Members: Alonzo Cortez, Geng Xia, Simon Pfeiffer, Madison Dowd

Introduction

Income inequality has long been a subject of interest and concern for policymakers, researchers, and the general public. Understanding the relationship between income and region is crucial to uncovering disparities in income and identifying factors that contribute to such disparities. This research paper aims to examine and predict the relationship between income and their associated United States region. More specifically, the final goal we want to achieve is to find trends for the states in which people have the highest income. To achieve this goal, we will analyze data from the United States Census Bureau and other relevant sources. By investigating the patterns and trends in income distribution across different regions, this study aims to provide insights into the factors that drive income disparities and inform policies that promote greater economic equality.

Data

For data collection and analysis, we will be using the dataset "usdata", a package in R, that includes the following variables:

- County Names
- State Names
- Population in 2000
- Population in 2010
- Population in 2017
- Population Change 2010 - 2017
- Percent of Population in Poverty in 2017
- Home ownership Rate 2006 - 2010
- Percent of Housing in Multi-Unit Structures 2006 - 2010
- Unemployment Rate in 2017
- Metro (Whether the county contains a Metropolitan Area)
- Median Education Level 2013-2017
- Per Capita Income per person 2013 - 2017
- Median Household Income
- Smoking Ban
    - Describes whether the type of county-level smoking ban in place in 2010, taking one of the values "none", "partial", or "comprehensive"

We will also use data from the ACS 2019 Survey file, with the following data points:

- State
- Median Household Income
- 25th Percentile Income

- 50th Percentile Income
- 75th Percentile Income

These data were collected from Census Quick Facts (no longer available as of 2020) and its accompanying pages. The smoking ban data were from a variety of sources. The data set “counties,” which we use for our analysis, contains data on 3142 counties throughout the United States. The set contains data collected from various years, which is important to keep track of throughout our analysis. For example, the poverty variable comes from the percent of population in poverty in 2017, while the per_capita_income measures the per capita income from 2013-2017.

While the data is not a simple random sample from a population, the US census bureau uses a complex sampling design to determine which households are interviewed. Specifically, it uses a two-stage sample design to select its sample: (1) the selection of primary sampling units (PSUs), and (2) the selection of address units within sample PSUs. (Source)

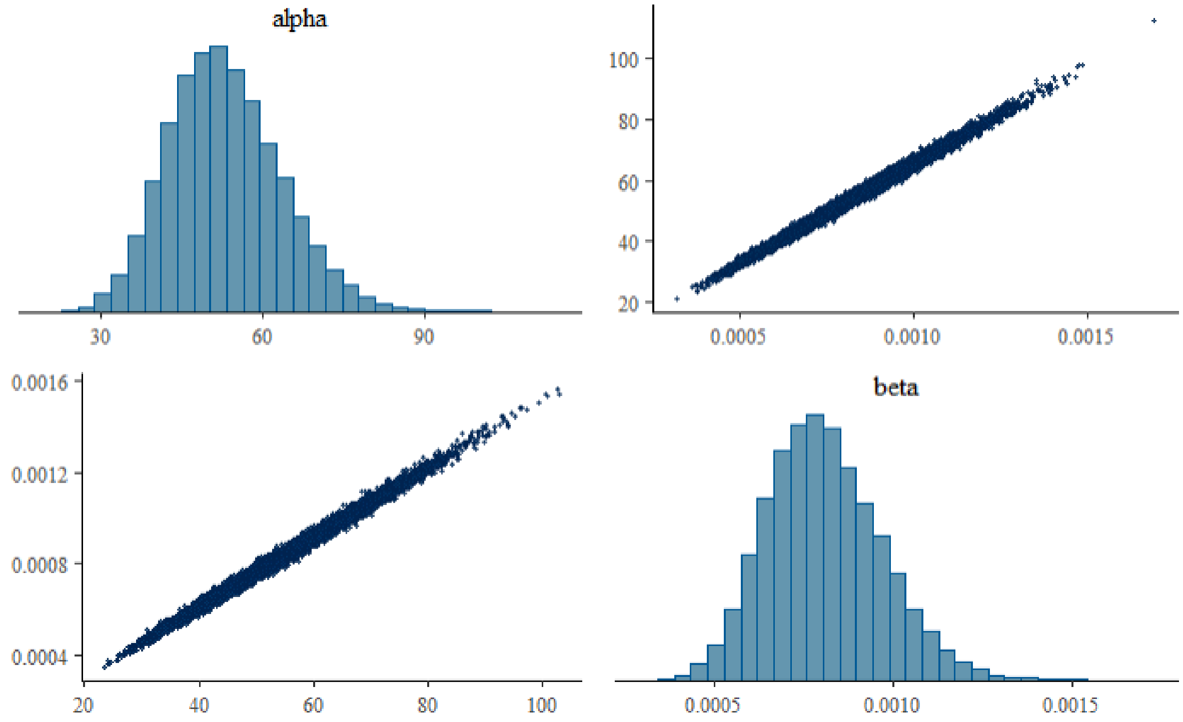
One potential defect in our analysis is that we’re working with survey data. With survey data, the responses will always have some sort of uncertainty. For example, some responders may not be able to recall certain information they are being asked for, and may be too lazy to accurately report on it. There is always bias and variation in survey responses – and therefore survey data – that we have to be wary of in our analysis.

In addition, people are not truly required to fill out the survey data. No one has been prosecuted for refusal to fill out the survey since 1970, so we may not be able to make causal claims associated with our findings.

To account for the data collecting process in your analysis, we’ll look at some of the similarities that may occur within the population that did choose to participate. Is the population skewed towards high income earners? Or skewed towards a specific region(s) of the country? We will use weights to try to even out some of the disparity.

If we were able to collect our own data, we would use a random sample to accurately represent the United States population, and make sure that we got 100% participation in these surveys. But, this method is not realistic when collecting survey data. 100% participation is generally never achievable, and this is something that we will have to look at in our analysis.

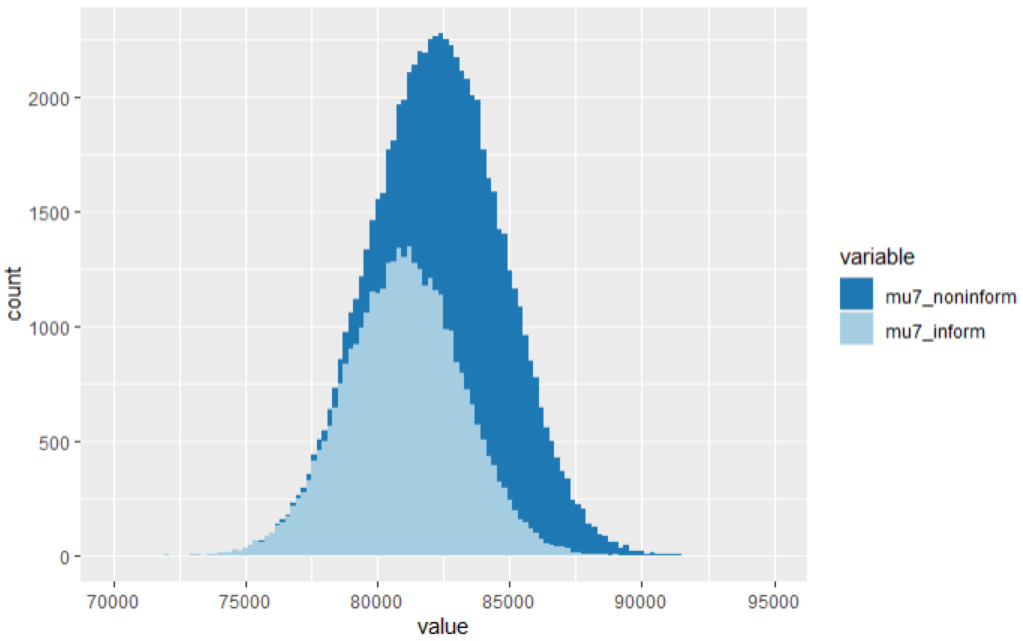
Methodology



Across the US, the alpha parameter for the gamma distribution of income ranges from 30 to 90, while the beta parameter ranges from 0.0005 to 0.0015. A higher alpha value (e.g., 90) would indicate a more skewed distribution with a longer tail on one end, suggesting that income values may be concentrated towards one end of the distribution (e.g., higher incomes) with a less likely occurrence of lower incomes. A lower alpha value (e.g., 30) would indicate a less skewed or more symmetric distribution with income values more evenly spread across the distribution. Similarly, a higher beta value (e.g., 0.0015) would indicate a steeper decline in the tail of the distribution, meaning that higher incomes are less likely to occur, while a lower beta value (e.g., 0.0005) would indicate a slower decline, suggesting that higher incomes are more likely to occur. It's important to interpret these parameter values in the specific context of the data and the modeling assumptions being used, as they can affect the shape, scale, and characteristics of the income distribution modeled with a gamma distribution.

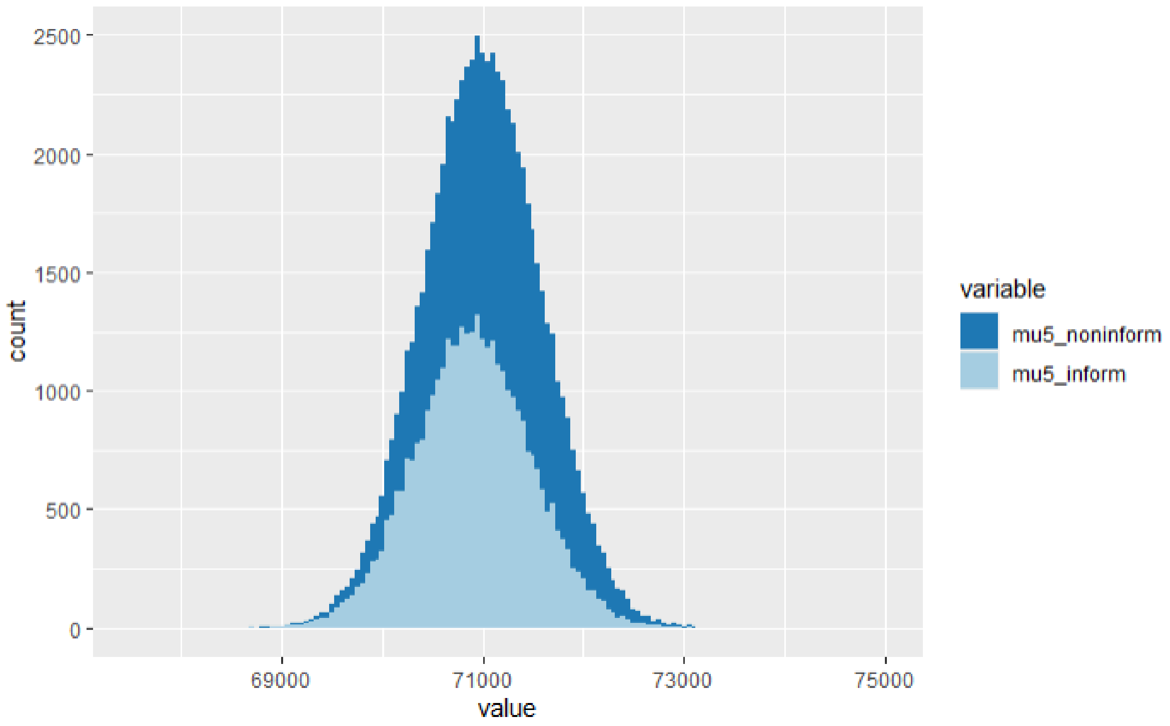
The linear relationship between alpha and beta may imply that certain states consistently exhibit similar income distribution characteristics. For instance, states with higher alpha values (indicating more skewed distributions) and higher beta values (indicating higher income frequencies) may consistently have income distributions that are concentrated towards higher incomes with less likelihood of lower incomes, and vice versa. This could provide insights into regional or state-specific income patterns and dynamics.

Connecticut (biggest sigma):

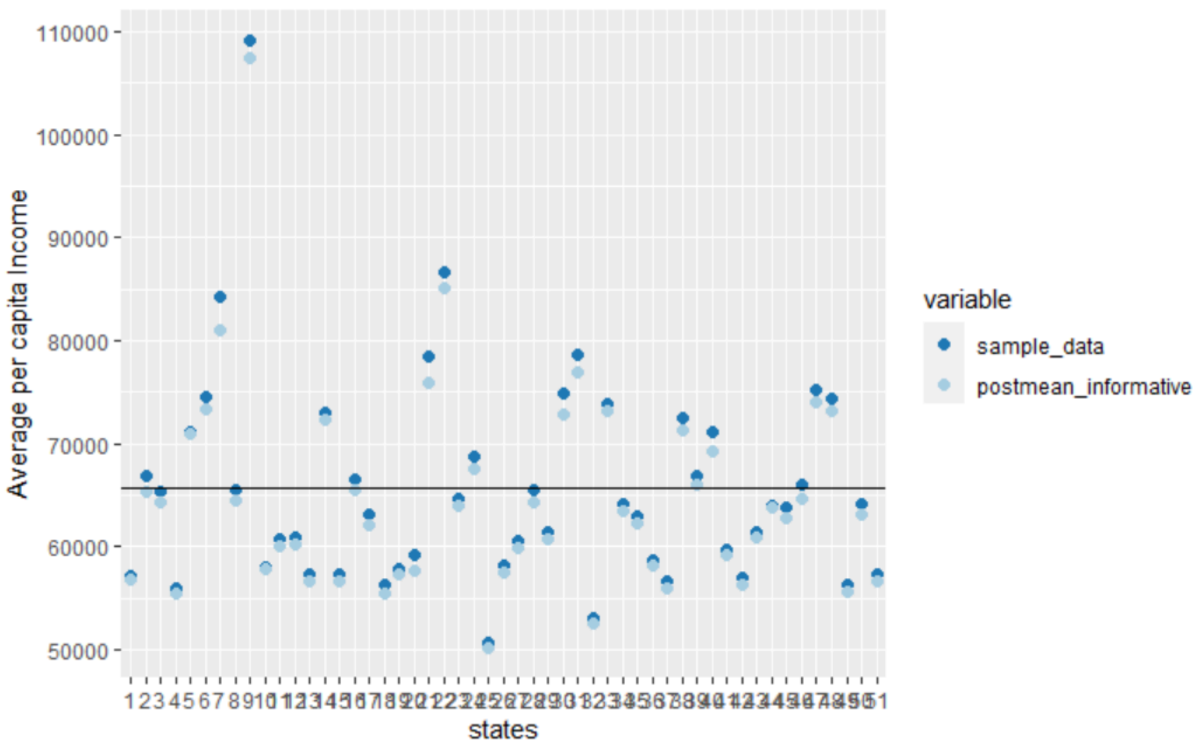


The use of non informative and informative priors in the two distributions for income suggests that different assumptions were incorporated into our analysis. The non-informative prior yielded a mean income of around $84,000 USD could suggest that the informative prior (mean income of about $81,000k USD) led to a narrower distribution with lower variability. In regards to the sample size, we see that the non-informative prior had about 2100 samples while the informative prior had about 1300, suggesting that the estimates of the non informative may be more reliable due to its larger sample size. However, it's also important to note that sample size alone is not the only determinant in reliability of our statistical estimates, and other factors such as methodology and representativeness come into play. Further, the observation that Connecticut had the largest standard deviation compared to the other states may indicate greater income variability within Connecticut. Further analysis, such as examining the factors contributing to the income variability, could provide meaningful insights into the economic and social dynamics within the state. This high standard deviation may be due to the small relative sample size for Connecticut at 805, which ranks as the 3rd lowest values for n in our dataset. Connecticut also ranks top 11 in 25th percentile income across the US while also ranking in the top 4 in 75th percentile income. This indicates that generally, most residents in Connecticut are top earners when compared to the US as a whole.

California (smallest sigma):

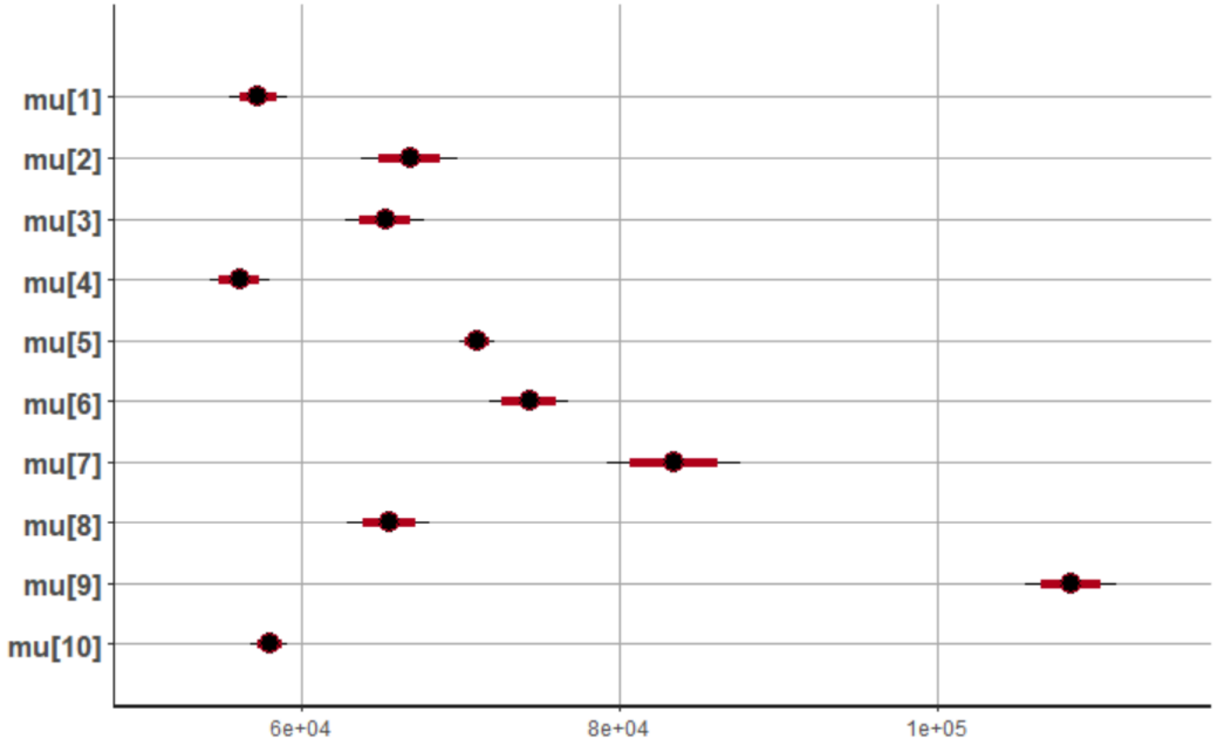


The above graph depicts the comparison between the informative and noninformative prior posterior simulation for the state of California. California had the smallest standard deviation from our entire data set, which suggests that most of their population is highly centered around the same number for per capita income. As depicted, although the count is obviously lower for our informative prior, they center around almost exactly the same number, around $71,000. The graph depicts the accuracy of our noninformative prior compared to our informative prior, and as we can see, the noninformative prior is just as accurate in predicting per capita income in California as the informative prior is.



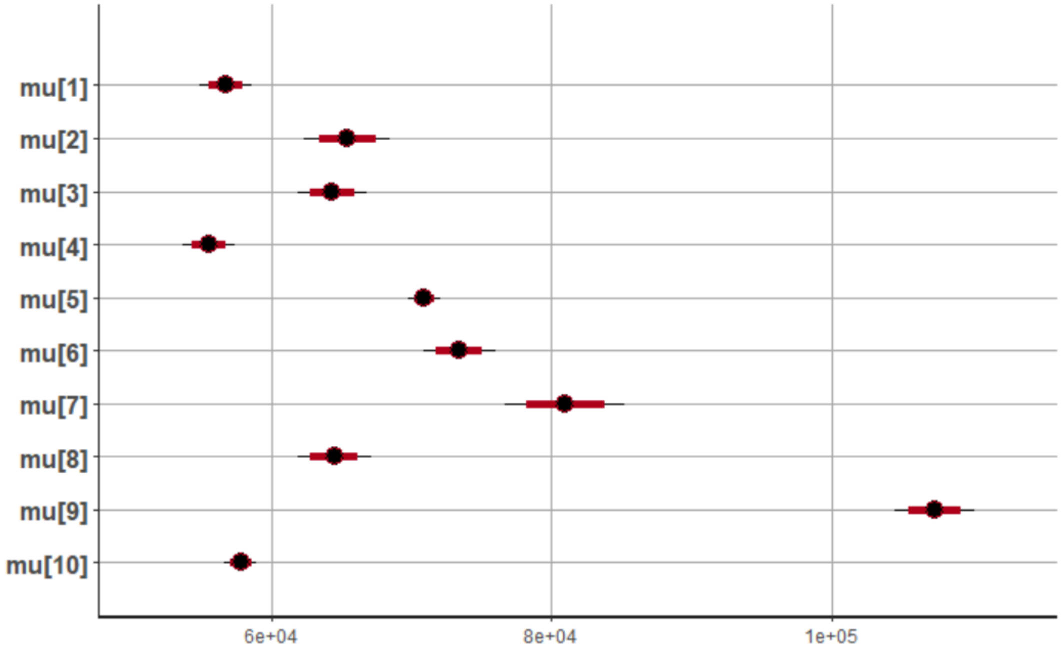
From our analysis which included all 50 states, as well as the District of Columbia, the above is a scatterplot comparing the data we collected from our sample compared to the estimate generated from our posterior distribution with an informative prior. From the plot, we can see our posterior distribution does a decently good job at predicting the average per capita income for each of the 51 states or district(s). The dots are very close together, and many even overlap. This is reassuring evidence that our posterior distribution predictive model converges to real life data and our simulation is accurate.

<u>Noninformative:</u>



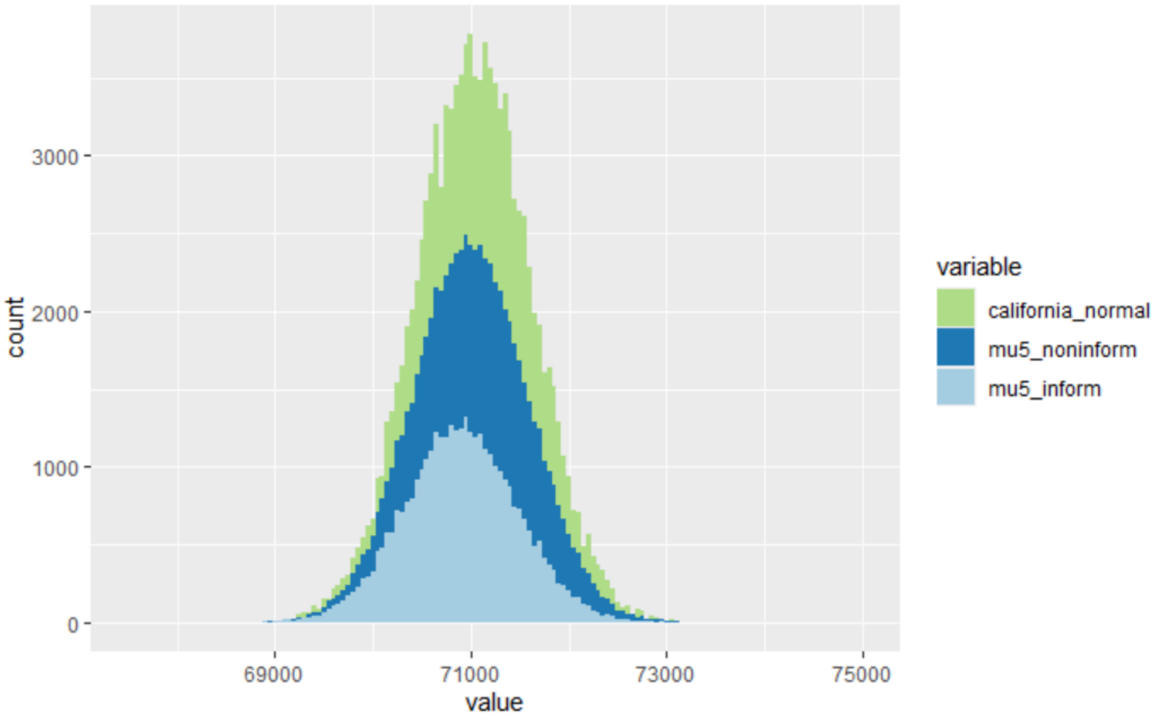
The above plot details our output from the posterior simulation for the estimated mean for the first 10 states in our analysis. The values range from just under $60,000 to just over $100,000. We used a noninformative, or flat, prior for this simulation. However, given the fact that our prior was not informative, it was heavily influenced by the input of new data. Initially, for our values of mu[x], it all ranged between $22,000 to $36,000, with the mean being centered around $31,000 for all 50 states. However, when plotted via histogram, the reality was that most states had an average income of around $60,000 USD. Although the noninformative prior generally allows for more flexibility, this was its main drawback in our findings.

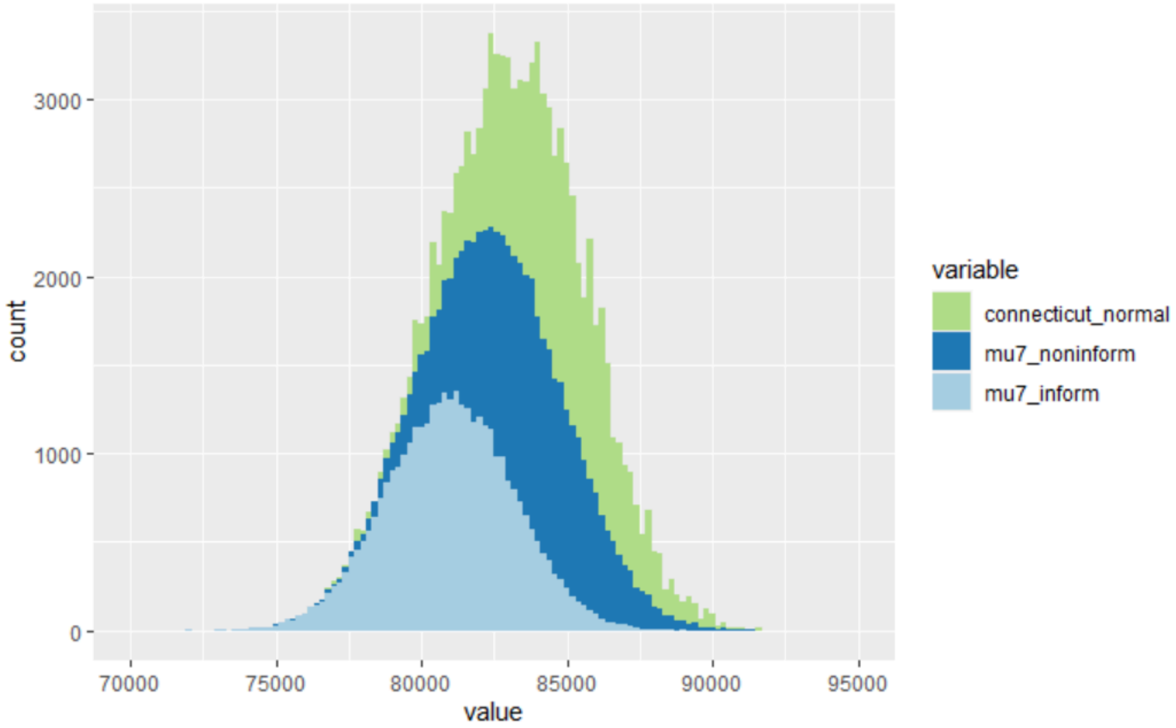
Informative



The above plot details our posterior simulation results from an informative prior. Compared to our results for the same 10 states, the results from the informative simulation are almost identical to the noninformative prior results we got (see above).

Our results from the informative and noninformative simulations could be so similar because there is a large amount of data available on per capita income for each state in the US. This means that the data itself provides a lot of information about the distribution of per capita income across states, and the prior has less influence on the posterior. In such cases, both informative and noninformative priors tend to converge to similar posterior distributions, as we see in this case.


count
3000
2000
1000
0
69000
71000
73000
75000
value
variable
california_normal
mu5_noninform
mu5_inform


count
3000
2000
1000
0
70000
75000
80000
85000
90000
95000
value
variable
connecticut_normal
mu7_noninform
mu7_inform