Bayesian Vaccine Coverage Rate Research in Nigeria

Qi Wang[1]
Department of Statistics, University of California, Santa Cruz[1]

Table 1: Example of Data from Abia

	cluster	region	y	n
1069	755	Abia	3	3
1070	756	Abia	6	6
1071	757	Abia	3	3
1072	758	Abia	3	4
1074	760	Abia	4	4

Abstract

In this report, we are going to explore the vaccination coverage for MCV1 among children in Nigeria in 2018 in Bayesian approach. Non-hierarchical and hierarchical models are both included.

KEY WORDS: Bayesian Hierarchical Model, Beta-Binomial Model, Model Selection

1. Data Description

The data in this report describes the vaccination coverage for the first dose of measles-containing-vaccine(MCV1) among children aged 12-24 months in Nigeria in 2018. It is clustered in each row, which describes one census enumeration area like a collection of households. And in each cluster, the number of children vaccinated is recorded as y, and the number of eligible children is recorded as n. Also, each cluster belongs to one of the 37 level 1 administrative areas. Take data from Abia as an example, here is some rows of data in table 1.

2. Descriptive Statistics And Exploratory Data Analysis

2.1 Overview of Data

Since we are interested in the overall vaccine coverage rate for all the regions, here is an overall plot of all the regions' coverage rate in figure 1. The difference vaccine coverage rate is obvious among different regions of the country. Some regions like Lagos, Anambra have a coverage rate more than 0.8, on the contrary, for other countries like Sokoto and Zamfara, the coverage rate is only around 0.2. Furthermore, as shown in figure 2, we can also tell difference between regions through the region wise box plot. However, we need more concrete statistical results, which will be discussed in the latter part of this report.

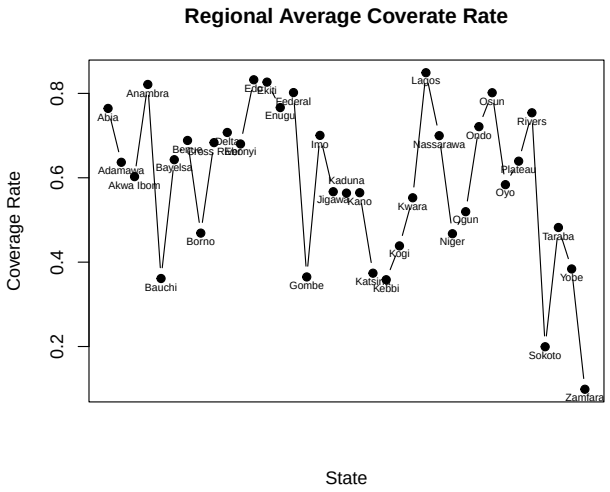

Figure 1: Overall Coverage Rate for Each Region

2.2 Logistic Regression

Since we are comparing the coverage probability of each region, I will use a basic logistic regression to check the coverage rate difference with Abia being the reference group. As shown in table 2, these are regions that have significant coverage rate difference from Abia at 0.05 significance level. Therefore, we need to consider the between group difference and within group difference at the same time. Furthermore, a t test to compare the probability is also conducted in table 3, which includes only a subset of t-test due to the limitation of the space, and 1 means there are significant difference, 0 means no significant difference at significance level 0.05. From the boxplot in the last section and the test or regression result in this subsection, we can know that there exists between group variation and within group variation.

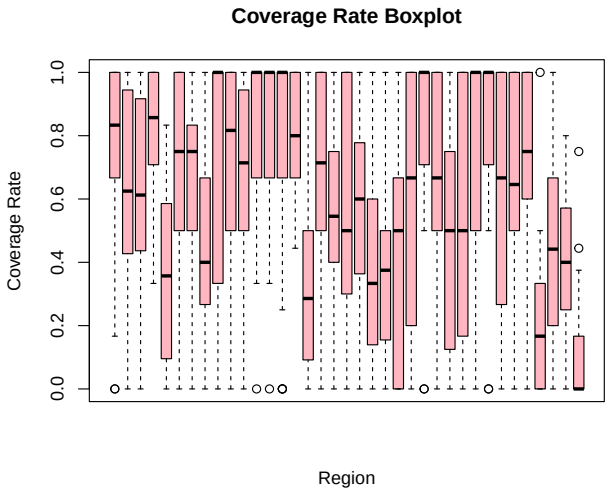

Figure 2: Boxplot of Overall Coverage

Table 2: Logistic Regression Coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
Bauchi	-0.7652	0.1886	-4.0570	0.0000
Borno	-0.6166	0.1952	-3.1594	0.0016
Gombe	-0.8745	0.1964	-4.4520	0.0000
Kaduna	-0.5362	0.1884	-2.8457	0.0044
Katsina	-0.8065	0.1911	-4.2199	0.0000
Kebbi	-0.8446	0.1956	-4.3174	0.0000
Kogi	-0.5521	0.2299	-2.4018	0.0163
Kwara	-0.4239	0.2103	-2.0154	0.0439
Niger	-0.6179	0.1919	-3.2204	0.0013
Sokoto	-1.4457	0.2329	-6.2071	0.0000
Taraba	-0.7025	0.1978	-3.5509	0.0004
Yobe	-0.5995	0.1921	-3.1212	0.0018
Zamfara	-2.0536	0.2763	-7.4320	0.0000

Table 3: t.test Result

	Oyo	Plateau	Rivers	Sokoto	Taraba
Oyo	1	0	1	1	0
Plateau	0	1	0	1	1
Rivers	1	0	1	1	1
Sokoto	1	1	1	1	1
Taraba	0	1	1	1	1

3. Bayesian Beta-Binomial Conjugate Model

3.1 Model Setting and Assumption

Since we have we have proved the existence of the within group variation and between group variation, we can propose a beta-binomial model for this setting. We are assuming that:

$$Y_i \mid N_i \sim \ Binomial(N_i \ , p_{s[i]})$$

where $s[i] \in \{1, 2, 3..., 37\}$ is the indicator of the area that the cluster i belongs to. Furthermore, for $p = (p_1, p_2, ..., p_J)$, we assume that:

$$p_j \sim_{iid} Beta(\alpha, \beta)$$

3.2 Getting Posterior Samples

According to Bayes Theorem, we can derive the posterior $\pi(p_i|Y_{i.})$ in which, $Y_{i.}$ is a reshape of data that for all the data that belongs to region i, and $i \in \{1, 2, 3..., 37\}$, that is:

$$\pi(p_i|Y_{i.}) \propto \pi(p_i) \times f(Y_{i.}|p_i)$$

Furthermore,

$$f(Y_{i.}|p_i) \propto \prod_{j=1}^{n_i} p_i^{Y_{ij}}(1-p_i)^{N_{ij}-Y_{ij}}$$

where the n_i is the number of clusters in region i. Therefore,

$$\pi(p_i|Y_{i.}) \propto p_i^{\alpha-1}(1-p_i)^{\beta-1} \times \prod_{j=1}^{n_i} p_i^{Y_{ij}}(1-p_i)^{N_{ij}-Y_{ij}}$$

$$= p_i^{\alpha+\sum_{j=1}^{n_i} Y_{ij}-1}(1-p_i)^{\beta+\sum_{j=1}^{n_i}(N_{ij}-Y_{ij})-1}$$

It is still a Beta distribution kernel, therefore,

$$p_i|Y_{i.} \sim Beta(\alpha + \sum_{j=1}^{n_i} Y_{ij} \ , \ \beta + \sum_{j=1}^{n_i}(N_{ij} - Y_{ij}))$$

So we can directly get the posterior distribution by applying R functions. In this model, I set a non-informative prior for p_i, which means I set hyper parameter α and β both to be 1. After sampling from the posterior distribution, the mean of the samples are shown in figure

There still seems to be difference in among the groups, it may be because that I set one non-informative prior so the posterior will be mostly affected by the likelihood, which seems to be similar to the picture in the descriptive statistics.

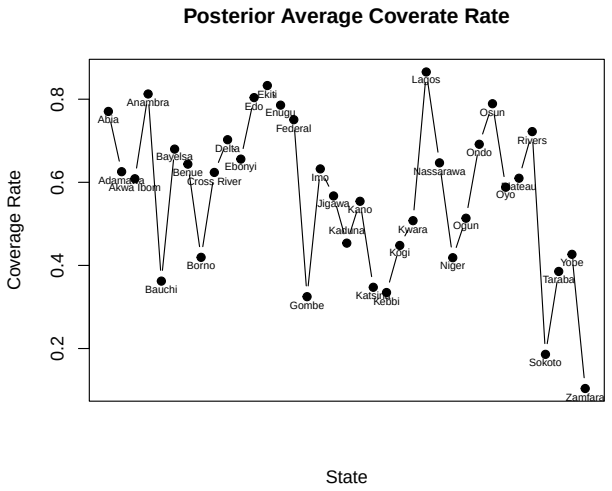

Figure 3: Posterior Mean of Each Region

3.3 Expected Ranking and Distribution

To begin with, with the posterior samples, the regions raking expectation is plotted in figure 4.Therefore, we get the 5 lowest vaccine coverage rate region:

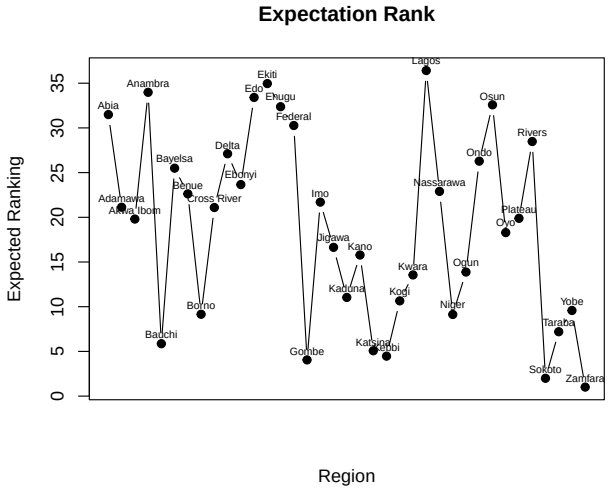

Figure 4: Expectation Rank of Each Region

```
##      [,1]
## [1,] "Zamfara"
## [2,] "Sokoto"
## [3,] "Gombe"
## [4,] "Kebbi"
## [5,] "Katsina"
```

And 5 highest vaccine coverage rate region:

```
##      [,1]
## [1,] "Lagos"
## [2,] "Ekiti"
## [3,] "Anambra"
## [4,] "Edo"
## [5,] "Osun"
```

For the lowest region, the expected ranking distribution for each of them are as presented in figure 5. Smaller rank means the lower coverage rate.

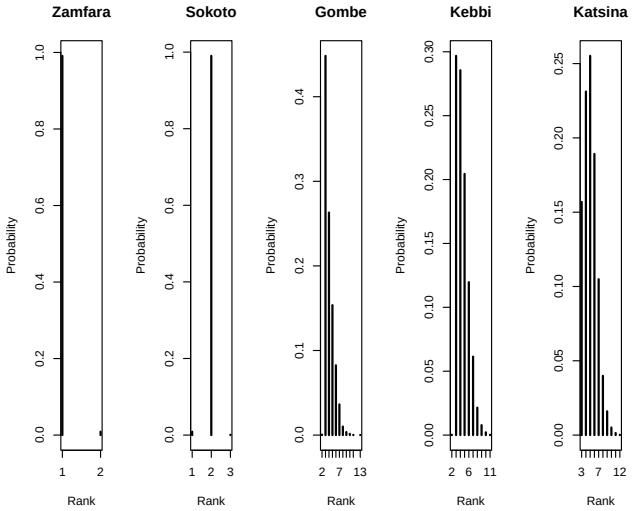

Figure 5: Rank Distribution of Lowest Rate Regions

For the highest region, the expected ranking distribution for each of them are as presented in figure 6. Similarly, higher rank means the higher coverage rate.

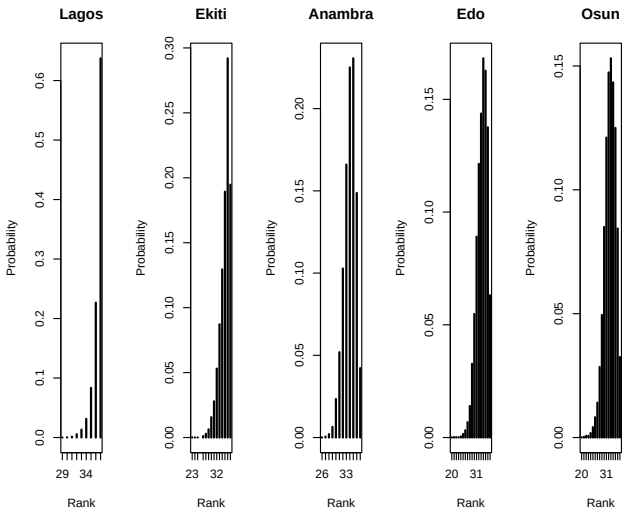

Figure 6: Rank Distribution of Lowest Rate Regions

4. Bayesian Hierachical Model

4.1 Model Setting

Now if we consider another model to represent the data as follows:

$$Y_{ij}|N_{ij} \sim Binomial(N_{ij}, p_i)$$
$$p_i|\mu_i, d \sim_{ind} Beta(\mu_i, d), \quad i = 1, 2, ..., J$$

$$logit(\mu_s) \sim N(0, \sigma_\mu^2)$$
$$logit(d) \sim N(0, \sigma_d^2)$$

After the reparameterization:

$$E(p) = \frac{\alpha}{\alpha+\beta} = \mu$$
$$Var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \mu(1-\mu)d$$

By the law of total expectation:

$$E(Y|\mu, d) = E(E(Y|p, \mu, d)) = nE(p) = n\mu$$

Also, by the law of total variance:

$$Var(Y|\mu, d) = E(Var(Y|p, \mu, d)) + Var(E(Y|p, \mu, d)) =$$
$$n\mu - n(\mu^2 + \mu(1-\mu)d) + n^2\mu(1-\mu)d$$

Therefore, in this way, we reduce the dependence of the mean and variance parameters.

First of all, we need to get the joint posterior distribution of all parameters, and here I will still use α_i and β_i then reparameterize later.

Because we have the equation for μ_i and d by equations:

$$\mu_i = \frac{\alpha_i}{\alpha_i+\beta}, \; d = \frac{1}{\alpha_i+\beta+1}$$

We have:

$$\alpha_i = \mu_i(\frac{1-d}{d}), \beta = (1-\mu_i)(\frac{1-d}{d})$$

Then put the reparameterized α_i and β to replace the μ_i and β in the distribution above will give us the posterior. The joint posterior will be:

$$\pi(p, \mu, d|Y_{ij}, Nij) \propto \prod_{i=1}^{J}\prod_{j=1}^{n_i} p_i^{Y_{ij}}(1-p_i)^{(N_{ij}-Yij)}$$

$$\times \prod_{i=1}^{J} \frac{1}{\beta(\alpha_i, \beta)} p_i^{\alpha_i-1}(1-p_i)^{\beta-1}$$

$$\times \pi(\mu) \times \pi(d)$$

Then I will get the full conditional distribution of each parameter:

$$\pi(p_i|others) \propto p_i^{\sum_{j=1}^{n_i} Y_{ij}+\alpha_i-1}(1-p_i)^{\sum_{j=1}^{n_i}(N_{ij}-Y_{ij})+\beta-1}$$

$$\sim Beta(\sum_{j=1}^{n_i} Y_{ij} + \alpha_i, \sum_{j=1}^{n_i}(N_{ij}-Y_{ij}) + \beta)$$

$$\pi(logit(\mu_i)|others) \propto exp(-\frac{logit(\mu_i)^2}{2\sigma_\mu^2}) \times \frac{1}{\beta(\alpha_i, \beta)} p_i^{\alpha_i-1}(1-p_i)^{\beta-1}$$

in which:

$$\alpha_i = logit^{-1}(logit(\mu_i))(\frac{1-logit^{-1}(logit(d))}{logit^{-1}(logit(d))}),$$

$$\beta = (1-logit^{-1}(logit(\mu_i)))(\frac{1-logit^{-1}(logit(d))}{logit^{-1}(logit(d))})$$

$$\pi(logit(d)|others) \propto exp(-\frac{logit(d)^2}{2\sigma_d^2}) \times$$

$$\prod_{i=1}^{J} \frac{1}{\beta(\alpha_i, \beta)} p_i^{\alpha_i-1}(1-p_i)^{\beta-1}$$

And here is the posterior histogram of d in figure 7. Also, here is a posterior for μ_i of a subset of the regions in figure 8.

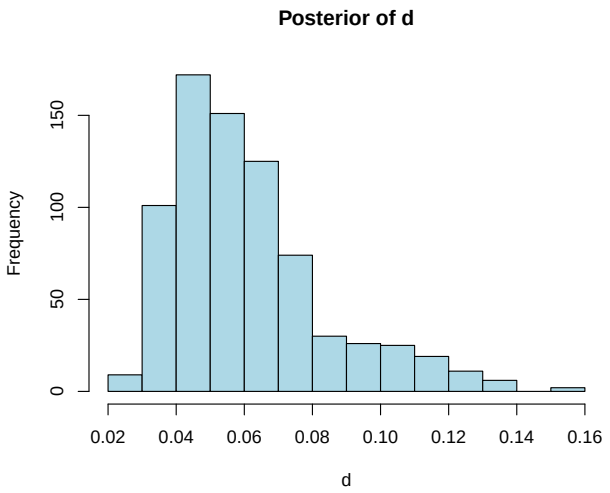

Figure 7: Posterior of d

4.2 Posterior Sample

After using Metropolis Hasting within Gibbs sampler, we have the posterior distribution for each μ_i compared with

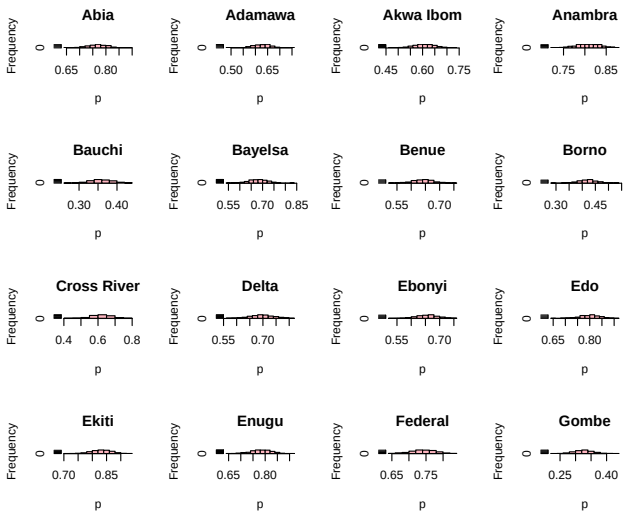

Figure 8: Hierarchical Posterior of p

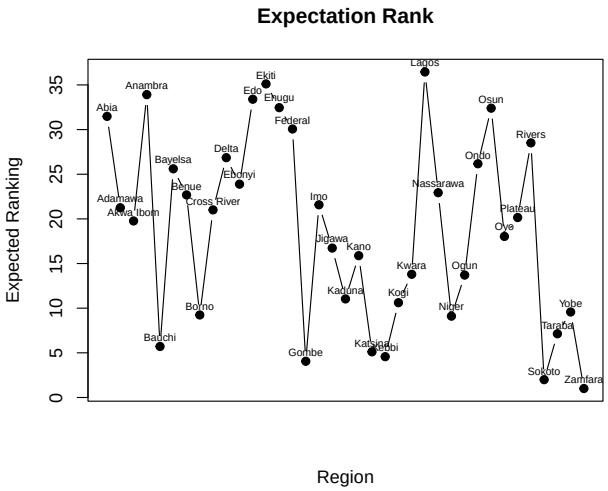

Figure 10: Expectation Rank of Each Region

the non-hierarchical model before in figure 9. However, actually, there is little difference between the first model and this one. But if we see the hyperparameters, there is a big difference between the distribution of μ_i and the distribution of p_i, which will be discussed later in the next section

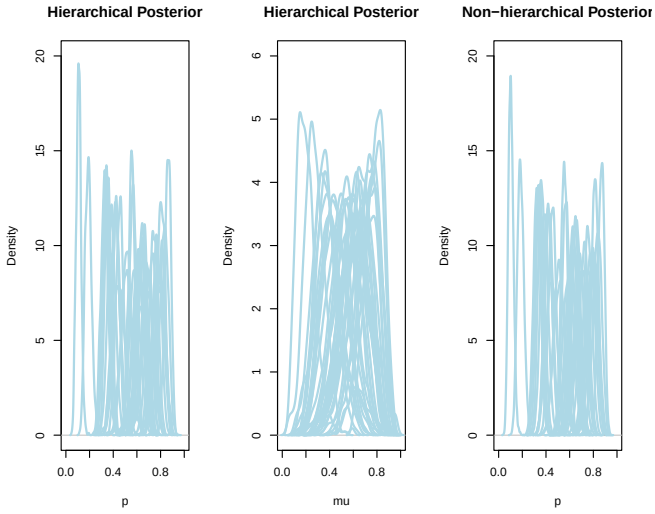

Figure 9: Comparisons

4.3 Posterior Ranking Inference

Here is the expectation of the rank for all the regions in our hierarchical model in figure 10. So we can get the 5 lowest coverage rate regions as follows:

```
##       [,1]
## [1,] "Zamfara"
## [2,] "Sokoto"
## [3,] "Gombe"
## [4,] "Kebbi"
## [5,] "Katsina"
```

And 5 highest vaccine coverage rate region:

```
##       [,1]
## [1,] "Lagos"
## [2,] "Ekiti"
## [3,] "Anambra"
## [4,] "Edo"
## [5,] "Enugu"
```

For the lowest region, the expected ranking distribution for each of them are as presented in figure 11. Smaller rank means the lower coverage rate.

For the highest region, the expected ranking distribution for each of them are as presented in figure 12. Similarly, higher rank means the higher coverage rate.

The ranking expectation plot is still almost telling similar information as the non-hierachical one. Detailed Discussion will be included later in the next section.

5. Discussion

As the results in the two models shown above, the non-hierarchical model mostly tells us about the difference among all the regions, which ignores the inner relationship. However, the hierarchical model will put an extra layer of prior distribution on the hyperparameters in the first model. The first model is conjugate and easy to sample since it is a beta-binomial model, that is really

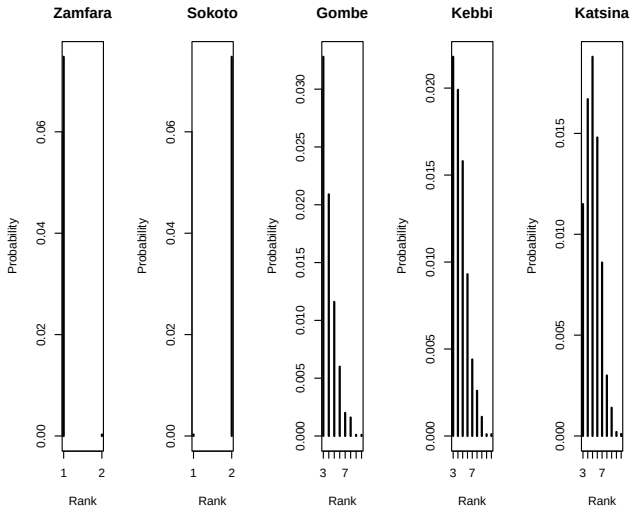

Figure 11: Rank Distribution of Lowest Rate Regions

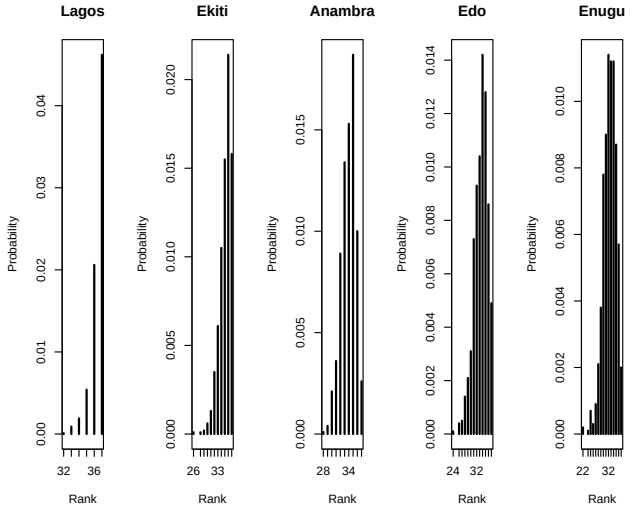

Figure 12: Rank Distribution of Lowest Rate Regions

a charming point to choose this model. However, considering all the observations come from a same country, we cannot ignore some inner relationship between the regions. For example, if we investigate other countries the regions in U.S will never have a same hyperparameter as the regions in other countries. Therefore, if we have more information about the country, or we have nice information about the hyperparameters in the hierarchical model, then we prefer the hierarchical one. However, the hierarchical model needs the Metropolis Hasting within Gibbs sampler to realize, which makes it sometimes hard to calculate.

From the posterior coverage rate distributions and posterior ranking distributions, we can see that the result of the posterior p_i is almost the same for them. But the first model has an assumption about the p_i, which is p_i's are i.i.d. However, with the hierarchical model of the distribution of μ_i, we can see that actually they are not i.i.d. under the hierarchical model. So the second model is more flexible and let us have the approach to find the distribution of the hyperparameters in the first model, which gave us more information about the country as a whole.

Also, I am confused why did we do the reparameterization, maybe my result is wrong, but according to the result, it actually didn't separate the variance and mean because if they share the same d, then the variance of p_i will still be $\mu_i(1-\mu_i)d$, which still depends on their mean. But the good news is that we made the variance of the μ_i not that dependent on the mean.

6. Potentially Better Modeling Approaches

To tell the truth, I spent lots of time doing the reparameterization part and derive the joint posterior distribution. So I am thinking of an easier way to model this case to avoid the logit function. Since:

$$\mu = \frac{\alpha}{\alpha+\beta} \in [0,1],\ d = \frac{1}{\alpha+\beta+1} \in [0,1]$$

If we put a uniform prior on both μ and d, then the calculation could be easier, and more reasonable. Because if we set $logit(\mu_i)$ to be a centered normal distribution, we have given the information that μ_i is concentrated around 0.5, which may not be a correct information. But if we set them following uniform distribution, there are several advantages. The first one is the calculation, the joint posterior will be easy to get. Also, even though we have to transform them again in to logit form when doing random walk MH, the Jacobian is also not hard to calculate. But the disadvantage is that although it's uniform distribution, I guess there could be a type of transformation of this pair of variables that is not uniform. But when calculating the joint posterior, it is much easier.