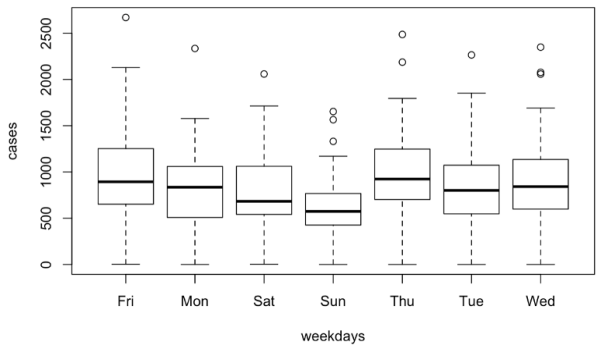
Donglai Xu

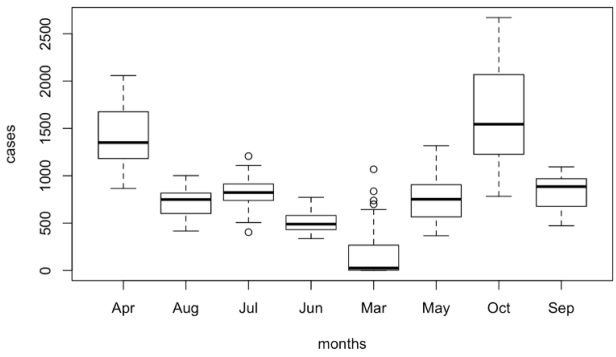
Zhi Li

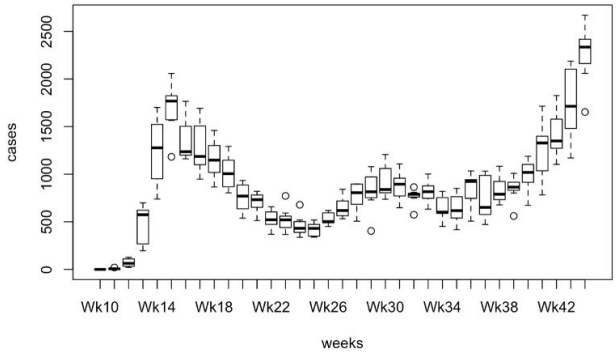
Final Project - Group 3

First Sampling Method: Stratified Random Sampling

Our population is the state of Pennsylvania from March 1st, 2020- October 31st, 2020. We are curious about the population mean COVID-19 cases per day, by calculation it is 859.404. Then we use boxplot to identify strata. There are three boxplots with stratification by weekdays, months and weeks below. We finally decide to stratify by months, since there are too many strata by weeks and it is not homogeneous if stratified by weekdays.







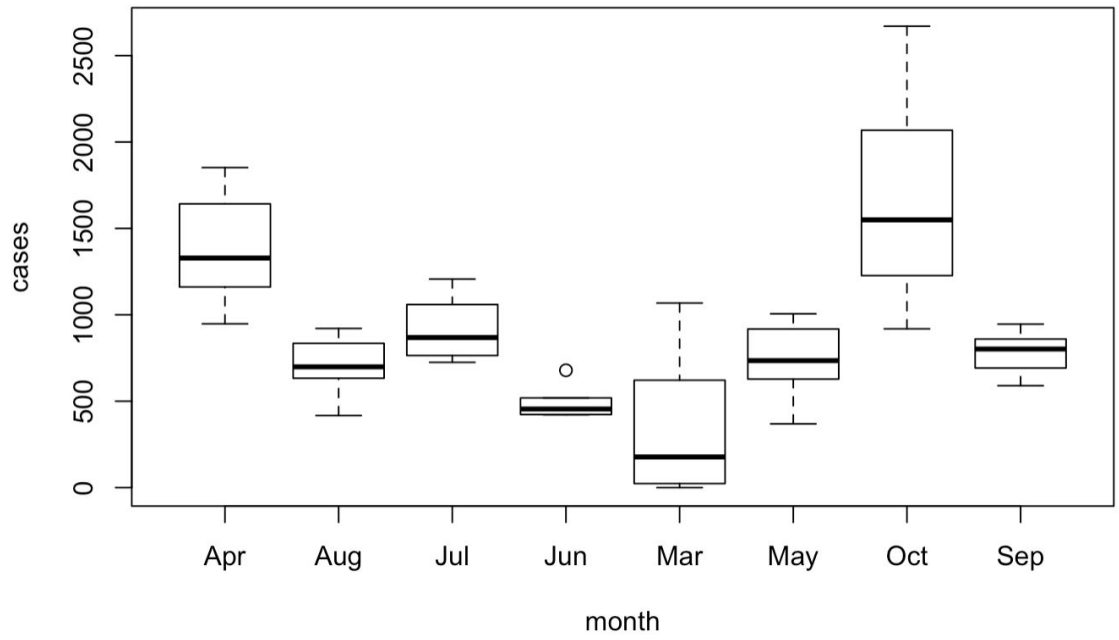
We decided to use a bound of 30 cases to estimate the sample size. For the estimate of standard deviation of each strata, we use the range divided by 4. After using Neyman Allocation, we have a total sample size of 112. We also obtain sample size for each strata by calculating the proportion of each strata.

month <fctr>	samplesize <dbl>
March	16
April	18
May	14
June	6
July	12
August	9
September	9
October	28

Stratified.Sampling <fctr>	Range <dbl>	sd <dbl>	var <dbl>	n <dbl>
March	1068	267.00	71289.00	31
April	1192	298.00	88804.00	30
May	950	237.50	56406.25	31
June	435	108.75	11826.56	30
July	803	200.75	40300.56	31
August	585	146.25	21389.06	31
September	622	155.50	24180.25	30
October	1887	471.75	222548.06	31

After obtaining the sample size for each strata, we randomly select samples from each strata according to the sample size. Then, we find the mean and the variance of the sample from each strata. By using the formula, we have the estimate for the population mean is 873.043 and the estimate for the variance of population mean is 210.545. Therefore, we have a bound of 29.02, which satisfies our goal of a bound smaller than 30 cases. Also, we have a confidence

interval from 844.023 to 902.063. The true population mean is 859.404, which is in our 95% confidence interval.

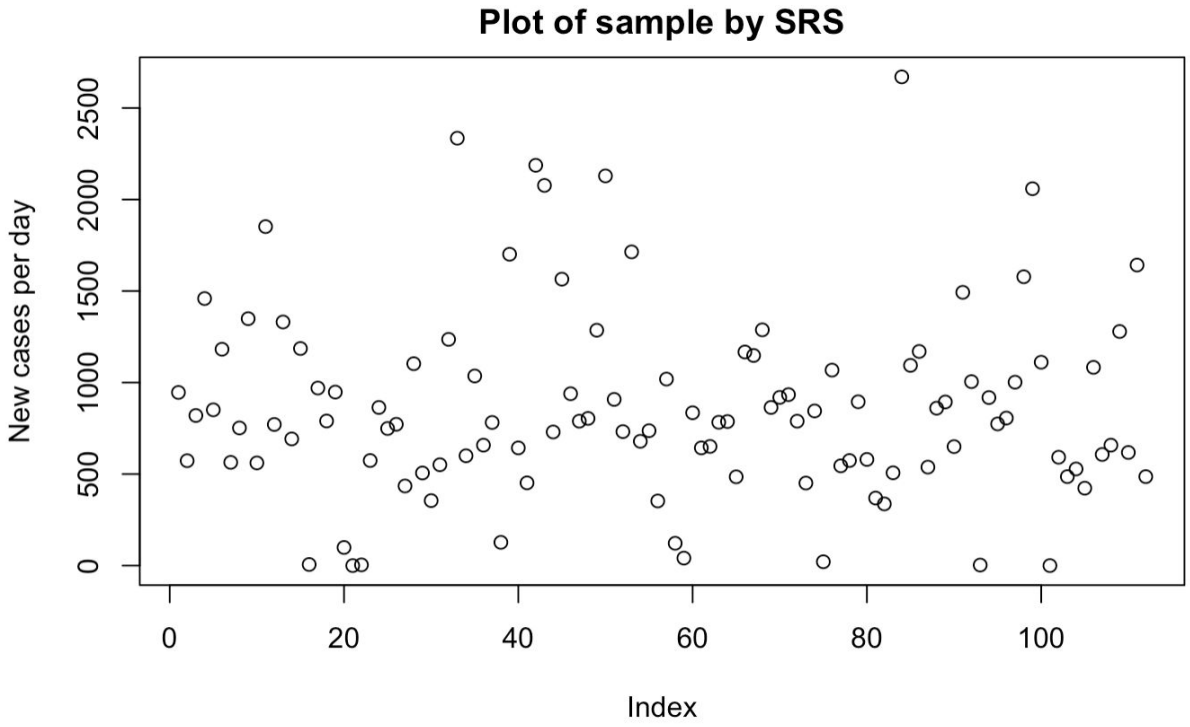


Stratified	n	mean	var	Ni
March	16	329.813	124375.1	31
April	18	1379.222	77514.54	30
May	14	741.857	36204.75	31
June	6	492	9654	30
July	12	916.417	28394.27	31
August	9	710.556	24209.03	31
Septembet	9	784.556	13347.78	30
October	28	1631.107	237600.8	31

Second Sampling Method: SRS

For the second sampling method, we decide to use simple random sampling. We have 245 data points and we randomly sample 112 data points from them. From the plot below, the points are randomly distributed. The mean number of new cases of the sample is 861.679 with a sample variance of 264287. By using the formula of estimation in SRS, we obtain a bound of

71.582, so the confidence interval is from 790.097 to 933.261. The true population mean is 859.404, which falls within our 95 percent confidence interval.

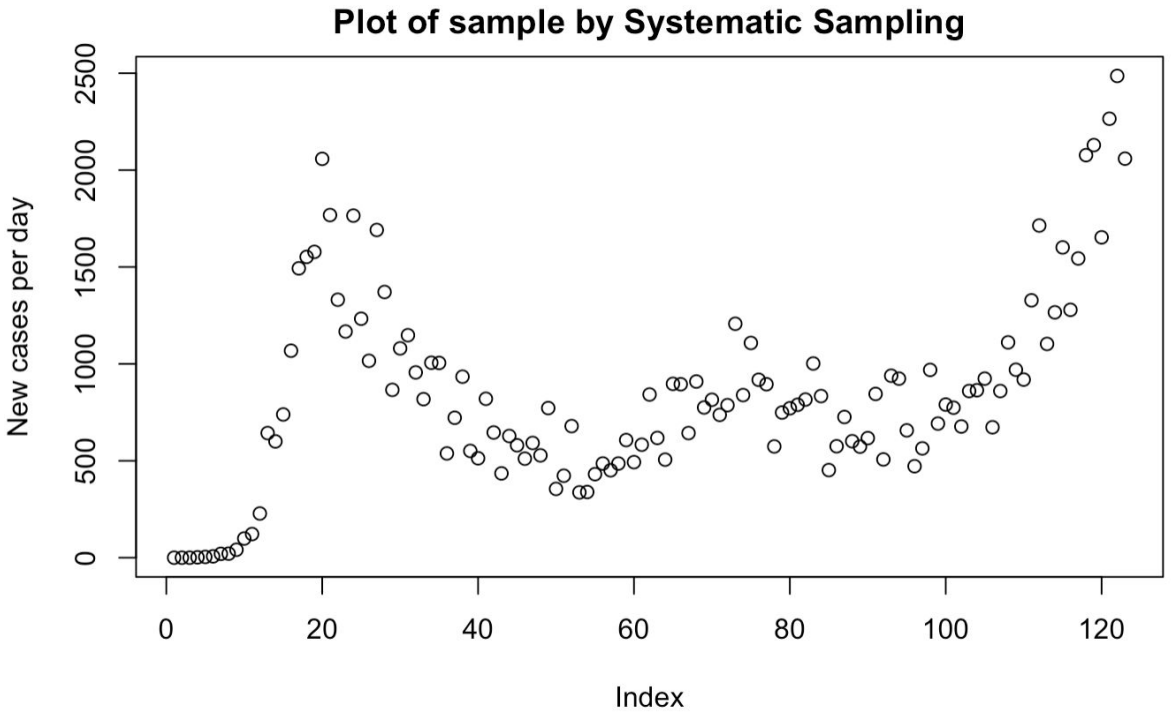


N_srs <dbl>	n_srs <dbl>	mean_srs <dbl>	var_srs <dbl>
245	112	861.6786	264287

Third Sampling Method: Systematic Sampling

For the third sampling method, we are using systematic sampling. We have 245 data points in our population and we want to get a sample size of 112. The value of k is 2, by using 245 divided by 112. Then, we randomly pick one element from the first two elements and select every second element thereafter in the population. Finally, since we randomly pick the first element, we get a sample size of 123. Based on the graph we got, the dots are not randomly distributed and it shows a trend, so we need to consider the successive differences in this situation. By using successive differences, the sample mean of the sample is calculated and the

estimate of population mean is 844.724 with a bound of 23.259. Therefore, the 95 percent confidence interval is 821.465 to 867.983. The true population mean is 859.404, which falls within our 95 percent confidence interval.



N_sys <dbl>	n_sys <dbl>	mean_sys <dbl>	sum_di_square <dbl>
245	123	844.7236	8151349

Discussion

After estimating the population mean through the three sampling methods above, all the estimated confidence intervals contain the true value of the population mean. From the materials we learned from class, stratified random sampling often needs less cost for obtaining the same precision as SRS. In our case, the bound calculated by stratified random sampling is smaller than by SRS, which does show that stratified random sampling is better than SRS with the same sample size. Also, from the lectures,

systematic sampling is often at least as precise as SRS and it is easy to perform. In our cases, SRS is less precise than systematic sampling, given the bound calculated by systematic sampling is the smallest.

Moreover, for an ordered population, SRS will overestimate the population variance and bound, which can be proved by the bound calculated by SRS is hugely larger than the other two.

Actually, before we perform the systematic sampling, we thought there may be an increasing pattern from March to October for the number of new cases per day, since the new cases per day should increase as more people are infected and the virus is more likely to live with a lower temperature. However, from the plot, it does not show a pattern like we think of. The number of new cases per day first had a surge from March to April, then slumped from April to June but increased again in August, followed by a small decrease in September and a surge in October. It is reasonable for the number of new cases per day to fluctuate a little bit and we think the surge in October because of the decrease of temperature. However, the surge during March followed by a slump does not seem reasonable for us. The surge may be explained by that it was the preliminary stage of the pandemic, so the difference is large. We may need further investigation to figure out the reason behind it.