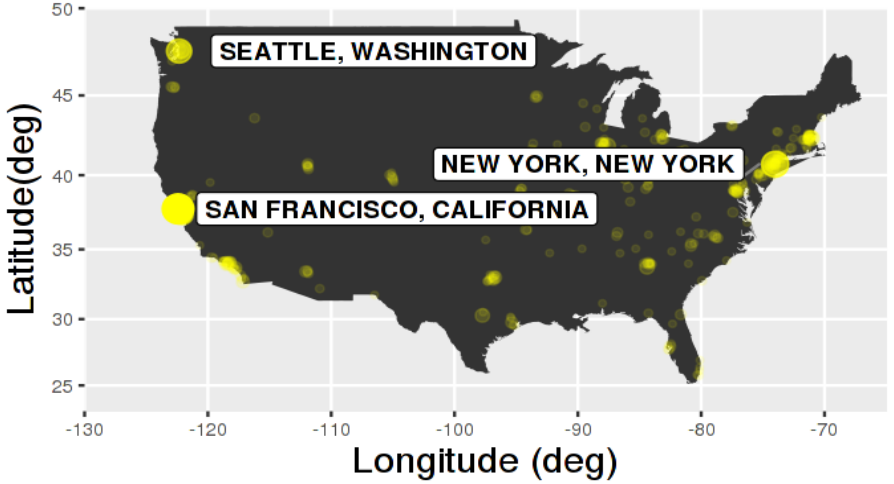
What happens before your H-1b application finds its way to the Visa lottery pool

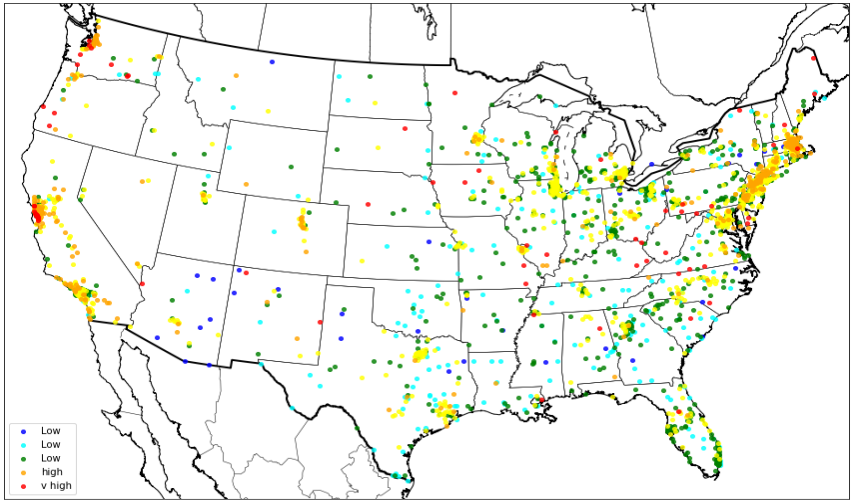
Background

- **The Big Question for every graduate**: which job will sponsor your future H-1b visa?
- For H-1b applicants, whether you would "win" the visa lottery means more than a lucky draw from the pool.

The Data

- The data is extracted from Kaggle H-1b Visa Petitions 2011-2016, the raw data could be found at the Office of Foreign Labor Certification (OFLC) website.
- The dataset contains 3 millions records of H-1b partitions from year 2011 to 2016 after LCA processing
- The LCA approved cases are ready for lottery



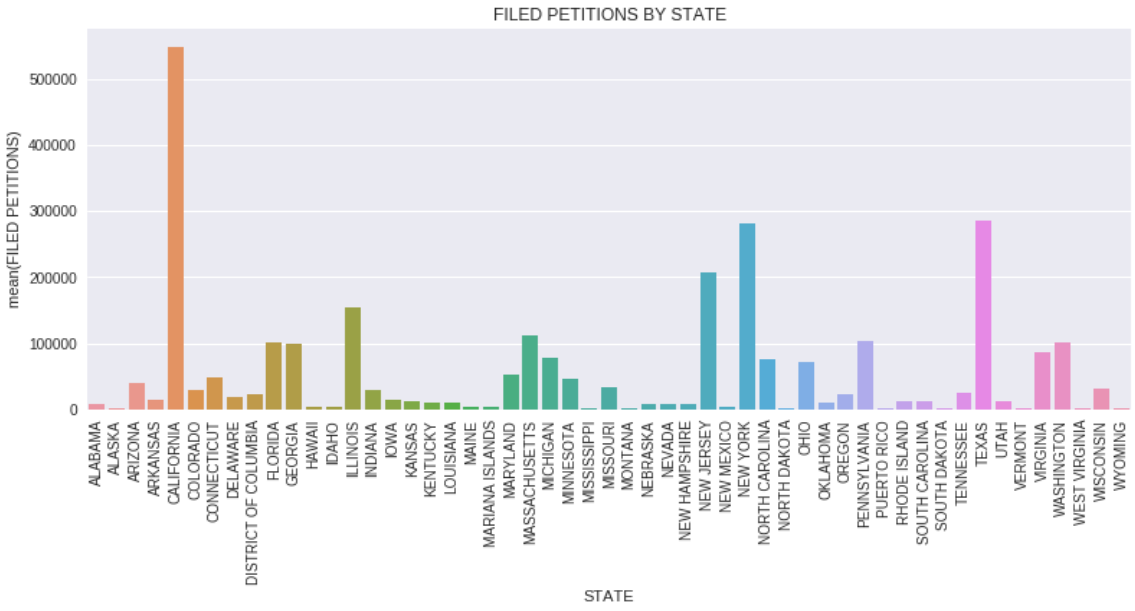


What happens before your H-1b application finds its way to the Visa lottery pool

The Questions Need Addressing

- Core question to address: the **probability of getting certified** after the LCA process.

- First, it's like buying a house: location**, location, LOCATION**
- Second, given location, can I do the job?



Variables Explained

- **CASE_STATUS**: Certified/Denied/ Withdrawn; **"CERTIFIED"** applicants are **eligible** to buy the ticket to lottery
- **SOC_NAM**E: Standard Occupational Classification (SOC), painfully detailed classification
- **WORKSITE**: City and State of your intended area of employment, we'll re-measure it to State level
- **PREVAILING_WAGE**: **average** wage paid to similarly employed workers in the requested occupation in the area of intended employment.

A hierarchical model to address the location problem, aka where should I move to after graduation

The Model

- The target: the probability of getting the stamp of Certified/Denied/Withdrawn
- It's easy to model the data with Multinomial distribution:

$$y \overset{iid}{\sim} Multinomial(\theta_{1j}, \theta_{2j}, \theta_{3j})$$

- where $\theta_{ij}, \quad i = 1, 2, 3$ is defined as the probability of getting certified/denied/withdrawn respectively

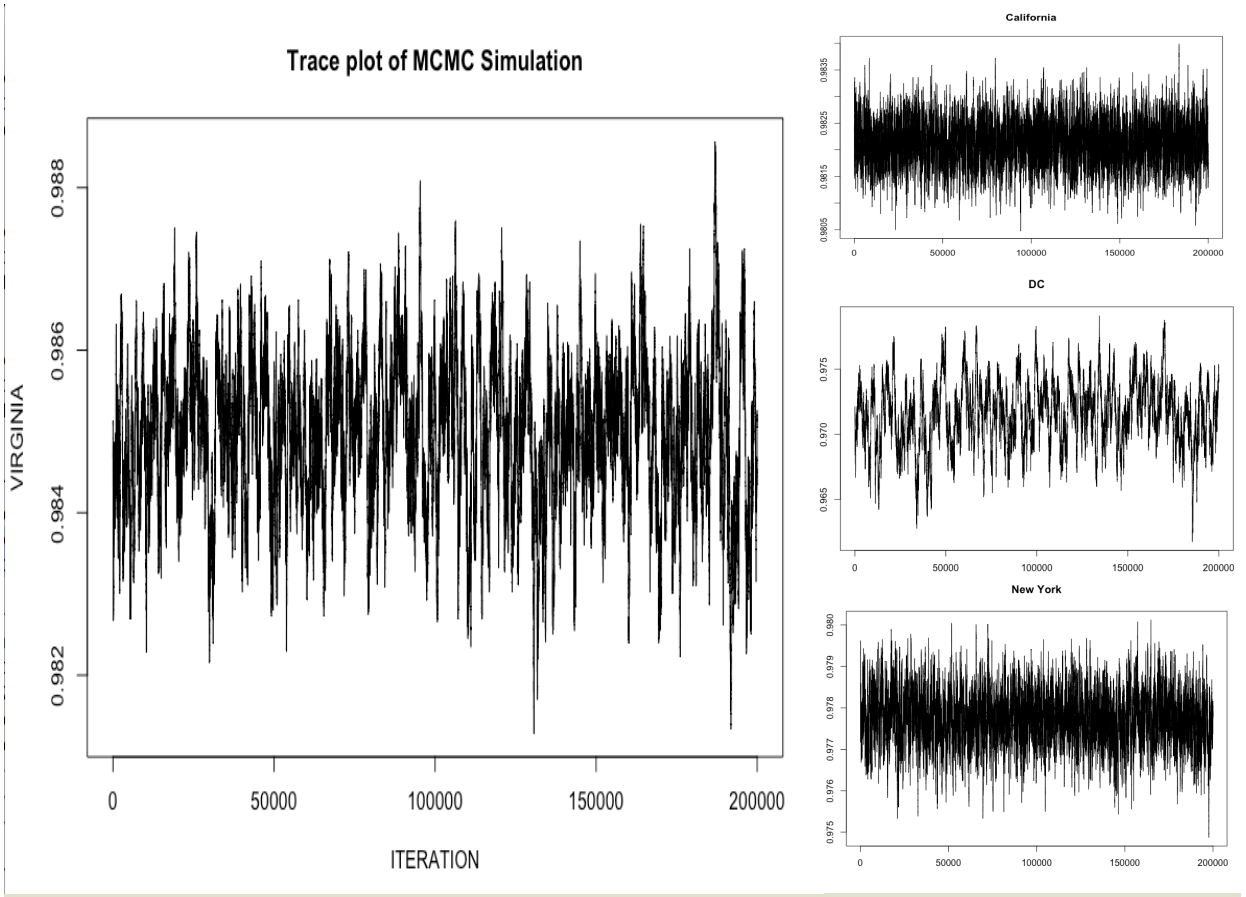
$$\text{let } \alpha_{ij} = \frac{\theta_{1j}}{\theta_{1j}+\theta_{2j}} \text{ and } \alpha_{2j} = 1 - \theta_{3j}$$
$$\text{and } \beta_{ij} = logit(\alpha_{1j}),\ \beta_{2j} = logit(\alpha_{2j})$$

- And we assign multivariate normal for the beta

	Certified	Denied	Withdrawn
ALBAMA	1733	41	279
ALASKA	154	13	32
ARIZONA	8692	105	937
ARKANSAS	3057	38	311
CALIFORNIA	104070	1892	13778
COLORADO	5507	113	882
CONNECTICUT	9100	100	835
DELAWARE	3081	25	416
DC	3566	106	462
…			

First 9 observations of the summary data

A hierarchical model to address the location problem, aka where should I move to after graduation



Simulation trace plot of alpha1, the probability of certified

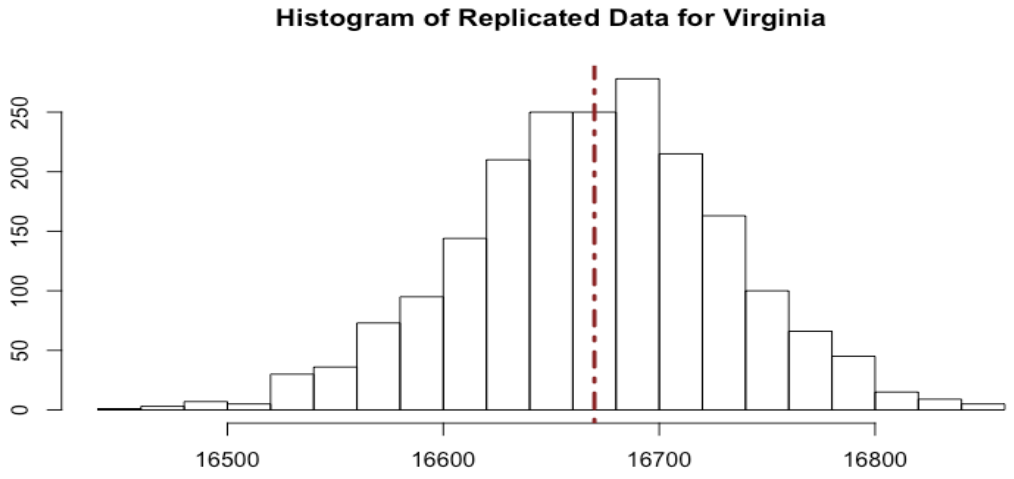
The Simulation

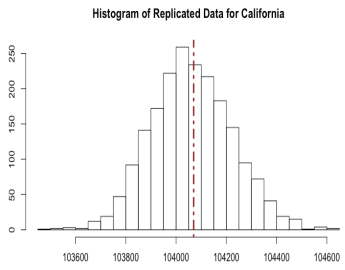
- The hierarchical contains 104 parameters plus 5 hyper-parameters, and the marginal posterior distribution is not well defined, so **Metropolis Algorithm** is used in simulation.
- The "random walk" **jumping distribution** is used for hyper-parameters, and multivariate normal is set as jumping distribution for each pair of beta.
- Realized average acceptance rate of 20%

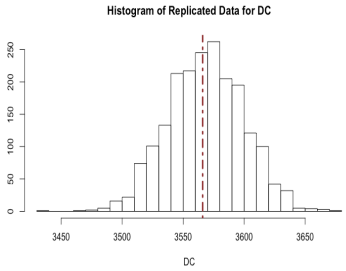
Posterior check suggests the model fits the data relatively well, and we can get to the fun part

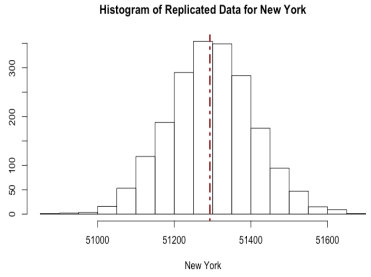
Checking the replicates

- 2000 samples are drawn from the simulated results, and the number of cases certified are generated and plotted.









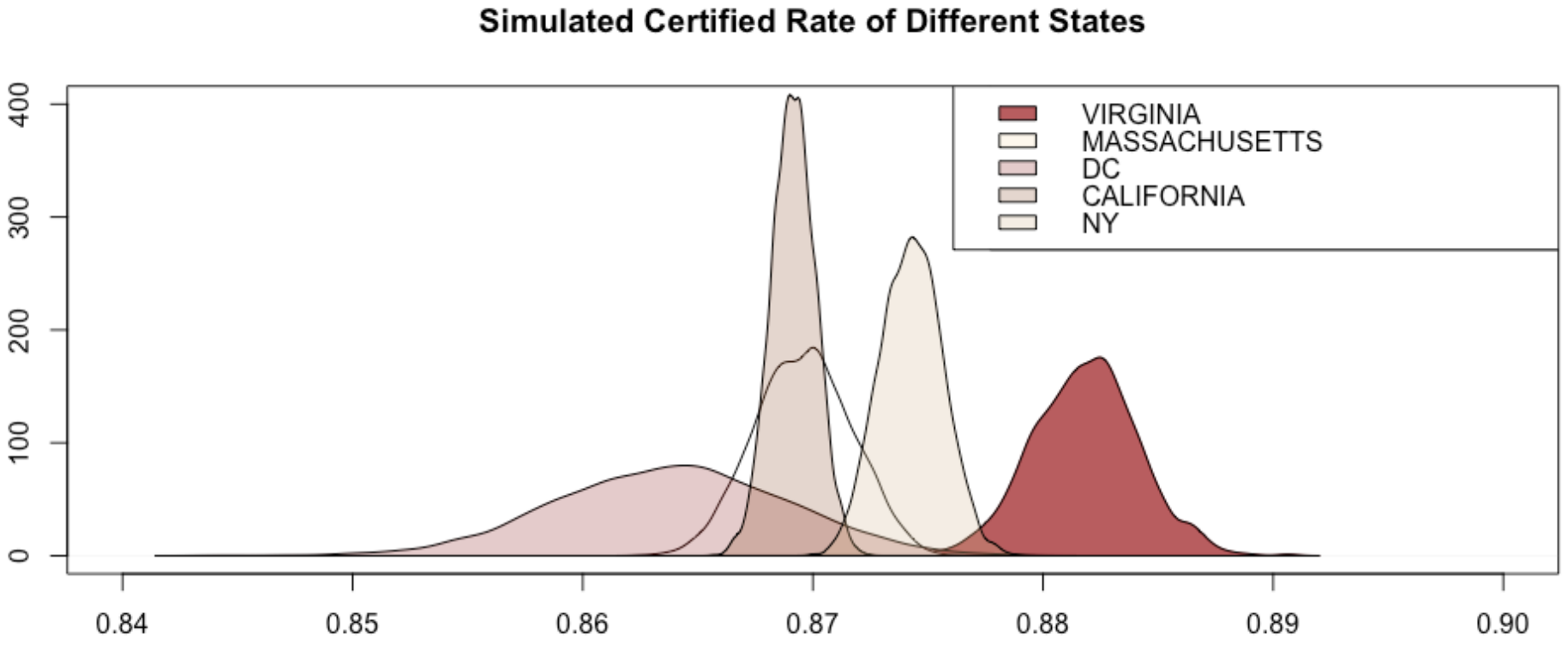
Quantiles of Simulated Parameters

	VA	DC	NY	CA
2.5%	0.877	0.854	0.872	0.867
25%	0.880	0.860	0.873	0.868
50%	0.881	0.864	0.874	0.869
75%	0.883	0.867	0.875	0.870
97.5%	0.886	0.874	0.877	0.871

Posterior check suggests the model fits the data relatively well, and we can get to the fun part

Check the simulated data: Virginia is GOOD!

- Simulated data suggests there is a difference between the certified rate between states, and so far staying in Virginia seems like a good choice.



After deciding that we stay put in Virginia, let's look at the job aspect in Virginia

A Simple Model

- Top 25 jobs with the largest amount of applicants for H-1b in year 2016 are choose, the rest is categorized as "other"
- Previous year's data is used as prior information
- Follow the Multinomial model in previous question, we have the similar set up for the model:

$$y \overset{iid}{\sim} Multinomial(\theta_{1j}, \theta_{2j}, \theta_{3j})$$

where $\theta_{ij}, \quad i = 1, 2, 3$ is defined as the probability of getting certified/denied/withdrawn respectively

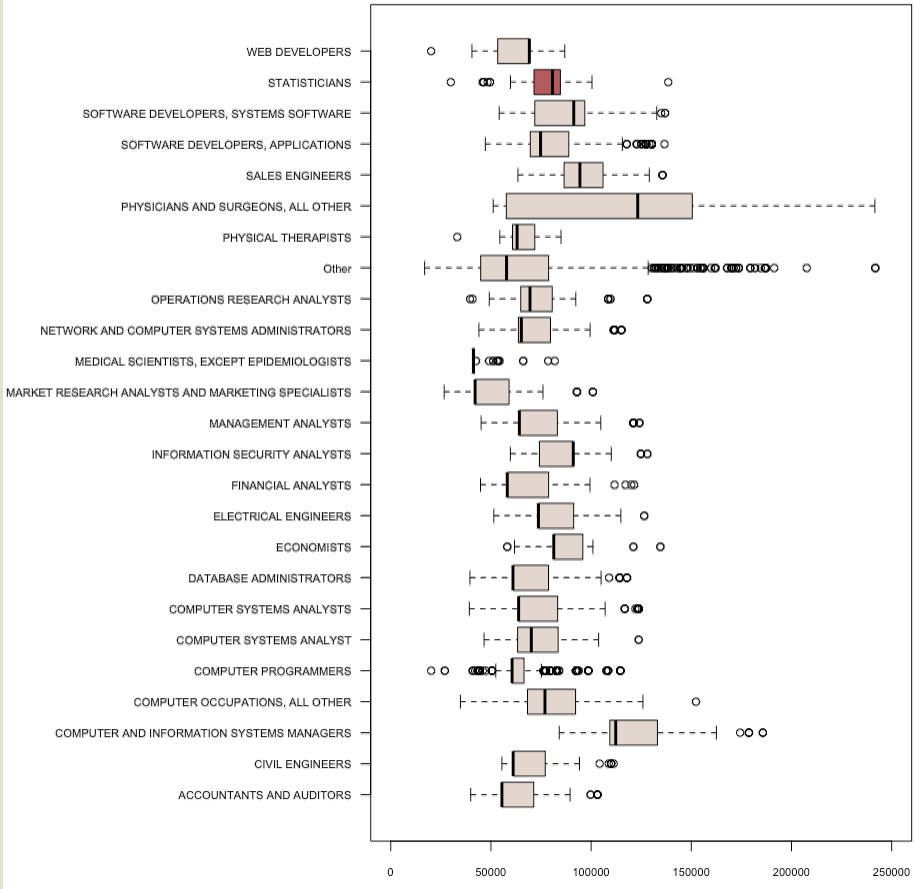
assign a conjugate prior

$$p(\theta_{.j} \sim Dir(\rho_{1j}\alpha_{1j}, \rho_{2j}\alpha_{2j}, \rho_{3j}\alpha_{3j})$$

where α_{ij} are previous year's data

ρ is a hyperparameter representing the weight of previous year's data

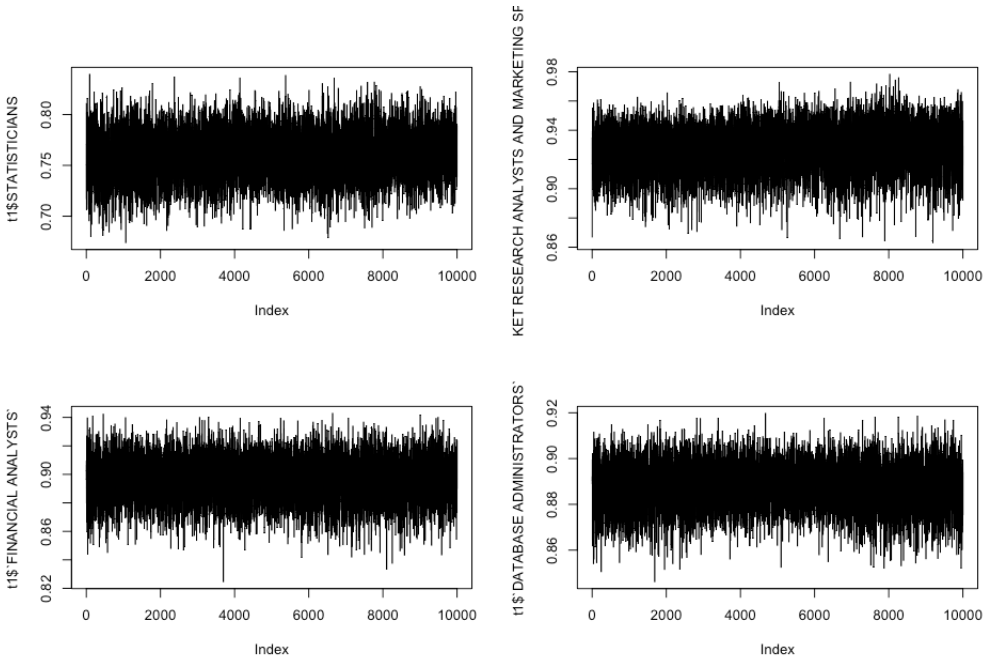
Print the PDF image

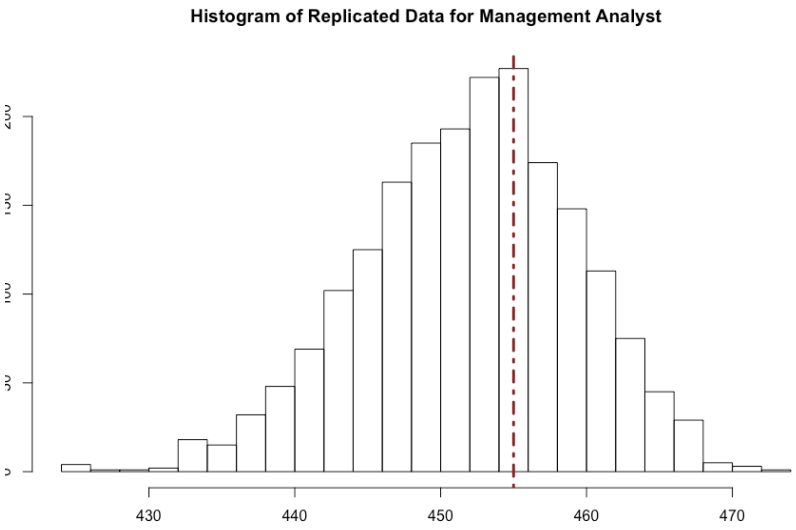


After deciding that we stay put in Virginia, let's look at the job aspect in Virginia

The Simulation

- Gibbs-Metropolis algorithm is used to simulate the parameters
- The thetas are sampled from Beta distribution (marginal of Dirichlet Distribution), and Metropolis algorithm is used for rho

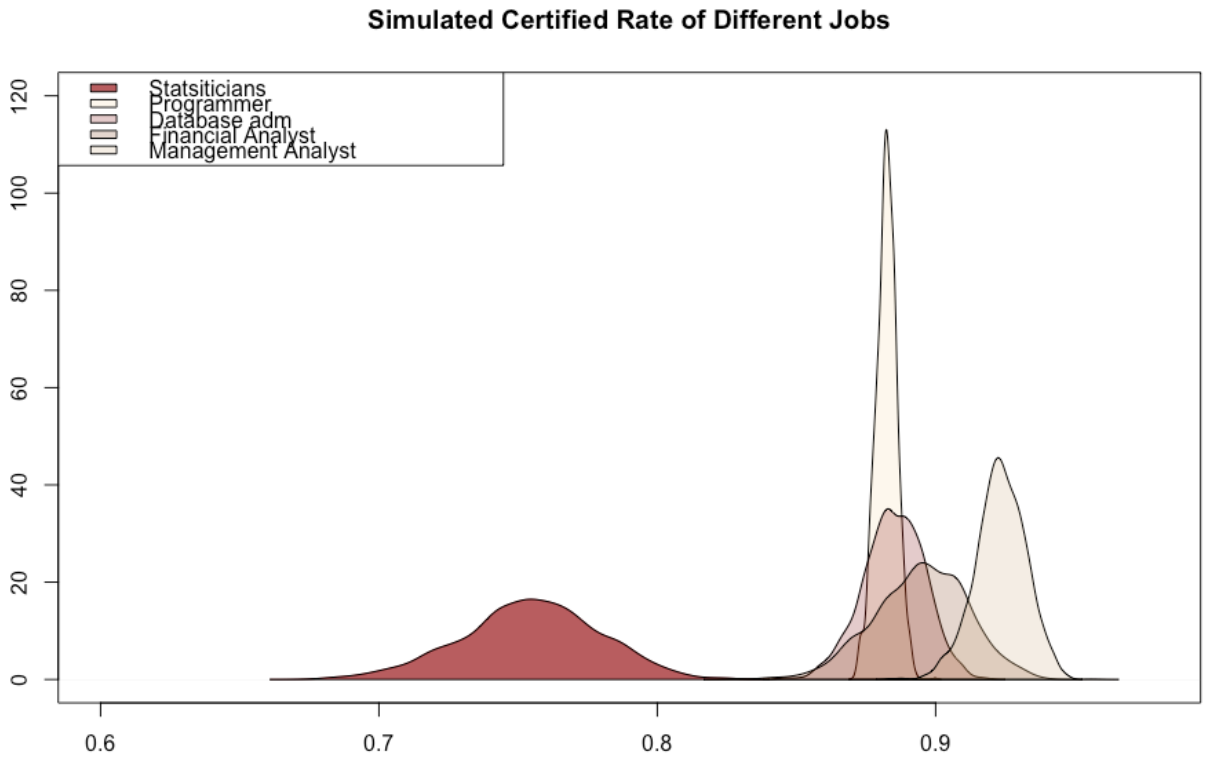




After deciding that we stay put in Virginia, let's look at the job aspect in Virginia

Simulate Posterior Distribution

- The simulated results suggest that there is a difference between certified rate among jobs, and the perspective for statisticians is not too optimal



Quantiles of Simulated Parameters

	Stat	FA	MA	CP
2.5%	0.703	0.860	0.904	0.875
25%	0.739	0.884	0.918	0.880
50%	0.755	0.895	0.923	0.882
75%	0.771	0.907	0.930	0.885
97.5%	0.801	0.925	0.940	0.890

Conclusion and further work

Conclusion

- The 2 Bayesian model offers some insights into the probability of your H-1b applications getting certified to be in the lottery
- So far as I can tell, stay in Virginia, and maybe look for jobs in Financial/Management analyst, or if programming works for you, in computer programming.

Further

- The two models, hierarchical as they are, are still simple models.
- Combined variables from other data source, e.g. economic development, demographic factors, etc, may offer more insight into the question.