Phenetic approach

Cluster analysis
Ordination methods
Discriminant analysis

Scales

•Nominal Scale

From mathematical operations, only equality (=) or inequality (≠) applies here.

•Ordinal Scale

In addition to equality and inequality, operators such as < and > also apply here.

•Interval Scale

In addition to the properties of the two previous scales, addition and subtraction are also possible here (values can also take the value 0).

•Ratio Scale

Allows expressing the ratio between objects (the division operator can also be used).

Classification of characters

(1) Qualitative:

- Binary (two-state, two-valued, alternative)
- Multistate (multivalued)

(2) Semiquantitative

(3) Quantitative:

- Discontinuous (discrete, meristic)
- Continuous

Conversion of a four-state qualitative trait/character into a system of binary traits/characters

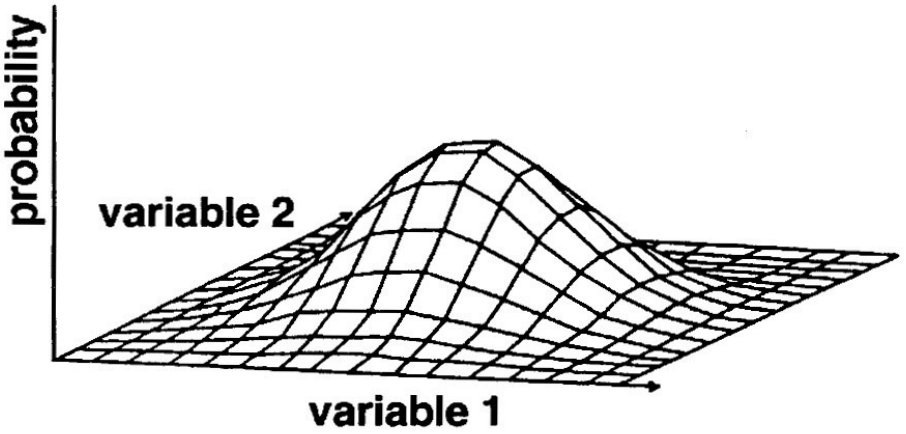
Qualitative character states	Artificial binary characters			
a	**1**	**0**	**0**	0
b	**0**	**1**	**0**	0
c	**0**	**0**	**1**	0
d	**0**	**0**	**0**	1

Transformation of data

Some multivariate methods do not require a **normal distribution of data**, or they are sufficiently robust in relation to deviations from the normal distribution of data (e.g., cluster analysis, PCA, etc.)

Other methods require a multivariate normal distribution of data (e.g., discriminant analysis).

By **transformation**, it is sometimes possible to **approximate the distribution of data to a normal distribution**.



Probability density diagram for a bivariate normal distribution

Transformation of data

• **Constants and functions independent of the analyzed data are used for transformation.**

• **Linear transformation** (e.g., multiplying variables by a constant) — if applied to all variables, the results of the analysis do not change; if applied to only one or a few variables, it leads to weighting them.

• **Nonlinear transformation** changes the structure of the data.

Transformation of data

Logarithmic transformation:
Measured values are replaced by their logarithms.

$$x'_{ij} = \log_c x_{ij}$$

Since the logarithm of zero is not defined, in such cases, a constant of 1 or 0.5 is added to each measured value of the given variable.
The formula then takes the form

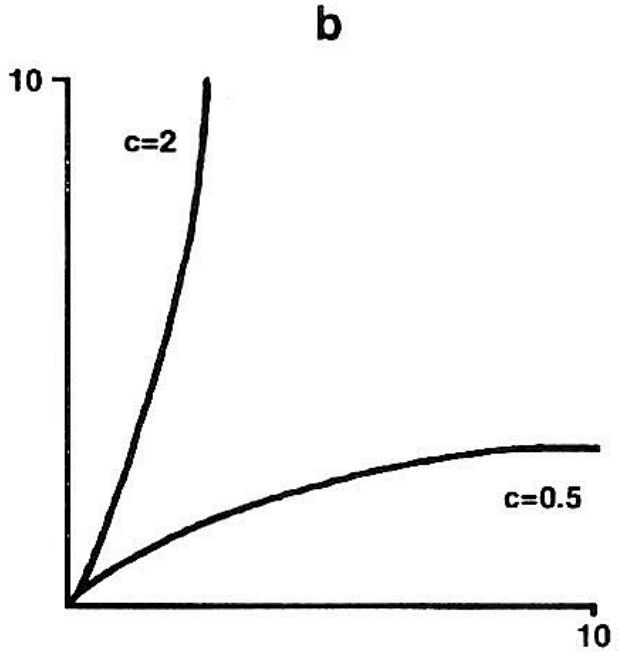
$$x'_{ij} = \log_c (x_{ij}+1)$$

Transformation of data

Square root transformation

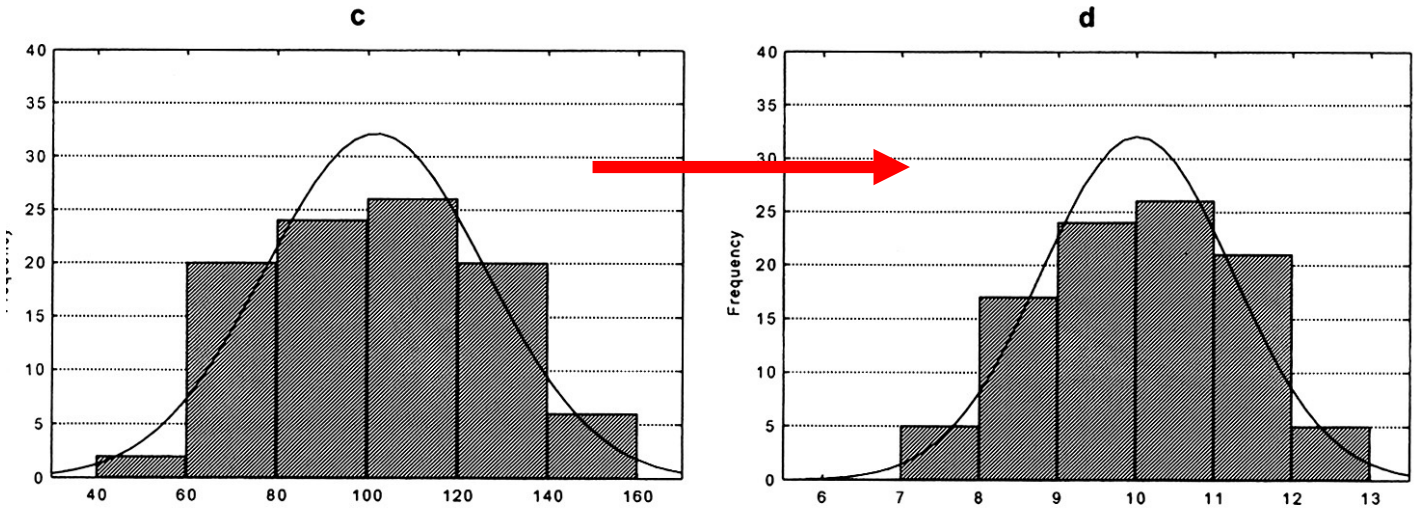
generally $x'_{ij} = x^{c}_{ij}$

c > 1 high numerical values are emphasized – used rarely
c < 1 high numerical values are underestimated
c = 0.5 - square root transformation



Transformation of data

Square root transformation



Variables must not reach zero values, therefore it is sometimes used in the form

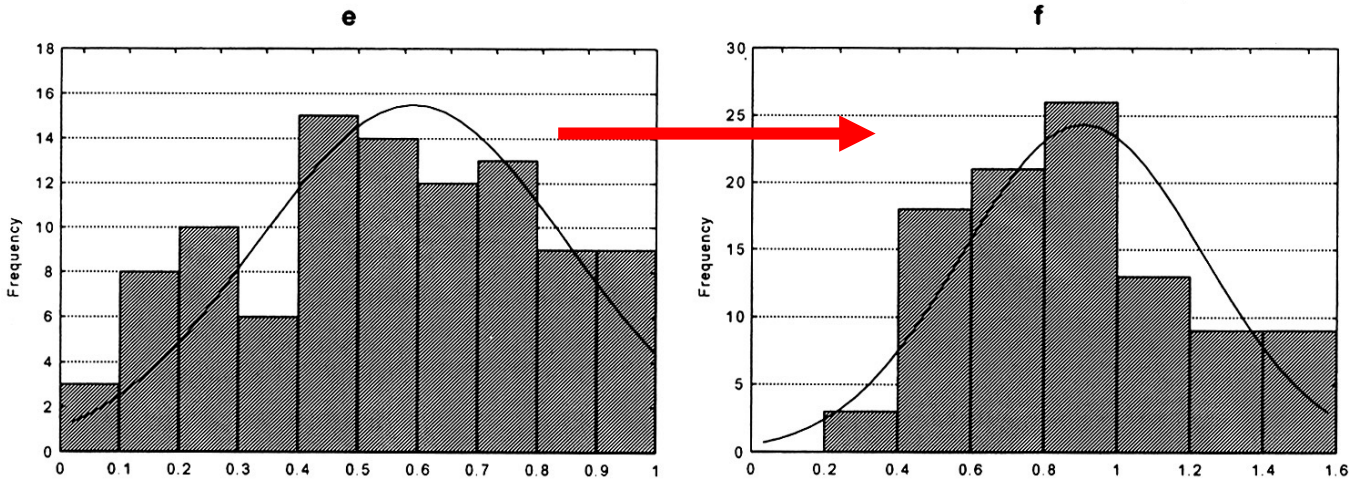
$$x'_{ij} = \sqrt{x_{ij} + 0.5}$$

Transformation of data

Arc sin transformation

$x'_{ij} = \text{arc sin } x_{ij}$

It is also used in combination with square root transformation;
the arcsine transformation assumes that the data are measured in the interval <0, 1>.
If this is not the case, the measured values can be divided by constants 10, 100, 1000, etc.



Standardization of data

For standardization, statistics derived from the analyzed data set are used (range, standard deviation, mean, maximum, etc.).
Variables are thus transformed to the same scale (in other words, the actual dimension of the corresponding variable no longer matters).

Centering (centering, standardization to a mean equal to zero):

$$x'_{ij} = x_{ij} - \underline{x_i}$$

Centering does not change the units in which the variables are measured; only the position of the zero point in the coordinate system is changed.

Standardization by range, ranging

$$x'_{ij} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}}$$

It is recommended to use in cases where the variables are measured on the same scale, but there are very large differences between their values; the values of the variables are converted into the interval [0,1].

Standardization of data

Standardization by standard deviation

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$$

where s_i is the standard deviation of variable *i*.

It is recommended to use in cases where variables are measured on different scales and units.

Resemblance/distance coefficients

(1) Distance coefficients for quantitative and binary variables (metric distances)

(2) Binary similarity coefficients

(3) Coefficients for mixed data

(4) Correlation coefficients

Metrics (distances)

If the distance coefficients meet the following requirements, they are considered metrics:

(1) Symmetry – for the distance between two objects (x, y), it holds that:

$$d(x,y) = d(y,x) \geq 0$$

(2) Triangular inequality – for the distance between three objects (x, y, z), it holds that:

$$d(x,y) \leq d(x,z) + d(y,z)$$

i.e., the distance between two objects is less than or equal to the sum of their distances from a third object.

(3) The distance between identical objects (and the distance of an object from itself) is 0:

$d(x,y) = 0$ in the case that $x=y$

(4) The distance between objects that are not identical is greater than 0 (it is positive):

$d(x,y) > 0$ v případě, že $x \neq y$.

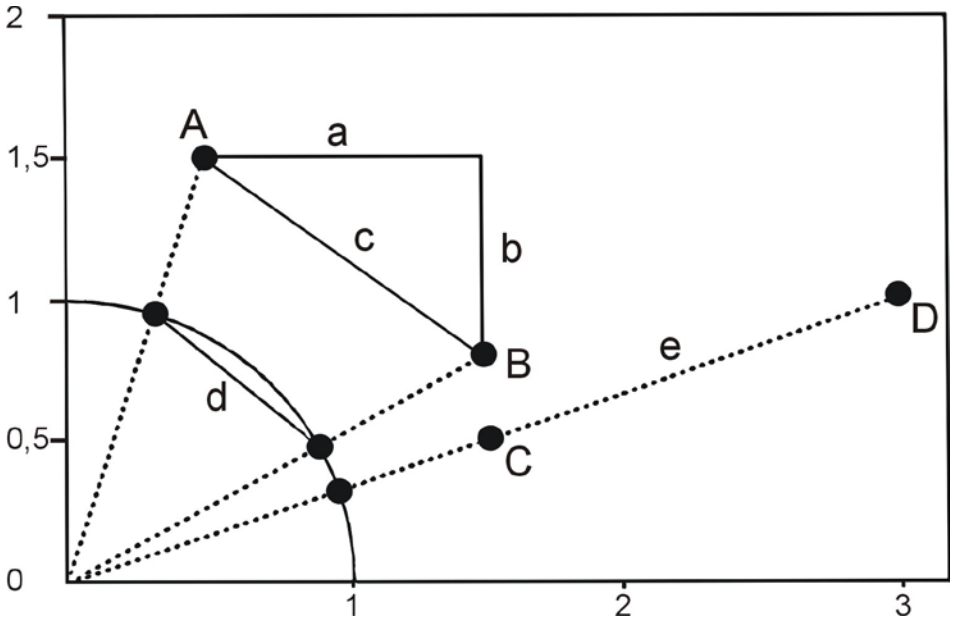
If the distance coefficients do not meet the triangular inequality criterion, they are considered pseudometrics (pseudometric, semimetric).

Metrics (distances)

Euclidean distance:

$EU=c$ $\qquad EU_{jk} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ik}\right)^2}$

where x_{ij} is the value of variable i for object j, x_{ik} is the value of variable i for object k, and n is the total number of variables.



Euclidean distance depends on the scale of variables

	Weight in pounds	Height in feet	Height in inches
A	60	3.0	36.0
B	65	3.5	42.0
C	63	4.0	48.0

$= (60 – 65)^2 + (3.0 – 3.5)^2 = 25.25$ $[(60 – 65)^2 + (36.0 – 42.0)^2 = 61]$

$= (60 – 63)^2 + (3.0 – 4.0)^2 = 10.00$ $[(60 – 63)^2 + (36.0 – 48.0)^2 = 153]$

$= (65 – 63)^2 + (3.5 – 4.0)^2 = 4.25$ $[(65 – 63)^2 + (42.0 – 48.0)^2 = 40]$

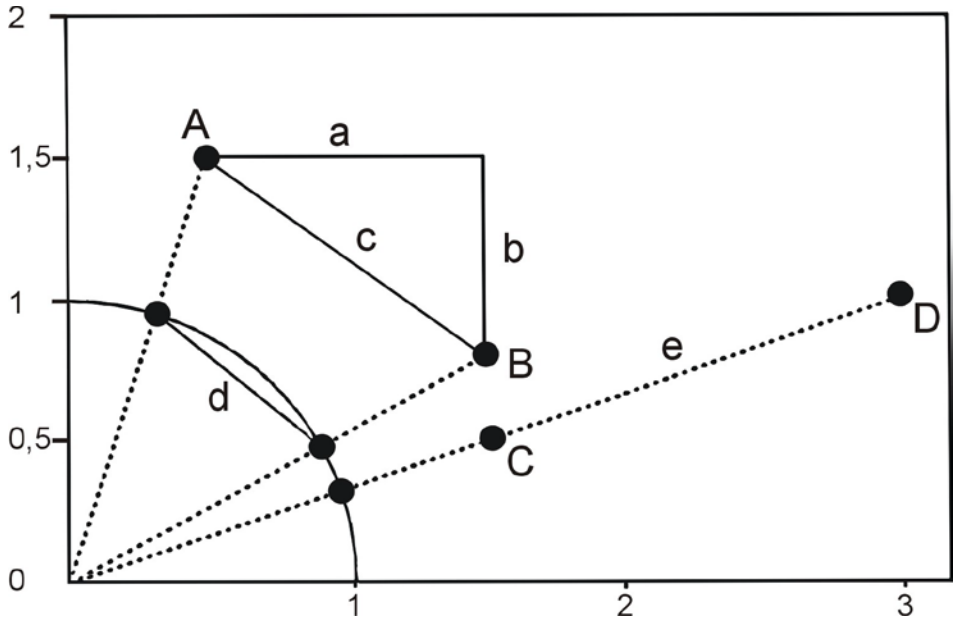
Metrics (distances)

Manhattan (city block) metric:

$CB = a + b$

$$CB_{jk} = \sum_{i=1}^{n} \left| x_{ij} - x_{ik} \right|$$

It resembles North American cities with perpendicular streets, where one has to walk around the blocks.



Minkowski metric:

$$MNK_{jk} = \sqrt[r]{\sum_{i=1}^{n} \left(x_{ij} - x_{ik} \right)^{r}}$$

kde $r \geq 1$;
pro $r=1$ …. CB
***Pro* $r=2$ … EU**

Distances

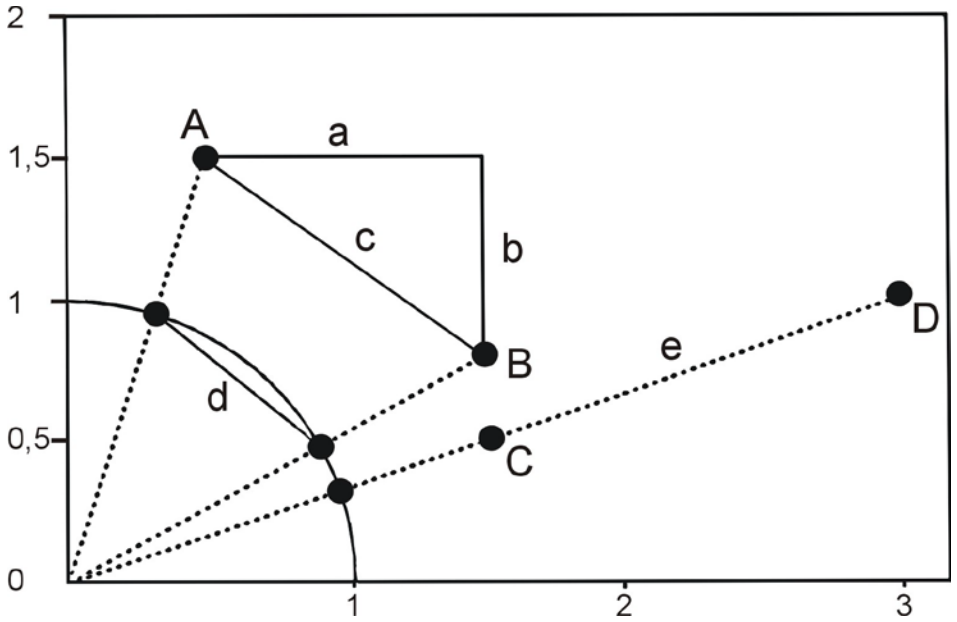
Chord distance

For two variables, the chord distance is the straight-line distance between the projections of the points on a circle with a unit radius

$$CH = d$$

$$CH_{jk} = \left(2\left[1 - \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2}} \right] \right)^{1/2}$$

The chord distance reaches the same values in cases where two or more objects exhibit proportionally the same values in all variables, without the specific values of these variables having to be the same for all objects (distance between points C and D).
It is not a true metric.



Similarity coefficients for binary data

Any function d is a **dissimilarity** if it meets at least the first three rules about metrics

(if $j = k$, then $d_{jk} = 0$; if $j \neq k$, then $d_{jk} > 0$; $d_{jk} = d_{kj}$);

Most dissimilarity functions have a lower bound of 0 and an upper bound of 1:

$$0 \leq d_{jk} \leq 1$$

Most dissimilarity functions, after transformation $(d_{jk})^{1/2}$, satisfy all the rules about metrics and then represent **distances**

We usually consider **similarity**: $s_{jk} = 1 - d_{jk}$

For identical objects, it holds that $s_{jk} = 1$

Similarity coefficients for binary data

Selection of similarity coefficient

	object 2		
object 1		1	0
	1	*a*	*b*
	0	*c*	*d*

a – the number of variables where both objects have the value + (or 1) (positive match)

b – the number of variables where object i has the value − (or 0) and object j has the value + (or 1)

c – the number of variables where object i has the value + (or 1) and object j has the value − (or 0)

d – the number of variables where both objects have the value − (or 0) (negative match)

The choice between coefficients depends primarily on whether negative matches make sense for the given variables, i.e., whether it makes sense to consider that the zero value of a variable has the same cause in the compared objects.

Similarity coefficients for binary data

Coefficients evaluating *a* and *d* symmetrically :

	object 2		
object 1		1	0
	1	*a*	*b*
	0	*c*	*d*

Simple matching coefficient:

Coefficient is close/similar to ED:

$ED^2 = n(1\text{-}SM)$

$n = a+b+c+d$

$ED = \sqrt{b+c}$

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

Rogers and Tanimoto coefficient

Disagreements are weighted twice;
values are always lower than with SM, except when b + c = 0

$$RT = \frac{a+d}{a+2b+2c+d}$$

Hamann index:

range [-1,1]

$$HAM = \frac{a+d-b-c}{a+d+b+c}$$

$$SM = \frac{HAM+1}{2}$$

Similarity coefficients for binary data

	object 2		
object 1		1	0
	1	*a*	*b*
	0	*c*	*d*

Coefficients evaluating *a* and *d* symmetrically:
d is taken into consideration, but *a* and *d* are not weighted equally

Baroni-Urbani – Buser II: $BB2 = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$

modified *SM*, *d* -> geometric average *a* and *d*
range [0,1]

Baroni-Urbani – Buser I: $BB1 = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c}$

modified *HAM*, *d* -> geometric average *a* and *d*
range [0,1]

Russell and Rao coefficient: $RR = \frac{a}{a + b + c + d}$

Increasing the value of d reduces the value of dissimilarity.

Similarity coefficients for binary data

Coefficients that do not take negative matches into account:

	object 2		
		1	0
object 1	1	*a*	*b*
	0	*c*	*d*

Jaccard coefficient: $$JAC = \frac{a}{a+b+c}$$

range [0,1]

conversion $$d_{jk} = \sqrt{1 - s_{jk}}$$
results in Euclidean distanca

Sørensen coefficient: $$SOR = \frac{2a}{2a+b+c}$$

Positive matches are weighted twice

Genetic distances according to Nei & Li (1979), Link et al. (1995) used in NJ, PCoA also correspond to this type of coefficients

Nei & Li (1979):

$$NL = 1 - \frac{2a}{2a+b+c}$$

Link et al. (1995):

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

Coefficients for mixed data

This category includes the Gower coefficient and distance for mixed data. They are used in cases where the matrix contains qualitative variables, quantitative variables, or binary variables (or all three types of variables).

Gower coefficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j– objects characterized by variable k,
n – total number of variables,
s_{ijk} – score of variable k

w_{ijk} is a weight that can take values of 1 or 0 depending on whether it is possible to compare the values of variable k for objects i and j (except for binary variables, it can only have a value of 0 if the value of variable kkk is unknown for one or both objects);
s_{ijk} is the score (value) for the corresponding variable k.
a) For binary variables:
$w_{ijk} = 1$ and $s_{ijk} = 0$ if $x_{ik} \neq x_{jk}$ $w_{ijk} = 1$ and $s_{ijk} = 0$ if $x_{ik} \neq x_{jk}$ (values of variable k for objects i and j)
$w_{ijk} = s_{ijk} = 1$ if $x_{ik} = x_{jk} = 1$ or if $x_{ik} = x_{jk} = 0$ and negative matching is taken into account (corresponds to the simple matching coefficient)
$w_{ijk} = s_{ijk} = 0$ if $x_{ik} = x_{jk} = 0$ and negative matching is not taken into account (corresponds to Jaccard's coefficient)

Coefficients for mixed data

Gower coefficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j– objects characterized by variable k,
n – total number of variables,
s_{ijk} – score of variable k

b) for nominal variables:

$w_{ijk} = 1$ if x_{ik} and x_{jk} are known; then
$s_{ijk} = 0$ if $x_{ik} \neq x_{jk}$; $s_{ijk} = 1$ if $x_{ik} = x_{jk}$ (the number of variable states is not taken into account)

Coefficients for mixed data

Gower coefficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objects characterized by variable k,
n – total number of variables,
s_{ijk} – score of variable k

c) for qualitative variables:
$w_{ijk} = 1$ if x_{ik} and x_{jk} are both known, and $s_{ijk} = 1 - \{|x_{ik} - x_{jk}| / (\text{range of variable } i)\}$
(corresponds to the Manhattan metric with data standardized by range)

Coefficients for mixed data

example:

Taxon / variable	Branching of stem	Colour of petals	Character of leaves	Average height of plants (cm)	Average length of petals(mm)
1	1	white (1)	simple (1)	30	2,6
2	1	red (2)	pinnate (2)	25	2,3
3	0	blue (3)	pinnate (2)	10	8,5
4	0	blue (3)	palmately lobed (3)	80	8,2

$$GOW(1,2) = \frac{[1\times 1]+[1\times 0]+[1\times 0]+\left[1\times\left(1-\frac{|30-25|}{80-10}\right)\right]+\left[1\times\left(1-\frac{|2.6-2.3|}{8.5-2.3}\right)\right]}{1+1+1+1+1} = 0.576$$

Correlation coefficients

Pearson correlation coefficient

$$r_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

n number of objects,
x_{i1} value of variable 1 for object i

Linear correlation assumes a normal distribution of data.

Spearman correlation coefficient (rank koeficient):

$$r_{12} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Specific values of the variables are not taken into account, but rather the rank order of the objects, where d_i is the difference in rank between objects;

Pearson correlation coefficient and Spearman correlation coefficient:
range [-1, +1], +1 indicates direct correlation, -1 indicates inverse correlation, 0 indicates no relationship