UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

Curso 2022/2023

DEPARTAMENTO DE ANÁLISIS MATEMÁTICO Y
MATEMÁTICA APLICADA

TRABAJO DE FIN DE GRADO

GRADO EN MATEMÁTICAS

Detección de tiempos vía modelos de difusión

David Sandoval Rodríguez

Dirigido por
Ana Carpio Rodríguez

Madrid, a 27 de Junio de 2023

Resumen

Este trabajo de fin de grado tiene como principal objetivo revisar y ampliar la información sobre los resultados obtenidos por Yuri Bakhtin en su artículo *Universal Statistics of Incubation Periods and Other Detection Times via Diffusion Models* [12]. En la publicación el autor demuestra como, bajo ciertas condiciones, la distribución del tiempo de incubación para enfermedades modelizadas como procesos de difusión presenta una característica asimetría positiva. El análisis está motivado por las conclusiones presentadas en *Evolutionary dynamics of incubation periods* [11] donde el problema se enfrenta con la modelización de organismos mediante grafos y el desarrollo de los síntomas como procesos estocásticos. El resultado es especialmente sorprendente porque las principales explicaciones acerca de la asimetría apelaban a la heterogeneidad de los individuos, mientras que estos dos artículos obtienen estas mismas observaciones modelizando organismos homogéneos.

Palabras clave: Modelos de difusión, camino aleatorio, tiempo de parada, asimetría positiva, modelos de nacimiento-muerte

Abstract

This final year project aims to review and expand the information regarding the results obtained by Yuri Bakhtin in his article *Universal Statistics of Incubation Periods and Other Detection Times via Diffusion Models*[12]. In the publication, the author demonstrates how, under certain conditions, the distribution of the incubation time for diseases modeled as diffusion processes exhibits a positive (right) skewness. The analysis is motivated by the conclusions presented in *Evolutionary dynamics of incubation periods* [11], where the problem is tackled by modeling organisms using graphs and symptom development as stochastic processes. The result is particularly surprising because the main explanations for the skewness relied on the heterogeneity of individuals, while these two articles obtain these same observations by modeling homogeneous organisms.

Keywords: difussion models, random walk, first passage time, right skew, birth-death model

Índice

1. Introducción

1.1. Presentación

El estudio de los tiempos de incubación de las enfermedades supone un pilar fundamental en nuestra capacidad tanto para prevenir contagios en una población como para tratar las enfermedades antes de que aparezcan los síntomas que pueden poner en peligro la vida del individuo.

Numerosas investigaciones han recopilado datos recogiendo los períodos de incubación de diversas poblaciones y enfermedades y todas ellas revelan que la exposición simultánea de una población a una enfermedad no resulta en un desarrollo simultáneo de los síntomas sino en una amplia distribución de tiempos [11]. Lo llamativo está en que las distribuciones son siempre similares: unimodales y positivamente asímetricas. Unimodales porque solo un valor se repite con muchá más frecuencia, que para una distribución continua significa tener un solo máximo local y positivamente asimétrica porque la distribución está "deformada"por la derecha, con un abrupto descenso por la izquierda y menos pronunciada por la derecha (del inglés *right skewed*).

El período de incubación lo podemos entender como el tiempo que tarda el agente dañino en multiplicarse dentro del individuo afectado hasta al alcanzar el umbral que supone el inicio de los síntomas. Los datos históricos de estudios donde se conoce el momento de exposición al agente nocivo permite reconstruir los períodos de incubación que llamaremos así aunque no se trate de un agente infeccioso. Naturalmente, por grande que sea la muestra, obtendremos de los datos reales un gráfico de columnas pero en él ya se puede intuir ese aspecto característico de la distribución que aparece recurrentemente.

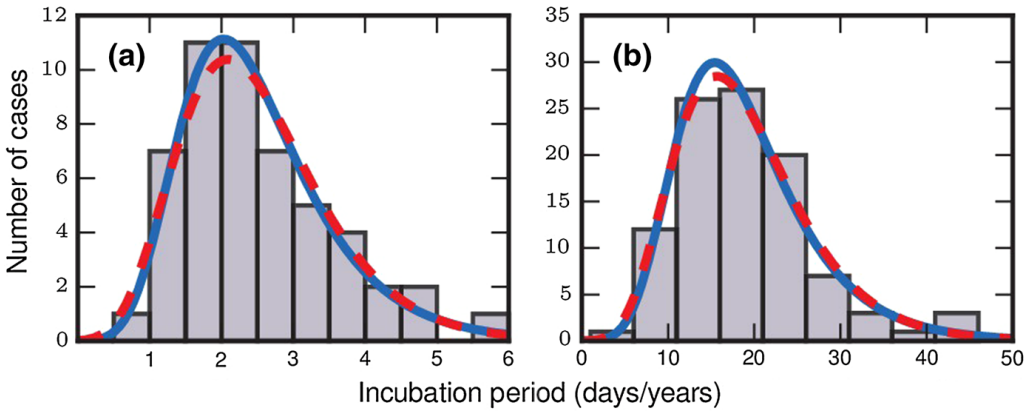

Figura 1: Tomado de Ottino-Löffler et al.[11]. Datos de un brote de faringitis estreptocócica transmitida por alimentos, reportado en Sartwell (1950), el tiempo se mide en días. b. Datos de un estudio de tumores de vejiga en trabajadores expuestos a un carcinógeno en una fábrica de tintes (Goldblatt 1949). El tiempo se mide en años. La figura muestra un gráfico de columnas con los datos reales y las aproximaciones a distribuciones continuas que pueden encontrarse en Ottino-Löffler et al. [11]

Ante este patrón, explicaciones previas buscaban la respuesta en individuos que en el momento del contagio tienen respuestas inmunitarias y estados de salud diferentes, a patógenos con diferentes capacidades, etc [1]. La naturaleza aleatoria de muchos de estos parámetros resulta poco satisfactoria para modelizar y predecir resultados. Es por ello que en Ottino-Löffler et al. [11] los autores tratan de estudiar la distribución de tiempos de incubación con un modelo que permita simular miles de infecciones y alterar cada parámetro. Para ello se utilizan grafos cuyos nodos y aristas representan las células del sistema. Este artículo se explicará en más detalle en el siguiente capítulo.

De manera similar, en el artículo de Yuri Bahktin se trata una clase de modelos más amplia partiendo del trabajo en Ottino-Löffler et al. [11].

1.2. Definiciones

Se detallan a continuación definiciones y conceptos que se consideran necesarios para seguir este trabajo y que podrían no ser elementales.

Definición 1.1. *Un camino aleatorio unidimensional es un proceso estocástico que modeliza la trayectoria en un solo eje de un partícula cuyo movimiento a lo largo del tiempo está determinado por una variable aleatoria.*

Para el caso discreto, la posición en el enésimo paso es igual a la suma de n variables aleatorias idénticamente distribuidas

$$X_n = X_0 + Z_1 + Z_2 + ... + Z_n$$

o lo que es lo mismo

$$X_n = X_{n-1} + Z_n$$

En el caso continuo, la posición tras un lapso de tiempo h está determinada por el estado inicial y una variable aleatoria Z

$$X(t+h) = X(t) + Z(h).$$

Definición 1.2. *Se dice que una variable aleatoria es un tiempo de parada τ para un camino aleatorio $X(t)$ si modeliza el momento en el que se alcanza en el camino un estado a, es decir, $\tau = \{t : X(t) = a\}$. En el caso de un camino discreto $\tau = \{n : X_n = a\}$.*

Definición 1.3. *Un camino aleatorio es un proceso de Markov si la posición en el instante inmediatamente futuro siguiente depende únicamente del estado presente. A esto se le llama Propiedad de Markov*

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

Definición 1.4. *Un proceso estocástico cumple la propiedad **fuerte** de Markov si condicionado a un tiempo de parada τ se cumple*

$$P(X_{\tau+m} = x_{\tau+m} | X_\tau = a) = P(X_m = x_{\tau+m} | X_0 = a).$$

La propiedad fuerte caracteriza el hecho de que el camino aleatorio se comporte de la misma manera tras pasar por un estado a que si hubiera comenzado en ese mismo punto. La propiedad fuerte implica la propiedad de Markov puesto que podemos fijar un tiempo de parada en cada estado $\tau_i = \{n : X_n = x_i\}$ luego para $m = 1$ podemos obtener en cada paso

$$P(X_{\tau_i+1} = x_{\tau_i+1} | X_{\tau i} = x_i) = P(X_1 = x_{\tau_i+1} | X_0 = x_i).$$

Definición 1.5. *Una ecuación diferencial estocástica es una ecuación donde aparecen uno o más procesos estocásticos.*

Definición 1.6. *Para una variable aleatoria X, su asimetría $\gamma(X)$ se define como*

$$\gamma(X) = \frac{E[(X - EX)^3]}{Var(X)^{3/2}} = \frac{\mu_3(X)}{\sigma(X)^3}$$

donde EX es la esperanza de X, $Var(X) = E(X - EX)^2$ su varianza y $\kappa_n(X)$ es el enésimo cumulante. Es igual la estandarización del tercer momento central.

Definición 1.7. *Los cumulantes son cantidades asociadas a una distribución que proporcionan una alternativa a los momentos. El enésimo cumulante de X se define como*

$$\kappa_n(X) = \frac{1}{i^k}\left[\frac{d^n}{d\lambda^n} ln\varphi_X(\lambda)\right]_{\lambda=0}$$

donde $\varphi_X(\lambda) = Ee^{i\lambda X}$ es la función característica y ln se refiere a la rama principal del logaritmo. El cumulante $\kappa_n(X)$ está bien definido si $E[X]^k < \infty$. Los cumulantes son los coeficientes del polinomio de Taylor de $ln\varphi(\lambda)$ evaluado en 0 y pueden ser expresados en términos de los momentos estándar de X

$$\kappa_n(X) = \mu_n(X) + P(\mu_1(X), ..., \mu_{n-1}(X)).$$

La asimetría la podemos calcular con los cumulantes de manera similar a como lo hemos mostrado antes ya que los primeros tres cumulantes coinciden con los tres primeros momentos centrales.

$$\gamma(X) = \frac{\kappa_3(X)}{\kappa_2^{3/2}(X)}.$$

1.3. Objetivos y estructura

En este trabajo aunaremos y expandiremos varios resultados sobre la asimetría positiva que aparece en las distribuciones de los tiempos de incubación de diversas enfermedades. Se comenzará con los trabajos ya citados de Bertrand Ottino-Löffler para introducir el tema a través de modelos más aplicados con el propósito acabar entendiendo como se parece este problema al problema del coleccionista de cupones bajo algunas condiciones y se continuará ahondando en el trabajo de Yuri Bahktin para llevar el análisis a un clase más amplia de problemas gracias a los modelos de difusión.

El trabajo cuenta con una aproximación a las simulaciones realizadas para Ottino-Löfler tratando de observar nosotros mismos si efectivamente emerge la característica forma en las funciones de densidad de las distribuciones de los periodos de incubación y después el análisis matemático riguroso del modelo utilizado. A continuación se presenta de manera muy detallada el teorema principal de *Universal Statistics of Incubation Periods and Other Detection Times via Diffusion Models* pretendiendo que el lector encuentre un trabajo mayoritariamente autocontenido para todo aquel iniciado en matemáticas aunque no haya estudiado antes los procesos estocásticos.

El primer artículo no solo motivó el segundo sino que el estudio de ambos resultados juntos permite tener una visión amplia sobre el tema, con diferentes enfoques desde lo discreto hasta lo continuo y haciendo patente una vez más cómo las matemáticas permiten al ser humano generar respuestas y predicciones para problemas de toda clase.

2. Aproximación a los resultados de Ottino-Loffler

Recordemos que un $grafo$ es un conjunto de vértices y aristas que representa un modelo mediante nodos (vértices) y las relaciones entre ellos (aristas). La idea principal de este artículo consiste en modelizar un organismo como un conjunto de nodos sanos conectados entre sí con diversas geometrías, es decir, con diferente número de aristas y extremos de las mismas; a continuación se sustituye algún nodo por un invasor que puede representar un virus, una bacteria o incluso una célula cancerígena.

Es importante recalcar que no se está tratando en ninguno de los dos artículos principales de modelizar el contagio de una población. Este modelo intenta entender como se extiende una enfermedad dentro de un organismo, no la manera en que la que se expande un virus de individuo en individuo. El modelo podría incluso no ser preciso para virus ya que un solo virus es capaz de infectar múltiples células en oposición a a las bacterias o células cancerosas que van desplazando o destruyendo más o menos al ritmo que se reproducen. Aún así, la tendencia a esa característica asimetría positiva aparece también en tiempos de incubación de procesos víricos como en el caso de la COVID-19 [13].

En Ottino-Löffler et al. [11] los autores exponen los resultados de simulaciones enormes donde juegan el número de nodos, la geometría de la red, la heterogeneidad de los organismos, la adaptabilidad o capacidad de reproducción del agente infeccioso y en esta sección, además de explicar y detallar sus resultados, comenzaremos con una recreación de estas simulaciones hechas por el estudiante. Pese a la menor escala debido a la limitación computacional, es sorprendente ver como estos resultados emergen desde que se modelizan apenas 100 individuos y no hacen más que hacerse más obvios cuando se aumenta el número de ellos.

2.1. Simulaciones en MATLAB

Para replicar los resultados de Ottino-Löffler et al. [11] vamos a usar el mismo modelo, el modelo de Moran.

En este modelo hablaremos de invasores y nodos sanos para referirnos a los n vértices de un grafo unidos entre sí por aristas que representan si dos nodos son adyacentes o más coloquialmente, vecinos. Tras colocar en uno de los nodos al invasor, en cada paso temporal siguiente se actualiza según unas reglas predefinidas el estado de la red. En este caso he usado el modelo de Moran con la regla Bd (Birth-death).

En cada paso siguiente a iniciar la simulación realizamos dos etapas:

- En primer lugar se selecciona un nodo que según nuestro modelo morirá o se reproducirá. Como la regla es $Birth$, nuestro nodo se reproducirá. Este paso puede ser totalmente aleatorio o en función de un parámetro llamado normalmente $fitness$ que se puede traducir como adecuidad. Un nodo tiene probabilidad de ser escogido

$$P(escoger\ nodo\ j) = \frac{fitness_j}{\sum_{i=1}^{n} fitness_i}$$

Un nodo sano será totalmente neutro para sus vecinos en el mismo sentido que una célula sana si bien no infecta a su vecina, tampoco hace nada para detener un invasor. De este modo tiene sentido pensar que un invasor tiene que ser más «adecuado», es decir, con mayor $fitness$ que un nodo sano, para causar síntomas en este sistema.

- En segundo lugar, de entre los vecinos del nodo anterior, de nuevo aleatoriamente o en base a su $fitness$ alguno es seleccionado para morir o nacer. En nuestro caso, *death* con minúscula quiere decir que este nodo adyacente será aleatorio y morirá para ser reemplazado por la descendencia del paso anterior.

(a) Grafo completo. (b) Grafo circular.

Figura 2: Geometrías empleadas en las simulaciones.

Con el modelo planteado, se han escrito varios programas de MATLAB de los que algunos se adjuntarán al final de esta sección y en un apéndice.

Se presentan a continuación las resultados de simular dos poblaciones de quince mil organismos conformados por cien nodos la primera y treinta nodos la segunda, conectados con geometrías diferentes: un grafo completo (todos los nodos conectados) y un anillo donde cada uno solo tiene de vecinos al siguiente y al anterior formando un anillo. Como es lógico, un invasor tiene que ser mucho más agresivo, de mayor $fitness$, en una red donde está conectado con menos nodos puesto que si bien es tan probable ser escogido a igual número de nodos sanos e invasores con respecto al grafo completo, estar conectado con menos nodos hace más probable no contagiar nuevos nodos mediante la regla *death*.

Recordemos que estamos buscando explicar la distribución en los tiempos de incubación de un invasor dentro del sistema. Este periodo es el tiempo que tardan en aparecer los síntomas desde la exposición. En estas simulaciones modelizamos la aparición de síntomas como el momento donde los invasores superan a los nodos sanos, los individuos son homogéneos. El número de pasos es siempre suficiente para que un individuo que no llegue a desarrollar a síntomas hasta muy tarde quede recogido y no se han representado aquellos que no los desarrollen porque el estudio no trata sobre la virulencia del agente sino sobre su tiempo de incubación en los individuos donde prospera. También podemos modelizar estados de salud diferentes donde los individuos tienen diferentes umbrales para la aparición de síntomas. Estamos introduciendo en el sistema diferencias que a priori se relacionaban con la característica forma de la distribución pero como se ha mostrado antes, incluso sin este tipo de factores, obtenemos estos datos.

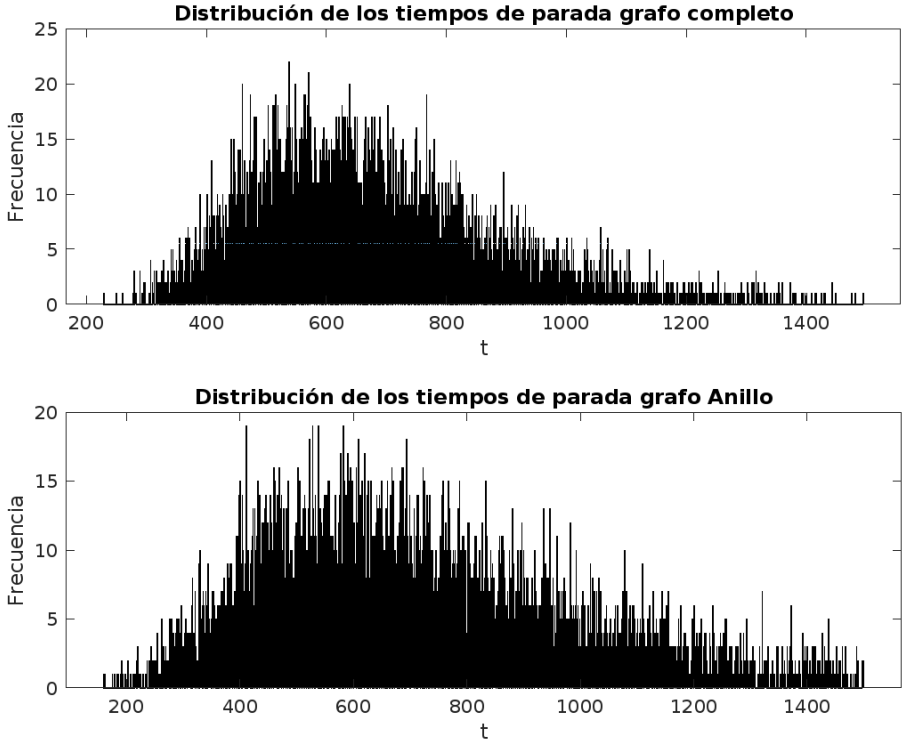

Figura 3: Pese a estar achatados para colocar ambos en una misma figura, las dos muestran una clara asimetría postiva.

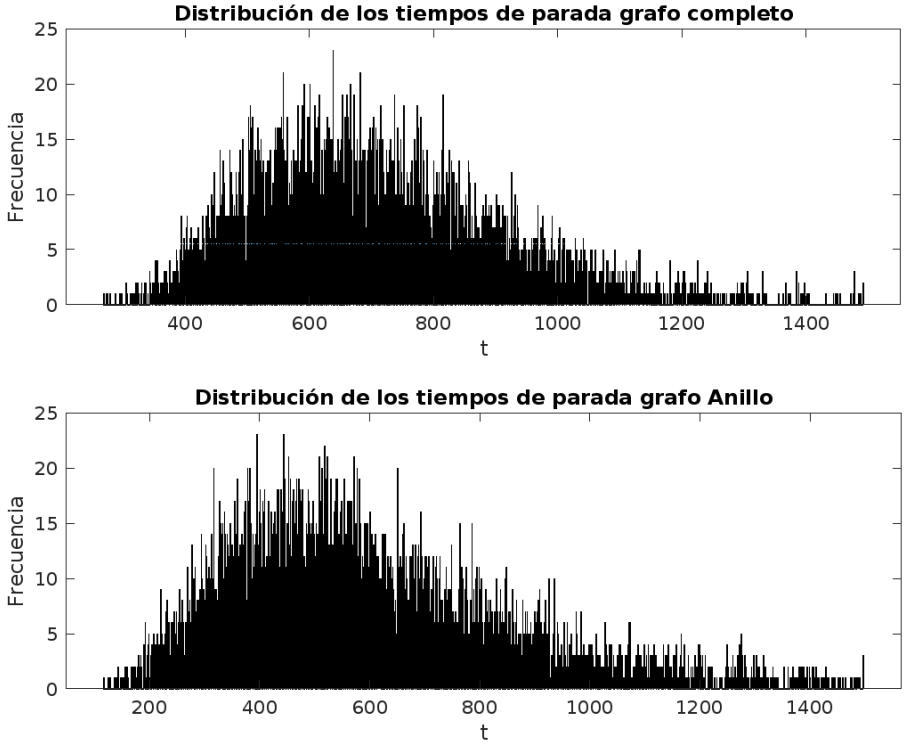

Figura 4: Con un rango de umbrales para la infección seguimos teniendo en ambas geometrías una notable asimetría.

El programa más sencillo es el caso del grafo completo con un umbral de síntomas preestablecido. En este caso umbral es la existencia de más nodos patógenos que sanos:

```
nodos = zeros(N, 1)+1;

x(randi(N)) = r; % Primer invasor, fitness r
hetN=randi(N/10);
for t = 1:1500
     % Paso 1: regla Birth

     fitness = nodos;
     fitness(nodos == 0) = 1; % nodos sanos fitness 1
     p = fitness / sum(fitness);
     i = randsample(N, 1, true, p);

     % Paso 2: regla death

     neighbors = setdiff(1:N, i);
     j = neighbors(randi(N-1));
     nodos(j) = nodos(i);

     if sum(nodos == 1) < N/2
          results(sim) = t;
          break;
          end
end
```

Este programa hace una simulación siguiendo el modelo descrito y permite calcular con pequeños cambios o añadidos las trayectorias de muchas simulaciones variando los parámetros y las condiciones.

Pasemos a ver si este comportamiento es explicable analíticamente o si por el contrario no es tan aventurado sugerir que emerge principalmente por heterogeneidad, a pesar de haber hecho simulaciones sin ella.

2.2. Resultados analíticos

En Ottino-Löffler et al. [11] se explica mediante dos mecanismos que la característica forma que aparece es muy similar a la distribución *lognormal*:

Definición 2.1. *Una distribución se dice log-normal cuando su logaritmo está normalmente distribuido. Dada una variable aleatoria X normalmente distribuida, e^X tendrá una distribución log-normal.*

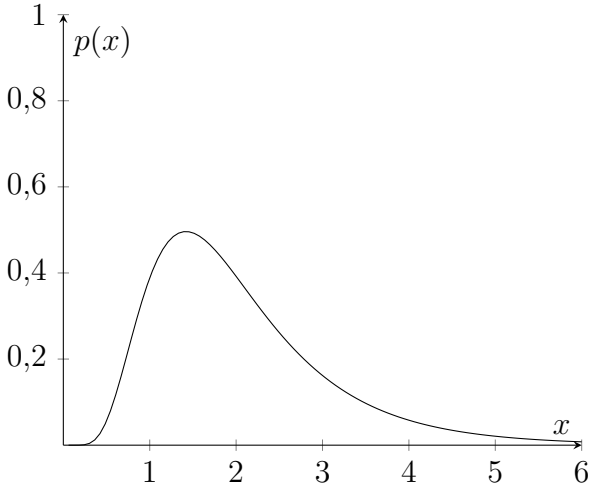

Figura 5: Función de densidad de una distribución log-normal con parametros $\mu = 0{,}6$ y $\sigma = 0{,}5$.

Estas dos observaciones emergen a medida que variamos el parámetro $fitness\ r$. Hemos explicado que Bd, el modelo escogido, tiene dos reglas: escogemos un nodo en base a su $fitness$ y después se reproduce reemplazando a alguno de sus vecinos escogido aleatoriamente. En un grafo completo de N nodos, la probabilidad de añadir un nuevo invasor tras tener m invasores es

$$p_m = P(seleccionar\ invasor)P(vecino\ escogido\ sano) = \frac{mr}{mr + (N-m)}\frac{N-m}{N-1}$$

que cuando $r \longrightarrow \infty$ tiende a $\frac{N-m}{N-1}$. El tiempo T que tarda un invasor en hacerse con la red es por tanto la suma de tiempo que tarda en pasar del estado donde hay m invasores a $m+1$ para $m = 1, 2, ..., N-1$. Para pasar de m a $m+1$ en t pasos, tiene que no haber sucedido en los $t-1$ pasos previos por tanto la probabilidad es

$$p_m(1-p_m)^{t-1}.$$

Observamos con esto que el tiempo que tarda en haber un nuevo invasor es una variable aleatoria geométrica y por tanto el tiempo de invasión total T es una suma de variables aleatorias

$$T = \sum_{m=1}^{N-1} Geo(p_m)$$

Veamos que este problema es idéntico al caso del coleccionista de cupones descrito ampliamente en P.Erdös, A. Rényi: On a classical problem of probability theory [2]. Este famoso problema se pregunta cuántos cromos indistinguibles (que se revelan al ser comprados) tenemos que comprar para tener los N cromos de una colección. Es decir, cual es la probabilidad de tener los N cromos habiendo comprado T paquetes.

Recordemos que habíamos modelizado el problema con la regla $Birth - death$ pero el análisis anterior revela que estudiar cuanto tarda un invasor en ocupar todos los nodos (o llegar a número de nodos que consideramos el umbral de los síntomas) escogiendo un nodo y contagiando un vecino suyo es igual que estudiar el tiempo que tardamos en añadir $N-1$ invasores con probabilidad $p_m(1-p_m)^{t-1}$ en cada paso pero, en cada instante hasta T, el número de invasores aumenta a lo sumo en uno por tanto es lo mismo que estudiar cuanto se tarda en pasar por los $N-1$ nodos escogiendo uno en cada paso y devolviendo siempre un invasor en su lugar que es exactamente el problema del coleccionista.

Adaptando por tanto los resultados de P.Erdös, A. Rényi [2] podemos encontrar la distribución asintótica cuando N se hace grande. Conocida su media $\nu = \sum_m p_m^{-1} = Nlog(N) + N\gamma$ podemos normalizar y encontrar que

$$\frac{T-\nu}{N} \xrightarrow{\mathcal{L}} Gumbel(-\gamma, 1)$$

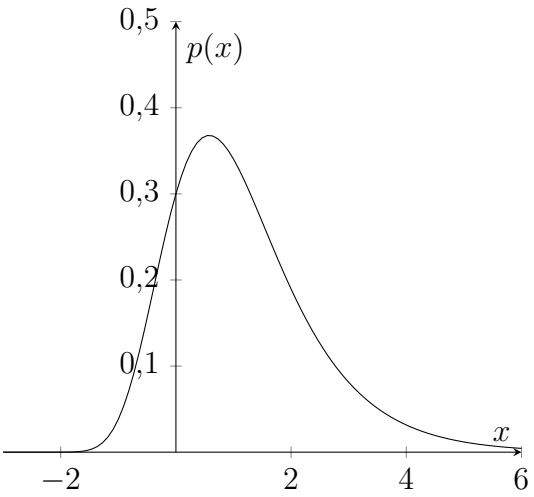

Figura 6: Función de densidad de una distribución Gumbel de parámetros $\mu = \gamma = 0.5772$, $\beta = 1$

A γ se la conoce como constante de Euler-Mascheroni. Tras esto, apreciamos que emerge de nuevo el mismo comportamiento asimétrico en la función de densidad.

Veamos ahora como se comporta el sistema cuando $r \longrightarrow 1$ y como aparece de nuevo el mismo patrón aunque por razones diferentes.

De nuevo, la probabilidad en cada paso de sumar un invasor es

$$p_m^+ = \frac{mr}{mr + (N-m)} \frac{N-m}{N-1}$$

y la probabilidad de perderlo es

$$p_m^- = \frac{N-m}{mr+(N-m)}\frac{m}{N-1}$$

así que la probabilidad de que en el siguiente paso aumente el numero de invasores es

$$q := \frac{p_m^+}{p_m^+ + p_m^-} = \frac{r}{r+1}$$

Esto define un camino aleatorio con $prob(+1\ invasor) = q$ y $prob(-1\ invasor) = 1-q$. Como buscamos estudiar la aparición de síntomas modelizándolo como el momento en el que el número de invasores supera un cierto umbral N y sin pasar por tener 0 invasores, en el caso límite de $fitness = r = 1$ donde tenemos un invasor neutral la probabilidad de aumentar un invasor es exactamente $1/2$. Con esto, definimos un camino aleatorio con tiempos de parada en 0 y N. Sea X_n el número de invasores tras n pasos. Con esto tenemos que

$$X_n = \sum_{i=1}^{n} X_i$$

donde $X_i \in \{-1, +1\}$. Recordando la introducción, definimos la variable aleatoria $T = \{n : X_n = N\}$ y el estudio de sus momentos [11]

$$\mu_1 = E[T] = \frac{N^2-1}{3}$$

$$\mu_2 = E[T] = (E[(T-E[T])^2])^{1/2} = \frac{7N^4 - 20^2 + 13}{45}$$

$$\mu_3 = E[(T)^3] = \frac{31N^6 - 147N^4 + 189N^2 - 73}{315}$$

permite calcular la asimetría sabiendo como se relacionan los momentos centrales con los momentos

$$\gamma(T) = \frac{\mu_3}{\sigma^3} = \frac{E[(T-ET)^3]}{Var(T)^{3/2}} = \frac{E[T^3] - 3E[T]\sigma(T) + 2E[T]^2 +}{Var(X)^{3/2}} \approx 1,8$$

Vemos por tanto que el comportamiento asintótico de la distribución también produce una asimetría positiva cuando el invasor es neutro.

Hasta ahora hemos visto modelos con pasos de tiempos discretos y el espacio de estados posibles que pueden tomar X_N también es discreto. Cabría preguntarse si podemos generalizar a una clase de modelos con tiempo continuo y trayectorias continuas. Nótese que si eso fuera posible, volver al caso discreto no consistiría más que estudiar la variable aleatoria $X(t)$ en instante concretos de $[0, T]$ equiespaciados para simular pasos discretos y tamaños del paso idénticamente distribuidos ya que en la definición 1.1 observamos que Z depende solo del lapso.

3. Resultados de Yuri Bahktin

En la sección anterior hemos expuesto las conclusiones de Ottino-Löffler et al. [11]. Si bien son muy rigurosas, puede surgir la duda sobre si sus resultados no se limitan a una serie de modelos demasiado rígidos y sobre qué condiciones del sistema son las que garantizan qué resultados. El objetivo que se plantea en el artículo de Yuri Bahktin es el de ampliar los resultados a una clase más amplia de modelos basadas en el tiempo de parada para modelos de difusión unidimensionales.

3.1. Modelos de difusión

De nuevo, sea también ahora $X(t)$ una variable aleatorio real que representa el nivel de infección de un individuo. Como antes, estamos interesados en el tiempo en que se alcanza un cierto nivel R y antes de que se anule la variable, por tanto estudiamos el tiempo durante el cual la variable se mueve en el intervalo $[0, R]$. 0 corresponderá a un individuo completamente sano mientras que R será la aparición de síntomas. Todos los valores intermedios representan estados del individuo donde el patógeno permanece latente. Se asume también que existe $x_0 \in (0, R)$ tal que el sistema inmunitario no detecta la infección hasta que $X(t) \geq x_0$. Tras este momento, X se comporta como un proceso de Markov homogéneo.

Definición 3.1. *Un proceso de Markov es homogéneo si la probabilidad de los estados tras el instante siguiente es independiente del paso, es decir*

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)$$

Definición 3.2. *Un proceso estocástico $X(t), t \geq 0$ es de Markov, homogéneo y continuo si para una secuencia de tiempos $0 \leq t_1 \leq ... \leq t_n$ y una serie de estados en esos instantes $i_0, i_1, ..., i_n$ se cumple que*

$$P(X(t+t_n) = j | X(0) = i_0, X(t_1) = i_1, ..., X(t_n) = i_n) = P(X(t+t_n) = j | X(t_n) = i_n)$$

Bajo las condiciones anteriores, X es una solución de una ecuación diferencial estocástica (EDE) de la forma

$$dX(t) = b(X(t))dt + \sigma(X(t))dW(t). \quad (1)$$

La función $b(x)$ la denominaremos deriva ($drift$ en inglés). En nuestro problema podría modelizarse como el efecto combinado de la infección y la respuesta inmunitaria. La idea sería muy similar a las reglas $birth-death$ del segundo capítulo de este trabajo: $B(x)$ modelizaría la expansión del invasor, $D(x)$ su decrecimiento de modo que $b(x) = B(x) - D(x), x \in [0, R]$. Indaguemos un poco para ver de donde sale (1) y entender por qué nuestro modelo es un buen modelo para el problema y es solución de dicha ecuación.

Recordemos que una camino aleatorio continuo es de la forma $X(t+h) = X(t) + Z(h)$ donde Z es una variable aleatoria que puede estar distribuida de cualquier manera siempre que sea independiente del tiempo, de la misma manera que lo era

el desplazamiento en cada nuevo paso en un camino aleatorio discreto. En realidad, Z puede ser una función de t siempre que esté idénticamente distribuida para todo instante. Idealmente lo que querríamos sería poder controlar los parámetros del modelo: el tamaño del paso es ahora la variación de invasores por unidad de tiempo por lo que consecuentemente, al ser un proceso aleatorio y en relación al tiempo transcurrido, no tendremos un variación fija sino en función de como esté distribuida Z. El tamaño del paso depende ahora de la variación en el tiempo, ya que $Z(t)$ solo depende del estado presente. Si suponemos por un segundo que Z está normalmente distribuida, con media 0 y varianza 1, la varianza de $X(t+h) - X(t)$, que es la variable aleatoria que representa el paso dado en cada lapso, para un paso h de tamaño pequeño tiene que ser exactamente h así que debemos multiplicar por la que sería la desviación típica. Con esto obtenemos

$$X(t+h) = X(t) + Z(t)\sqrt{h}.$$

Ahora nos damos cuenta de que para un tamaño del paso cada vez más pequeño, lo que obtenemos es una ecuación diferencial

$$\Delta X(t) = Z(t)\sqrt{\Delta t}.$$

Si $Z(t)$ estuviese distribuida de cualquier otra manera, su media μ se convertiría en lo que antes hemos llamado deriva, la tendencia general de los pasos de una trayectoria de la partícula o en nuestro caso el número de invasores, a moverse una cantidad determinada por unidad de tiempo. Añadiendo esto al modelo obtenemos y recordando que la varianza tiene que coincidir con la de Z

$$\Delta X(t) = \mu\Delta t \ + \ \sigma(Z)Z(t)\sqrt{\Delta t}$$

que formalizado queda

$$dX(t) = \mu dt \ + \ \sigma Z(t)\sqrt{dt}.$$

Definición 3.3. *Se dice que un proceso estocástico* $\{W(t) : t \geq 0\}$ *es un proceso de Wiener si:*

(I) $W(t), t \geq 0$ *tiene incrementos independientes estacionarios. Esto es, para dos intervalos disjuntos* $(t_1, t_2), (t_3, t_4)$*, las variables aleatorias* $W(t_1) - W(t_2)$*,* $W(t_3) - W(t_4)$ *son independientes.*

(II) W *tiene incrementos normales:* $W(t+h) - W(t) \sim N(0, h)$.

(III) W *tiene trayectorias continuas.*

(IV) $W(0) = 0$.

Recapitulando, si queremos introducir aleatoriedad en el sistema, igual que lo haríamos para el movimiento Browniano de una partícula, utilizamos aquí un proceso de Wiener y su derivada dW es la componente de su influencia en el número de invasores en cada instante. A este término se le suele llamar ruido blanco. $\sigma(X(t))$

es el coeficiente de amplitud de esta componente aleatoria e igual que hemos hecho antes y sabiendo que la varianza de un proceso de Wiener estándar es precisamente t, la multiplicamos por este término ya que no es otra cosa que la varianza de las trayectorias.

También como anteriormente, el periodo de incubación será el tiempo de parada de esta variable aleatoria que hemos descrito

$$\tau = \{t : X(t) = 0 \ o \ X(t) = R\}$$

En particular nos interesará el primer momento tal que $X(\tau) = R$, que será el periodo de incubación.

Definición 3.4. *Un modelo de difusión es un proceso estocástico sobre un espacio continuo que posee la propiedad de Markov y para los que las trayectorias son continuas casi seguro. Más formalmente, un proceso de Markov que toma valores en unn intervalo I que cumple*

$$\lim_{h \longrightarrow 0^+} \frac{1}{h} P(\{|X(t+h) - x| > \epsilon | X(t) = x\}) = 0 \ \forall x \in I.$$

Bajo las condiciones explicadas, $X(t)$ es modelo de difusión.

3.2. Teoremas Centrales

Llamaremos Γ a la aparición de síntomas, es decir $\Gamma = \{X(\tau) = R\}$ para el τ primer tiempo de parada. Γ es el espacio de las trayectorias que pasan por R antes que por 0.

Teorema 3.5. *Bajo las condiciones sobre Γ anteriormente descritas, la distribución es positivamente asimétrica, es decir, $\gamma(\tau) > 0$.*

Si observamos las definiciones de simetría y cumulantes, nos daremos cuenta de que esto es una consecuencia directa del signo del tercer cumulante y de manera más amplia se cumple que

Teorema 3.6. *Bajo las condiciones sobre Γ anteriormente descritas,*

$$\kappa_n(\tau) > 0 \ n \in \mathbb{N}.$$

Demostración. La distribución del proceso de difusión condicionado a Γ coincide con la de la solución X de una nueva ecuación diferencial estocástica

$$dX(t) = \tilde{b}(X(t))dt + \sigma(X(t))dW(t) \quad (2)$$

En esta nueva ecuación, σ es la misma función que en (1) y

$$\tilde{b}(x) = b(x) + \sigma^2(x)\frac{h'(x)}{h(x)}, \ 0 < x < R,$$

donde $h(x)$, $x \in [0, R]$ denota la probabilidad de Γ cuando la difusión (1) haya comenzado en x. Más formalmente, en un intervalo de tiempo finito $[0, T]$,

$$h(x) = P(X(T) \in \Gamma | X(t) = x).$$

A esto se le conoce como transformación h de Doob.

Si lo que se pretende es entender el comportamiento de estos modelos, su fundamento y los resultado que emergen de su estudio pormenorizado, conviene pausar la prueba un segundo para entender esto último. Esta transformación surge de que el espacio de estados que puede tomar $X(t)$ ya no es $\{X(t) : t \geq 0\}$ puesto que nos quedamos solo con las trayectorias de Γ. La pregunta sería si el proceso $X(t)$ condicionado por Γ sigue siendo una difusión. Bajo las condiciones que hemos descrito y por el teorema 6.6 de Bakhtin and Święch 2016 [9] resulta que estamos ante un modelo de difusión que es solución de (2). Para tener una intuición sobre la aparición de esta nueva $\tilde{b}$ veamos quienes son las probabilidades condicionadas a $X(t) = x$. Añadimos el tiempo como variable porque aunque sabemos que la distribución del paso no depende del instante sino de la posición, nos es útil para tener una notación más clara. Buscamos conocer como se comporta ahora la difusión cuanto transcurre un tiempo s pequeño

$$P(X(t+s) = y | X(t) = x, X(T) \in \Gamma) = \frac{P(X(t+s) = y; X(T) \in \Gamma | X(t) = x)}{P(X(T) \in \Gamma | X(t) = x)} =$$

$$= \frac{P(X(t+s) = y \cap X(T) \in \Gamma \cap X(t) = x)}{P(X(T) \in \Gamma | X(t) = x)} = P(X(t+s) = y | X(t) = x) \frac{h(t+s, y)}{h(t, x)}.$$

Como hacíamos para llegar a la ecuación diferencial, estamos interesados el cambio en la deriva por unidad de tiempo del modelo difusión luego aparecerá en el nuevo término su derivada en x que es una función de t. Realmente esto no es del todo riguroso pero puede resultar menos abrumador para seguir la prueba. Para ser más académicos, conviene explicar y adelantar para después, que modelos como los que son solución de una EDE como (1) constituyen un tipo de ecuación diferencial estocástica conocidos como difusiones de Itô. Para esta clase de modelos se puede calcular lo que se denomina como *generador infinitesimal* que es un operador diferencial que describe como evoluciona un proceso estocástico de Markov con tiempo continuo. En el caso unidimensional es

$$\mathcal{L}(f(x)) = b(x) f'(x) + \frac{1}{2} \sigma^2(x) f''(x)$$

Para los detalles consultar el capítulo 15 de Karlin y Taylor [5]. La fórmula general para calcular el generador de un proceso estocástico es

$$\mathcal{G} f = \lim_{t \to^- 0} \frac{E[f(X(t)) | X(0) = x] - f}{t}.$$

Conocidos $\mathcal{L}$ y la definición, podemos calcular el generador del condicionado [5] [8]

$$\mathcal{G} f(x) = \mathcal{L} f(x) + \sigma^2(x) \frac{h'(x)}{h(x)} f'(x) = b(x) f'(x) + \frac{1}{2} \sigma^2(x) f''(x) + \sigma^2(x) \frac{h'(x)}{h(x)} f'(x)$$

Naturalmente se parece a la intuición que se daba anteriormente. El nombre de generador no es casual ya que a partir del generador podemos describir la ecuación: la nueva deriva será lo que acompañe a f', obteniendo (2) como se buscaba.

Volvamos a la prueba. Para todo $x \in (0, R)$ llamaremos P_x a la distribución de la solución de (2) con condiciones iniciales $X(0) = x$. A la esperanza respecto de P_x la nombraremos E_x. Bajo estos supuestos, los tiempos de parada son finitos para la ecuación inicial y por tanto también lo son para la ecuación condicionada. Además, si definimos

$$\tau_y = inf\{t \geq 0 : X(t) = y\}$$

entonces para cualquier $y \in (0, R]$ las funciones

$$\alpha_n(x, y) = E_x \tau_y^n, \ n \in \{0\} \cup \mathbb{N}, \ 0 < x \leq y,$$

son suaves en $x \in (0, y]$ hasta y, satisfaciendo para cualquier n natural las ecuaciones en derivadas parciales

$$\mathcal{L}\alpha_n(x, y) = -n\alpha_{n-1}(x, y), \ 0 < x \leq y$$

donde $\mathcal{L}$ es el generador del semigrupo asociado a la difusión (2). Denotemos ahora al n-ésimo cumulante de τ_y bajo P_x, es decir el enésimo cumulante de la variable aleatoria τ_y de un proceso estocástico que resuelve la ecuación diferencial estocástica con distribución P_x, como $\kappa_n(x, y)$, $0 < x \leq y \leq R$. Como$\kappa_n(R, R) = 0$, podemos escribir

$$\kappa_n(x, R) = -(\kappa_n(R, R) - \kappa_n(x, R)) = -\int_x^R \frac{d}{dy}\kappa_n(y, R) \quad (3)$$

La propiedad fuerte de Markov implica que bajo P_x los tiempos de para $\tau_{x \leq y \leq R}$ forman un proceso con incrementos independientes y si $x \leq y_1 \leq y_2 \leq R$, entonces la distribución de $\tau_{y_2} - \tau_{y_1}$ bajo P_x no depende de $x \in (0, y_1)$. Combinado con la suavidad de κ_n, obtenemos

$$\frac{d}{dy}\kappa_n(y, R) = \frac{d^-}{dy}\kappa_n(y, R) = \left.\frac{d^-}{dz}\kappa_n(z, y)\right|_{z=y}.$$

Usando la definición 1.6 y la definición de los α_n obtenemos que

$$\kappa_n(z, y) = \alpha_n(z, y) + P_n(\alpha_1(z, y), ..., \alpha_n - 1(z, y))$$

donde cada monomio que forma P_n es al menos de segundo orden. Como $\alpha_1(y, y) = ... = \alpha_{n-1}(y, y) = 0$, tenemos que la derivada de cada uno de los monomios con respecto a z en $z = y$ es igual a 0 y por tanto

$$\left.\frac{d^-}{dy}\kappa_n(y, R)\right|_{z=y} = \left.\frac{d^-}{dz}\alpha_n(z, y)\right|_{z=y} \leq 0,$$

puesto que $\alpha_n(z, y)$ es al menos no creciente en z. Veamos que la desigualdad es estricta. Sea un $z_0 \in (0, y)$ y nótese que para $z \in (z_0, y)$ se tiene que

$$\alpha_n(z, y) \geq u(z)\alpha_n(z_0),$$

donde $u(z)$ es la probabilidad de que el proceso comenzado en z pase por z_0 antes que por y. Esta función satisface la ecuación

$$b(z)u'(z) + \frac{1}{2}\sigma(z)u''(z) = 0,\ x \in [z_0, y] AQUI\ NO\ SERIA\ \tilde{b}??$$

con condiciones de contorno

$$u(z_0) = 1$$
$$u(y) = 0.$$

Como u es no negativa y se hace 0 en su extremo derecho, se debe tener que su derivada en y es, al menos, no positiva, $u'(y) \leq 0$, pero si $u'(y) = 0$, por la unicidad de soluciones de las ecuaciones diferenciales de segundo orden regulares, esto implica que $u \equiv 0$. Esto naturalmente no puede ser porque u no cumpliría las condiciones de contorno descritas así que $u'(y) < 0$. De esto concluímos que

$$\frac{d^-}{dz}\alpha_n(z, y)\bigg|_{z=y} < 0$$

luego volviendo a (3)

$$\kappa_n(x, R) = -\int_x^R \frac{d}{dy}\kappa_n(y, R) = -\int_x^R \frac{d}{dy}\alpha_n(y, R) > 0.$$

□

4. Conclusiones

Modelizar algo tan complicado como la aparición de síntomas de una determinada enfermedad es sin duda ambicioso. Más aún cuando se hace con reglas sencillas que parecen ignorar muchos factores internos y externos que pueden darse. Además el estudio está motivado por las observaciones empíricas que de nuevo pueden agregar muchos errores desde el momento que se considera que ha sucedido la inoculación hasta que se recoge el dato con el tiempo transcurrido hasta la aparición de síntomas. No obstante eso no tiene que impedir tratar de arrojar algo de luz sobre problemas complicados mediante herramientas que sabemos controlar y que una vez desarrolladas ciertamente parecen estar bien encaminadas para permitirnos conocer mejor el mundo e incluso tomar decisiones entendiendo por ejemplo, en qué momento podemos tener un número elevado de personas contagiadas desarrollando síntomas de una enfermedad.

En todas las fuentes que se han usado para este trabajo hay más resultados y se podría detallar *ad infinitum* cada definición, cálculo, lema o paso de demostración pero se ha intentado seleccionar aquello que permita seguir el hilo conductor marcado por los objetivos sin descuidar el rigor que caracteriza a nuestra facultad.

5. Apéndice

5.1. Simulaciones en Matlab

Aquí mostramos algunos ejemplos de los códigos usados para las simulaciones. El intérprete de MATLAB dentro de LATEX está inglés por lo que se han quitado las tildes de los comentarios. Código para grafo completo con N nodos.

```
x = zeros(N, 1)+1; % vector que inicia un individuo como un
                     % conjunto de nodos sin enfermedad

% Bucle para simular individuos
results = zeros(num_simulations, 1);
for sim = 1:num_simulations

    x(randi(N)) = r; % colocamos aleatoriamente el invasor
    hetN=randi(N/10);
    for t = 1:1500 % simulamos el paso del tiempo con ciclos donde el
                     invasor puede o no afectar a otros nodos

        % Paso 1: seleccionamos aleatoriamente un nodo en funcion
                     de su fitness
        fitness = x;
        %fitness(x == 0) = 1; % los nodos sanos tienen adaptibilidad 1
        p = fitness / sum(fitness);
        i = randsample(N, 1, true, p);
        % Paso 2: "contagiamos" alguno de sus vecinos
        % Como el grafo es completo, sus vecinos son todos

        neighbors = setdiff(1:N, i);
        j = neighbors(randi(N-1));
        x(j) = x(i);

        % Modelizamos la aparicion de sintomas como la existencia de mas
            % invasores que nodos sanos en el grafo. Este es el tiempo
            % de parada
        if sum(x == 1) < N/2
            results(sim) = t;
            break;
        end
    end
    x = zeros(N, 1)+1;
end
```

Código para grafo en forma de anillo con umbrales aleatorios dentro de un rango.

```
x = zeros(N, 1)+1;

results = zeros(num_simulations, 1);
for sim = 1:num_simulations
    x(randi(N)) = r; % colocamos aleatoriamente el invasor

    for t = 1:1500
        fitness = x; % Paso 1
        p = fitness / sum(fitness);
        i = randsample(N, 1, true, p);
        % Paso 2
         % Solo consideramos sus vecinos anterior y siguiente
        j = mod(i + randi([-1 1]), N);
        if j == 0 % Si j es 0, lo tratamos como el ultimo nodo
            j = N;
        end
        x(j) = x(i);
        % Umbrales aleatorios dentro de un rango del 10% de nodos

        if sum(x == 1) < (N/2-N/10+hetN)
            results(sim) = t;
            break;
        end
    end
    x = zeros(N, 1)+1;
end
```

Simulación de las dos geometrías propuestas con estados de salud diversos.

```
% Grafo Completo
% Parametros del modelo
N = 100; %nodos del grafo que representa cada individuo
r = 1.68; % adaptabilidad del patogeno (fitness)
num_simulations = 15000; % numero de individuos contagiados

% Simulamos y obtenemos los resultados
SimCompletoHet;

% Graficamos el histograma
subplot(2,1,1);
histogram(results(results>0), 'BinMethod', 'integers',
          'Normalization', 'count');
xlabel('t');
ylabel('Frecuencia');
title('Distribucion de los tiempos de parada grafo completo');

% Anillo
% Parametros del modelo
N = 30; %nodos del grafo que representa cada individuo
r = 2; % adaptabilidad del patogeno (fitness)
num_simulations = 15000; % n mero de individuos contagiados

SimCircularHet;

% Graficamos el histograma
subplot(2,1,2);
histogram(results(results>0), 'BinMethod', 'integers',
          'Normalization', 'count');
xlabel('t');
ylabel('Frecuencia');
title('Distribucion de los tiempos de parada grafo Anillo');

% Ajustamos el tama o de la figura y guardamos el resultado
fig = gcf;
fig.Position(3:4) = [800 600];
saveas(fig, 'histogramas15000het.png');
```

Referencias

[1] Philip E. Sartwell. "The distribution of incubation periods of infectious disease". En: *Oxford Journals* 51.3 (1950), págs. 310-318. DOI: 10.1093/oxfordjournals.aje.a119397.

[2] Paul Erdős y Alfréd Rényi. "On a classical problem of probability theory". En: *Publication of The Mathematical Institute of the Hungarian Academy of Sciences* 6 (1961), págs. 215-220.

[3] Emanuel Parzen. *Procesos estocásticos*. Madrid: Paraninfo, 1972.

[4] David R. Cox y Halbert D. Miller. *The theory of stochastic processes*. CRC Press, 1978.

[5] Samuel Karlin y Howard M. Taylor. *A Second Course in Stochastic Processes*. London: Academic Press, 1981.

[6] Mark I. Friedlin y Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wisseschaften 260. Berlin: Springer-Verlag, 1984.

[7] Thomas M. Liggett. *Continuous Time Markov Processes: An Introduction*. Los Angeles, CA: Department of Mathematics, UCLA, 1991. URL: https://www.math.ucla.edu/~tml/liggett_first24.pdf.

[8] Alexandre Thiéry. *Doob H-transforms*. https://linbaba.wordpress.com/2010/06/02/doob-h-transforms/. 2010.

[9] Samuel A. A. Monter y Yuri Bakhtin. "Normal forms approach to diffusion near hyperbolic equilibria". En: *Nonlinearity* 24.6 (2011), págs. 1883-1907. DOI: 10.1088/0951-7715/24/6/011.

[10] Yuri Bakhtin y Andrzej Swiech. "Scaling limits for conditional diffusion exit problems and asymptotics for nonlinear elliptic equations". En: *Transactions of the American Mathematical Society* 368.9 (2016), págs. 6487-6517. DOI: 10.48550/arXiv.1310.6023. arXiv: arXiv:1310.6023 [math.PR].

[11] Bertrand Ottino-Loffler, Jacob G Scott y Steven H Strogatz. "Evolutionary dynamics of incubation periods". En: *eLife* 6 (2017), e30212. DOI: https://doi.org/10.7554/eLife.30212.

[12] Yuri Bakhtin. "Universal statistics of incubation periods and other detection times via diffusion models". En: *Bulletin of mathematical biology* 81.5 (2019), págs. 1070-1088. DOI: https://doi.org/10.1007/s11538-018-00558-w.

[13] Conor McAloone et al. "Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research". En: *BMJ Open* 10.8 (2020). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2020-039652. URL: https://bmjopen.bmj.com/content/10/8/e039652.

[14] *Itô diffusion*. https://en.wikipedia.org/wiki/It%C3%B4_diffusion. A 25 de abril de 2023.