Edie and Dari

R Project Written Summary (SHORT)

Dari coded a population simulation (<https://math-23c-final.netlify.app/>) which models generations of the populations that consider adjustable variables, including dominant and recessive phenotypes that influence survivability. For our dataset, we used the specific inputs: 0% for the start percentage of dominance, 50% chance for the dominant phenotype to live, 75% chance for the recessive phenotype to live, and a 3% mutation rate.

We extracted the columns of the amount of dominant and recessive genes and created a new column for the total number of cells. From this data, we generated various bar plots. Using maximum likelihood estimation (MLE) of minus log-likelihood (MLL), we modeled the total percentage of recessive genes as a logistic function with the coefficients of -2.90 and 0.211 which resulted in an MLL of 13.7. For this and the following tests, we used two χ2 tests. One gives a low-ball P-value and the other should be more accurate but give a much higher P-value. The more accurate test gave a P-value of 1 and the other a value of 0.1.

After centering our y-values, we were able to then center our x-values (based on the 0 of the y-values) and normalize the data. We then performed a linear regression and took the arctangent of those values. After de-normalizing our y-values and overlaying the fitted arctangent function, we saw that this model is a fairly good fit. This is substantiated by the low-ball p-value of ~0.78 (which is much larger than the 0.1 from the fitted logistic curve) and the more accurate p-value of 1. Since both of the distributions had an accurate p-value of 1, it implies they are both excellent fits.

Next, we created a matrix where each column is representative of the change of the population from the previous generation and added boolean columns of whether the population changed. The histogram of the population changes shows an interesting spike around -8 and 5, but a gap near 8.

A contingency table of the increase of recessive and heterozygous phenotypes showed the two variables are correlated. A permutation test gave a ~0.0478 p-value, which confirmed our beliefs. Note: Since we were operating over boolean values, a permutation test is better than a χ2 test or other classical methods to test our hypothesis.

Looking at a scatterplot of the recessive genotype vs. dominant genotype, we expect that the dominant and recessive phenotypes should be negatively correlated and have a high covariance because they have a large spread. When we analyzed the correlation and covariance of the variables, we found that they indeed matched our expectations.

Subsequently, we looked at the above plot and fit a Γ distribution to it. To do this, we recreated the initial “dataset” and found its expectation and variance, which we used to find the shape and rate parameters for the Γ distribution function. We then overlaid our Γ distribution function over our original plot. We performed a t-test which returned a p-value of 0.9994, and a χ2 test which gave a P-value of 1.4\*10-17.

Our simulation is based on standard models used in biology. Therefore, we would expect the simulation should accurately model the real world. Nevertheless, we got some unintuitive results. These results may be genuine, or they could be due to a fault inherent to the simulation and/or randomness and the result of the parameters we set at the beginning of the simulation.