MATH 392 Problem Set 3

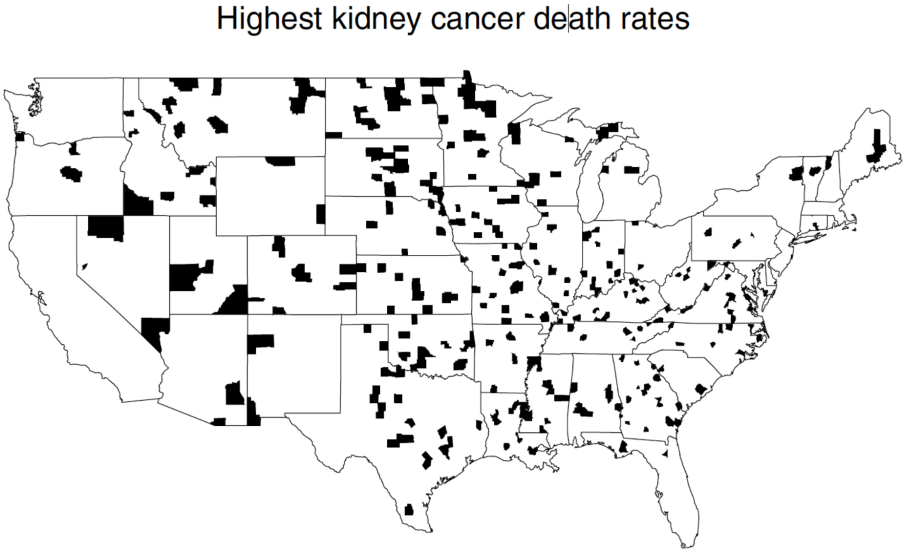
Exercises from the book

7.5: 5, 8, 11, 12

7.6: 3, 12, 14, 23

Case Study: Bayes vs. Frequentist Estimators

The map below identifies the counties in the US with the highest kidney cancer rates in the US from 1980 - 1989.

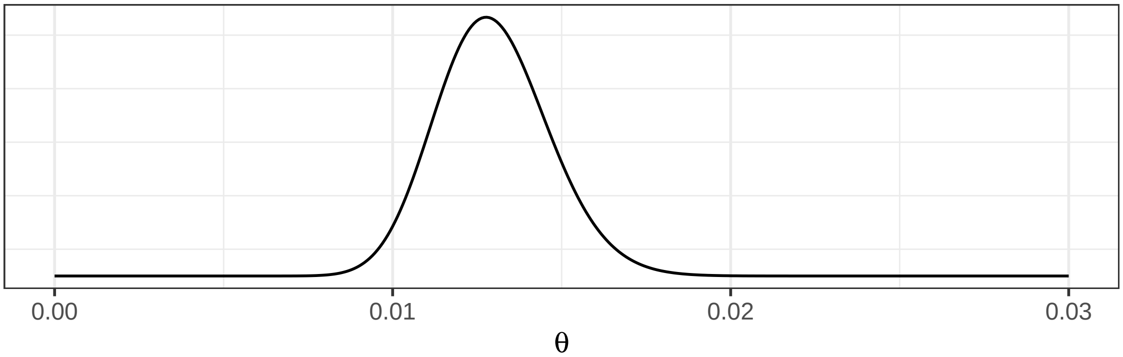


As we discussed in class, it is difficult to identify a meaningful geographic pattern because many of these rates may have been caused by the high variability inherent in counties with very small populations. We will use simulation to evaluate how the picture would change if we were to use a Bayes Estimator.

Formulating a prior

A Bayes Estimator requires that we specify a loss function and a prior/posterior. For the loss function, we'll use the standard squared loss. The prior is open to more debate, but a sensible place to start would be to coalate all of the information that we have about the variability in cancer rates across counties in the US. Recent data and expertise suggest that cancer rates average around and have a distribution well-described by the Gamma distribution.

Let θ_i be the cancer rate in county i (cases per 100,000). $\theta_i \sim \text{Gamma}(\alpha = 17.87, \beta = .7144)$.

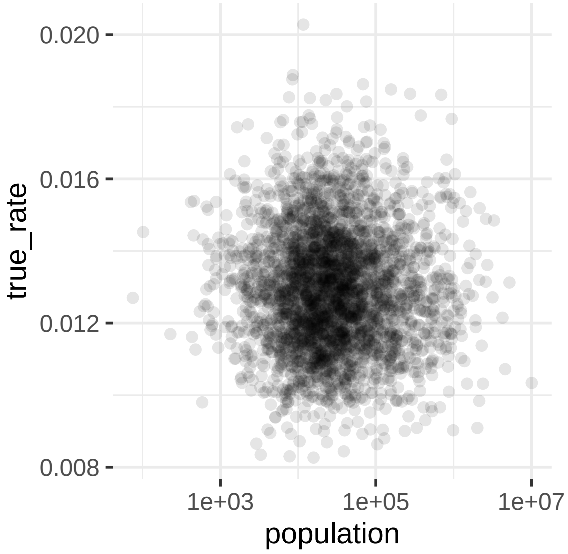


Let's start out simulation by assuming that each county has a cancer rate drawn at random from this prior distribution. We can append those rates to a dataframe of county population from the `tidycensus` package (consult the Rmd for this assignment to harvest this code) and print out the first 10 counties.

	NAME	population	true_rate
1	Autauga County, Alabama	55200.00	0.02
2	Blount County, Alabama	57645.00	0.01
3	Chambers County, Alabama	33826.00	0.01
4	Chilton County, Alabama	43930.00	0.01
5	Colbert County, Alabama	54495.00	0.01
6	Dale County, Alabama	49255.00	0.01
7	Elmore County, Alabama	81212.00	0.01
8	Hale County, Alabama	14887.00	0.01
9	Lawrence County, Alabama	33171.00	0.01
10	Limestone County, Alabama	93052.00	0.01

Even though these are simulated, let's think of them as the true cancer rates of these counties.

Exercise 1: Construct a plot that shows the relationship between the size of the population in a county and the corresponding cancer rate. You may need to use transformations of the scales so that the visualization is informative. How would you describe the relationship between these two variables?



There is no relationship between the true rate and the population. This was by design: each rate was drawn i.i.d. from the same gamma distribution.

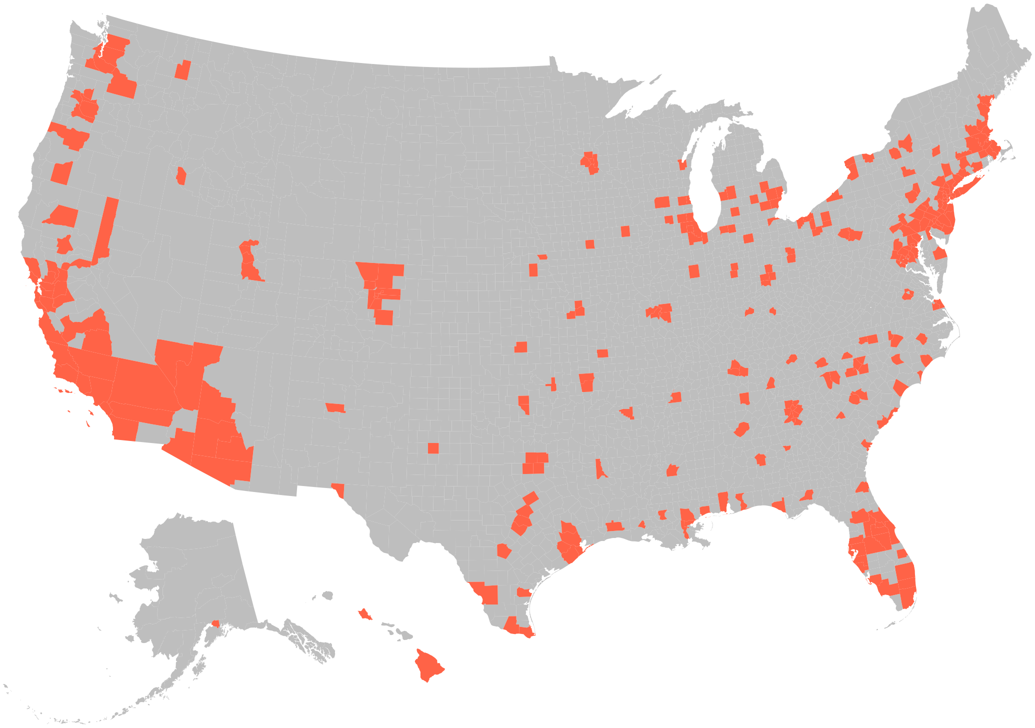
A model for the data

The number of cases, X_i, that actually materialize in county i could be sensibly modeled using the Poisson distribution, $X_i \sim \text{Poisson}(n_i \times \theta_i/100{,}000)$, where n is the population of county i.

Exercise 2: For each county in `acs`, use the Poisson distribution to simulate the number of cases according to that county's underlying rate. Add these counts as a new column in the dataframe called `n_cases`.

```
acs <- acs %>%
  mutate(n_cases = rpois(n(), true_rate * population))
```

Exercise 3: Construct a county map of the US that shades in red the counties that rank in the top 10% in terms of number of cases (there is code in the Rmd that you are encouraged to utilize). Describe the pattern that emerges and propose an explanation for this structure.



Estimating θ_i

It is clear that better than simply visualizing the raw number of cases would be to estimate each county's underlying rate.

Exercise 4: For each county, come up with the maximum likelihood estimate of θ_i. Note that for each county, we only observe a single observation. First lay out the general form of the MLE in this setting, then compute it for each county and add these estimates as a new column in `acs`.

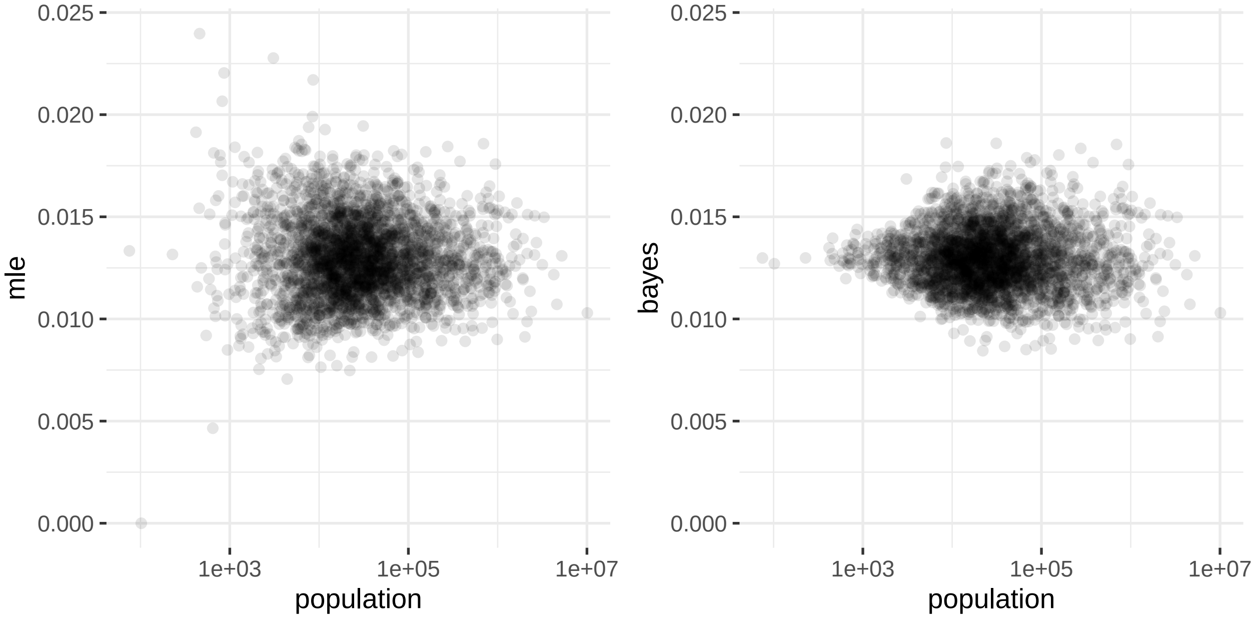
```
acs <- acs %>%
  mutate(mle = (n_cases / population))
```

Exercise 5: As an alternative, lay out the general form of the Bayes Estimator using the squared loss and the Gamma prior outlined above. Then compute this estimate for each county and add it as a column to `acs`. Using `xtable()` as we did above, print out this final table with both columns of estimates.

```
acs <- acs %>%
  mutate(bayes = ((n_cases + alpha)/(population + beta)))
```

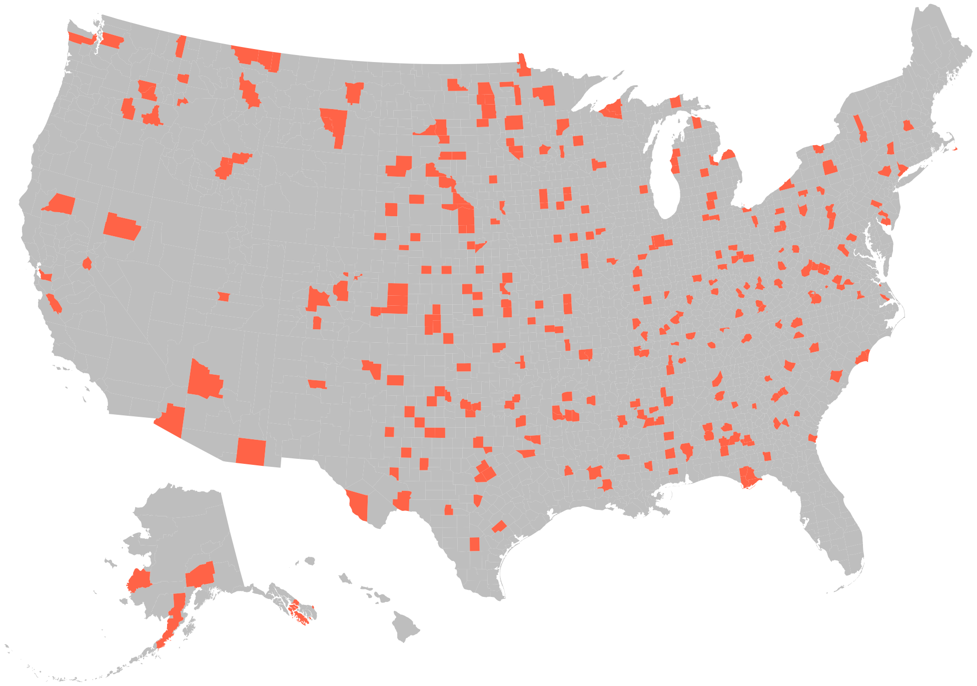
	NAME	population	true_rate	n_cases	mle	bayes
1	Autauga County, Alabama	55200.00	0.02	909	0.02	0.02
2	Blount County, Alabama	57645.00	0.01	750	0.01	0.01
3	Chambers County, Alabama	33826.00	0.01	357	0.01	0.01
4	Chilton County, Alabama	43930.00	0.01	454	0.01	0.01
5	Colbert County, Alabama	54495.00	0.01	786	0.01	0.01
6	Dale County, Alabama	49255.00	0.01	720	0.01	0.01
7	Elmore County, Alabama	81212.00	0.01	1017	0.01	0.01
8	Hale County, Alabama	14887.00	0.01	226	0.02	0.01
9	Lawrence County, Alabama	33171.00	0.01	465	0.01	0.01
10	Limestone County, Alabama	93052.00	0.01	1156	0.01	0.01

Exercise 6: What is the relationship between each of these estimates and the population size of each county? Construct two scatterplots side by side (see code for example), with population size on the x-axis of both and each of the estimates on the y-axes. Again, be sure to transform the scales to better reveal the structure. Describe the trend that you see in each plot.



Exercise 7: Remake the US map two ways: one plotting the MLE and the other with the Bayes Estimator. What do you think is the cause of the dominant spatial pattern in the former? What about for the latter?

MLE Estimate



Bayes Estimate