HW3: Linkage Disequilibrium

CS 2840 Spring 2025

Released: Thursday, April 10, 2025

Due: Thursday, April 24, 2025, **11:59pm**

Overview

All homework assignments in this course will be submitted on gradescope. For this assignment, your submission should include a PDF file with your answers to 1, 2, 3, and 4 (no coding this time!).

1 r^2 and Pearson Correlation

Let A and B be two biallelic loci with respective alleles a, a' and b, b'. Consider we select a random individual from a large population with these loci. Let X be 1 if the individual has allele a at locus A and Y be 1 if the individual has allele b at locus B; both variables are 0 if their respective conditions are not met. In a sort of rough sense, $X = \mathbb{1}_a$, for example.

Let $\mathbb{P}(X = 1) = p_a$ and $\mathbb{P}(Y = 1) = p_b$. These two definitions fully define the individual allele frequencies of A and B in the population. Let $\mathbb{P}(X = 1, Y = 1) = f_{ab}$. Show that the LD metric

$$r^2 = \frac{D^2}{p_a(1-p_a)p_b(1-p_b)} = \rho^2$$

Where $D = f_{ab} - p_a p_b$ and ρ is the correlation coefficient between X and Y, i.e. $\rho = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}$.

2 Recombination and LD

Retain the setup from problem 1, assuming random mating, no mutation, and constant recombination fraction c throughout generations.

(a) How does f_{ab} change from generation $t = 0$ to generation $t = 1$, assuming an infinitely large population? Your answer should include c. *Note that D should appear nowhere in your answer to this subpart.*

(b) Using this result, determine how fast r^2 and D approach 0 given c, starting from D_0 and r_0^2 respectively. Which approaches 0 faster? *Hint: start by deriving D_1 from D_0 using the evolution of f_{ab} from a.*

3 r^2 and D'

Continue on with the same setup as in problems 1 and 2. Come up with an example $f_{ab}, f_{a'b}, f_{ab'}, f_{a'b'}$ and frequencies $p_a, p_b, p_{a'}, p_{b'}$ such that:

(a) $D' = r^2 = 1$

(b) $r^2 < D' = 1$

(c) $r^2 < 1$ and $D' < 1$

(d) Then, prove that

$$r^2 \leq D'$$

Note that $D' = \frac{D}{D_{\mathsf{norm}}}$, where

$$D_{\mathsf{norm}} = \begin{cases} \max\{-p_A p_B,\ -(1-p_A)(1-p_B)\}, & \text{if } D < 0 \\ \min\{p_A(1-p_B),\ p_B(1-p_A)\}, & \text{if } D > 0 \end{cases}$$

4 Fisher's Exact Test

Fisher's exact test computes the exact p-value for a 2×2 contingency table by enumerating all possible tables with fixed margins (i.e, we know the allele frequencies) and summing up all probabilities smaller than that of our observed table. Consider the following table for alleles at Locus 1 (alleles a and a') and Locus 2 (alleles b and b'):

	b	b'	Row Totals
a	w	x	$w + x = k_a$
a'	y	z	$y + z = k_{a'}$
Column Totals	$w + y = k_b$	$x + z = k_{b'}$	$n = w + x + y + z$

(a) Derive the probability of obtaining any particular set of values w, x, y, and z in the table under the null hypothesis of no association between the alleles, assuming we have tabulated the allele frequencies (and thus have fixed k_x for all x).

(b) Why does Fisher's exact test not scale well with number of samples? Remember that Fisher's test enumerates all possible tables given our observed marginals, and then sums the probability we see a table "as or more extreme" than the table we observed (according to some notion of "extreme").