Optimal Data Analysis
Vol. 3 (November 9, 2014), 115-116

2155-0182/10/$3.00

UniODA *vs.* Bray-Curtis Dissimilarity Index for Count Data

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

The Bray-Curtis dissimilarity index is widely-used as an index of the magnitude of compositional dissimilarity in the count of different categories between two samples, yet it fails to address the statistical reliability of the dissimilarity, the precise nature of the dissimilarity, and the potential cross-generalizability of the findings. All of these shortcomings are remedied by the use of UniODA in this application.

The Bray-Curtis dissimilarity index (BCDI) is a widely-used index of the degree or magnitude of difference (i.e., the *compositional dissimilarity*) in the number or count of different categories between two samples.[1] When used with raw count data the BCDI is obtained by computing the sum of the absolute differences between counts across categories, and dividing this value by the sum of the abundances of the counts across categories. The BCDI is bounded at the extremes by 0 when the samples are identical and by 1 when the samples are "completely disjoint" (i.e., for each category the count is zero for one sample and non-zero for the other sample). The result is conventionally multiplied by 100 and expressed in terms of a percentage. Subtracting this value from 100 yields a measure of the similarity between the two samples, called the Bray-Curtis Index (BCI).[1] For the data presented in Table 1, for example, the BCDI is computed as 56.8%. The magnitude of the computed BCDI index begs three questions: (a) whether the magnitude of the between-sample difference is statistically reliable, or if the noted differences might be attributable to chance-based variation in the category counts; (b) specifically what categories differentiate the two samples and in what manner; and (c) whether a similar result might accrue if a second independent random assessment of the categories was conducted.

Table 1: Number (Count) of Five Ecological Categories for Two Sampling Sites

	Sampling Site	
Ecological Category	S29	S30
A	11	24
B	0	37
C	7	5
D	8	18
E	0	1

It is straightforward to demonstrate that *uni*variate *o*ptimal (maximum-accuracy) *d*ata *a*nalysis (UniODA) may instead be used to assess inter-sample differences in a manner that

2155-0182/10/$3.00

addresses all three questions.[2,3] For UniODA no distributional assumptions are made concerning the counts; either a non-directional (exploratory or *post hoc*) or directional (confirmatory or *a priori*) hypothesis regarding the nature of the compositional dissimilarity may be evaluated; and the inter-sample dissimilarity is expressed on a normed scale called the *e*ffect *s*trength for *s*ensitivity (ESS) that ranges from 0 (level of dissimilarity that is expected by chance) to 100 (the samples are completely disjoint).

For the data in Table 1 the UniODA model is: if sample=S29 then predict that ecological category=A, C, or D; and if sample= S30 then predict that ecological category=B or E [addressing question (b)]. For this model $p< 0.00016$ [addressing question (a)], and ESS= 44.7. By convention, for every analysis ESS<25 reflects a relatively weak effect; ESS< 50 is a moderate effect; ESS<75 is a relatively strong effect; and ESS≥75 is a strong effect.[3] The model correctly classified 26/26 (100%) of the counts from sample S29, and 38/85 (44.7%) of the counts from sample S30: these are indices of model sensitivity. When the model predicted that the sample was S29 it was correct for 26/73 (35.6%) of the counts, and then the model predicted that the sample was S30 it was correct for 38/38 (100%) of the counts: these are indices of model predictive value. To address question (c) a directional jackknife ("leave-one-out") validity analysis is conducted: here p<0.000004 and ESS=43.5, indicating that this finding is expected to cross-generalize to an independent random sampling of the ecological categories at the two sampling sites.

References

[1]http://www.econ.upf.edu/~michael/stanford/maeb5.pdf

[2]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[3]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

Author Notes

E-mail: Journal@OptimalDataAnalysis.com.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646