measure_option — Option for similarity and dissimilarity measures

Description Syntax Options References Also see

Description

Several commands have options that allow you to specify a similarity or dissimilarity measure designated as *measure* in the syntax; see [MV] **cluster**, [MV] **mds**, [MV] **discrim knn**, and [MV] **matrix dissimilarity**. These options are documented here. Most analysis commands (for example, `cluster` and `mds`) transform similarity measures to dissimilarity measures as needed.

Syntax

command ... , ... `measure(`*measure*`)` ...

or

command ... , ... *measure* ...

measure	Description
cont_measure	similarity or dissimilarity measure for continuous data
binary_measure	similarity measure for binary data
mixed_measure	dissimilarity measure for a mix of binary and continuous data

cont_measure	Description
`L2`	Euclidean distance (Minkowski with argument 2)
`Euclidean`	alias for `L2`
`L(2)`	alias for `L2`
`L2squared`	squared Euclidean distance
`Lpower(2)`	alias for `L2squared`
`L1`	absolute-value distance (Minkowski with argument 1)
`absolute`	alias for `L1`
`cityblock`	alias for `L1`
`manhattan`	alias for `L1`
`L(1)`	alias for `L1`
`Lpower(1)`	alias for `L1`
`Linfinity`	maximum-value distance (Minkowski with infinite argument)
`maximum`	alias for `Linfinity`
`L(#)`	Minkowski distance with # arguments
`Lpower(#)`	Minkowski distance with # arguments raised to # power
`Canberra`	Canberra distance
`correlation`	correlation coefficient similarity measure
`angular`	angular separation similarity measure
`angle`	alias for `angular`

binary_measure	Description
`matching`	simple matching similarity coefficient
`Jaccard`	Jaccard binary similarity coefficient
`Russell`	Russell and Rao similarity coefficient
`Hamann`	Hamann similarity coefficient
`Dice`	Dice similarity coefficient
`antiDice`	anti-Dice similarity coefficient
`Sneath`	Sneath and Sokal similarity coefficient
`Rogers`	Rogers and Tanimoto similarity coefficient
`Ochiai`	Ochiai similarity coefficient
`Yule`	Yule similarity coefficient
`Anderberg`	Anderberg similarity coefficient
`Kulczynski`	Kulczyński similarity coefficient
`Pearson`	Pearson’s ϕ similarity coefficient
`Gower2`	similarity coefficient with same denominator as `Pearson`

mixed_measure	Description
`Gower`	Gower’s dissimilarity coefficient

Options

Measures are divided into those for continuous data and binary data. *measure* is not case sensitive. Full definitions are presented in *Similarity and dissimilarity measures for continuous data*, *Similarity measures for binary data*, and *Dissimilarity measures for mixed data*.

The similarity or dissimilarity measure is most often used to determine the similarity or dissimilarity between observations. However, sometimes the similarity or dissimilarity between variables is of interest.

Similarity and dissimilarity measures for continuous data

Here are the similarity and dissimilarity measures for continuous data available in Stata. In the following formulas, p represents the number of variables, N is the number of observations, and x_{iv} denotes the value of observation i for variable v.

The formulas are presented in two forms. The first is the formula used when computing the similarity or dissimilarity between observations. The second is the formula used when computing the similarity or dissimilarity between variables.

`L2` (aliases `Euclidean` and `L(2)`)

requests the Minkowski distance metric with argument 2. For comparing observations i and j, the formula is

$$\left\{\sum_{a=1}^{p}(x_{ia}-x_{ja})^2\right\}^{1/2}$$

and for comparing variables u and v the formula is

$$\left\{\sum_{k=1}^{N}(x_{ku}-x_{kv})^2\right\}^{1/2}$$

L2 is best known as Euclidean distance and is the default dissimilarity measure for discrim knn, mds, matrix dissimilarity, and all the cluster subcommands except for centroidlinkage, medianlinkage, and wardslinkage, which default to using L2squared; see [MV] **discrim knn**, [MV] **mds**, [MV] **matrix dissimilarity**, and [MV] **cluster**.

L2squared (alias Lpower(2))

requests the square of the Minkowski distance metric with argument 2. For comparing observations i and j, the formula is

$$\sum_{a=1}^{p}(x_{ia}-x_{ja})^2$$

and for comparing variables u and v, the formula is

$$\sum_{k=1}^{N}(x_{ku}-x_{kv})^2$$

L2squared is best known as squared Euclidean distance and is the default dissimilarity measure for the centroidlinkage, medianlinkage, and wardslinkage subcommands of cluster; see [MV] **cluster**.

L1 (aliases absolute, cityblock, manhattan, L(1), and Lpower(1))

requests the Minkowski distance metric with argument 1. For comparing observations i and j, the formula is

$$\sum_{a=1}^{p}|x_{ia}-x_{ja}|$$

and for comparing variables u and v, the formula is

$$\sum_{k=1}^{N}|x_{ku}-x_{kv}|$$

L1 is best known as absolute-value distance.

Linfinity (alias maximum)

requests the Minkowski distance metric with infinite argument. For comparing observations i and j, the formula is

$$\max_{a=1,\dots,p}|x_{ia}-x_{ja}|$$

and for comparing variables u and v, the formula is

$$\max_{k=1,\dots,N}|x_{ku}-x_{kv}|$$

Linfinity is best known as maximum-value distance.

`L(#)`

requests the Minkowski distance metric with argument #. For comparing observations i and j, the formula is

$$\left(\sum_{a=1}^{p} |x_{ia} - x_{ja}|^{\#}\right)^{1/\#} \qquad \# \geq 1$$

and for comparing variables u and v, the formula is

$$\left(\sum_{k=1}^{N} |x_{ku} - x_{kv}|^{\#}\right)^{1/\#} \qquad \# \geq 1$$

We discourage using extremely large values for #. Because the absolute value of the difference is being raised to the value of #, depending on the nature of your data, you could experience numeric overflow or underflow. With a large value of #, the `L()` option will produce results similar to those of the `Linfinity` option. Use the numerically more stable `Linfinity` option instead of a large value for # in the `L()` option.

See Anderberg (1973) for a discussion of the Minkowski metric and its special cases.

`Lpower(#)`

requests the Minkowski distance metric with argument #, raised to the # power. For comparing observations i and j, the formula is

$$\sum_{a=1}^{p} |x_{ia} - x_{ja}|^{\#} \qquad \# \geq 1$$

and for comparing variables u and v, the formula is

$$\sum_{k=1}^{N} |x_{ku} - x_{kv}|^{\#} \qquad \# \geq 1$$

As with `L(#)`, we discourage using extremely large values for #; see the discussion above.

`Canberra`

requests the following distance metric when comparing observations i and j

$$\sum_{a=1}^{p} \frac{|x_{ia} - x_{ja}|}{|x_{ia}| + |x_{ja}|}$$

and the following distance metric when comparing variables u and v

$$\sum_{k=1}^{N} \frac{|x_{ku} - x_{kv}|}{|x_{ku}| + |x_{kv}|}$$

When comparing observations, the Canberra metric takes values between 0 and p, the number of variables. When comparing variables, the Canberra metric takes values between 0 and N, the number of observations; see Gordon (1999) and Gower (1985). Gordon (1999) explains that the Canberra distance is sensitive to small changes near zero.

`correlation`

requests the correlation coefficient similarity measure. For comparing observations i and j, the formula is

$$\frac{\sum_{a=1}^{p}(x_{ia} - \overline{x}_{i.})(x_{ja} - \overline{x}_{j.})}{\left\{\sum_{a=1}^{p}(x_{ia} - \overline{x}_{i.})^2 \sum_{b=1}^{p}(x_{jb} - \overline{x}_{j.})^2\right\}^{1/2}}$$

and for comparing variables u and v, the formula is

$$\frac{\sum_{k=1}^{N}(x_{ku}-\overline{x}_{.u})(x_{kv}-\overline{x}_{.v})}{\left\{\sum_{k=1}^{N}(x_{ku}-\overline{x}_{.u})^2\sum_{l=1}^{N}(x_{lv}-\overline{x}_{.v})^2\right\}^{1/2}}$$

where $\overline{x}_{i.} = (\sum_{a=1}^{p} x_{ia})/p$ and $\overline{x}_{.u} = (\sum_{k=1}^{N} x_{ku})/N$.

The correlation similarity measure takes values between -1 and 1. With this measure, the relative direction of the two vectors is important. The correlation similarity measure is related to the angular separation similarity measure (described next). The correlation similarity measure gives the cosine of the angle between the two vectors measured from the mean; see Gordon (1999).

`angular` (alias `angle`)

requests the angular separation similarity measure. For comparing observations i and j, the formula is

$$\frac{\sum_{a=1}^{p} x_{ia}x_{ja}}{\left(\sum_{a=1}^{p} x_{ia}^2 \sum_{b=1}^{p} x_{jb}^2\right)^{1/2}}$$

and for comparing variables u and v, the formula is

$$\frac{\sum_{k=1}^{N} x_{ku}x_{kv}}{\left(\sum_{k=1}^{N} x_{ku}^2 \sum_{l=1}^{N} x_{lv}^2\right)^{1/2}}$$

The angular separation similarity measure is the cosine of the angle between the two vectors measured from zero and takes values from -1 to 1; see Gordon (1999).

Similarity measures for binary data

Similarity measures for binary data are based on the four values from the cross-tabulation of observation i and j (when comparing observations) or variables u and v (when comparing variables).

For comparing observation i and j, the cross-tabulation is

		obs. j	
		1	0
obs.	1	a	b
i	0	c	d

a is the number of variables where observations i and j both had ones, and d is the number of variables where observations i and j both had zeros. The number of variables where observation i is one and observation j is zero is b, and the number of variables where observation i is zero and observation j is one is c.

For comparing variables u and v, the cross-tabulation is

		var. v	
		1	0
var.	1	a	b
u	0	c	d

a is the number of observations where variables u and v both had ones, and d is the number of observations where variables u and v both had zeros. The number of observations where variable u is one and variable v is zero is b, and the number of observations where variable u is zero and variable v is one is c.

Stata treats nonzero values as one when a binary value is expected. Specifying one of the binary similarity measures imposes this behavior unless some other option overrides it (for instance, the `allbinary` option of `matrix dissimilarity`; see [MV] **matrix dissimilarity**).

Hubálek (1982) gives an extensive list of binary similarity measures. Gower (1985) lists 15 binary similarity measures, 14 of which are implemented in Stata. (The excluded measure has many cases where the quantity is undefined, so it was not implemented.) Anderberg (1973) gives an interesting table where many of these measures are compared based on whether the zero–zero matches are included in the numerator, whether these matches are included in the denominator, and how the weighting of matches and mismatches is handled. Hilbe (1992b, 1992a) implemented an early Stata command for computing some of these (as well as other) binary similarity measures.

The formulas for some of these binary similarity measures are undefined when either one or both of the vectors (observations or variables depending on which are being compared) are all zeros (or, sometimes, all ones). Gower (1985) says concerning these cases, "These coefficients are then conventionally assigned some appropriate value, usually zero."

The following binary similarity coefficients are available. Unless stated otherwise, the similarity measures range from 0 to 1.

`matching`

requests the simple matching (Zubin 1938, Sokal and Michener 1958) binary similarity coefficient

$$\frac{a+d}{a+b+c+d}$$

which is the proportion of matches between the 2 observations or variables.

`Jaccard`

requests the Jaccard (1901, 1908) binary similarity coefficient

$$\frac{a}{a+b+c}$$

which is the proportion of matches when at least one of the vectors had a one. If both vectors are all zeros, this measure is undefined. Stata then declares the answer to be one, meaning perfect agreement. This is a reasonable choice for most applications and will cause an all-zero vector to have similarity of one only with another all-zero vector. In all other cases, an all-zero vector will have Jaccard similarity of zero to the other vector.

The Jaccard coefficient was discovered earlier by Gilbert (1884).

`Russell`

requests the Russell and Rao (1940) binary similarity coefficient

$$\frac{a}{a+b+c+d}$$

`Hamann`

requests the Hamann (1961) binary similarity coefficient

$$\frac{(a+d)-(b+c)}{a+b+c+d}$$

which is the number of agreements minus disagreements divided by the total. The Hamann coefficient ranges from -1, perfect disagreement, to 1, perfect agreement. The Hamann coefficient is equal to twice the simple matching coefficient minus 1.

`Dice`

requests the Dice binary similarity coefficient

$$\frac{2a}{2a+b+c}$$

suggested by Czekanowski (1932), Dice (1945), and Sørensen (1948). The Dice coefficient is similar to the Jaccard similarity coefficient but gives twice the weight to agreements. Like the Jaccard coefficient, the Dice coefficient is declared by Stata to be one if both vectors are all zero, thus avoiding the case where the formula is undefined.

`antiDice`

requests the binary similarity coefficient

$$\frac{a}{a+2(b+c)}$$

which is credited to Anderberg (1973) but was shown earlier by Sokal and Sneath (1963, 129). We did not call this the Anderberg coefficient because there is another coefficient better known by that name; see the `Anderberg` option. The name `antiDice` is our creation. This coefficient takes the opposite view from the Dice coefficient and gives double weight to disagreements. As with the Jaccard and Dice coefficients, the anti-Dice coefficient is declared to be one if both vectors are all zeros.

`Sneath`

requests the Sneath and Sokal (1962) binary similarity coefficient

$$\frac{2(a+d)}{2(a+d)+(b+c)}$$

which is similar to the simple matching coefficient but gives double weight to matches. Also compare the Sneath and Sokal coefficient with the Dice coefficient, which differs only in whether it includes d.

`Rogers`

requests the Rogers and Tanimoto (1960) binary similarity coefficient

$$\frac{a+d}{(a+d)+2(b+c)}$$

which takes the opposite approach from the Sneath and Sokal coefficient and gives double weight to disagreements. Also compare the Rogers and Tanimoto coefficient with the anti-Dice coefficient, which differs only in whether it includes d.

`Ochiai`

requests the Ochiai (1957) binary similarity coefficient

$$\frac{a}{\{(a+b)(a+c)\}^{1/2}}$$

The formula for the Ochiai coefficient is undefined when one or both of the vectors being compared are all zeros. If both are all zeros, Stata declares the measure to be one, and if only one of the two vectors is all zeros, the measure is declared to be zero.

The Ochiai coefficient was presented earlier by Driver and Kroeber (1932).

`Yule`

requests the Yule (see Yule [1900] and Yule and Kendall [1950]) binary similarity coefficient

$$\frac{ad - bc}{ad + bc}$$

which ranges from -1 to 1. The formula for the Yule coefficient is undefined when one or both of the vectors are either all zeros or all ones. Stata declares the measure to be 1 when $b + c = 0$, meaning that there is complete agreement. Stata declares the measure to be -1 when $a + d = 0$, meaning that there is complete disagreement. Otherwise, if $ad - bc = 0$, Stata declares the measure to be 0. These rules, applied before using the Yule formula, avoid the cases where the formula would produce an undefined result.

`Anderberg`

requests the Anderberg binary similarity coefficient

$$\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d} \right) \Big/ 4$$

The Anderberg coefficient is undefined when one or both vectors are either all zeros or all ones. This difficulty is overcome by first applying the rule that if both vectors are all ones (or both vectors are all zeros), the similarity measure is declared to be one. Otherwise, if any of the marginal totals ($a+b$, $a+c$, $c+d$, $b+d$) are zero, then the similarity measure is declared to be zero.

Though this similarity coefficient is best known as the Anderberg coefficient, it appeared earlier in Sokal and Sneath (1963, 130).

`Kulczynski`

requests the Kulczyński (1927) binary similarity coefficient

$$\left(\frac{a}{a+b} + \frac{a}{a+c} \right) \Big/ 2$$

The formula for this measure is undefined when one or both of the vectors are all zeros. If both vectors are all zeros, Stata declares the similarity measure to be one. If only one of the vectors is all zeros, the similarity measure is declared to be zero.

`Pearson`

requests Pearson's (1900) ϕ binary similarity coefficient

$$\frac{ad - bc}{\{(a+b)(a+c)(d+b)(d+c)\}^{1/2}}$$

which ranges from -1 to 1. The formula for this coefficient is undefined when one or both of the vectors are either all zeros or all ones. Stata declares the measure to be 1 when $b + c = 0$, meaning that there is complete agreement. Stata declares the measure to be -1 when $a + d = 0$, meaning that there is complete disagreement. Otherwise, if $ad - bc = 0$, Stata declares the measure to be 0. These rules, applied before using Pearson's ϕ coefficient formula, avoid the cases where the formula would produce an undefined result.

`Gower2`

requests the binary similarity coefficient

$$\frac{ad}{\{(a+b)(a+c)(d+b)(d+c)\}^{1/2}}$$

which is presented by Gower (1985) but appeared earlier in Sokal and Sneath (1963, 130). Stata uses the name `Gower2` to avoid confusion with the better-known Gower coefficient, which is used with a mix of binary and continuous data.

The formula for this similarity measure is undefined when one or both of the vectors are all zeros or all ones. This is overcome by first applying the rule that if both vectors are all ones (or both vectors are all zeros) then the similarity measure is declared to be one. Otherwise, if $ad = 0$, the similarity measure is declared to be zero.

Dissimilarity measures for mixed data

Here is one measure that works with a mix of binary and continuous data. Binary variables are those containing only zeros, ones, and missing values; all other variables are treated as continuous.

`Gower`

requests the Gower (1971) dissimilarity coefficient for a mix of binary and continuous variables. For comparing observations i and j, the formula is

$$\frac{\sum_v \delta_{ijv} d_{ijv}}{\sum_v \delta_{ijv}}$$

where δ_{ijv} is a binary indicator equal to 1 whenever both observations i and j are nonmissing for variable v, and zero otherwise. Observations with missing values are not included when using `cluster` or `mds`, and so if an observation is included, $\delta_{ijv} = 1$ and $\sum_v \delta_{ijv}$ is the number of variables. However, using `matrix dissimilarity` with the `Gower` option does not exclude observations with missing values. See [MV] **cluster**, [MV] **mds**, and [MV] **matrix dissimilarity**.

For binary variables v,

$$d_{ijv} = \begin{cases} 0 & \text{if } x_{iv} = x_{jv} \\ 1 & \text{otherwise} \end{cases}$$

This is the same as the `matching` measure.

For continuous variables v,

$$d_{ijv} = \frac{|x_{iv} - x_{jv}|}{\left\{\max_k(x_{kv}) - \min_k(x_{kv})\right\}}$$

d_{ijv} is set to 0 if $\max_k(x_{kv}) - \min_k(x_{kv}) = 0$, that is, if the range of the variable is zero. This is the `L1` measure divided by the range of the variable.

For comparing variables u and v, the formula is

$$\frac{\sum_i \delta_{iuv} d_{iuv}}{\sum_i \delta_{iuv}}$$

where δ_{iuv} is a binary indicator equal to 1 whenever both variables u and v are nonmissing for observation i, and zero otherwise. If there are no missing values, $\sum_i \delta_{iuv}$ is the number of observations; otherwise, it is the number of observations for which neither variable u nor v has a missing value.

If all the variables are binary,

$$d_{iuv} = \begin{cases} 0 & \text{if } x_{iu} = x_{iv} \\ 1 & \text{otherwise} \end{cases}$$

If at least one variable is continuous,

$$d_{iuv} = \frac{|x_{iu} - x_{iv}|}{\left\{\max_v(x_{iv}) - \min_v(x_{iv})\right\}}$$

d_{iuv} is set to 0 if $\max_v(x_{iv}) - \min_v(x_{iv}) = 0$, that is, if the range of the observation is zero.

The Gower measure interprets binary variables as those with only 0, 1, or missing values. All other variables are treated as continuous.

In [MV] **matrix dissimilarity**, missing observations are included only in the calculation of the `Gower` dissimilarity, but the formula for this dissimilarity measure is undefined when all the values of δ_{ijv} or δ_{iuv} are zero. The dissimilarity is then set to missing.

❑ Technical note: Matrix dissimilarity and the Gower measure

Normally the commands

```
. matrix dissimilarity gm = x1 x2 y1, Gower
. clustermat waverage gm, add
```

and

```
. cluster waverage x1 x2 y1, measure(Gower)
```

will yield the same results, and likewise with `mdsmat` and `mds`. However, if any of the variables contain missing observations, this will not be the case. `cluster` and `mds` exclude all observations that have missing values for any of the variables of interest, whereas `matrix dissimilarity` with the `Gower` option does not. See [MV] **cluster**, [MV] **mds**, and [MV] **matrix dissimilarity** for more information.

Note: `matrix dissimilarity` without the `Gower` option does exclude all observations that have missing values for any of the variables of interest.

❑

❑ Technical note: Binary similarity measures applied to averages

Some cluster-analysis methods (such as Stata's kmeans and kmedians clustering) need to compute the similarity or dissimilarity between observations and group averages or group medians; see [MV] **cluster**. With binary data, a group average is interpreted as a proportion.

A group median for binary data will be zero or one, except when there are an equal number of zeros and ones. Here Stata calls the median 0.5, which can also be interpreted as a proportion.

In Stata's `cluster kmeans` and `cluster kmedians` commands for comparing a binary observation to a group proportion (see *Partition cluster-analysis methods* in [MV] **cluster**), the values of a, b, c, and d are obtained by assigning the appropriate fraction of the count to these values. In our earlier table showing the relationship of a, b, c, and d in the cross-tabulation of observation i and observation j, we replace observation j by the group-proportions vector. Then when observation i is 1, we add the corresponding proportion to a and add one minus that proportion to b. When observation i is 0, we add the corresponding proportion to c and add one minus that proportion to d. After the values of a, b, c, and d are computed in this way, the binary similarity measures are computed using the formulas as already described.

❑

John Clifford Gower (1930–2019) was born in London. He studied mathematics and statistics at the Universities of Cambridge and Manchester. From 1955 until his retirement in 1990, he worked at Rothamsted Experimental Station in Hertfordshire (where R. A. Fisher, W. G. Cochran, F. Yates, J. A. Nelder, and R. W. M. Wedderburn also worked at various times). Gower's initial focus was on computing: the Elliott 401 computer then at Rothamsted (now visible in the Science Museum in London) was probably the first in the world to be devoted entirely to agricultural and biological research. From the mid 1960s, his main emphasis was applied multivariate analysis, especially classification problems and graphical methods. That led to first-authored books on biplots and Procrustes problems and to several highly cited papers. In retirement, Gower was long associated with the Open University. He traveled and collaborated widely and actively supported several learned societies.

Paul Jaccard (1868–1944) was a Swiss botanist who was born in Sainte-Croix (Vaud) and died in Zürich. He studied at Lausanne, Zürich, and Paris before being appointed to posts at Lausanne in 1894, where he specialized in plant geography, undertaking fieldwork in Egypt, Sweden, and Turkestan. In 1903, Jaccard returned to Zürich to a chair in general botany and plant physiology at ETH. His interests there centered on the microscopic analysis of wood, and anatomical and physiological studies of the growth of trees.

Robert Reuven Sokal (1926–2012) was born in Vienna to a Jewish family. He gained degrees from St. John's University in Shanghai and the University of Chicago. Sokal worked at the University of Kansas–Lawrence and (from 1969) the State University of New York–Stony Brook. He was one of the leaders in the development of numerical taxonomy (Sokal and Sneath 1963; Sneath and Sokal 1973) and was prominent in the application of statistical methods within biological systematics. With F. J. Rohlf, he authored one of the leading biometrics texts (Sokal and Rohlf 2011). In the latter stages of his career, his interests centered on genetic variation in human populations, European ethnohistory, and spatial statistics. Sokal was a member of the US National Academy of Sciences.

Peter Henry Andrews Sneath (1923–2011) was born in Ceylon (now Sri Lanka) and studied medicine in Cambridge and London. After military service, he specialized in microbial systematics and the application of computers to biomedical science, working for the Medical Research Council in the UK and the University of Leicester. With Robert Sokal, Sneath wrote the two initial texts on numerical taxonomy. He is a Fellow of the Royal Society. A bacterial taxon, the genus *Sneathia*, was named after him in 2002.

References

Anderberg, M. R. 1973. *Cluster Analysis for Applications.* New York: Academic Press. https://doi.org/10.1016/C2013-0-06161-0.

Cox, D. R. 2015. A conversation with John C. Gower. *International Statistical Review* 83: 339–356. https://doi.org/10.1111/insr.12094.

Czekanowski, J. 1932. "Coefficient of racial likeness" und "durchschnittliche Differenz". *Anthropologischer Anzeiger* 9: 227–249.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26: 297–302. https://doi.org/10.2307/1932409.

Driver, H. E., and A. L. Kroeber. 1932. Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology* 31: 211–256.

Futuyma, D. J. 2012. Robert R. Sokal (1926–2012). *Science* 336: 816. https://doi.org/10.1126/science.1224101.

Gilbert, G. K. 1884. Finley's tornado predictions. *American Meteorological Journal* 1: 166–172.

Gordon, A. D. 1999. *Classification.* 2nd ed. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9780367805302.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857–871. https://doi.org/10.2307/2528823.

———. 1985. "Measures of similarity, dissimilarity, and distance". In *Encyclopedia of Statistical Sciences*, edited by S. Kotz, N. L. Johnson, and C. B. Read, vol. 5: 397–405. New York: Wiley.

Hamann, U. 1961. Merkmalsbestand und Verwandtschaftsbeziehungen der Farinosae. Ein Beitrag zum System der Monokotyledonen. *Willdenowia* 2: 639–768.

Hilbe, J. M. 1992a. sg9.1: Additional statistics to similari output. *Stata Technical Bulletin* 10: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, p. 132. College Station, TX: Stata Press.

———. 1992b. sg9: Similarity coefficients for 2 × 2 binary data. *Stata Technical Bulletin* 9: 14–15. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 130–131. College Station, TX: Stata Press.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews* 57: 669–689. https://doi.org/10.1111/j.1469-185X.1982.tb00376.x.

Jaccard, P. 1901. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37: 241–272. https://doi.org/10.5169/seals-266440.

———. 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles* 44: 223–270. https://doi.org/10.5169/seals-268384.

Kaufman, L., and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley. https://doi.org/10.1002/9780470316801.

Kulczyński, S. 1927. Die Pflanzenassoziationen der Pieninen [In Polish, German summary]. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B (Sciences Naturelles)* Suppl. II: 57–203.

Ochiai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions [in Japanese, English summary]. *Bulletin of the Japanese Society of Scientific Fisheries* 22: 526–530.

Pearson, K. 1900. Mathematical contributions to the theory of evolution—VII. On the correlation of characters not quantitatively measureable. *Philosophical Transactions of the Royal Society,* A ser., 195: 1–47. https://doi.org/10.1098/rsta.1900.0022.

Rogers, D. J., and T. T. Tanimoto. 1960. A computer program for classifying plants. *Science* 132: 1115–1118. https://doi.org/10.1126/science.132.3434.1115.

Ross, G. J. S. 2019. John Clifford Gower, 1930–2019. *Journal of the Royal Statistical Society,* A ser., 182: 1639–1641. https://doi.org/10.1111/rssa.12518.

Russell, P. F., and T. R. Rao. 1940. On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of the Malaria Institute of India* 3: 153–178.

Sneath, P. H. A. 1995. Thirty years of numerical taxonomy. *Systematic Biology* 44: 281–298. https://doi.org/10.1093/sysbio/44.3.281.

———. 2010. Reflections on microbial systematics. *Bulletin of Bergey's International Society for Microbial Systematics* 1: 77–83.

Sneath, P. H. A., and R. R. Sokal. 1962. Numerical taxonomy. *Nature* 193: 855–860. https://doi.org/10.1038/193855a0.

———. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco: Freeman.

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 28: 1409–1438.

Sokal, R. R., and F. J. Rohlf. 2011. *Biometry*. 4th ed. New York: Freeman.

Sokal, R. R., and P. H. A. Sneath. 1963. *Principles of Numerical Taxonomy*. San Francisco: Freeman.

Sørensen, T. J. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Royal Danish Academy of Sciences and Letters, Biological Series* 5: 1–34.

Yule, G. U. 1900. On the association of attributes in statistics: With illustrations from the material of the Childhood Society, etc. *Philosophical Transactions of the Royal Society*, A ser., 194: 257–319. https://doi.org/10.1098/rsta.1900.0019.

Yule, G. U., and M. G. Kendall. 1950. *An Introduction to the Theory of Statistics*. 14th ed. New York: Hafner.

Zubin, J. 1938. A technique for measuring like-mindedness. *Journal of Abnormal and Social Psychology* 33: 508–516. https://doi.org/10.1037/h0055441.

Also see

[MV] **matrix dissimilarity** — Compute similarity or dissimilarity measures

[MV] **cluster** — Introduction to cluster-analysis commands

[MV] **clustermat** — Introduction to clustermat commands

For suggested citations, see the FAQ on citing Stata documentation.