Analysis of the stochastic spread of Stride simulation outcomes.

Episim

April 22, 2018

Abstract

The goal of this paper is to determine the stochastic spread of Stride simulation outcomes and find boundaries for testing purposes. We will discuss the following use cases (diseases).

- Influenza A
- Influenza C
- Measles (R0 is 12)
- Measles (R0 is 16)

1 Configuration and data files

These files are used in all our test runs.
The files are found in our repo and under the provided link.

- GeoGen config Config of geogen.

- Flanders Cities data for relative pop of flanders.

- Flanders Commuting commuting info of flander cities.

- Household Flanders household sample data.

2 Scripts

We used 3 extra scripts to gather data and compile the gathered data into images and results.

2.1 SeedGenerator.py

Found here. Helper script to generate a given amount of random seeds. Written to a .txt file. Same format as Seedtester will use to generate data off seeds.

2.2 SeedTester.py

Found here.
This script will run stride with the config files and seed.txt file provided. Running stride for n given days. It will output for each seed the seed and amount of infected to a .dat files (referenced later).

2.3 Plotter.py

Found here. This will generate for each .dat file in seedTesterResults a:

- box plot

- histogram

- QQ-plot

- Scatter plot

- A data file containing useful data like variance, mean and boundaries...

3 Quantifying boundaries

In our previous paper we already concluded that the number of threads and choice of random number generators did not have a large impact on end results. In this paper we will not attempt to analyze all data with every different thread/generator configuration. Instead we used lcg64 and 1 thread (sequential) in all our test cases.

In this section we will discuss the results of each of the above given use cases and try to decide boundaries of our data. Notable parameters are. The population is of size 600 000 and we used 30 steps (simulated days) for each run.
Plotter.py compiles some useful information about the runs on Influenza A. First off we inspect our QQ plot. The data tends to follow the line so we could assume if our sample set is bigger the data converges to a normal distribution for all use cases.

A naive interpretation to set boundaries is to take the lower and upper bound each sample. This is insufficient since we only tested a small sample size and it is still possible there are outliers.

Another approach we attempted is to make use of a BTI to take as data points. So we could assume 99% of the cases would fall with in the BTI. But in practice, in some cases the tests still gave False negatives.

The last approach, is to calculate a "margin" parameter (that is also used in BatchRuns.cpp for use cases). Margin represents a fraction of how much a result can be off the average or "expected result". This method seemed to be fool proof on False negatives. And it is reasonable to say that the given margin is not to big for False positives.

Formula to calculate the margin:

$$ml = (mean - lower_bound)/mean$$
$$mh = (higher_bound - mean)/mean$$
$$margin = max(ml, mh)$$

What follows are the results our plotter script compiled for each of the use cases. What's most important is to take a look at the distribution and the calculated "Batchruns.cpp" margin value.
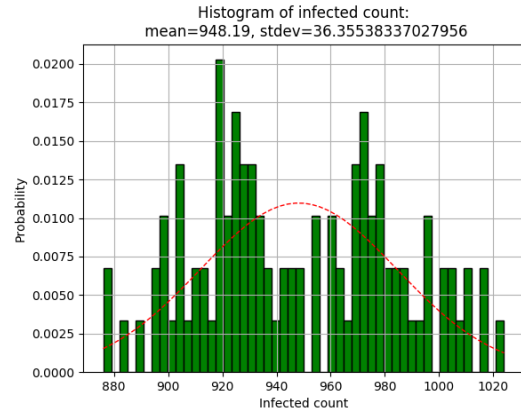
3.1 Influenza A

Stride config file found here.
Use of this seed file.
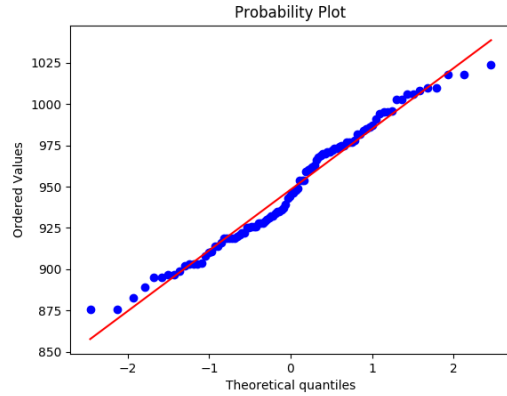
```
Mean:                    948.19
Standard deviation:      36.53853507537686
Lower bound:             876
Higher_bound:            1024
Shapiro-wilk test-stat:  0.9734843373298645
Shapiro-Wilk p-value:    0.041086576879024506
99% confidence interval: (923.5157171969315, 972.8642827171041)
Batchrun.cpp margin:     0.07995233022917342
```
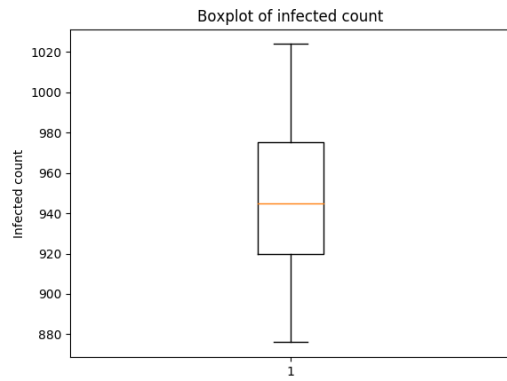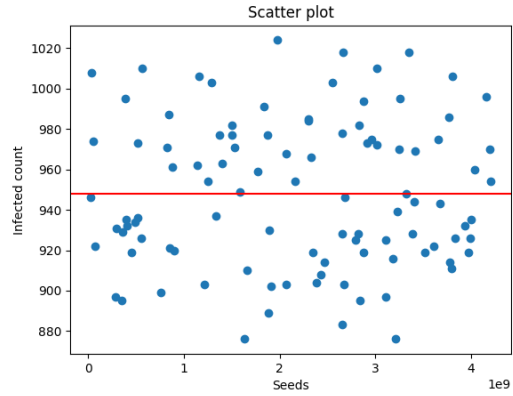

(a) Boxplot of the Influenza A runs

(b) Histogram of the Influenza A runs

(c) QQ-plot of Influenza A

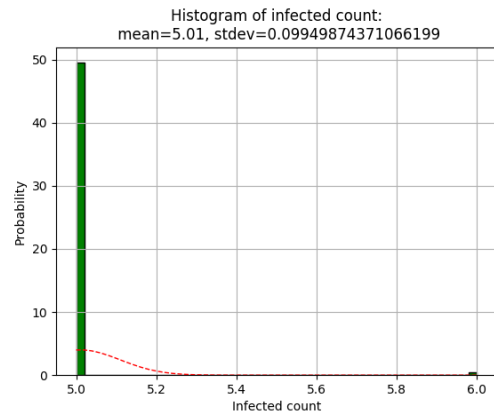(d) Scatter plot of Influenza A

Figure 1: Figures of Influenza A

3.2 Influenza C

Stride config file found here.
Use of this seed file.
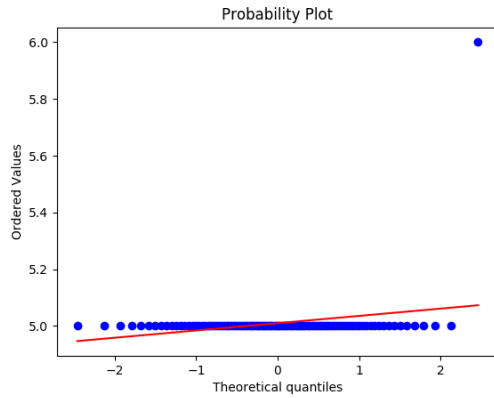
```
Mean:                     5.01
Standard deviation:       0.1
Lower bound:              5
Higher_bound:            6
Shapiro-wilk test-stat:  0.07519251108169556
Shapiro-Wilk p-value:    3.3711737281791994e-22
99% confidence interval: (4.9424705198164975, 5.077529479948231)
Batchrun.cpp margin:     0.19760479041916174
```
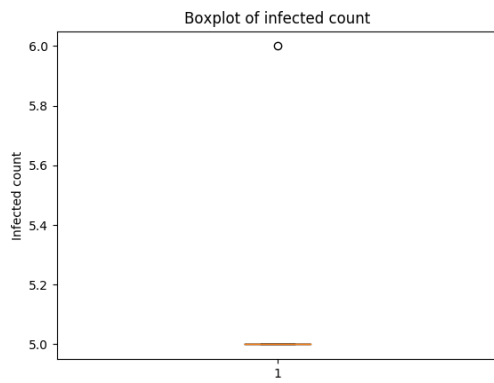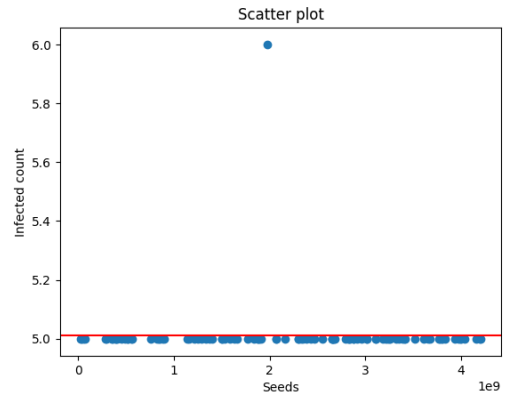

(a) Boxplot of the Influenza C runs

(b) Histogram of the Influenza C runs


(c) QQ-plot of Influenza C

(d) Scatter plot of Influenza C

Figure 2: Figures of Influenza C

3.3 Measles 12

Stride config file found here.
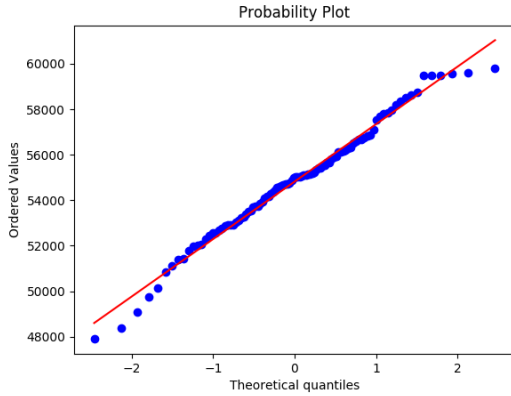Use of this seed file.

```
Mean:                   54819.31
Standard deviation:     2503.0075876308038
Lower bound:            47897
Higher_bound:           59786
Shapiro-wilk test-stat: 0.9848060607910156
Shapiro-Wilk p-value:   0.3075026273727417
99% confidence interval: (53129.0419871193, 56509.57800699185)
Batchrun.cpp margin:    0.12627502972948762
```
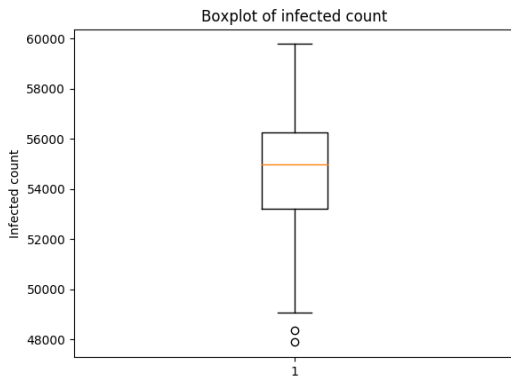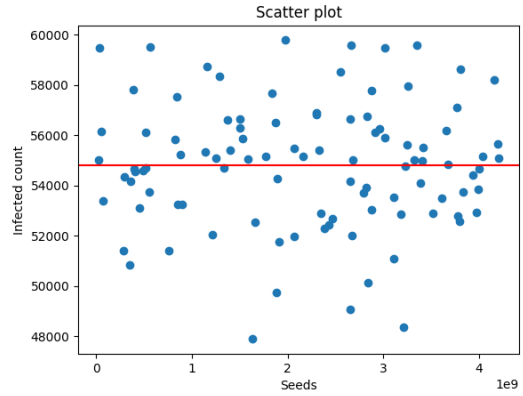

(a) Boxplot Measles 12 runs
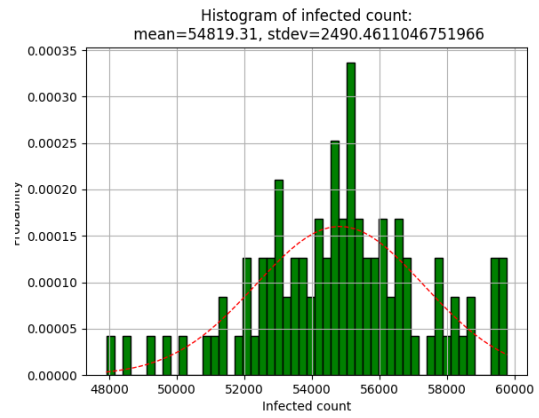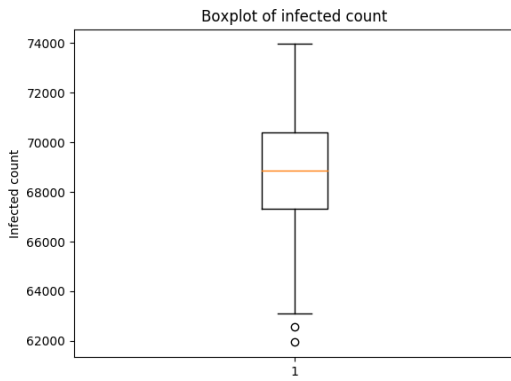
(b) Histogram of the Meaz1les 12

(c) QQ-plot of Meazles1 12

(d) Scatter plot of Measles 12
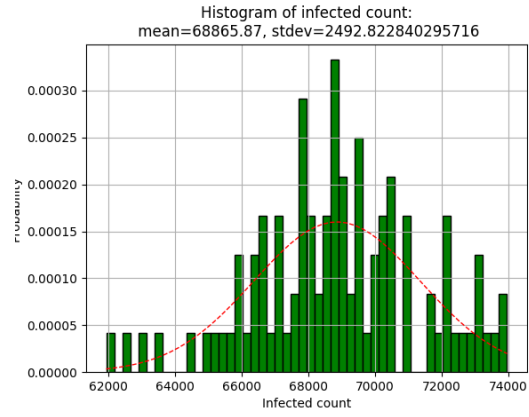
Figure 3: Figures of Measles 12

3.4 Measles 16

Stride config file found here.
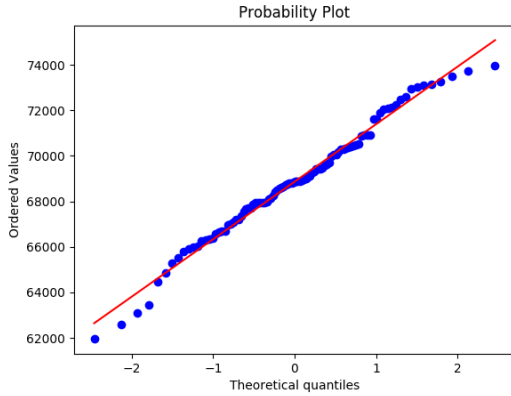Use of this seed file.

```
Mean:                  68865.87
Standard deviation:    2505.3812212390703
Lower bound:           61947
Higher_bound:          73964
Shapiro-wilk test-stat: 0.9847463369369507
Shapiro-Wilk p-value:  0.30445000529289246
99% confidence interval: (67173.99908468217, 70557.74090942339)
Batchrun.cpp margin:   0.10046878083439585
```
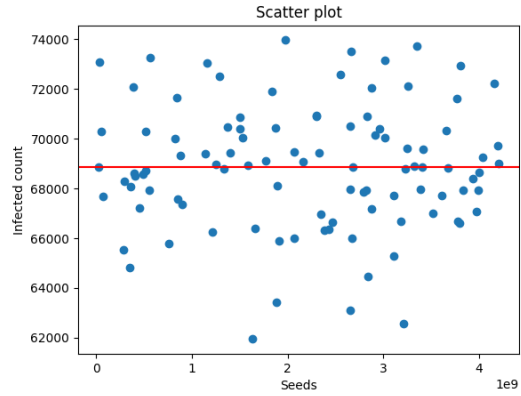

(a) Boxplot Measles 16 runs

(b) Histogram of the Meaz1les 16

(c) QQ-plot of Meazles1 16

(d) Scatter plot of Measles 16

Figure 4: Figures of Measles 16

Conclusion

Something goes wrong when we try to run the Influenza C use case. (Or only five people are susceptible to this disease but this is rather unlikely). The margins we found that we can use in the scenario tests are

- Influenza A : 0.080

- Measles 12 : 0.126

- Measles 16 : 0.100

Further we can conclude the stochastic spread is according to a normal distribution for the three representative use cases.We do this by checking the QQ plots.