Analysis of Covid-19 cases in California Counties

Grant Hutchings

Abstract

California has been one of the states hit hardest by Covid-19. This analysis aims to understand differences in mortality rate and to predict death at the county and state level. We propose three models to do so; a binomial, a beta-binomial, and a hierarchical model. The first model, using a binomial likelihood and a beta prior, assumes a common mortality rate among the counties. The second model uses a Beta-Binomial likelihood as means of dealing with over-dispersion, and similarly assumes a common mortality rate. Finally, we consider a Hierarchical model with a binomial likelihood, beta prior, and non-informative hyperprior, which does not assume a common mortality rate. We will see that the first model predicts the most death, while the Beta-Binomial model predicts the least. We will also see that prediction becomes more difficult in counties with lower case counts.

Given the continued rise of Covid-19 cases, it is not unreasonable to assume that 20% of California residents may eventually contract Covid-19. Using this assumption, we will give posterior probabilities of more than 200,000 deaths state-wide for each of the three models.

Key Words: Beta-Binomial, Hierarchical, Leave-One-Out

1. Covid-19 in California

1.1 Data

We begin by familiarizing ourselves with the data at hand. Figure 1. shows that Los Angeles is an extreme outlier in terms of total cases and total deaths. While one might dismiss this as merely a function of high population, we also note that LA is in the 96th percentile for incidence rate (1 case in 1066 people), defined as total cases divided by population and the 68th percentile for mortality rate (3.40%), seen in Figure 2. We therefore expect that Los Angeles county will have a large impact on our inference as it sits well above average in all 4 of these categories and contributes over 25% of the total population, 38% of the total cases, and 44% of the total deaths. We suspect this will make accurate prediction for smaller counties challenging.

We also note a few other outliers that may be of importance later. Figure 2. shows the incidence rate and mortality by county. We see most of the mortality incidence mass lies below 4% and 7% respectively with only a handful of counties exceeding those thresholds. We note Shasta County with a 12% mortality rate. Even though Shasta has an unusually high mortality rate, we do not expect it to pull the overall expected mortality rate up too much given its low case count of 25. On the other hand, San Francisco has a very high incidence rate, but a low mortality rate, given its large population we expect this may pull the overall mortality rate down significantly. We also note Mono

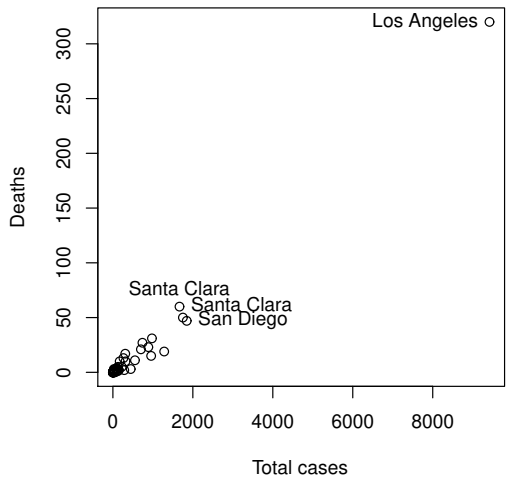

Figure 1: Total cases and total death by county. Most of the mass falls near the origin.

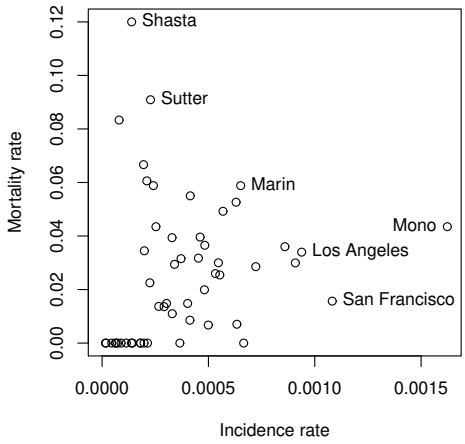

Figure 2: Mortality and Incidence rates by county. Most of the mass lies under 4% mortality and .05% incidence.

and Marin counties with relatively high rates in both cases. This makes them likely candidates for being influential in our analysis.

Another important aspect of the data is the large range between deaths in the counties. There are a handful of counties that have experienced huge problems with Covid-19, and have seen a tragically high number of deaths. On the other hand, there are many counties in California that have very low case counts. 16 of the 58 counties have had zero deaths as seen by the cluster in the lower left of figure 2. This contrast adds to the challenge of accurate prediction, especially in areas with low case counts.

2. Methods and Analysis

2.1 Model 1: Binomial

2.1.1 Methods

The first model we consider treats the deaths in each county as following a binomial likelihood, governed by a common mortality rate θ. We then assume a beta prior on θ for convenience of conjugacy. Hyperparameter choice of $\alpha = \beta = 1/2$ was chosen to put more prior mass at low mortality rate.

$$y_i \sim Bin(n_i, \theta),\ \theta \sim Be(1/2, 1, 2) \quad (1)$$

We take advantage of Beta-Binomial conjugacy and conveniently sample from the posterior distribution

$$\theta|Y \sim Be(1/2 + Y, 1/2 + N - Y) \quad (2)$$

where Y and N are the pooled death and case counts respectively. Clearly, a shortcoming of

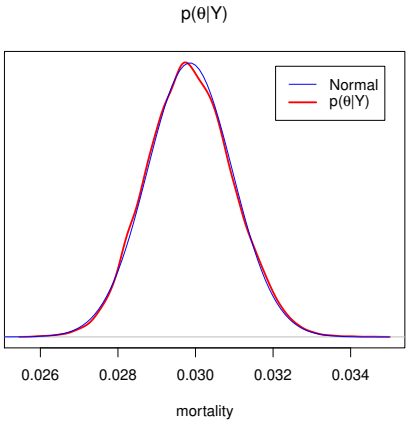

Figure 3: Posterior distribution of θ with normal approximation given. Parameters for the normal were taken to be the expected value and variance of the posterior samples.

this approach is that the data are pooled, and any information at the county level is lost. As mentioned earlier, outliers like Los Angeles will have a relatively large impact on this pooled estimate of θ.

2.1.2 Analysis

The posterior distribution, $p(\theta|y)$, is very close to normal with mean 2.99% and standard deviation 0.11%. This is slightly higher than the pooled mortality estimate from the data of 2.72%. Our posterior expectation is no doubt influenced by Los Angeles' aforementioned high mortality rate of 3.40%. This is shown in figure 3.

Figure 4. shows boxplots generated using the binomial expectation $\theta * y_i$, using samples from the posterior distribution of θ with y_i the number of deaths in each county. These boxplots have whiskers at the 2.5% and 97.5% quantiles. We see that using a pooled estimate of θ results in 95% posterior intervals which frequently miss the expected number of deaths from the binomial distribution shown as blue x's.

We conclude our analysis of model 1 by analyzing draws from the posterior predictive distribution for each county. We show that this more accurately represents the within county uncertainty. For each county, the posterior predictive distribution is Beta-Binomial.

$$\tilde{y_i} \sim BeBi(n_i, y_i + 1/2, n_i - y_i + 1/2) \quad (3)$$

Figure 5. shows there is much more uncertainty in the predictive distribution. This is because the Beta-Binomial distribution allows for differences in mortality rate between counties. We see that expected deaths are much more likely to fall within the 95% posterior credible interval.

Recall our previous question regarding a 20% total infection of California; If 20% of California contracts Covid-19, what is the posterior probability that more than 200,000 people will die. To answer this question, we use a new case count, defined as 20% of the county population. We take samples from the predictive distribution for each county and find the proportion of samples leading to this many deaths. We find that the binomial model predicts more than 200,000 deaths with a posterior probability of 1.

2.2 Model 2

2.2.1 Methods

We additionally propose modeling each counties deaths with a Beta-Binomial distribution and a non-informative prior.

$$y_i \sim BeBi(n_i, \mu, \tau) \quad (4)$$

$$p(\mu, \tau) \sim ((\mu(1-\mu)(1+\tau)^2)^{-1} \quad (5)$$

Where $\mu = \alpha/\alpha + \beta$ and $\tau = \alpha + \beta$ come from the traditional Beta distribution parameters

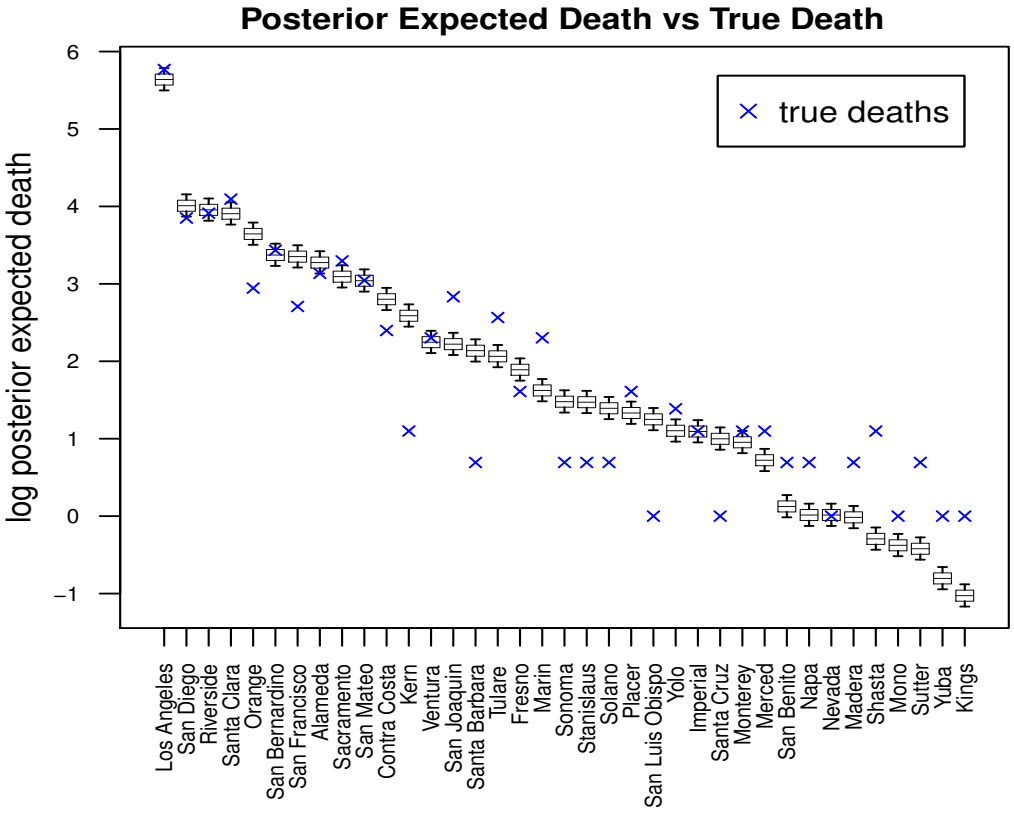

Figure 4: Data is shown on the log scale to better visualize counties with low case counts and so uncertainty is visible. Given the log scale, counties with zero deaths ($y_i = 0$) are left out.

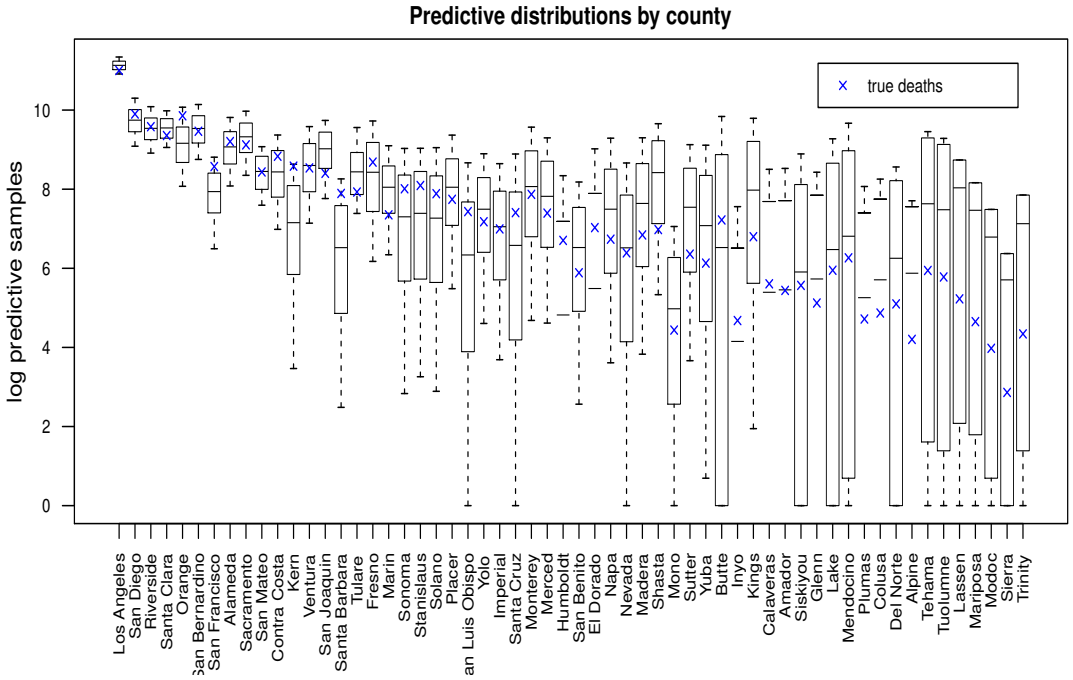

Figure 5: Data is shown on the log scale to better visualize counties with low case counts and so uncertainty is visible.

α and β. This transformation results in a Beta distribution with expectation μ, a single parameter.

This model is frequently used as an alternative to the binomial model in the previous section to account for over-dispersion in the data. We believe this model may be a good alternative because the binomial model does under-predict the sample variance in the data (deaths). The standard deviation in the data is 43.1 deaths, while the binomial predicts a standard deviation of 26.5 deaths. However, it is worth noting that the binomial model only under-predicts the variance in the data because of Los Angeles. The sample variance is extremely inflated by the number of deaths in LA compared to other counties. If we remove LA, the sample and binomial predicted standard deviations are 13.1 and 19.8 deaths respectively, the binomial actually over-predicts variation in this case.

To sample from the posterior distribution $p(\mu, \tau|y)$ we first transform the parameters μ,τ to the real line using the transformations $\theta_1 = logit(\mu)$ and $\theta_2 = log(\tau)$. We use rejection sampling to obtain out posterior samples of μ and τ. This is made significantly easier thanks to our transformation. With the new unbounded parameters θ_1, θ_2, we can easily use a multivariate student-t as our proposal distribution. We use the posterior mode and 2 times the curvature at the posterior mode for the mean and variance of the proposal t distribution. We double the curvature in an attempt to better explore the tails of our posterior. We choose 4 degrees of freedom so out tails are sufficiently fat for the rejection sampling algorithm. Fat tails in the proposal distribution help to accept more samples.

2.2.2 Analysis

We turn our attention to looking for especially influential counties in the inference on μ, the mortality rate. We do this by preforming a leave-one-out analysis. We will generate samples of μ 58 different times, each time leaving out the data from one of the counties. After doing this, we can look at the distribution in expected value of μ, seen in figure 6. The green dashed lines indicate the 2.5% and 97.5% quantiles.

In figure 6, the two counties in the left tail are San Joaquin and Marin. What we immediately notice is both of these counties have an unusually high mortality rate at 5.9% and 5.5% respectively. This shows that removing these counties leads to a large reduction in posterior expected mortality rate. The two counties in the right tail are Orange and Kern, both of which have abnormally low mortality rates. Leaving these out tends to increase posterior expected mortality.

While both this model and the previous model suffer from a pooled estimate for the posterior mortality. The Beta-Binomial model does help the over-dispersion problem brought about by Los Angeles. The expected posterior standard deviations for both models are compared to the sample standard deviation in table 1.

The most important difference between this model and the binomial model of the previous section is that the beta-binomial distribution allows for the probability of death to change from county to county, while the binomial model assumes this probability to be fixed. We expect better inference from the beta-binomial model given the high variability in our data. We see that this difference helps account for over-dispersion in the data. See table 1 for predicted standard deviations by the two models

Table 1: Standard deviation predicted by models 1 and 2 compared to sample standard deviation. See that the Beta-Binomial model does better than the binomial at accounting for variation in the data.

Model	Std Dev
Sample	43.09
Beta-Bin	32.43
Binomial	26.53

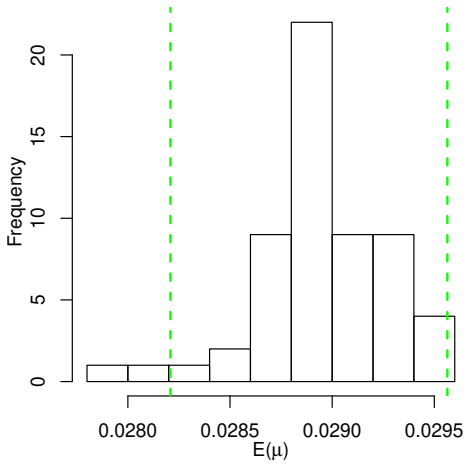

Figure 6: Leave-One-Out analysis of the posterior expectation of μ

2.3 Model 3

2.3.1 Methods

The final model considered is a hierarchical extension of the previous models.

$$y_i \sim Bin(n_i, \theta_i),\ \theta_i \sim Be(\mu\tau, (1-\mu)\tau) \quad (6)$$

$$p(\mu, \tau) \sim ((\mu(1-\mu)(1+\tau)^2)^{-1} \quad (7)$$

The main benefit of this model is the differentiation of mortality rates, which should improve prediction at the county level.

The joint posterior $p(\boldsymbol{\theta}, \mu, \tau|\boldsymbol{y})$ can be decomposed for ease of sampling.

$$p(\boldsymbol{\theta}, \mu, \tau|\boldsymbol{y} \propto p(\boldsymbol{\theta}|\mu, \tau, \boldsymbol{y})p(\mu, \tau|\boldsymbol{y}) \quad (8)$$

$$p(\boldsymbol{\theta}|\mu, \tau, \boldsymbol{y}) \propto \prod_{We=1}^{m} \frac{\theta_i^{\mu\tau+y_i-1}(1-\theta_i)^{(1-\mu)\tau+n_i-y_i-1}}{\beta(\mu\tau+y_i, (1-\mu)\tau+n_i-y_i)} \quad (9)$$

$$p(\mu, \tau|\boldsymbol{y}) \propto (\mu(1-\mu)(1+\tau)^2)^{-1} \times \prod_{We=1}^{m} \frac{\beta(\mu\tau+y_i, (1-\mu)\tau+n_i-y_i)}{\beta(\mu\tau, (1-\mu)\tau} \quad (10)$$

We proceed as in the previous model with a transformation $\theta_1 = logit(\mu), \theta_2 = log(\tau)$. We then use this decomposition to sample from the joint posterior my first performing sampling importance resampling to get samples of μ, τ then using the fact that the full conditional distribution of θ is a beta distribution.

2.3.2 Analysis

Now that we have attained samples from the joint posterior we examine a similar plot to figure 4. In figure 8 we see that our model has much more accurately captured the uncertainty in the data. Counties with less cases and deaths have very high posterior uncertainty. Thus, our 95% posterior intervals much more often capture the true death rate seen in the data. We also notice that we do very well at finding the true number of deaths for the first few counties, as they have the most cases to make inference with. Figure 4 shows that in model 2 we are unable to

capture the uncertainty and as a result, we had no confidence in our ability to recover the true values. On the other hand Figure 8 shows that model 3 does a much better job at capturing this uncertainty.

It is clear from this model that there are differences between the counties, but the analysis has brought them all closer to a grand mean. Figure 9 shows the distribution of posterior mortality rate by county. This shows that a-posteriori counties with abnormally high mortality rates were brought down by the analysis and counties with abnormally low rates were brought up. We also notice that variation increases as we move to the right, which represents lower case count.

3. 20% total infection study

Recall our interest in the question: If 20% of California residents contract Covid-19, what is the probability that more than 200,000 people will die? As mentioned in section 2.1, model 1 predicts more than 200,000 people will die with probability 1. This probability was determined using samples from the Beta-Binomial posterior predictive distribution. Model 2 predicts this probability at 79.1% which was calculated using posterior samples of μ, τ then predicting new values for the number of deaths using the Beta-Binomial distribution, the likelihood for model 2. Prediction was made for model 3 using posterior samples of theta and sampling new values of the number of deaths from the a binomial distribution. Model 3 estimates this probability as 99.9%. I cannot say why model 2, with a Beta-Binomial likelihood predicts such a different probability than the other two models, but as model 3 is the most robust, it better captures the uncertainty, and we should trust its prediction more than the other models. This information is summarized in table 2.

Table 2: Posterior probabilities of more than 200,000 deaths under the three models given that 20% of California contracts Covid-19

Model	Posterior Prob
1	1
2	0.7909
3	0.9995

4. Conclusions

We have fit three models to Covid-19 data from California. The first two models assume one mortality rate for the whole state, while the third allows this rate to vary by county. Covid-19 data is incredibly hard to model due to the extreme heteroskedasticity. We cannot expect to make accurate predictions for counties with extremely low counts when the analysis is easily overshadowed by larger ones. We hope to have convinced the reader that predictions using a hierarchical model should be more accurate and robust to outliers than simpler models, as it better captures the uncertainty in the data.

REFERENCES

Gelman, A., Carlin, J. B., Stern, H. S., Dunson D. B., Vehtari, A., and Rubin, D. B. (2014), "Bayesian Data Analysis, Third Edition".

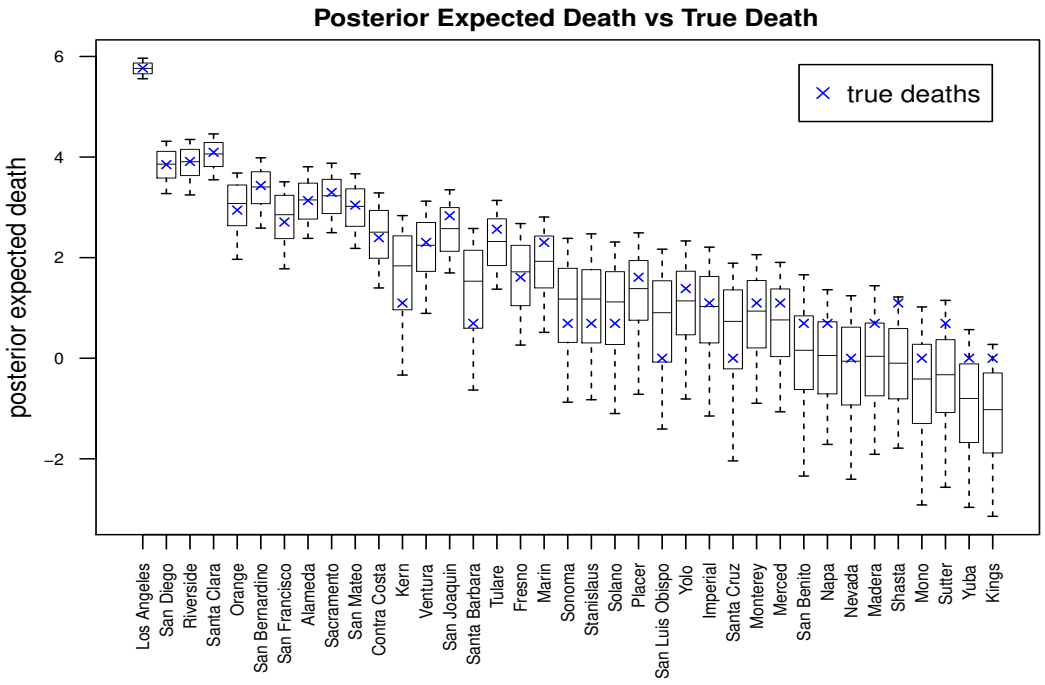

Figure 7: Expected deaths county from the hierarchical model. Uncertainty generally increases as case counts go down, and we are less likely to recover the true data. Data is shown on a log scale so uncertainty is more easily seen.

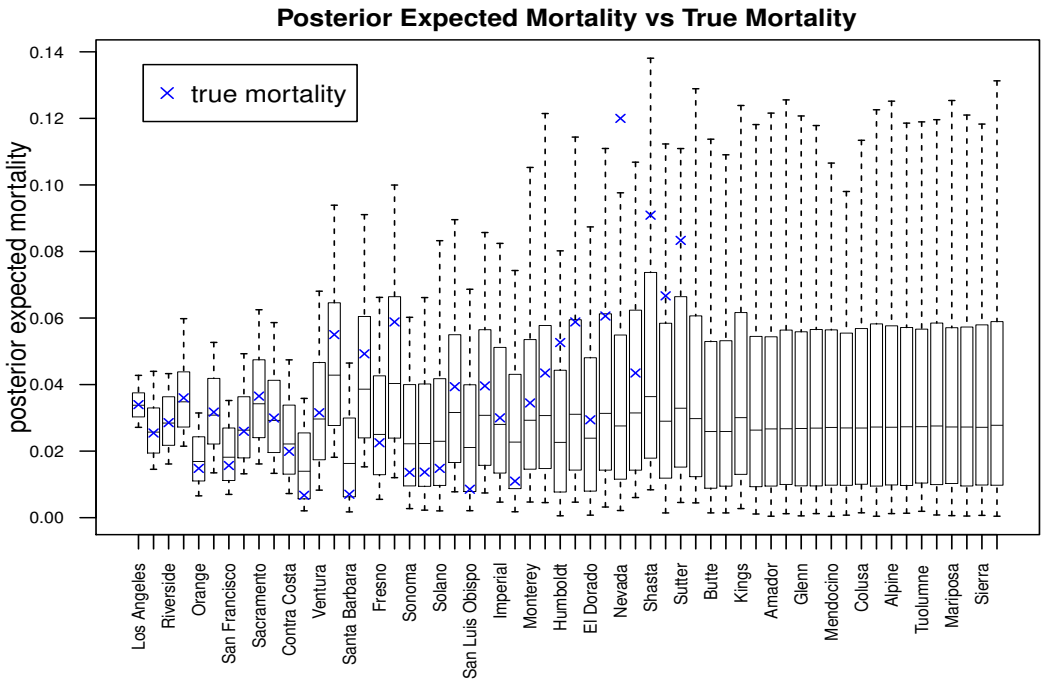

Figure 8: Expected mortality rate by county from the hierarchical model. Uncertainty is increasing as case counts go down, all a-priori mortality rates have converged towards a grand mean and counties with zero deaths have all been placed around the same mortality rate.