Some Definitions

- n_t: The number of timepoints.
- f_{ai}: The frequency of mutation a at time point i
- f_i: The mean of the two mutations at time point i
- σ_i^2: The variance of the paired time series at time i
- $\bar{d}$: The average of the differences between both time series
- σ_p^2: Variance for the pair of mutational time series.

The Math

The script calculates the relative similarity between all pairs of mutational time series with length n_t containing frequencies $0 <= f <= 1$. These are consistent with n independent draws from a normal distribution, assuming a variance of

1. $\sigma_i^2 = n_{binom} f_i (1 - f_i)$

where f_i is the mean frequency of both mutations at each time point and n_{binom} is picked arbitrarily as a value that gives reasonable uncertainties given our data (in this case, $n_{binom} = 1/5$). A time series with n_t random draws will have a variance of

2. $\sigma_{tot}^2 = \sum_i^{n_t} \sigma_i^2 = n_{binom} \sum_i^{n_t} f_i (1 - f_i)$

Since we are interested in the similarity between both mutational time series, let $d_i = |_{ai} - f_{bi}|$ for all time points yielding a mean and variance of

3. $\bar{d} = \frac{1}{n_t} \sum_i^{n_t} d_i$
4. $\sigma_p^2 = \frac{\sigma_{tot}^2}{n_t^2} = \frac{n_{binom}}{n_t^2} \sum_i^{n_t} f_i (1 - f_i)$

Finally, given σ_d and $\bar{d}$ we can calculate the probability that this pair of mutations belong to the same genotype using the cumulative probability distribution of the normal distribution:

5.

$$p_{pair} = 1 - \int_{-\bar{d}/\sigma_p}^{\bar{d}/\sigma_p} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \equiv 1 - erf[\frac{\bar{d}}{\sqrt{2\sigma_p^2}}]$$

Notes

For each pair we discarded a time point if both f_{ai} and f_{bi} failed the condition $f_{detected} < f < f_{fixed}$, where $f_{detected} = 0.03$ and $f_{fixed} = 0.97$. Note the these filters check "greater than" and "less than", and NOT "equals to". This won't affect the results that much, but should be clarified since it affects repeatability.