Measures of population divergence for binary data: limitations and improvements via simulations

E. Nikita[1], P. Nikitas[2]

[1]Science and Technology in Archaeology and Culture Research Centre, The Cyprus Institute, 20 Konstantinou Kavafi street, 2121 Aglantzia, Nicosia, CYPRUS; e.nikita@cyi.ac.cy
[2] Department of Chemistry, Aristotle University of Thessaloniki, University Campus, 54124 Thessaloniki, GREECE; nikitas@chem.auth.gr

Keywords: *biodistance, population affinities, simulations, binary data*

Abstract

Measures of divergence are used extensively in biodistance studies in order to examine past population history and gene flow. Such measures predominantly employ skeletal and dental traits recorded as binary dichotomies (present/absent). The performance of the measures of divergence for binary data as unbiased estimators of population divergence is examined using 10 datasets of nonmetric traits taken from the literature. The measures are based on the application to the sample proportions the probit, logit, and arcsine transformations, the latter via the Smith, Anscombe, and Freeman-Tukey formulas. The case of the untransformed data is also examined. It is shown that the main source of biased estimations at low and high proportions is the data transformation adopted in these measures that results in variances that are asymptotically valid within a range of proportion values around 0.5. The estimation of variances via simulations improves their performance and the measures become nearly unbiased estimators but again provided that low and high proportions are excluded. For further improvement, a new modification of these measures is proposed, which makes them strict unbiased estimators of population divergence throughout the proportion range provided that the variances are estimated via simulations. This modification does not alter the measure of divergence based on untransformed data, which is also a strict unbiased estimator irrespective of the use simulated or non-simulated variances. The computation of the proposed measures, their p-values and confidence intervals are implemented using custom functions in R.