Freie Universität Berlin
Fachbereich Mathematik und Informatik
Institut für Mathematik

An Epidemic as a Sequence of Random Events

Seminar Report in the Module “Stochastics in Action”

Johannes Schade
5564109
10.07.2023

Contents

1 Introduction

In 2023, one could argue that an essay covering an epidemic does not require an introduction. On the other hand, it is worth pointing out the stochastics going on in an epidemic, since this provides the groundwork for simulating such events and thus opening up the possibility to make predictions.

As stochastics is the field of studying randomness, making such simulations relies on simulating randomness. In the digital age, this of course should include generating randomness on computers. This is not as straightforward task as one might think, because computers are deterministic in nature – thus, true randomness is impossible to generate on computers.

In practice, randomness in form of drawing random numbers from a distribution is substituted by generating *pseudo-random numbers*, which are not genuinely randomly drawn numbers, but a collection of numbers that *appear* to be randomly drawn [Law15, Chapter 7.1], which is in most cases an approach fit for purpose.

In section 2, we will investigate where and how randomness can be found in an epidemic. Furthermore, we will develop a naïve mathematical model for simulating the number infections caused by one infected individual using Poisson processes.

Having a computer simulation of the model in mind, section 3 demonstrates how continuously uniformly distributed pseudo-random numbers can be generated on a computer. We proceed by presenting the *alias method* introduced by [Wal77], allowing to generate pseudo-random numbers from discrete finite distributions using the uniform distribution generator. Furthermore, an outlook is given how this can be extended to infinite discrete distributions.

Finally, section 4 an overveiw over an actual simulation using the alias method is given.

2 An Epidemic as a sequence of random events

This section follows [Cas21]. The goal of the section is to point out the randomness of the number of contacts made, which can greatly influence the number of individuals infected by an contagious individudal. We will identify Poisson-processes as a suitable model and thus provide a motivation for the Alias method, described in later sections.´

A starting point for modelling any epidemic are the states individuals from a population go through. These could, for example, be:

1. S: susceptible to infection (not yet infected)

2. E: exposed (infected, but not yet infectious)

3. A: asymptomatic (infectious, but not yet showing symptoms)

4. I: infectious (infectious, and showing symptoms)

5. R: recovered (uninfectable by previous infection or other means like vaccination)

Here, we restrict ourselves to modelling the number of transmissions one infected individual might cause. This, of course, depends on a) he number of contacts made during the infection and b) the probability, a contact gets infected.

2.1 Number of contacts as Poisson processes

In this subsection (follwoing [Geo02, Chapter 2.4]), we want to find a suitable distribution of a random variable X_i modelling the number of contacts made on day i. A question for an answer to the problem could be stated as "How likely do k events happen in a given period of time?". This leads to the construction of the *Poisson distribution* (**Pois**), where the considered time interval T and is partitioned into n equidistant subintervals of length T/n. If $n \longrightarrow \infty$, these subintervals become arbitrarily small, leading to the conclusion that at most one event is taking place in one of these subintervals. Futhermore, it is reasonable to assume that this probability is proportional to the length of the subinterval, say $\alpha \frac{T}{n}$. Furthermore, we assume that that an event happens in subinterval i independently of whether an event happens in subinterval j for $i \neq j$. Thus, the probability of k events happening in n subintervals can be assumed to be $\lim_{n\to\infty} \mathbf{Bin}(n, \alpha T/n)$ distributed. This leads to convergence to the following probability function:

Theorem 1. Let $T > 0$, $p_n = \alpha \frac{T}{n}$ and $\lim_{n\to\infty} np_n = n\alpha T/n = \alpha T =: \lambda$. Then

$$\lim_{n\to\infty} \Pr_n(k) = \lim_{n\to\infty} \binom{n}{k} p_n^k \left(1 - p_n\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Proof.

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{k} p_n^k \left(1-p_n\right)^{n-k} &= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} p_n^k \left(1-p_n\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} p_n^k \left(1-p_n\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k \left(1-p_n\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{\mathcal{O}(n^k)}{k!} p_n^k \mathcal{O}\left(\left(1-p_n\right)^n\right) \\
&= \lim_{n\to\infty} \frac{n^k}{k!} p_n^k \left(1-p_n\right)^n \\
&= \lim_{n\to\infty} \frac{\lambda^k}{k!} \left(1-\frac{np_n}{n}\right)^n \\
&= \frac{\lambda^k}{k!} e^{-np_n} = \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

□

This probability function describes the **Pois**(λ) distribution with *intensity* λ.

Theorem 2. Let $X \sim \mathbf{Pois}(\lambda)$. Then $\mathrm{E}[X] = \lambda$.

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}$$

□

Thus we can simulate a table of the number of contacts c_i on day i by sampling from **Poi**(λ), with a sensible average number of contacts λ, e.g. $\lambda = 4$:

i	1	2	3	4	5	6	7	8	9	…
c_i	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	…

where $X_i \overset{\text{iid}}{\sim} \mathbf{Pois}(\lambda)$.

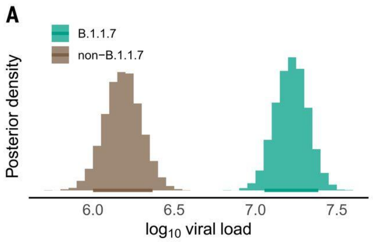

Figure 1: From [Jon+21]

2.2 Infection Probability

We still need as input (e.g. from statistical observations). Usually, an individual remains relatively uncontagious shortly after infection, with contagiousness climaxing after a couple of days and then decreasing to 0 again (compare to figure 1).

This could be given as a table:

i	1	2	3	4	5	6	7	8	9	≥ 10
r_i	0	0	0	0	0.1	0.3	0.4	0.4	0.2	0

Now we can simulate the average number of people infected by one person by calculating

$$\begin{aligned}
\sum_i r_i c_i &= \sum_i r_i X_i \\
&= 0 \cdot X_1 + 0 \cdot X_2 + 0 \cdot X_3 + 0 \cdot X_4 + 0.1 \cdot X_5 + 0.3 \cdot X_6 \\
&\quad + 0.4 \cdot X_7 + 0.4 \cdot X_8 + 0.2 \cdot X_9 \\
&= \mathbf{r}^\intercal \mathbf{X}
\end{aligned} \tag{1}$$

where we denote $\mathbf{X}$ as a 10-dimensional vector of iid $\mathbf{Pois}(\lambda)$ random variables and $\mathbf{r} \in [0,1]^{10}$.

3 Sampling from a discrete distribution

In this section, we will show how pseudo-random numbers of finite discrete distributions can be generated. First, we will introduce a procedure to pseudo-sample from the continuous uniform distribution in the interval $[0,1)$ ($\mathcal{U}_{[0,1)}$). Then we will use this result by using the alias method of [Wal77] to pseudo-sample from arbitrary finite discrete distributions. Finally, we will discuss how this might be extended to infinite discrete distributions.

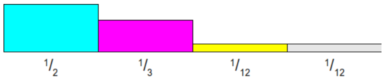

Figure 2: Distribution $\mathcal{D}$. From [Sch11]

3.1 Uniform distribution generator

This subsection picks from [Law15, Chapter 7.2]. We will introduce the *Linear congruential generator* (LCG). One can sample from $\mathcal{U}_[0, 1)$ by proceeding with the following: For a seed (initial value) $Z_0 \in \mathbb{N}_{>0}$, obtain z_i by recursive formula

$$Z_i = aZ_{i-1} + c \mod m$$

with $a, c, m \in \mathbb{N}_{>0}$, $a < c < Z_0 < m$ and return $U_i = \frac{Z_i}{m}$.´

The seed Z_0 is commonly chosen in dependence of the computer clock or temperature to introduce an element of randomness. But it should be mentioned that a, c, m, Z_0 determine $Z_i, \forall i > 0$ completely, which reminds of the pseudo-random nature of this procedure (i.e. the sample only *appears* to be uniformly distributed). Moreover, , be algebraic nature of the modulus, addition and multiplication operations, we have a periodicity of at most m, when the pseudo-random numbers start to repeat. This means that, depending the size of the sample one desires, a large m should be chosen.

More advanced generators are available (compare [Law15, Chapter 7]), which are often inspired by LCGs.

3.2 Walker's trick

One way (amongst many) to sample from finite discrete distributions given a $\mathcal{U}_{[0,1)}$ generator is given by Alastair Walker [Wal77], also called the *alias method.* Here we will use [Cas21] and a post of Keith Schwarz [Sch11] to illustrate the approach.

This can be descirebed with "putting the probabilities in a box". As an illustrating example, consider a finite distribution $\mathcal{D}$ with $n = 4$ possible values (compare figure 2), each probability bar with same width: The goal of this procedure is to create 4 equal-heights column with at most 2 probabilites (figure 5).

For this example, that could be achieved by the following steps:

1. Scale by $n = 4$ (figure 3).

2. Redistribute some mass from first probability $\Pr[X = 1], X \sim \mathcal{D}$ column to the columns 3 and 4 (i.e. the columns the probabilities $\Pr[X = 3], \Pr[X = 4]$) (figure 4).

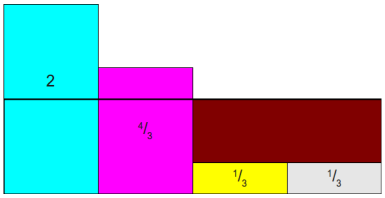

Figure 3: From [Sch11]

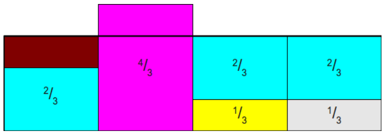

Figure 4: From [Sch11]

3. Fill up the free space of the first column with the overshoot of the second column (figure 5)

Now we can sample from $\mathcal{D}$ by

1. sampling a value w from $\mathcal{U}_{[0,1)}$ and choose column $\lfloor nw \rfloor$
2. sampling a value h from $\mathcal{U}_{[0,1)}$. Choose Prob if $h \leq$ Prob, else choose Alias from the table in figure 5.

We can interpret this the following way: We will sample the "original" value $\lfloor nw \rfloor$, but with the probability $1 - \text{Prob}\left[\lfloor nw \rfloor\right]$ we will sample its "alias" $\text{Alias}\left[\lfloor nw \rfloor\right]$ instead.

The creation of the alias table (steps 1-3) can be described as matching a column a of height bigger than 1 with a column b of height smaller than 1, and then filling up b with mass from a.

This generation of an alias table can be generalized with an algorithm for an arbitrary finite distribution taking n values (according to [Sch11]), which is done in algorithm 1.

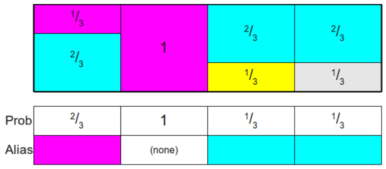

Figure 5: From [Sch11]

Algorithm 1: Creating an alias table

```
Input  : $d \in \mathbb{R}^n_{\geq 0}$ s.t. $\sum_{i=1}^{\infty} d_i = 1$
Output: probability array $p \in [0,1]^n$, alias array $a \in [0 : n-1]^n$
$d \leftarrow n \cdot d$;
initialize $p :=$ None$^n$, $a := \mathbf{0} \in \mathbb{R}^n$;
for $i = 0$ to $n-2$ do
    find $g, \ell \in [0 : n-1]$ s.t. $d_g \geq 1, d_\ell < 1$;
    $p_\ell \leftarrow d_\ell$ ;                       // assign probability to $p_\ell$
    $a_\ell \leftarrow g$ ;                         // assign alias value of $\ell$ to $a_\ell$
    $d_g \leftarrow d_g - (1 - d_\ell)$ ;          // redistribution of probability mass
    $d_\ell \leftarrow$ None ;            // remove value with alias assigned from $d$
end
find $k$ s.t. $p_k =$ None ;                      // one value left to consider
$p_k \leftarrow 1$;
```

It is not hard to see for every finite probability distribution, columns of equal width can always be distributed into such a box, since probability distributions are bounded. Concluding that generating a box with contributions of at most two columns is not trivial. This is safeguarded by the following theorem.

Theorem 3 (Schwarz)**.** Let $h_0, ..., h_{n-1}$ be the heights of n rectangles of width 1, such that $\sum_{i=0}^{n-1} h_i = n$. Then there exists a distribution of the rectangles such that the height is always 1, the i-th column has non-zero mass of the i-th rectangle and each column has mass of at most two rectangles; This can be achieved by algorithm 1

Proof. We induct on n.

The case $n = 0$ is trivial. Assume there exists an n for which the statement holds, and that there are $n+1$ rectangles with heights $h_1, ..., h_n$ such that $\sum_{i=0}^{n} h_i = n+1$. Then there exist g such that $h_g \geq 1$ and an ℓ such that $g_\ell \leq 1$. This can be seen by contradiction: if there wasn't such an ℓ, then $h_i > 1\ \forall i$, which implies $\sum_{i=0}^{n} h_i > n+1$, contradicting the assumption. By an analogous logic, such an ℓ , because otherwise $h_i < 1\ \forall i \neq \ell$, which would result in the contradiciton $\sum_{i=0}^{n} h_i < n+1$.

Thus such an redistribution must exist.

Now consider the following construction. Consider column ℓ having height h_ℓ, which leaves $1 - h_\ell$ unfilled. Fill the rest of column ℓ with a section of h_g and reduce column g by the mass $1 - h_\ell$ redistributed to column ℓ. Then n rectangles with heights $\{h_i : i \in [0:n] \setminus \{\ell\}\}$ remain and using the induction step, we well find such a desired redistribution. $\square$

3.3 Sampling from infinite discrete distributions

Generally, Walker's trick only works for finite discrete distributions, since it requires as input a distribution specified as a finite list. For our application, this does not form a big problem, since for **Pois**(4), we have $\Pr[X \geq 20]$ is practically 0. If one nevertheless wishes to sample from a truly infinite distribution, one can still use the alias method at least partly. [Wal77, Chapter 8.4.3] suggests the following:

Find an n such that $q := \sum_{i=0}^{n} \Pr(i) \approx 1$, i.e. $\Pr[X \leq n] \approx 1$ It holds that

$$\Pr(i) = q \left[\frac{\Pr(i)}{q} \mathbb{1}_{[0:n]}(i) \right] + (1-q) \left[\frac{\Pr(i)}{1-q} (1 - \mathbb{1}_{[0:n]}(i)) \right]$$

and that

$$\sum_{i=0}^{\infty} \frac{\Pr(i)}{q} \mathbb{1}_{[0:n]}(i), \sum_{i=0}^{\infty} \frac{\Pr(i)}{1-q} (1 - \mathbb{1}_{[0:n]}(i)) = 1$$

since $\sum_{i=0}^{\infty} \Pr(i) \nVdash_{[0:n]}(i) = q, \sum_{i=0}^{\infty} \Pr(i)(1 - \nVdash_{[0:n]}(i)) = 1 - q)$ by construction. This gives rise to the following procedure:

1. Sample h from $\mathcal{U}_{[0,1)}$
2. If $h \leq q$, return the alias method for $\frac{\Pr(x)}{q}$
3. If $h > q$, return other method for $\frac{\Pr(x)}{1-q}$

A method for $x > q$ could e.g. be the inversion method (see [Wal77, Chapter 8.4.3]). The conversion method for a probability function $\Pr(\cdot)$ essentially consists of sampling $u \sim \mathcal{U}_{[0,1)}$ and returning the smallest x such that

$$\sum_{i=0}^{x} \Pr(i) > u.$$

This is a discretized version of returning $p^{-1}(u)$ for a continuous distribution p.

4 Simulation

This section provides a brief overview over the simulation of equation 1 using the alias method. We first implemented a function `poisson(i, lamb)` implementing the probability function of **Pois**(`lamb`) for value `i`. Using this function with `lamb=4`, we generated an array `d` with 20 elements, where `d[i]`$=$ Pr[i] for $i < 18$, and `d[19]` $=$ Pr[$i \geq 20$]. We further implemented algorithm 1 in a function `alias_table(d)` taking an array `d` specifying a probability distribution and returning arrays `probs` and `alias` forming the alias table. The function `alias_sample(probs, alias)` samples from an alias table specified as two arrays `probs` and `alias` and returning an integer bigger or equal than 0, as described in section 3.2.

We sampled 10^4 **Pois**(4) pseudo-random numbers to check this method for consistency and found the sample mean to be 4.0024, and the sample variance to be 3.92139424, which is close to the true mean and variance, which are both 4.

For simulating equation 1, we sampled 10^3 arrays of size 10, where for each element of the arrays the alias method was used. Computing the dot product with an array `r` corresponding to r of section 2.2 for each array. The expected number of infections over all arrays was calculated to be 5.6216, which is close to the true value of 5.6.

References

[Cas21] Bill Casselman. *An epidemic is a sequence of random events*. May 2021. URL: `https://mathvoices.ams.org/featurecolumn/2021/05/01/may-fc/`.

[Geo02] Hans-Otto Georgii. *Stochastik*. de Gruyter Lehrbuch. de Gruyter, 2002.

[Jon+21] Terry C Jones et al. “Estimating infectiousness throughout SARS-CoV-2 infection course”. In: *Science* 373.6551 (2021), eabi5273.

[Law15] Averill Law. *Simulation Modeling and Analysis*. 5th ed. McGraw Hill Higher Education, 2015. ISBN: 9781259010712.

[Sch11] Keith Schwarz. *Darts, Dice, and Coins: Sampling from a Discrete Distribution*. 2011. URL: `https://www.keithschwarz.com/darts-dice-coins/`.

[Wal77] Alastair J Walker. “An efficient method for generating discrete random variables with general distributions”. In: *ACM Transactions on Mathematical Software (TOMS)* 3.3 (1977), pp. 253–256.