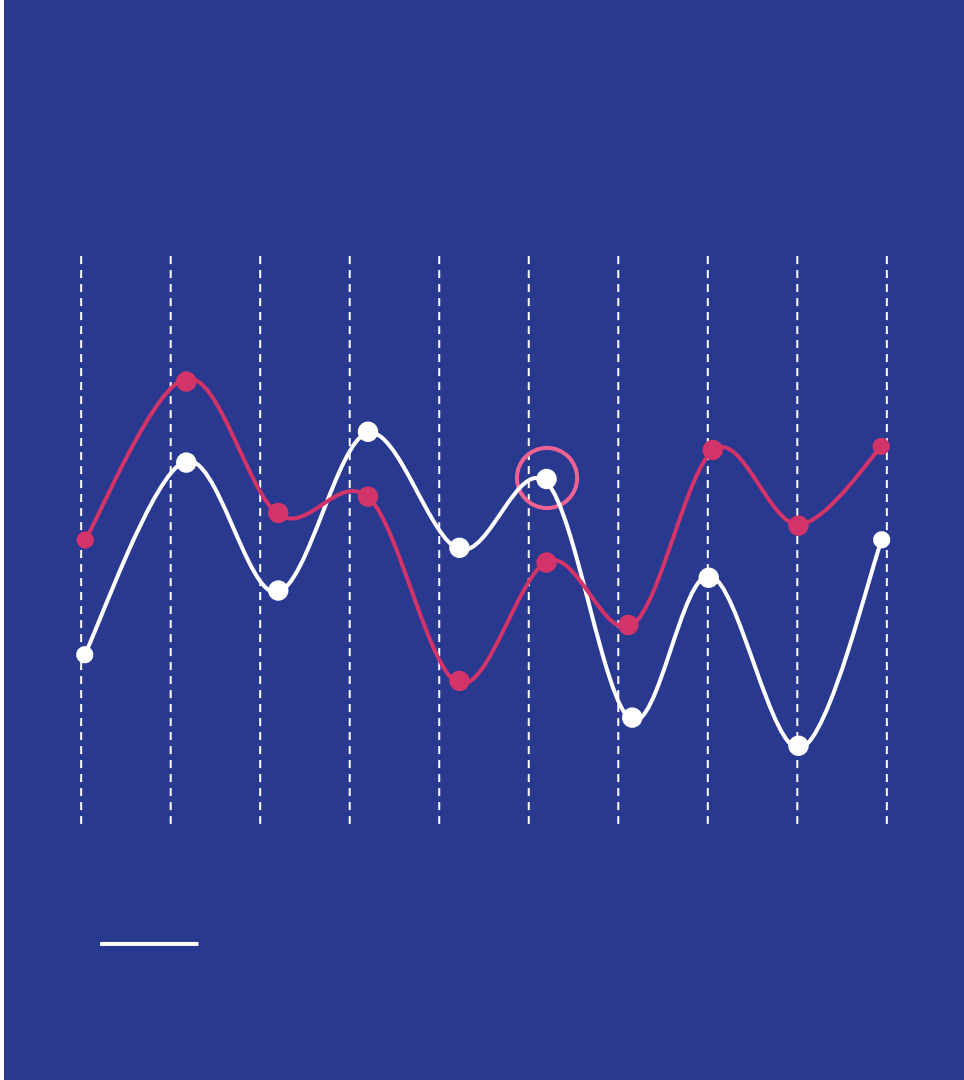
MTH5411 Final Project Presentation

Group 1 - New Jersey

Pascal Dao
Tarique Alam
Dec 2, 2020



Outline

- Intro
- Task 1
  - Method
  - Results
- Density Estimation with KDE
  - Method
  - Results
- Task 2
  - Method
  - Results
- Task 3
  - Method
  - Results

Intro

Covid-19 is an abbreviation for the infamous coronavirus disease known as SARS-CoV-2 that has swept the world in the past year killing almost 1.5 million people worldwide at the time of this report.

The focus of this project was to collect existing data of number of cases and fatalities in a particular state and counties of that state assigned to us in USA. Following that we processed the data and found out the temporal distribution of cases and deaths in our assigned state (New Jersey).

Next we found the correlations between total cases in counties of New Jersey (till October 2020) and student infections in and average income of the counties. A similar correlation was calculated but between cases and average county income. Added to that, correlation between number of daily cases and daily deaths for each county was also calculated. We also computed the correlation between the total cases and average daily temperatures of the respective counties.

Finally the correlation between total cases(total number at risk) and total number of risks recovered and fatalities have been calculated.

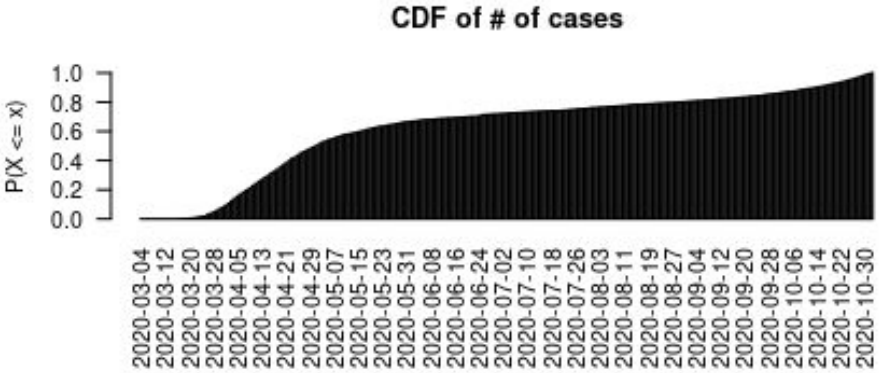
Task 1- Methods

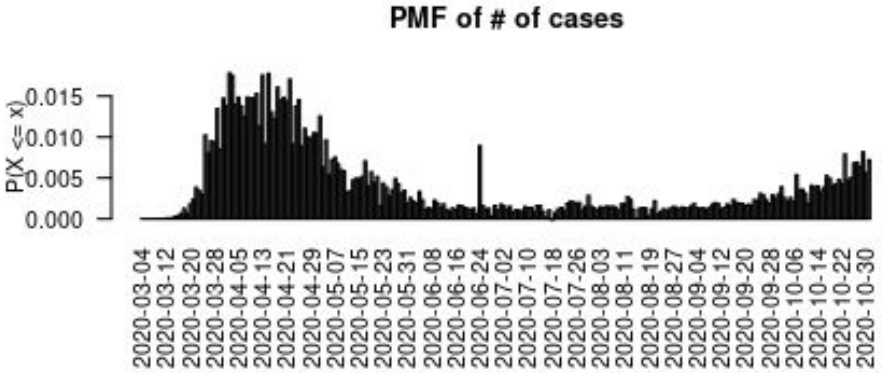
Cases/Deaths

- **CDF**
  - Time is discrete RV (dates from March 3rd to Oct 30th 2020)
  - Value for each date is cases/deaths on that day over sum of number of cases/deaths
  - $P(X \leq x)$ = deaths[x]÷sum_of_deaths
  - Data is naturally a CDF
- **PMF**
  - Time is discrete RV (dates from March 3rd to Oct 30th 2020)
  - Value for each date is cases on that day minus cases on previous day over sum of number of cases/deaths
  - $P(X=x)$ = (cases[x]-cases[x-1])÷sum_of_cases

Task 1 - Results

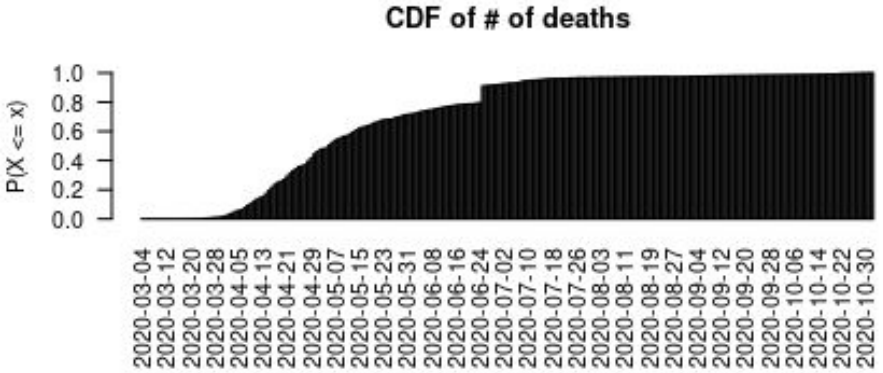
Cases

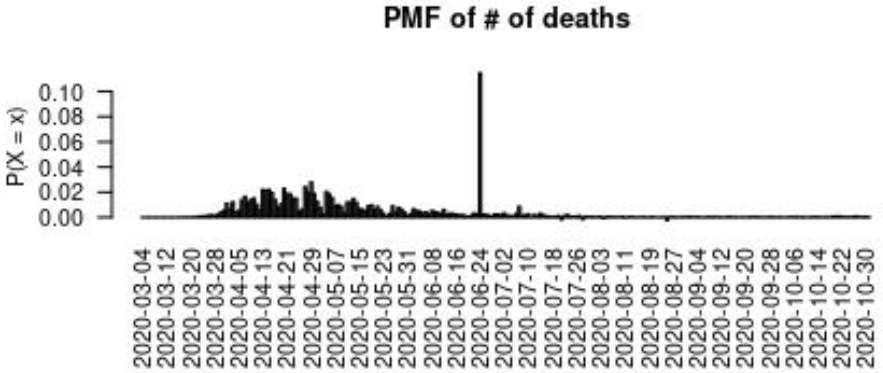




Task 1 - Results

Deaths





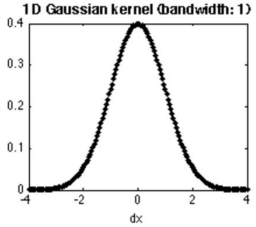
Estimating Density Distribution - Methods

- Kernel Density Estimator

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right).$$

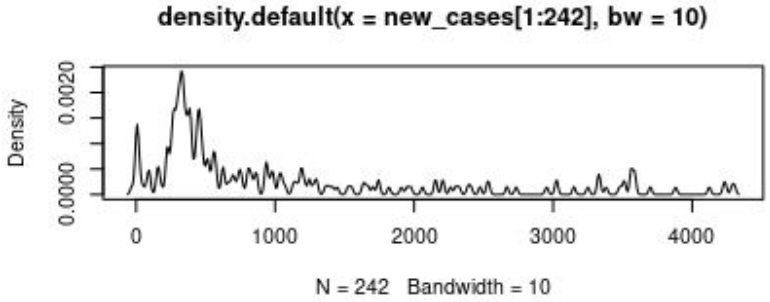
- Gaussian Kernel

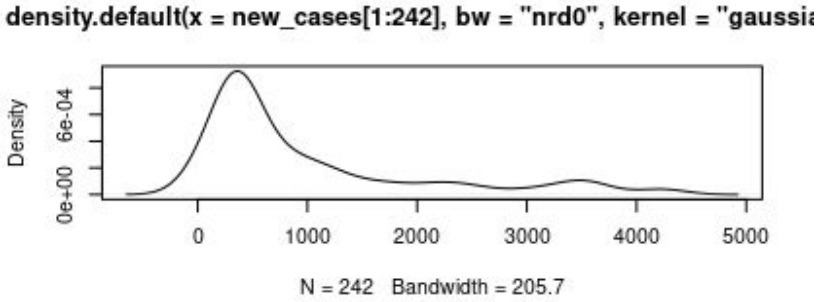
$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

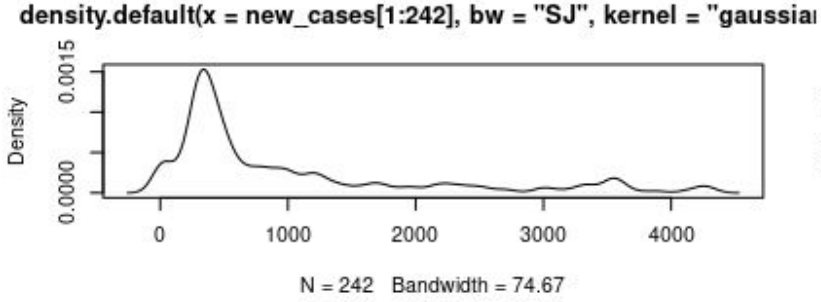


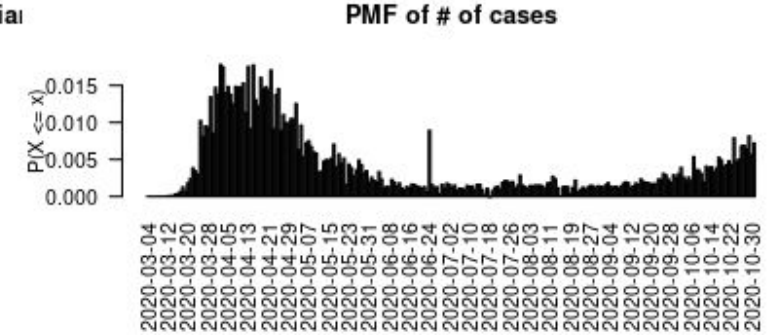
- Bandwidth computed using Silverman Rule-of-Thumb, and then Sheater Jones

Estimating Density Distribution - Results









Estimating Density Distribution - Results (SJ, bw = 74.67)

Between Jan 21, 2020 & Oct 31, 2020

- 1st Qu (25th percentile)
  - Day 60
  - May 2nd
- Mean
  - Day 121
  - Jul 2nd
- 3rd Qu (75th percentile)
  - Day 181
  - Aug 31st

Task 2 - Methods

- Cases/Deaths-Students corr
  - From counties dataset get cases by end Oct & fips of county
  - Match fips in ElementrarySecondary.csv to get num of students in that county
  - 21 Counties in NJ
- Cases/Deaths-Income corr
  - Income data from nj.gov [1]
    - 2010 Census
  - Mean Nonfamily Income
- Optional result - Cases-avg_temp corr
  - Temp data from ncdc.noaa.gov [2]
  - Data from Jan to Sept 2020
  - Avg per county
- Cases-Death corr
  - 21 Correlation values
  - From Jan 21, 2020 to Oct 31, 2020

Task 2 - Results

- Cases-Students corr
  - 0.9352122
- Cases-Income corr
  - 0.1185551
- Deaths-Students correlation
  - 0.9312863
- Deaths-Income correlation
  - 0.1881319
- Optional result - Cases-avg_temp corr
  - 0.1361953
- Cases-Deaths corr table

County	CDcor
1: Atlantic	0.290668
2: Bergen	0.498078
3: Burlington	0.28948
4: Camden	0.369828
5: Cape May	0.258436
6: Cumberland	0.467296
7: Essex	0.548938
8: Gloucester	0.159115
9: Hudson	0.541692
10: Hunterdon	0.42183
11: Mercer	0.617317
12: Middlesex	0.417866
13: Monmouth	0.350254
14: Morris	0.391314
15: Ocean	0.303778
16: Passaic	0.502623
17: Salem	0.284065
18: Somerset	0.502371
19: Sussex	0.469669
20: Union	0.431644
21: Warren	0.517914

Task 3 - Methods

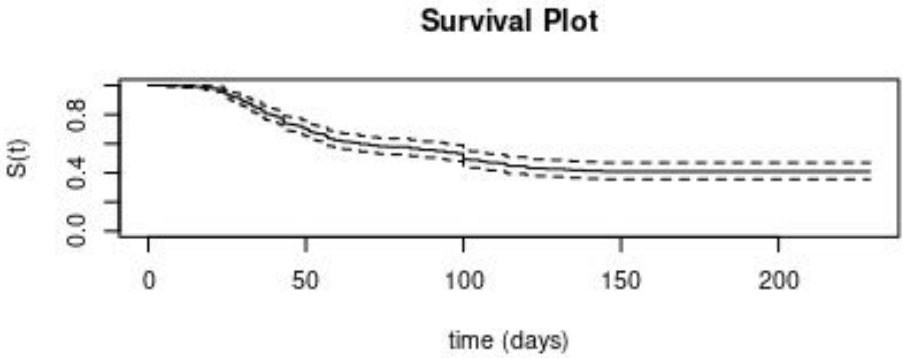
Survival & Hazard Analysis

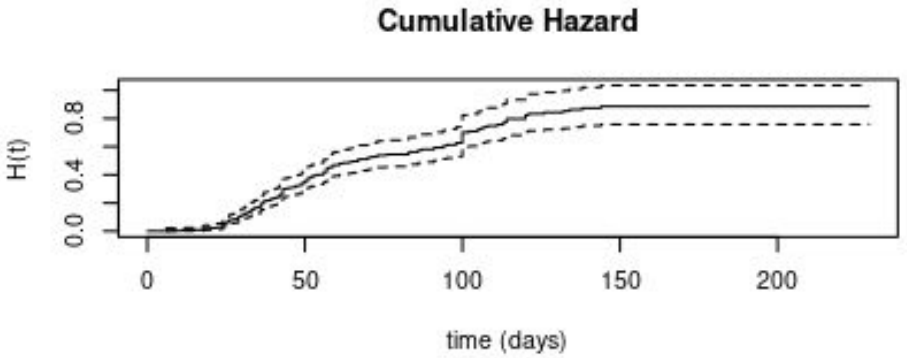
- Used survfit from survival package
- Events had to be 0(censored) or 1(died)
- Stretched time and deaths vector
- Cumulative hazard plot obtained by enabling cumhaz option on plot function

days	deaths
1	0
2	0
3	0
4	0
5	0
6	1
7	0
8	0
9	0
10	1
11	0
12	1
13	0
14	2

days	deaths
1	0
2	0
3	0
4	0
5	0
6	1
7	0
8	0
9	0
10	1
11	0
12	1
13	0
14	1
14	1

Task 3 - Results





Conclusion

In conclusion, data gathered from our calculations suggest an increasing rate of new cases towards end of october with a gradual increase in fatalities as well. The data shows Essex county being the most hard hit in number of cases and fatalities with both values increasing as of October 2020.

A stark observation, was the correlation found was that the higher the number of students, the higher the number of cases and deaths in the respective counties for example in the case of Essex county.

Apart from the observations above there was no strong evidence to support any other strong correlation between income, county temperature and cases, fatalities. Interestingly, there was a positive correlation between income and number of cases. Intuitively it can be assumed that for higher income counties the cases would be lower. However, the data do not support this intuition.

Citations

1. Wnyakanga. (n.d.). Department of Labor and Workforce Development: 2010 American Community Survey Income Page. Retrieved from https://www.nj.gov/labor/lpa/industry/incpov/2010income.html

2. NCEI.Monitoring.info@noaa.gov. (n.d.). Climate at a Glance. Retrieved from https://www.ncdc.noaa.gov/cag/county/mapping/28/tavg/202010/9/value