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 frequencies given in the population. 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:

$$\mathcal{A}_I = \{A_1, \ldots, A_I\}.$$

The space of *ordered* draws of size K is the product $\mathcal{A}_I^K$. The space of unordered draws is

$$\mathcal{G}_I^K = \mathcal{A}_I^K / S_K$$

where S_K acts on $\mathcal{A}_I^K$ by permuting the order of coordinates. We uniquely represent $G \in \mathcal{G}_I^K$ with a vector

$$\mathbf{g} = (c_1, \ldots, c_I) \qquad \sum_{i=1}^{I} c_i = K$$

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

$$\mathcal{D}(\mathbf{g}_1, \mathbf{g}_2) = 1 - \frac{1}{K^2} \langle \mathbf{g}_1, \mathbf{g}_2 \rangle$$

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
A_1	A_1A_1	A_2A_3	A_1A_3	A_1A_4
A_2	A_1A_2	A_2A_2	A_2A_3	A_2A_4
A_3	A_1A_3	A_2A_3	A_3A_3	A_3A_4
A_4	A_1A_4	A_2A_4	A_3A_4	A_4A_4

$\mathcal{D}(G_1, G_2) = \frac{3}{4}$

Table: Computation of $\mathcal{D}$ in $K = 4$, $I \geqslant 4$ case.

$\mathcal{D}$ is not a distance metric on our space, as it does not obey the triangle inequality.

Enumerating Cases

We represent a pair of draws $(\mathbf{g}_1, \mathbf{g}_2) \in \mathcal{G}_I^K \times \mathcal{G}_I^K$ by concatenating them into a $2 \times I$ matrix

$$\hat{\mathbf{g}} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} = \begin{pmatrix} g_1^1 & \cdots & g_1^I \\ g_2^1 & \cdots & g_2^I \end{pmatrix}$$

By the symmetry of $\mathcal{D}$ in its arguments, the order of our pair of draws does not matter, so we take

$$\mathcal{P}_I^K = \mathcal{G}_I^K \times \mathcal{G}_I^K / S_2$$

where S_2 permutes the rows of each $\hat{\mathbf{g}} \in \mathcal{G}_I^K \times \mathcal{G}_I^K$. To reduce our cases further, we take

$$\mathcal{C}_I^K = \mathcal{P}_I^K / S_I$$

where S_I acts by permuting the columns of $\hat{\mathbf{g}}$. We denote equivalence classes of a matrix $\hat{\mathbf{g}}$ as

$$[\hat{\mathbf{g}}]_\sim \in \mathcal{P}_I^K$$
$$[\hat{\mathbf{g}}] \in \mathcal{C}_I^K$$

$\mathcal{C}_I^K$ is the set of $2 \times I$ nonnegative integer matrices with rows summing to K, up to permutation of rows and columns.

Draws	$\hat{\mathbf{g}}$	$\mathcal{D}$
$A_1A_1,\ A_1A_1$	$\begin{pmatrix} 2&0&0&0 \\ 2&0&0&0 \end{pmatrix}$	0
$A_1A_1,\ A_1A_2$	$\begin{pmatrix} 2&0&0&0 \\ 1&1&0&0 \end{pmatrix}$	$\frac{1}{2}$
$A_1A_1,\ A_2A_2$	$\begin{pmatrix} 2&0&0&0 \\ 0&2&0&0 \end{pmatrix}$	1
$A_1A_1,\ A_2A_3$	$\begin{pmatrix} 2&0&0&0 \\ 0&1&1&0 \end{pmatrix}$	1
$A_1A_2,\ A_1A_2$	$\begin{pmatrix} 1&1&0&0 \\ 1&1&0&0 \end{pmatrix}$	$\frac{1}{2}$
$A_1A_2,\ A_1A_3$	$\begin{pmatrix} 1&1&0&0 \\ 1&0&1&0 \end{pmatrix}$	$\frac{3}{4}$
$A_1A_2,\ A_3A_4$	$\begin{pmatrix} 1&1&0&0 \\ 0&0&1&1 \end{pmatrix}$	1

Table: Enumeration of elements in $\mathcal{C}_I^K$ for $K = 2, I \geqslant 2$, each given by its representative draws and matrix.

Proposition For $I \geqslant 2K$ this enumeration is independent of I.

The size of $\mathcal{C}_I^K$, denoted M_K, for $I \geqslant 2K$ is the OEIS sequence A331722:

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

Probability of Cases

Let the drawing probability of A_i be p_i and q_i in each draw, respectively. As vectors,

$$\mathbf{p} = (p_1, \ldots, p_I) \qquad \mathbf{q} = (q_1, \ldots, q_I)$$

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

$$P\big([\hat{\mathbf{g}}]\big) = \sum_{H \in \mathrm{Orbit}_{S_I}([\hat{\mathbf{g}}]_\sim)} P(H)$$
$$= C(\hat{\mathbf{g}}) \sum_{i_1 \neq \cdots \neq i_{N(\hat{\mathbf{g}})}} \Big(\prod_{j=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_1^j} q_{i_j}^{g_2^j} + \prod_{i=1}^{N(\hat{\mathbf{g}})} p_{i_j}^{g_2^j} q_{i_j}^{g_1^j} \Big)$$

where the first $N(\hat{\mathbf{g}})$ columns of $\hat{\mathbf{g}}$ are nonzero (we can always find such an element in each S_I-equivalence class). The coefficient $C(\hat{\mathbf{g}})$ is

$$C(\hat{\mathbf{g}}) = \frac{(K!)^2}{(1 + \mathbb{1}_{r_1 = r_2})(\prod_{\ell=1}^{L} |\{c_\ell\}|!)} \prod_{i=1}^{I} \frac{1}{g_1^j!\ g_2^j!}$$

where $c_1, \ldots, c_L$ are the unique nonzero columns in $\hat{\mathbf{g}}$, $|\{c_\ell\}|$ is the count of each c_ℓ in $\hat{\mathbf{g}}$, and r_i is the set of nonzero entries in the ith row of $\hat{\mathbf{g}}$.

Expected Dissimilarity

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

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

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 $\mathbf{p}$, $\mathbf{q}$:
$\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{p})] \leqslant \mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{q})]$
iff $\langle \mathbf{p}, \mathbf{q} \rangle \leqslant \langle \mathbf{p}, \mathbf{p} \rangle$.

Therefore, this inequality does not always hold, such as when $\mathbf{p} = (0.8, 0.2, 0, \ldots, 0)$ and $\mathbf{q} = (0.9, 0.1, 0, \ldots, 0)$. Nonetheless, we have

Theorem For any K, I and $\mathbf{p}$, $\mathbf{q}$:
$\frac{1}{2}\big(\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{p})] + \mathbb{E}[\mathcal{D}(\mathbf{q}, \mathbf{q})]\big) \leqslant \mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{q})]$.

Thus, for any $\mathbf{p}$ and $\mathbf{q}$, $\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{q})]$ is bounded below by either $\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{p})]$ or $\mathbb{E}[\mathcal{D}(\mathbf{q}, \mathbf{q})]$.

Connections

The combinatorial nature of our problem connects to multiple settings:

- $\mathbb{E}(\mathcal{D})$ measures genetic difference between populations. We find conditions for when intrapopulation genetic variation ($\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{p})]$ and $\mathbb{E}[\mathcal{D}(\mathbf{q}, \mathbf{q})]$) exceeds interpopulation difference ($\mathbb{E}[\mathcal{D}(\mathbf{p}, \mathbf{q})]$).
- When drawing without replacement from large population of objects, samples behave similarly to random unordered draws with replacement. In this instance, $\mathbb{E}[\mathcal{D}]$ 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 $\mathcal{C}_I^K$ 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

[1] Joanna L Mountain and Uma Ramakrishnan. Impact of human population history on distributions of individual-level genetic distance. *Human Genomics*, 2:4–19, 2005.

[2] Xiran Liu, Zarif Ahsan, Tarun Martheswaran, 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.

[3] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at http://oeis.org.

Acknowledgements

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