**STAT 466 DUE December 11, 2020**

**Final Project**

Below are the steps you should follow for your STAT 466 final project. Your project, including all components below, should be uploaded to Canvas as a PowerPoint slide deck. Also be prepared to show your final 3 sample data sets.

In Canvas > Files > FinalProject, you will find available data sets that you can use.

Data source site: <https://data.pa.gov/Health/COVID-19-Aggregate-Cases-Current-Daily-County-Heal/j72v-r42c>

Or, you may find your own COVID-19 data if you wish (e.g. you could get data from the PSU Dashboard for University Park and other campuses). Please site your source.

1. Define your population of interest. Examples include:

* State of Pennsylvania from March 20, 2020 – November 20, 2020
* Centre County from June 2020 – August 2020
* Etc.

1. Define and compute your parameter of interest. Examples include:

* Population mean COVID-19 cases per day (= \_\_\_\_)
* Population total COVID-19 cases (= \_\_\_\_)
* Population proportion of residents testing positive for COVID-19 per day (*p* = \_\_\_\_)
* Etc.

1. Compute the sample size for a stratified random sample (e.g. using Neyman allocation). The data sets include some potential variables to explore for stratification or you can create your own (homogeneous) strata.

YOU CHOOSE a bound that is reasonable – the bound should not be so large that it is uninformative, nor should it be so small that it requires a large sample size relative to *N*.

To estimate the standard deviation for each stratum, you can take a small initial sample. Or, you may use a reasonable range to estimate standard deviation (s ≈ range/4).

1. Estimate the parameter of interest using a stratified random sample:
   1. Take a stratified random sample using the sample size computed above.
   2. Graphically display the data (e.g. boxplots by strata).
   3. Use your stratified random sample to compute the parameter estimate with a bound.
   4. Does the true value for the parameter of interest (per part #2) fall within the ~95% confidence interval?
2. Estimate the parameter of interest using a second sampling method we learned (e.g. SRS, systematic sampling, cluster sampling or two-stage cluster sampling)
   1. Take a sample using your sampling method of choice, using the same sample size from part #3 – hold *n* constant for this entire project.
   2. Graphically display the data.
   3. Use your sample to compute the parameter estimate with a bound.
   4. Does the true value for the parameter of interest (per part #2) fall within the ~95% confidence interval?
3. Estimate the parameter of interest using a third sampling method we learned (e.g. SRS, systematic sampling, cluster sampling or two-stage cluster sampling)
   1. Take a sample using your sampling method of choice, using the same sample size from part #3 – hold *n* constant for this entire project.
   2. Graphically display the data.
   3. Use your sample to compute the parameter estimate with a bound.
   4. Does the true value for the parameter of interest (per part #2) fall within the ~95% confidence interval?
4. Write some final remarks summarizing key aspects of your analysis. For example, did all three confidence intervals capture the true value of the parameter (where there is a ~5% chance each interval did not)? What was interesting to you? What parts of this project did you find most valuable?

NOTES:

* At the end of this project, you will have 3 SAMPLE DATA SETS and 3 PARAMETER ESTIMATES with BOUNDS.
* Choose your strata (and clusters, if you use cluster sampling) wisely.
* If you use systematic sampling, consider if successive differences should be used or not.
* There will be milestones during the last two weeks of class, where you will need to share various parts of your project, per the details provided above.
* If you worked in a group, you may be asked to provide an evaluation for your group members.