Tracking COVID-19 Vaccine Hesitancy Within Demographic Subgroups through Bayesian Multinomial Modeling

Stanford STATS 271 Project
June 4, 2021

Ishan Shah
Department of Statistics
Stanford University
ijshah@stanford.edu

1 Introduction

Rapidly decreasing case and mortality rates have demonstrated that significant uptake of COVID-19 vaccines is conclusively the United States' way out of the pandemic. However, since its peak in April, vaccination rates have declined over 70% with 37% of adults still needing a shot. This can be attributed to the lack of access as well as hesitancy among the population for various reasons. I intend to model the latter because there is more definitive data on people's opinions and it is important to track changes in this over time to see whether certain interventions may or may not have helped.

This work was in part inspired by Figuerido's work using public opinion surveys and Bayesian methods to map global trends in vaccine confidence [1]. The study employed a three-category Bayesian multinomial logit Gaussian process and Gibbs Sampling to model this confidence in 149 countries and found significant increases and decreases in many countries. This analysis could be useful on a smaller scale within the United States among subgroups who often are spotlighted in hesitancy conversations, such as young people, racial minorities, and conservatives.

This problem can also be viewed similarly to election forecasting, particularly the Economist's [2]. The authors model polling data as binomial responses, and use prior knowledge relating to various regions and polls to adjust their estimates through bias terms. Paired with an informative prior based on economic fundamentals, they employ Bayesian simulation methods to predict the two major candidates' vote share.

As in [1], my model will spotlight the three categories of:

1. People who are already vaccinated or eager to receive the vaccine (subscripted by E thoroughout the paper)
2. People who are unsure, hesitant, or "waiting on others" (subscripted by H)
3. People who refuse or are strongly hesitant of the vaccine (subscripted by N)

I will focus specifically on Black and Hispanic Americans in this analysis due to time constraints from data extraction, but this analysis is intended to be easily expandable to other subgroups if the data is present.

2 Data

The data for this project comes from two sources. First are national public opinion surveys specifically asking whether people have already gotten the vaccine, are eager to get it, are unsure or hesitant about it, or are strictly against it. All polls came from the FiveThirtyEight election polling

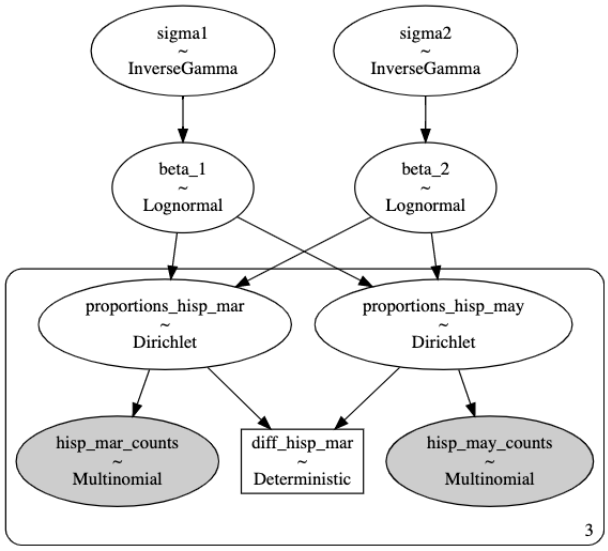

Figure 1: Example Model Hierarchy

database [4]. Economist/YouGov polls were the primary source of data as they are released weekly and consistently ask the same question with the same response choices. I extracted data from 28 polls between February and May 2021 and grouped them into these four months.

The second source of data is CDC county-level vaccine hesitancy estimates paired with corresponding demographic data [5]. While they do not explicitly provide the links between demographics and hesitancy, I can use the counties with significant populations of Black and Hispanic Americans as an informative prior for modeling the proportions.

3 Model

The goal of this analysis is to model the distribution of the proportions π for Multinomial Distribution y with for any given subgroup s and time t as shown below.

$$y_{st} \sim \text{Mult}(N_{st}; \pi = (\pi_E, \pi_H, \pi_N))$$

where N_{st} is the number of people sampled for that subgroup and month. The rest of the model will be hierarchical, with eventually a Dirichlet distribution being used as the final prior for the expected frequencies. The CDC hesitancy data can provide a relatively informative prior if the county demographics are already known. With multinomials, it is often more practical to work in ratios to transform the values to a continuous space. Therefore, I define the two coefficients which are unique to each subgroup:

$$\beta_{1s} = \frac{\pi_E}{\pi_N} \tag{1}$$

$$\beta_{2s} = \frac{\pi_H}{\pi_N} \tag{2}$$

After exploration of the CDC hesitancy data, I found that the log of the data's β_1 and β_2 equivalents is approximately normally distributed. Therefore,

$$\log(\beta_{1s}) \sim \mathcal{N}(\mu_{1s}, \sigma_{1s}) \tag{3}$$

$$\log(\beta_{2s}) \sim \mathcal{N}(\mu_{2s}, \sigma_{2s}) \tag{4}$$

In this case, I simply chose the top 100 counties with the highest percentage of Black and Hispanic Americans, found the ratios, applied a log transformation and took the median to serve as the initial distribution means μ_{1s} and μ_{2s}. Lastly, I used a standard uninformative prior for the σ terms:

$$\sigma_{1s}, \sigma_{2s} \sim \text{InvGamma}(1, 1)$$

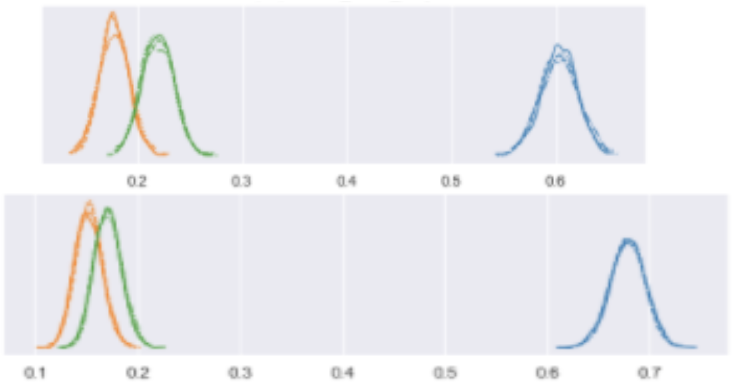

Figure 2: Black (top) vs. Hispanic (bottom) posterior distributions - May

Putting all this together in a Dirichlet distribution prior:

$$\text{Dirichlet}(\frac{w * \beta_{1s}}{\beta_{1s} + \beta_{2s} + 1}, \frac{w * \beta_{2s}}{\beta_{1s} + \beta_{2s} + 1}, \frac{w}{\beta_{1s} + \beta_{2s} + 1}) \tag{5}$$

I considered the weight term w to be a hyperparameter essentially controlling the influence the prior has on the model. Put another way, it's if the prior represents w samples.

Monte Carlo methods from Python's PyMC3 library were used to build and sample from the model, generating the posterior distributions. Initially, I tried the Metropolis sampler, but the sampling chains were fairly noisy and inconsistent, so I ended up primarily using the package's default sampler, NUTS (the No-U-Turn Sampler) [3]. It is essentially an adaptation of Hamiltonian Monte Carlo that avoids random walk behavior by taking more informed steps through first-order gradient information, efficiently creating smooth posterior distributions.

An example of the model hierarchy is shown visually in Figure 1. The example represents finding the difference in posterior proportions for Hispanics between March and May 2021.

4 Results

4.1 Current Hesitancy

First, I wanted to find the posterior distributions of both Black and Hispanic Americans' vaccine attitudes as they stand currently in the month of May using their corresponding priors (weight 500). After implementing the NUTS sampler with 1000 draws and 4 chains the visual distributions are shown in Figure 2 (blue is eager, orange is hesitant, green is no). Hispanics seem to be more eager at the moment, with a mean proportion of 69.1% whereas Black Americans have a mean of 60.3%. Their 90% highest density intervals (HDIs) are mutually exclusive, indicating a significant difference.

4.2 Tracking Hesitancy Over Time

The next step was to utilize all the polling data from February until now, and see how attitudes have changed over time in the two demographic groups. Figure 3 covers much of the relevant information. There seems to be a steady increase in confidence over all 3 months among Black Americans, whereas the majority of the increase among Hispanic Americans happened in March. We have enough survey responses to keep the 90% HDI intervals (represented by the bands around the lines) relatively narrow. I specifically picked out the March-May Hispanic difference to examine in more detail (right side of Figure 3). Interestingly, nearly all of the increase in eagerness came from the hesitant group, and the "hard no" group did not budge at all, which is not the case among Black Americans.

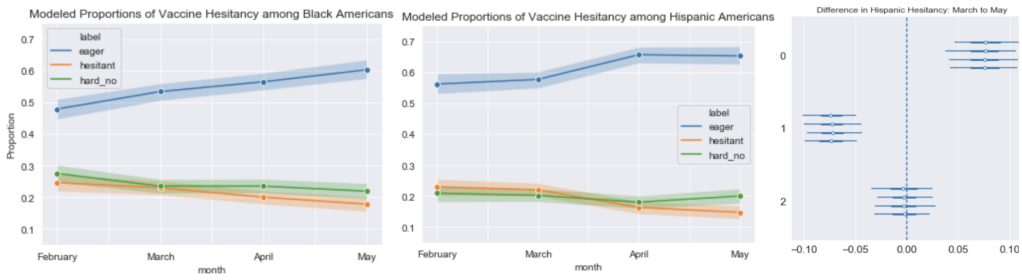

Figure 3: Black (left) vs. Hispanic (middle) change over time; March vs. May Hispanic Change (right)

4.3 Varying Priors

A main point of interest in this analysis is what happens to the posterior distributions if we vary the prior weight on the Dirichlet distribution. I chose the four weights of 50, 200, 500, and 1000, which ranges from nearly meaningless to very influential. Among Black Americans, we can see from Table 1 that the prior increases the percent hesitant but also decreases the standard deviations. The latter is to be expected as we are de-facto increasing the amount of data in the model by having a more influential prior. However, if we look at the proportions in February, the opposite happens. Hesitancy goes down as the weight increases and eagerness significantly increases. This makes sense because the prior is uniform across all months, and increasing the prior weight causes the likelihood to have less influence, meaning the estimates will converge. In fact, the CDC hesitancy estimates were created in April, so it would make sense that the prior could heavily skew February and May estimates. In fact, the April means have a much smaller variation, from 0.577 (weight 50) to 0.559 (weight 1000), helping verify the prior's validity.

Prior weight	Eager	Hesitant	No
50	0.656 $\pm$0.022	0.145	0.199
200	0.630 $\pm$0.021	0.161	0.209
500	0.603 $\pm$0.018	0.178	0.219
1000	0.583 $\pm$0.016	0.191	0.226

Table 1: May proportions for Black Americans with different Prior Weights

Prior weight	Eager	Hesitant	No
50	0.408 $\pm$0.025	0.281	0.311
200	0.444 $\pm$0.022	0.263	0.294
500	0.479 $\pm$0.019	0.246	0.275
1000	0.504 $\pm$0.017	0.234	0.262

Table 2: February proportions for Black Americans with different Prior Weights

I examined the same data for Hispanic Americans as well and encountered similar results.

5 Discussion

I have found that Black and Hispanic Americans' attitudes have become more favorable to the vaccine over time, and the Bayesian framework allowed me to capture the inherent uncertainty surrounding these estimates. There are certainly ways to improve this analysis such as adding temporal components to the prior, or adding in correlations between different subgroups as they are not inherently independent. Also, further model evaluation with, for example, a posterior predictive distribution and/or cross validation would help further assess the model's predictive power rather than just its descriptive power.

While this analysis is imperfect, I believe it could be an important roadmap to analyzing other subgroups' attitudes in more detail. There was enough data for both Black and Hispanic people that the error bands remained rather narrow, but applying this to those with more limited data such as Asian Americans, Native Americans, or possibly combinations of multiple subgroups could prove to be a more tact use of these Bayesian techniques. The public could be more well-informed with increased nuance, and policymakers could have a better idea on how to target certain interventions, which undeniably would be a big step towards ending this pandemic.

References

[1] Alexandre Figueiredo et al. "Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study". In: *Lancet (London, England)* 396 (Sept. 2020). DOI: 10.1016/S0140-6736(20)31558-0.

[2] Merlin Heidemanns, Andrew Gelman, and G. Elliott Morris. "An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election". In: *2.4* 2.4 (Oct. 2020). DOI: 10.1162/99608f92.fc62f1e1. URL: https://doi.org/10.1162%5C%2F99608f92.fc62f1e1.

[3] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. 2011. arXiv: 1111.4246 [stat.CO].

[4] *Latest Polls*. June 2021. URL: https://projects.fivethirtyeight.com/polls/.

[5] *Vaccine Hesitancy for COVID-19*. URL: https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/.