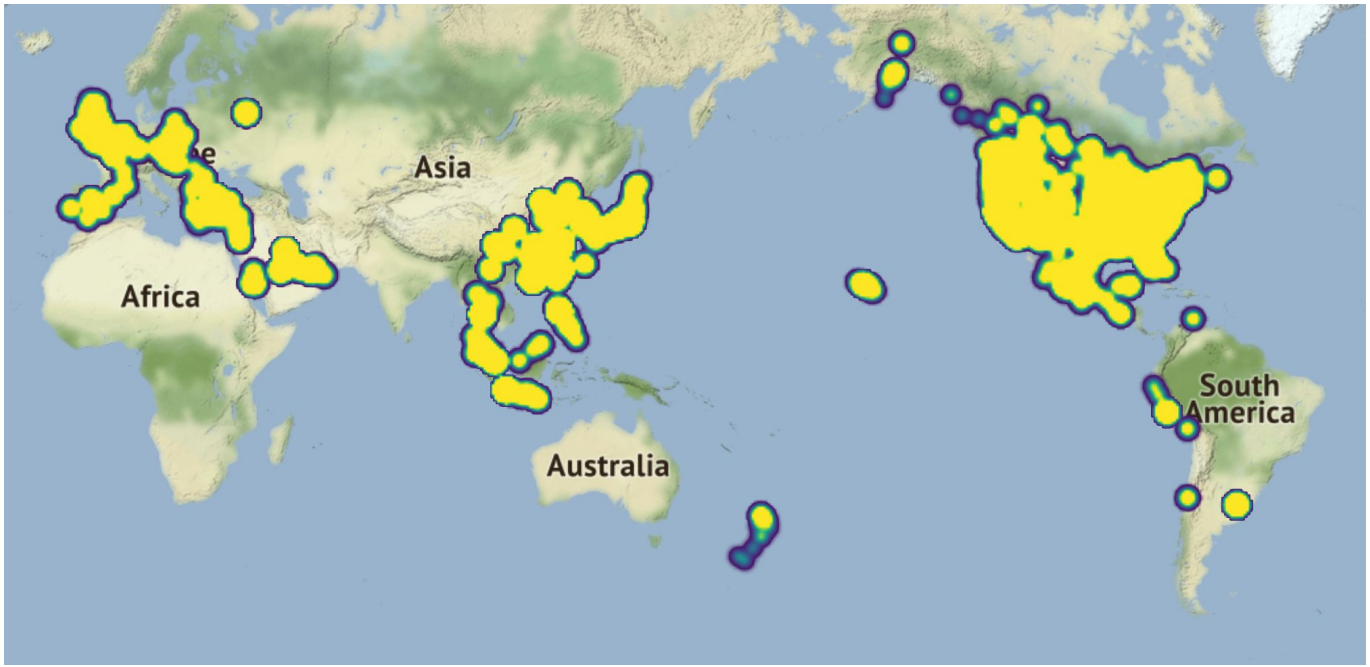
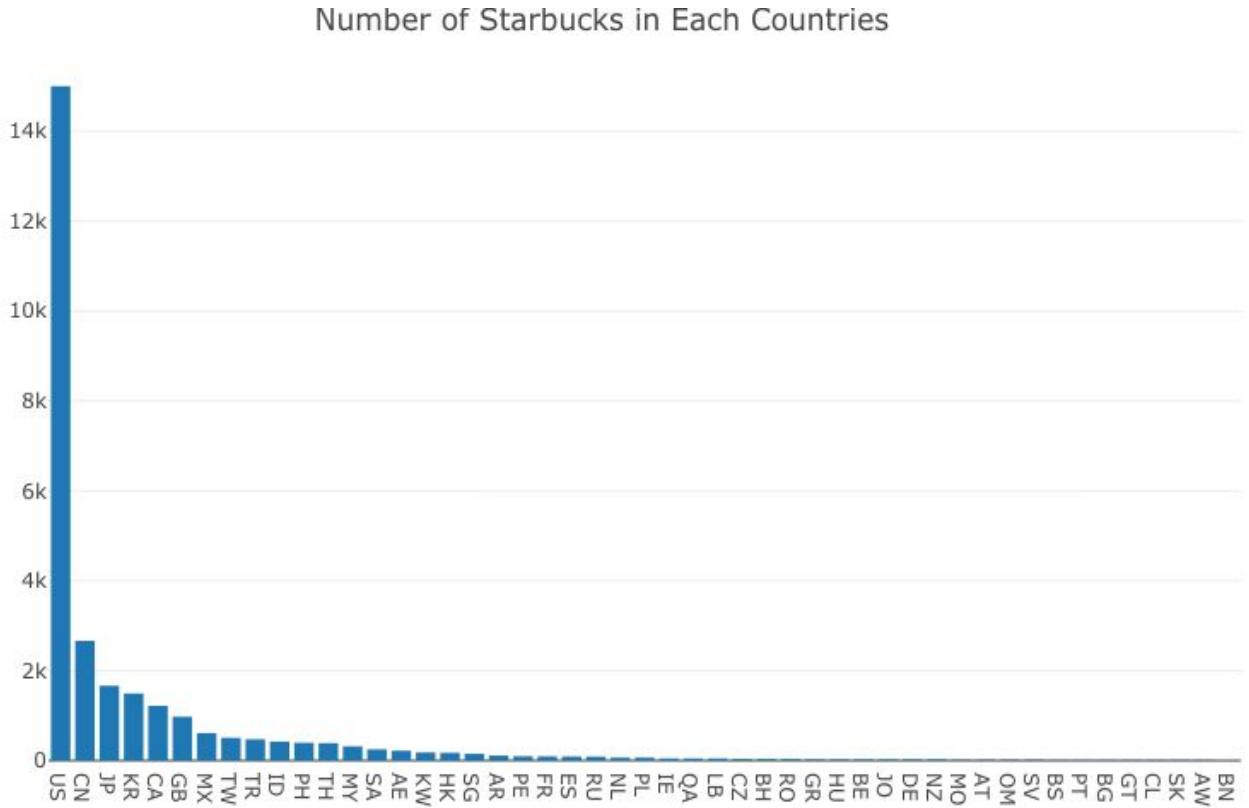
Analysis of Population on Starbucks

By Jisun Lee, Yuepeng Chen, Huiyu Yi

Overview of the World



Overview of the World



Extremely Popular in These Countries.

```
> top5<-sort(data1,decreasing = TRUE)[1:5]
> top5

   US    CN    JP    KR    CA
15003  2665  1665  1493  1219
```

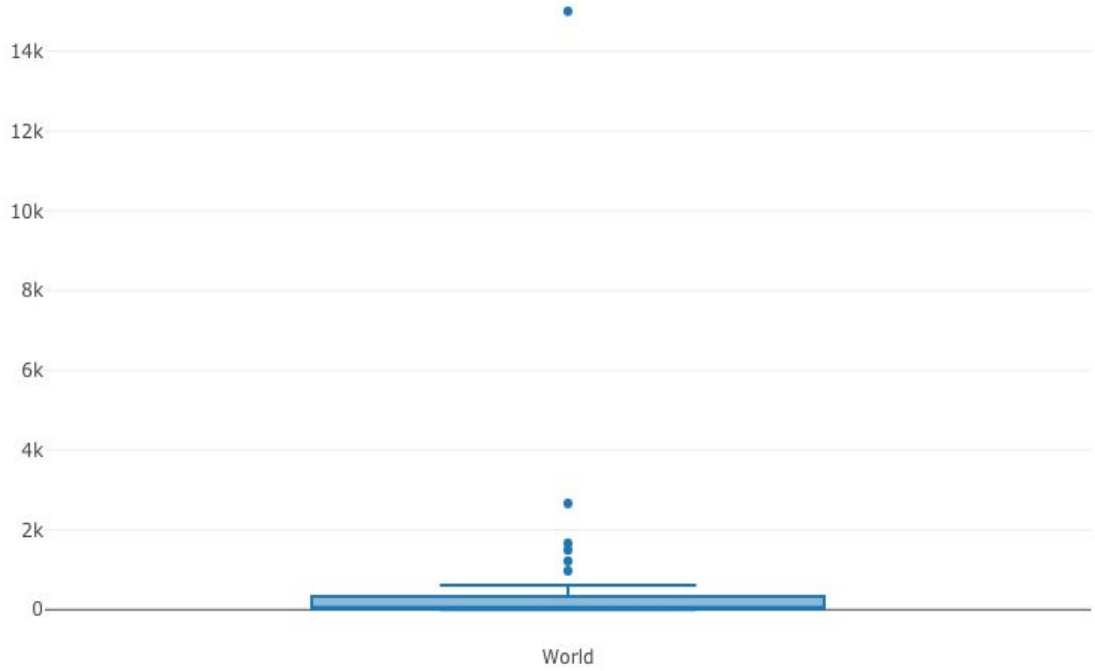
We can see that the US, China, Japan, Korea and Canada are the top 5.

US has extremely high number of Starbucks since US is the hometown of Starbucks.
For Canada, it is close to the US.

Summary of Stores Around the World

```
## [1] "The Min. number of Starbucks around the world is 1"
## [2] "The 1st Qu. number of Starbucks around the world is 26"
## [3] "The Median number of Starbucks around the world is 62"
## [4] "The Mean number of Starbucks around the world is 577.326530612245"
## [5] "The 3rd Qu. number of Starbucks around the world is 317"
## [6] "The Max. number of Starbucks around the world is 15003"
```

Find out the outliers



This box plot clearly shows the distribution of these data, and see there so many outliers and one of them is really far away.

Dealing with outliers

```
outliers <- boxplot.stats(dataforsum$Freq)$out
Out <- which(dataforsum$Freq %in% c(outliers)) #position of outliers in the dataframe
min(dataforsum[Out,]$Freq)
```

```
## [1] 975
```

To better analyze these data, we planned to divide all of the data into four categories. We find the outliers and find the minimum of them. This will be the boundary of the extremely high density category.

Dealing with Outliers - dividing into four categories

Based on the minimum of outlier, we set the range of each category as below:

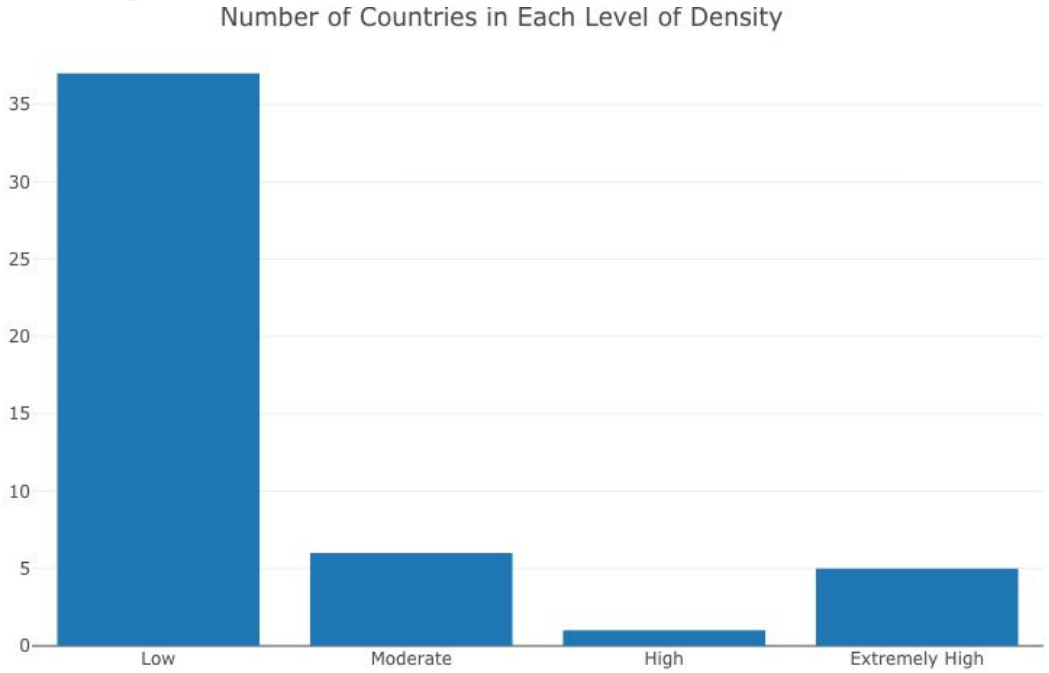
<= 375 : low

Between 375-675 : moderate

Between 675-975 : high

>= 975 : extremely high

Four Categories




```
##                  number
## low                  37
## moderate              6
## high                  1
## Extremly_high         5
```

Focusing on low density category

```
## [1] "The Min. Number of Starbucks in Low Denisity Country is 1"
## [2] "The 1st Qu. Number of Starbucks in Low Denisity Country is 17"
## [3] "The Median Number of Starbucks in Low Denisity Country is 34"
## [4] "The Mean Number of Starbucks in Low Denisity Country is 66.7567567567568"
## [5] "The 3rd Qu. Number of Starbucks in Low Denisity Country is 91"
## [6] "The Max. Number of Starbucks in Low Denisity Country is 317"
```

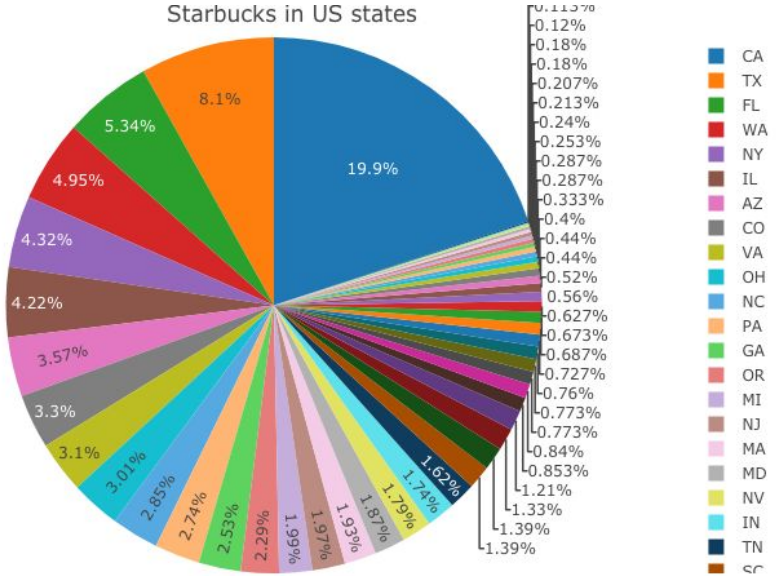
We can see that the median is 34 but the mean is about 66.

Even though they are belong to low density category, they are still huge different.

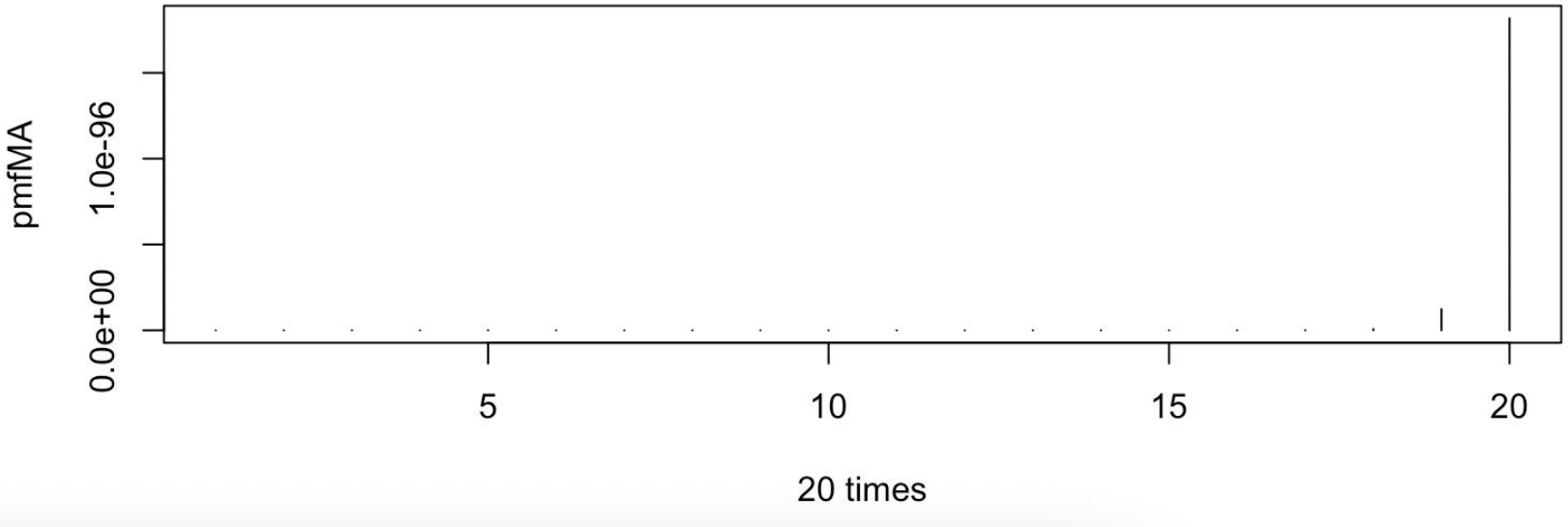
Focus on the US

Focus in the US

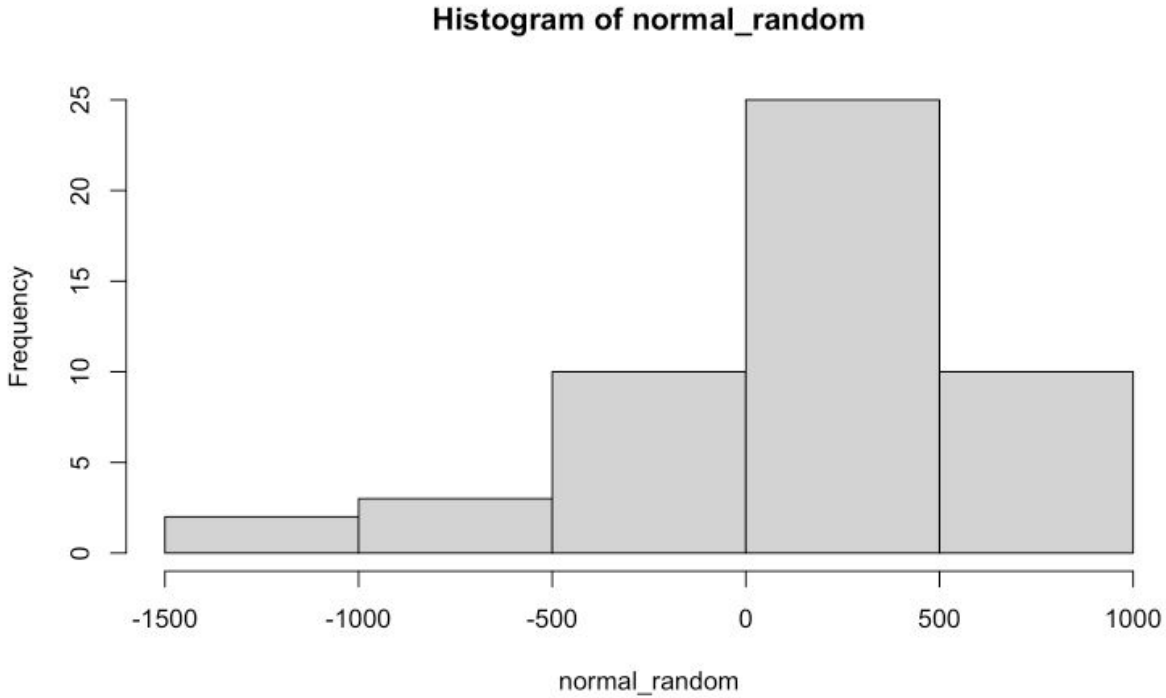
To make it more clear, we made a pie chart to show the percentage of each state.



Binomial distribution in MA



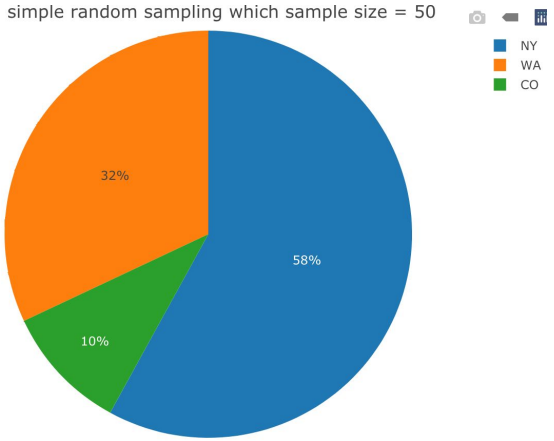
Sampling by normal distribution with sample size 50

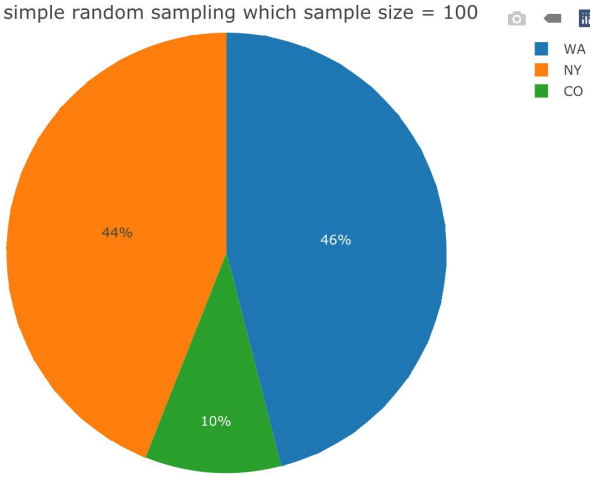


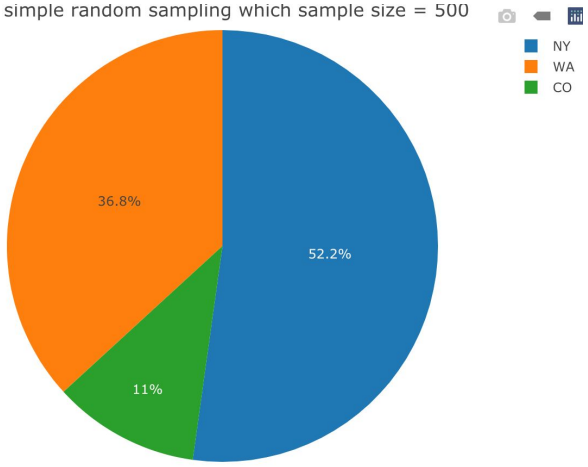

```
> mean
[1] 294.1765
> sd
[1] 456.3604
```

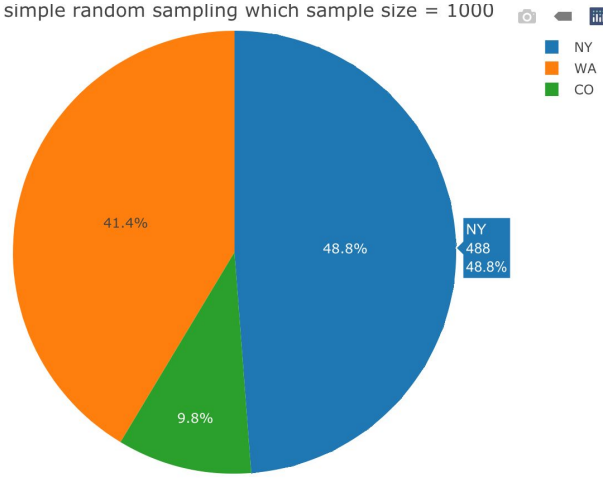
Simple random sample with different sample sizes

```
The simple random sample layout with sample size 50 is
 CO NY WA
 8 28 14
The simple random sample layout with sample size 100 is
 CO NY WA
 7 52 41
The simple random sample layout with sample size 500 is
 CO NY WA
 37 244 219
The simple random sample layout with sample size 1000 is
 CO NY WA
 80 500 420
```


simple random sampling which sample size = 50
NY
WA
CO
32%
58%
10%

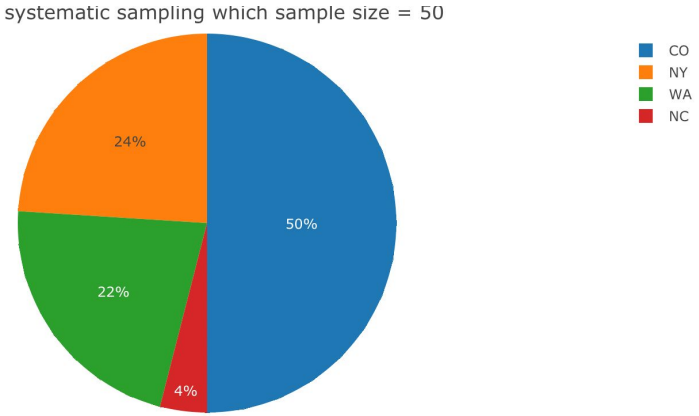

simple random sampling which sample size = 100
WA
NY
CO
44%
46%
10%

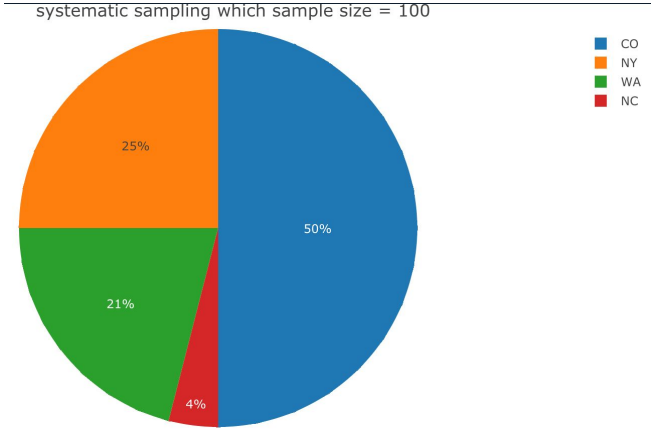

simple random sampling which sample size = 500
NY
WA
CO
36.8%
52.2%
11%

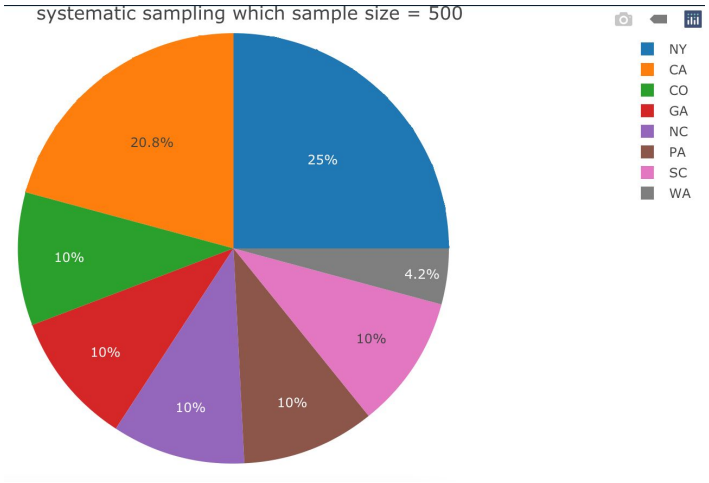

simple random sampling which sample size = 1000
NY
WA
CO
41.4%
48.8%
9.8%
NY
488
48.8%

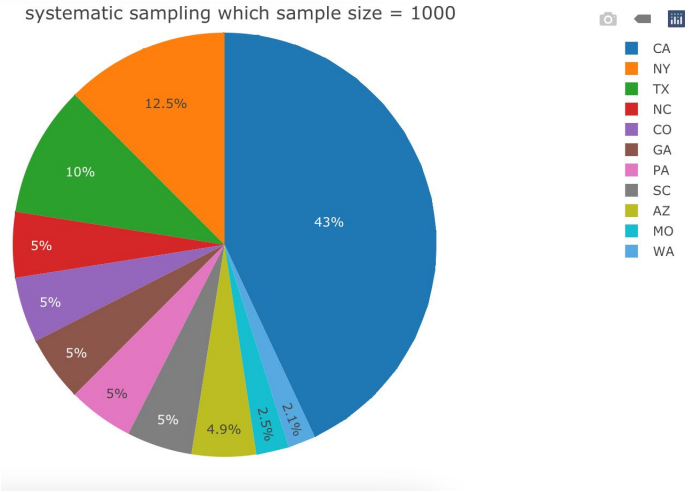
Systematic sample with different sample sizes

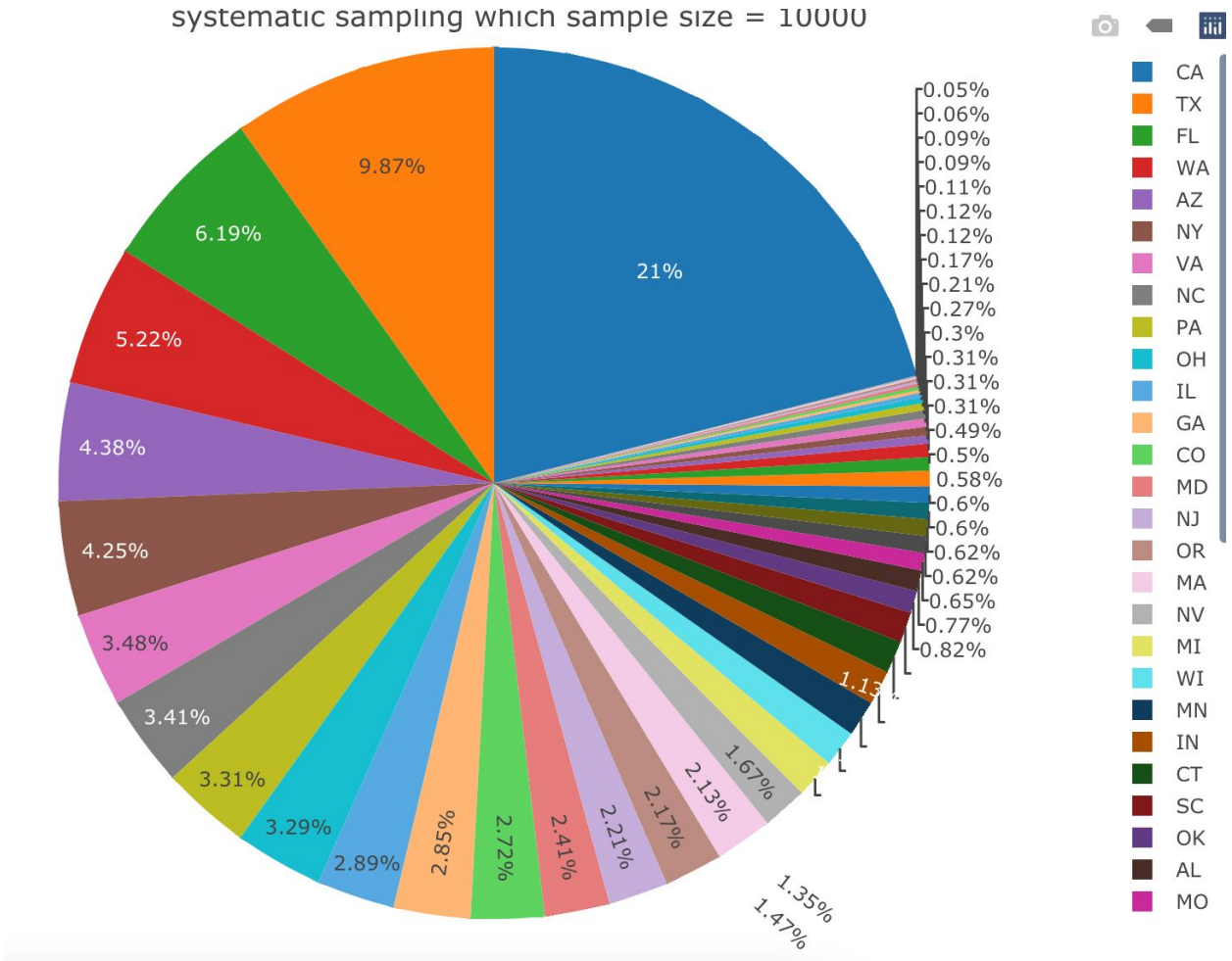
```
The systematic sample layout with sample size 50 is
 CO NC NY WA
 25 2 13 10
The systematic sample layout with sample size 100 is
 CO NC NY WA
 50 4 25 21
The systematic sample layout with sample size 500 is
 CA CO GA NC NY PA SC WA
 104 50 50 50 125 50 50 21
The systematic sample layout with sample size 1000 is
 AZ CA CO GA MO NC NY PA SC TX WA
 49 430 50 50 25 50 125 50 50 100 21
```


systematic sampling which sample size = 50
CO
NY
WA
NC
24%
50%
22%
4%


systematic sampling which sample size = 100
CO
NY
WA
NC
25%
50%
21%
4%


systematic sampling which sample size = 500
NY
CA
CO
GA
NC
PA
SC
WA
20.8%
25%
10%
4.2%
10%
10%
10%
10%


systematic sampling which sample size = 1000
CA
NY
TX
NC
CO
GA
PA
SC
AZ
MO
WA
12.5%
43%
10%
5%
5%
5%
5%
5%
4.9%
2.5%
2.1%


systematic sampling which sample size = 10000
21%
9.87%
6.19%
5.22%
4.38%
4.25%
3.48%
3.41%
3.31%
3.29%
2.89%
2.85%
2.72%
2.41%
2.21%
2.17%
2.13%
1.67%
1.35%
1.47%
0.05%
0.06%
0.09%
0.09%
0.11%
0.12%
0.12%
0.17%
0.21%
0.27%
0.3%
0.31%
0.31%
0.31%
0.49%
0.5%
0.58%
0.6%
0.6%
0.62%
0.62%
0.65%
0.77%
0.82%
CA
TX
FL
WA
AZ
NY
VA
NC
PA
OH
IL
GA
CO
MD
NJ
OR
MA
NV
MI
WI
MN
IN
CT
SC
OK
AL
MO

Visualization and conclusion

The influence of population on Starbucks is very complicated.
In the US scale, although California has their own Coffee brand, most Starbucks in the US are also in California.
For world scale, it is really interesting that Starbucks are so popular in East Asian countries. The reason may combine economic, culture, and population elements.

Thank you