BIM 105
Fall 2024

Lab 9: Geometric Random Walks

Instructions: Read the prompts and use MATLAB to answer the questions. Submit your solution as a single script. A template script is available on Canvas.

Part 1. Modeling Population Changes

Consider a population with some population $N(0)$. Let $R(t)$ be a random variable that denotes the coefficient of population growth from time $t-1$ to t. In other words, the population at time t depends on the previous population with the following relationship:

$$N(t) = R(t)N(t-1),$$

where $R(t)$ is a random variable following some probabilistic model. In other words, $N(1) = R(1)N(0)$, $N(2) = R(2)R(1)N(0)$, $N(3) = R(3)R(2)R(1)N(0)$, and so on. For the purposes of this lab, the random variable R is independent with respect to time. In terms of $N(0)$, $N(t)$ can be written in product notation as:

$$N(t) = N(0)\prod_{t=1}^{t} R(t)$$

Now consider the population $N(t)$ in the log domain ($\log N(t)$). Recall that the log of a product can be written as the sum of logs, and therefore

$$\log N(t) = \log\left[N(0)\prod_{t=1}^{t} R(t)\right] = \log N(0) + \sum_{t=1}^{t} \log R(t)$$

From this perspective, we observe that the multiplicative changes in population over time become additive changes in the log domain. In fact, considering this model in the log domain simplifies how we can characterize some parts of the population's behavior. Because the changes in population are a sum of random variables in the log domain, computing the expected value of $\log N(t)$ at some time t becomes a relatively simple computation. It's just the initial population size ($\log N(0)$) plus the average growth at each time period, which is expected value of $\log R(t)$. For an arbitrary $\log R(t)$, the expected value is

$$\mathrm{E}[\log R(t)] = p_1 \log R_1 + p_2 \log R_2 + \cdots + p_K \log R_K$$

$$= \sum_{k} p_k \log R_k$$

where K is the number of values in R and p_k are the corresponding probabilities of sampling each value in R. Putting this together with the initial size $\log N(0)$ and the assumption of R's time-

independence, we can get an expression for the expected population size at time t, in the log domain as

$$\mathrm{E}[\log N(t)] = \log N(0) + t\,\mathrm{E}[\log R(t)]$$

The expected variance of $\log N(t)$ can be formulated in a similar way. Recall that the variance of a random variable can be written as $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$, meaning that the variance of $\log N(t)$ can be expressed as

$$\mathrm{Var}(\log N(t)) = t\,\mathrm{Var}(\log R(t)) = t\,(\mathrm{E}[\log R(t)^2] - \mathrm{E}[\log R(t)]^2)$$

To see how these equations work out in practice, let's do an example with a simple model of $R(t)$. Let $R(t)$ be a random variable with possible values $\{2, 0.6\}$ that occur with equal probability ($[0.5, 0.5]$). The expected value of $\log R(t)$ is $\mathrm{E}[\log R(t)] = (0.5)(\log 2) + (0.5)(\log 0.6) = 0.091$. Because this is a positive value, the population is expected to grow, on average. We can see the behavior when considering the expected log population size, which is

$$\mathrm{E}[\log N(t)] = \log N(0) + 0.091t$$

From the starting population size $\log N(0)$, we see that the population (on average) tends to increase over time.

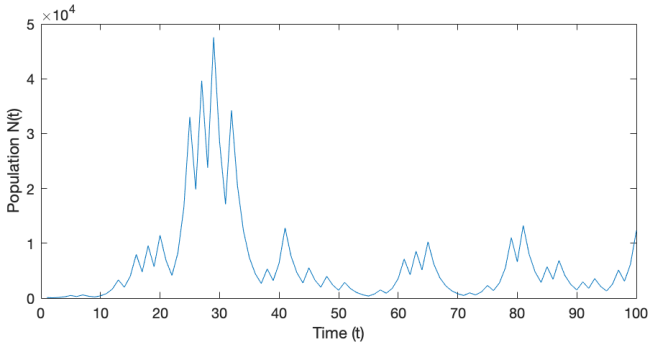
The population changes are random, however, so individual realizations of the system will deviate from this average. We can use MATLAB to generate population profiles over time that stem from this distribution of R. Starting with an initial population size, say, $N(0) = 100$, we need only to sample a growth coefficient at each time point and multiply the current population size by that coefficient. A convenient function that does this for an arbitrary discrete distribution R is given here as **`sim_geometric_population(N0, R, probs, t)`** where **`N0`** is the initial population size, **`R`** is the values of R to sample, **`probs`** contains the probabilities of each value in R, and **`t`** is the duration of the simulation.

```
function N = sim_geometric_population(N0, R, probs, t)

C = cumsum(probs);
N = NaN(t, 1);
N(1) = N0;

for i = 1:t-1
    Rn = R(1+sum(C(end)*rand>C));
    N(i+1) = Rn*N(i);
end

end
```

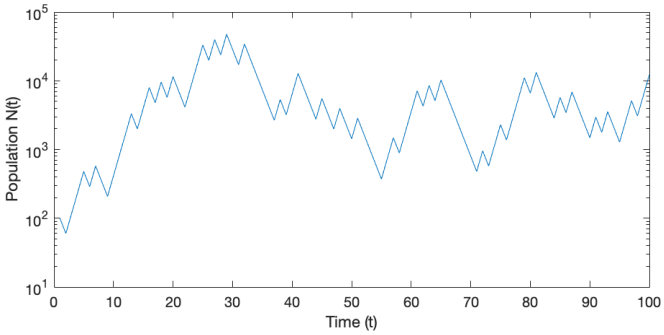
Then, for our example, we could perform a simulation using our values as:

```
N0 = 100;
t = 100;
R = [2, 0.6];
probs = [0.5, 0.5];
N = sim_geometric_population(N0, R, probs, t);
```

We then plot the trajectory with **`plot(N)`**, which gives a plot that will look something like this:



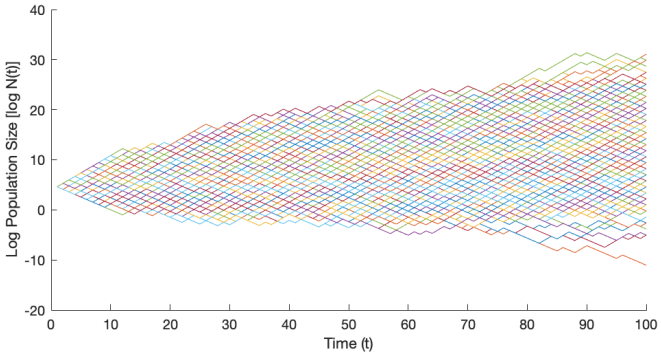
We can observe that the population sees large fluctuations of growth and depletion depending on which values of R are sampled. As we saw earlier, it's also helpful to plot the population size on a log-scale. Plotting the log of the population size with **`plot(log(N))`** gives us something that looks like:



In the log domain, we observe that that changes to population size are sequential additions of $\log R$ corresponding to which value of R was sampled.

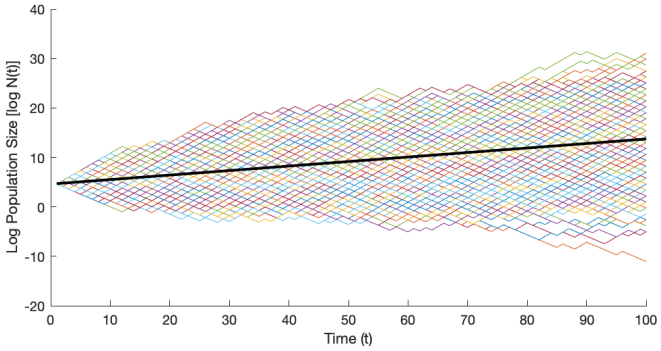
We can scale up the number of simulations to get a better view of the ensemble behavior of populations under this model. Let's use 1000 simulations of the population starting at $N(0) = 100$ and plot all of the trajectories together.

```
T = 100; Ns = 1000; N0 = 100;
R = [2, 0.6]; probs = [0.5, 0.5];
Nt = NaN(Ns, T);

for i = 1:Ns
    Nt(i,:) = sim_geometric_population(N0, R, probs, T);
end
plot(log(Nt)')
xlabel('Time (t)')
ylabel('Log Population Size [log N(t)]')
```



We see that, on average, the population size is slightly increasing, as we predicted from the expected value of $\log R$. How well do the simulations agree with our formula for $\mathrm{E}[\log N(t)]$? Let's add this line to the plot. We can call:

```
hold on
EX = sum(probs.*log(R)); % Expected value of R(t), slope of growth
y0 = log(N0); % Y-intercept
plot(1:T, y0+EX*(1:T), 'k', 'LineWidth', 2);
```



We see that the simulated populations are growing, on average, by the expected rate.

Question 1. Let $R(t)$ assume $\{1.2, 0.8\}$ with equal probability.

1a: What is the expected value of $\log R(t)$?

1b: What is the variance of $\log R(t)$?

1c: Simulate 1000 populations over 100 generations from this model starting with $N(0) = 100$ and plot the log-population over time.

1d: Add to your plot from Question 1c the expected growth of $\log N(t)$ (i.e., plot $\mathrm{E}[\log N(t)] = \log N(0) + t\,\mathrm{E}[\log R(t)]$).

1e: What proportion of simulations arrive at populations less than $\log N(t) = 0$ at time $t = 100$?

Question 2. Let $R(t)$ assume $\{5, 0.25\}$ with equal probability.

2a: What is the expected value of $\log R(t)$?

2b: What is the variance of $\log R(t)$?

2c: Simulate 1000 populations over 100 generations from this model starting with $N(0) = 100$ and plot the log-population over time.

2d: Add to your plot from Question 1c the expected value of $\log N(t)$.

2e: What proportion of simulations arrive at populations less than $\log N(t) = 0$ at time $t = 100$?

Question 3: Which model from Questions 1 and 2 produces a higher proportion of populations greater than $\log N(t) = 0$ at time $t = 100$?

Question 4: From the simulations produced in Question 2, extract the vector of the log-population sizes at time $t = 100$.

4a: Plot a histogram of the log-population sizes at $t = 100$.

4b: Compute the expected value and variance of $\log N(t)$ at $t = 100$.

4c: Generate a Normal probability density with mean and variance as computed in Question 4b on the interval of your plot from Question 4a. Overlay a plot of the density onto your histogram. How well does the Normal fit match the results of the simulations?

(*Hint: Set the* `'Normalization'` *property of* `histogram` *to* `'pdf'` *when you call the function in Question 4a)*

(Hint: Use the function `normpdf(points, mu, sigma)` *to produce the Normal density. Be sure to use the square root of the variance, as* `sigma` *indicates the standard deviation)*

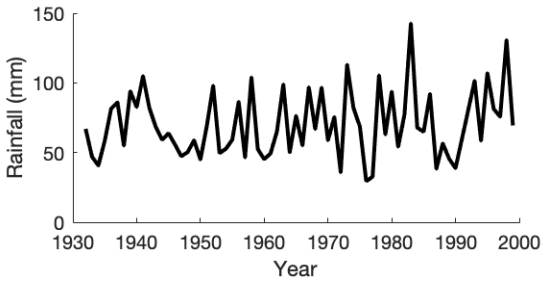
Part 2. Bay Checkerspot Butterflies

The Bay checkerspot butterfly. (Image from the Bureau of Reclamation[1])

The Bay checkerspot butterfly (*Euphydryas editha bayensis*) is a butterfly species local to the San Francisco Bay Area. The subspecies has been the subject of extensive study since the 1960s when Stanford researchers recognized its vulnerability in light of the rapid land development taking place in the Bay Area. In 1987, the U.S. Fish and Wildlife Service designated the species threatened. Since the designation, however, all populations outside of Santa Clara county have disappeared. Today, the butterfly is the focus of numerous private and public conservation projects which aim to protect several core habitat areas. Ongoing projects are also aiming to reintroduce the population to previous habitats[1]. Nevertheless, the species remains threatened.

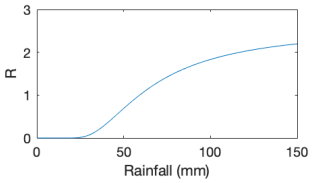
The Jasper Ridge Biological Preserve in San Mateo County deemed their local population extinct in 1998. Stanford Biologist Paul Ehrlich, who studied the population there since the 1960s, examined 70 years of climate data and concluded that large fluctuations in the local climate likely lead to the demise of the population. In this part of the lab, we are going to use a model linking local climate to Bay checkerspot population growth to investigate what effect changes in climate may have on the subspecies' growth.

Rainfall data for San Jose over 1932-1999 is available on Canvas as **`checkerspots.csv`**. We can read the file to use the data in MATLAB by calling **`data = csvread('checkerspots.csv', 1, 1)`**. The second and third arguments tell MATLAB how many rows and columns it can skip when loading the data, respectively. Plotting the rainfall data shows us the tendency of annual rainfall in the Bay Area to be highly variable.



Correlating the amount of rainfall in a given year to changes in Bay checkerspot population size has led to the following model of population growth,

$$R = e^{a - bx^{-2}}$$



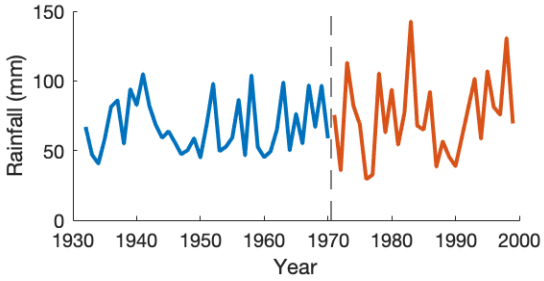
where $A = 0.932171$, $B = 3262.42$, and x is the amount of rain in a year (in millimeters). Generally speaking, more rain leads to a larger population. However, when it rains less than 30 mm, the population is predicted to be in significant danger (evidenced by $R < 0.1$). This model tells us how we can convert the observed rainfall amounts into predicted growth coefficients for each year.

[1] https://www.usbr.gov/newsroom/newsrelease/detail.cfm?RecordID=61851

Because the Bay checkerspot population has been decreasing dangerously since the 1970s, researchers are specifically interested in assessing changes in local climate between pre-1971 data and post-1971 weather data. We can separate the rainfall data with the indexing commands

```
x_pre1971 = data(data(:,1)<1971, 2);
x_post1971 = data(data(:,1)>=1971, 2);
```

Then, plotting the rainfall of the two time periods separately allows us to visualize the differences in rainfall before and after 1971.



Question 5: Load the rainfall data in MATLAB and compute the average annual rainfall for the pre-1971 and post-1971 data. In which time period is the average annual rainfall greater?

Question 6: Use the model between rainfall and population growth given above to produce the vectors of population growth coefficients for the time periods pre- and post-1971. Which time period has a greater average $\log R$?

Question 7: Use `sim_geometric_population` to simulate 100 generations of 1000 populations by assuming R is founded on a random uniform sampling of the growth coefficient from the period before 1971. Use $N(0) = 100$.

(Hint: `probs = ones(1, length(R))/length(R)`*)*

> **7a**: Plot the trajectories of log-population size with the corresponding estimate of the model's average growth, $\mathrm{E}[\log N(t)]$.
>
> **7b**: What proportion of simulations are below $\log N(t) = 0$ at time $t = 100$?

Question 8: Simulate 100 generations of 1000 populations by assuming R is founded on a uniform random sampling of the growth rates from the period after 1971. Use $N(0) = 100$.

> **8a**: Plot the trajectories of log-population size with the corresponding linear estimate of the model's average growth, $\mathrm{E}[\log N(t)]$.
>
> **8b**: What proportion of simulations are below $\log N(t) = 0$ at time $t = 100$?

Question 9: What explanation can you give for differences in your answers to Question 7b and Question 8b? Do the rainfall data and population model support the hypothesis that large fluctuations in local climate can be damaging for the Bay checkerspot?