cross institution draft

```
## Loading required namespace: rjags
```

```
## Note: Summary statistics were not produced as there are >50 monitored
## variables
## [To override this behaviour see ?add.summary and ?runjags.options]
## FALSEFinished running the simulation
```

Methods

We decide to use Beta-Binomial models to study the positive COVID test rates across these institutions. In our setting, we assume that the number of people tested as positive follows a binomial distribution with each test being an independent Bernoulli trial. We also think that there exist potential factors that could affect the positive test rate of these institutions and the relationship between the positive test rate and these factors are linear. Some other assumptions we have are independent and constant variance. The factors considered include the institution type (University versus LAC), institution enrollment, and the test rate (calculated as the total number of tests administrated from 2020 to 2021 divided by enrollment). We also assume that there exists some baseline positive test rate for all educational institutions in the US. These factors will be included in our model as regressors in the link function of positive test rate p_j, where the subscript j denotes the specific institution.

For the first model (Model1), we used the logistic regression as a link function on p_j, and the regressors included are the intercept, the institution type, the standardized enrollment, and the standardized test rate. We assume that the intercept β_0, the coefficient for institution type β_1, the coefficient for standardized enrollment β_2, and the coefficient for standardized test rate β_3 all follows weakly informative priors $N(0, 100)$. Below are the equations for Model1:

- Sampling: for j in 1,…30, n_j:

$$\begin{gathered}\mathbf{Y}_j|\mathbf{p}_j,\mathbf{n}_j \overset{i.i.d.}{\sim} \binom{\mathbf{n}_j}{\mathbf{p}_j} \\ \mathbf{p}_j = \beta_{\mathbf{0}} + \beta_{\mathbf{1}} * Type + \beta_{\mathbf{2}} * Zscore\ Enrollment + \beta_{\mathbf{3}} * Zscore\ TestRate\end{gathered} \tag{1}$$

- Prior for $\mathbf{p}_j$, for j= 1,…30:

$$\begin{gathered}\beta_{\mathbf{0}} \sim \mathcal{N}(0,\, 100) \\ \beta_{\mathbf{1}} \sim \mathcal{N}(0,\, 100) \\ \beta_{\mathbf{2}} \sim \mathcal{N}(0,\, 100) \\ \beta_{\mathbf{3}} \sim \mathcal{N}(0,\, 100)\end{gathered} \tag{2}$$

In the second model (Model2), we assume that our positive test rate p_j follows a Beta distribution with parameters a_j and b_j. We will use reverse elicitation to make posterior inferences on a_j and b_j. We define the mean of the Beta distribution as $\mu_j = \frac{a_j}{a_j+b_j}$ and the sample size as $\eta = a_j + b_j$. Then we use the logistic regression as a link function on μ_j with regressors as the intercept, the institution type, and the test rate. Similarly, we assume the intercept β_0, the coefficient for institution type β_1, and the coefficient for test rate β_3 all follows weakly informative priors $N(0, 100)$. The equations for Model2 are shown below:

- Sampling: for j in 1,. . . 30, n_j:

$$\begin{aligned} \mathbf{Y}_j|\mathbf{p}_j, \mathbf{n}_j &\overset{i.i.d.}{\sim} \binom{\mathbf{n}_j}{\mathbf{p}_j} \\ \mathbf{p}_j &\sim Beta(a_j, b_j) \end{aligned} \tag{3}$$

- Prior for $\mathbf{p}_j$, for j= 1,. . . 30:

$$\begin{gathered} \mathbf{a_j} = \phi_j * \mu_j \\ \mathbf{b_j} = \phi_j * (1 - \mu_j) \\ logit(\mu_j) = \beta_0 + \beta_1 * type_j + beta3 * test_rate_j \\ \phi_j = exp(logeta_j) \\ logeta_j \sim logis(logn, 1) \end{gathered} \tag{4}$$

- Hyperprior

$$\begin{gathered} \beta_{\mathbf{0}} \sim \mathcal{N}(0,\, 100) \\ \beta_{\mathbf{1}} \sim \mathcal{N}(0,\, 100) \\ \beta_{\mathbf{3}} \sim \mathcal{N}(0,\, 100) \\ logn = log^{100} \end{gathered} \tag{5}$$

For both Model1 and Model2, we made an additional 10000 draws after an adaptation period of 1000 draws and a burn-in period of 5000 draws. We keep every 5th draws to reduce the effect of temporal correlation between consecutive MCMC draws. For convergence consistency, we ran 3 chains for both models.

- Diagnostics: For Model1, the trace plots and ACF plots show that our draws are well-mixed and our parameters have converged to their posterior region. The overlaid density plots show that the draws from 3 separate chains all converged to the same distribution with roughly the same density curves. To make sure that Model1 gives realistic predictions as what we would see in real life, we simulated data sets using Model1 and compared them to the observed data. We first checked the overlayed density plot; of the 100 simulated data sets we randomly selected, they all have similar density curves as the density curve of the observed data. Then we compared the observed mean number of the positive test of each institution to the distribution of the 10000 mean number of the positive test of the simulated data to check if the posterior estimates of the mean of number the positive test using p_j are reasonable. We discovered that the observed mean lies in the center of the distribution of the simulated data mean for almost all institutions, which indicates that a data set like the one we observed is not uncommon compared to the data simulated using Model1, which confirmed that Model1 is adequate.

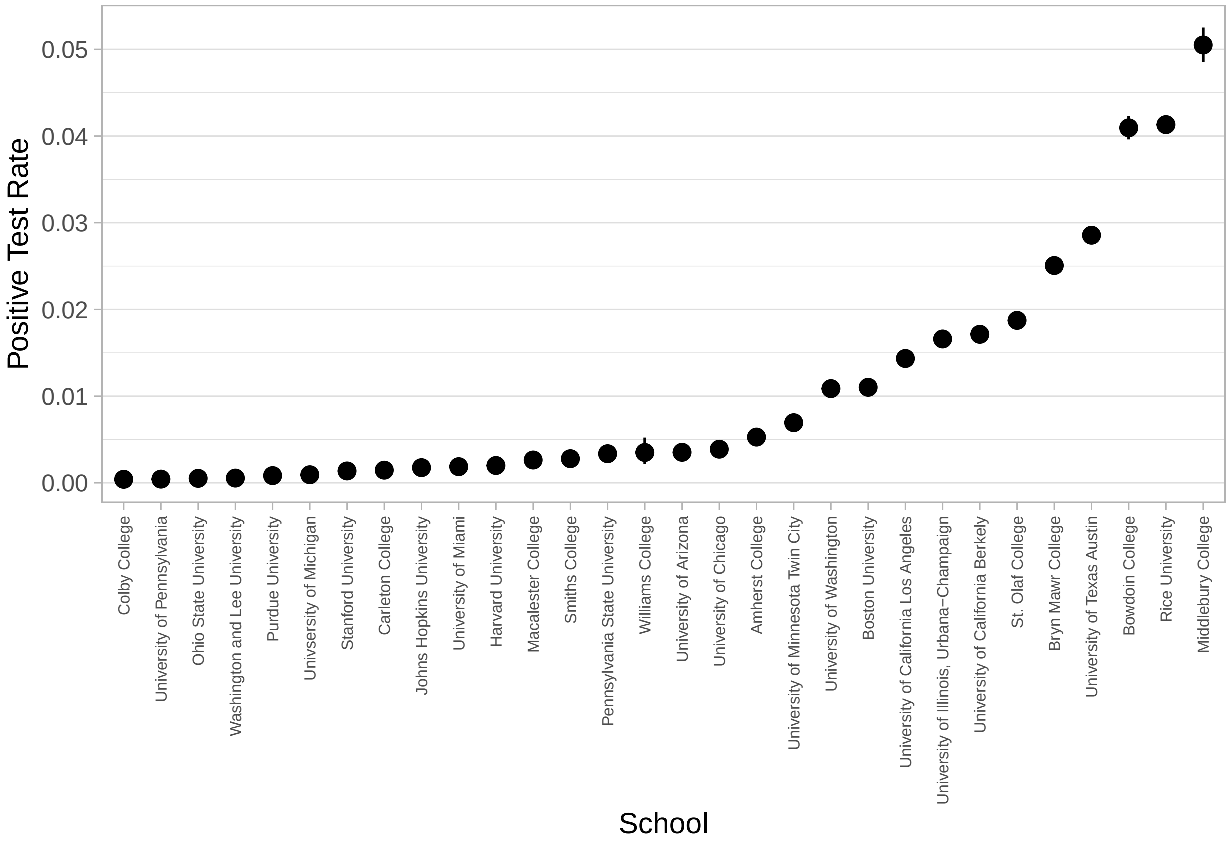
For Model2, we also see that our draws are well-mixed and our parameters have converged to their posterior region from checking the trace plots and ACF plots. The overlaid density plots also show that the draws from three separate chains all converged to the same distribution with roughly the same density curves. Similarly, we also conducted the posterior predictive check on Model2. The 100 simulated data have a very similar density curve to the density curve of the observed data. As in Model1, checking the posterior estimates of the mean number of positive tests confirmed the fact that our model is able to generate data that mimic the observed data, and therefore we will be able to conclude that Model2 is adequate as well.

Results

With model1, we know that if the enrollment number and test rate are at their average level among these institutions, the positive test rate of a University is 0.0142 (β_0), and for a LAC is 0.0071 ($\beta_0 + \beta_1$. Looking at the credible intervals for these coefficients, we also know there is a 95% chance that the baseline positive test rate β_0 for University is between 0.0140 to 0.0143; for β_1, we know that the baseline positive test rate of LAC is from # to # lower than the baseline positive test rate of University.

Holding the institution type and test rate constant, with every one standard deviation increase in the enrollment, there will be a 16.0% increase in the positive test rate (β_2), and we know that this increase will be from # to # with a 95% probability. Holding the institution type and enrollment constant, with every one standard deviation increase in the test rate, there will be a 72.3% decrease in the positive test rate (β_3), and we know that this increase will be from # to # with a 95% probability.

The posterior inference made on p_j using Model1 is shown in Fig.# below:



Using Model2, we know that if the test rate is at their average level among these institutions, the mean of the distribution of positive test rate of a University is 0.0452 (β_0), and for LAC is 0.0152 ($\beta_0 + \beta_1$. There's a 95% chance that the baseline positive test rate for University is from # to #, and for LAC is from # to #. Holding the institution type constant, with every one standard deviation increase in the test rate, there will be a 95.2% increase in the positive test rate (β_3). This increase could range from # to #, with a 95% possibility.