Winter 18

The American Dream - Upward Economic Mobility

Jack Schaffer

STATS 485 – Capstone Seminar

Table of Contents

Introduction

In the United States, economic divide has been on the forefront of national topics, especially during the most recent election. Due to the increasing financial inequality in America, many believe the American Dream does not exist. An article by the New York Times, *The American Dream, Quantified at Last,* suggests that since the 1940s (when the term "American Dream" was first coined during the Great Depression), an individual's chance at making more money than their parents has declined considerably (Leonhardt 17).

This paper investigates household incomes of individuals living in various commuting zones (local labor markets) around the United States in the 1980s and 90s and compares it to their personal income in 2010. In doing so, we are able to identify whether individuals achieved economic upward mobility by the time they turned 30 years old. The data for this study comes from the United States Department of Agriculture's 1980 and 1990 commuting zones and labor market areas dataset and federal tax records from the Department of the Treasury. Utilizing this data, individuals with low household income in the 1990s were identified and their personal income by 2010 was observed to determine upward mobility. For this study, we determined an individual achieved upward mobility if their income was among the top quintile of personal incomes in 2010 and investigated the population proportions of the general population, the four U.S. Census Bureau regions, and forty commuting zones to decipher whether or not the American Dream is still apparent in American society.

Methods

Sampling Method

The commuting zones that were used as a part of this study were selected according to a probability proportional to size scheme, with the very largest commuting zones being sampled with probability 1. The reason this sampling technique was implemented was because the populations of the zones vary significantly. Overall, it is likely to provide a more efficient estimate of the total population.

Within each zone, sub-sampling was done uniformly at random resulting in a simple random sample of the sub-population corresponding to that commuting zone. In other words, each individual had an equal chance at being selected. This technique assumes that each of the four regions defined by the U.S. Census Bureau have a roughly equal size in population. However, the Northeast is 20% smaller and the South is 50% larger, while the West and Midwest are similar. This results in a representative sample for each region, however there is a "skewed" sample of the general U.S population due to overrepresentation of the Northeast and underrepresentation of the South. While the general population may be slightly misrepresented, the benefit of this technique is that it is the easiest and quickest to apply and no further steps must be taken to define the sample.

Statistical Analysis

The methods used to analyze economic mobility were the Clopper-Pearson Confidence Interval, due to the fact that it has no sample requirements, is an exact interval, and performs effectively when evaluating probabilities near 0 and 1, and the log likelihood ratio. By stating that the Clopper-Pearson interval has no sample requirements, we mean to say that for any number of individuals in the sample, we can construct an exact interval that holds (1-α)% confidence, while various other confidence intervals have sampling requirements. An example of this is the standard confidence interval (or Wald Interval), which requires that n*p and n*(1−p) are ≥5 (or 10), where n is the number of individuals and p is the population proportion. Additionally, this interval is especially suitable because our population proportions for upward mobility range from about 0 to 0.16, where most values are very close to 0. The increased stability of the Clopper-Pearson interval with population proportions around 0, compared to other confidence intervals such as the Wald and Wilson Intervals, can be seen from figure 1.

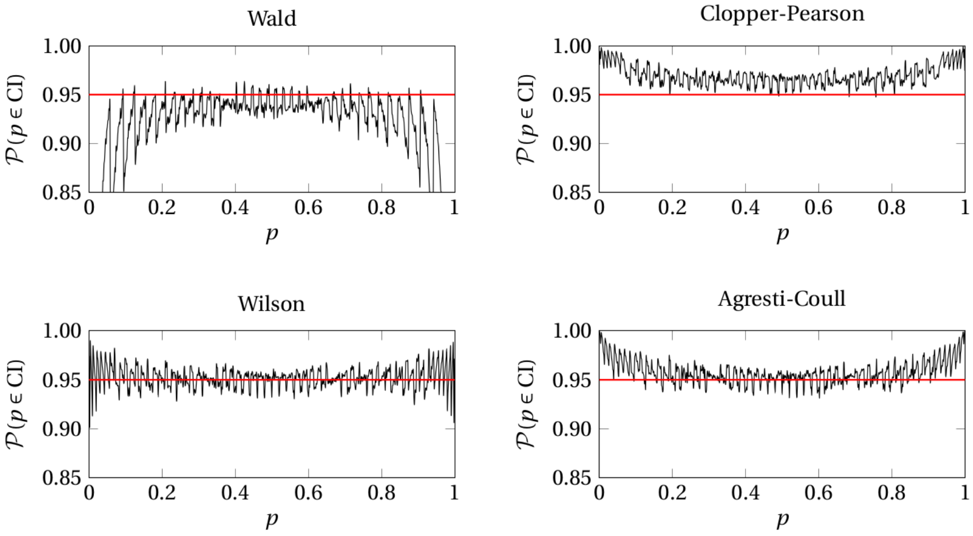


Figure 1: Confidence Interval Coverage Probability for n = 50

In a journal entry called *Interval Estimation for a Binomial Proportion* by Brown, Cai and DasGupta (BCD), the authors claim the Clopper-Pearson interval is wastefully conservative because it ensures the coverage probability is always equal to or above the nominal confidence interval (Brown 113). However from *figure 1*, we can see that the Wald and Wilson intervals have reduced coverage probability below 95% for various population proportions close to 0, while the Clopper-Pearson interval ensures at least 95% confidence. Conversely, the Clopper-Pearson interval results in a deficiency in length compared to other intervals, which is BCDs main argument against the Clopper-Pearson interval. While length is something to consider, in a response to BCD, Corcorana and Mehta refute BCD's statement by inferring that statisticians desire reliable, accurate solutions and support the use of the Clopper–Pearson interval because of its exactness (Brown 122).

In order to interpret whether or not each of the 40 zones and each of the four regions has similar upward mobility or vary in true upward mobility, the log likelihood ratio test was conducted for both cases. This method was selected over other proportion equality tests due to its effectiveness with smaller sample sizes and our zone sample sizes range from 12 to 138.

Results

The results of the Clopper - Pearson Confidence intervals for each zone are shown in *figure 2.*

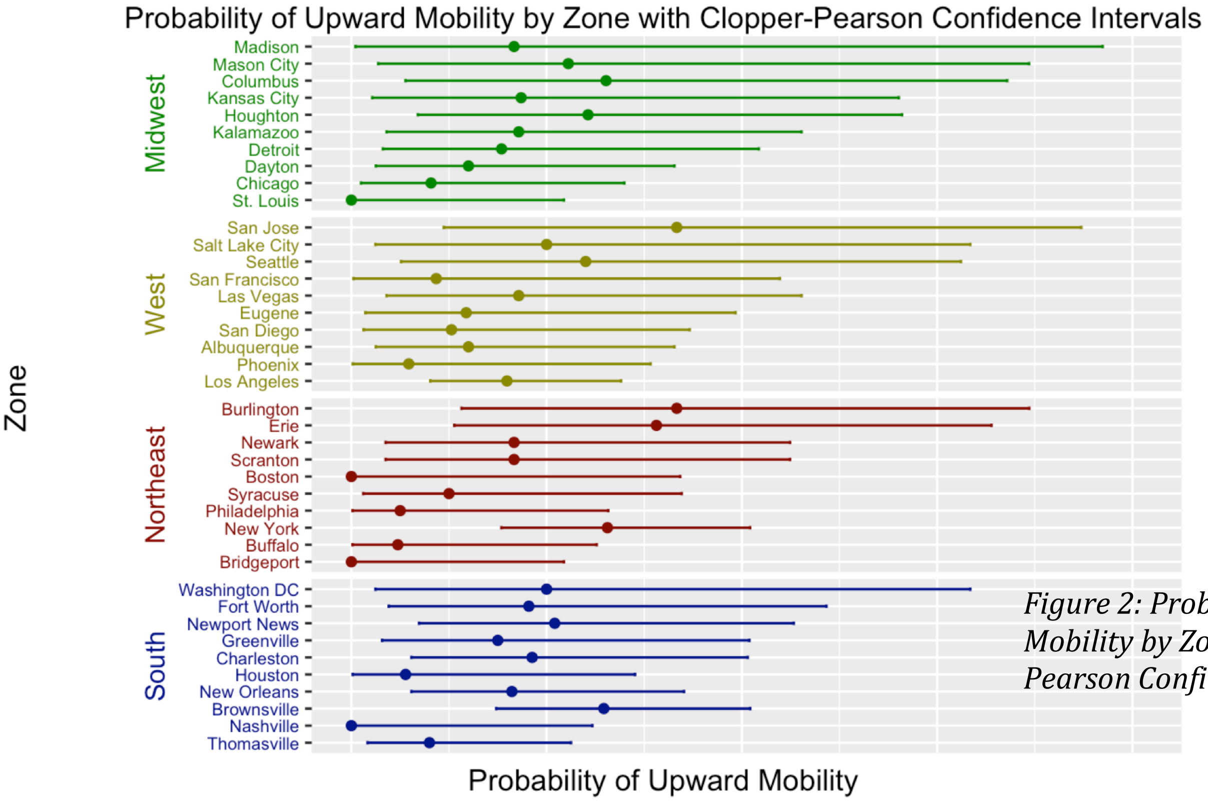


Figure 2: Probability of Upward Mobility by Zone with Clopper-Pearson Confidence Intervals

Figure 2 shows the observed probability of upward mobility for each zone and the corresponding confidence intervals. This graph shows us, for each corresponding zone, we are 95% confident that these intervals contain the true population proportion, and reject probabilities outside of these bounds. For example, in Madison, Wisconsin the limits for the confidence interval are [.002,0.385]. Thus we claim, with 95% confidence, the true probability of upward mobility for this zone is between [.002, 0.385].

As for regions, the results are showcased in *table 1* and *figure 3*.

Region	Probability of Upward Mobility	Lower Limit	Upper Limit
Midwest	0.073	0.047	0.108
West	0.076	0.052	0.105
South	0.082	0.060	0.109
Northeast	0.084	0.059	0.114

Table 1: Probability of Upward Mobility and Confidence Interval Limits by Region

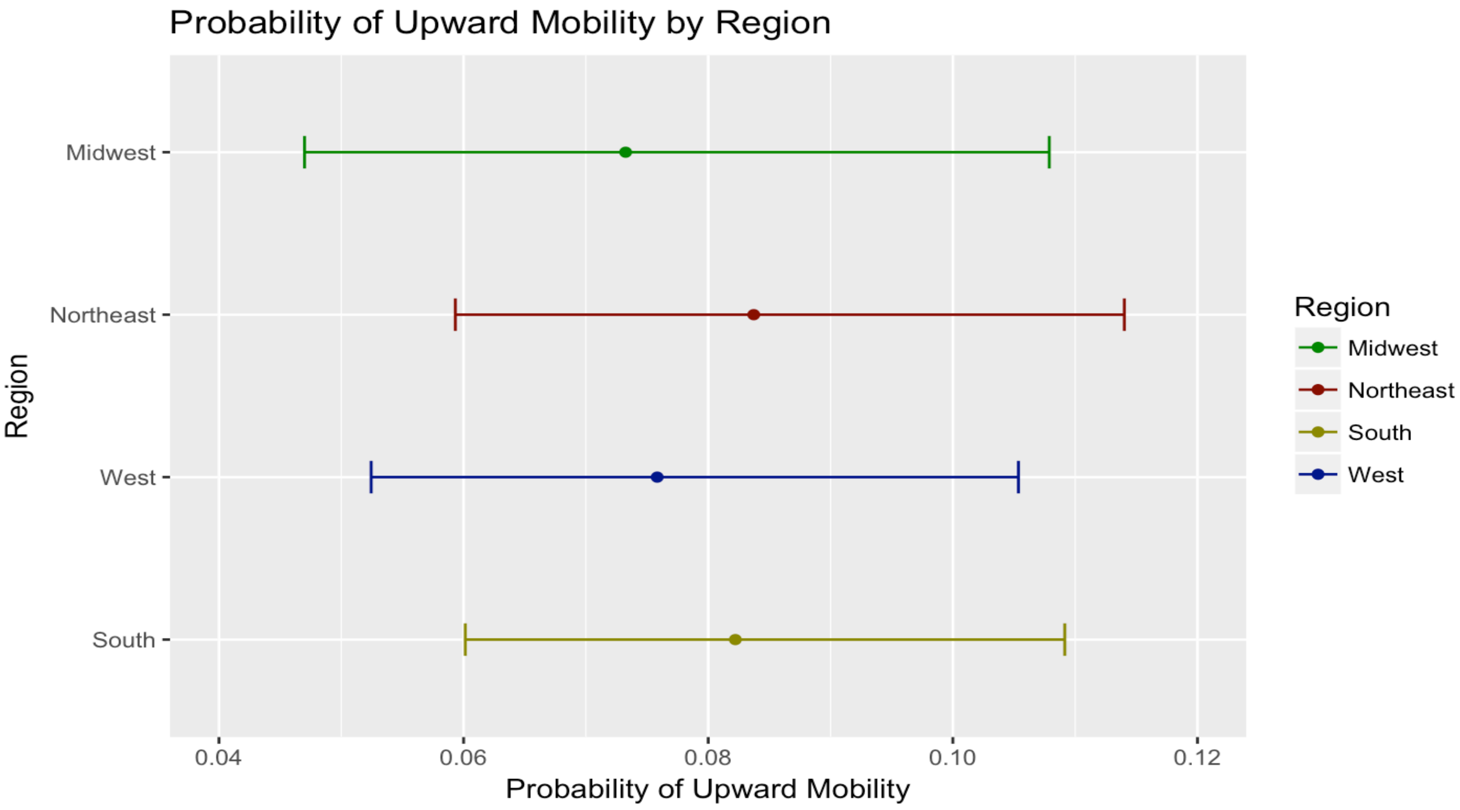


Figure 3: Probability of Upward Mobility and Confidence Interval Limits by Region

According to *table* 1, the Northeast has the highest observed probability of upward mobility at 8.4%, while the Midwest has the lowest at 7.3%. Interestingly, the confidence intervals of each region in *figure 3* do not suggest any significant difference in true population proportion among regions.

Further analysis was conducted to observe the general population and conclude whether or not specific zones and or regions have a larger true population proportion of upward mobility compared to each other. The general population has an observed population proportion of 0.079 and a confidence interval of [0.067, 0.093]. Thus, we are 95% confident that the true population proportion for upward mobility among the general population is between [0.067, 0.093]. From our observed sample, only 7.9% of the population achieved economic upward mobility, bounded by 6.7% and 9.3%. From this, we can infer that less than 1 in 10 individuals in low socioeconomic households reach the top quintile of incomes in 2010. Thus, it is misleading to suggest that the American Dream is easily obtainable for those at the lowest end of the financial spectrum in America.

While less than 10% of individuals in low socioeconomic households reached the top quintile of incomes in 2010, the log likelihood tests conducted for zone and region suggest that living in certain zones may increase an individual's probability of realizing upward mobility compared to others. The likelihood ratio test for zones resulted in a p-value of 7.01e-11, thus we reject the null hypothesis and infer zones vary in probability of upward mobility. However, the calculation for regions resulted in a p-value of .919, which is greater than α at a 95% confidence level thus we do not reject the null hypothesis and conclude the regions are similar in upward mobility.

Discussion

As indicated in the results, less than 10% of individuals in low socioeconomic households achieve economic upward mobility, however the zone in which an individual resides increases ones potential of doing so. The American Dream as we know it is diminishing through the indication that few Americans are growing financially and reaching the top quintile of incomes in the United States. However, while this study suggest the American Dream is no longer apparent, our interpretation of the American dream and upward mobility may not be true as it relates to the realities of society. Perhaps the American Dream should not be defined as reaching the top quintile of incomes, but rather climbing socioeconomic status generationally, consistently making more than your parents.

With respect to the statistical methods in this study it is important to note that there are other methods suitable for this analysis. While the Clopper-Pearson interval ensures exactness there are a few downsides that should be noted. Clopper-Pearson intervals are known to over fit the confidence level (shown in *figure 1*), resulting in even greater coverage than necessary, which produces deficiencies in length. Additionally, this increased conservatism can be heightened with smaller sample sizes. An alternative method, such as the Jeffrey's interval (which is suggested by BCD for small sample sizes) would be a suitable substitute for the Clopper-Pearson method, especially when calculating the intervals for the 40 zones because some sample sizes are below 30 (Brown 1). However, the Jeffrey's interval has limitations in coverage probability with population proportions around

0 (shown in *figure 4*), which is why the Clopper – Pearson interval was chosen for this analysis.

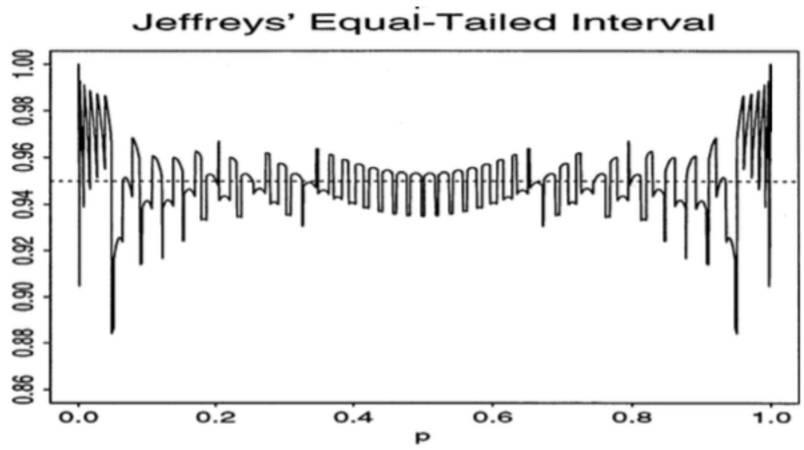


Figure 4: Coverage Probability of the Jeffrey's Equal-Tailed Confidence Interval for n = 50

The findings of this study state that the American Dream is no longer prevalent among those struggling financially, however that does not mean it cannot be revitalized. Further investigation of the zones that have a larger proportion of individuals achieving economic mobility can be performed to indicate what systematic opportunities, market conditions, or legislation contribute to financial improvement for their residing citizens. Additionally, observations of the individuals specifically can be conducted to pinpoint attributes of their lifestyle or achievements, such as education level, that may correlate to their improved financial standing.

Works Cited

Brown, Lawrence D., et al. “Interval Estimation for a Binomial Proportion.” *Statistical Science*, Institute of Mathematical Statistics, projecteuclid.org/euclid.ss/1009213286.

Leonhardt, David. “The American Dream, Quantified at Last.” *The New York Times*, The New York Times, 8 Dec. 2016, www.nytimes.com/2016/12/08/opinion/the-american-dream-quantified-at-last.html.

Appendix

The following code is required in order to carry out the computations of this paper.

```r
# Load Package Dependencies
library(PropCIs) # Clopper-Pearson Confidence Interval Method
library(ggplot2) # Visualization Tool
library(knitr) # Used to Create Tables

# Read mobility0.csv and store it as dataframe variable named mobility
mobility = read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/mobi
lity0.csv")

# Set rownames
row.names(mobility) <- mobility$zone
```

*The data used in this analysis can be found at link provided above in the chunk of code.

Overview

The computations in this document explore upward economic mobility of individuals within the four U.S. Census Bureau regions: Midwest, Northeast, South, and West. The following code consists of statistical analysis on each of the 40 individual zones and four U.S Census Bureau regions, which includes the calculation for Clopper-Pearson confidence intervals. In order to carry out this analysis, various data structures and functions were used to structure our data and compute effectively in addition to the generation of visuals and figures to represent data.

R Code

Confidence Intervals and Statistical Analysis

To begin the investigation of upward mobility in each zone, a new column vector named k is added to our mobility data frame. The *k* column variable, represents the # of individuals that achieved upward mobility in each zone and is calculated with the following equation:

$$k = n * phat$$

This is shown in code below:

```
# Column vector k is created by multiplying each row by n.lowstart by p.upmover
k = mobility$n.lowstart * mobility$p.upmover

# Next, we will add this vector to our mobility dataframe to associate each zone with the # of individuals that achieved upward mobility.
mobility = data.frame(mobility, k)
```

By utilizing the modified mobility data frame, two column vectors that contain the lower and upper bounds of the confidence interval, denoted as *l.conf_i* and *u.conf_i* respectively, are constructed.

Clopper Pearson Confidence Interval - PropCIs Package Method

The code below uses the function exactci(…) from the PropCIs package to create the lower and upper bounds of the Clopper-Pearson confidence intervals for each zone:

```
conf.int = c()
l.conf_i_PropCIs = c()
u.conf_i_PropCIs = c()

for (row in 1:nrow(mobility))
{
  conf.int = c(conf.int, exactci(mobility[row,"k"], mobility[row,"n.lowstart"], conf.level = 0.95))
}

for (i in 1:40){
  l.conf_i_PropCIs = c(l.conf_i_PropCIs, conf.int[[i]][1])
  u.conf_i_PropCIs = c(u.conf_i_PropCIs, conf.int[[i]][2])
}

# Now that we have two column vectors of the lower and upper bounds of the interval, we initialize a new data.frame instance and include these two column vectors
mobility = data.frame(mobility, l.conf_i_PropCIs, u.conf_i_PropCIs)
```

Clopper Pearson Confidence Interval - Quantile of Beta Method

In order to test the validity of the packages calculations the values for the lower and upper bounds of each zones confidence interval are calculated through the Clopper-Pearson equation, which is:

$$ClopperPearson\ CI = [alpha/2 * quantile\ of\ Beta(n * phat, n * (1 - phat) + 1), (1 - alpha/2 * quantile\ of\ Beta(n * phat + 1, n * (1 - phat))]$$

In base R, the qbeta(…) function utilizes the above equation to calculate confidence intervals and is shown below:

```
# First we set out alpha value to indicate a 95% confidence interval
alpha = rep(0.05, length(mobility$zone))
```

```
# Next we calculate the lower and upper CP CI using the qbeta(...) func
tion which utilizes the equation shown above.

l.conf_i_calc = qbeta(alpha/2,mobility$k, mobility$n.lowstart - mobilit
y$k + 1)

u.conf_i_calc = qbeta(1-alpha/2, mobility$k + 1, mobility$n.lowstart -
mobility$k)


# Now that we have two column vectors of the lower and upper bounds of
the interval, we initialize a new data.frame instance and include these
two column vectors
mobility = data.frame(mobility, l.conf_i_calc, u.conf_i_calc)
```

By viewing the mobility data matrix columns, we can see that both of the CIs are the same for the package calculation and qbeta(...) calculation:

```
head(mobility[,9:12])

##                  l.conf_i_PropCIs u.conf_i_PropCIs l.conf_i_calc
## Albuquerque          0.0125485878        0.1654819  0.0125485878
## Phoenix              0.0007443642        0.1532677  0.0007443642
## Salt Lake City       0.0123485272        0.3169827  0.0123485272
## San Jose             0.0473536266        0.3738417  0.0473536266
## San Francisco        0.0011001686        0.2194866  0.0011001686
## Las Vegas            0.0180376398        0.2305750  0.0180376398
##                  u.conf_i_calc
## Albuquerque          0.1654819
## Phoenix              0.1532677
## Salt Lake City       0.3169827
## San Jose             0.3738417
## San Francisco        0.2194866
## Las Vegas            0.2305750
```

Log Likelihood Ratio Test

Here is how the log likelihood ratio tests were calculated:

```
# Log Likelihood Ratio Test
n_l = mobility$n.lowstart
n_l = as.numeric(n_l)
phat_u = mobility$p.upmover
phat_u = as.numeric(phat_u)
phat_null = sum(mobility$n.lowstart * mobility$p.upmover)/sum(mobility$
n.lowstart)

log_lik <- function(p) {
  sum(dbinom(as.integer(n_l * phat_u), size = n_l, prob = p, log = T))
}

test_statistic = log_lik(phat_u) - log_lik(phat_null)
test_statistic

p_value = pchisq(2*(test_statistic),3,lower=F)
p_value
```

Formulation of Region Data

Furthering our analysis, we observe the differences of the four commuting zones.

```
# Create a data matrix for each commuting zone: west, midwest, northeas
t, south
midwest = subset(mobility, region %in% "midwest")
northeast = subset(mobility, region %in% "northeast")
south = subset(mobility, region %in% "south")
west = subset(mobility, region %in% "west")
```

Now that we have created subsets of zones with respect to their specific region, we can create variables for *n* (# of indviduals with a lowstart), *phat* (observed probability of upward mobility), and *k* (# of individuals that achieved upward mobility). After computing these values for each region, we will generate a mobility dataframe for regions.

```
# Initialize regionMobility dataframe variable
regionsName = c("midwest", "northeast", "south", "west")
regionMobility = data.frame(matrix(nrow = 4, ncol = 3))
rownames(regionMobility) = regionsName
colnames(regionMobility) = c("n", "phat", "k")

# Next, we will create a list that contains our region dataframes and r
un a for loop to calculate n, phat, and k for each region.
regions = list(midwest, northeast, south, west)
i = 1
for (region in regions){
  regionMobility[i,] = list(sum(region$n.lowstart), sum(region$k)/sum(r
egion$n.lowstart), sum(region$k))
  i = i + 1
}

regionMobility = data.frame(regionMobility, regionsName)
```

Next, we generate Clopper-Pearson confidence intervals using the Beta method:

```
# First we set out alpha value to indicate a 95% confidence interval
alpha = rep(0.05, 4)

# Next we calculate the lower and upper CP CI using the qbeta(...) func
tion which utilizes the equation shown above.

l.conf_i_calc = qbeta(alpha/2,regionMobility$k, regionMobility$n - regi
onMobility$k + 1)

u.conf_i_calc = qbeta(1-alpha/2, regionMobility$k + 1, regionMobility$n
- regionMobility$k)


# Now that we have two column vectors of the lower and upper bounds of
the interval, we initialize a new data.frame instance and include these
two column vectors
regionMobility = data.frame(regionMobility, l.conf_i_calc, u.conf_i_cal
c)
```

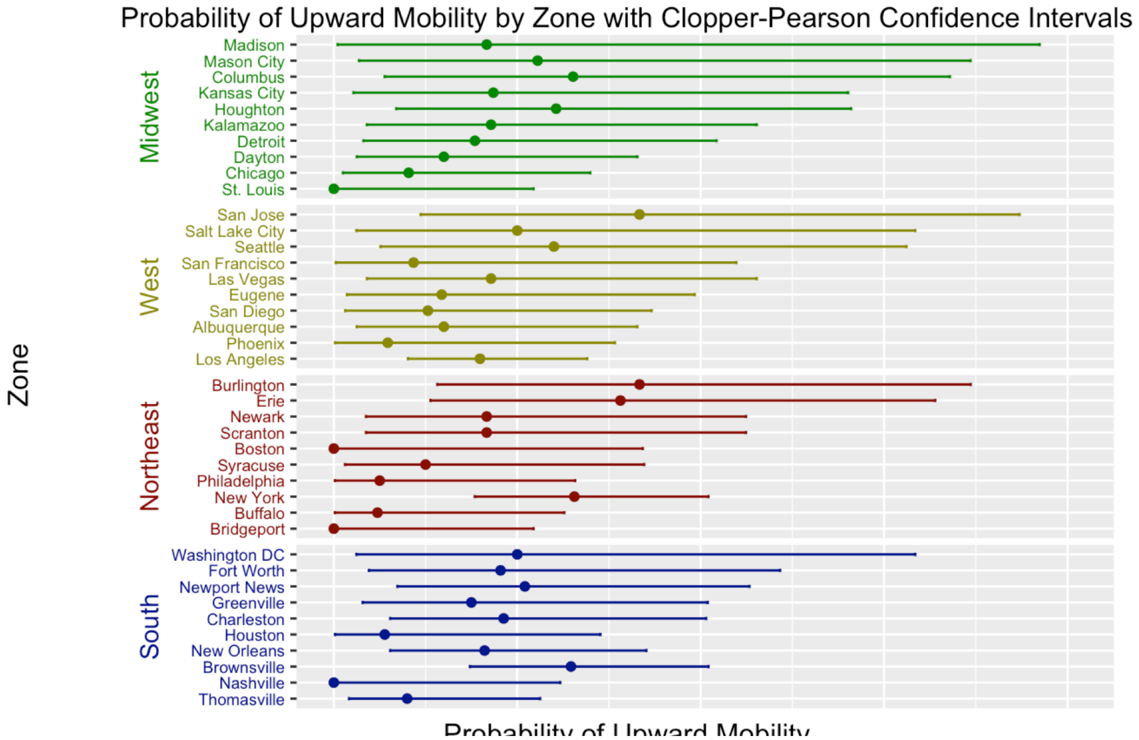
Figures and Visualizations

This code block contains a sorting of the regions to allow graphing in ascending order by confidence interval length and the creation of plots for each region:

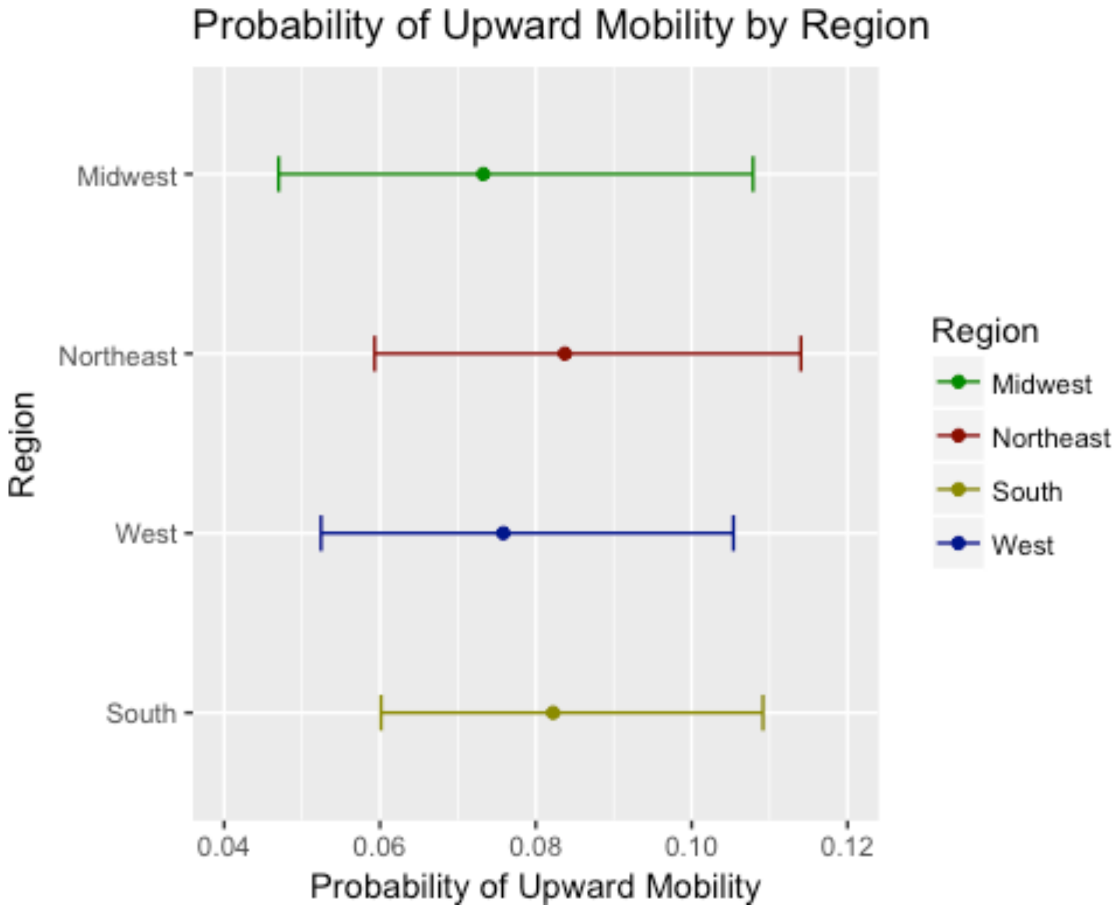
```
### midwest
conf_i_length = vector("numeric")
conf_i_length = midwest$u.conf_i_calc - midwest$l.conf_i_calc
midwest = data.frame(midwest, conf_i_length)
midwest <- midwest[order(midwest$p.upmover,midwest$conf_i_length),]
midwest$zone <- factor(midwest$zone, levels = midwest$zone[order(midwes
t$conf_i_length)])

### northeast
conf_i_length = vector("numeric")
conf_i_length = northeast$u.conf_i_calc - northeast$l.conf_i_calc
northeast = data.frame(northeast, conf_i_length)
northeast <- northeast[order(northeast$conf_i_length),]
northeast$zone <- factor(northeast$zone, levels = northeast$zone[order(
northeast$conf_i_length)])

### south
conf_i_length = vector("numeric")
conf_i_length = south$u.conf_i_calc - south$l.conf_i_calc
south = data.frame(south, conf_i_length)
south <- south[order(south$conf_i_length),]
south$zone <- factor(south$zone, levels = south$zone[order(south$conf_i
_length)])

### west
conf_i_length = vector("numeric")
conf_i_length = west$u.conf_i_calc - west$l.conf_i_calc
west = data.frame(west, conf_i_length)
west <- west[order(west$conf_i_length),]
west$zone <- factor(west$zone, levels = west$zone[order(west$conf_i_len
gth)])
```

```
### mobility
conf_i_length = vector("numeric")
conf_i_length = mobility$u.conf_i_calc - mobility$l.conf_i_calc
mobility = data.frame(mobility, conf_i_length)
mobility <- mobility[order(mobility$region,mobility$p.upmover, mobility
$conf_i_length),]
mobility$zone <- factor(mobility$zone, levels = mobility$zone[order(mob
ility$conf_i_length)])
mobility$region <- factor(mobility$region, levels = c("midwest", "north
east", "south", "west"))

# Midwest Plot
midwestPlot <- ggplot(data=midwest, aes(x= reorder(zone, conf_i_length)
, y=p.upmover, group = 1)) + geom_point(color = "green4") + labs(x = "M
idwest" , y = "Probability of Upward Mobility") + coord_flip() + geom_e
rrorbar(aes(ymin = midwest$l.conf_i_calc, ymax = midwest$u.conf_i_calc)
, width = 0.2, color = "green4") + theme(axis.text.y = element_text(siz
e = 7, color = "green4")) + theme(plot.margin = (unit(c(.001, .001, .00
1, .001), "cm"))) + theme(aspect.ratio=.2) +  theme(axis.title.x=elemen
t_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank(
)) + theme(axis.title = element_text(color="green4")) + scale_y_continu
ous(limits = c(0,0.405), breaks = c(0,0.1,0.2,0.3,0.4))

# Northeast Plot
northeastPlot <- ggplot(data=northeast, aes(x= reorder(zone, conf_i_len
gth), y=p.upmover, group = 1)) + geom_point(color = "red4") + labs(x =
"Northeast" , y = "Probability of Upward Mobility") + coord_flip() + ge
om_errorbar(aes(ymin = northeast$l.conf_i_calc, ymax = northeast$u.conf
_i_calc), width = 0.2, color = "red4") + theme(axis.text.y = element_te
xt(size = 7, color = "red4")) + theme(plot.margin = (unit(c(.001, .001,
.001, .001), "cm"))) +  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) + theme(aspect.ratio=.2) + theme(
axis.title = element_text(color="red4")) + scale_y_continuous(limits =
c(0,0.405), breaks = c(0,0.1,0.2,0.3,0.4))

# South Plot
southPlot <- ggplot(data=south, aes(x= reorder(zone, conf_i_length), y=
p.upmover, group = 1)) + geom_point(color = "blue4") + labs(x = "South"
, y = "Probability of Upward Mobility") + coord_flip() + geom_errorbar(
aes(ymin = south$l.conf_i_calc, ymax = south$u.conf_i_calc), width = 0.
2, color = "blue4") + theme(axis.text.y = element_text(size = 7, color
= "blue4")) + theme(plot.margin = (unit(c(.001, .001, .001, .001), "cm"
))) +  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) + theme(aspect.ratio=.2) + theme(
axis.title = element_text(color="blue4")) + scale_y_continuous(limits =
c(0,0.405), breaks = c(0,0.1,0.2,0.3,0.4))

# West Plot
westPlot <- ggplot(data=west, aes(x= reorder(zone, conf_i_length), y=p.
upmover, group = 1)) + geom_point(color = "yellow4") + labs(x = "West"
, y = "Probability of Upward Mobility") + coord_flip() + geom_errorbar(
aes(ymin = west$l.conf_i_calc, ymax = west$u.conf_i_calc), width = 0.2,
color = "yellow4") + theme(axis.text.y = element_text(size = 7, color =
"yellow4")) + theme(plot.margin = (unit(c(.001, .001, .001, .001), "cm"
))) + theme(aspect.ratio=.2) +  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) + theme(axis.title = element_text
(color="yellow4")) + scale_y_continuous(limits = c(0,0.405), breaks = c
(0,0.1,0.2,0.3,0.4))

midwestPlot <- ggplot_gtable(ggplot_build(midwestPlot))
northeastPlot <- ggplot_gtable(ggplot_build(northeastPlot))
southPlot <- ggplot_gtable(ggplot_build(southPlot))
westPlot <- ggplot_gtable(ggplot_build(westPlot))

maxWidth = unit.pmax(midwestPlot$widths[2:3], northeastPlot$widths[2:3]
,southPlot$widths[2:3],westPlot$widths[2:3])

midwestPlot$widths[2:3] = maxWidth
northeastPlot$widths[2:3] = maxWidth
southPlot$widths[2:3] = maxWidth
westPlot$widths[2:3] = maxWidth

grid.newpage()
grid.arrange(midwestPlot, westPlot, northeastPlot, southPlot, ncol = 1,
left="Zone", bottom="Probability of Upward Mobility", top ="Probability
of Upward Mobility by Zone with Clopper-Pearson Confidence Intervals")
```

```
#Region Plot
```



```
regionMobility$regionsName = c("South", "West", "Northeast", "Midwest")

regionGraph <- ggplot(data=regionMobility, aes(x=reorder(regionsName, c
onf_i_length), y=phat, group = 1, color = regionsName)) + geom_point()
+
  labs(x = "Region" , y = "Probability of Upward Mobility") +
  geom_errorbar(aes(ymin = regionMobility$l.conf_i_calc, ymax = regionM
obility$u.conf_i_calc), width = 0.2) +
  coord_flip() + theme(plot.margin = (unit(c(.001, .01, .001, .01), "cm
"))) + scale_y_continuous(limits = c(0.04,0.12), breaks = c(0.04, 0.06,
0.08,0.10, 0.12)) +
  scale_color_manual("Region",values=c("green4", "red4", "yellow4", "bl
ue4")) + labs(title = "Probability of Upward Mobility by Region")

regionGraph
```


Probability of Upward Mobility by Region
Region
Midwest
Northeast
West
South
Probability of Upward Mobility
Region
Midwest
Northeast
South
West