Pandemic Flu Spread Project Report

- for ISYE-6644

Group Number - 277
Members - Samuel Wade Wang

Contents

Abstract

In this project, I consider the problem of finding the expectation / variance of people being contaminated each day / date of flu-end in a flu outbreak under some modeling assumptions. Under such assumptions, we can associate a Markov chain to this problem, and use the tools of Markov chains to solve this problem mathematically. The Markov chain formulation also enabled us to design simulations.
In this report, I will first do the maths in the first chapter, then summarize the experiments conducted with my Python package "flusim".

1 Problem Definition and the Maths

In this section, I will define of our problem of interest, lay down some mathematical notions, as well as mentioning some target of interests. After that, I will establish a math method for calculating them.

Before we proceed, let us set up some mathematical notations:

Notation 1.0.1. *We define:*

- $[n]$ *to be the set* $\{0, 1, \ldots, n\}$ *for any* $n \in \mathbb{N}$ *(here we use the convention* $0 \in \mathbb{N}$*).*
- $[n]_k$ *to be the set of k-subsets of* $[n]$ *for* $k \in [n]$*; that is,* $[n]_k = \{K \subseteq [n] : |K| = k\}$.
- *The Kronecker delta symbol* $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$
- *A generalized symbol* $\delta_{iK} = \begin{cases} 1 & \text{if } i \in K \\ 0 & \text{if } i \notin K \end{cases}$ *for* $i \in [n], k \in [n], K \in [n]_k$.
- *The Bernoulli random variable* $\mathbf{I}_p$ *for given* $p \in [0, 1]$*, with pmf* $\mathbf{I}_p(1) = 1 - \mathbf{I}_p(0) = p$.
- *The random variable* $\mathbf{B}_{n,p}$ *with distribution given by sum of n-many i.i.d.* $\mathbf{I}_p$*, for* $n \in \mathbb{N}, p \in [0, 1]$.

1.1 Problem Definition

Suppose there is a flu outbreak within $n + 1$ people - call these people $\{S_i\}_{i\in[n]}$. Suppose also that the flu stays on some one for a fixed period after infection - say for ℓ days. We are interested in the expected number of people having the flu (here we call them *ill*; those that are not victims are *healthy*) for each day. The spread of disease is assumed to follow the following:

Assumption 1.1.1. *We make the following assumptions:*

- ***(Initial Condition)*** *There are no victims before day* 0*, and that there is only one victim* S_0 *at day* 0.
- ***(Rule of Flu-Spread)*** *Given a day D, if* (S_i, S_j) *is a pair of ill-healthy people on that day, the probability that* S_i *would infect* S_j *is given by a fixed number p; these events are independent.*
- ***(Rule of Flu-Recovery)*** *Suppose* S_i *is ill on day D, then* S_i *is healthy on day* $D + 1$ *iff* S_i *is ill on day* $D, D - 1, \ldots, D - \ell + 1$*; otherwise* S_i *is ill on day* $D + 1$.
- ***(Rule of Independence)*** *The probability that an* S_i *is sick on day D for* $i \in [n]$ *are mutually conditionally independent on knowledge of past events.*

For what follows, let us write $q = 1 - p$*.*

Mathematically, let us introduce the following random variables.

Definition 1.1.2. *Let* I_i^D *be the RV (random variable) indicating whether* S_i *is ill or not on day D; precisely:*

$$I_i^D = \begin{cases} 0 & \text{if } S_i \text{ is healthy on day } D \\ 1 & \text{if } S_i \text{ is ill on day } D \end{cases}$$

Given $K \in [n]_k$*, we write* J_K^D *as the event:*

$$J_K^D := \{I_i^D = \delta_{iK}\}_{i\in[n]}$$

That is, S_i *is sick on day D precisely when* $i \in K$*. Let* Σ^D *be the RV of victim counts each day, and T be the RV of end date of flu; precisely:*

$$\Sigma^D = \sum_i I_i^D,\ T = \min_{D>0}\{\Sigma^D = 0\}$$

The goal of this report is to tackle the following problem:

Problem 1.1.3. *How to:*

- *Simulate I_i^D, Σ^D.*
- *Calculate $E[\Sigma^D]$, $E[T]$, as well as their variances.*

1.2 Count of Newly-ill People and Markov Chains

In this subsection, we will show that how this problem can be regarded as a Markov chain.

It would be convenient to formulate 1.1.1 in terms of the notations in 1.1.2. We can first rewrite the first three conditions of 1.1.1 as follows

Lemma 1.2.1. *We have:*

- ***(Initial Condition)*** *$I_i^D = \delta_{i0}\delta_{D0}$ for $D \leq 0$.*
- ***(Rule of Flu-Spread)*** *Given $k \in [n]$, $K \in [n]_k$, $i \in [n] \setminus K$, we have:*

$$\Pr(I_i^{D+1} = 1 \mid J_K^D) = \Pr(\mathbf{B}_{k,p} > 0) = 1 - q^k$$

  where $q = 1 - p$ as in 1.1.1.

- ***(Rule of Flu-Recovery)*** *Given $I_i^D = 1$, we have $I_i^{D+1} = 0$ iff $I_i^D = I_i^{D-1} = \ldots = I_i^{D-\ell+1} = 1$.*

Remark 1.2.2. *The right hand side of the flu-recovery condition is equivalent to $I_i^{D-\ell} = 1 - I_i^{D-\ell+1} = 0$.*

These conditions alone allows us to make a simulation where we want to keep track of each person's healthy and sick days. The independence condition in 1.1.1 is more trickier to formulate. I have tried to find what are the minimal requirements for this to provide both a rigorous and a feasible calculation, but couldn't find an elegant one.

Suppose we want to know Σ^D. The essential random RV to encapsulate the whole situation is the **subset / number of newly sick people in each day**. They are defined as follows.

Definition 1.2.3 (Newly-Ill Count)**.** *Let N^D be the random variable:*

$$N^D = \{i \in [n] : I_i^D = 1 - I_i^{D-1} = 1\}$$

That is: N^D is the number of people that are newly-ill on day D.

Lemma 1.2.4 (Newly-ill determines Total-ill)**.** *We have $\Sigma^D = \sum_{d \in [\ell-1]} N^{D-d}$.*

Proof. For a fixed day D, we can divide $\{S_i\}_{i \in [n]}$ into the following four groups:

	Ill on day $D-1$	Not ill on day $D-1$
Ill on day D	A	B
Not ill on day D	C	D

The value of Σ^D is the sum of number of people in group A and B.

B is the group of "newly-ill" people. The number of people in group B is just N^D.

A is the group of "still-ill" people. They are the same group of people that are recently sick. This is given by the sum of people that are newly-ill on day $D-1, D-2, \ldots, D-\ell+1$.

This proves the identity. □

In this sense, computations of Σ^D reduces to computations of N^D. The good thing is that unlike Σ^D, the computation of N^D depends on a globally bounded amount of past dates.

Proposition 1.2.5 (Near-Markovity of Newly-ill)**.** *We have:*

$$\Pr\left(N^D = m \mid (N^{D-1-d})_{d\in[\ell-1]} = (k^d)_{d\in[\ell-1]}\right) = \binom{n+1-k}{m}(1-q^k)^m(q^k)^{n-k-m}$$

for any $(k^d)_{d\in[\ell-1]}$, m, D, where $q = 1-p$, $k = \sum_{d\in[\ell-1]} k^d$.

Proof. We have:

$$\begin{aligned} &\Pr\left(N^D = m \mid (N^{D-1-d})_{d\in[\ell]} = (k^d)_{d\in[\ell]}\right) \\ =& \Pr\left(N^D = m \mid \Sigma^{D-1} = k, (N^{D-1-d})_{d\in[\ell]} = (k^d)_{d\in[\ell-1]}\right) \\ =& \binom{n+1-k}{m}(1-q^k)^m(q^k)^{n-k-m} \end{aligned}$$

where the first equality is by 1.2.4. The second equality is rather intuitive - the number of people not sick on day $D-1$ is $n-k$, each of them has a chance of $(1-q^k)$ of getting sick on day D, so the formula follows. □

Remark 1.2.6. *If one wants to show 1.2.5 rigorously straight from the rules, one would need to use the law of total probability to condition on various configurations of disjoint subsets of $[n]$ of size $k^0, \dots, k^{\ell-1}$; this is a messy calculation that I don't think is relatively important for what follows.*

In this sense, suppose our goal is to simulate Σ^D, it suffices to keep track of a bounded number of N^D, and the transitional probabilities are given by a binomial distribution. A simulation design is outlines in ADDREF.

Another direct consequence of this is that we can define an assocaited Markov chain associated to this process which allows us to do some analytical calculations. The Markov chain is defined as follows.

Theorem 1.2.7 (Associated Markov Chain)**.** *Define the RV $(\mathcal{N}^D)_{D\in\mathbb{Z}} := ((N^{D-d})_{d\in[\ell-1]})_{D\in\mathbb{Z}}$*

- *We can read-off Σ^D via $\Sigma\mathcal{N}^D = \Sigma^D$.*
- *We can read-off T via $\min_{D\geq 0}\{\mathcal{N}^D = (0)_{d\in[\ell-1]}\}$.*
- *The RV $\mathcal{N}$ is a Markov chain with state space $\times_{d\in[\ell-1]}[n+1]$, and transition probability described as follows: take states $\sigma := (s^d)_{d\in[\ell-1]}, \tau := (t^d)_{d\in[\ell-1]}$.*
  - *If $s^d \neq t^{d+1}$ for $d \in [\ell-2]$, we have $\Pr\left(\mathcal{N}^{D+1} = \tau \mid \mathcal{N}^D = \sigma\right) = 0$.*
  - *Otherwise, the transition probability is $\binom{n+1-\Sigma\sigma}{t^0}(1-q^{\Sigma\sigma})^{t^0}(q^{\Sigma\sigma})^{n+1-\Sigma\sigma-t^0}$.*
- *This chain has an only absorbing state $(0, 0, \dots, 0) \in \times_{d\in[\ell-1]}[n+1]$.*

As a reality check, in the case $d = 1$, we have $\mathcal{N}^D = \Sigma^D$, and Σ^D solely depends on Σ^{D-1}.

Remark 1.2.8. *How big is the "actual" state space of this chain?*

- *Each N^D takes value in $[n+1]$, so it is bounded above by $(n+2)^\ell$.*
- *However, Σ^D also takes value in $[n+1]$, so in view of 1.2.4, the size of state space is bounded by the number of ℓ-nonnegative integers with sum $\leq n+1$ (distinguishing orders). Combinatorially, this number is:*

  $$\binom{n+1+\ell}{\ell} = \frac{(n+1+\ell)!}{(n+1)!\ell!}$$

  where H is the number of combination with replacement.
- *Conversely, for sufficiently large D, all the non-absorbing states are reachable.*
- *As a reality check, for the case $n, \ell = 30, 3$, $(n+2)^\ell = 32768$, while $C_\ell^{n+1+\ell} = 5984$.*

1.3 Number of Sick People each Day

After establishing the fact that $\{\mathscr{N}^D\}_{D\in\mathbb{Z}}$ is a Markov chain, in this subsection, we show how to compute the probability distribution of Σ^D. Due to the size of transition arrays, this might not lead to a practical algorithm, but is still sufficient for problems of smaller scale.

Consider the possible states of $\mathscr{N}^D$ as vector indices. We get a transition matrix $\mathscr{M}$ with entries given via the formula in 1.2.7 - that is, suppose σ, τ are two possible states of $\mathscr{N}^D, \mathscr{N}^{D+1}$, then:

$$\mathscr{M}_{\sigma,\tau} = \Pr(\mathscr{N}^{D+1} = \tau \mid \mathscr{N}^D = \sigma)$$

Here are some properties of $\mathscr{M}$.

Remark 1.3.1. *The following properties of the matrix $\mathscr{M}$ is clear:*

- *This matrix $\mathscr{M}$ is sparse: Since $\mathscr{N}^D = (N^{D-d})_{d\in[\ell-1]}$, $\mathscr{N}^D, \mathscr{N}^{D+1}$ has $\ell-1$-many repeated entries, so every state of $\mathscr{N}^D$ can only travel to at most $(n+2)$-different states at time $D+1$. In this sense, the number of nonzero entries is at most:*
$$\binom{n+1+\ell}{\ell}(n+2)$$
- *Again, we may get a better bound by imposing the constraint $\Sigma^{D+1} \in [n+1]$.*
- *However, by the third property listed in the remark following 1.2.7, a big power of this matrix is dense. In this sense, we should expect that the number of nonzero entries after a period of time inflates to:*
$$\binom{n+1+\ell}{\ell}^2$$
For example, when $n, \ell = 30, 3$, this quantity is $5984^2 = 35808256$.

Let ε be the initial state $(1, 0, \ldots, 0, 0)$, and for a state σ of $\mathscr{N}^D$, let $\mathbf{e}(\sigma)$ be the vector with entries 0 except at component σ that equals 1. Then we get the following:

Theorem 1.3.2. *Consider the RVs $\mathscr{N}^D, \Sigma^D$ with $D \geq 0$.*

- *The distribution of $\mathscr{N}^D$ is given by $\mathscr{M}^D\mathbf{e}(\varepsilon)$.*
- *The distribution of Σ^D is given by:*
$$\Pr(\Sigma^D = k) = \sum_{\sigma:\Sigma\sigma=k} \Pr(\mathscr{N}^D = \sigma) = \sum_{\sigma:\Sigma\sigma=k} \left(\mathscr{M}^D\mathbf{e}(\varepsilon)\right)_\sigma$$
- *The expectation of Σ^D is:*
$$E[\Sigma^D] = \sum_{k\in[n+1]} k \left(\sum_{\sigma:\Sigma\sigma=k} \left(\mathscr{M}^D\mathbf{e}(\varepsilon)\right)_\sigma\right) = \sum_{\sigma}(\Sigma\sigma)\left(\mathscr{M}^D\mathbf{e}(\varepsilon)\right)_\sigma$$
- *The variance of Σ^D is:*
$$E[(\Sigma^D)^2] - E[\Sigma^D]^2 = \sum_{k\in[n+1]} k^2 \left(\sum_{\sigma:\Sigma\sigma=k} \left(\mathscr{M}^D\mathbf{e}(\varepsilon)\right)_\sigma\right) - \left(\sum_{k\in[n+1]} k \sum_{\sigma:\Sigma\sigma=k} \left(\mathscr{M}^D\mathbf{e}(\varepsilon)\right)_\sigma\right)^2$$

1.4 Date of Flu End

For Markov chains with absorbing states, it is often fun to ask the question on the expected time of getting absorbed, and the variance of the absorbing time. In our scenario, this corresponds to the time when flu ends (since the only absorbing state is $(0)_{[\ell-1]}$).

The definition of the RV T for time of flu end is given in 1.1.2.

These are most often calculated via **first step analysis**. There are standard formulas on the expectation, variation of this RV, and we derive these here, as I can't find a source with complete derivation of the formula for variance.

For what follows, we use ω to denote the this absorbing state. For each state σ, consider the vector $\mathbf{T}$ with components defined by:

$$(\mathbf{T})_\sigma = E[T \mid \mathscr{N}^0 = \sigma]$$

We want to compute $(\mathbf{T})_\varepsilon$. By conditioning on first step, we get for $\sigma \neq \omega$:

$$E[T \mid \mathscr{N}^0 = \sigma] = \mathscr{M}_{\sigma,\omega} + \sum_{\tau\neq\omega} \mathscr{M}_{\sigma,\tau}(1 + E[T \mid \mathscr{N}^0 = \tau]) = 1 + \sum_{\tau\neq\omega} \mathscr{M}_{\sigma,\tau} E[T \mid \mathscr{N}^0 = \tau]$$

since the rows of $\mathscr{M}$ sums to 1. By writing $\sum_\sigma \mathbf{e}(\sigma)$ as $\mathbf{1}$, we get a matrix-vector representation:

Proposition 1.4.1. *The expected date of flu-end with initial state σ is given by the formula:*

$$\big((\mathbf{I} - \mathscr{M})_-^{-1}\mathbf{1}_-\big)_\sigma$$

where $\mathbf{I}$ is the identity matrix.

One can also approximate $(\mathbf{I} - \mathscr{M})_-^{-1}$ via truncating the series expansion $\sum_{i\geq 0} \mathscr{M}_-^i$.

The matrix $\mathbf{I} - \mathscr{M}$ can be ill-conditioned, in which case a direct numerical solution via **??** might not be feasible.

Now let us use first-step analysis to calculate variance. We have recursion (assuming $\sigma \neq \omega$):

$$\begin{aligned} E[T^2 \mid \mathscr{N}^0 = \sigma] &= \mathscr{M}_{\sigma,\omega} + \sum_{\tau\neq\omega} \mathscr{M}_{\sigma,\tau} E[(T+1)^2 \mid \mathscr{N}^0 = \tau] \\ &= 1 + \left(\sum_{\tau\neq\omega} \mathscr{M}_{\sigma,\tau} E[T^2 \mid \mathscr{N}^0 = \tau]\right) + 2\left(\sum_{\tau\neq\omega} \mathscr{M}_{\sigma,\tau} E[T \mid \mathscr{N}^0 = \tau]\right) \end{aligned}$$

Now let us write $\mathbf{t}$ as $(\mathbf{I} - \mathscr{M})_-^{-1}\mathbf{1}_-$ - which is the vector with entries $E[T^2 \mid \mathscr{N}^0 = \sigma]$ where $\sigma \neq \omega$ - and let $\mathbf{t}'$ be vector with entries $E[T^2 \mid \mathscr{N}^0 = \sigma]$ where $\sigma \neq \omega$. What we want to compute is:

$$\mathbf{t}' - \mathbf{t}_{\mathrm{sq}}$$

where $\mathbf{t}_{\mathrm{sq}}$ is $\mathbf{t}$ with each entry squared. The above recurrence is the same as:

$$(\mathbf{I} - \mathscr{M})_- \mathbf{t}' = \mathbf{1}_- + 2\mathbf{t}$$

Therefore, we get the formula:

Proposition 1.4.2. *The variance of flu-end date with initial state σ is given by the formula:*

$$\big((\mathbf{I} - \mathscr{M})_-^{-1}(\mathbf{1}_- + 2\mathbf{t}) - \mathbf{t}_{\mathrm{sq}}\big)_\sigma$$

where the notations are similar as above.

2 Experiments with "flusim"

"flusim" is a simple Python package I wrote using the maths developed in the previous section, that can be used to simulate, visualize, calculate the related statistics of this flu problem. The package is available at the testpypi website or at the GitHub repository. The package can also be installed from command line via:

```
pip install -i https://test.pypi.org/simple/ flusim==0.0.1
```

2.1 Simulation

In order to calculate the statistics, I ran 30000 trials with cut-off day set to 20000, and calculated the 95% confidence interval for normal means and variance via the formulas in [DP21]. Although the trials are i.i.d. but not normal, for very large number of trials the answer shouldn't be too off from the formula. The results:

```
Simualted Day 1-3 statistics and Flu End Date Statistics
(using default alpha=0.05, which stands for 95% confidence):
- Day 1:
    - Expectation of Sick: (1.5890896344033556, 1.6089103655966444)
    - Variance of Sick:    (0.57891751165741, 0.5977467300370205)
- Day 2:
    - Expectation of Sick: (2.511890245423345, 2.5396430879099885)
    - Variance of Sick:    (2.2251940147909797, 2.297568166890038)
- Day 3:
    - Expectation of Sick: (2.8692246733829467, 2.904708659950386)
    - Variance of Sick:    (5.9465849114057265, 6.139996828742952)
- Flu End Date:
    - Expectation of End:  (689.7894987139955, 690.4691679526711)
    - Variance of End:     (800439.4003986634, 826473.5900133488)
```

2.2 Calculation

In order to calculate the statistics, I used the formulas derived in the previous chapter. The results:

```
Calculated Day 1-3 statistics and Flu End Date Statistics
(using default alpha=0.05, which stands for 95% confidence):
- Day 1:
    - Expectation of Sick: 1.6000000000000005
    - Variance of Sick:    0.5880000000000005
- Day 2:
    - Expectation of Sick: 2.5203542928556217
    - Variance of Sick:    2.243334133156104
- Day 3:
    - Expectation of Sick: 2.8796008587930224
    - Variance of Sick:    6.033295056236872
- Flu End Date:
    - Expectation of End:  687.399934583829
    - Variance of End:     822054.0766941558
```

2.3 Conclusion

When the problem size is fairly tractable, computations, simulations can be made. In the case where the population size is 31, day of recovery is 3, our simulations and calculations doesn't differ too much.

The Jupyter notebook and the output html file associated to the results above are included in the attached files `package_demo.ipynb`, `package_demo.html`.

References

[DP21] Goldsman David and Goldsman Paul. *A First Course in Probability and Statistics*. Lulu, 2021.