DS-GA 1002 - Homework 10

Eric Niblock

November 16th, 2020

1. (a) Amongst academy award nominees, why would we expect award winners to tend to have longer lifespans than those that haven't won awards? [Hint: The reason is statistical.]

It is likely that academy award nominees are nominated more than once. Winners have likely been nominated many times in the past, before actually winning an award after having built a career. So, it seems reasonable that the award winners are those which have longer lifespans, and longer careers.

(b) We want to use your quantitative skills to trade stocks. We go online and find 500 large US companies you might want to trade today. We train a model using data from 2011-2018, and then validate the performance on data from 2019-2020. The model does unrealistically well! Explain what went wrong. [Hint: What bias did we introduce?]

This is a case of selection bias. The top companies today were likely not the top companies during the period 2011-2018 (or at least some of them have changed). We have had prior knowledge that the companies of today would do well, and hence we have committed an act of selection bias by choosing them.

(c) In the 1950s two studies were performed to test the effectiveness of the polio vaccine. The studies focused on young children (grades 1-3) that were at the highest risk for polio. In study N, the students in grades 1 and 3 were used as the control, and were not given any treatment. In grade 2, students whose parents consented were vaccinated, and the others did not receive treatment. In study R, grades 1-3 were combined. All of the parents were asked for consent. Of the children that received consent, half were randomly given a placebo, and half were randomly given the vaccine. Neither the doctors nor the children knew which shot they were given. The results are given below.

Study N

	Size	Rate
Grade 2 (vaccine)	225,000	25
Grades 1 and 3 (control)	725,000	54
Grade 2 (no consent)	125,000	44

Study R

	Size	Rate
Treatment	200,000	28
Control	200,000	71
No consent	350,000	45

All rates are quoted in cases per 100,000 children. Interestingly, children coming from higher income families were more likely to receive consent, and more likely to be infected with polio.

For each of the following questions, your answer should be a short explanation.

i. **In study R, why was the control group given a placebo instead of nothing?**

If the control group wasn't given a placebo, then we would not be able to control for the placebo effect, in which patients react to a treatment simply because they believe it will help, and not because of any therapeutic value. Furthermore, the study would no longer be double-blind; every patient would know what group they were apart of.

ii. **Using study N, someone wants to compare the "no consent" group with rate 44 to the "vaccine" group with rate 25 to show the vaccine is effective. What is wrong with this idea?**

Again, we would not be able to control for the placebo effect. The "vaccine" group may be experiencing the effects of either the vaccine or the placebo effect, or both. Without comparison to the placebo-controlled group, we cannot discern the effect of the vaccine.

iii. **Assuming there wasn't a large disparity of polio incidents between the grades, why did the "control" group in study N have a much lower rate than the "control" group in study R?**

In study R, those who consented were split between control and treatment groups, while in study N, those in the control group did not require consent. From the information provided, we also know that consent and infection rate are correlated, which explains the higher rate within the control group in study R.

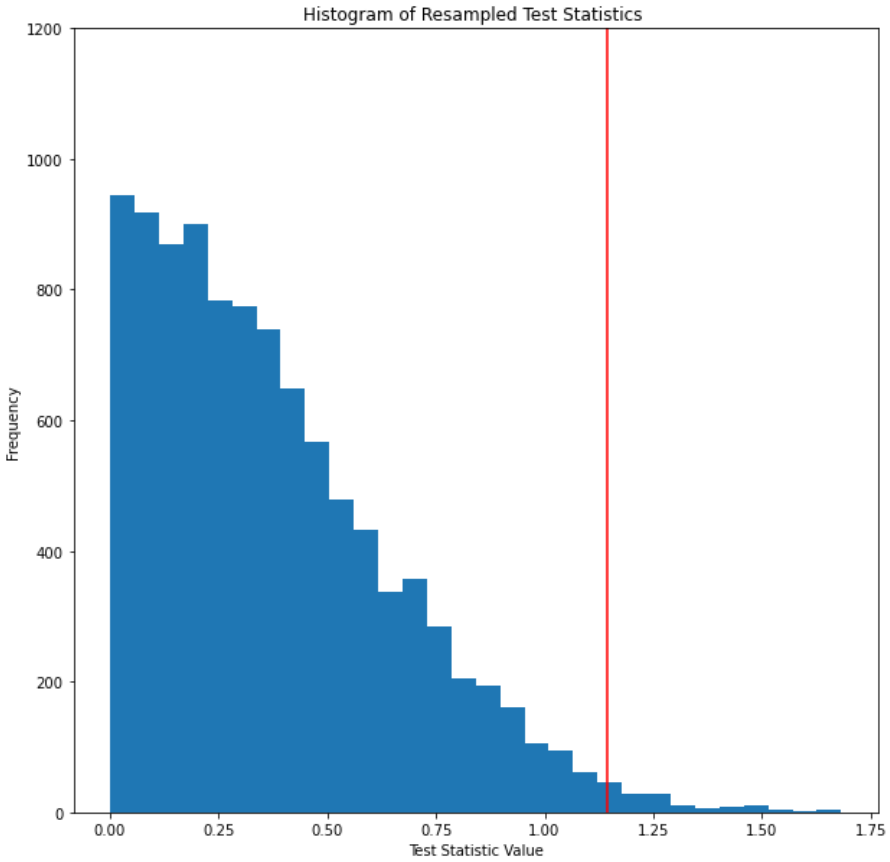
iv. Using this data, what pair of numbers would you compare to best conclude that the polio vaccine is highly effective at preventing polio?

We should compare the treatment and control groups from study R, because the control group in study R received a placebo, and thus we can control for the placebo effect within the treatment group of study R.

2. You have been given a file *rain.txt* (same as from HW 7) containing rainfall data from 52 clouds. Half of the 52 clouds were chosen at random, and treated with a chemical to increase precipitation. We assume all of the clouds are independent. Perform a permutation test using 10,000 resamples of the log-rainfalls.

(a) Plot a histogram of your resampled test statistics, and overlay a vertical line where the test statistic computed from the data lies.

The histogram was generated using data from the simulation shown in part (b). The red line indicates the value of the test statistic calculated from the data.



(b) Report your p-value.

The p-value was generated in the following script. We found how many of the permuted test statistics fell above the actual test statistic.

```
import numpy as np
data = np.loadtxt('rain.txt',skiprows=1)
```

```
untreated = data[:,0]
treated = data[:,1]
clouds = data.flatten('F')
logclouds = np.log(clouds)

T = abs(np.mean(logclouds[26:]) - np.mean(logclouds[:26]))
```

```
T_list = []
for i in range(10000):
    np.random.shuffle(logclouds)
    mean1 = np.mean(logclouds[:26])
    mean2 = np.mean(logclouds[26:])
    T_list.append(abs(mean2-mean1))
```

```
p = 0
for t in T_list:
    if t > T:
        p += 1
print('p-value: ', p/len(T_list))
```

```
p-value:  0.0132
```

3. **Suppose that 50 people are given a placebo and 50 are given a new treatment for headaches. For each of the 100 patients we determine whether their symptoms have improved. We find that 30 of the 50 people given the placebo improved, and 40 of the 50 people given the new treatment improved. Let $\tau = p_2 - p_1$ where p_2 is the probability of improving under the new treatment, and p_1 is the probability of improving under the placebo.**

   (a) **Estimate τ, find the standard error of your estimator and a 90% confidence interval for τ using the parametric bootstrap with 10,000 samples.**

   The following code generates the parametric bootstrap using 10,000 samples. The output shows our estimate for τ, the standard error of our estimator and a 90% confidence interval.

```
test = np.concatenate((np.zeros(10), np.ones(40)))
control = np.concatenate((np.zeros(20), np.ones(30)))
```

```
p_1 = sum(control)/len(control)
p_2 = sum(test)/len(test)
```

```
from scipy.stats import bernoulli
ts = []
for i in range(10000):
    test_samp = sum(bernoulli.rvs(p_2, size=50))/50
    control_samp = sum(bernoulli.rvs(p_1, size=50))/50
    ts.append(test_samp - control_samp)
```

```
tau = np.mean(ts)
sterr = np.std(ts, ddof=1)
L,U = np.quantile(ts, [0.05,0.95])
print('Estimate of Tau: ', round(tau,5))
print('Standard Error of Sample Mean: ', round(sterr,5))
print('Confidence Interval for Tau: ', (round(L,5),round(U,5)))
```

```
Estimate of Tau:  0.20052
Standard Error of Sample Mean:  0.08935
Confidence Interval for Tau:  (0.059, 0.34)
```

   (b) **Suppose p_1, p_2 have a uniform joint prior with PDF $f(p_1, p_2) = 1$ for $p_1, p_2 \in [0, 1]$ and 0 otherwise.**

      i. **Compute the posterior mean of τ.**

      Take X_i to be a Bernoulli random variable with parameter p_2 to represent individual i in the new treatment group. Take Y_i to be a Bernoulli random variable with parameter p_1 to represent individual j in the new treatment group. We know from Bayes' Theorem that,

$$f(p_1, p_2|x_1, ..., x_{n_x}, y_1, ..., y_{n_y}) \propto f(x_1, ..., x_{n_x}, y_1, ..., y_{n_y}|p_1, p_2) f(p_1, p_2) \quad (1)$$

And by conditional independence, we have,

$$\begin{aligned} f(x_1, ..., x_{n_x}, y_1, ..., y_{n_y}|p_1, p_2) &= \left(\prod_{i=1}^{n_x} f(x_i|p_1, p_2) \prod_{j=1}^{n_y} f(y_j|p_1, p_2) \right) \\ &= \left(\prod_{i=1}^{n_x} f(x_i|p_2) \prod_{j=1}^{n_y} f(y_j|p_1) \right) \\ &= \prod_{i=1}^{50} f(x_i|p_2) f(y_i|p_1) \end{aligned} \quad (2)$$

Given that $x_1, ..., x_{n_x}, y_1, ..., y_{n_y}$ is provided, and that $f(p_1, p_2) = 1$, we find,

$$f(p_1, p_2|x_1, ..., x_{n_x}, y_1, ..., y_{n_y}) \propto p_1^{30}(1-p_1)^{20} p_2^{40}(1-p_2)^{10} \quad (3)$$

Where the factor preventing equality is simply a normalizing constant. Furthermore, we can simplify the posterior, again through conditional independence,

$$f(p_1|y_1, ..., y_{n_y}) f(p_2|x_1, ..., x_{n_x}) \propto p_1^{30}(1-p_1)^{20} p_2^{40}(1-p_2)^{10} \quad (4)$$

Meaning that the equations are seperable as below,

$$f(p_1|y_1, ..., y_{n_y}) \propto p_1^{30}(1-p_1)^{20} \quad (5)$$

$$f(p_2|x_1, ..., x_{n_x}) \propto p_2^{40}(1-p_2)^{10} \quad (6)$$

Where again, the factor preventing equality is simply the absence of a normalizing factor. Finally, concerning τ, we have,

$$f(\tau|x_1, ..., x_{n_x}, y_1, ..., y_{n_y}) = f(p_2|x_1, ..., x_{n_x}) - f(p_1|y_1, ..., y_{n_y}) \quad (7)$$

$$E[\tau|x_1,...,x_{n_x},y_1,...,y_{n_y}] = E[p_2|x_1,...,x_{n_x}] - E[p_1|y_1,...,y_{n_y}] \tag{8}$$

$$\begin{aligned} E[\tau|x_1,...,x_{n_x},y_1,...,y_{n_y}] &= \frac{\alpha_2}{\alpha_2+\beta_2} - \frac{\alpha_1}{\alpha_1+\beta_1} \\ &= \frac{41}{41+11} - \frac{31}{31+21} \\ &= 0.1923 \end{aligned} \tag{9}$$

ii. By simulating 10,000 draws from the posterior distribution of τ, find a posterior 90% (credible) interval for τ.

Below, we sample from the posterior and generate an estimate for τ,

```
from scipy.stats import beta
ps_1 = beta.rvs(31,21,size=10000)
ps_2 = beta.rvs(41,11,size=10000)
```

```
taus = ps_2 - ps_1
tau = np.mean(taus)
sterr = np.std(taus, ddof=1)
L,U = np.quantile(taus, [0.05,0.95])
print('Estimate of Tau: ', round(tau,5))
print('Standard Error of Sample Mean: ', round(sterr,5))
print('Confidence Interval for Tau: ', (round(L,5),round(U,5)))
```

```
Estimate of Tau:  0.19393
Standard Error of Sample Mean:  0.08686
Confidence Interval for Tau:  (0.05096, 0.33499)
```

4. **The file *pew_data.dta* (can be read using *read_stata* in *pandas*) has data from Pew Research Center polls taken during the 2008 election campaign. The file *2008.csv* has the number of votes cast in each state for Obama and McCain.**

   (a) **Using the poll data, compute for each of the 50 states the proportion of people polled in that state that identified as 'very liberal', 'liberal', or 'moderate' (i.e., the union of these 3 categories). This will use the *state* and *ideo* columns. Note that each row in the given data represents a separate person that was polled, so you will need to aggregate the data by state (see *groupby* in *pandas*). List the proportions for Hawaii, South Dakota, and Montana.**

   Below, we generate the proportions of people who were polled that possessed a left leaning ideology. This proportion is printed concerning the requested states.

```
import pandas as pd
pew = pd.read_stata(r'pew_data.dta')
votes = pd.read_csv(r'2008.csv')
```

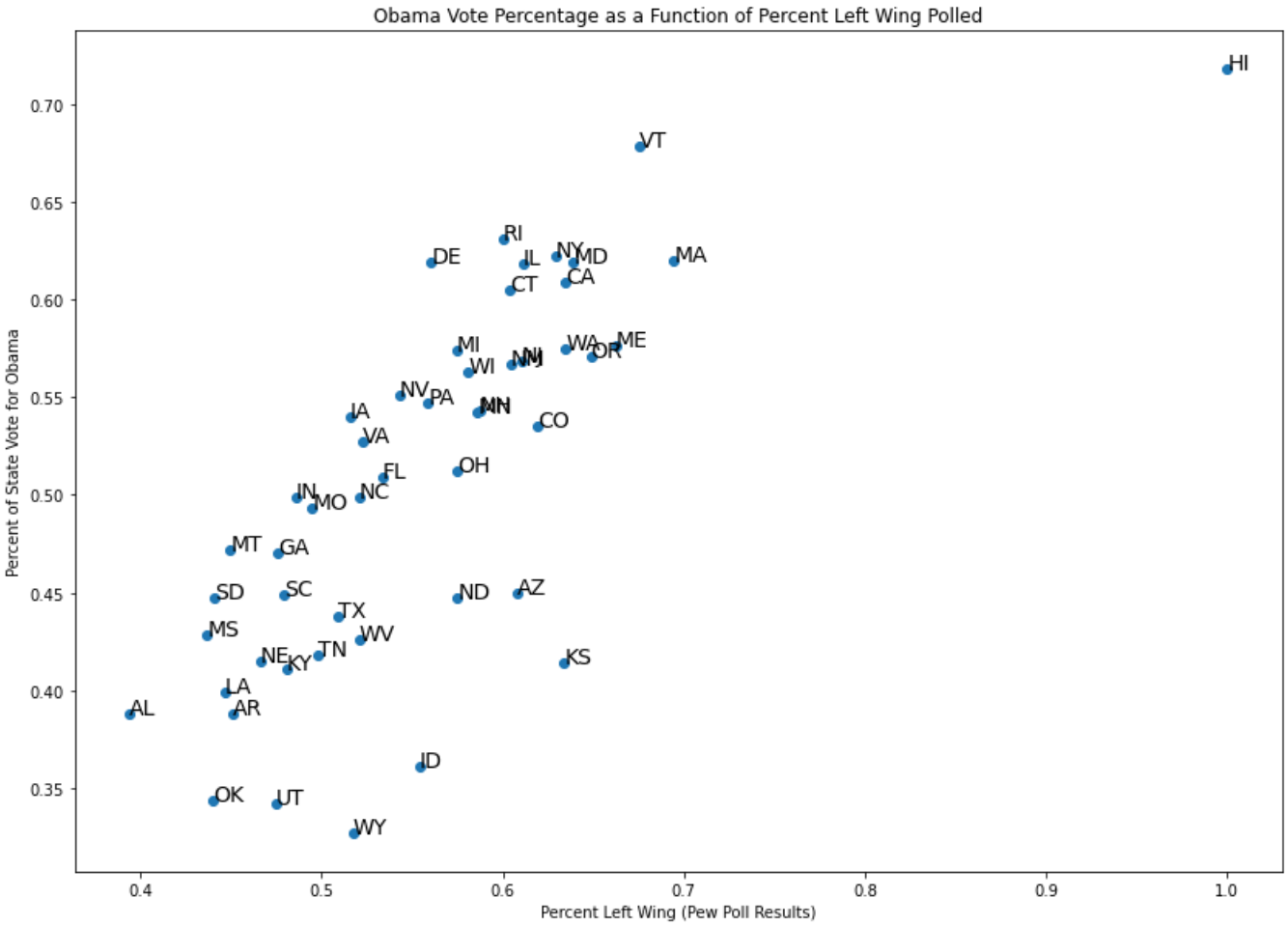
```
temp = pew.groupby(['state']).count()
states = temp.index.values.tolist()
grouped = pew.groupby(['state','ideo']).count()
```

```
pew_dict = {}
for state in states:
    current = grouped.loc[state]
    current = current.fillna(0)
    proportion = (current.loc['very liberal'][0] + \
                  current.loc['liberal'][0] + \
                  current.loc['moderate'][0])/current['survey'].sum()
    pew_dict[state] = (proportion,current['survey'].sum())
print('Proportion of Left Wing - Hawaii:         ', pew_dict['hawaii'][0])
print('Proportion of Left Wing - South Dakota:   ', pew_dict['south dakota'][0])
print('Proportion of Left Wing - Montana:        ', pew_dict['montana'][0])
```

```
Proportion of Left Wing - Hawaii:          1.0
Proportion of Left Wing - South Dakota:    0.44086021505376344
Proportion of Left Wing - Montana:         0.449438202247191
```

   (b) **Produce a scatter plot of the proportion polled that are 'very liberal', 'liberal', or 'moderate' against *vote_Obama_pct* for all states with polling data (each point on the plot is a single state). Annotate each point on the plot with the two letter abbreviation of the corresponding state. [Hints:**

   - **Install the *us* Python package, and use *us.states.lookup(statename).abbr* to get the two letter abbreviation. Use *plt.annotate* to annotate the plot.**
   - **Washington D.C. (District of Columbia) is not a state.]**



(c) Consider the following Bayesian model for the polling data. Let θ_j denote the proportion of the people polled in state j that identify as 'very liberal', 'liberal', or 'moderate.' Assume each θ_j has prior distribution $Gamma(\alpha, \beta)$ with $\alpha = 30$ and $\beta = 54$ and that the θ_j-values are independent. Let n_j denote the total number of people polled in state j. Assume the number of people X_j that identify as 'very liberal', 'liberal', or 'moderate' in state j follows the model

$$X_j \sim Poisson(n_j \theta_j)$$

for $j = 1, ..., 50$, with each state independent. Give a formula for the posterior mean $E[\theta_j | X_j = x_j]$ of θ_j in terms of n_j and x_j.

From the already derived result, we know that the posterior mean given a $Gamma(\alpha, \beta)$ prior and data from a $Poisson(\lambda)$ is

$$E[\lambda|X_1, ..., X_n] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n} \tag{10}$$

So, in our present scenario, we have,

$$E[\theta_j|X_j = x_j] = \frac{\alpha + x_i}{\beta + n_j} \tag{11}$$

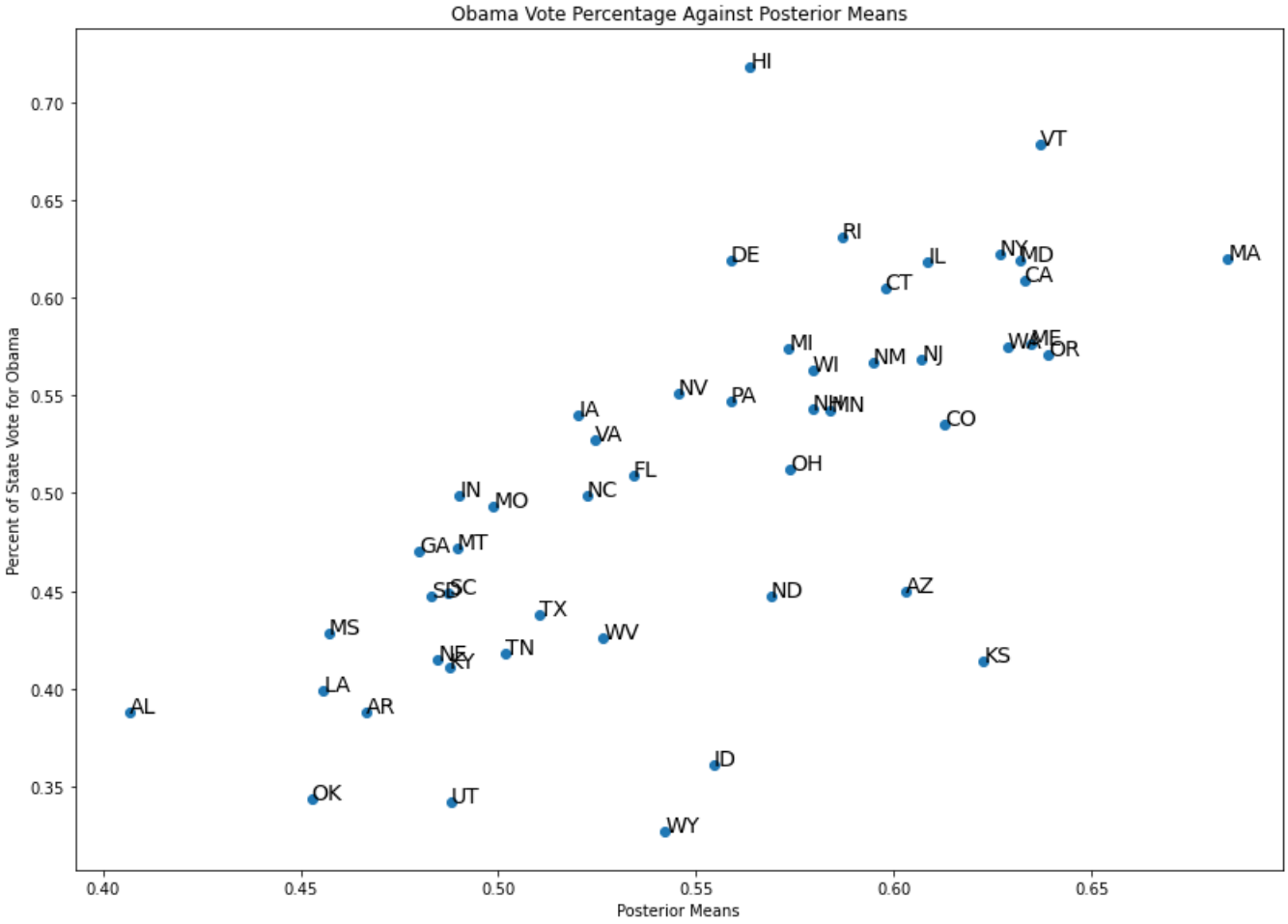
(d) List the posterior means for Hawaii, South Dakota, and Montana.

The following code was used to generate the posterior means of every state. The results for the desired states are printed,

```
ps = []
for state in states:
    post = (30 + pew_dict[state][0]*pew_dict[state][1])/(54 + pew_dict[state][1])
    ps.append(post)
    if state in ['hawaii','south dakota','montana']:
        print('Posterior Mean - ', state,':', post)
```

```
Posterior Mean -  hawaii : 0.5636363636363636
Posterior Mean -  montana : 0.48951048951048953
Posterior Mean -  south dakota : 0.48299319727891155
```

(e) Produce a scatter plot of the posterior means for θ_j against *vote_Obama_pct* for each state (each point on the plot is a single state). If a state has no polling data, use the prior mean. Annotate each point on the plot with the two letter abbreviation of the corresponding state.


Obama Vote Percentage Against Posterior Means
Percent of State Vote for Obama
Posterior Means
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.40
0.45
0.50
0.55
0.60
0.65
HI
VT
RI
DE
IL
NY
MD
MA
CT
CA
MI
WA
ME
OR
WI
NM
NJ
NV
PA
NH
MN
IA
CO
VA
OH
FL
IN
NC
MO
GA
MT
SD
SC
ND
AZ
TX
MS
WV
NE
TN
KY
KS
LA
AL
AR
ID
OK
UT
WY