedentary minutes per week.

   Column 7: `obese`: The patients obesity category, with values `normal`, `overweight`, `obese`.

   Column 8: `marriage`: The patients marriage category, with values `other`, `married`, `divorced`, `widowed`, `nevermarried`.

   Column 9: `gender`: `M` or `F`, denoting Male or Female.

   Consider your response variable (Y) to be the patients systolic blood pressure. It is a good idea to plot your response variable by itself first, to see the range, skew, etc.

   (a) Create a histogram of systolic blood pressure. Be sure to add labels to your axes (when appropriate) as well as a main title.

   (b) Does the histogram suggest the data is left skewed, right skewed, or approximately symmetric?

   (c) Create a boxplot of systolic blood pressure. Be sure to add labels to your axes (when appropriate) as well as a main title.

   (d) Are there any outliers (unusually small or large observations) in the data? If so, are they unusually large or small? What is the smallest data point (approximately), and the largest?

2. Continue with `patients101.csv`. Consider your response variable (Y) to be the patients systolic blood pressure. When you have a numeric response variable, and numeric explanatory variables, scatter plots are often useful plots to make.

   (a) Consider your first explanatory variable to be X_1 = weight. Create a scatter plot with systolic blood pressure on the y axis, and weight on the x axis. Be sure to add labels to your axes as well as a main title.

   (b) What trend do you see weight having on blood pressure, if any?

   (c) Consider your second explanatory variable to be X_2 = sedmins. Create a scatter plot with systolic blood pressure on the y axis, and sedentary minutes on the x axis. Be sure to add labels to your axes as well as a main title.

   (d) What trend do you see weight having on blood pressure, if any?

3. Continue with `patients101.csv`. Consider your response variable (Y) to be the patients systolic blood pressure. When you have a numeric response variable, and categorical explanatory variables, grouped box-plots or grouped histograms are often useful.

   (a) Consider your third explanatory variable to be X_3 = gender. Create a grouped box-plot by gender.

   (b) Does there appear to be a difference in systolic BP based on gender?

   (c) Consider your fourth explanatory variable to be X_4= obese. Create a grouped histogram by obesity category.

   (d) Does there appear to be a difference in systolic BP based on obesity category? Explain.

4. Continue with `patients101.csv`.

   (a) Find the average systolic blood pressure.

   (b) Find the average systolic blood pressure by marriage category.

   (c) Find the standard deviation of systolic blood pressure by marriage category.

   (d) Find the number of people in each marriage category.

   (e) Find the five number summary of weight.