Modeling Genetic Drift With Random Walk

Natalie Hawley

December 2020

Abstract

Genetic drift is a change in the allele frequency of a small population due to chance. In this project, random walk is used as the basis of two models for genetic drift, one in which the allele frequency of the whole population varies and another in which the number of offspring each individual has varies. The first model was used for two experiments predicting losses and fixations in populations of varying sizes and varying initial allele frequencies. The model predicted that as population size increases, the number of generations required to reach a loss or fixation also increases, and that the closer the initial allele frequency is to 0 or 1, the fewer generations required to reach a loss or fixation.

Objective

To build a model of genetic drift using the principles of random walk and determine the amount of time it takes to reach a loss or fixation of a trait under different conditions.

Introduction

Genetic drift occurs when there is a chance distribution of a trait because the trait has a neutral effect on fitness. Loss of a trait occurs when it is eliminated from a population, and fixation occurs when there is a 100% chance of the trait in new offspring because every individual in the population has the allele. There are several different parameters which may affect the amount of time it takes to reach a loss or fixation. I want to be able to predict the amount of time it takes for populations of varying sizes and with varying fractions of the population exhibiting the trait to reach a loss or fixation.

Theory and Methods

Because a trait with no effect on fitness would have no effect on natural selection either, genetic drift depends on probability, and can therefore be modeled with random walk. In random walk, there is some probability p that the walker, in this case the population, will go left, and some probability q that the walker will go right. The allele frequency of the population is allowed to walk between 0 and 1, modeling how the proportion of individuals in the population with the given trait can change. In Model 1, there is an equal probability that the allele frequency will increase or decrease by 1/n (where n is the population size) in each generation.

Additionally, Model 2 also uses randomization; however, this model tracks individuals, and the number of offspring each individual produces can vary. The reproduction is asexual, so if the parent carried the allele, all of the offspring will carry the allele too, or if the parent did not carry the allele, none of its offspring will carry the allele either. Therefore, the allele frequency can vary each generation depending how many offspring are produced by carriers and non-carriers.

Results and Discussion

Figure 1 and Figure 2 show sample walks for each model. Because growth is exponential, and every individual is accounted for, the Model 2 requires more computing power, and it is difficult to run the model for as many generations. Still, both samples demonstrate how an individual walk may deviate from the average. In random walk, average displacement is

represented by the equation <m> = N(p-q)Δx. In a model like this, where the probability of allele frequency increasing or decreasing (*p* and *q*) are equal, the average displacement should be 0, and on average, the final position of the walks should be 0.5 on average. This result is illustrated by Figure 3 and Figure 4. For each model, I ran 50 trials, and the final allele frequency of each trial is recorded in the histogram. The average of all trials was 0.480725 for the first model and 0.488526 for the second. The histogram for the Model 2 was usually similar to a normal distribution. Perhaps because this model made it impossible to reach a loss or fixation (having 0 offspring was not an option, so individuals with both traits would always exist in the population), the allele frequency would never get stuck at 0 or 1. While this does happen in Model 1, it is expected that the bins containing 0 and 1 would be inflated which is apparent when you compare Figure 3 to Figure 4.

Since the first model is able to run through hundreds of generations easily, I selected it for answering the questions of how population size and starting allele frequency affect the amount of time for a loss or fixation to be reached. For populations from 10 to 40, the generation at which a loss or fixation occurred was recorded for 50 trials, and the average was calculated. As seen in Figure 5, as population size grows, it takes longer for a loss or fixation to be reached on average. For populations up to 20 individuals, it usually takes less than 100 generations, but for a population of 40, it can take over 400 generations on average. For the second experiment, the initial allele frequency in the population varied between 0.1 and 0.9, and for 50 trials each, the generation at which a loss or fixation was reached was recorded. Intuitively, one might assume that the closer the initial allele frequency is to 0 or 1, the easier it will be to reach a loss or fixation. These results can be seen in Figure 6 which shows an almost bell-shaped curve, indicating the intermediate starting frequencies cause it to be harder to reach a loss or fixation. While the intermediate frequencies took over 150 generations to see a loss or fixation on average, population which started with allele frequencies of 0.1 or 0.9 would usually take around 50 generations.

Conclusion

This model of genetic drift using random walk provided results which would be expected in nature. As population size increased, it was more difficult for the population to experience a loss or fixation. Also, when the initial allele frequency was close to 0 or 1, it was easier for the population to experience a loss or fixation. This model can be used to predict whether populations are at risk of losing or fixing a trait which is important because of its effects on biodiversity.

For future work, an option for individuals to have zero offspring could be added to Model 2, so the loss and fixation questions could be tested for that model also. However, that would likely require high performance computing because it took my personal computer several minutes to run less than 10 generations. Furthermore, incorporating dominance patterns to the traits could also be an interesting extension of the project.

Figures

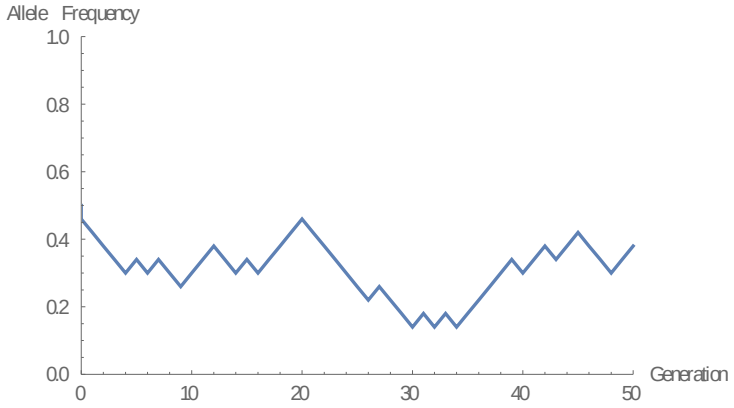

Figure 1 - Sample of Model 1 random walk for 50 generations.

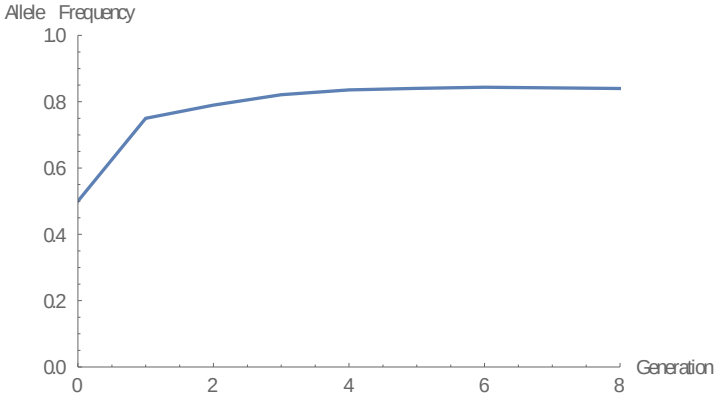

Figure 2 – Sample of Model 2 random walk for 8 generations.

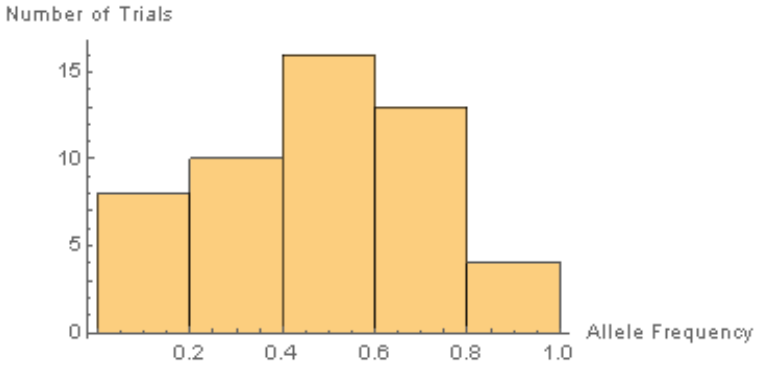

Figure 3 - Histogram of final allele frequencies for 50 trials of Model 1.

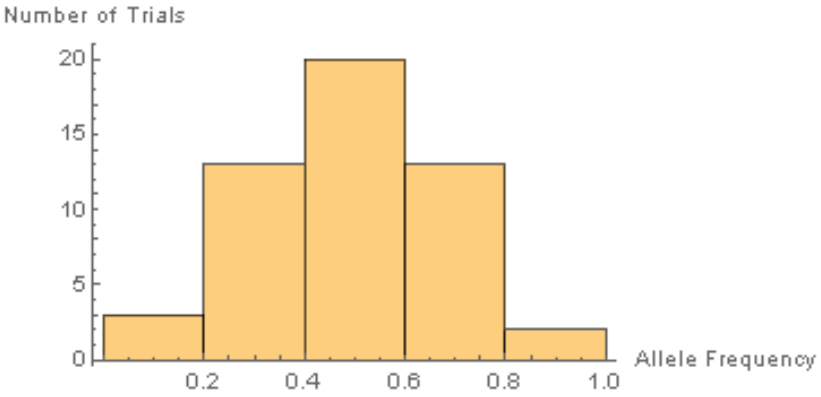

Figure 4 - Histogram of final allele frequencies for 50 trials of Model 2.

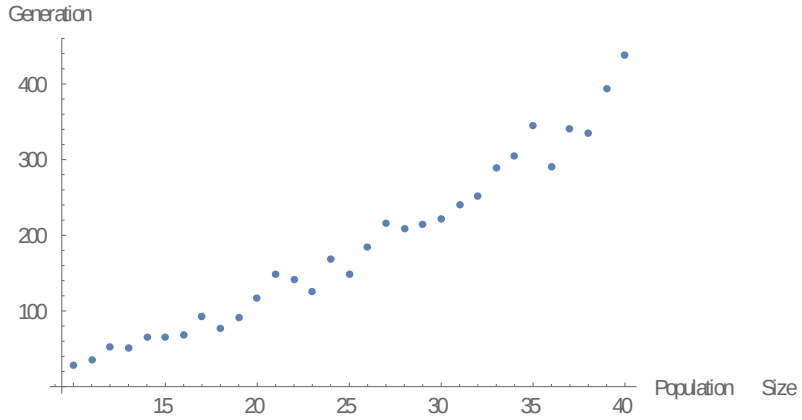

Figure 5 - Varying population sizes and the average generation at which a loss or fixation was reached.

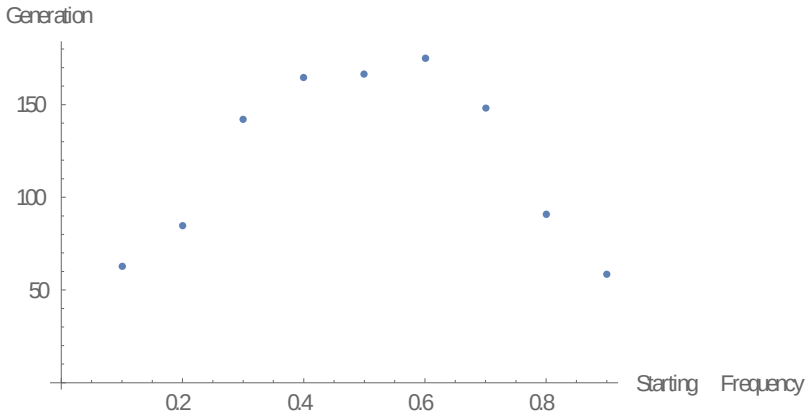

Figure 6 - Varying initial allele frequencies and the average generation at which a loss or fixation was reached.