Dissimilarity Between Random Unordered Draws with Replacement

Zarif Ahsan, Xiran Liu, Noah Rosenberg
Rosenberg Lab, Department of Biology, Stanford University

Motivation

In genetics, individuals inherit copies of a gene from each of their parents; some organisms show polyploidy, where a gene is represented in more than two copies in an individual. We can measure genetic dissimilarity between individuals by considering the pairwise dissimilarity of vectors of individual alleles. Assuming random mating, each vector can be regarded as a random, unordered draw with replacement. Our measure of genetic dissimilarity thus becomes a more general measure of dissimilarity between random, unordered draws with replacement. We examine the mathematical properties of this dissimilarity, which has multiple combinatorial applications.

Definitions

We denote a collection of $I$ distinct objects by:

$$A_I = \{A_1, \ldots, A_I\}.$$  

The space of ordered draws of size $K$ is the product $A^K_I$. The space of unordered draws is

$$S_K = A^K_I / \mathbb{S}_K$$

where $\mathbb{S}_K$ acts on $A^K_I$ by permuting the order of coordinates. We uniquely represent $G \in G^K_I$ with a vector

$$ \mathbf{g} = (g_1, \ldots, g_K) \quad \sum_{i=1}^K c_i = K$$

where $c_i$ is the count of $A_i$ in $G$. We define our dissimilarity measure $D$ as

$$D(G_i, G_j) = \frac{1}{K^2} \mathbb{P}(\mathbf{g} = \mathbf{g}')$$

This computes the proportion of all pairs taking one item from each draw that are not matches:

$$A_1 \ A_2 \ A_3 \ A_4$$

where $A_1, A_2, A_3, A_4$ are the elements of $A_I$. For each draw, we define its probability by summing across each element in the orbit of $[\mathbf{g}]$:

$$P_i = \sum_{G_i \in \mathbb{O}([\mathbf{g}])} P(G_i)$$

where $\mathbb{O}([\mathbf{g}])$ is the set of all elements of $G^K_I$ in the orbit of $[\mathbf{g}]$.

Probability of Cases

Let the drawing probability of $A_i$ be $p_i$ and $q_i$ in each draw, respectively. As vectors, $\mathbf{p} = (p_1, \ldots, p_I)$ and $\mathbf{q} = (q_1, \ldots, q_I)$.

For each $[\mathbf{g}]$, we find its probability by summing across each element in the orbit of $[\mathbf{g}]$:

$$P([\mathbf{g}]) = \sum_{G_i \in \mathbb{O}([\mathbf{g}])} P_i$$

where $\mathbb{O}([\mathbf{g}])$ is the set of all elements of $G^K_I$ in the orbit of $[\mathbf{g}]$.

Expected Dissimilarity

We use these probabilities and their respective $D$ values to compute an expectation for a given $K$. We prove that, in general,

$$E[D(\mathbf{p}, \mathbf{q})] = 1 - \langle \mathbf{p}, \mathbf{q} \rangle$$

We expect the mean dissimilarity to be minimized when two draws are taken with the same probabilities (i.e. $\mathbf{p} = \mathbf{q}$). However,

**Theorem** For any $K, I$ and $\mathbf{p}, \mathbf{q}$:

$$E[D(\mathbf{p}, \mathbf{p})] = E[D(\mathbf{p}, \mathbf{q})]$$

**Proposition** For $I = 2K$ this enumeration is independent of $I$.

The size of $G^K_I$, denoted $M_K$, for $I = 2K$ is the OEIS sequence A331722:

$$M_K = 2, 7, 21, 66, 192, \ldots$$

Connections

The combinatorial nature of our problem connects to multiple settings:

- $E[D]$ measures genetic difference between populations. We find conditions for when interpopulation genetic variation $E[D(\mathbf{p}, \mathbf{p})]$ and $E[D(\mathbf{q}, \mathbf{q})]$ exceeds interpopulation difference $E[D(\mathbf{p}, \mathbf{q})]$.

- When drawing without replacement from large population of objects, samples behave similarly to random unordered draws with replacement. In this instance, $E[D]$ provides a measure of variability among these samples.

- The probability expressions determined can be used to compute expected values of other measures on the space of draws.

Open Questions

- Can we determine a generating function for the elements of $G^K_I$ and algorithmically enumerate its elements?

- Can we show that the probability of our single-population dissimilarity exceeding our two-population dissimilarity approaches 0 as $K$ grows large?

References


[2] Xiran Liu, Zarif Ahsan, Tarun Matheswaran, and Noah Rosenberg. When is the allele-sharing dissimilarity between two populations exceeded by the allele-sharing dissimilarity of a population with itself? Statistical Applications in Genetic and Molecular Biology, in revision.


Acknowledgements

The authors would like to thank the Stanford University Research Institute in Mathematics (SURIM) 2023 Program and Lernik Aessner for supporting this project, Tarun Matheswaran for his contributions to the predecessor of this project, Chloe Schiff and Egor Lapko for poster edits and template, and the rest of Rosenberg Lab for their mentorship and guidance.