Assignment #3 for Public Health Surveillance

Due by **5pm Mar 10 (UPDATE:** WE WILL ACCEPT SUBMISSIONS UNTIL END OF THE DAY MARCH 27 BUT WILL NOT FIELD QUESTIONS ON THE ASSIGNMENT AFTER MARCH 10).

Submit via Canvas. **You should work on this assignment alone and should not discuss with your classmates.** Feel free to consult the internet and class notes. Your answers should be brief (1-2 sentences per question) and should be typed. Jiye will answer questions about material covered in class in office hours but will not provide assistance with the assignment itself. Clarifying questions can be emailed to Jiye (jiye.kwon@yale.edu), and Dan will respond to the whole class via Canvas. Jiye will not respond to emails sent on the weekend or immediately before the assignment is due. This is graded out of 100 points.

Part 1:

For the following proposed analyses, which SATSCAN version(s) (ie Poisson, Bernoulli, space-time permutation) and analysis type (spatial, temporal, space-time) would you choose. Briefly explain why.

* 1. Reports of foodborne illness, with date of onset and census tract of residence (a census tract is like a ZIP code but covers a smaller area)
  2. Childhood leukemia, with year of diagnosis, and ZIP code
  3. Syndromic surveillance for respiratory conditions in emergency departments in NYC, with a goal of detecting spikes in cases related to novel coronavirus.
  4. Legionellosis cases in Connecticut with home address and date of testing
  5. Bacterial infections, with information on resistance profile, date of testing, ZIP of residence and an interest in detecting pockets of resistance

Part 2: As an analyst working with the Connecticut Department of Health, you have been asked to identify areas that have an unusually high incidence of respiratory infections in the state that might be associated with the novel coronavirus. The data run from February 27, 2020 to March 27, 2020. You have data on the number of respiratory hospitalizations each day in each ZIP code. These data can be found in the file ‘ct\_resp.csv’ (disclaimer: these are simulated data). date=date (m/d/Y); zip=ZIP code of residence; lat=latitude of the center of the ZIP, long=longitude of the centroid of the ZIP; resp=number of hospitalizations due to a respiratory cause in the ZIP code for that day, diar=number of hospitalizations for diarrhea for that day; pop=population size for the ZIP; hh.size=household size in the ZIP. When running SATSCAN, make sure to check the box “HTML file for Google Map” on the results tab to automatically generate a map

1. Run a purely spatial Poisson analysis, adjusted for population size
   1. How many clusters do you detect?
   2. What are some potential explanations for why there are clusters of respiratory infections in these census tracts? What does it mean statistically? if you find a cluster?
   3. Paste in a screenshot of your map from your HTML file
2. Run an analysis where the cases are hospitalization due to a respiratory cause, and the control is hospitalization due to diarrheal disease (note: select ‘Bernoulli’ under the probability model on the analysis tab)
   1. Where are the most likely clusters, and how unusual are they?
   2. Why do some clusters that were found in the analysis adjusted for household size not appear here?
   3. Paste in a screenshot of your map from your HTML file
3. Run a space-time permutation analysis, using the number of hospitalization due to a respiratory cause as the case file
   1. Where (what city) and when (what dates) did the most significant cluster occur?
   2. How unusual is the most likely cluster (e.g. how many more cases do we see than expected?)
   3. How often would we expect to see a cluster of this size in similar data?
   4. Paste in a screenshot of your map from your HTML file
4. Run a Poisson regression in r, controlling for household size and using population size as an offset term. Use the predicted values form this regression as your ‘population’ estimate (paste your code for partial credit)
   1. How many clusters do you detect now, and where are they?
   2. How strong are the most likely clusters?
   3. Explain why the analysis here yields a different result from the analysis where we only adjust for population size (question 3)?
   4. Paste in a screenshot of your map from your HTML file
5. Based on the analyses you have run, which of the clusters will you prioritize for further investigation, and why?
6. What are the most likely explanations for why we see the space-time cluster that we see?

Part 3:

1. We have the number of diarrheal disease cases for 5 weeks from 5 different ZIP codes (Region 1-5), and you want to detect unusual activity using a space-time permutation test. For the data in the spreadsheet diarrheal\_disease.xlsx:
2. calculate the expected number of cases in each week and each ZIP code, and calculate a rate ratio for each week and ZIP code. (paste in your table)
3. Which week/ZIP code has the most unusually high activity?
4. Optional Bonus (+5 points): Using a permutation test, determine how likely it is that the entry (week/ZIP-code) with the largest rate ratio occurred by chance (wat is the P-value?).