# STAT 402 Project Proposal (Sketch)

## Goal

To complete the first case study and answer the following questions:

* Which risk factors are associated with hypertension? Do these associations hold with and without the survey design information (survey weight, bootstrap weights, specifying the 11 degrees of freedom)?
* Does the prevalence of hypertension and selected risk factors vary between men and women?  Across age groups?  How does your interpretation of these results change when the analysis is run with and without the survey design information?
* How would you summarize the impact of including and not including the survey design information in your analysis?  Did you see a greater impact for certain estimates and not others?

## Process

**Values Below the LOD**

There are not too many missing values (about 5% of the observations have missing data), but there are many observations that are below the Limit of Detection (LOD). However, while only about 2% of the LAB\_BCD values are below the LOD, a large 20% of the LAB\_BHG values fall below the LOD. Thus, if these values below the LOD are estimated poorly, this can have a significant effect on the accuracy of our analysis.

Popular methods of estimating these values below the LOC include removing observations that contain these LOD values or substituting these values with some constant (such as 0, LOC/2, etc.).

This method does not seem to have any statistical basis however.

While this is still a topic under investigation, it appears possible to instead use the Expectation-Maximization (EM) algorithm to find maximum likelihood estimates of the LOD values.

**Missing Values**

After the observations below the LOD have been estimated, the missing values in the data set must be estimated. One method under consideration is to simply replace these missing values with the sample mean of the variable they come from (or with the mode for categorical variables). Another possible method is to replace these missing values with randomly selected values from their respective variable. This second method has the advantage of better considering the variability of the data.

At this point in the analysis, there are no missing values in the data set, and the rest of the analysis can proceed.

**Theoretical Model**

The response HIGHBP will be modeled as independent Bernoulli random variables, while the initial link function will be the logit, or log-odds, function.

**Variable Selection**

The process that determines which variables go into the final model will be based off backwards selection, but will use Analysis of Deviance to test multiple terms at once to save time.

**Checking Assumptions**

We will use standard procedures such as residual plots to determine the validity of the model.

**Next Steps**

At this point, a model including information about the survey (weights, bootstrap weights, etc.), will be created in a similar fashion.

The first research question can be answered simply by comparing the variables in each of the two models.