MATH 392 Problem Set 3: Case Study

Sam D. Olson

Time Taken on Case Study

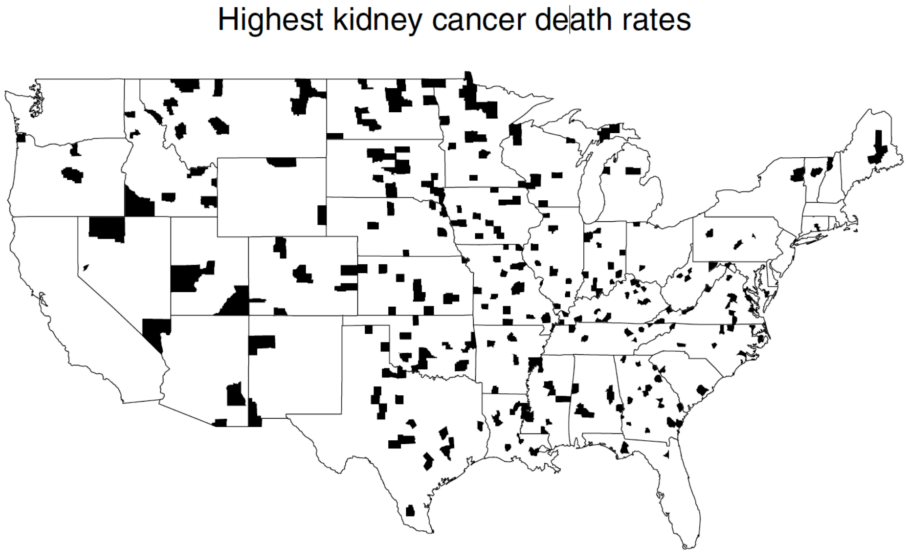
5 1/2 hours

4 1/2 hours coding/proofreading

1 hour on commentary/analysis

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.



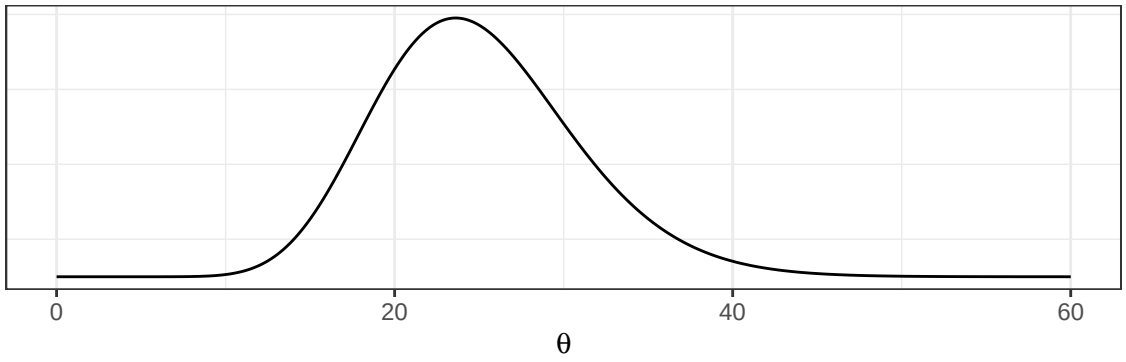
As we discussed in class, it is difficult to identify a meaningful geographic pattern because many of these rates may have been caused by the high variability inherent in counties with very small populations. We will use simulation to evaluate how the picture would change if we were to use a Bayes Estimator.

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). $\theta_i \sim \text{Gamma}(\alpha = 17.87, \beta = .7144)$.



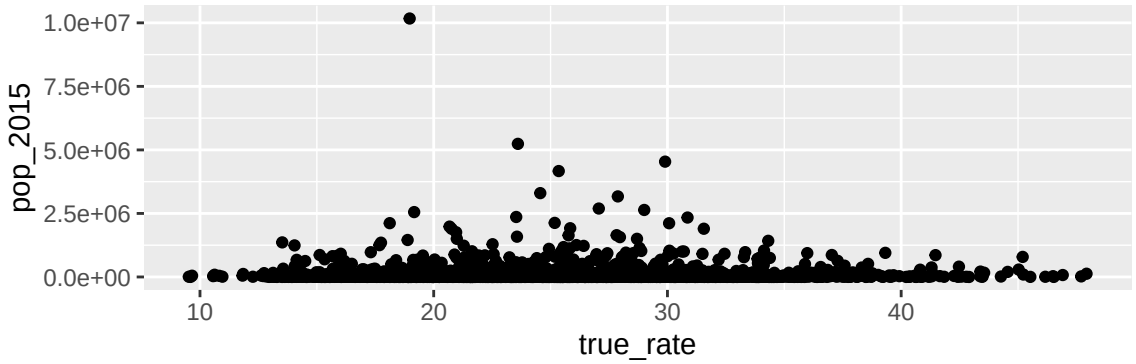
Let's start out simulation by assuming that each county has a cancer rate drawn at random from this prior distribution. We can append those rates to a dataframe of county population from the `usmap` package (consult the Rmd for this assignment to harvest this code) and print out the first 10 counties.

	fips	abbr	county	pop_2015	true_rate
1	01001	AL	Autauga County	55347.00	33.47
2	01003	AL	Baldwin County	203709.00	28.18
3	01005	AL	Barbour County	26489.00	24.55
4	01007	AL	Bibb County	22583.00	20.82
5	01009	AL	Blount County	57673.00	17.98
6	01011	AL	Bullock County	10696.00	36.56
7	01013	AL	Butler County	20154.00	27.09
8	01015	AL	Calhoun County	115620.00	25.94
9	01017	AL	Chambers County	34123.00	17.06
10	01019	AL	Cherokee County	25859.00	26.01

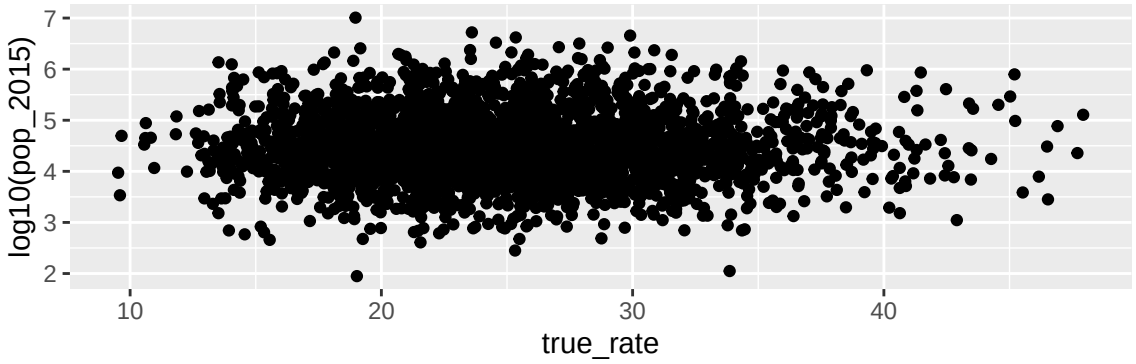
Even though these are simulated, let's think of them as the true cancer rates of these counties.

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?

Plots & Commentary on Exercise 1



As this plot looks rather messy, we may note the wide dispersion of values in the population variable 'pop_2015'. Similar to other population-wide analysis, we transform the population variable with the log scale for ease of digestion (with particular note of exponential growth between rural and urban areas).



Even with a log transformation, the above visual is messy. However, this is important to note, as it indicates—on first glance—that cancer rates have a wide dispersion generally, i.e. we don't have a strong prior that smaller (or larger) counties would have higher (or lower) incidence of cancer rates.

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 `countypop`, 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`.

```
incd <- read_csv(here("data", "incd.csv"),
                 col_names = FALSE,
                 skip = 9,
                 na = c("", "NA", "*")) %>%
  select(1, 2, 4) %>%
  slice(2:3142)
names(incd) <- c("county", "fips", "cancer_incd")
incd <- incd %>%
  mutate(county = tolower(county) %>%
           str_replace(pattern = "\\(.*\\)",
                       replacement = "")) %>%
  separate(county, into = c("county", "state"), sep = ", ") %>%
  mutate(county = str_replace(county,
                              pattern = " county",
                              replacement = ""),
         cancer_rate = cancer_incd/100000) %>%
  drop_na(cancer_rate)
x <- incd$cancer_rate
n <- length(x)
lamby <- n * x*100
```

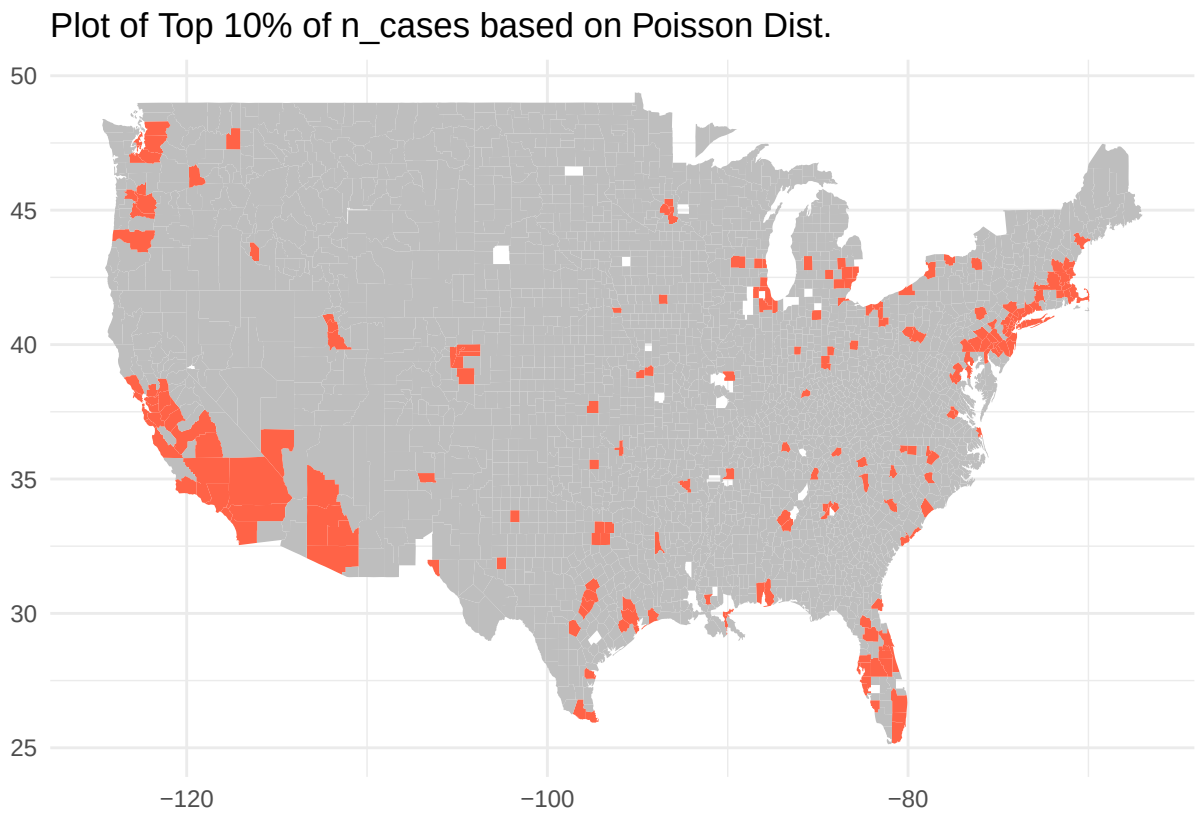
```r
set.seed(847)
library(usmap)
data(countypop)

countypop <- countypop %>%
  mutate(true_rate = rgamma(n(), alpha, beta), n_cases = rpois(n(), x*pop_2015))

library(xtable)
print(xtable(slice(countypop, 1:10)), comment = FALSE)
```

	fips	abbr	county	pop_2015	true_rate	n_cases
1	01001	AL	Autauga County	55347.00	33.47	21
2	01003	AL	Baldwin County	203709.00	28.18	74
3	01005	AL	Barbour County	26489.00	24.55	7
4	01007	AL	Bibb County	22583.00	20.82	2
5	01009	AL	Blount County	57673.00	17.98	11
6	01011	AL	Bullock County	10696.00	36.56	4
7	01013	AL	Butler County	20154.00	27.09	11
8	01015	AL	Calhoun County	115620.00	25.94	28
9	01017	AL	Chambers County	34123.00	17.06	7
10	01019	AL	Cherokee County	25859.00	26.01	6

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.



Commentary on Exercise 3

The above visual is striking in comparison to the first image on page 1, insomuch as they present very different pictures—the first indicated rates in central United States are high, but this is much *less* so when we construct rates using the Poisson distribution. That being said, there appears to be some clustering of high rates on the coasts, though not to a great extent.

One more note on the above image: It appears the top 10% of cases occur in high population areas, particularly in NE United States and California area. This intuitively makes sense, as a higher population would likely have a higher *number of cases* of the disease, though not necessarily a higher incidence rate.

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 (per 100,000 people).

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 `countypop`.

M.L.E. of Poisson:

For $X_1, X_2, ..., X_n \sim Poisson$ iid, then:

$f\left(x \mid \theta\right) = \frac{\theta^x e^{-\theta}}{x!}$

Similarly, for the likelihood function, we have:

$$f_n\left(\bar{x} \mid \theta\right) = \frac{\theta^{\sum_{i=1}^{n} x_i} \quad e^{-\theta}}{\sum_{i=1}^{n} x_i!}$$

Taking the log likelihood function, $L(\theta)$, we then have:

$$L(\theta) = \sum_{i=1}^{n} \left(x_i log\left(\theta\right) - \theta - log\left(x_i\right)\right)$$

Thus:

$$L(\theta) = log\left(\theta\right) \sum_{i=1}^{n} \left(x_i\right) - n\theta - \sum_{i=1}^{n} log\left(x_i\right)$$

Taking the derivative and setting equal to zero gives us:

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{1}{\theta} \sum_{i=1}^{n} x_i - n = 0$$

$$\rightarrow n\theta = \sum_{i=1}^{n} x_i \rightarrow \theta = \frac{1}{n} \sum_{i=1}^{n} x_i$$

As $\frac{1}{n} \sum_{i=1}^{n} x_i$ is the mean, the M.L.E. in this setting is $\hat{\theta} = \bar{X}$.

Said differently, the M.L.E. of θ is the sample mean of the observed values of incidence of county i.

Using this estimate, we then have:

```
x <- incd$cancer_rate
y <- countypop$pop_2015
```

```r
set.seed(847)
library(usmap)
data(countypop)

countypop <- countypop %>%
  mutate(true_rate = rgamma(n(), alpha, beta),
         n_cases = rpois(n(), x*pop_2015), mle = x * pop_2015)

library(xtable)
print(xtable(slice(countypop, 1:10)), comment = FALSE)
```

	fips	abbr	county	pop_2015	true_rate	n_cases	mle
1	01001	AL	Autauga County	55347.00	33.47	21	14.28
2	01003	AL	Baldwin County	203709.00	28.18	74	62.95
3	01005	AL	Barbour County	26489.00	24.55	7	6.17
4	01007	AL	Bibb County	22583.00	20.82	2	4.90
5	01009	AL	Blount County	57673.00	17.98	11	19.32
6	01011	AL	Bullock County	10696.00	36.56	4	2.25
7	01013	AL	Butler County	20154.00	27.09	11	5.50
8	01015	AL	Calhoun County	115620.00	25.94	28	36.30
9	01017	AL	Chambers County	34123.00	17.06	7	6.72
10	01019	AL	Cherokee County	25859.00	26.01	6	7.84

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 `countypop`. Using `xtable()` as we did above, print out this final table with both columns of estimates.

First, note the loss function is given by:

$L(\theta, x) = (\theta - x)^2$

Thus, the general form of the Bayes Estimator using the squared loss is generally given by:

$E\left(L(\theta, x)\right) = \int_{\Omega}(\theta - x)^2 \xi\left(\theta \mid x\right) d\theta$

Where:

$\xi\left(\theta \mid x\right) = \frac{f(x|\theta)\xi(\theta)}{g(x)}$

And:

$g(x)$ is the marginal joint p.d.f. or p.f. of $X_1, ..., X_n$

As we are using the Gamma distribution, note, for $x > 0$,

$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

The Jist

We've made this computation previously in Exercise #5 from Section 7.4.

At a high-level, given the initial specifications of the Gamma distribution, we know with the squared error loss, the Bayes estimator is the mean of the posterior distribution. As the original mean of the Gamma distribution was $\frac{\alpha_{old}}{\beta_{old}}$, our new (Bayes) Estimates are given by the Gamma distribution with new parameters parameters α_{new} and β_{new}, Where:

$\alpha_{new} = \alpha_{old} + n\bar{x}_n$

$\beta_{new} = \beta_{old} + n$

Using the above form, we then compute this estimate for each county, given in the below table.

Note, $n = 1$ for each county, so $\beta_{new} = .7144 + 1 = 1.7144$.

Similarly, $\alpha_{new} = 17.87 + MLE_i$, where MLE_i is the MLE of county i.

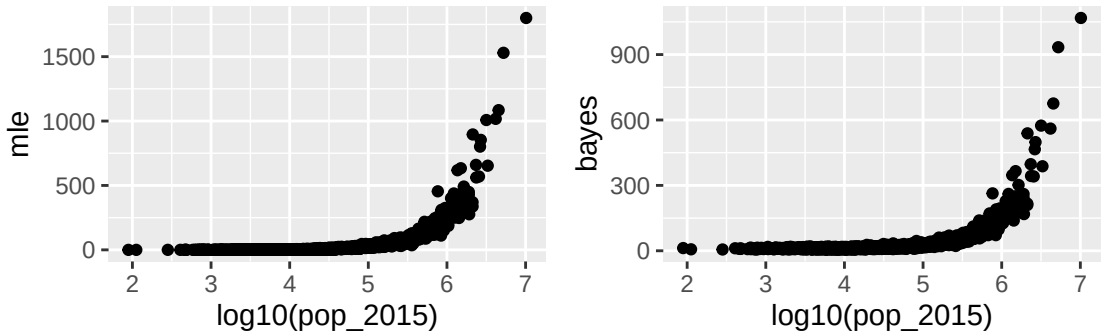
```
set.seed(847)
library(usmap)
data(countypop)

countypop <- countypop %>%
  mutate(true_rate = rgamma(n(), alpha, beta), n_cases = rpois(n(), x*pop_2015),
         mle = x * pop_2015, bayes = rgamma(n(), alpha+mle, beta+1))

library(xtable)
print(xtable(slice(countypop, 1:10)), comment = FALSE)
```

	fips	abbr	county	pop_2015	true_rate	n_cases	mle	bayes
1	01001	AL	Autauga County	55347.00	33.47	21	14.28	19.32
2	01003	AL	Baldwin County	203709.00	28.18	74	62.95	55.62
3	01005	AL	Barbour County	26489.00	24.55	7	6.17	11.32
4	01007	AL	Bibb County	22583.00	20.82	2	4.90	10.52
5	01009	AL	Blount County	57673.00	17.98	11	19.32	21.79
6	01011	AL	Bullock County	10696.00	36.56	4	2.25	7.73
7	01013	AL	Butler County	20154.00	27.09	11	5.50	10.62
8	01015	AL	Calhoun County	115620.00	25.94	28	36.30	31.40
9	01017	AL	Chambers County	34123.00	17.06	7	6.72	14.98
10	01019	AL	Cherokee County	25859.00	26.01	6	7.84	18.11

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 axis on both and each of the estimates of the y-axes. Again, be sure to transform the scales to better reveal the structure. Describe the trend that you see in each plot.



Commentary on Exercise 6

Plot 1: MLE

The MLE estimates compared to population (log normalized) is given on the left-hand side. The MLE plot appears to have a non-linear (quadratic or polynomial) fit to the data.

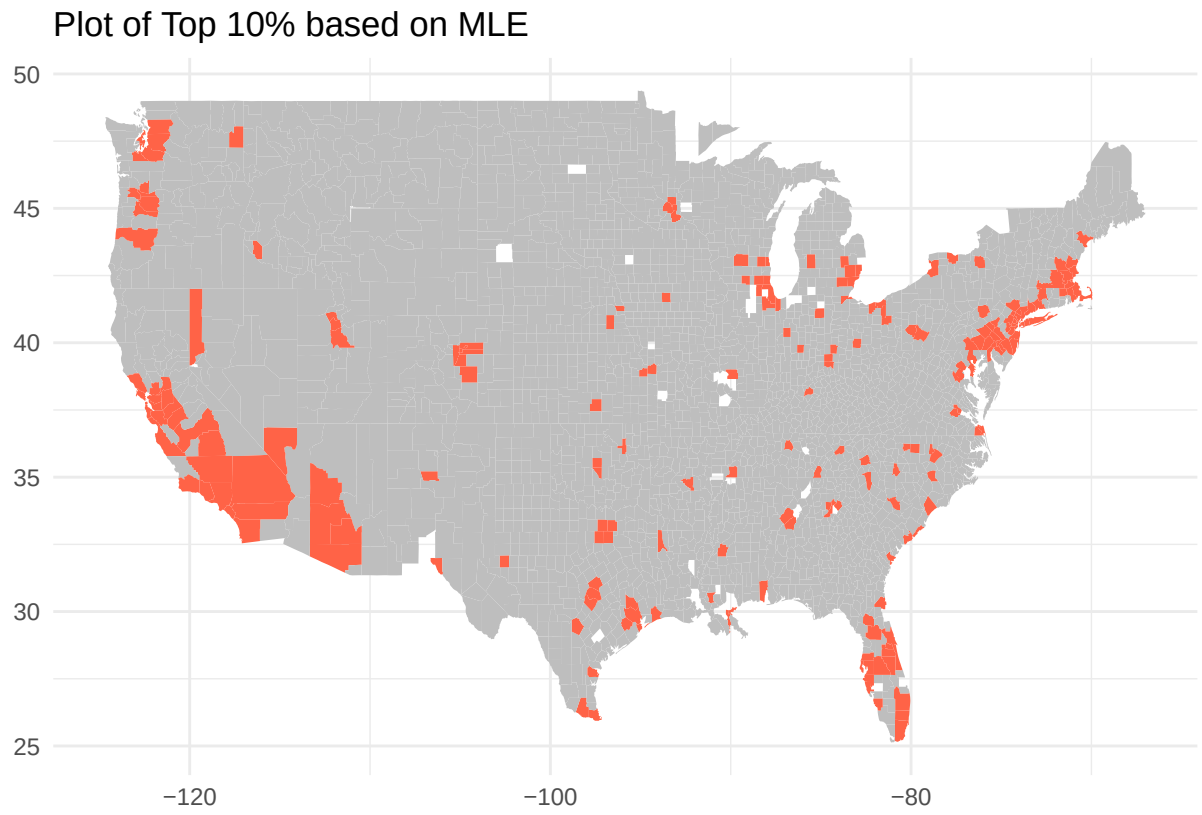
Plot 2: Bayes

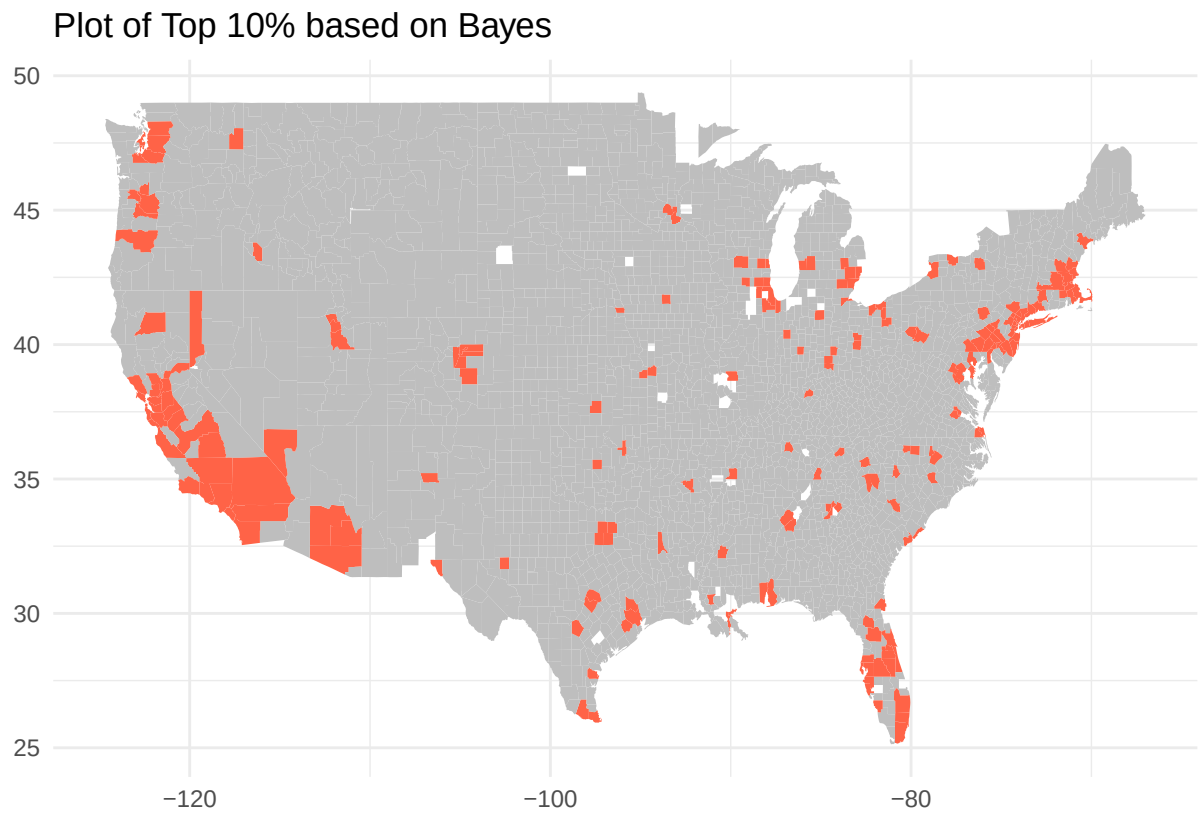
Similar to the MLE, the Bayes estimates compared to population (log normalized) is given on the right-hand side. The Bayes plot appears to have a non-linear (quadratic or polynomial) fit to the data as well.

Comparison of Both Plots

Both the MLE and Bayes estimates appear to have a non-linear (quadratic or polynomial) relationship to population (log normalized). Taken together, these estimates appear to indicate areas of the United States tend with high population (10^5 and greater) tend to have increasing estimates of having the cancer.

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?





Commentary on Exercise 7

Plot 1: MLE

It appears the cause of the dominant spatial pattern is a result of where populations are largest, e.g. Bay area California, Seattle. More broadly, though not shown, the dominant trend appears to cluster around state capitals, which tend to be highly populated relative to other counties in a state, though California is understandably an exception.

Plot 2: Bayes

Similar to the MLE spatial plot, the dominant trend of the Bayes plot is a clustering around highly populated (and urban) areas.

Given the prompt motivating this commentary, I have a lingering fear the Bayes estimate was constructed inappropriately. However, if we interprete the Bayes estimate as given in this case study, it is similar (but not the same as) to the MLE.

Comparison of Both Plots

Interestingly, the two maps have very similar shading. This is relatively unsurprising when reviewing the construction of both estimates, which utilize a simulation formulation. Nonetheless, it is worthwhile to note that some counties are dropped from the MLE to the Bayes plot, indicating that some estimates of the number of cases were overestimated in the MLE estimation compared to the Bayes estimation.

Overall, both maps indicate areas of high estimates of the number of cancer cases occur in counties with high populations, and that a number of these are located in coastal areas (in particular California, Florida, and the NE United States).