PROXIMITIES

Standardizing Cases or Variables

Either cases or variables can be standardized. The following methods of standardization are available:

Z

PROXIMITIES subtracts the mean from each value for the variable or case being standardized and then divides by the standard deviation of the values. If a standard deviation is 0, PROXIMITIES sets all values for the case or variable to 0.

RANGE

PROXIMITIES divides each value for the variable or case being standardized by the range of the values. If the range is 0, PROXIMITIES leaves all values unchanged.

RESCALE

From each value for the variable or case being standardized, PROXIMITIES subtracts the minimum value and then divides by the range. If a range is 0, PROXIMITIES sets all values for the case or variable to 0.50.

MAX

PROXIMITIES divides each value for the variable or case being standardized by the maximum of the values. If the maximum of a set of values is 0, PROXIMITIES uses an alternate process to produce a comparable standardization: it divides by the absolute magnitude of the smallest value and adds 1.

MEAN

PROXIMITIES divides each value for the variable or case being standardized by the mean of the values. If a mean is 0, PROXIMITIES adds one to all values for the case or variable to produce a mean of 1.

SD

PROXIMITIES divides each value for the variable or case being standardized by the standard deviation of the values. PROXIMITIES does not change the values if their standard deviation is 0.

Transformations

Three transformations are available for the values PROXIMITIES computes or reads:

ABSOLUTE

Take the absolute values of the proximities.

REVERSE

Transform similarity values into dissimilarities, or vice versa, by changing the signs of the coefficients.

RESCALE

RESCALE standardizes the proximities by first subtracting the value of the smallest and then dividing by the range.

If you specify more than one transformation, PROXIMITIES does them in the order listed above: first ABSOLUTE, then REVERSE, then RESCALE.

Proximities Measures

Measures for Continuous Data

EUCLID

The distance between two items, *x* and *y*, is the square root of the sum of the squared differences between the values for the items.

$$\mathrm{EUCLID}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

SEUCLID

The distance between two items is the sum of the squared differences between the values for the items.

$$\mathrm{SEUCLID}(x, y) = \sum_i (x_i - y_i)^2$$

CORRELATION

This is a pattern similarity measure.

$$\mathrm{CORRELATION}(x, y) = \frac{\sum_i \left(Z_{xi} Z_{yi}\right)}{N}$$

where Z_{xi} is the (standardized) *Z*-score value of *x* for the *i*th case or variable, and *N* is the number of cases or variables.

COSINE

This is a pattern similarity measure.

$$\mathrm{COSINE}(x, y) = \frac{\sum_i (x_i y_i)}{\sqrt{\left(\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)\right)}}$$

CHEBYCHEV

The distance between two items is the maximum absolute difference between the values for the items.

$$\mathrm{CHEBYCHEV}(x,y) = \max_i \left| x_i - y_i \right|$$

BLOCK

The distance between two items is the sum of the absolute differences between the values for the items.

$$\mathrm{BLOCK}(x,y) = \sum_i \left| x_i - y_i \right|$$

MINKOWSKI(*p*)

The distance between two items is the *p*th root of the sum of the absolute differences to the *p*th power between the values for the items.

$$\mathrm{MINKOWSKI}(x,y) = \left(\sum_i \left| x_i - y_i \right|^p \right)^{1/p}$$

POWER (p,r)

The distance between two items is the *r*th root of the sum of the absolute differences to the *p*th power between the values for the items.

$$\mathrm{POWER}(x,y) = \left(\sum_i \left| x_i - y_i \right|^p \right)^{1/r}$$

Measures for Frequency Count Data

CHISQ

The magnitude of this similarity measure depends on the total frequencies of the two cases or variables whose proximity is computed. Expected values are from the model of independence of cases (or variables), *x* and *y*.

$$\mathrm{CHISQ}(x,y)=\sqrt{\sum_i \frac{\left(x_i-E(x_i)\right)^2}{E(x_i)}+\sum_i \frac{\left(y_i-E(y_i)\right)^2}{E(y_i)}}$$

PH2

This is the CHISQ measure normalized by the square root of the combined frequency. Therefore, its value does not depend on the total frequencies of the two cases or variables whose proximity is computed.

$$\mathrm{PH2}(x,y)=\frac{\mathrm{CHISQ}(x,y)}{\sqrt{N}}$$

Measures for Binary Data

PROXIMITIES constructs a 2×2 contingency table for each pair of items in turn. It uses this table to compute a proximity measure for the pair.

		Item 2	
		Present	Absent
Item 1	Present	*a*	*b*
	Absent	*c*	*d*

PROXIMITIES computes all binary measures from the values of *a*, *b*, *c*, and *d*. These values are tallies across variables (when the items are cases) or tallies across cases (when the items are variables).

Russel and Rao Similarity Measure

This is the binary dot product.

$$\mathrm{RR}(x, y) = \frac{a}{a + b + c + d}$$

Simple Matching Similarity Measure

This is the ratio of the number of matches to the total number of characteristics.

$$\mathrm{SM}(x, y) = \frac{a + d}{a + b + c + d}$$

Jaccard Similarity Measure

This is also known as the similarity ratio.

$$\mathrm{JACCARD}(x, y) = \frac{a}{a + b + c}$$

Dice or Czekanowski or Sorenson Similarity Measure

$$\mathrm{DICE}(x, y) = \frac{2a}{2a + b + c}$$

Sokal and Sneath Similarity Measure 1

$$\mathrm{SS1}(x, y) = \frac{2(a + d)}{2(a + d) + b + c}$$

Rogers and Tanimoto Similarity Measure

$$\mathrm{RT}(x, y) = \frac{a+d}{a+d+2(b+c)}$$

Sokal and Sneath Similarity Measure 2

$$\mathrm{SS2}(x, y) = \frac{a}{a+2(b+c)}$$

Kulczynski Similarity Measure 1

This measure has a minimum value of 0 and no upper limit. It is undefined when there are no nonmatches $(b = 0 \text{ and } c = 0)$. Therefore, PROXIMITIES assigns an artificial upper limit of 9999.999 to K1 when it is undefined or exceeds this value.

$$\mathrm{K1}(x, y) = \frac{a}{b+c}$$

Sokal and Sneath Similarity Measure 3

This measure has a minimum value of 0, has no upper limit, and is undefined when there are no nonmatches $(b = 0 \text{ and } c = 0)$. As with K1, PROXIMITIES assigns an artificial upper limit of 9999.999 to SS3 when it is undefined or exceeds this value.

$$\mathrm{SS3}(x, y) = \frac{a+d}{b+c}$$

Conditional Probabilities

The following three binary measures yield values that you can interpret in terms of conditional probability. All three are similarity measures.

Kulczynski Similarity Measure 2

This yields the average conditional probability that a characteristic is present in one item given that the characteristic is present in the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1.

$$\mathrm{K2}(x, y) = \frac{a/(a+b) + a/(a+c)}{2}$$

Sokal and Sneath Similarity Measure 4

This yields the conditional probability that a characteristic of one item is in the same state (present or absent) as the characteristic of the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1.

$$\mathrm{SS4}(x, y) = \frac{a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)}{4}$$

Hamann Similarity Measure

This measure gives the probability that a characteristic has the same state in both items (present in both or absent from both) minus the probability that a characteristic has different states in the two items (present in one and absent from the other). HAMANN has a range of –1 to +1 and is monotonically related to SM, SS1, and RT.

$$\mathrm{HAMANN}(x, y) = \frac{(a+d) - (b+c)}{a+b+c+d}$$

Predictability Measures

The following four binary measures assess the association between items as the predictability of one given the other. All four measures yield similarities.

Goodman and Kruskal Lambda (Similarity)

This coefficient assesses the predictability of the state of a characteristic on one item (presence or absence) given the state on the other item. Specifically, lambda measures the proportional reduction in error using one item to predict the other, when the directions of prediction are of equal importance. Lambda has a range of 0 to 1.

$$t_1 = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$$
$$t_2 = \max(a + c, b + d) + \max(a + b, c + d)$$
$$\mathrm{LAMBDA}(x, y) = \frac{t_1 - t_2}{2(a + b + c + d) - t_2}$$

Anderberg's *D* (Similarity)

This coefficient assesses the predictability of the state of a characteristic on one item (presence or absence) given the state on the other. *D* measures the actual reduction in the error probability when one item is used to predict the other. The range of *D* is 0 to 1.

$$t_1 = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$$
$$t_2 = \max(a + c, b + d) + \max(a + b, c + d)$$
$$\mathrm{D}(x, y) = \frac{t_1 - t_2}{2(a + b + c + d)}$$

Yule's *Y* Coefficient of Colligation (Similarity)

This is a function of the cross-product ratio for a 2×2 table. It has a range of –1 to +1.

$$\mathrm{Y}(x, y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Yule’s *Q* (Similarity)

This is the 2×2 version of Goodman and Kruskal’s ordinal measure *gamma*. Like Yule’s *Y*, *Q* is a function of the cross-product ratio for a 2×2 table and has a range of –1 to +1.

$$\mathrm{Q}(x, y) = \frac{ad - bc}{ad + bc}$$

Other Binary Measures

The remaining binary measures available in PROXIMITIES are either binary equivalents of association measures for continuous variables or measures of special properties of the relation between items.

Ochiai Similarity Measure

This is the binary form of the cosine. It has a range of 0 to 1 and is a similarity measure.

$$\mathrm{OCHIAI}(x, y) = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)}$$

Sokal and Sneath Similarity Measure 5

This is a similarity measure. Its range is 0 to 1.

$$\mathrm{SS5}(x, y) = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Fourfold Point Correlation (Similarity)

This is the binary form of the Pearson product-moment correlation coefficient. Phi is a similarity measure, and its range is 0 to 1.

$$\mathrm{PHI}(x, y) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Binary Euclidean Distance

This is a distance measure. Its minimum value is 0, and it has no upper limit.

$$\mathrm{BEUCLID}(x, y) = \sqrt{b + c}$$

Binary Squared Euclidean Distance

This is also a distance measure. Its minimum value is 0, and it has no upper limit.

$$\mathrm{BSEUCLID}(x, y) = b + c$$

Size Difference

This is a dissimilarity measure with a minimum value of 0 and no upper limit.

$$\mathrm{SIZE}(x, y) = \frac{(b - c)^2}{(a + b + c + d)^2}$$

Pattern Difference

This is also a dissimilarity measure. Its range is 0 to 1.

$$\mathrm{PATTERN}(x, y) = \frac{bc}{(a + b + c + d)^2}$$

Binary Shape Difference

This dissimilarity measure has no upper or lower limit.

$$\mathrm{BSHAPE}(x, y) = \frac{(a + b + c + d)(b + c) - (b - c)^2}{(a + b + c + d)^2}$$

Dispersion Similarity Measure

This similarity measure has a range of –1 to +1.

$$\mathrm{DISPER}(x,y)=\frac{ad-bc}{(a+b+c+d)^2}$$

Variance Dissimilarity Measure

This dissimilarity measure has a minimum value of 0 and no upper limit.

$$\mathrm{VARIANCE}(x,y)=\frac{b+c}{4(a+b+c+d)}$$

Binary Lance-and-Williams Nonmetric Dissimilarity Measure

Also known as the Bray-Curtis nonmetric coefficient, this dissimilarity measure has a range of 0 to 1.

$$\mathrm{BLWMN}(x,y)=\frac{b+c}{2a+b+c}$$

References

Anderberg, M. R. 1973. *Cluster analysis for applications*. New York: Academic Press.

Romesburg, H. C. 1984. *Cluster analysis for researchers*. Belmont, Calif.: Lifetime Learning Publications.