Data 102, Fall 2023
Final Project Proposal

Due: 11:59 PM Friday, November 10, 2023

1. What dataset are you using?
   - Election dataset
2. If you're using any supplementary data, what are you using and why (1-2 sentences)? If you aren't, justify why you don't need any (1-2 sentences).
   - We will use some supplementary data on the results of the 2016 election. This is in order to gain a greater understanding of the existing political dynamics in each district as we anticipate these will have an effect on the primary election environment for each candidate.

General Research Question (Duplicate this section and choose the relevant version of question 4 for each one)

1. What is your research question?
   - Does the amount of endorsements from various different groups affect a candidate's ability to attract donors and/or win elections?
2. Which of the four techniques will you primarily be using for this question?
   - Multiple Hypothesis Testing
3. (Optional) If you're using more than one technique for this question, what other(s) are you using? Your answer here.
4. Depending on the technique you've selected for this question, answer one (or more) of the following:

   (a) **Multiple Hypothesis Testing:** Describe your hypothesis tests ($\leq$ 6), and how you plan to test them (2-3 sentences).
   - We will conduct hypothesis tests relating the endorsements of various groupings of endorsers (ex. party endorsements, presidential candidates, the eventual party nominee for president, and outside interest groups) to two outcome variables, percent of vote achieved and amount of funds raised. The individual hypothesis tests will consist of fitting a linear regression model with one-hot-encoded features for endorsement or anti-endorsement, which will generate coefficients with p-values that should tell us whether or not such coefficients are significant.

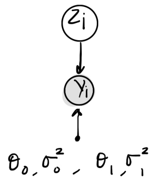
General Research Question (Duplicate this section and choose the relevant version of question 4 for each one)

1. What is your research question?
   - How can we separate the candidates in the dataset into those who won and those who did not win, based on their number of unique donors?
2. Which of the four techniques will you primarily be using for this question?
   - Bayesian Hierarchical Modeling with a Gaussian Mixture Model
3. (Optional) If you're using more than one technique for this question, what other(s) are you using? Your answer here.
4. Depending on the technique you've selected for this question, answer one (or more) of the following:

   (a) **Bayesian Inference:** Draw a graphical model (a hand-drawn whiteboard picture is fine) and briefly describe the variables in it. What is/are the unknown variable(s) you are trying to infer?

z_i = won primary $\in \{0,1\}$
y_i = # unique donors

$Y | z_i=0 \sim N(\theta_0, \sigma_0^2)$
$Y | z_i=1 \sim N(\theta_1, \sigma_1^2)$



We are trying to infer the parameters of the distributions of the number of unique donors. This is a Gaussian Mixture Model where the distribution of the number of unique donors is normal, where the mean and variance depend on whether or not the candidate won the election. Z represents a binary variable of whether or not the candidate won, Y represents the number of unique donors, and θ and σ represent the parameters of the Normal distributions for each of the cases (winning or losing).