1 Progress

We have assembled two large datasets for the purposes of the project:

- **Pennsylvania voter file:** This file contains a row for every registered voter in Pennsylvania, as well as their party registration, limited demographic information (age, gender), and their voting participation over a set of recent primaries and general elections.
- **Precinct election results:** This file, sourced from the Open Election project, gives the total number of votes received by each candidate in the 2016 presidential election.

Because these files were sourced from distinct locations, there are some challenges in matching up every precinct in every county. For the milestone, we have focused on four of Pennsylvania's 67 counties: Adams, Allegheny, Bedford, and Chester. These counties are diverse: Allegheny and Chester were won by Clinton, while Bedford and Chester were won by Trump. Together, they represent 1,639 precincts and about 950,000 total votes for Clinton and Trump.

We have written code to build and train simple models over this dataset. To compute our objective – the Poisson Binomial log likelihood – at a specific value of β, we use the following procedure for a precinct k:

- Compute the number of Trump and Clinton votes within k (the average precinct having around 600 votes)
- Look up all the voters in k who voted in the 2016 presidential election. Using their data from the voter file, construct the design matrix X_k.
- For a given value of β, compute a vector of probabilities $p_k = \sigma(X_k\beta)$ where $\sigma(\cdot)$ is the sigmoid function. Compute the log-likelihood of the Poisson Binomial using this vector of probabilities p_k and the true number of Trump/Clinton votes.

These values are summed over all precincts to get an overall log-likelihood.

Note that because the data is imperfect – and because individuals may vote for third party candidates or fill out a ballot but only vote in other races – the Clinton/Trump vote total sum is not exactly equal to the number of voters who participated in a given precinct. A quick visual check confirmed that the numbers are within 10% in most cases. To get around this issue, we take the percentage of Clinton/Trump votes that went to Clinton, multiply this by the length of p_k, and round the resulting number to estimate the number of voters in our file who cast a ballot for Clinton or would have done so if they "had to" choose one of the candidates.

We want the model to learn the correct values of β so as to maximize the overall likelihood of the observed data. But we have a problem: the Poisson Binomial likelihood involves sums over all possible configurations of votes ASDFSDFSDFSDF

To address this problem, we make use of the Lyapunov CLT, to observe that the asymptotic distribution of d_k, the number of Democratic votes in precinct k, is given by:

$$d_k \xrightarrow{d} N\left(\sum_i p_{k,i}, \sum_i p_{k,i}(1-p_{k,i})\right)$$

where $p_{k,i}$ is the i^{th} entry of p_k. This result is proven in the appendix. It allows us to *estimate* the likelihood with a much simpler function of β. In this case, the contribution of precinct k to the overall log-likelihood is approximately:

$$\ell_k = -log\left(\sum_i p_{k,i}(1-p_{k,i})\right) + \frac{1}{\sum_i p_{k,i}(1-p_{k,i})}\left(d_k - \sum_i p_{k,i}\right)^2$$

where $p_{k,i} = \sigma(\beta^T x_{k,i})$