# 3

# Távolság, hasonlóság, korreláció...

## *(Az adatmátrixból egy másik mátrixba)*

Az elemzés első, meghatározó szakasza a mintavétellel és az adatok esetleges átalakításával lezárult. Ezután már arra kell összpontosítanunk, hogy miként "hámozhatjuk ki" a sokdimenziós adattérben rejlő információt, hogyan tárhatjuk fel az objektumok közötti kapcsolatrendszert. Az első lépés ebben a meglehetősen komplikált folyamatban a pontok közötti távolságok – vagy más, rokon jellegű összefüggések (hasonlóság, különbözőség, korreláció) – kiszámítása. (Megjegyzendő persze, hogy bizonyos módszerek egyszerűen megkerülik ezt a lépcsőfokot, amint arra a 0.1 ábra is utalt.)

### 3.1 Alapfogalmak

*3.1.1 Metrikák, az euklidészi távolság*

Mielőtt áttekintenénk a cím alapján első látásra is sokrétű terminológiát, tisztáznunk kell: mit is értünk valójában *távolságon?* Köznapi értelemben nincs különösebb gond: két pont távolsága a közöttük meghúzható egyenes szakasz hosszúsága. Ez az úgynevezett *euklidészi távolság* kiterjeszthető akármennyi dimenzióra is (lásd a 3.47 formulát). Még ha a sokdimenziós esetet nem is tudjuk elképzelni, a köznapi távolságfogalom jelenti a legjobb kiindulópontot a többi távolság és hasonlóság tárgyalásához. Ha *n* pontunk van, akkor a közöttük minden lehetséges párosításban kiszámított távolságok egy újabb mátrixba, a *távolságmátrixba* írhatók be. A 2.1 adatmátrix három oszlopára (egyedére) nézve a távolságmátrix a következő lesz:

| | 1. egyed | 2. egyed | 3. egyed |
|---|---|---|---|
| 1. egyed | 0 | 3,0 | 3,0 |
| 2. egyed | 3,0 | 0 | 5,1 |
| 3. egyed | 3,0 | 5,1 | 0 |

azaz, "hivatalos" formában

$$D_{3,3} = \begin{bmatrix} 0 & 3.0 & 3.0 \\ 3.0 & 0 & 5.1 \\ 3.0 & 5.1 & 0 \end{bmatrix} \qquad (3.1)$$

Az euklidészi távolság csak egy – bár kiemelt jelentőségű esete – egy általános függvénycsoportnak, a *metrikáknak*. Adataink feldolgozásában nagyon sokféle metrika jöhet számításba. Metrikának tekintünk minden olyan $d_{jk}$ függvényt, amely az összes pontra nézve megfelel a következő feltételeknek (metrikus axiómák):

1) Amennyiben két pont egybeesik, azaz $j=k$, akkor $d_{jk} = 0$. ($d_{jk}$ akkor és csak akkor 0, ha $j=k$.)

2) Ha két pont különböző, azaz $j \neq k$, akkor $d_{jk} > 0$.

3) A *szimmetriaaxióma* szerint $d_{jk} = d_{kj}$ (azaz mindegy, hogy a távolságot melyik irányból mérjük).

A fenti három axióma jól láthatóan "érvényesül" a 3.1 mátrixban. Az átlóban 0-k, az átlón kívül pozitív értékek szerepelnek, az egész mátrix pedig az átlóra nézve szimmetrikus. Így elegendő lenne a bal alsó sarokban levő három értéket megadni ("alsó félmátrix"), amint azt gyakran meg is tesszük (pl. a 3.2 mátrix).

4) A metrikus sajátság igen fontos, megkülönböztető kritériuma a *háromszög-egyenlőtlenség* axiómája. Eszerint $d$ csak akkor metrika, ha bármely három $i, j, k$ pontra igaz a következő összefüggés: $d_{ij} + d_{ik} \geq d_{jk}$. Szavakban: két pont távolsága nem lehet nagyobb, mint egy harmadik ponttól vett távolságaik összege.

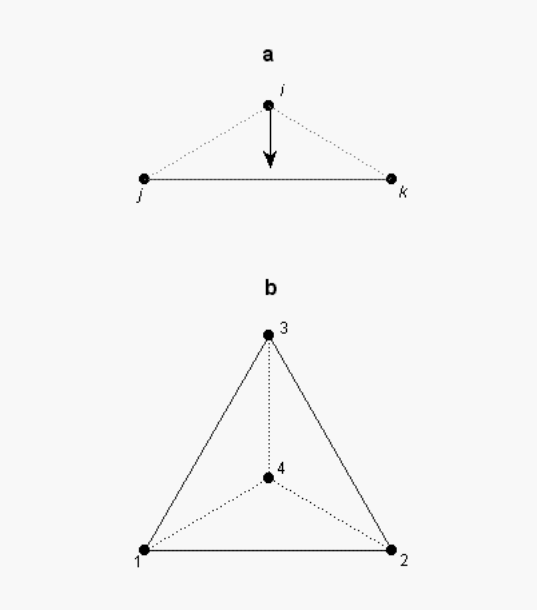
Ezt könnyen beláthatjuk a kétdimenziós esetre a 3.1a ábra segítségével. Adottak az $j$ és a $k$ pontok, és ekkor kellene a harmadik, $i$ pontot úgy megkeresni, hogy a másik kettőtől vett távolságainak összege kisebb legyen $d_{jk}$-nál. Látjuk, hogy az euklidészi távolság esetén ez lehetetlen, a $d_{ij}+d_{ik}$ összeg akkor lesz a legkisebb, ha az $i$ pont éppen ráesik a $jk$ egyenesre. Bárhová is mozgatjuk az $i$ pontot, a távolságösszeg szükségképpen növekszik, a háromszög-egyenlőtlenség tehát fennáll.

Felmerülhet persze mindenkiben a kérdés, hogy tudunk-e olyan egyszerű példát szerkeszteni, amelyben teljesül a háromszög-egyenlőtlenség, és az euklidészi távolságok felrajzolása mégsem sikerül. A 3.1b ábra segítségével, négy pont alapján elképzelhetünk egy ilyen szituációt is. Legyen a négy pont közötti összes lehetséges távolságok alsó félmátrixa a következő:

$$D_{4,4} = \begin{bmatrix} 0 & & & \\ 3.0 & 0 & & \\ 3.0 & 3.0 & 0 & \\ 1.6 & 1.6 & 1.6 & 0 \end{bmatrix} \qquad (3.2)$$

A 3.1b ábra mutatja, hogy az 1., a 2. és a 3. pont egy egyenlő (3 egységnyi) oldalú háromszöget alkot. A 3.1a ábra példáját kiterjesztve gyorsan belátható, hogy a 4. pont akkor lesz a legközelebb a többihez, ha egy síkba kerül velük, mégpedig éppen a háromszög súlypontjába. De még ebben az esetben is $\sqrt{3} = 1{,}73$ távolságnyira van mindegyik ponttól, azaz a fenti "távolságmátrix" nem euklidészi. A metrikus feltételeknek viszont eleget tesz, hiszen 1,6 + 1,6 > 3,0; a háromszögegyenlőtlenség tehát teljesül.

Egy $d$ függvényről tehát az az erősebb állítás, hogy euklidészi, mert akkor metrikus is, míg ez fordítva – mint láttuk – nem feltétlenül igaz. A 3.2 mátrixot ugyan teljesen önkényesen




**3.1 ábra. a**: Három pontot nem tudunk úgy felrajzolni a papír síkjában, hogy ne teljesülne a háromszög-egyenlőtlenség. **b**: A négy pont azt szemlélteti, hogy a 3.2 mátrix nem euklidészi. A 4. pont távolsága bármelyik másiktól ugyanis nem lehet kisebb √3-nál.

töltöttük ki, az illusztráció végett, de valóban léteznek nem euklidészi metrikák is (lásd a 3.4 táblázatot). Mi megelégedhetünk azzal az egyszerű megfogalmazással, hogy minden *d* metrika euklidészi, ha a pontok elhelyezhetők egy olyan térben, amelyben *d* éppen a közöttük levő euklidészi távolság.

Az euklidészi tulajdonság precíz, mátrixalgebrai megfogalmazását pl. Gower & Legendre (1986) cikkében találhatjuk meg (lásd még Telegdi 1986). Ennek lényege az, hogy *d* euklidészi, ha a $\Delta_{m,m} = [-d_{jk}^2]$ mátrixra és egy tetszés szerinti $\mathbf{x}_m$ vektorra (azzal a feltétellel, hogy $\mathbf{x}'\mathbf{1}=0$) az alábbi összefüggés érvényes:

$$Q(\Delta) = \mathbf{x}'\Delta\mathbf{x} \geq 0 \qquad (3.3)$$

(kvadratikus alak, lásd a C függeléket). A 3.2 mátrixra a fenti egyenlőtlenség nem áll fenn, $\mathbf{x}' = [\,1\;\;1\;\;1\;\;-3\,]$ mellett például $Q(\Delta) = -3{,}96$. Ha az 1,6-ok helyére $\sqrt{3}$ at írunk a mátrixba, azaz a tér "éppen" euklidészi, akkor $Q(\Delta)=0$, ha pedig még nagyobb értéket, akkor $Q(\Delta)>0$.

Mindenképpen fel kell hívni a figyelmet egy, eddig jó néhányszor elkövetett "pongyolaságra". Már az előző fejezetben emlegettük az "adattér" különböző formáit, amelyben a változók v. az objektumok egyaránt tengelyek lehetnek. Ezt az adatteret jócskán illusztráltuk is különféle szórásdiagramokkal (2.1, 2.3-5 és 2.9 ábrák). Kimondatlanul is az euklidészi távolságot tekintettük érvényesnek a pontok közötti távolságok kifejezésére. Nem elegendő azonban csak pontokról és tengelyekről beszélni, hiszen a térfogalom szerves része a pontok közötti távolság definíciója is. Ennek megfelelően egy tér akkor euklidészi, ha a pontok között euklidészi távolságokat értelmezünk. A tér akkor metrikus, ha a távolságokra érvényesek a metrikus axiómák, míg egyéb esetekben a tér nem-metrikus.

Miután a tér fogalmát tisztáztuk, még mindig nem eléggé világos: miért fontos már az elején beszélni arról, hogy mikor tekinthetünk egy teret euklidészinek? Miért előnyös az euk-

lidészi tér a többiekkel szemben? Csak néhány alapvető "mentséget" sorolunk fel az euklidészi tér használata mellett:

- A pontok elrendeződését egy nem-euklidészi térben nemigen tudjuk elképzelni (pláne sok dimenzióban). Az adatmátrixot kiindulásképpen mindig egy euklidészi térben ábrázolható pontsereg koordinátáiként fogjuk fel. Eredményeinket, s itt elsősorban az ordinációs módszerek szórásdiagramjaira gondolunk (7. fejezet), is euklidészi térben ábrázoljuk (általában a papír síkjában). E mentális és gyakorlati kötöttségek miatt is érdemes ragaszkodni az euklidészi feltételekhez.
- A többváltozós módszerek jelentős része feltételezi, hogy a pontok euklidészi, de legalább metrikus térben helyezkednek el. Az osztályozó módszerek közül például az eltérésnégyzet-összeggel és a varianciával számoló eljárások, vagy a centroid módszer említhető meg. A legtöbb ordinációs eljáráshoz is teljesülniük kell a metrikus feltételeknek (kivétel pl. a nem-metrikus többdimenziós skálázás, mint a neve is mutatja). Emiatt tisztában kell lennünk azzal, hogy egy adott távolság vagy hasonlósági függvény milyen többváltozós értékelő módszerben alkalmazható egyáltalán, s ha igen, milyen formában.

Az euklidészi tér nyilvánvaló előnyei ellenére persze megpróbálhatunk egy nem-metrikus térben is dolgozni. A biológus számára elég sok olyan "értelmes" függvény áll rendelkezésre, amit nem-metrikus jellege ellenére is alkalmazni szeretne. Ekkor azonban vigyáznia kell, hogy milyen módszert választ a későbbiek során. Ezt a választást majd táblázatok segítségével igyekszünk megkönnyíteni (pl. 3.2 táblázat).

A függvények metrikus tulajdonságai mellett természetesen más szempontokat is figyelembe kell vennünk, mielőtt eldöntjük, hogy melyiket alkalmazzuk. Megvizsgálható még, hogy adataink szisztematikus megváltoztatásakor miként változnak a függvényértékek, hogy kiszűrhessük a kevéssé megfelelőeket. Lamont & Grant (1979), Wolda (1981) és Hajdú (1981) szolgáltatja a legfigyelemreméltóbb példákat egy ilyen típusú összehasonlításra. Kötetünkben azonban nincs hely minden részletre kiterjedő értékelésre, csak néhány alapesetet mutathatunk be.

### *3.1.2 Különbözőség*

A metrika és az euklidészi távolság definícióját követően most már itt az ideje, hogy meghatározzuk a *különbözőség* (*"dissimilarity"*) fogalmát is. Minden olyan $d$ függvényt különbözőségnek nevezünk, amelyre az 1-3 metrikus axiómák teljesülnek, a 4. viszont nem feltétlenül. A különbözőség tehát általánosabb, mint a metrika és az euklidészi távolság; ezeket speciális esetként tartalmazza. Különbözőség például az euklidészi távolság négyzete, amelyről a 3.1 mátrix alapján is könnyen belátható, hogy nem metrika.

A különbözőségi függvények jelentős része nemcsak alulról, hanem felülről is korlátos, és sok esetben a különbözőség elnevezést kizárólag ezekre alkalmazzák. A felső határ rendszerint 1 (maximális különbözőség), az alsó határ pedig 0 (azaz $0 \leq d_{jk} \leq 1$). Ilyen típusú különbözőségi indexekre bőven találunk példát a 3.5 részben. A különbözBUBUségi indexek egy részére megmutatható, hogy a $\sqrt{d_{jk}}$ formában teljesítik csak a metrikus axiómákat, ekkor nevezhetők igazán *távolság*nak.

*3.1.3 Hasonlóság*

A biológus általában nem annyira távolságokban, mint *hasonlóságokban* (*"similarity"*) gondolkodik. A sokdimenziós térbeli pontelrendezést ritkán képzeli el, és távolságok helyett az objektumok intuitíve is felfogható hasonlóságát szeretné valamilyen kvantitatív formában kifejezni. Erre a célra számos hasonlósági függvény közül választhatunk. A teljesen megegyező objektumok adják a maximális hasonlóságot (rendszerint $s_{jj}$=1), míg a lehető legnagyobb mértékben különbözők a minimálisat ($s_{jk}$=0). A hasonlóság tehát komplementer a [0,1] intervallumban mért különbözőséggel; a kettő egymásból kifejezhető:

$$s_{jk} = 1 - d_{jk} \tag{3.4}$$

A hasonlóságok nyilvánvalóan nem teljesítik a metrikus axiómákat. Számos, a [0,1] intervallumban értelmezett hasonlósági függvényre megmutatható azonban, hogy az alábbi átalakítás után:

$$d_{jk} = \sqrt{1 - s_{jk}} \tag{3.5}$$

már metrikusak, és többnyire euklidésziek is (vö. Gower & Legendre 1986).

Az **S** hasonlósági mátrixból a 3.5 formulával történő átalakítással biztosan euklidészi távolságot kapunk, ha $0 \leq s_{jk} \leq 1$, és az **S** mátrix pozitív szemidefinit (C függelék).

*3.1.4 Korreláció, asszociáltság*

A különbözőségeknek, távolságoknak és – a komplementaritás miatt – a hasonlóságoknak is közvetlen geometriai értelmük van: a sokdimenziós tér pontjainak relatív helyzetét fejezik ki. A függvények egy másik csoportja viszont, a pontok konfigurációját figyelembe véve, a tengelyek közötti kapcsolatokat tárja fel. Ide tartoznak a különféle korrelációs és asszociáltsági koefficiensek. Amennyiben a pontok egy véletlen mintából származó mintavételi egységeket képviselnek, a korreláció vagy az asszociáltság erőssége a hagyományos statisztikai tesztekkel is megvizsgálható. Formailag kiszámíthatók akkor is, ha mintavételi egységek a tengelyek, de ennek nehézségeire már a 2.1 részben rámutattunk. Csak emlékeztetőül: az attribútum dualitás elve csak óvatosan érvényesíthető az ilyen függvények esetében, különösen ha – a jelen kötetben egyébként nem tárgyalt – statisztikai próbákat is alkalmazni szeretnénk.

A korreláció és asszociáltsági együtthatók rendszerint a [–1,1] intervallumban mérik a kapcsolat erősségét (kivéve pl. kovariancia). A szélső értékek maximális erősségű, de ellentétes irányú kapcsolatra utalnak. A 3.5 összefüggés segítségével ezek is sok esetben euklidészi távolsággá alakíthatók.

## 3.2 Együtthatók bináris adatokra

A biológiában igen gyakoriak a bináris (prezencia/abszencia) típusú adatok, nemritkán a mintát leíró összes változó ilyen. Ennek megfelelelően általánosan elterjedtek és közismertek a bináris adatokra kidolgozott hasonlósági koefficiensek is. Matematikai tulajdonságaikat tekintve rendkívül sokfélék lehetnek, s kizárólag a közös adattípus miatt kerülnek egy fejezetbe. A függvényeket a legismertebb formájukban adjuk meg, még akkor is, ha csak távol-

sággá alakítva jöhetnek számításba az adatfeldolgozásban. Előrebocsátjuk, hogy az $1-s_{jk}$ átalakítással egyik említett hasonlóság sem tehető euklidészivé, míg a $\sqrt{1-s_{jk}}$ átalakítással már egy jelentős részük euklidészivé válik (3.2 táblázat). A többváltozós elemzésben tehát elsősorban az utóbbiakat javasoljuk.

Több, e részben ismertetett hasonlósági függvény csupán speciális, prezencia/abszencia adatokra leegyszerűsített formája a 3.5 részben bemutatandó függvényeknek. Látszólag felesleges ismétlésekbe bocsátkozunk tehát. A párhuzamosság azonban sokak számára nem mindig nyilvánvaló, így célszerű, ha a függvények mindkét változatát megadjuk. A jelölésnél nem az *s* ("*similarity*") rövidítést fogjuk alkalmazni különféle indexeléssel, hanem a függvények elnevezésére utaló betűszavakat használunk (pl. *SM, Y*1, stb.).

Prezencia/abszencia adatokra az alábbi, ún. négymezős (2×2-es) kontingenciatábla jelöléseivel nagymértékben leegyszerűsödik a képletek felírása:

| | | 2. objektum: 1 | 2. objektum: 0 | |
|---|---|---|---|---|
| 1. objektum | 1 | $a$ | $b$ | $a+b$ |
| | 0 | $c$ | $d$ | $c+d$ |
| | | $a+c$ | $b+d$ | $n$ |

ahol

*a:* az olyan változók száma, amelyek mindkét összehasonlítandó objektumban megvannak (közös prezencia);

*b*: azon változók száma, amelyek csak az 1. objektumot jellemzik, a másikból hiányzanak;

*c:* a csak a 2. objektumot jellemző, az 1-ből hiányzó változók száma; és

*d:* azoknak a változóknak a száma, amelyek mindkét szóbanforgó objektumból hiányoznak ugyan, de legalább egy objektumot jellemeznek a mintában (közös abszencia).

Az *a, b, c* és *d* értékek alsó indexeit (pl. $a_{12}$) az egyszerűsítés kedvéért elhagytuk. Nyilván $a+b+c+d=n$, azaz a mintában szereplő változók száma. A táblázat peremösszegei az egyes objektumokat jellemző ill. nem jellemző változók számának felelnek meg.

A hasonlósági együttható kiválasztásában a legkritikusabb mozzanat a *d* érték figyelembe vétele vagy mellőzése. *d*, mint említettük, a mindkét összehasonlítandó objektumból hiányzó változók száma. Rögtön felvetődik a kérdés: vajon a duplán hiányzó változók növeljék-e a hasonlóságot, s ha igen, mely esetekben? Bár ezt a problémát a bináris változókról szóló 1.4.2 részben egyszer már érintettük, nem árt most visszatérni rá. Amennyiben valóban prezencia/abszencia adatokról van szó, azaz 1 minőségileg többet jelent a 0-nál (fajok jelenléte szemben az abszenciával, bizonyos morfológiai tulajdonságok megléte azok hiányával szemben, stb), akkor a *d* értéke figyelmen kívül hagyható. Mondhatjuk ugyanis, hogy hasonlóságot csak azon változók alapján értelmezhetünk, amelyek legalább az egyik objektumot jellemzik, s annak nincs szerepe, hogy még milyen változók szerepelnek a mintában. Ha dichotomizált nominális változóink is vannak (lásd 1.4.1 rész), akkor pedig *d* értéke bizonyosan mellőzendő,

hiszen ezzel csak a dichotomizált változók erőteljesebb súlyozását érnénk el. Ez ellentétes azzal az általános felfogással, hogy *a priori* az összes változó egyformán fontosnak tekintendő.

Milyen esetekben dönthetünk mégis úgy, hogy a *d*-t, mint hasonlóságot növelő tényezőt is figyelembe vesszük? Klasszikus példa a mikroorganizmusok hasonlósága azon az alapon, hogy az egyes törzsek mely szubsztrátumokat képesek bontani ill. nem bontani. Mindkét típusú reakciót egyformán fontosnak tekinthetjük, és kimondhatjuk: két törzs hasonlóságát az is növelje, ha egy adott szubsztrátumot egyikük sem bont. Azaz *a* és *d* értéke egyformán fontos információt hordoz. Egy cönológiai vizsgálatban, kvadrátok flórájának összevetésében is lehet értelme a *d*-nek, hiszen a fajok hiánya értelmes információ: az adott niche-t valamilyen más, kompetitív faj foglalta el. Amíg azonban a fajok prezenciája bizonyosan azt jelenti, hogy azok életképesek az adott területen, az abszencia nem feltétlenül jelenti ennek az ellenkezőjét. Egy faj éppen véletlenszerűen is hiányozhat adott területről (Green 1971). Persze ennek az érvelésnek a fordítottja is igaz lehet, mint Goodall (1973a) megjegyzi, hiszen a rendkívül gyakori, ubikvista fajok együttes előfordulása is lehet véletlenszerű hatások eredménye. Látjuk tehát: a kérdés meglehetősen komplikált ahhoz, hogy most egy általánosan érvényes receptet adhassunk. Mindenesetre kimondható: ha nagyon sok ritka faj van a mintában, amelyek együttes előfordulása valóban egy véletlenszerű eseménynek tekinthető, akkor nem indokolt a *d* figyelembe vétele, mert az túlságosan megnövelné a hasonlóságokat. Egy viszonylag "kiegyenlítettebb", a fajok gyakoriságában kisebb ingadozásokat mutató mintában viszont értelmes lehet a *d*. Azok számára pedig, akik pedig végképp nem tudnak dönteni, jó szívvel ajánlható a *d*-t egyfajta köztes módon figyelembe vevő, "kompromisszumképes" 3.19 és 3.20 koefficiens.

Ha a bináris adatok csupán látszólag prezencia/abszencia típusúak, de valójában kétállapotú nominális adattípusnak felelnek meg, az 1-gyel és 0-val való kódolás önkényes (1 nem jelent minőségileg többet, mint a 0). Nyilván ekkor *d* értéke teljesen egyenrangú az *a*-val, és csak olyan koefficienseknek van értelme, amelyek *a*-t és *d*-t szimmetrikusan kezelik. A részletes tárgyalást ezekkel a hasonlósági függvényekkel kezdjük a 3.2.1 részben.

A prezencia/abszencia koefficiensek közötti választás megkönnyítésére egy grafikus módszert is alkalmazunk. A kiindulás a 3.1 táblázat adatmátrixa lesz (következő oldal), amelyben 9 objektumot 18 változó jellemez. Az objektumok az 1→9 irányban fokozatosan, egyenlő lépésekben alakulnak át egy-egy változó kiesésével ill. belépésével. A 17-18. változó szándékosan csupa 0 értékű, hogy *d* ne legyen 0 a közös prezenciát már nem felmutató 1/9 párosításban sem. Így láthatjuk, hogy a hasonlóságok elérik-e ilyenkor a 0-t. Az 1. objektum összehasonlítása önmagával és a többi nyolccal minden egyes függvényre kilenc értéket ad, amelyek vonaldiagramos ábrázolása megmutatja, hogy "szabályosan" reagálnak-e a függvények az adatok szisztematikus megváltoztatására.

### *3.2.1 Az a és d értékekre nézve szimmetrikus hasonlósági együtthatók*

A legegyszerűbb függvények egy hányados ("index") segítségével fejezik ki az objektumok hasonlóságát, tehát bizonyos értelemben százalékos jelentésük van, sok esetben pedig valószínűségi interpretációjuk is.

**3.1 táblázat.** Mesterséges adatok mátrixa a prezencia/abszencia koefficiensek értékeléséhez. Az objektumok egy "grádiens" mentén egyenletesen távolodnak a kiinduló 1. objektumtól.

| | Objektumok | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Változók | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 9 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 11 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 13 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Elsőként az *egyezési koefficienst* ("*simple matching coefficient*", Sokal & Michener 1958) mutatjuk be:

$$SM = \frac{a+d}{a+b+c+d} = \frac{a+d}{n} \tag{3.6}$$

amely az egyezések száma osztva a változók számával. Teljes egyezés esetén $SM$=1, teljes különbözőség esetén $SM$=0. $SM$ voltaképpen annak a valószínűsége, hogy egy véletlenszerűen kiválasztott változóra nézve a két objektum megegyező. Ugyanakkor $SM$ rokonságban van a bináris adatokra felírható *euklidészi távolsággal* is:

$$ED = \sqrt{b+c} \tag{3.7}$$

mivel

$$ED^2 = n\,(1 - SM) \tag{3.8}$$

Az euklidészi távolság értéktartománya [0, $\sqrt{n}$ ]. A 3.8 kapcsolat miatt lényegében véve mindegy, hogy melyiket választjuk. Miután a 3.6 függvény értéktartománya nem függ $n$-től, ennek használata ajánlható elsősorban, hiszen különböző $n$-ekre kapott elemzések is összevethetők egymással.

A 3.8 összefüggés miatt nem is kell hangsúlyozni, hogy $\sqrt{1-SM}$ euklidészi (3.2 táblázat). $SM$ másik előnyös tulajdonsága, hogy lineárisan követi az objektumok fokozatos megváltoztatását, és viszonylag egyenletesen változik távolsággá alakítva is (3.2ab ábra).

Az egyezési koefficiens (3.6) egy változatának tekinthető a *Rogers - Tanimoto* (1960) index:

**3.2 táblázat**. Prezencia-abszencia koefficiensek metrikus ill. euklidészi tulajdonságai. Jelölések: N=nem-metrikus, M=metrikus, E=euklidészi. Minden hasonlósági függvényt a 3.5 egyenlet szerint transzformálni kell, mielőtt e tulajdonságokat vizsgáljuk. Kivétel a Mountford és az Ochiai index, amelyeket a 3.32 exponenciális függvény, ill. a húrtávolság helyettesít.

| Függvény neve | Tulajdonság | Függvény neve | Tulajdonság |
|---|---|---|---|
| *a és d-re szimmetrikus függvények* | | *a és d-re nem szimmetrikus függvények* | |
| Egyezési koefficiens, *SM* (3.6) | E | Baroni-Urbani - Buser I, *BB*1 (3.20) | E |
| euklidészi távolság, *ED* (3.7) | E | Baroni-Urbani - Buser II, *BB*2 (3.19) | E |
| Rogers - Tanimoto, *RT* (3.9) | E | Russell - Rao, *RR* (3.23) | N |
| Sokal - Sneath I, *SS*1 (3.11) | M | Faith I, *FA*1 (3.21) | N |
| Anderberg I, *A*1 (3.12) | E | Faith II, *FA*2 (3.22) | N |
| Anderberg II, *A*2 (3.13) | N | *d-t ignoráló függvények* | |
| korreláció, *PHI* (3.14) | E | Jaccard, *JAC* (3.24) | E |
| Yule I, *Y*1 (3.16) | N | Sorensen, *SOR* (3.25) | N |
| Yule II, *Y*2 (3.17) | N | Húrtávolság, *CH* (3.28) | E |
| Hamann, *HAM* (3.18) | E | Kulczynski, *KUL* (3.29) | N |
| | | Sokal - Sneath | N |
| | | Mountford, *MFD* (3.32) | M? |

$$RT = \frac{a+d}{a+2b+2c+d} \tag{3.9}$$

amely tehát kétszeresen veszi figyelembe a különbözőséget okozó változókat, így értéke *SM*-nél mindig alacsonyabb (kivéve természetesen a $b+c=0$ esetet). Anderberg (1973) értelmezése szerint a nevező az $n$ változóra kapott összes megvalósult karakterállapot száma, a számláló pedig azon állapotok száma, amelyben az összehasonlított objektumok meg is egyeznek.

Gower & Legendre (1986) megmutatta, hogy az

$$s = \frac{a+d}{a+d+\theta(b+c)} \tag{3.10}$$

általános alakban felírható függvénycsaládra $\sqrt{1-s}$ mindenképpen euklidészi, ha $\theta \geq 1$. Ugyanakkor, ha az egyezéseket súlyozzuk kétszeresen, mint az alábbi, Sokal & Sneath-nek (1963) tulajdonított együtthatóban

$$SS1 = \frac{2a+2d}{2a+b+c+2d} \tag{3.11}$$

akkor annak távolságmegfelelője egy *nem-euklidészi metrika*.

Első látásra valószínűségi alapon értelmezhetjük az alábbi két hasonlósági függvényt (Anderberg 1973). Az első formulában:

$$A1 = \left( \frac{a}{a+b} \frac{a}{a+c} \frac{d}{b+d} \frac{d}{c+d} \right)^{1/2} \tag{3.12}$$

az egyes tagok feltételes valószínűségként foghatók fel. Pl. *a/(a+b)* annak a valószínűsége,

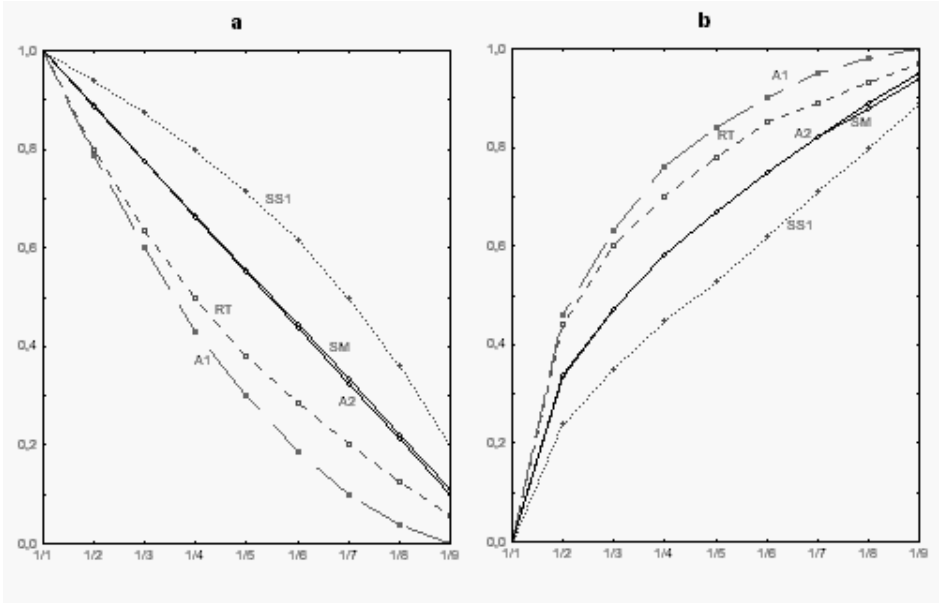
hogy egy véletlenszerűen kiválasztott változó a 2. objektumra 1-es értéket vesz fel feltéve, hogy az 1 objektumra is 1-es az értéke. A 3.12 függvény tehát négy feltételes valószínűség geometriai közepének a négyzete (a mértani középhez a szorzatból negyedik gyököt kellene vonnunk).

A függvény jelentése talán jobban megérthető a következők szerint. Mint később látni fogjuk, a 3.26 hasonlósági függvény – amely a 3.12 összefüggés első két tagját tartalmazza – az 1. ill. 2. objektumokhoz mutató vektorok szögének a cosinusa. Egybeeséskor, $0^{o}$-nál értéke 1 (teljes hasonlóság), a legnagyobb elérhető szögnél, $90^{o}$-nál pedig 0 az értéke (teljes különbözőség). A 3.26 összefüggés persze nem szimmetrikus *a*-ra és *d*-re nézve, így a kódolás felcserélésével egészen különböző eredményekre vezethet. Nos, a 3.12 függvény éppen a 3.26 függvénnyel és kétféle kódolással kiszámított két cosinus érték geometriai közepének négyzetgyöke lesz. *A*1 olyan esetekben használható tehát, amikor nem tudjuk eldönteni, hogy milyen kódolást alkalmazzunk.

*A*1 lehetséges értékei a [0,1] intervallumba esnek. Teljes hasonlóság esetén $b=c=0$, azaz az összes tag értéke 1 lesz, így a végeredmény is 1. Teljes különbözőség mellett $a=d=0$, így az összefüggés értéke is 0.

Sokal & Sneath (1963:130) és Anderberg (1973) javasolt egy rokon formulát is, amelyben a négy feltételes valószínűségnek az aritmetikai közepét számítjuk ki az alábbiak szerint:

$$A2 = \frac{1}{4}\left(\frac{a}{a+b}+\frac{b}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right) \tag{3.13}$$




**3.2 ábra.** Az *a* és *d* értékét szimmetrikusan tekintő, [0,1] intervallumban működő hasonlósági függvények változása egy összehasonlítási sorban (a 3.1 táblázat 1. objektumát összevetve mindegyikkel). **a**: eredeti függvény, **b**: a 3.5 összefüggés alapján távolsággá alakított függvény.

Ez a függvény a Kulczynski-indexszel (3.29) számított két, a kódolásban eltérő hasonlóságértéknek az átlaga, tehát a 3.12 formulához hasonlóan ugyancsak a kódolási problémák "kivédésére" alkalmas. Ezek komplementje azonban, akárcsak a Kulczynski indexé, nem euklidészi.

A *szorzat-momentum korrelációs koefficiens* (3.70) bináris esetben kifejezhető a 2×2-es kontingenciatáblázat jelöléseivel is:

$$PH = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{3.14}$$

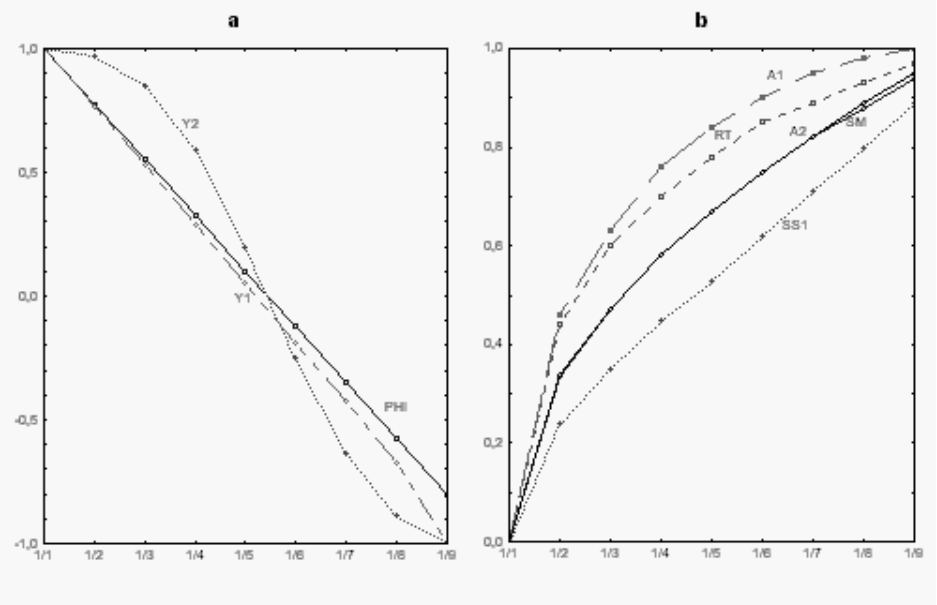
Tulajdonságait a későbbiekben, a korreláció (3.70) tárgyalásakor ismertetjük. Ehelyütt csak annyit érdemes megjegyeznünk, hogy ha a számlálóból elvesszük a *bc* tagot, akkor a 3.12 egyenletet kapjuk. A *PHI* koefficiens és a változók függetlenségét kifejező khi-négyzet statisztika között szoros összefüggés van:

$$PHI^2 = \chi^2/n. \tag{3.15}$$

Ugyancsak változók kapcsolatának mérésére alkalmas elsősorban a *Yule* féle predikta-bilitási index is

$$\gamma 1 = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \tag{3.16}$$

amely azt méri, hogy mennyiben "jósolható meg" az egyik változó egy adott megvalósulása a másik ismeretében. *Y*1= 1 ill. *Y*1=—1 értékekre lesz teljes a megjósolhatóság. Az első esetben




**3.3 ábra.** Az *a* és *d* értékét szimmetrikusan tekintő, [–1,1] intervallumban működő hasonlósági függvények változása egy összehasonlítási sorban (a 3.1 táblázat 1. objektumát összevetve mindegyikkel). **a:** eredeti függvény, **b**: a 3.5 alapján távolsággá alakított függvény.

*bc*=0 tehát a két változó minden objektumra megegyező. A második esetben *ad*=0, tehát ha az egyik változó 1-et vesz fel, akkor a másik 0-t és fordítva, minden objektumban. *Y*1 nincs definiálva arra az esetre, amikor a 2×2-es kontingenciatábla bármelyik peremösszege 0 (azaz az egyik változó konstans értékű). Ugyanez elmondható a *PHI* korrelációról is. A 3.16 függvényből leszármaztatható Yule másik együtthatója:

$$Y2 = (ad - bc) / (ad + bc) \qquad (3.17)$$

Egyik Yule függvény sem transzformálható euklidészi távolsággá, és főképpen az *Y*2 változása tűnik elfogadhatatlannak, mivel nem lineáris (3.3 ábra).

Mivel ugyancsak a [–1,1] intervallumban fejezi ki a hasonlóságot, itt említjük meg a *Hamann* indexet is:

$$HAM = (a + d - b - c) / (a + d + b + c) \qquad (3.18)$$

A függvény azonban nem mond semmi újat az egyezési együtthatóval (3.6) szemben, hiszen csak annak értéktartományát szélesíti ki a [–1,1] intervallumba. Érvényes ui. az *SM* = (*HAM* + 1) / 2 összefüggés.

A 3.2-3 ábrák összesítő értékeléséből kiderül, hogy hasonlósági függvény formájában az *SM*, a *PHI* és – megközelítőleg – az *A*2 változik lineárisan a 3.1 táblázatbeli 1/1 □ 1/9 összehasonlítási sorban. Távolsággá alakítva ezt a tulajdonságukat elveszítik, bár az első lépést kivéve változásuk közel lineáris marad. Ezek közül *SM* és *PHI* euklidészi, így kétségkívül ők tűnnek a legelőnyösebbeknek. *A*1 és *RT*, ill. *SS*1 már hasonlóság formában sem lineáris, s ez a sajátság a távolsággá alakítást követően az első kettő esetében még tovább fokozódik, míg *SS*1 kerül legközelebb a linearitáshoz (kár, hogy *SS*1 nem euklidészi). Rendkívül sajátságos az *Y*2 lefutása a középtájt mutatkozó inflexiós ponttal. Az *Y1, Y2* és *A1* függvények elérik a 0 hasonlóságot, ehhez ui. elég *a* vagy *d* értékének 0-ra csökkennie, s ez kétségkívül előnytelen lehet.

A prezenciát és abszenciát ugyancsak szimmetrikusan kezelik a különféle információelméleti függvények is, amelyeket majd a 3.7 részben, a kettőnél több objektumra alkalmazható heterogenitási függvények között ismertetünk.

### *3.2.2 Az a és d értékekre nézve nem-szimmetrikus hasonlósági koefficiensek*

Az alábbi két index – mintegy kompromisszumként – átmenetet képez az előző rész függvényei és a *d* értékét teljesen mellőző hasonlóságok között. Baroni-Urbani & Buser (1976) szerint *d*-t nem lenne szabad teljesen figyelmen kívül hagyni, ugyanakkor eredeti formájában a *d* érték túlhangsúlyozza a közös abszenciát. A megoldást az jelenti, ha *d* helyett az *a* és *d* geometriai közepével számolunk. Ekkor az egyezési koefficiens Baroni-Urbani - Buser-féle módosítása a következő lesz:

$$BB2 = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c} \qquad (3.19)$$

a Hamann indexé pedig

$$BB1 = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c} \qquad (3.20)$$

A két formula csupán értéktartományában tér el egymástól, hasonlóan a kiindulásként használt

*SM* és *HAM* indexekhez: *BB*2=(*BB*1+1)/2. A kettő közül a [0,1] értéktartományú *BB*2 használata a kényelmesebb. Bár a szerzők részletes eloszlásvizsgálatot mellékeltek indexeik előnyeinek érzékeltetésére, a *BB* formulákat eddig még viszonylag ritkán használták. Figyelemre méltó viszont, hogy Kenkel & Booth (1987) egyértelműen a *BB*1-et találták a legmegfelelőbbnek egy biogeográfiai összehasonlító vizsgálatban.

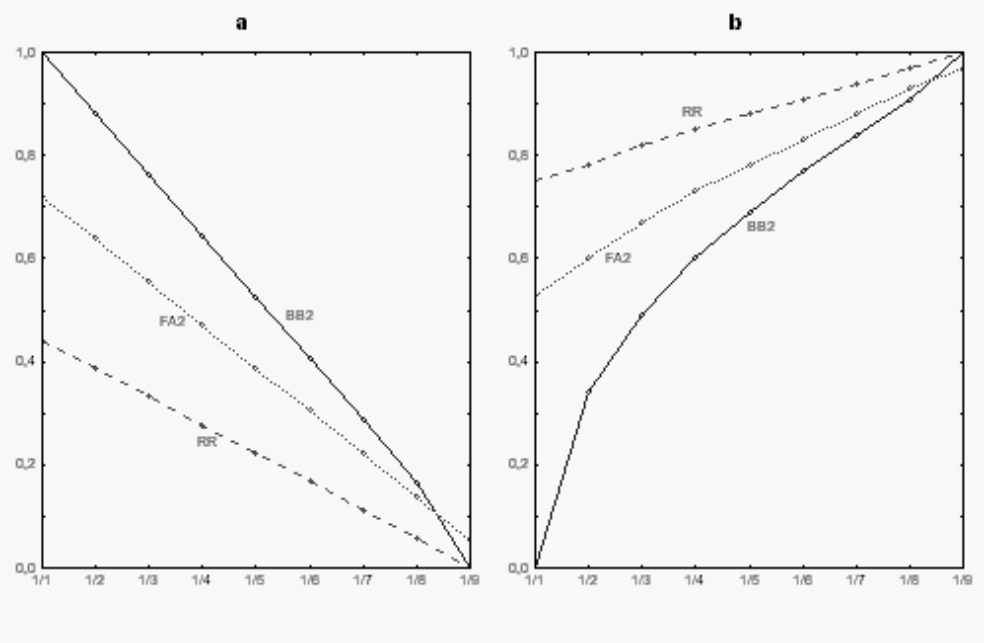
Faith (1983) megmutatta, hogy a *BB*2 hasonlóság csekély mértékben növekedhet is, ha *d* értéke nő *a* rovására (pl. ha *a*=10, *d*=1 és *b*+*c*=5, akkor *BB*2=0,247, míg *a*=9, *d*=2 és *b*=*c*=5 mellett *BB*2=0,259). Azaz, bár szándékunk szerint a dupla 0-k kisebb súllyal részesednek, egy dupla 1-es felváltása dupla 0-val nemkívánatos változást eredményezett. Ennek kiküszöbölésére Faith a következő hasonlósági indexet javasolta:

$$FA1 = (a - b - c) / (a + b + c + d) \qquad (3.21)$$

amelyben *a* értéke növeli, *b* és *c* értéke pedig csökkenti a hasonlóságot, *d*-nek pedig csupán a nevezőben jut hely. Ha tehát *d* nő *a* rovására, a hasonlóság mindenképpen csökken. A 3.21 függvény a [–1,1] intervallumban méri a hasonlóságot, s ezért kényelmesebb lehet az alábbi módosított képlet, mely a *FA*2=(*FA*1+1)/2 összefüggés jobboldalának átalakításával kapható meg:

$$FA2 = (a + d/2) / (a + b + c + d) \qquad (3.22)$$

ahol *d* jelenléte a számlálóban kissé félrevezető lehet az első látásra. A függvény tulajdonképpen negatívan veszi *b*-t és *c*-t figyelembe, hiszen azok nem szerepelnek a számlálóban. *a* súlyozása egyszeres, *d* pedig köztes súlyozású.




**3.4 ábra.** Az *a* és *d* értékekre nem szimmetrikus hasonlósági függvények változása egy összehasonlítási sorban (vö. 3.2 táblázat). **a**: eredeti függvény, **b:** távolsággá alakítva 3.5 szerint.

A *Russell & Rao index* is figyelembe veszi *d* értékét a nevezőben:

$$RR = a/(a+b+c+d) \quad (3.23)$$

így *d* értéke nem közömbös a hasonlóság kiszámításában, sőt: növekedése csökkenti két objektum hasonlóságát. A formula valójában egy egyszerű relatív gyakoriság: annak az eseménynek a becsült valószínűsége, hogy egy véletlenszerűen kiválasztott tulajdonság mindkét objektumban megvan. *d* viszonylag magas értéke túlzott és nemkívánatos befolyással lehet *RR*-re.

Az *FA*2 és *RR* koefficiensek kedvezőtlen tulajdonsága, hogy bár elméletileg a [0,1] intervallumban fejezik ki a hasonlóságot, az objektumok önmagukkal vett hasonlósága rendszerint *nem* 1 (3.4 ábra). Ennek fontos következménye, hogy komplementjeik semmiképp sem metrikusak (ellentétben Gower & Legendre 1986 2. táblázatával), ha a többi feltételt be is tartják.

A három függvény grafikus értékelése a 3.4 ábrán látható (*BB*1 és *FA*1, a *BB*2-vel ill. *FA*1-el fennálló összefüggés miatt, nem szerepel a rajzon). *BB*2 csaknem lineáris, a másik kettő teljes mértékben lineáris (3.4a ábra) a vizsgált objektumsorozatra. A teljes [0,1] intervallumot csupán a *BB*2 használja ki, *FA*2 viszont sem a felső, sem az alsó határt nem éri el (azaz $a$=0 esetén sem 0). Távolsággá alakítva *RR* még mindig közelítően lineáris, de egy nagyon szűk intervallumba beszorítva.

### *3.2.3 A d értéket figyelmen kívül hagyó együtthatók*

A további formulákban *d* már egyáltalán nem szerepel, így a dupla nullák (közös abszenciák) száma természetesen semmiféle hatással sincs az eredményre. Elsősorban az ökológusok körében népszerűek. A legismertebb és legegyszerűbb a *Jaccard index*

$$JAC = a / (a+b+c) \quad (3.24)$$

amely annak az eseménynek a becsült valószínűsége, hogy két objektum megegyezik egy, legalább az egyiküket jellemző változóban. Ez tehát egy feltételes valószínűség, így a lehetséges értékek a [0,1] intervallumba esnek. A 3.5 átalakítással a Jaccard index euklidészi távolsággá alakítható (3.2 táblázat), széleskörű alkalmazásának tehát geometriai korlátai nincsenek.

A *Sorensen (Dice) index* annyiban különbözik az előzőtől, hogy *a* értékét duplán veszi figyelembe mind a számlálóban, mind a nevezőben:

$$SOR = 2a / (2a+b+c) \quad (3.25)$$

A dupla súlyozás a prezenciák "közös részére" utal, míg a $b+c$ összeg a különbözőséget okozza (hasonlítsuk össze a 3.59 formulával). A súlyozás következménye, hogy *SOR* nem konvertálható euklidészi távolsággá.

Az *Ochiai koefficiens* (más források szerint *Otsuka* volt a javaslattevő) a következő:

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (3.26)$$

amelynek geometriai értelmezése a nyilvánvalóbb: *OCH* a két pontra mutató vektorok hajlásszögének a cosinusa (emlékeztetőül: a kódolás felcserélésével kapott másik cosinus értékkel vett geometriai közép volt az *A*1 formula (3.12)). Teljes egyezés esetén értéke 1, maximális

különbözőségre pedig *OCH*=0. A 3.26 függvény a 3.55 egyenlet prezencia/abszencia esetre egyszerűsített alakja. Fager & McGowan (1963) javasolta egy korrekciós tényező alkalmazását is:

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\max\{(a+b),(a+c)\}}} \tag{3.27}$$

amely azonban nem befolyásolja lényegesen az eredményt, s csak az önhasonlóságot viszi 1 alá, így az 1. axióma nem teljesülhet.

A *húrtávolság* közvetlen kapcsolatban áll a 3.26 formulával:

$$CH = \left[ 2\left( 1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right) \right]^{1/2} \tag{3.28}$$

amely tehát az egységsugarú hipergömbre vetíti a két pontot (2.22 standardizálás) és ezután méri a közöttük lévő euklidészi távolságot (összehasonlítandó a 3.54 formulával).

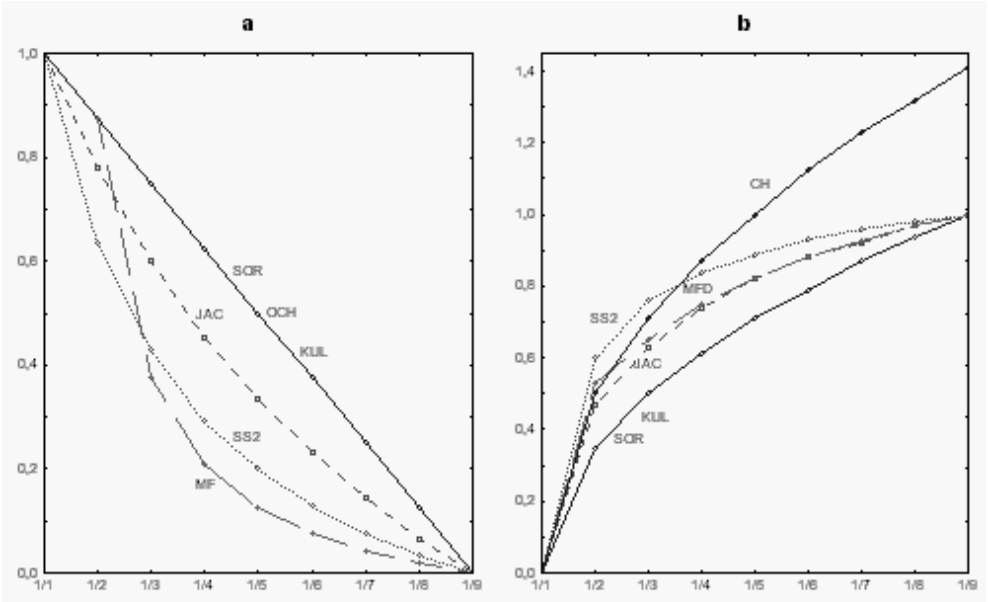
Az *a*-ra vonatkozó kétnűség aritmetikai középértéke a Kulczynski index:

$$KU = \frac{1}{2}\left( \frac{a}{a+b} + \frac{b}{a+c} \right) \tag{3.29}$$

A *Sokal - Sneath* (1963) -féle második hasonlósági együttható pedig a következő:

$$SS2 = a / (a+2b+2c) \tag{3.30}$$

A fenti két formula egyike sem ajánlható jó szívvel, mivel nem euklidészieknek. Egyéb tulajdonságaikat tekintve lásd a 3.5 ábrát, ill. a következő oldalon található értékelést.




**3.5 ábra.** A *d* értékeket figyelmen kívül hagyó hasonlósági függvények változása egy összehasonlítási sorban (vö. 3.2 ábra). **a**: eredeti függvény, **b:** távolsággá alakítva 3.5 szerint, kivéve *MFD* és *CH*.

A fajok és egyedszámok logaritmikus eloszlásából kiindulva javasolt *Mountford* (1962) egy speciális hasonlósági formulát. A logaritmikus eloszlás egy paramétere az $\alpha$, amely diverzitási mértékszámként is számításba jöhet (vö. Pielou 1975:43-45). Két mintaterület fajösszetétel szerinti összehasonlítására Mountford szerint jól használható az $MF=1/\alpha$ függvény, amely relatíve független a mintanagyságtól (s így a ritka fajoktól). *MF* becslésére a négymezős kontingenciatábla adatai alapján az alábbi formula alkalmas:

$$MF = 2a / (ab + ac + 2bc) \qquad (3.31)$$

Ennek azonban súlyos hibája, hogy a két objektum teljes egyezésekor 0-val kellene osztanunk. Teljes különbözőségnél *MF* értéke 0. Orlóci (1978) szerint *MF* egy relatív távolsággá alakítható, amely – szimulációs tapasztalatok alapján – metrikus:

$$MFD = e^{-MF} \qquad (3.32)$$

azzal a megjegyzéssel, hogy $b=c=0$ esetre *MF* értékét kellően nagy pozitív számnak vesszük, hogy $MFD = 0$ legyen. A 3.31 hasonlóság Kenkel & Booth (1987) és Wolda (1981) értékelése szerint a fajösszetétel megváltozásával először hirtelen csökken, majd egyre kevéssé változik, ami kétségtelenül nem kívánatos a többváltozós elemzésben (l. még a 3.5b ábrát).

A 3.2.3 részben ismertetett hasonlósági függvények grafikus értékelése a 3.5 ábrán látható. *SOR, OCH és KUL*, a példaadatokra legalábbis, teljesen egybeesik és lineáris lefutású. A többi három függvény a *JAC*, *SS*2, *MF* sorrendben egyre jobban eltér a lineáristól. Távolsággá alakítva azonban a *CH* viselkedése tűnik a legideálisabbnak, s ugyanakkor ez euklidészi is. A diagramok tanúsága szerint a Mountford index használható a legkevésbé.

Nincs különösebb jelentősége a többváltozós elemzésben azoknak az együtthatóknak, amelyek nem teljesítik a szimmetria-axiómát sem. Csupán a teljesség kedvéért említjük meg tehát Kulczynski másik indexét ($s=a/(b+c)$), a Simpson indexet ($s=a/(a+b)$) és a Braun-Blanquet indexet ($s=a/(a+c)$). Ezen indexek csak speciális esetekben alkalmazhatók, amikor az összehasonlítás iránya kitüntetett (pl. aszimmetrikus mátrixokra kidolgozott többváltozós módszerekben, Gower 1977).

## 3.3 Koefficiensek nominális változókra

Ha az adatmátrixban lévő összes változó nominális és 2-nél több állapotú (="*multistate nominal*"), akkor az objektumok összehasonlítása legegyszerűbben az *a* és *d* értékét szimmetrikusan kezelő prezencia/abszencia koefficiensek több állapotra általánosított változataival történhet. Ha *u* jelöli azoknak a változóknak a számát, melyre mindkét összehasonlítandó objektum megegyezik, akkor az egyezési index a következő

$$SM = u/n \qquad (3.33)$$

A Rogers - Tanimoto index megfelelője:

$$RT = u / (2n-u) \qquad (3.34)$$

Míg a Sokal - Sneath I. koefficiens (3.11) az alábbi formulával írható fel:

$$SS1 = 2u / (n+u) \qquad (3.35)$$

A Gower-féle általánosított hasonlósági formula (3.103) az objektumok összehasonlítását a 3.33 szerint végzi el a nominális változókra.

A *PHI* együttható értelmezésének kiterjesztéséhez egy újabb kontingencia-táblázatot kell felírnunk. Mivel a *PHI* függvényt általában statisztikai értelemben vett változók között számoljuk, a táblát 2 változóra mutatjuk be:

| | | 2. változó | | | |
|---|---|---|---|---|---|
| | | 1 | *j* | *q* | |
| 1. változó | 1 | $f_{11}$ | | | $f_{1.}$ |
| | *i* | | $f_{ij}$ | | $f_{i.}$ |
| | *p* | | | $f_{pq}$ | $f_{p.}$ |
| | | $f_{.1}$ | $f_{.j}$ | $f_{.q}$ | |

A táblázatban $f_{ij}$ jelöli annak a gyakoriságát, hogy az 1. változó *i* állapota és a 2. változó *j* állapota együtt fordult elő a mintában. $f_{i.}$ és $f_{.j}$ a marginális gyakoriságok, míg $f_{..}=m$, azaz a mintanagyság. *p* az 1. változó, *q* pedig a 2. változó lehetséges értékeinek a száma. A két változó kapcsolata a jól ismert khi-négyzet statisztika alapján kifejezve a következő:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(f_{ij} - \frac{f_{i.}f_{.j}}{f_{..}}\right)^2}{\frac{f_{i.}f_{.j}}{f_{..}}} \tag{3.36}$$

$\chi^2$ értéke nyilvánvalóan nő, ha $f_{..}$ nő; megoldásként a 2×2-es táblára alkalmazott 3.15 átalakítás juthat először eszünkbe. $\chi^2/f_{..}$ maximális értéke azonban $\min\{(p-1),(q-1)\}$ (lásd pl. Anderberg 1973:76) ezért ezzel még le kell osztanunk, hogy egy általános esetre alkalmas függvényt kaphassunk:

$$CR = \left(\frac{\chi^2 / f_{..}}{\min[(p-1),(q-1)]}\right)^{0.5} \tag{3.37}$$

amely *Cramér-index* (Cramér 1946) néven ismeretes a szakirodalomban. Ennek értéke tehát a [0,1] intervallumban mozog *p* és *q* bármely értékére. *CR* alkalmazását viszont a többváltozós elemzésben sokan megkérdőjelezik: a standard intervallum ellenére ugyanis nem biztos, hogy – mondjuk – a 0,5-ös *CR* érték ugyanolyan erősségű kapcsolatra utal egy 5×5-ös táblázat valamint egy 3×6-os táblázat esetén. (Ilyen jellegű problémára majd konkrét példát is látunk a 9. fejezetben.) *CR* tehát csak akkor ajánlható hasonlósági mátrixok kiszámítására, ha *p* ugyanaz minden változóra, s ez nem mindig teljesül. Általános esetre a megoldást a Goodman & Kruskal (1954) féle prediktabilitási index jelenti, amely az egyik változó adott értékének ismeretében a másikra vonatkozó megjósolhatóságot méri. Tételezzük fel először a következőket: ki szeretnénk találni, hogy egy objektum a 2. változóra milyen értéket vesz fel úgy, hogy az 1. változóra vonatkozó értéket nem ismerjük. Nyilván a legjobb tipp a legnagyobb gyakoriságú érték lesz, azaz $\max_j [f_{.j}]$ keresendő, mert ez minimalizálja a rossz találat valószínűségét. Ha azonban azt már tudjuk, hogy az 1. változóra az objektum konkrét értéke a változó *i*-edik állapota, akkor csak a táblázat *i*-edik sorát kell néznünk, s ekkor $\max_j [f_{ij}]$-t kell kikeresnünk a rossz találat valószínűségének minimalizálásához. A találati hiba csökkenése tehát arányos a két érték különbségével. Azaz, ha az 1. változót figyelembe vesszük a 2. változó konkrét értékének megjóslására az átlagos hibacsökkenés (relatív prediktabilitás) a következő:

$$LAS = \frac{\sum_{i=1}^{p} \max_j[f_{ij}] - \max_j[f_{.j}]}{f_{..} - \max[f_{.j}]} \tag{3.38}$$

Ennek értéke 0 ha az 1. változó nem ad semmiféle információt a másikról (függetlenség), ill. *LAS*=1, ha az 1. változó értékének ismeretében a 2. változóra már csak egyetlen érték jöhet számításba. (Ez utóbbi esetben a kontingencia-táblázat minden sorában és oszlopában csak egy nem nulla értékű cella van.) *LAS* azonban egy nem-szimmetrikus mértékszám, az 1. változó ismerete a 2.-ra vonatkozó megjósolhatóságot nem ugyanolyan mértékben növeli, mint a 2. ismerete az 1.-re vonatkozóan. A szimmetriafeltételnek is eleget teszünk azonban, ha a két prediktabilitás-értéket átlagoljuk, azaz

$$\Lambda = \frac{\sum_{i=1}^{p} \max_j[f_{ij}] + \sum_{j=1}^{q} \max_i[f_{ij}] - \max_i[f_{i.}] - \max_j[f_{.j}]}{2f_{..} - \max_i[f_{i.}] - \max_j[f_{.j}]} \tag{3.39}$$

(Goodman - Kruskal lambda). A prezencia/abszencia esetben e képlet a már ismertetett *Y*1 indexre egyszerűsödik. Mivel *Y*1-ről tudjuk, hogy nem metrika, nyilván Λ sem az. Ebből a szempontból Λ tehát hátrányban van a *CR* indexszel szemben, amely viszont teljesíti az euklidészi feltételeket.

### *3.3.1 Szekvenciák összehasonlítása*

A biológiában központi fontosságúak a szekvenciák, mint például a nukleinsavak bázissorrendje, vagy a fehérjék aminosav szekvenciái. A bennük rejlő információ alapjában véve nominális jellegű – még akkor is biológiailag a sorrendiség *is* lényeges – és az összehasonlításokban csak a pozicionális megegyezéseket vesszük figyelembe. A kiindulópont nem a szokványos adatmátrix, hanem közvetlenül az alapegységek sorozata. A távolságfüggvények tárgyalása azonban semmiképpen sem lenne teljes, ha nem említenénk meg néhány fontosabb módszert biológiai szekvenciák összehasonlítására.

Az összehasonlítás legkritikusabb lépése a két szóban forgó szekvencia maximális illeszkedésének, átfedésének a megkeresése. Az egyik legismertebb eljárás Needleman & Wunsch (1970) optimalizációs algoritmusa, amely a következőket veszi figyelembe:

- maximális legyen a *pozicionális egyezések* száma, *M*;
- minimális legyen a *pozicionális eltérések* száma, *U*, amikor ugyanabban a pozícióban nem egyforma alapegység található a két szekvenciában (=*Hamming távolság*);
- a két szekvencia hossza nem feltétlenül azonos, a különbséget jelölje *G*. A legjobb illeszkedés megkereséséhez *G*>0 esetén a láncok valamelyikét meg kell szakítanunk, így bizonyos alapegységeknek nem lesz megfelelő párjuk a másik szekvenciában. A hasonlóság kiszámításában a megszakításokat (*"indel"*) valamilyen "büntetőponttal" vesszük figyelembe, azaz egy *w* számmal súlyozzuk. A módszer egyes változatai ezen – egyébként bevallottan önkényes – súlyértékben térnek el egymástól. A megszakításokat akár figyelmen kívül is hagyhatjuk, ekkor *w*=0. Egyes szerzők viszont a megszakításokat pozicionális eltérésnek tekintik, azaz *w*=1. Swofford & Olsen (1990)

szerint az illesztésből egyenesen ki kell hagynunk a nagy megszakításokat, mert ezek erősen eltorzíthatják a jól illeszkedő szakaszokra vonatkozó eredményt. Rövidebb megszakításokra a $w$=0,5 tekinthető jó kompromisszumnak.

- az *effektív lánchossz* a következő:

$$L = M + U + wG \tag{3.40}$$

amelyből, az egyezési indexszel (3.6) analóg hasonlóság a következő:

$$S = M / L \tag{3.41}$$

Az illesztés algoritmusa, azaz $S$ maximalizálása, számítógépet igényel, bár kisebb szekvenciákra az elemzést magunk is végrehajthatjuk. Az eljárás részletezésére itt nem vállalkozhatunk; lásd pl. Kruskal (1983), Weir (1990) vagy Waterman et al. (1991). Az $S$ együttható általános érvényű, egyaránt használható bázis vagy aminosav-szekvenciák összehasonlításában. $S$ értékét közvetlenül is felhasználhatjuk a további elemzésben, leginkább különbözőségként az 1–$S$ komplement formájában.

A CTGTATC és CTATAATCCC bázissorendekre több egyenértékű megoldást ad az algoritmus, mindegyikre $M$=6, $U$=1 és $G$=3. Egy lehetséges maximális illeszkedés a következő:

CTGTA T C
CTATAATCCC

$w$=1 esetén a szekvenciák hasonlósága $S$=6/(6+1+1□3) = 0,6.

Bázisszekvenciák esetén a szekvenciák időbeli változására az $S$ értékének a csökkenése a jellemző, amennyiben feltételezzük, hogy a négy bázis egyforma valószínűséggel cserélődik bármelyik másik bázisra pontmutáció révén. Ha μ a mutációs ráta és $t$ az eltelt idő, akkor a

$$2\mu t = K = \frac{3}{4}\ln\left(\frac{3}{4S-1}\right) \tag{3.42}$$

mennyiség használható az *evolúciós távolság* becslésére (Jukes & Cantor 1969). $K$ tehát megközelítően lineárisan növekszik az idővel, de nem minden határ nélkül: ha $S$ eléri a 0,25-öt, akkor valójában a teljesen véletlenszerűen előállított két bázissorrend várható hasonlóságát kapjuk, és $K$-nak már nincs értelme. A függvény kétségtelen hátránya, hogy nem veszi figyelembe: egy ponton több mutáció is végbemehet. Attól is eltekint, hogy az A□G és T□C átalakulások (tranzíciók, lásd még a 6.3-4 alfejezeteket) jóval gyakoribbak, mint a többi (ezt a Kimura-távolság viszont figyelembe veszi, lásd Waterman et al. 1991). Fehérjékre a fenti összefüggésbe 3 helyett 19, 4 helyett pedig 20 írandó, ha megengedjük azt az egyszerűsítést, hogy minden aminosav egyformán gyakori. Tekintve, hogy hányadosuk közel van 1-hez, a 3.42 függvény a $K$ = –ln $S$ alakra redukálódik.

A fenti tárgyalás éppen csak érintette a szekvenciák összehasonlításának szerteágazó témakörét. Egyéb formulákat, amelyek pl. megengedik a populáción belüli variabilitást is, Weir (1990) könyvében találhatunk.

## 3.4 Az ordinális skálán mért adatok esete

Ordinális típusú *változók* összehasonlítására jól ismert és kipróbált *rendstatisztikák* állnak rendelkezésre, s ezek a többváltozós analízisben is számításba jöhetnek. Akármilyen formában is

kódoltuk az adatokat, az eredeti értékeket először rangokká kell alakítanunk. A változó legkisebb értéke kapja az 1-es rangot, a következő a 2-est és így tovább. Az $x_{ij}$ adatot tehát egy $r_{ij}$ rangszám váltja fel, amely kifejezi: az $i$ változónak a $j$ objektumban megfigyelt értéke hányadik az $i$ változóra vonatkozó rangsorban. Két változó – most: két rangsor – megegyezése a szorzat-momentum korrelációs együtthatóval (3.70) analóg *Spearman-féle rang-korrelációval* számítható ki a legegyszerűbben:

$$RHO_{hi} = 1 - \frac{6\sum_{j=1}^{m}(r_{hj} - r_{ij})^2}{m(m^2 - 1)} \qquad (3.43)$$

amely teljesen megegyező sorrendekre 1, éppen ellentétes rangszámokra pedig –1 értéket vesz fel. *RHO* értéke 0 körüli amikor a két sorrend között semmiféle összefüggés nincs. A rangkorreláció használhatóságát nagymértékben korlátozzák az egyező (kapcsolt) rangok, amelyek mindenképpen jelentkeznek amikor a változó kevesebb, mint $m$-féle különböző értéket vesz fel. Viszonylag kevés számú kapcsolt rang még kezelhető ún. korrekciós formulák segítségével, de túl sok egyezés már lerontja az együttható érzékenységét, s inkább a *TAU* használata ajánlható. A rangkorrelációt leginkább olyan esetekben érdemes alkalmazni, amikor megfigyeléseink eleve bizonyos sorrendiséget jelentenek (pl. állatfajok érkezési sorrendje egy csapdára stb). A függvény levezetése megtalálható pl. Yule & Kendall (1964, p. 272) és Legendre & Legendre (1983:206-207) könyvében.

A Spearman-féle rangkorreláció erősen súlyozza a nagy rangszámbeli különbségeket, s így a kis eltérések nemigen jutnak érvényre az eredményben. Ez akár előnyös is lehet, hiszen sokszor a kis rangszámbeli eltérések csupán a kevéssé megbízható mintavételezésnek vagy megfigyelésnek tudhatók be. Ha minden rangszámbeli eltérést egyenlően akarunk figyelembe venni, mert a rangsorban a kis eltérések is jelentősek és megbízhatóak, a *Kendall-féle koefficiens* alkalmazható:

$$TAU_{hi} = \frac{4\sum_{j=1}^{m} C_j - m(m-1)}{m(m-1)} \qquad (3.44)$$

$C_j$ a következőképpen határozható meg: az 1. változó értékeit növekvő rangszám szerint felsoroljuk és melléírjuk a második változó megfelelő rangszámait. A második változó minden egyes rangszámára megszámoljuk, hogy utána hány darab nála nagyobb rangszám szerepel a sorban. Ezek összege teljes egyezésnél $m(m–1)/2$, és ezért kell az összeget 4-gyel szorozni, hogy *TAU* az 1 értéket vegye fel. Teljesen ellentétes két rangsorra viszont a $C_j$-k összege 0, a hányados tehát –1 lesz. A képlet bonyolultabb alakot ölt, ha a sorbarendezendő értékek között azonosak is vannak (lásd a következő oldalon bemutatott alternatív számításmódot).

*TAU* kiszámítását egy példával is illusztráljuk. Legyen a két összehasonlítandó változóra és 6 objektumra vonatkozó adattáblázat a következő:

1. változó: 12 16 18 14 17 20
2. változó: 15 18 19 13 12 17

A számoláshoz az alábbi kis táblázatot készítjük el:

| Az 1. változó értékei sorba rendezve | A 2. változó megfelelő értékei | Rangszámok a 2. változóra | A rangszámot követő, nagyobb rangszámok darabszáma |
|---|---|---|---|
| 12 | 15 | 3 | 3 (5, 6, 4) |
| 14 | 13 | 2 | 3 (5, 6, 4) |
| 16 | 18 | 5 | 1 (6) |
| 17 | 12 | 1 | 2 (6, 4) |
| 18 | 19 | 6 | 0 |
| 20 | 17 | 4 | 0 |
| | | | Összeg: 9 |

majd *TAU* értékét a 3.44 képlet alapján kiszámítjuk: *TAU*=(4□9 – 6□5) / (6□5 )= 0,2.

Ordinális skálán mért változók alapján az *objektumok* páronkénti összehasonlítása nehézkesebb, meglehetősen elhanyagolt téma. Sok esetben ugyanis ordinális adatokat közvetlenül elemeznek intervallum v. arányskálán mért adatokra kidolgozott eljárásokkal. Mondanunk sem kell, hogy ez nem korrekt, hiszen *ordinális változóknál az értékek közötti különbségeket nem értelmeztük, nem is beszélve a hányadosokról.* Persze be kell vallani, a változó lehetséges értékeinek sorrendi viszonyait nehezen tudjuk érvényesíteni objektumok közötti hasonlóságokban. Alkalmazhatók ugyan a nominális változókra kidolgozott indexek, de ekkor nyilvánvalóan információt veszítünk: az ordinális skála "lefelé" konvertálása nominálissá szükségképpen ezzel a következménnyel jár. Ha pedig a 3.5 rész függvényeit használjuk, azzal implicit módon áttérünk az intervallum skálára, hiszen az egyes állapotok közötti különbségnek is értelmet adunk. Elképzelhető a fenti rendstatisztikák formális alkalmazása objektumokra is – az attribútum dualitás értelmében – , mégpedig elsősorban a Kendall-féle *TAU* komplementje jöhet számításba (Diday & Simon 1976). A függvény a 3.44 formulával is kiszámítható objektumokra is (persze *m* helyett *n* írandó ekkor), de a szemléltetés kedvéért egy másik, a nyers adatokon alapuló számításmódot is bemutatunk. Legyen a két összehasonlítandó objektum *j* és *k*, és definiáljunk egy $\Delta_{hi}^{j}$ segédváltozót a következőképpen:

$$\Delta_{hi}^{j} = \begin{cases} 1 & \text{if} \quad x_{hj} > x_{ij} \\ -1 & \text{if} \quad x_{hj} < x_{ij} \\ 0 & \text{if} \quad x_{hj} = x_{ij} \end{cases}$$

Legyen $T_j$ azon változó-párok száma, amelyekre $\Delta_{hi}^{j} = 0$ a *j* objektum esetében, s definiáljuk $T_k$-t hasonlóképpen a *k*-adik objektumra. Ezek felhasználásával a keresett különbözőség:

$$DTAU_{jk} = 1 - \frac{2}{\sqrt{[n(n-1)-T_j][n(n-1)-T_k]}} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} \Delta_{hi}^{j} \Delta_{hi}^{k} \qquad (3.45)$$

Vagyis $DTAU_{jk} = 1 - TAU_{jk}$. A függvény nincs definálva arra az esetre, ha valamelyik – vagy mind a kettő – objektumban az összes változó azonos értéket vesz fel az ordinális skálán, mert ekkor $T=n(n-1)$ és a nevező 0-vá válik.

A (3.45) függvény viszonylag könnyedén kezeli az egyezéseket, amelyek ordinális változókkal jellemzett objektumok esetén nagyszámúak lehetnek. Gondoljunk például egy cönológiai adattáblázatra, amelyben a kvadrátokat a fajok Braun-Blanquet féle AD értékeivel jellemezzük (ez ui. egy ordinális változó). Egy faj 6-féle értéket vehet fel, mégpedig többnyire a skála elején lévőket (ugyanis egy kvadrátban eleve csak kevés nagy tömegességű faj lehet). Adott kvadrát fajösszetétele a fajok AD rangsorával írható le. Teljes rangsorról azonban a lehetséges értékek kis száma miatt nem beszélhetünk, a kvadrátban talált fajok AD értékei az ún. részlegesen rangsorolható adatokra jelentenek példát. Critchlow (1985) ilyen típusú adatokra sorol fel további mérőszámokat. Biológiai alkalmazásukra pl. Dale (1989) tett javaslatot: az ún. Levenshtein távolság egy speciális esetét (Ulam távolság) alkalmazta cönológiai adatok többváltozós elemzésében. E távolságmérték úgy értelmezhető, hogy hány cserét kell az egyik kvadrát (részleges) fajsorrendjében végrehajtani, hogy megkapjuk a másik kvadrát (részleges) fajsorrendjét.

A segédváltozók felhasználásával felírható egy másik formula is, amelyet Goodman & Kruskal javasolt (vö. Rudas 1986) ordinális változók asszociáltságának mérésére. A fenti jelölésekkel, objektumok összehasonlítására a Goodman-Kruskal γ a következő alakot ölti:

$$\gamma_{jk} = \frac{\sum_{h=1}^{n-1}\sum_{i=h+1}^{n}\Delta_{hi}^{j}\Delta_{hi}^{k}}{\sum_{h=1}^{n-1}\sum_{i=h+1}^{n}|\Delta_{hi}^{j}||\Delta_{hi}^{k}|} \qquad (3.46)$$

Ez valójában egy egyszerű arányszám. A nevezőben azoknak a változópároknak a száma szerepel, amelyek mind a *j*, mind pedig a *k* objektumban sorba rendezettek (nem egyezőek). A számláló pedig az 1×1 és 1×−1 szorzatok számának egymáshoz való viszonya alapján eldönti, hogy ez a sorbarendezés inkább azonos vagy eltérő irányú volt-e a két objektumban. Teljes azonosság esetén $\gamma_{jk}$=1, teljesen ellentétes sorbarendezésre pedig $\gamma_{jk}$= –1. Különbözőséget a komplementképzéssel állíthatunk elő.

### 3.5 Koefficiensek arány- és intervallumskálán mért változókra

Mivel az intervallum- és az arányskála között a formulák szempontjából a legtöbb esetben nincs különbség, az ilyen típusú adatokra alkalmas függvényeket együtt tárgyaljuk. A kivételt egyébként azok a koefficiensek jelentik, amelyek az adatok "eltolására" (egy konstans hozzáadására) nem invariánsak (húrtávolság, szögeltérés, geodéziai távolság, keresztszorzat, kovariancia). Ezeket ne alkalmazzuk olyan változókra, amelyek 0 pontja önkényes! A függvények "viselkedését" a 3.3 táblázat adatai alapján, a prezencia/abszencia koefficiensekhez hasonló módon illusztráljuk. A például szolgáló mesterséges adatok 9 objektum fokozatos megváltozását írják le egy képzeletbeli grádiens mentén olymódon, hogy minden változó viselkedését egy optimumgörbe jellemez (7.9a ábra). Ennyi elegendő ahhoz, hogy az Olvasó némi áttekintő képet kapjon a függvényekről. Részletesebb – bár nem minden koefficiensre kiterjedő – értékelésre Hajdu (1981) mutat be más adatsorokat. A példa alapján viszont magunk is elkészíthetjük a függvények bármilyen, esetleg egészen speciális célú értékelését.

A távolságfüggvények bemutatásához a legjobb kiindulópont az *euklidészi távolság*:

**3.3 táblázat.** Mesterséges adatok mátrixa a koefficiensek értékeléséhez. Az objektumok egy "grádiens" mentén egyenletesen távolodnak a kiinduló 1. objektumtól úgy, hogy a grádiensre egy optimumgörbe szerint reagálnak.

| | Objektumok | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Változók | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 |
| 6 | 3 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 |
| 7 | 2 | 3 | 4 | 4 | 3 | 2 | 1 | 0 | 0 |
| 8 | 1 | 2 | 3 | 4 | 4 | 3 | 2 | 1 | 0 |
| 9 | 0 | 1 | 2 | 3 | 4 | 4 | 3 | 2 | 1 |
| 10 | 0 | 0 | 1 | 2 | 3 | 4 | 4 | 3 | 2 |
| 11 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 4 | 3 |
| 12 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 4 |
| 13 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

$$EU_{jk} = \left[ \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 \right]^{1/2} \tag{3.47}$$

amely megfelel a mindennapi, intuitív távolságfogalomnak (3.6 ábra) s kiszámítása a jól ismert Pitagorasz-tétel általánosítása sok dimenzióra. Az euklidészi távolság a referencia-alap minden egyéb hasonlóság, különbözőség és távolság megítélésekor, mint azt a fejezet elején már említettük. $d_{jk}$ egyébként – a négyzetre emelés miatt – a nagy eltéréseket emeli ki elsősorban. Alsó határa 0, míg felső korlátja nincsen.

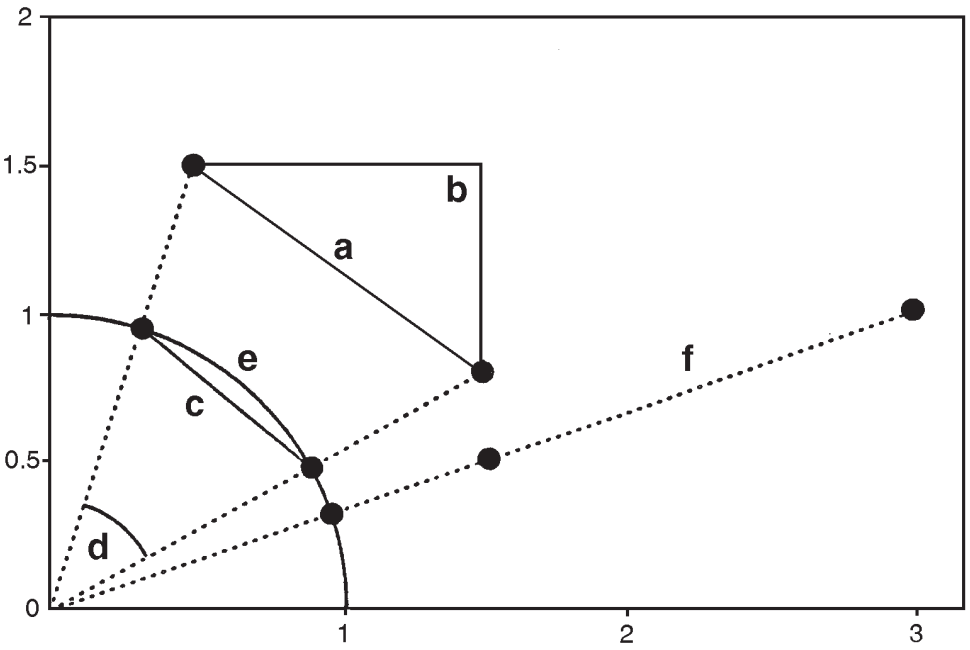
A 3.3 táblázat adataira az euklidészi távolság az 1. objektumtól először gyorsan nő, majd a közös fajok számának csökkenésével egyre kevésbé változik (3.7a ábra). Ha tovább folytatnánk a sorozatot, az 1/9 távolságérték szintjén maradnánk.

A *Manhattan-metrika* egyszerűen a két objektum közötti különbségek abszolút értékeinek az összege:

$$CB_{jk} = \sum_{i=1}^{n} |x_{ij} - x_{ik}| \tag{3.48}$$

amelyet "háztömb" ("*city block*") metrikának is neveznek, mindkét névvel utalva arra, hogy egy amerikai típusú, szabályos alaprajzú városban két pont között általában nem az euklidészi távolság a megteendő út, mert kénytelen-kelletlen meg kell kerülnünk a háztömböket (3.6 ábra). Mint a nevében is benne van, a 3.48 függvény metrika, de nem euklidészi (3.4 táblázat).

Az euklidészi távolság és a Manhattan-metrika speciális esetei egy általános függvénycsoportnak, a *Minkowski-metrikáknak:*




**3.6 ábra**. Távolságfüggvények illusztrációja kétdimenziós térben. **a**: euklidészi távolság, **b**: Manhattan-metrika, **c:** húrtávolság, **d:** szögeltérés, **e:** geodéziai távolság, **f:** a húrtávolság 0, ha a két objektumot leíró változók aránya megegyező.

$$MNK_{jk}^{(r)} = \left[\sum_{i=1}^{n} |x_{ij} - x_{ik}|^r\right]^{1/r} \tag{3.49}$$

ahol $1 \geq r$. $r$ = 1-re kapjuk a Manhattan-metrikát, $r$ = 2-re pedig az euklidészi t ávolságot. Az $r > 2$ esetben a nagy különbségek már rendkívül erős hangsúlyt kapnak; ez a többváltozós elemzésbeli alkalmazásukat nem indokolja.

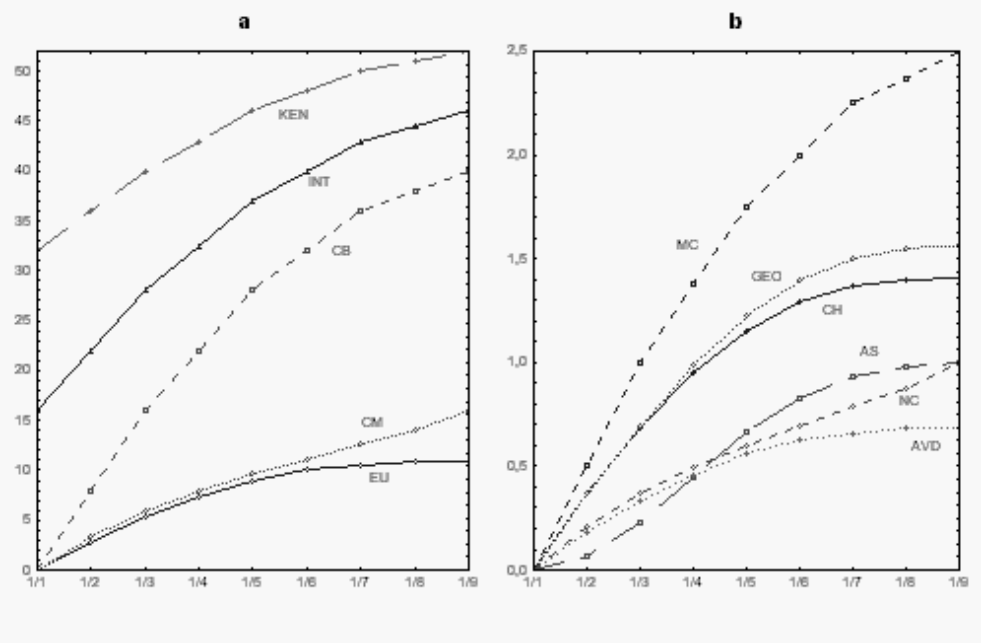
A változók számával leosztva megkapjuk, hogy egy változó átlagosan mennyivel járul hozzá a távolsághoz

$$AVD_{jk} = \frac{1}{n}\left[\sum_{i=1}^{n} (x_{ij} - x_{ik})^2\right]^{1/2} \tag{3.50}$$

illetve az abszolút eltérések összegéhez:

$$MC_{jk} = \frac{1}{n}\sum_{i=1}^{n} |x_{ij} - x_{ik}| \tag{3.51}$$

Az utóbbi függvényt numerikus taxonómusok javasolták, és *átlagos karaktereltérés* ("*mean character difference*") néven ismeretes (Cain & Harrison 1958). Ez az a formula, amit Czekanowski alkalmazott antropológiai vizsgálataiban ("*durchschnittlische Differenz*"; ezt




**3.7 ábra.** Az intervallum típusú változókra alkalmas távolságfüggvények egy csoportjának grafikus összehasonlítása. A skála módosulása miatt az *EU* és *CM* relativizált változatai *(AVD* ill. *NC*) a **b** ábrán szerepelnek.

azért jegyeztük meg, mert sok könyvben egy másik, a 3.59 formulára hivatkoznak Czekanowski index néven).

A 3.7a-b ábrákon látható, hogy a osztás művelete nem változtat a görbe alakján, viszont így a függvény jobban összehasonlítható a többi különbözőséggel.

A Manhattan metrikából származtatható a *Canberra metrika*

$$CM_{jk} = \sum_{i=1}^{n} \frac{|x_{ij} - x_{ik}|}{|x_{ij}| + |x_{ik}|} \qquad (3.52)$$

(Lance & Williams 1967b) melynek révén az egyes változók hatása jóval kiegyenlítettebbé válik. Cönológiai kvadrátok esetében például ugyanaz a különbség ritka fajok esetén sokkal nagyobb mértékben járul az eredményhez, mint a gyakori fajok esetén. Az abszolútérték jelek alkalmazásával a nevezőben, Gower & Legendre (1986) javaslata szerint, a függvény negatív értékekre is használható (pl. amikor az adatokat előzetesen a szórással standardizáltuk). Az összehasonlításból nyilván ki kell zárnunk azokat a változókat, amelyek mindkét objektumra nézve 0 értékűek.

*CM* nem euklidészi, de előnyös tulajdonsága – legalábbis a példaadatok alapján –, hogy megközelítően lineárisan változik (3.7a ábra).

**3.4 táblázat.** Intervallum-típusú adatokra alkalmas együtthatók metrikus ill. euklidészi tulajdonságai. N: nem-metrikus, M: metrikus, E: euklidészi..

| Függvény neve | Tulajdonság | Függvény neve | Tulajdonság |
|---|---|---|---|
| Euklidészi távolság | E | Pinkham - Pearson | |
| Manhattan-metrika | M | Gleason | |
| Canberra-metrika | M | Ellenberg | N |
| Húrtávolság | N | Pandeya | |
| Szögeltérés | N | khi-négyzet távolság | E |
| Geodéziai távolság | N | 1 – korreláció | N |
| Clark | E | 1 – hasonlósági hányados | |
| Bray - Curtis | N | 1 – *DKEN* | N |
| Marczewski - Steinhaus | M | Faith átmeneti koefficiens | N |
| 1 – Kulczynski | N | Uppsala koefficiens | N? |

A Canberra metrika lehetséges értékei a [0,*n*] intervallumban mozognak, ezért az *n*-nel történő osztással kapott, ún. *normált Canberra-metrika*:

$$NC_{jk} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{ij}-x_{ik}|}{|x_{ij}|+|x_{ik}|} \tag{3.53}$$

már a standard, [0,1] intervallumban vesz fel csak értékeket. Clifford & Stephenson (1975) megfontolandó javaslata szerint *n* helyett csupán azoknak a változóknak a számával kell osztanunk, melyek értéke legalább az egyik objektumban nem 0.

Amennyiben az objektumokra, mint pontokra mutató vektorokat előzetesen egységnyi hosszúságúra normáljuk (2.22 átalakítás) és ezután számítjuk ki a közöttük lévő euklidészi távolságot, akkor az ún. *húrtávolságot* (Orlóci 1978) kapjuk. A normálás az alábbi formulába be van építve, így ha ezt alkalmazzuk, előzetes standardizálásra nincs szükség:

$$CH_{jk} = \left[2\left(1-\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\sqrt{\sum_{i=1}^{n}x_{ij}^2\sum_{i=1}^{n}x_{ik}^2}}\right)\right]^{1/2} \tag{3.54}$$

Ez a távolság, mint a bináris adatokra alkalmas változatnál már említettük, az egységsugarú, origó-középpontú hipergömb felületére vetített pontok között kifeszülő húr hosszának felel meg (3.6 ábra: c). Amennyiben tehát a változók arányát tekintve a két objektum megegyezik, a húrtávolság 0 lesz (3.6 ábra: f). Emiatt a húrtávolság az eredeti pontokra nézve nem metrika, hiszen az 1. axióma nem teljesül.

A húrtávolság képletébe "beépítve" találjuk a *szögeltérést*:

$$AS_{jk} = 1 - \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2}} \tag{3.55}$$

amely a két vektor közötti szög (3.6 ábra: d) cosinusának a komplementje. Azaz, *AS* értéke 0 ha a vektorok közötti szög $0^o$ ($\cos 0^o = 1$), illetve 1 a derékszög esetében ($\cos 90^o = 0$).

A *geodéziai távolság* rokon az előző kettővel, és a két pont közötti *körív* hosszának felel meg:

$$GEO_{jk} = \arccos \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\left( \sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2 \right)^{1/2}} \tag{3.56}$$

(3.6 ábra: e). *GEO* értéke 0 és $\pi/2$ között lehet. Neve onnan származik, hogy a Föld felületén mérve két pont között valójában ezt, és nem az euklidészi a távolságot kell megtenni. A húrtávolság és a geodéziai távolság, mint a képleteikből is látható, összefügg egymással (3.7b ábra), ezért a könnyebben értelmezhető húrtávolság használata feleslegessé teszi a másikét.

Az euklidészi és a húrtávolság egy-egy függvénycsoport képviselői voltak, amelyek a változók közötti *eltéréseket* (3.47-53), ill. a változók *arányosságát* (3.54-56) veszik alapul. Az első csoportba még nagyon sokféle függvény tartozik, amelyek az eddigiek változatainak tekinthetők. A Canberra-metrikához legközelebb a Clark-féle (1952) *divergencia-koefficiens* ("*coefficient of divergence*") áll:

$$CL_{jk} = \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_{ij} - x_{ik}}{x_{ij} + x_{ik}} \right)^2 \right)^{1/2} \tag{3.57}$$

Az összegben szereplő tagok négyzetét vesszük figyelembe, e függvény tehát lényegében véve úgy viszonyul a Canberra-metrikához, mint az euklidészi távolság a Manhattan-metrikához (ui. a nagyobb eltérések jobban kifejeződnek az eredményben). A függvény értéke, az *n*-nel történő osztás miatt, teljes egyezés esetén 0, maximális különbözőség esetén pedig 1.

Az alábbi formula viszont már lényegesebben különbözik a Canberra-metrikától: az összegzés külön-külön történik mind a számlálóra, mind a nevezőre.

$$BC_{jk} = \frac{\sum_{i=1}^{n} | x_{ij} - x_{ik} |}{\sum_{i=1}^{n} (x_{ij} + x_{ik})} \tag{3.58}$$

A függvény voltaképpen egy egyszerű index formájában adja meg, hogy az összegzett értékek hányadrészében van eltérés a két objektum között. Ezt a különbözőségi formulát *Bray - Curtis*

(1957) index néven ismerik elsősorban, bár Pielou (1984) 100-zal szorzott alakban *százalékos különbözőség* (*"percentage difference"*) néven ismerteti. A formula a – tévesen – Czekanowski index néven ismert hasonlósági függvénynek a komplementje, amelyet a teljesség kedvéért külön is bemutatunk:

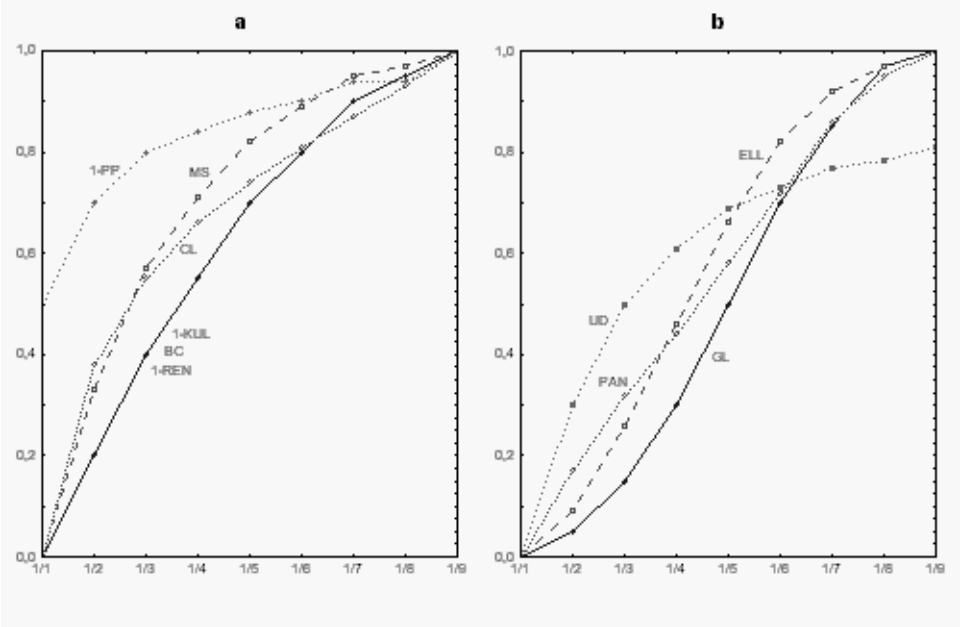
$$1 - BC_{jk} = \frac{2\sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}}{\sum_{i=1}^{n} (x_{ij} + x_{ik})} \tag{3.59}$$

Prezencia/abszencia esetben 1-*BC* a Sorensen indexszel (3.25) egyezik meg, vagyis *BC* sem metrika (3.4 táblázat). Előnyös viszont, hogy változása közelítően lineáris jellegű (3.8a ábra).

A *Marczewski - Steinhaus koefficiens* (Holgate 1971, Lewandowsky 1972) az eltérések összegét az objektumpár maximális értékeinek összegéhez viszonyítja:

$$MS_{jk} = \frac{\sum_{i=1}^{n} | x_{ij} - x_{ik} |}{\sum_{i=1}^{n} \max\{x_{ij}, x_{ik}\}} \tag{3.60}$$

A függvényt a halmazelmélet alapján is értelmezhetjük. A számláló a *j* és *k* objektumot reprezentáló halmazok szimmetrikus differenciája, a nevező pedig a halmazelméleti unió (Or-




**3.8 ábra.** Az intervallum típusú változókra alkalmas különbözőségi függvények egy csoportjának (**a**) és a növénycönológiában alkalmazott négy speciális függvény (**b**) értékelése a 3.3 táblázat adataira.

lóci 1978). *MS* metrika de nem euklidészi. Komplementje *Ruzicka index* néven ismeretes, és a következő formákban találkozhatunk vele:

$$RUZ_{jk} = 1 - MS_{jk} = \frac{\sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}}{\sum_{i=1}^{n} \max\{x_{ij}, x_{ik}\}} = \frac{\sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}}{\sum_{i=1}^{n} x_{ij} + \sum_{i=1}^{n} x_{ik} - \sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}} \tag{3.61}$$

A Ruzicka index prezencia/abszencia esetben a Jaccard-indexszel (3.24) ekvivalens.

Intervallum skálán mért változókra a *Kulczynski index* (3.29) a következő alakot ölti:

$$\frac{1}{2}\left(\frac{1}{\sum_{i=1}^{n} x_{ij}} + \frac{1}{\sum_{i=1}^{n} x_{ik}}\right)\sum_{i=1}^{n} |x_{ij} - x_{ik}| = 1 - KUL_{jk} = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}}{\sum_{i=1}^{n} x_{ij}} + \frac{\sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\}}{\sum_{i=1}^{n} x_{ik}}\right) \tag{3.62}$$

A példaadatokra $1 - KUL = BC$ (3.8a ábra), mert az összeg minden objektumra azonos.

A minimum és maximum viszonyát úgy is kifejezhetjük, hogy a hányadost még az összegzés előtt képezzük. Ekkor a maximális különbözőség *n* lesz, így *n*-nel osztva kapunk [0,1] intervallumba eső különbözőségi együtthatót:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x_{ij} - x_{ik}|}{\max\{x_{ij}, x_{ik}\}}\right) = 1 - PP_{jk} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\min\{x_{ij}, x_{ik}\}}{\max\{x_{ij}, x_{ik}\}}\right) \tag{3.63}$$

A hasonlósági függvény *Pinkham & Pearson koefficiens* néven ismeretes. Hasonlatokkal élve: $1 - PP$ úgy viszonyul *MS*-hez, mint a normált Canberra-metrika (*NC*, 3.53) a Bray - Curtis indexhez (*BC*, 3.58). $1 - PP$ azonban nem metrika, hiszen egy objektum önmagától vett különbözősége nem 0, s lefutása is elég szabálytalan (3.8a ábra). Az első problémán úgy segíthetünk, ha nem *n*-nel, hanem a nem dupla 0-ás változók számával osztunk.

A téma iránt jobban érdeklődő Olvasók kedvéért megemlítünk néhány, a növénycönológusok körében ismert formulát. Ezek a két állomány vagy kvadrát közötti hasonlóság kiszámításában a közös fajok esetleges mennyiségi eltéréseit akár figyelmen kívül is hagyhatják, azaz rájuk nézve a függvények prezencia/abszencia koefficiensként működnek. Ilyen hasonlósági index a Gleason-féle (1920) formula

$$GL_{jk} = \frac{\sum_{i \in A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik})} \tag{3.64}$$

ahol *A* azon fajok halmaza, amelyek mind *j*-ben, mind pedig *k*-ban jelen vannak. A számlálóban az összegzés tehát a közös fajokra vonatkozik. A nevező így annyival több a számlálónál, amekkora a nem közös fajokban mutatkozó mennyiségi különbség. Ellenberg ezt a mennyiségi különbséget kétszeresen veszi figyelembe:

$$EL_{jk} = \frac{\sum_{i \in A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik}) + \sum_{i \notin A} x_{ij} + \sum_{i \notin A} x_{ik}} \tag{3.65}$$

(vö. Goodall 1973a), a különbséget a példaadatok esetében a gradiens közepén emeli ki jobban (3.8b ábra). Rokon jellegű a *Pandeya koefficiens*,

$$PAN_{jk} = \frac{\sum_{i \in A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik}) + \sum_{i \in A} | x_{ij} - x_{ij} |} \tag{3.66}$$

amely azonban már a mindkét helyen meglévő fajok mennyiségi különbségeit is figyelembe veszi különbözőséget növelő tényezőként.

Az eltéréseket mérő koefficiensek közül megemlítjük az ún. $\chi^2$*-távolságot*, ami az adattáblázat sorainak és oszlopainak összegével való kettős standardizálás után számított euklidészi távolságnak felel meg:

$$CHISQ_{jk} = \left[ \sum_{i=1}^{n} \frac{1}{\sum_{h=1}^{m} x_{ih}} \left( \frac{x_{ij}}{\sum_{s=1}^{n} x_{sj}} - \frac{x_{ik}}{\sum_{s=1}^{n} x_{sk}} \right)^2 \right]^{1/2} \tag{3.67}$$

a $\chi^2$-távolság fontossága a korreszpondencia-elemzéssel kapcsolatosan (7.3 alfejezet) válik nyilvánvalóvá. Távolságfüggvényként önmagában ritkán jön számításba.

Az arányokra érzékeny együtthatók közül hármat (*AS, CH, GEO*) – más típusú távolságok társaságában – már említettünk. Most sor kerülhet még néhány hasonló célú, s nem kevésbé fontos mérőszám bemutatására is. Minden formulában vektorok skaláris szorzata szerepel (vö. C függelék) s ennek alapján már ránézésre felismerhető, hogy mely függvény érzékeny a változók közötti arányokra. Az adatmátrix két oszlopára felírhatjuk az ún. *keresztszorzatot* ("*cross product*"):

$$CP_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \tag{3.68}$$

amelyet nyers adatokra ritkán alkalmazunk (pl. nem-centrált PCA, 7.1.5 rész). Rendszerint az adatmátrixot előzőleg oszlopok szerint centráljuk, és az így módosított értékekből számolunk a 3.68 egyenlet alapján. A kapott *eltérésszorzat-összeg*et *m*–1-gyel osztva adódik a *kovariancia.* Ennek képlete a nyers adatokból kiindulva a következő:

$$COV_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1} \tag{3.69}$$

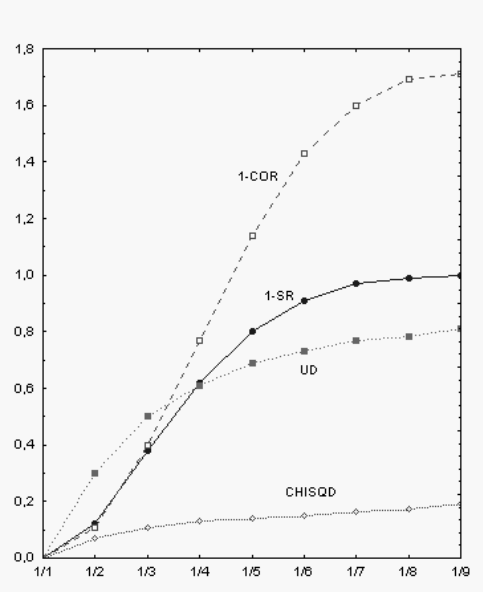
amely jól ismert a standard statisztikából is változók kapcsolatának a mérésére. Mivel a kovariancia nem korlátos mértékszám, azaz felső és alsó határa nincs, helyette inkább a *korreláció* jön számításba. Ez is kiszámítható a 3.68 egyenlet alapján, ha az adatokat előzetesen *oszlopok szerint* a szórással standardizáltuk. Közvetlen kiszámítására az alábbi – jól ismert – képlet szolgál:

$$COR_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \qquad (3.70)$$

Távolsággá alakítva – figyelembe véve a már említett (2.1 alfejezet) gondokat – objektumok között is alkalmazható. A 3.68-70 függvények igazi felhasználási területe azonban az, amikor változók közötti összefüggéseket mérünk velük a főkomponens vagy kanonikus korreláció analízis kezdetén (7.1-2 alfejezet). A korreláció különbözőséggé alakítva sem metrika, hiszen 0 értéket kapunk két nem egyenlő objektumra is, ha az egyik adatait a másiknak valamilyen konstans értékkel való szorzásával megkaphatjuk. Az indexelés megfelelő átalakításával a korrelációt a sorokra (változókra) is felírhatjuk.

Ebbe a csoportba tartozik a *hasonlósági hányados* ("*similarity ratio*", Wishart 1969, van der Maarel 1979) is

$$SR_{jk} = 1 - \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - \sum_{i=1}^{n} x_{ij}x_{k\,i}} \qquad (3.71)$$




**3.9 ábra**. Néhány távolság és különbözőségi index változása a példaadatokra (3.3 táblázat).

amelynek értékei a [0,1] intervallumba esnek, 1 jelöli a teljes egyezést. Prezencia/abszencia adatok esetén *SR* megegyezik a Jaccard indexszel. A korrelációval fennálló erős rokonsága a 3.9 ábráról is leolvasható.

Az eltéréseket ill. az arányosságot vizsgáló függvényeken kívül megemlítendő egy harmadik függvénytípus is. Ezek a két összehasonlított objektumot leíró változók *minimális egyezésére* érzékenyek (Faith 1984). A függvénycsalád alaptípusa a Kendall (1970) féle minimális egyezési együttható:

$$KEN_{jk} = \sum_{i=1}^{n} \min[x_{ij}, x_{ik}] \tag{3.72}$$

amely különbözőséggé alakítva a következőképpen is felírható:

$$DKEN_{jk} = \sum_{i=1}^{n} \{\max_h[x_{ih}] - \min[x_{ij}, x_{ik}]\} \tag{3.73}$$

A Kendall-féle hasonlóság a halmazelméleti metszetnek felel meg. Nem korlátos mérték (nincs felső határa), és ezért elsősorban akkor célszerű használni, ha az adatokat előzetesen standardizáltuk. Az oszlopok (objektumok) összege szerinti standardizálást tartalmazza például az állatökológusok körében népszerű *Renkonen index:*

$$REN_{jk} = \sum_{i=1}^{n} \min \left\{ \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}} \right\} = 1 - 0.5 \sum_{i=1}^{n} \left| \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} - \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}} \right| \tag{3.74}$$

Egyik gyakori elnevezése ("*percentage similarity of distribution*", Whittaker & Fairbanks 1958) magyarázza meg e függvény jelentését: a standardizálás ugyanis egyedszámadatok esetén pl. azzal az eredménnyel jár, hogy egy relatív gyakoriságeloszlást kapunk mindkét objektumra, és 100×*REN* ezek százalékos megegyezését jelenti. A standardizálás révén egyébként a változók objektumon belüli aránya válik fontossá, s ezáltal elmosódik az arányosságra, ill. minimumra érzékeny koefficiensek közötti – ezek szerint nem is olyan éles – határ. A példaadatokra 1–*REN* megegyezik *BC*-vel (3.8a ábra), de ebből nem szabad általános következtetéseket levonni, mert eme egyezés az oszlopösszegek azonosságának a következménye.

*Átmeneti formák.* A különböző érzékenységű koefficiensek között közvetlen átmeneteket képezhetünk, s ezáltal mindkettő hatása jelentkezik az eredményben. Faith (1984) és Faith et al. (1987) javasolták például a Manhattan-metrika és a Kendall koefficiens egyszerű átlagát ("*intermediate coefficient*"):

$$INT_{jk} = \frac{1}{2}\left[ \sum_{i=1}^{n} | x_{ij} - x_{ik} | + \max_h[x_{ih}] - \min[x_{ij}, x_{ik}] \right] \tag{3.75}$$

E függvénynek nincs felső korlátja, bár ez *n*-nel való osztással megoldható. Egy másik átmeneti jellegű formula az "*Uppsala koefficiens*" (Noest & van der Maarel 1989):

$$UD_{jk} = \frac{1}{n - z_{jk}} \sum_{i=1}^{n} \frac{1}{2} \left[ \frac{|x_{ij} - x_{ik}|}{x_{ij} + x_{ik}} + \frac{|x_{ij} - x_{ik}|}{x_{\max} - x_{\min}} \right] \qquad (3.76)$$

ahol $z_{jk}$ a *j* és *k* objektumból egyaránt hiányzó változók száma (az osztás tehát nem *n*-nel történik, ellentétben más függvényekkel!) és $x_{\max} - x_{\min}$ pedig a változók által felvehető értékek tartománya. A függvény a Bray-Curtis index és a terjedelemmel standardizált Manhattan-metrika (l. Gower-index, 3.103) közötti átmenet. E függvény jellemzője, hogy a skála elején levő eltérések súlyozottabban járulnak a különbözőséghez, mint a skála végén levők. Például, ha $x_{\max} - x_{\min}=9$, akkor a 0 és 1 eltérése 0,566-tal járul az összeghez, a 8 és 9 eltérése pedig csak 0,085-tel. A nagyobb értékek eltéréseinek fontosságát csökkentve implicit módon ugyanazt csináljuk, mintha az adatokat előzőleg logartimikus transzformációval módosítottuk volna.

*Genetikai távolságok.* Az intervallum, ill. arányskálán mérhető változók speciális eseteit jelentik az *allélgyakoriságok.* Az objektumok ekkor *populációk,* a változók pedig annyi csoportba oszthatók, ahány *lókuszt* vizsgálunk. Az allélgyakoriságokat minden egyes lókuszra az összeg szerint standardizálni kell, s így a táblázatban lókuszonkénti relatív gyakoriságok szerepelnek. A relatív géngyakoriság-adatokra számos speciális távolságfüggvény áll rendelkezésre, amelyek figyelembe veszik a változók csoportosulását és genetikailag többé-kevésbé értelmezhetőek is. (Ha a lókuszokat "összemosnánk", akkor az előzőekben bemutatott függvények nagy része megfelelne a távolság mérésére, de ez nem lenne "genetikai"). Az értelmezhetőség arra utal, hogy a genetikai távolság a populációk szétválása óta eltelt idővel van összefüggésben, s ezért a változást okozó mutációról és sodródásról egy jól megfogalmazott modellre van szükség. Természetesen enélkül is számítható távolság, de ekkor ennek csupán geometriai jelentése lehet és nem képezheti alapját pl. az evolúciós folyamatok értelmezésének (Weir 1990).

A távolságmértékek viselkedését az egyszerűség kedvéért egy egylókuszos/kétalléles esetre fogjuk illusztrálni. Az allélgyakoriságok példamátrixában az 1. populációtól való fokozatos távolodás tükröződik (ennek hátterét most nem firtatjuk), míg végül az egyik allél teljesen lecserélődik a másikra:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. allél: | 1,0 | 0,9 | 0,8 | 0,7 | 0,6 | 0,5 | 0,4 | 0,3 | 0,2 | 0,1 | 0,0 |
| 2. allél: | 0,0 | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 | 1,0 |

A továbbiakban a lókuszok számát *L* jelöli, az adatmátrix egy értékére pedig $x_{hij}$ utal, amely a *h* lókusz *i* alléljének a relatív gyakorisága a *j* populációban. $n_h$ jelöli az allélek számát a *h* lókuszon. Miután relatív gyakoriságokat használunk, a populációt képviselő pontok egy hipersíkon vannak minden egyes lókuszra nézve (két allél esetén a 2.9c és a 3.11 ábrán látható egyenesen).

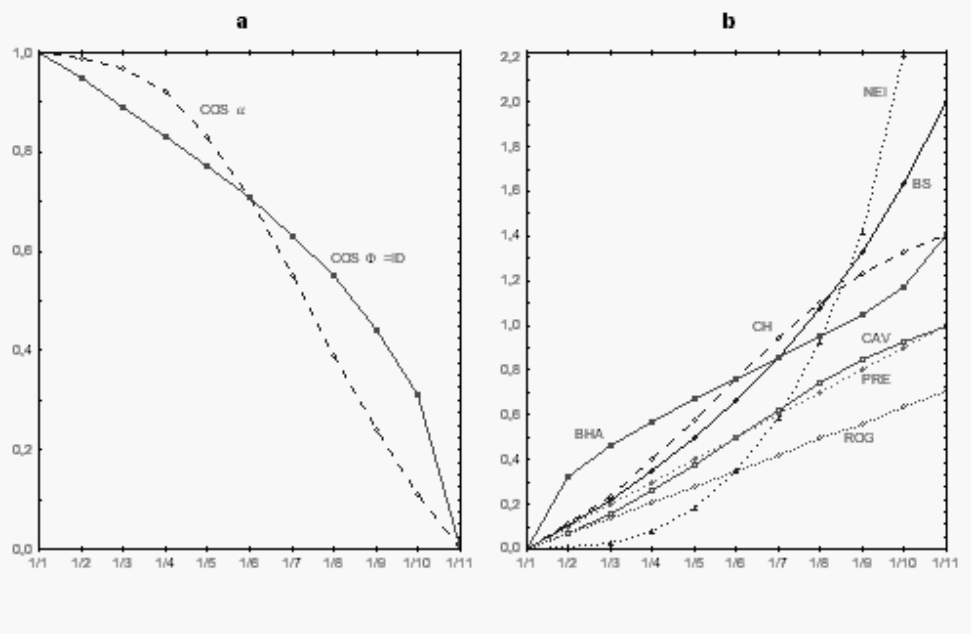
Lényegében véve az átlagos euklidészi távolságnak felel meg a *Rogers-féle* (1972) *genetikai távolság*:

$$ROG_j = \frac{1}{2L} \sum_{h=1}^{L} \left[ \sum_{i=1}^{n_h} (x_{hij} - x_{hik})^2 \right]^{1/2} \qquad (3.77)$$

melyet a populáción belüli heterozigócia jelentősen befolyásolhat. Legfőbb hátrányául ugyanis azt hozhatjuk fel, ami az euklidészi távolság ökológiai alkalmazásának is fő akadálya: előfordulhat, hogy kisebb a távolság közös alléllel nem is rendelkező két populáció között, mint két másik, néhány allélban megegyező populáció között. Hasonlóan kritizálható a *Prevosti-féle genetikai távolság* (cf. Wright 1978), azaz az átlagos karaktereltérés lókuszonként:

$$PRE_{jk} = \frac{1}{2L}\sum_{h=1}^{L}\sum_{i=1}^{n_h} | x_{j\,i} - x_{k\,i} | \qquad (3.78)$$

A relatív gyakoriságok közötti különbségek alkalmazása a távolság kifejezésére geometriailag jól interpretálható ugyan, a fenti nehézség miatt azonban a genetikusok többre tartják az arányosságra érzékeny függvényeket. Ezek közé tartozik a leggyakrabban használt együttható, a *Nei-féle genetikus azonosság* ("*genetic identity*", Nei 1972, 1978) és több származéka. Az identitást egylókuszos esetre voltaképpen a 3.55 függvénnyel mérhetjük (az 1-ből való kivonás nélkül), amely a két populációra mutató vektor hajlásszögének ($\alpha$, 3.11 ábra) a cosinusa. Ennek értéke teljes azonosság esetén 1, teljes különbözőség esetén pedig 0 (3.10a ábra). Miután a képletben relatív gyakoriságok szerepelnek, az eredménynek *valószínűségi interpretációja* is van. A számláló azon valószínűségnek a becslése, hogy a két populációból származó egy-egy egyed a lókuszon azonos allélt hordoz ($\hat{q}_{jk}$). A nevezőben szereplő két négyzetösszeg pedig annak az eseménynek a valószínűségét becsli, hogy az ugyanabból a populációból származó két egyed azonos allélt hordoz ($\hat{q}_j$, ill. $\hat{q}_k$). A nevező értéke a két populációra vonatkozó valószínűségek mértani közepe:




**3.10 ábra.** Genetikai szögfüggvények (**a**) és távolságmértékek (**b**) változása egy lókusz és két allél esetén a két populáció teljes eltávolodásával (allélgyakoriságok a szövegben). *CH* a 3.54 szerinti húrtávolság.

$$ID_{jk} = \frac{\sum_i x_{ij} x_{ik}}{\left( \sum_i x_{ij}^2 \sum_i x_{ik}^2 \right)^{1/2}} = \frac{\hat{q}_{jk}}{\sqrt{\hat{q}_j \hat{q}_k}} = \cos a \qquad (3.79)$$

A formula tehát a *j* és *k* populációk közötti génazonosság és a populációkon belüli génazonosság hányadosaként fogható fel. A függvény *L* lokuszra a következőképpen általánosítható:

$$ID_{jk} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h\,i} x_{k\,i}}{\left( \sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h\,i}^2 \sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{k\,i}^2 \right)^{1/2}} \qquad (3.80)$$

amely azonban nem ad torzítatlan becslést, s ezért kis mintanagyság (*m*, amely most azonos minden populációra) esetén korrigálni kell (Nei 1978):

$$IDC_{jk} = \frac{(m-1) \sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{h\,i} x_{k\,i}}{\left( \sum_{h=1}^{L} (2m \sum_{i=1}^{n_h} x_{h\,i}^2 - 1) \times \sum_{h=1}^{L} (2m \sum_{i=1}^{n_h} x_{k\,i}^2 - 1) \right)^{1/2}} \qquad (3.81)$$

A Nei-féle génazonosság akkor válik igazán genetikailag értelmezhetővé, ha a populációk szétválása óta eltelt *időt* tudjuk vele kifejezni. Ekkor sokféle modell jöhet számításba. A legegyszerűbb esetben az adott allélből bármely másik allélba való mutációt tételezünk fel $\mu$ mutációs ráta mellett. Ekkor fennáll az alábbi összefüggés:

$$NEI_{jk} = -\ln ID \approx 2\mu t \qquad (3.82)$$

amely a *Nei-féle genetikai távolság*. Ez nincs definiálva arra az esetre, amikor minden allél csak az egyik populációt jellemzi (3.10b ábra). A Nei-távolság lényeges leegyszerűsítéseket tartalmaz, mert feltételezi, hogy a populációk elválása óta a mutáció egyformán valószínű minden lókuszon és mindkét leszármazási vonalon (Hillis 1984). Ezt a problémát Hillis a lókuszonként vett genetikus azonosságok aritmetikai átlagával hidalja át:

$$HIL_{jk} = \ln \left[ \frac{1}{L} \sum_{h=1}^{L} \frac{\sum_{i=1}^{n_h} x_{h\,i} x_{k\,i}}{\left( \sum_{i=1}^{n_h} x_{h\,i}^2 \sum_{i=1}^{n_h} x_{k\,i}^2 \right)^{1/2}} \right] \qquad (3.83)$$

Ennek is megadható a torzítatlan becslése, a 3.81 formulához analóg módon (Swofford & Olsen 1990). Egylókuszos esetre *HIL* megegyezik *NEI*-vel.

A Nei-féle távolság nem alkalmas arra az esetre sem, amikor a populációk eltávolodását kizárólag a sodródás okozza. Ekkor egy geometriai jellegű mértékszám, a *Balakrishnan - Shangvi távolság* jöhet számításba (Weir 1990):

$$BS_{jk}^2 = \frac{1}{\sum_{h=1}^{L} n_h - 1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{(x_{ij} - x_{ik})^2}{x_{ij} + x_{ik}} \tag{3.84}$$

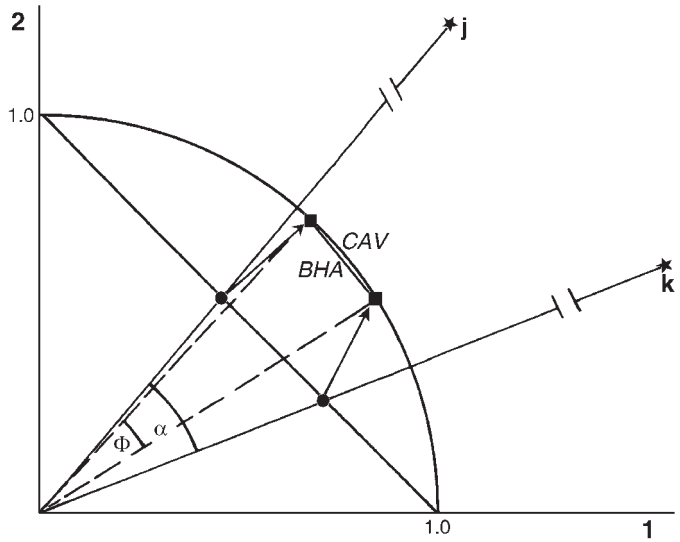
A genetikai távolság definiálása és értelmezése meglehetősen komplikált terület, s ezt legjobban a Cavalli-Sforza és munkatársai által használt formulákkal illusztrálhatjuk. Egy adott *h* lókuszon az allélek relatív gyakoriságait *négyzetre emelve* a populációkat képviselő pontok közvetlenül rákerülnek az egységsugarú hipergömb felszínére (3.11 ábra). A *j* és *k* pontokra mutató vektorok hajlásszöge ekkor egyszerűen megkapható:

$$\cos\Phi = \sum_{i=1}^{n_h} \sqrt{x_{hij} x_{hik}} \tag{3.85}$$

Ennek figyelembevételével Cavalli-Sforza & Edwards (1967) az egyes lókuszokra kapott geodéziai távolságok (ívhosszok, lásd a 3.56 egyenlet és a 3.11 ábra) standardizálásával és átlagolásával definiálta a genetikai távolságot:

$$CAV_{jk} = \left[ \frac{1}{L} \sum_{h=1}^{L} \left( \frac{2}{\pi} \arccos \sum_{i=1}^{n_h} \sqrt{x_{hij} x_{hik}} \right)^2 \right]^{1/2} \tag{3.86}$$

vagy pedig a hipergömbre vetített pontok közötti húrtávolságot mérte:




**3.11. ábra.** Néhány genetikai távolság geometriai értelmezése két populáció között egy lókuszra és két allélre. ✯ jelöli az eredeti gyakoriságértékeket (*j*-re 10; 12, míg *k*-ra 20; 8). ● jelöli a relatív gyakoriságokat, ■ pedig a négyzetre emelt, így a körívre került relatív gyakoriságokat.

$$BHA_{jk} = \left[ 2 - 2\sum_i \sqrt{x_{ij}x_{ik}} \right]^{1/2} = \left[ \sum_i \left( \sqrt{x_{ij}} - \sqrt{x_{ik}} \right)^2 \right]^{1/2} \qquad (3.87)$$

(*Bhattacharyya távolság*, vö. Mardia et al. 1979, 3.11 ábra), és ezt átlagolta a lókuszok szerint. Weir (1990) úgy véli, hogy ezek kizárólag geometriai mértékszámok, mindennemű genetikai jelentés nélkül. A gond azonban az, hogy Φ és a Nei-féle genetikus azonosságban szereplő α nem azonos (3.11 ábra), s úgy tűnik, hogy az utóbbinak van még geometriailag is könnyebben érthető jelentése. A Φ szög alkalmazása mellett Tóthmérész (1986) értelmezésében az szól, hogy cos Φ – a végső szakasztól eltekintve – közelítőleg lineárisan csökken az allélgyakoriságok közötti eltérés növekedésével, míg ez nem áll fenn a cos α -ra (3.10a ábra). Swofford & Olsen (1990) határozottan a Cavalli-Sforza-féle mértékek mellett áll, s genetikai interpretációt is ad. Eszerint a sodródási szituációt a 3.86 függvény jól magyarázza, mivel a távolság értéke független a kezdeti géngyakoriságoktól. Mardia et al. (1979: 379) mutatja be egy lókuszra, hogy Weirrel szemben Swofford & Olsennek lehet igaza, hiszen a Balakrishnan - Shangvi távolság és a Bhattacharyya távolság között egyszerű matematikai összefüggés áll fenn.

*A niche-átfedés mérőszámai.* Fajok ökológiai nichének mérése és a niche-átfedés számolása alkalmas kiindulópontot jelenthet a fajok közötti kapcsolatok többváltozós elemzésére. A *niche-átfedés* mérőszámai ugyanis távolság- v. hasonlósági függvénynek is felfoghatók, s talán már nem is kell mondanunk, hogy máshonnan már ismerős függvények a "niche zsargonban" akár külön néven is szerepelhetnek. Ilyen például a *Schoener* (1970) index, amely a fajokra alkalmazott Renkonen indexnek felel meg (standardizálás tehát az egyes fajok egyedszámösszege szerint!) Megemlíthető még a *Horn formula* is (Horn 1966), amely információelméleti megfontolásokon alapszik. Legyen most *n* a mintavételi helyek száma, és ezek az adatmátrix soraiban szerepeljenek. A *j* oszlopvektort a *j* faj gyakoriság-eloszlásaként foghatjuk fel, s a faj *niche-szélességét* a Shannon-féle entrópiával fejezhetjük ki:

$$\hat{H}_j = -\sum_i \frac{x_{ij}}{\sum_h x_{hj}} \log \frac{x_{ij}}{\sum_h x_{hj}} \qquad (3.88)$$

A *j* és *k* fajok teljes átfedésben vannak, ha a két oszlopvektor összeadásával a fenti entrópia nem változik. Az egyesített vektorokra ez a minimális érték, melyet $\hat{H}_{\min}$ jelöl. A két faj a lehető legnagyobb mértékben különbözik, azaz az átfedés 0, ha sohasem fordulnak elő együtt. Ekkor az egyesített oszlopvektorokra számított entrópia legyen $\hat{H}_{\max}$. Minden aktuális érték, $\hat{H}_{obs}$, e két szélsőség közé esik. Az alábbiak szerint standardizálva:

$$HN_{jk} = \frac{\hat{H}_{\max} - \hat{H}_{obs}}{\hat{H}_{\max} - \hat{H}_{\min}} \qquad (3.89)$$

a függvény a 0 értéket veszi fel teljes különbözőség, 1-et pedig teljes egyezés esetén. A számolásra alkalmas formula a következő:

$$HN_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}+x_{ik})\log(x_{ij}+x_{ik}) - \sum_{i=1}^{n} x_{ij}\log x_{ij} - \sum_{i=1}^{n} x_{ik}\log x_{ik}}{(x_{ij}+x_{ik})\log(x_{ij}+x_{ik}) - x_{.j}\log x_{.j} - x_{.k}\log x_{.k}} \qquad (3.90)$$

ahol $x_{.j}$ és $x_{.k}$ a $j$ és $k$ oszlopok összegét jelöli. Más interpretációval a formula mintavételi helyek között hasonlósági indexként is alkalmazható.

*Alakbeli hasonlóság és távolság.* Penrose (1954) szerint az euklidészi távolság két összetevőre bontható fel, az egyik rész tisztán a "*méretbeli*" különbségeknek tudható be, a másik pedig az "*alakbeli*" eltérések eredménye:

$$d_{jk}^2 = (n-1)SHAPE_{jk}^2 + nNSIZE_{jk}^2 \tag{3.91}$$

Ha két objektum összehasonlításában a méretbeli különbségeket nem akarjuk figyelembe venni, csak az alakbeli egyezés az érdekes, akkor a Penrose javasolta formula alkalmazható:

$$SHAPE_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-x_{ik})^2 - \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(x_{ij}-x_{ik})\right]^2 \tag{3.92}$$

Ez lényegében véve a két összehasonlított objektumra az egyes tulajdonságokban mutatkozó eltérések *varianciája* (négyzetek átlaga – átlag négyzete). Várhatóan akkor nagy az értéke, ha az eltérések nagyságrendjében és irányában nagy különbségek mutatkoznak a két objektum között. A méretbeli koefficiens:

$$SIZE_{jk} = \left[\frac{1}{n^2}\left[\sum_{i=1}^{n}(x_{ij}-x_{ik})\right]^2\right]^{1/2} \tag{3.93}$$

viszont akkor lesz nagy, ha a különbségek általában egyirányúak.

> A Penrose féle *SHAPE* függvénnyel szemben a korreláció (3.70) az alakbeli hasonlóság kifejezésére jobban használható (Rohlf & Sokal 1965). A főkomponens analízis speciális változatai pedig (vö. 7.6 alfejezet) még árnyaltabb elemzési lehetőséget nyújtanak a modern morfometriában, így a Penrose koefficienseknek ma már kisebb a jelentősége.

*Általánosított távolság.* Ha az euklidészi távolságot alkalmazzuk, akkor az egymással korreláló változók hatását valójában túlhangsúlyozzuk. A belső súlyozás egy speciális esetéről beszélhetünk, amellyel gyakorlatilag mindig találkozhatunk, hiszen a biológiai változók rendszerint korrelálnak egymással. Az alábbi kis adatmátrix illusztrálja a belső súlyozás hatását:

| | | | | |
|---|---|---|---|---|
| 1. változó | 5,1 | 6,2 | 7,1 | 8,0 |
| 2. változó | 4,0 | 5,0 | 6,2 | 7,3 |
| 3. változó | 3,0 | 2,0 | 9,0 | 6,0 |

Az első két változó között erős pozitív korreláció van, s lehetséges, hogy ezek voltaképpen egy harmadik, nem vizsgált háttérváltozó hatását tükrözik. Mindkettőt figyelembe véve megnöveljük a háttérváltozó jelentőségét a 3. változóhoz képest. Ez nemkívánatos *lehet* az eredmények interpretációjában. A belső súlyozást azonban a *Mahalanobis-féle* (1936) *általánosított távolság* ("*generalized distance*") alkalmazásával kiküszöbölhetjük:

$$GEND_{jk}^2 = \sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(x_{hj}-x_{hk})(x_{ij}-x_{ik}) \tag{3.94}$$

vagy mátrixalgebrai felírásban

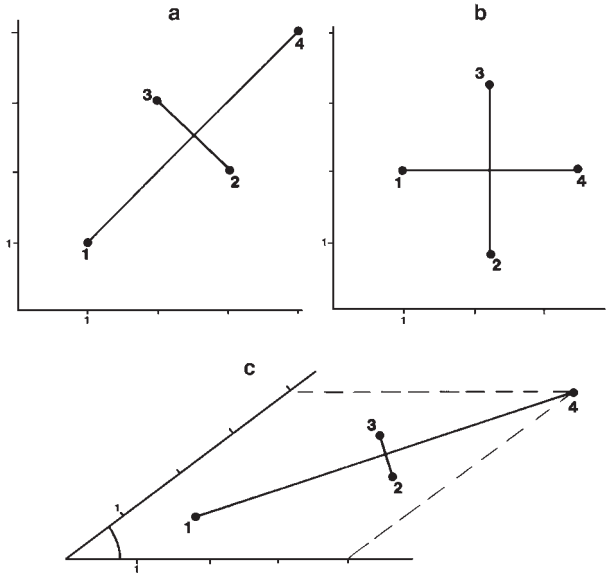
$$GEND_{jk}^2 = (\mathbf{x}_j - \mathbf{x}_k)'\mathbf{W}^{-1}(\mathbf{x}_j - \mathbf{x}_k) \tag{3.95}$$

ahol $\mathbf{x}_j$ és $\mathbf{x}_k$ a $j$ és $k$ objektumoknak megfelelő oszlopvektorok, $\mathbf{W}^{-1}$ az $n$ változó varianciakovariancia mátrixának az inverze (C függelék), $w_{hi}$ annak egy eleme. A Mahalanobis távolság a változók szórását egységnyire standardizálja. Emiatt, ha az eredeti változók teljesen korrelálatlanok, akkor 3.95 eredménye megegyezik a standardizált adatokból számított euklidészi távolság négyzetével. Az általánosított távolságok mátrixa metrikus információt tartalmaz standardizált és egymásra merőleges tengelyekre. Egy ilyen mátrixból végrehajtott főkoordináta-elemzés (lásd 7.4.1 rész) tehát teljesen egyformán "fontos" tengelyeket hoz létre (azaz a variancia arányosan oszlik meg a tengelyek között).

A Mahalanobis távolság változókat "kiegyenlítő" hatása a 3.12a-b ábra alapján érzékelhető. A négy pont euklidészi ill. Mahalanobis távolságai a következő félmátrixokba foglalhatók össze:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | | | | | 0 | | | |
| 2,23 | 0 | | | illetve | 1,73 | 0 | | |
| 2,23 | **1,41** | 0 | | | 1,73 | **2,45** | 0 | |
| **4,24** | 2,23 | 2,23 | 0 | | **2,45** | 1,73 | 1,73 | 0 |

A Mahalanobis távolság a pontok elrendeződéséből a változás két fő "irányára" érzékeny (az 1-4 és a 2-3 pontok elhelyezkedése szerint), s ezeket azonos fontosságúnak tekinti. Következésképpen a $d_{14}$ és $d_{23}$ távolságok azonosak lesznek (3.12b ábra). Ezekről a fő irányokról jóval többet fogunk látni a 7. fejezetben.




**3.12 ábra**. Pontok euklidészi távolságai *derékszögű* koordináta rendszerben (**a**), általánosított távolságai egy *önkényes,* derékszögű koordináta rendszerben *(***b**) és euklidészi távolságai egy *ferdeszögű* koordináta rendszerben (**c**). $COR_{12}$ = 0,8, a tengelyek közötti szög tehát arc cos 0,8 = 36,8$^0$. A szaggatott vonal a 4. pont helyének meghatározását segíti a ferdeszögű koordináta rendszerben.

Az általánosított távolságot objektumok csoportjai (pl. populációk) közötti távolság mérésére is alkalmazhatjuk (ez valójában a tradicionális felhasználási terület). Ekkor a következő formulával dolgozunk:

$$GEND_{jk}^2 = (\mathbf{x}_j - \mathbf{x}_k)'\mathbf{W}^{-1}(\mathbf{x}_j - \mathbf{x}_k) \tag{3.96}$$

ahol $\bar{\mathbf{x}}_j$ és $\bar{\mathbf{x}}_k$ a *j* és *k* csoport átlagvektorai (azaz: az egyes változók átlagai oszlopvektorban összesítve), és $\mathbf{W}^{-1}$ pedig a **W** egyesített variancia-kovariancia mátrix inverze (az összes csoportra, az adatokat összevonva kell ezt kiszámítanunk). A távolságnak csak akkor van értelme, ha a csoportonként számítható kovarianciák azonosak (helyesebben: ugyanazon közös kovarianciának a becslései) és a változók többváltozós normális eloszlásúak. Sneath & Sokal (1973) véleménye szerint azonban a távolság nem túl érzékeny e feltételek megsértésére ("*robusztusság*"). Megjegyzendő, hogy az általánosított távolság kiszámítása csak akkor lehetséges, ha az objektumok száma nem kisebb a változók számánál. Ellenkező esetben a **W** mátrix szinguláris (C függelék) és nem invertálható. Ugyancsak ez a helyzet, ha bármely két változó között –1 vagy 1 a korreláció értéke, illetve ha valamelyik változó varianciája 0.

*Távolság nem derékszögű koordináta-rendszerekben.* Mindeddig nem mondtuk ki, annyira egyértelmű volt, hogy adatainkat egy olyan koordinátarendszer segítségével ábrázoljuk, ahol a tengelyek közötti szög mindig $90^o$. A derékszögű koordinátarendszerből áttérve egy ferdeszögűbe, ahol a tengelyek közötti szögek cosinusa a korrelációnak felel meg, a pontok közötti távolságban a változók közötti kapcsolatoknak is szerep jut. A távolságformula a következő:

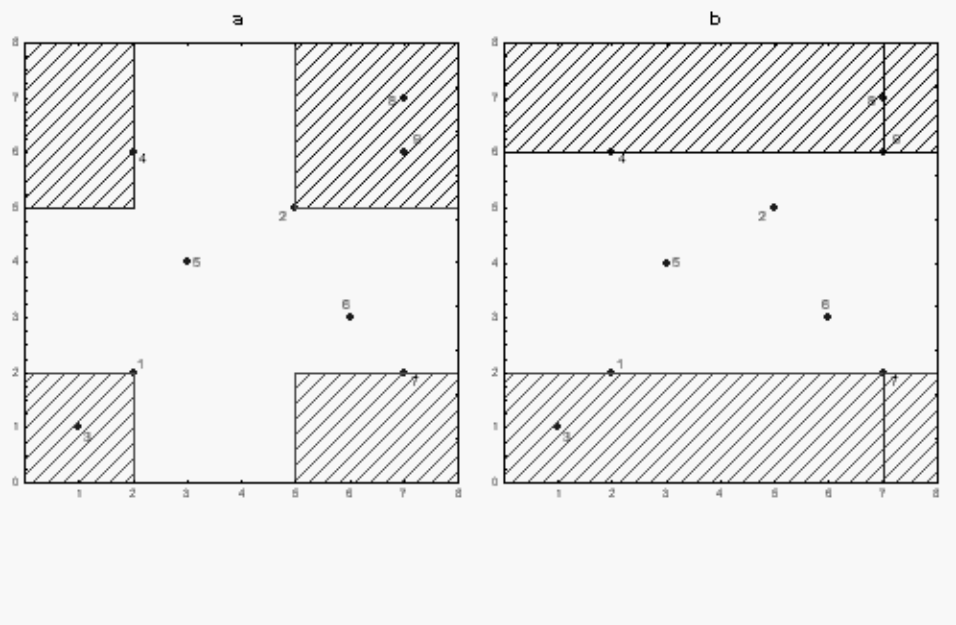
$$OBL_{jk} = \left[ \sum_{h=1}^{n} (x_{hj} - x_{hk})^2 + 2 \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} (x_{hj} - x_{hk})(x_{ij} - x_{ik}) COR_{hi} \right]^{1/2} \tag{3.97}$$

ahol $COR_{hi}$ a *h* és *i* változók 3.70 szerint számított korrelációja (Orlóci 1978: 49). A függvény összetevői a négyzetes euklidészi távolság és egy korrekciós második tag. Ez utóbbi értéke pozitív, ha a *j* és *k* objektum "megfelel" a változók közötti korrelációknak (mint pl. az 1. és 4. pont a 3.12a ábrán). Ekkor az új távolságérték nagyobb lesz, mint a derékszögű koordinátarendszerben mért euklidészi távolság. Ha a két objektum relatív helyzete "ellentmond" a változók korrelációinak (2. és 3. pontok, 3.12a és c ábra) akkor a korrekciós tag negatív, és az új távolságérték az euklidészinél kisebb lesz. Ferdeszögű koordináta rendszerbe áttérve tehát a korreláló változók hatása – az objektumpártól függően – vagy kidomborodik vagy pedig elenyészik.

Összegezve, a Mahalanobis távolsággal ellentétben a ferdeszögű koordináta rendszerben mért távolság a két eredeti változó korrelációja alapján kimutatható fő irányt emeli ki, s az erre merőleges hatást negligálja. A 3.12c ábrán a pontok távolságmátrixa a következő:

| | | | |
|---|---|---|---|
| 0 | | | |
| 2,92 | 0 | | |
| 2,92 | **0,64** | 0 | |
| **5,69** | 2,92 | 2,92 | 0 |

*Speciális mértékek.* Az arány- vagy intervallum-skálán mért változókra alkalmas mérőszámok között több olyan is akad, amelyek sehogyan sem oszthatók be logikusan az előzőekben tár-




**3.13 ábra.** A Calhoun távolság megállapítása két változóra. **a:** az 1. és 2. pontra, **b**: a 7. és 9. pontra.

gyalt csoportok egyikébe sem. Ilyen mértékszám az objektumok topológiai viszonyaira, relatív elhelyezkedésére érzékeny *Calhoun távolság* (Bartels et al. 1970). A távolság alapja az, hogy két adott pont között a sokdimenziós térben hány további pont helyezkedik el (azaz a 3.94 és 3.97 távolságokhoz hasonlóan, sőt annál közvetlenebbül, a többi pont is befolyásolja két pont távolságát). A Calhoun mértékszám kiszámítását a 3.13 ábra illusztrálja az alábbi adatok segítségével :

| | |
|---|---|
| 1. változó | 2 5 1 2 3 6 7 7 7 |
| 2. változó | 2 5 1 6 4 3 2 7 6 |

Adott objektumpárt kiválasztva minden egyes változóra egy intervallum határozható meg, ezen intervallumok a sokdimenziós térben egy hiperfelületet jelölnek ki. Az 1. és 2. pontra a fenti példában a 3.13a ábra *nem* árnyalt részeiről van szó.

A Calhoun távolság kiszámításához a következőket kell figyelembe venni:

$n_1$ = azon pontok száma, amelyek a két pont által meghatározott hipersík belsejébe esnek (5. és 6. pontok a 3.13a ábrán);

$n_2$ = a hipersík peremére eső pontok száma, ezek legalább egy változóban megegyeznek a *j* vagy a *k* objektummal (a 4. és 7. pontok, 3.13a ábra);

$n_3$ = azon pontok száma, amelyek legalább egy változóban mindkét ponttal megegyeznek és a hipersíkon kívül esnek (a 3.12a ábrán ilyen pont nem látható; ha azonban a 7. és 9. pontok közötti Calhoun távolságot keressük, akkor a 8. pont ilyen pozícióban van, 3.13b ábra).

Ezek után a keresett távolság:

$$CAL_{jk} = w_1 n_1 + w_2 n_2 + w_3 n_3 \qquad (3.98)$$

amelyben $w_1$, $w_2$ és $w_3$ önkényesen megadott súlyok (Bartels et al. eredeti javaslata szerint értékük 6, 3 ill. 2). Orlóci (1978) szerint logikus lenne a $CAL_{jk} = n_1$ definíció (ekkor $w_2$=$w_3$=0), hiszen csak $n_1$ pont esik ténylegesen a $j$ és $k$ pont közé s így elkerülhetjük az önkényes súlyozást is. A Calhoun távolság nem metrika (pl. két pont távolsága 0 lehet akkor is ha nem esnek egybe), ennek ellenére érdemes megpróbálkozni vele, mert skálabeli eltérésekre nem érzékeny.

Goodall (1964, 1966) javasolta a *valószínűségi hasonlóságot* ("*probabilistic similarity index*"), amely két objektum hasonlóságát a többi hasonlóság függvényeként definiálja. A páronkénti hasonlóságot tehát az egész minta befolyásolja, s ebben emlékeztet a 3.94, 3.97 és 3.98 függvényekre. Az alapkoncepció azonban lényegesen eltér az előzőektől, mint az alábbi számításmenet is mutatja.

1. Legyen $d_{i,jk} = | x_{ij} - x_{ik} |$, azaz a $j$ és $k$ objektum *Manhattan távolsága* az $i$ változó szerint. Az $m$ elemű mintában nyilván $m(m–1)/2$ ilyen érték van minden egyes változóra. Rendezzük nagyság szerinti sorba a $d_{i,jk}$ értékeket.

2. Az $i$ változóra definiáljuk a $j$ és $k$ objektumok különbözőségét aszerint, hogy milyen arányban szerepelnek a $d_{i,jk}$-nál kisebb vagy azzal megegyező értékek a mintában. Azaz, legyen

$$p_{i,jk} = \frac{\#(d \le d_{i,jk}}{m(m-1)/2} \qquad (3.99)$$

Minél nagyobb ez az érték, annál nagyobb eltérés mutatkozik a két objektum között a minta egészéhez viszonyítva. $p_{i,jk}$ annak a valószínűsége, hogy az adott objektumpárra az $i$ változó értékei legfeljebb $d_{i,jk}$ mértékben térnének el, ha $x_{ij}$-t és $x_{ik}$-t véletlen módon választanánk ki az összes ($m$) érték közül. A valószínűség tehát fordított arányban van a hasonlósággal.

3. Miután $p_{i,jk}$ értékét minden változóra kiszámítottuk, álítsuk elő a következő szorzatot:

$$q_{jk} = \prod_{i=1}^{n} p_{i,jk} \qquad (3.100)$$

4. Most egy másik sorbarendezés következik: az $m(m–1)/2$ darab $q$ értéket rangsoroljuk. A $j$ és $k$ objektumok hasonlóságát a $q_{jk}$-nál nagyobb értékek aránya adja meg:

$$GD_{jk} = \frac{\#(q > q_{jk}}{m(m-1)/2} \qquad (3.101)$$

Ez annak az eseménynek a valószínűsége, hogy a $j$ és $k$ objektum legalább olyan hasonló egymáshoz, mint az adott esetben, hogyha a változók értékeit teljesen véletlenszerűen választanánk ki a mintaösszletből.

A Goodall-féle index – nem vitatható – ötletes kifejezése a mintán belüli relatív hasonlóságoknak. Ez ugyanakkor hátránynak is bizonyulhat, mert a hasonlóságok csak az adott mintára érvényesek: egyetlen egy új objektum vagy változó hozzáadása teljesen felboríthatja a hasonlósági struktúrát. Bár a sorbarendezésnél elvész a metrikus információ, a 3.101 index mégis hasznos lehet a biológiai osztályozásokban. Kiemelendő a mérési skálától való függetlensége. A 3.99 függvény megfelelő átalakítással kiterjeszthető pl. a nominális és ordinális változókra is.

Ha a változókat teljesen függetlennek tekinthetjük, akkor nem csupán formális hasonlóság számolható. Fisher (1963) megmutatta, hogy az alábbi mennyiség

$$X^2 = -\ln \sum_{i=1}^{n} \ln p_{i,jk} \qquad (3.102)$$

$\chi^2$ eloszlást követ $2n$ szabadságfok mellett. Minél nagyobb 3.102 értéke, annál nagyobb a hasonlóság a két objektum között.

A fenti index egyik kiterjesztése az *affinitási index* (Goodall 1968), amely egy objektum "vonzódását" méri egy csoporthoz, figyelembe véve a csoporthoz nem tartozó összes egyéb objektumhoz való hasonlóságát is. Ennek alapján eldönthető, hogy ezt az objektumot beosszuk-e a csoportba. A *deviancia index* (Goodall 1966) viszont ellentétesen jár el: kifejezi, hogy az objektumok mennyire térnek el attól a populációtól, amelybe beosztottuk őket.

## 3.6 Koefficiensek kevert adattípusokra

A többféle változótípust tartalmazó adathalmazra nem használható egyik eddig említett távolság- és hasonlóságfüggvény sem. Ez a probléma ugyan a változók átalakításával megoldható lenne, de ez részben információ-veszteséggel jár vagy pedig valamilyen külső információ figyelembe vételével lehetséges csupán. Ha adatainkat eredeti formában szeretnénk hagyni (s ez a gyakoribb eset), akkor a megoldást a kevert adattípusra kidolgozott speciális formulák jelentik. Legismertebb közülük a Gower (1971b) index, amelynek további előnye, hogy hiányzó adatokat is megenged. A képlet a következő:

$$GOW_{jk} = \frac{\sum_{i=1}^{n} w_{kj} s_{kj}}{\sum_{i=1}^{n} w_{kj}} \qquad (3.103)$$

ahol $w_{ijk} = 0$ ha a $j$ és $k$ objektumok összehasonlítása nem lehetséges az $i$ változóra, mivel az $x_{ij}$ vagy $x_{ik}$ értéke ismeretlen. Ezen kívül

a) bináris változókra:

$w_{ijk} = 1$ és $s_{ijk} = 0$ ha $x_{ij} \neq x_{ik}$
$w_{ijk} = s_{ijk} = 1$ ha $x_{ij} = x_{ik} = 1$ vagy ha $x_{ij} = x_{ik} = 0$ és a dupla nullákat (közös abszenciákat) figyelembe vesszük;
$w_{ijk} = s_{ijk} = 0$ ha $x_{ij} = x_{ik} = 0$ és a dupla nullákat kizárjuk az összehasonlításból;

b) nominális változókra:

$w_{ijk} = 1$ ha $x_{ij}$ és $x_{ik}$ ismert; ekkor
$s_{ijk} = 0$ ha $x_{ij} \neq x_{ik}$
$s_{ijk} = 1$ ha $x_{ij} = x_{ik}$

c) intervallum és arányskálán mért változókra:

$w_{ijk} = 1$ ha $x_{ij}$ és $x_{ik}$ ismertek; s ekkor $s_{ijk} = 1 - \{ | x_{ij} - x_{ik} | / (\text{az } i \text{ változó terjedelme}) \}$.

A Gower index sem tudja azonban kezelni az ordinális típusú változókat. Komplementje különbözőségi indexként jöhet számításba. Megjegyzendő, hogy a bináris esetre, ha a dupla nullákat figyelembe vesszük, a Gower index az egyezési koefficienssel (3.6), ha pedig mellőzzük, akkor a Jaccard indexszel (3.24) azonos. Nominális változókra a 3.33 indexnek felel meg, intervallum és arányskála esetén pedig a változók terjedelmével történő standardizálás alapján számított Manhattan távolsággal (3.48) arányos.

A fenti koefficiens egyik alternatívája a következő távolságformula (Podani 1980):

$$DM_{jk} = \left( \sum_{i=1}^{n} w_{kj} \left[ \frac{x_{ij} - x_{ik}}{q_{kj}} \right]^2 \right)^{1/2} \tag{3.104}$$

ahol $w_{ijk} = 0$ ha a $j$ és $k$ objektumok összehasonlítása az $i$ változóra hiányzó adatok miatt nem lehetséges, egyébként $w_{ijk} = 1$;

a) bináris változókra:
$q_{ijk} = 1$.

b) nominális változókra:
$q_{ijk} = x_{ij} - x_{ik}$ ha $x_{ij} \neq x_{ik}$
$q_{ijk} = 1$ ha $x_{ij} = x_{ik}$

c) intervallum és arányskálán mért változókra:
$q_{ijk} = \max(x_{ih}) - \min(x_{ih})$ ; $h$=1,..., $m$.

## 3.7 Távolságok általánosítása kettőnél több objektumra (heterogenitási mértékszámok)

Számos klasszifikációs eljárás nem az objektumok között páronként értelmezett távolságok alapján számol, hanem két v. több objektum alkotta objektumhalmaz valamilyen belső tulajdonságát fejezi ki. Ezekre a belső sajátságokra – jobb szó híján – *heterogenitás* néven utalunk (ennek komplementje lesz a *homogenitás*). Objektumok csoportjainak heterogenitását részben a szokványos statisztika mérőszámaival, részben pedig információelméleti függvényekkel fejezhetjük ki.

A legismertebb heterogenitási mértékszám az objektumhalmazra vonatkozó *eltérésnégyzet-összeg* ("sum of squares"):

$$SSQ_A = \sum_{i=1}^{n} \sum_{j \in A} (x_{ij} - \bar{x}_{iA})^2 \tag{3.105}$$

ahol $\bar{x}_{iA}$ az $i$ változó átlaga az $A$ objektumhalmazban. A 3.105 képlet az $A$-n belüli objektumok között mért páronkénti euklidészi távolságok segítségével is kifejezhető:

$$SSQ_A = \frac{\sum_{j=1}^{m_A} \sum_{k=1}^{m_A} d_{jk}^2}{2m_A} \tag{3.106}$$

ahol $m_A$ az $A$-ban levő objektumok száma. Ennek alapján két objektumra az eltérésnégyzet-összeg a közöttük értelmezett euklidészi távolság négyzetének a fele:

$$SSQ_{jk} = d^2_{jk} / 2 \tag{3.107}$$

Az eltérésnégyzet-összeget az objektumok számával elosztva a *varianciát* kapjuk:

$$VAR_A = SSQ_A / m_A = \frac{\sum_{i=1}^{n} \sum_{j \in A} (x_{ij} - \bar{x}_{iA})^2}{m_A} \tag{3.108}$$

amelyet a következőképpen is felírhatunk:

$$VAR_A = \frac{\sum_{j=1}^{m_A} \sum_{k=1}^{m_A} d^2_{jk}}{2m^2_A} \tag{3.109}$$

Két objektumra pedig a variancia a következő

$$VAR_{jk} = d^2_{jk} / 4 \tag{3.110}$$

Az objektumok közötti távolságok vagy különbözőségek, ($DIS_{jk}$), *átlagával* is kifejezhető a heterogenitás:

$$AVG_A = \frac{\sum_{j=1}^{m_A - 1} \sum_{k=1}^{m_A} DIS_{jk}}{(m^2_A - m_A)/2}, \quad j,k \in A \tag{3.111}$$

Ennek előnye, hogy bármilyen távolságfüggvényre alkalmazható nem-euklidészi térben is, míg az eltérésnégyzet-összeg és a variancia az euklidészi távolságkoncepcióhoz kapcsolódik.

Ha egy $m_A$ objektum alkotta $A$ halmazt $n$ nominális változó ír le, ahol az $i$ változó állapotainak a száma $p_i$, akkor az objektumhalmaz heterogenitása a *súlyozott entrópiaösszeggel* is kifejezhető:

$$H_A = nm_A \log m_A - \sum_{i=1}^{n} \sum_{h=1}^{p_i} f_{hi} \log f_{hi} \tag{3.112}$$

ahol $f_{hi}$ az $i$ változó $h$ állapotának a gyakorisága az $A$ halmazban. A 3.112 függvény voltaképpen az objektumok *rendezetlenségének* a mértékszáma. Minimális a rendezetlenség, ha az objektumok minden egyes változóra nézve teljesen egyöntetűek, s maximális, ha minden egyes változóra és annak minden $h$ állapotára $f_{hi} = m_A / p_i$. A $p$=2 esetre és két objektumra a fenti formula a 2×2-es kontingenciatábla jelöléseivel is felírható:

$$H = 2\,(b+c) \log 2 \tag{3.113}$$

amely tovább egyszerűsödik a 2-es alapú logaritmus alkalmazásával:

$$H = 2\,(b+c) \tag{3.114}$$

Az $A$ objektumhalmaz jellemzésére alkalmas másik információelméleti mérőszámot a változók közötti *kölcsönös információ* jelenti. Alacsony érték a változók közötti egyezésre utal, következésképpen az objektumok közötti hasonlóságok nagyok. Bináris adatokra felírva a kölcsönös információ a következő:

$$I_A = (n-1)m_A \log m_A - \sum_{i=1}^{n}\left[ f_i \log f_i - (m_A - f_i)\log(m_A - f_i) + \sum_{g=1}^{\omega} f_g \log f_g \right] \quad (3.115)$$

ahol $f_i$ az $i$ változó előfordulásainak a száma az $A$ csoportban, $f_g$ pedig a $g$ változó-kombináció gyakorisága az $A$ csoportban. A lehetséges változó-kombinációk száma $\omega = 2^n$. Két objektumra az alábbi kifejezést kapjuk:

$I = 2\ (b + c - 1) \log 2$ ha $b+c > 0$; (3.116)

illetve

$I = 0$ ha $b+c = 0$. (3.117)

Megjegyzendő, hogy a 3.115 mérőszám kiemelt fontosságú a sokfajú pontmintázatok elemzésében (Juhász-Nagy 1976).

## 3.8 Irodalmi áttekintés

Éles ellentétben a mintavételezést és az adatátalakítást tárgyaló szűkös szakirodalommal, a távolságfüggvényekről könyvtárnyi terjedelmű anyag áll rendelkezésünkre. Az adott problémához leginkább illő függvény kiválasztása számos könyvfejezet és nagyon sok cikk tárgya. Emellett szinte hetente "fedeznek fel" új, speciális igényeket kielégítő formulákat is. Az alábbi összesítésben emiatt csak a legfontosabb, a témát egy-egy szempontból részletesen áttekintő forrásokat említjük.

A növényökológiában használatos függvényekről a legteljesebb összeállítást Goodall (1973a) és Orlóci (1978) adja. Pielou (1984) és Greig-Smith (1983:194-195) már inkább csak néhány fontosabb függvényre összpontosít, de azokat alaposabban megvizsgálja. Legendre & Legendre (1983:170-215) sok függvényt sorol ugyan fel, de néhány megállapításukkal nehéz egyetérteni. Mindenütt beleütközhetünk az R és Q mód megkülönböztetésébe (azaz fajok, ill. mintavételi helyek az objektumok). A prezencia/abszencia koefficiensekről az első értékelő összesítést, paleontológiai szempontból, Cheetham & Hazel (1969) közölte. Kenkel & Booth (1987) viszont a prezencia/abszencia koefficiensek biogeográfiai alkalmazhatóságát vizsgálta meg. Megjegyzendő, hogy a Baroni-Urbani- Buser féle index mellett érvelnek, bár az Ochiai és a Jaccard együtthatókat is elfogadhatónak találják. Lamont & Grant (1979) és Hajdu (1981) számos együtthatót hasonlított össze, megvizsgálva, hogy miképpen változik az értékük különböző szituációkban. Grafikus értékelési módszerük adta az ötletet az itt használt szemléltetéshez is. Ezt a módszert vette át Shi (1993) is nem kevesebb, mint 39 különböző prezencia/abszencia koefficiens vizsgálatában. További összehasonlító értékeléseket találunk Campbell (1978), Janson & Vegelius (1981), Hubálek (1982), Wolda (1981), Jackson et al. (1989) és – legújabban – Batagelj & Bren (1995) cikkeiben. Taxonómusok számára Sneath & Sokal (1973) monográfiájában található a mindmáig legjobb értékelés, bár ezt a könyvet nemcsak rendszertanosoknak ajánljuk. Egyik nagy értéke a könyvnek a csaknem teljes bibliográfia a numerikus taxonómia kezdeti korszakából. A mikrobiológusok figyelmét Austin & Colwell (1977) prezencia/abszencia koefficienseket értékelő cikkére hívjuk fel.

A matematika eszköztárát is figyelembe véve Anderberg (1973) úttörő könyve ma is nagy haszonnal forgatható. Az egyes függvények euklidészi és metrikus sajátságait Gower & Legendre (1986) vizsgálta meg részletesen. Az információelméleti módszerek legrészletesebb összefoglalása Feoli et al. (1984) monográfiájában lelhető fel.

Szólnunk kell a speciális területekről is. Szekvenciák összevetésében pl. ma már nemcsak a könyvünkben említett módszerek jöhetnek számításba (lásd pl. a Miyamoto & Cracraft [1991] szerkesztette kötetet). A niche-átfedés mérőszámairól Abrams (1980), Hurlbert (1982) and Ganis (1991) nyújt további információt. Az alakbeli hasonlóságot, mint említettük, már nemigen szokták objektumok távolságaival definiálni. A biológiai formák értékelésében az utóbbi tíz évben jelentős fejleményeknek lehettünk tanúi. Eme új, geometriai morfometria eredményeiről még olvashatunk a 7.6 alfejezetben.

Aki fellapozza a fent említett művek akár egy részét is, megállapíthatja: meglehetősen ingoványos területre tévedt. Szinte alig akad olyan függvény, amelyet egyformán ítélne meg a szakirodalom. Különböző célok, különböző objektumok, más és más szempontok keverednek időnként nagy összevisszaságban. Könnyen lehet az is, hogy egy-egy függvényt teljesen ellentétesen ítélnek meg, mint pl. a húrtávolságot, amelyet Kenkel & Orlóci (1986) kifejezetten előnyösnek tekint ökológiai ordinációkban, míg Faith et al. (1987) ökológiailag irrelevánsnak vél. Nagy szükség lenne tehát egy modern, áttekinthető, a témát alaposan feltáró elemzésre, de ez még várat magára. Ugyancsak sok ellentmondásra, sőt hibákra bukkanhatunk a függvények metrikus, ill. euklidészi tulajdonságait illetően. A Russell - Rao indexet például több cikk is a metrikus formulák közé sorolja, bár ennek komplementje nyilván nem metrika, hiszen egy objektum önmagától vett távolsága csak akkor 0, ha $d$=0. Azaz az első metrikus axióma nem teljesül!

*3.8.1 Számítógépes programok*

A nagy, kommerciális programcsomagok általában kevés számú, de általánosan ismert, és a legtöbb probléma megoldásában alkalmazható függvényt tartalmaznak. Ezzel szemben számos, kevésbé elterjedt program ismeretes, amelyek sokkal szélesebb választékot nyújtanak (3.5 táblázat), "feleslegesen megnehezítve" – mondhatnánk ironikusan – a felhasználó dolgát. Ezeket tehát akkor ajánlhatjuk, ha a speciálisabb függvényeket szeretnénk alkalmazni elsősorban.

A táblázatban nem jutott hely minden említésre érdemes programnak. Szekvenciák elemzésére például számos programcsomag készült, közülük csak néhányat emelhetünk ki. Nukleinsav bázissorrendek illesztésére, és a Jukes - Cantor távolság számítására alkalmas pl. a University of Wisconsin Genetics Computer Group (Devereux et al. 1984) programcsomagja. nukleinsav szekvenciák közötti távolságok számítására és ezek további elemzésére fejlesztett ki a téma elismert szakértője, Nei (1991) egy programcsomagot.

Goodall valószínűségi indexe és sok rokon jellegű függvény szerepel a Goodall et al. (1991) kidolgozta programokban. A Calhoun távolság kiszámítására Orlóci (1978) közöl egy BASIC nyelvű programlistát sok más, jól használható programmal egyetemben. Ludwig & Reynolds (1988) ugyancsak BASIC nyelvű programcsomagja is tartalmazza az ismertebb hasonlósági és távolságfüggvényeket. Információelméleti mértékszámokra Feoli et al. (1984) könyvében találunk FORTRAN nyelvű programokat.

Sok – a kötetben is szereplő – formula nem található meg a táblázatban, és nincs tudomásunk olyan programcsomagokról sem, amelyek tartalmaznák ezeket (pl. Gleason, Ellenberg függvények, stb.). Ha ezekre van szükségünk, célszerű egy saját programot készíteni, pl. BASIC nyelven, majd az így kiszámított távolságmátrix már beolvasható lesz további elemzésekre, pl. a **SYN-TAX** és a **NuCoSA** (Tóthmérész 1994) esetében.

**3.5 táblázat.** Hasonlósági és távolságfüggvények különböző programcsomagokban. A táblázatban nem szerepel olyan függvény, amelyet a jelen kötet nem tárgyal.

| | BMDP 7 | Statistica | NT-SYS | SYN-TAX | NuCoSA |
|---|---|---|---|---|---|
| egyezési koefficiens | + | + | + | + | + |
| Rogers - Tanimoto | | | + | + | + |
| Anderberg I | | | | + | |
| Anderberg II | | | + | + | |
| PHI | + | | + | + | + |
| Yule II | | | + | + | + |
| Baroni-Urbani - Buser I | | | | + | + |
| Baroni-Urbani - Buser II | | | | + | + |
| Russell - Rao | | | + | + | + |
| Kulczynski (p/a) | | | + | + | + |
| Jaccard | + | | + | + | + |
| Sorensen/Dice | + | | + | + | + |
| Ochiai | + | | + | + | + |
| Fager | + | | | | |
| Spearman Rho | | | | | + |
| Kendall Tau | | | | | + |
| Jukes - Cantor | | | + | | |
| euklidészi távolság | + | + | + | + | + |
| Manhattan-metrika | | + | | + | + |
| Minkowski általános formula | + | + | | | |
| átlagos távolság | | | + | + | |
| átlagos karaktereltérés | | | + | + | |
| Canberra-metrika | | | + | + | + |
| normált Canberra-metrika | | | | + | |
| húrtávolság | | | | + | |
| szögeltérés | + | | + | + | + |
| geodéziai távolság | | | | | |
| Pinkham - Pearson | | | | | + |
| Bray-Curtis/százalékos kül. | + | | + | + | |
| Marczewski-Steinhaus/Ruzicka | | | | + | |
| Kulczynski | | | | | |
| khi$^2$ távolság | + | | + | | + |
| keresztszorzat | | | | + | |
| kovariancia | + | + | | + | |
| korreláció | + | + | + | + | + |
| hasonlósági hányados | | | | + | + |
| Kendall/Renkonen | | | + | | |
| Rogers | | | + | | |
| Prevosti | | | + | | |
| Nei | | | + | | |
| Balakrishnan - Shangvi | | | + | + | |
| Cavalli-Sforza - Edwards | | | + | | |
| Horn | | | | + | |
| Penrose size | | | + | + | |
| Penrose shape | | | + | + | |
| általánosított távolság | | | | + | |
| távolság ferdeszög• koord. rend. | | | | | + |
| Gower kevert adatokra | | | | + | |
| Távolság kevert adatokra | | | | + | |

### 3.9 Kérdezz – válaszolok!

**K:** *Meg kell hagyni, jól elárasztottál ezekkel a különféle koefficiensekkel. Teljesen megfájdult a fejem, mire végigolvastam ezt a fejezetet, és a sok-sok név bizonnyal éjszaka sem hagy majd nyugton.*

**V:** El kell ismernem, hogy egy elég fárasztó, bár igen fontos részen vagy túl, – de ezt nem lehetett megkerülni. A módszertani sokféleséget bizonyára sikeresen érzékeltettem. Egyébként nem véletlen, hogy a most bemutatott függvények jelentős részét biológusok vagy biológiai problémákkal szembenéző statisztikusok "agyalták ki". S ha tudnád, hogy még milyen sok van, amelyre itt már nem jutott hely!? A hasonlóság- és távolságfüggvények legnagyobb és legáttekinthetetlenebb irodalma talán éppen a biológiával kapcsolatos.

**K:** *Már az elejétől zavart egy kissé, hogy hol távolságról, hol különbözőségről, hol pedig hasonlóságról beszéltél. Bár tudom, hogy mi közöttük az eltérés, azért jó lenne ha ezekre a függvényekre valamilyen gyüjtőnévvel együttesen utalhatnánk.*

**V:** Egyetértek: sok esetben nem volt könnyű az egyértelmű fogalmazás, és néha a terminológiába is belebonyolódtam. Egyébként létezik ilyen gyűjtőfogalom, a *"resemblance"*, amelyet – ha jól tudom – Orlóci (1972, 1978) használt először ezzel a céllal. Bár a resemblance szó eredeti jelentése leginkább a *hasonlóság*, általános gyűjtőnévként is jól meghonosodott a szakirodalomban. A "komparatív függvény" elnevezés (Podani 1980) is alkalmazható, bár eddig nem is használtam. Ha valakinek jobb ötlete adódna, azt szívesen vennénk.

**K:** *Ha már olyan jól elárasztottál bennünket a komparatív vagy nem is tudom mi néven nevezendő függvényekkel, akkor legalább adnál némi útmutatót, hogy mikor melyiket lehet alkalmazni! A szövegből, a táblázatokból és a rajzok alapján elég nehéz eldönteni, mikor mit használjak!*

**V:** Egyértelmű választ, hogy ekkor és ekkor márpedig csak ez és csak ez a függvény jöhet számításba én nem adhatok, s tartok tőle: ilyen tanácsot senkitől sem fogsz kapni. A függvényt magadnak kell kiválasztanod, s ehhez bizony meg kell értened az egyes függvények jelentését, s látnod kell, hogy bizonyos esetekben ezek miként viselkednek. Egy nagyon általános útmutatót persze össze tudok állítani, Legendre & Legendre (1983) és Gower & Legendre (1986) után "szabadon", hiszen csak az alapötlet származik tőlük. Az eddig leírtak figyelembevételével a következő "koefficiens-határozókulcsot" adhatom a kezedbe, amely a legtöbb fent említett formulát tartalmazza (a speciálisakat nem):

**1a** A változók nem egyforma típusúak, az adatokban nem szerepel ordinális változó ..... Gower (3.103), távolság (3.104)

**1b** Az összes változó azonos típusú ..... **2**

**2a** A változók nominális típusúak (bináris esetben is, azaz a kódolás önkényes) ..... **3**

**2b** A változók más típusúak ..... **7**

**3a** Egyszerű hányadosok, elsősorban objektumok összehasonlítására ..... **4**

**3b** Függetlenséget v. megjósolhatóságot mérik, elsősorban változók összevetésére alkalmasak . **5**

**4a** Az egyezést és a különbözőséget okozó vátozókat egyformán súlyozzuk .....egyezési index (3.33)

**4b** Az egyezéseket kétszeresen súlyozzuk .....Sokal - Sneath I (3.35)

**4c** A különbözőséget kétszeresen súlyozzuk ..... Rogers - Tanimoto (3.34)

**5a** Metrika, változók függetlenségét méri .....Cramér (3.37)

**5b** Nem-metrika, kölcsönös megjósolhatóságot mér ..... **6**

**6a** Adataink binárisak .....Yule I (3.16)

**6b** A változók többállapotúak .....Goodman - Kruskal lambda (3.39)

**7a** A változók ordinálisak ..... **8**

**7b** A változókat intervallum vagy arányskálán mérjük (binárisak is lehetnek!) ..... **9**

**8a** Elsősorban változók összevetésére, kevés egyezéssel, a nagy eltérések erőteljes kiemelésével ..... Spearman rho (3.43)

**8b** Változók és objektumok összehasonlítására is, sok egyezést is megenged, az eltéréseket egyformán súlyozza ....Kendall tau (3.44-45), Goodman - Kruskal gamma (3.46)

**9a** Változóink bináris típusúak ..... **10**

**9b** A változók nem binárisak ..... **17**

**10a** A közös abszenciák száma befolyásolja az eredményt ..... **11**

**10b** A közös abszenciákat (*d*) egyáltalán nem vesszük figyelembe ..... **16**

**11a** A közös abszenciák éppen olyan fontosak, mint a közös prezenciák ..... **12**

**11b** A közös abszenciák és prezenciák nem egyformán hatnak az eredményre ..... **15**

**12a** Az egyezések és az eltérések súlyozása azonos ..... **13**

**12b** Az egyezések ill. eltérések eltérő fontosságúak ..... **14**

**13a** A függvény metrika .....egyezési index (3.6), euklidészi távolság (3.7), Anderberg I (3.12), PHI (3.15)

**13b** A függvény nem metrika .....Yule I, II (3.16-17), Anderberg II (3.13)

**14a** Az egyezések duplán számítanak .....Sokal - Sneath I (3.11)

**14b** Az eltérések számítanak duplán ..... Rogers -Tanimoto (3.9)

**15a** A közös abszenciák száma (*d*) csökkenti a hasonlóságot .....Russell - Rao (3.23)

**15b** A közös abszenciák köztes hatásúak ..... Baroni-Urbani - Buser I, II és Faith I, II (3.19-22)

**16a** A függvény metrika .....Jaccard (3.24), Ochiai (3.26)

**16b** A függvény nem teljesíti a metrikus feltételeket.....Sorensen (3.25), Kulczynski (3.29), Mounford (3.31)

**17a** Adott konstans hozzáadása az értékekhez nem változtatja meg az eredményt (intervallum skálára csak ezek alkalmasak, de természetesen arányskála esetén is használhatók) ..... **18**

**17b** Adott konstans hozzáadása minden értékhez befolyásolja az eredményt (csak arányskálára jók, intervallum skálára semmiképpen sem ajánlhatók) ..... **21**

**18a** A függvény implicit standardizálást tartalmaz ..... **19**

**18b** Az értékeket nem standardizáljuk ..... **20**

**19a** Standardizálás a sor- és az oszlopösszegek szerint .....$khi^2$ távolság (3.67)

**19b** Standardizálás egységnyi szórásra ..... korreláció (3.70)

**20a** Az értékek közötti különbségek
számítanak ..... euklidészi távolság (3.47), Manhattan-metrika (3.48)

**20b** A minimális egyezések
összegződnek..... Kendall függvény (3.72), Renkonen (3.74)

**21a** A változók közötti arányokra érzékeny mértékszámok ..... **22**

**21b** A változók abszolút mennyiségi eltéréseire érzékeny függvények ..... **24**

**22a** A vektorok közötti
szöggel arányosak ..... húrtávolság (3.54), szögeltérés (3.55), geodéziai mérték (3.56)

**22b** Nincsenek közvetlen kapcsolatban a vektorok közötti szöggel ..... **23**

**23a** Értelmezési tartományuk végtelen ..... keresztszorzat (3.68), kovariancia (3.69)

**23b** A lehetséges értékek 0 és 1 közé esnek ..... hasonlósági hányados (3.71)

**24a** Az objektumpár egyezését (vagy különbözőség esetén az eltérését) először
összegzik, majd az adott párra megadható lehetséges maximumhoz viszonyítják;
értékük 0 és 1 közé esik ..... **25**

**24b** Az egyezést és a lehetséges maximumot az összegzés előtt
viszonyítják egymáshoz ..... Canberra (3.52), Clark (3.57)

**25a** A mindkét objektumban meglevő változók közötti
eltérés nem számít ..... Gleason (3.64), Ellenberg (3.65)

**25b** Az eltérések mindenképpen számítanak
..... Bray - Curtis (3.58), Marczewski - Steinhaus (3.60), Kulczynski (3.62), Pandeya (3.66).

Ha a fenti útmutatás során eljutottál valamelyik függvénycsoporthoz, a továbbiakban finomabb dolgok számítanak. Döntésedhez már a konkrét megoldandó probléma ismerete szükséges, és ekkor a szóba jöhető függvényeket érdemes egy kicsit alaposabban áttanulmányozni, megvizsgálni a viselkedésüket az e kötetben leírt módon, egy számodra értelmes adatsor alapján, s csak azután dönteni. Célszerű egyébként több koefficienst is kipróbálni ugyanarra az adathalmazra, s az eredményeket később összehasonlítani. Ebből minden kezdő adatelemző sokat tanulhat!

**K:** *Ha már választottam a koefficiensek közül, és tudom, hogy változóim intervallum- és arányskálán mozognak, akkor még mindig bizonytalan maradok: milyen standardizáló módszerek alkalmazhatók az adott különbözőség vagy hasonlóság kiszámítása előtt!*

**V:** Igen, jogos az aggodalmad, hiszen – a koefficiens ismeretében – számos adatátalakítási művelet eleve kizárható. Máskor pedig a standardizálás művelete benne van a formulában, mint erre néhány példát már láthattál is. Mindenesetre segítségül szolgálhat az értelmes kombinációkat feltüntető kompatibilitási táblázat, amely utal a megjósolhatatlan eredménnyel járó, értelmetlen vagy nem logikus kombinációkra is (3. 6 táblázat). Az bizonyos, hogy minél speciálisabb célú egy koefficiens, annál kevésbé "viseli el" az adatok átalakítását. Vigyázat, a táblázatbeli + nem jelenti azt, hogy a standardizálás után a metrikus sajátságok is feltétlenül megmaradnak!

**K:** *Mennyire súlyos az a probléma, hogy egy nekem nagyon tetsző koefficiens nem euklidészi?*

**3.6 táblázat**. Egyes távolságfüggvények és standardizálási módszerek kompatibilitása. Jelmagyarázat: + = elfogadható kombináció, N = a standardizálás nem változtatja meg az eredményt, így felesleges, E = kizárható, bármely oknál fogva nem ajánlott (pl. nincs értelme, 0-val történ• osztáshoz vezethet, stb). Számok jelölik azokat a kombinációkat, amelyek külön megjegyzést érdemelnek: (1) húrtávolság, (2) Whittaker-távolság néven ismert, (3) lineáris korreláció, (4) Renkonen index. Ezeket még egy további standardizálással már nem célszerű kombinálni.

| | Változók szerint | | | | | Objektumok szerint | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Terjedelem | Szórás | Összeg | Maximum | Normálás | Terjedelem | Összeg | Maximum | Normálás |
| euklidészi távolság | + | + | + | + | + | + | 2 | + | 1 |
| Manhattan metrika | + | + | + | + | + | + | + | + | + |
| Canberra metrika | E | ? | N | N | N | E | + | + | + |
| Clark | E | E | N | N | N | E | + | + | + |
| Bray-Curtis | + | E | + | + | + | + | + | + | + |
| Marczewski-Steinhaus | + | E | + | + | + | + | + | + | + |
| Kulczynski | + | E | + | + | + | + | + | + | + |
| Pinkham-Pearson | E | E | N | N | N | E | + | + | + |
| Gleason | + | E | + | + | + | + | + | + | + |
| Ellenberg | + | E | + | + | + | + | + | + | + |
| Pandeya | + | E | + | + | + | + | + | + | + |
| kovariancia | + | 3 | + | + | + | + | + | + | + |
| Hasonlósági arányosság | + | E | + | + | + | + | + | + | + |
| Kendall | + | E | + | + | + | + | 4 | + | + |

**V:** Nagyon sokszor kiderülhet, hogy a nem-euklidészi sőt nem-metrikus mértékek olyan távolságokat adnak, amelyek euklidészi térben is érvényesek. Egyesek gyakorlatilag sosem, csak speciálisan "szerkesztett" esetekben sértik meg a feltételeket. Ez a "megsértés" sem mindig jelentékeny, tehát eltekinthetünk a dologtól. Ennek mértékét a főkoordináta-elemzés alkalmazásával lehet megállapítani, mégpedig a negatív sajátértékek száma és nagysága alapján. A későbbiek során erre utalni fogunk (7.4.1 rész).

**K:** *A mintavételnél és az adatátalakításnál is meggyőzőek voltak azok a példáid, amikor kis változtatások alkalmazásával egy sorozatot képeztünk, s ennek tanulmányozásával többet tudtunk meg a vizsgált objektumokról, mintha csak egy kiragadott értéknél maradtunk volna. Jól emlékszem pl. a kvadrátnagyságra, vagy pedig a Clymo-transzformáció paraméterére. Képezhető-e hasonló sorozat (térsor) a hasonlósági függvényekre is?*

**V:** Ne mondd, hogy az eddigiek alapján nem is sejted a választ: persze, hogy képezhető. A Minkowski metrikaosztályról már szóltunk, bár ennek igazándiból csak két lépése érdekes, a Manhattan és az euklidészi metrika; a magasabb hatványok már túlhangsúlyozzák a nagy eltéréseket. Általános sorozatot alkothat a Faith-féle "*intermediate coefficient*" (3.75) is, ha a

következőképpen írjuk fel:

$$INT_{jk} = \sum_{j=1}^{n}\left[\alpha|x_{ij} - x_{ik}|+(1-\alpha)(\max\{x_{ih}\} - \min\{x_{ij}, x_{ik}\})\right] \text{ ahol } 0 \le \alpha \le 1 \qquad (3.118)$$

ekkor $\alpha$ változtatásával egy folytonos átmeneti sor állítható elő a Manhattan-metrika ($\alpha$=1) és a Kendall koefficiens ($\alpha$=0) között. Gondolkodom azon, hogy az euklidészi távolság és a húrtávolság között is lehetne hasonló módon átmeneteket képezni. Ekkor a mennyiségbeli ill. az aránybeli eltérések között "egyensúlyoznánk".

**K:** *Elismerted, hogy még a lényegesebb függvények közül is kimaradhatott néhány. Én például hallottam valahol a Pearson-féle kontingencia-együtthatóról. Ha van még helyed, bemutatnád ezt nekem?*

**V:** A kontingencia-együttható a – Cramér indexhez (3.37) hasonlóan – azt a problémát próbálja megoldani, hogy a $\chi^2$ maximális értéke a mintanagysággal változik:

$$KK = \left(\frac{\chi^2}{f_{..} + \chi^2}\right)^{1/2} \qquad (3.119)$$

Ha feltételezzük, hogy mindkét változó értékei sok kategóriára oszthatók (*p* és *q* nagy), és sok megfigyelés alapján a gyakoriságeloszlás közelít a kétváltozós normális eloszláshoz, akkor *KK* négyzete a két változó közötti korrelációs koefficiens (3.7) négyzetéhez közelít. Ez azonban csak elméletileg érdekes, mert ezek a feltételek igen ritkán teljesülnek (Anderberg 1973), s ezért nem is említettem ezt a lehetőséget.

Ezen kívül van még egy, amely inkább emlékeztet a Cramér indexre, de a minimum helyett a *p*–1 és *q*–1 mértani közepével oszt:

$$CS = \left(\frac{\chi^2 / f_{..}}{[(p-1)(q-1)]^{1/2}}\right)^{1/2} \qquad (3.120)$$

(Csuprov formula, vö. Anderberg 1973). A normálás akkor ad a Cramér indextől jelentősen eltérő eredményt, ha *p* és *q* értéke nagyon különböző.

**K:** *Nem részletezted ugyan, de említetted, hogy a genetikai távolságnál fontos a biológiai interpretálthatóság. Hogy van ez másutt, például az ökológiában?*

**V:** Igen, a genetikai távolság analógiájára ökológiai (vagy akár taxonómiai) távolságról is beszélhetünk. Az alapprobléma mindig az, hogy a geometriailag szemléletes távolságfüggvények mennyire értelmesek ökológiailag is. Gondolj arra, hogy valahol a mérsékelt övben, a tengerpartról elindulunk a part mentén húzódó hegységbe, egészen 2500 m tengerszintfeletti magasságig. A parton egy szegényes, sótűrő fajokból álló flóra van. 2000 m fölött is csak kevés fajból áll a vegetáció, míg a montán növényzet, 1000 m körül, rendkívül fajgazdag. Prezencia/abszencia adatokból számolt euklidészi távolságok alapján így a magashegységi növényzet közelebb van a tengerpartihoz, mint a montánhoz, ami viszont ökológiailag nyilvánvaló képtelenség. A geometriai interpretálhatóság tehát nem minden, emellett ügyelnünk kell arra is, hogy az alkalmazott függvények biológiailag is értelmesek legyenek.

**K:** *Hogyan lehetne a változók eltérő fontosságát is érvényesíteni a komparatív függvények megszerkesztésében?*

**V:** Nyilván a súlyozásra gondolsz, mert ez valóban beépíthető sok formulába. Prezencia/ abszencia típusú ökológiai adatok esetében például kimondhatjuk, hogy a gyakori fajban mutatkozó eltérés lényegesebb információt hordoz, mint a ritka fajra jutó eltérés ("súlyozott különbözőségi index", Podani 1978):

$$WDI_{jk} = \frac{\sum_{i=1}^{n} p_i |x_{ij} - x_{ik}|}{\sum_{i=1}^{n} p_i} \qquad (3.121)$$

A súly, $p_i$, az *i* faj prezenciájának a mintából becsült valószínűsége. A súlyérték persze más is lehet, pl. a faj entrópiája, amely a köztes gyakoriságú fajokat emeli ki (Tóthmérész 1997).