1. Adatok

1.1. Az adat fogalma

Az adat valamely vizsgált *objektum* mért vagy megfigyelt *tulajdonságát* megadó, többnyire numerikus érték. Az *objektum* (object, observation, case, individual, Merkmalsträger) és a *tulajdonság* (variable, descriptor, Merkmal) fogalmakat elvontan értelmezzük, gyakorlatban azok ugyanis legkülönbözőbb alakban jelenhetnek meg. Példákat ad az 1.1 táblázat:

1.1 táblázat. Objektumok és tulajdonságok

objektum	*tulajdonság*
anyagminták	összetétel
oldatok	komponens koncentrációk
spektrumok	csúcsmagasságok
páciensek	lelet eredmények
emberek	testméretek, hajszín stb.
folyók	vízhozamok
országok	népesedési adatok

1.2. Az adatok fajtái

Az adatok az objektumok tulajdonságai jellegének megfelelően *kategorikus* (nem metrikus, kvalitatív, osztályozó) és *metrikus* (kvantitatív) adatok lehetnek. A kategorikus adatok objektumok csoportjainak (kategóriáinak) *megnevezései*, esetleg *kódjai*, vagy a csoportokhoz önkényesen rendelt *rangszámok,* a metrikus adatok pedig *mérések vagy leszámlálások eredményei.*

A kategorikus adatok következőképpen csoportosíthatók:

a) *nevesítő* vagy *nominális* adatok, amelyek az objektumoknak egy olyan minőségi tulajdonságát írják le, amelyek az objektum valamely egyértelmű osztályozását teszik lehetővé (nem, név, foglalkozás, szín, íz stb). Két objektum a nominális tulajdonságban vagy megegyezik, vagy nem, az ilyen adatokra tehát az A = B, avagy A $\neq$ B művelet értelmezett. Ha az objektumok olyan csoportokba oszthatók, amelyek egymás komplementumai (A és nem A), a nominális adat *igen* avagy *nem* (y/n, 1/0) értékű lehet. Ezek a *dichotomikus* vagy

bináris adatok. A nominális adatok esetén a gyakoriság-, és a módusz számolásának lehet értelme.

b) *rendező* vagy *ordinális* adatok, amelyek az objektumoknak olyan minőségi tulajdonságát írják le, amelyek nagyság szerint sorbaállíthatók, rendezhetők. (például iskolai érdemjegy, rendfokozat, betegség foka, vasuti kocsiosztály stb). A rendező adatokra már már az $A \leq B \leq C$ ... reláció is érvényes. Rendező adatok az objektumok minősítő osztályozására használhatók, és belőlük már egyes leíró statisztikai jellemzők is számíthatók (gyakoriság, módusz, medián, kvantilisek, terjedelem stb).

A mérhető vagy metrikus adatok két csoportba oszthatók: *különbség (intervallum) skálán* és *arányos skálán* értelmezett adatok.

A *különbség (intervallum) skálán* értelmezett adatoknak önkényes 0-pontja van, így csak *különbségüknek* van értelme, arányuknak azonban nem. Ilyen mennyiség például az energia, a Celsius fokban mért hőmérséklet. Az értelmezett műveletek: $=$, $\neq$, $<$, $>$ $\leq$, $\geq$ és a $+$ és $-$.

Az *arányos skálán* értelmezett adatoknak, lévén valódi 0-pontjuk, arányuk is értelmes. Ilyen a legtöbb fizikai, kémiai jellemző: hossz, térfogat, anyagmennyiség, abszolut hőmérséklet. Esetükben az $=$, $\neq$, $<$, $>$ $\leq$, $\geq$, $+$ és $-$ műveletek mellett a multiplikativ műveletek is alkalmazhatók. Az adatok ilyen csoportrosítása az *1.2 táblázatban* látható.

1. 2 táblázat. Adatok csoportosítása

Csoport	Fajta	Értelmezhető művelet	Érték
Kategorikus	nevesítő, (nominális)	$A = B$, $A \neq B$	verbális név, kód
	dichotomikus (bináris)		igen / nem, y / n, 1 / 0
	rendező (ordinális)	$A \leq B \leq C$	1,2,3,...I,II.III,... +, ++ ,+++ ,...
Mérhető vagy metrikus	különbség (intervallum) skálán értelmezett	$=$, $\neq$, $<$, $>$ $\leq$, $\geq$ $+$ és $-$.	diszkét vagy folytonos valós számok
	arányos skálán értelmezett	$=$, $\neq$, $<$, $>$ $\leq$, $\geq$ $+$, $-$,* és /	diszkét vagy folytonos valós számok

Í

Technikai okokból figyelembe szokás venni, hogy a numerikus adatok *diszkrétek*-e vagy *folytonosak*.

1.3. Az adatok elrendezése

Adatokat táblázatokba, sorokba és oszlopokba szokás elrendezni. A kapott táblázatot a mátrixszámítás szabályai szerint kezelhető mátrixnak, *adatmátrixnak* tekintjük:

$$\underset{I\ J}{\mathbf{D}} = \left(d_{ij}\right) = \begin{bmatrix} d_{11}...d_{1j}...d_{1J} \\ ... \\ d_{i1}...d_{ij}...d_{iJ} \\ ... \\ d_{I1}...d_{Ij}...d_{IJ} \end{bmatrix} \qquad i = 1,2,...,I; \qquad j = 1,2,...,J \tag{1.1.}$$

A szokás, a konvenció szerint az adatmátrix egy *sora* egy *objektum* (megállapított sorrendben elrendezett) tulajdonságait tartalmazza, következésképpen a mátrix egy *oszlopa* egy *tulajdonság* (különböző objektumoknál megvalósult) értékeiből áll. Egy-egy adatot ezért két index, a sorra jellemző, (szokásosan) *i*-vel jelölt objektumindex, és a tulajdonságra jellemző, (szokásosan) *j*-vel jelölt tulajdonságindex azonosít, az adat jele tehát d_{ij}.

Legyen az *i* objektum index utolsó értékének jele *I,* a *j* tulajdonságindexé *J.* Egy *I* darab objektumra és *J* számú tulajdonságra kiterjedő vizsgálat adatait egy *I* x *J* méretű **D** adatmátrix fogja tartalmazni, amelynek i-edik sora

$$\mathbf{d}^i = \left[d_{i1}\ d_{i2}\ ...\ d_{iJ}\right] \tag{1.2}$$

az *i*-edik *objektumvektor*, az *i*-edik objektum *J* darab leíró adatát tartalmazza. A *j*-edik oszlop pedig

$$\mathbf{d}_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ . \\ . \\ d_{Ij} \end{bmatrix} \tag{1.3}$$

a *j*-edik *tulajdonságvektor*, a *j*-edik tulajdonságnak az összes *I* objektum esetén megfigyelt értékeiből áll.

A gyakorlatban szokásosan *I* > *J*, azaz több objektumot vizsgálunk, mint tulajdonságot. Az adatmátrix tehát általában álló téglalap alakú. Egy-egy adatmátrix-oszlop (tulajdonságvektor) elemeiből az egyváltozós statisztika módszereivel a tulajdonság számos jellemzőjének becsült értéke (az átlag, a tapasztalati szórás, a terjedelem) kiszámítható [1,2]. Egy-egy adatmátrix sor az objektumot jellemzi.

1.4. Léptékváltás (skálázás)

A tulajdonságvektorok elemeit első lépésben olyan számértékkel adják meg, amely ahhoz a mértékegységhez tartozik, amellyel a tulajdonságot mérték. Ennek megfelelően a tulajdonságok (többnyire dimenzióval és mértékegységgel bíró) számértékei egészen eltérő nagyságrendűek lehetnek. Esetenként akár elvi (valószínűségszámitási), akár gyakorlati (számítástechnikai, kémiai) okokból a tulajdonságok eredeti léptéke hátrányos lehet, és léptékváltásra, skálázásra van szükség.

Az adatmátrix elemei skálázhatók

a) *oszloponként* (egy-egy tulajdonság skáláját módosítva, pl. áttérve dimenziómentes egységekre),

b) *soronként* (egy-egy objektum különböző tulajdonságait új egységekre cserálve, pl. áttérve anyagmennyiségekről móltörtekre),

c) egyszerre, *mindkettő szerint,* (kettős skálázás)

d) *elemenként* (globálisan, tekintet nélkül mátrixbeli helyzetükre).

A skálázás a mátrixelemek skálájának *eltolását*, *zsugorítását* vagy egyidejüleg *mindkettőt* jelentheti.

Eltolás konstansnak az adatokhoz való *hozzáadását* (kivonását) jelenti, zsugorítás (nyujtás) konstassal való *szorzást* (osztást).

A leggyakoribb skálázásokat az 1.3 táblázat tartalmazza.

A skálázásokhoz megjegyezhető, hogy ismert matematikai és valószínűségszámítási tételekből következik, hogy

a) az eltolás, így a centrálás az adatok szórását nem változtatja meg,

b) a zsugorított, igy a standardizált változó dimenziómentessé válik,

c) az eltolás és a zsugoritás pozitiv számmal az adatok sorrendjét nem változtatja.

d) a standardizált változó szórása 1,

e) azok a skálázott változók, amelyek összege minden esetben konstans, pl. 0 vagy 1, „zárt” változóvá válnak, amelyek közül egy (vagy több) már nem független a többitől, azokból kiszámítható. Ilyenek pl. a centrált, az oszlopösszeggel zsugorított és a standardizált változók, de ilyen tulajdonsága van az egységnyi hosszra normáltaknak is.

f) a kovariancia definíciójában centrált adatok szerepelnek. Következésképpen egy tetszőleges I x J méretű ($I \geq J$) *centrált adatokat tartalmazó* $\mathbf{D}_c$ adatmátrixból a

$$\mathbf{C} = \frac{1}{I-1}\mathbf{D}_c^T\mathbf{D}_c, \qquad (1.4)$$

a J darab tulajdonság kovariancáit tartalmazó *kovarianciamátrix* számítható, miközben a standardizált, nulla közepű és egységszórású változókat tartalmazó $\mathbf{D}_{st}$-ből a korrelációs mátrix:

$$\mathbf{R} = \frac{1}{I-1}\mathbf{D}_{ST}^{T}\mathbf{D}_{ST} \tag{1.5}$$

kapható.

1.3 táblázat. Skálamódosítások

nem	fajta	az új érték	módosító állandó	az új skála jellege
centrálás (centering)	sorátlaggal (row centering)	$d_{ij}^{(c)} = d_{ij} - \bar{d}_i$	$\bar{d}_i = \sum_{j=1}^{J} d_{ij} / J$	+ és – értékek, 0 sorösszeg
	oszlopátlaggal (column centering)	$d_{ij}^{(c)} = d_{ij} - \bar{d}_j$	$\bar{d}_j = \sum_{i=1}^{I} d_{ij} / I$	+ és – értékek, 0 oszlopösszeg
	teljes átlaggal (global centering)	$d_{ij}^{(c)} = d_{ij} - \bar{d}$	$\bar{d} = \sum_{i=1}^{I}\sum_{j=1}^{J} d_{ij} /(I,J)$	+ és – értékek, 0 elemösszeg, nem 0 sor és oszlopösszeg.
	mindkét átlaggal (double centering)	$d_{ij}^{(c)} = d_{ij} - \bar{d}_j - \bar{d}^i + \bar{d}$		+ és – értékek, 0 sor-és oszlopösszeg
	logaritmizálás után	$\ln d_{ij}^{(c)} = \ln d_{ij} - C_j$	$C_j = \sum_{i=1}^{I} \ln d_{ij} / I$	(log skálán) zérus közép, + és - értékek
zsugorítás / (nyujtás) (expansion/ contraction)	sorszórással (row scaling)	$\frac{d_{ij}}{s_i}$	$s_i = \sqrt{\frac{\sum_{j=1}^{J}\left(d_{ij} - \bar{d}_i\right)^2}{J}}$	az új értékek átlaga $\bar{d}_i / s_i$, szórása 1.
	oszlopszórással (column scaling)	$\frac{d_{ij}}{s_j}$	$s_j = \sqrt{\frac{\sum_{i=1}^{I}\left(d_{ij} - \bar{d}_j\right)^2}{I}}$	az új értékek átlaga $\bar{d}_j / s_j$, szórása 1.
	teljes szórással (global scaling)	$\frac{d_{ij}}{s}$	$s = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(d_{ij} - \bar{d}\right)^2}{I.J}}$	az új értékek átlaga $\bar{d} / s$, szórása 1.
	maximumra (maximum scaling)	$d_{ij} / \max_j$	$\max_j$ a *j*-edik oszlop legnagyobb eleme	az új értékek nem nagyobbak, mint 1.
	egységnyi hosszra	d_{ij} / l_j	$l_j = \sum_{i=1}^{I} d_{ij}^2 = \mathbf{d}_j^T \mathbf{d}_j$	az új értékek négyzetösszege 1.
	oszlopösszeggel	d_{ij} / sum	$sum = \sum_{i=1}^{I} d_{ij} = \mathbf{1}_j^T \mathbf{d}_j$	az új értékek összege 1.
centrálás és zsugorítás	standardizálás (autoscaling)	$\frac{d_{ij}^{(c)}}{s_j} = \frac{d_{ij} - \bar{d}_j}{s_j}$	$s_j = \sqrt{\frac{\sum_{i=1}^{I}\left(d_{ij} - \bar{d}_j\right)^2}{I-1}}$	új értékek várhatóan a -3...+3 tartományban, átlaguk 0.
	terjedelemre (range scaling)	$\frac{d_{ij} - \min_j}{\max_j - \min_j}$	$\min_j$ a *j*-edik oszlop legkisebb eleme	új értékek a 0...1 tartományban

1.5 A léptékváltás hatásai

A skála eltolása vagy léptékváltása természetesen bonyodalmakkal is járhat. Információ veszthető, adatok sorozatának egyes elemei függővé, a többi függvényévé válhatnak. Sokaságok számos statisztikai jellemzője (középérték, szórás, terjedelem) nyilvánvalóan függ a változok számértékétől. Várható tehát, hogy miközben bizonyos többváltozós statisztikai módszerek invariánsak a skálázásra, mások más eredményt adnak más nagyságrendű, vagy függő elemeket is tartalmazó adatvektorok feldolgozása során. Az adatvektorok ilyen természetű hatására az adott statisztikai módszer ismertetésénél felhívjuk a figyelmet.

Jellemző példa lehet a kovariancia. A kovariancia mátrix ismeretesen meghatározza a tulajdonságértékek kovariancia ellipszoidját. Az adatmátrix egy vagy több, alkalmasint minden vektoroszlopának zsugorítással járó léptékváltoztatása tehát közvetve megváltoztatja a kovariancia ellipszoid helyzetét és méreteit.

A léptékváltás végül –miután megváltoztatja a tulajdonság számértékeket– természetesen módosítja az objektumok között lévő „*távolságokat*” (l. 2 fejezet). Ez a tény különösen fontos lesz az alakfelismerés és a faktoranalízis során.

Irodalom

[1] Sachs, L.: Statistische Methoden. Planung und Auswertung. 7. Auflage. Springer, Berlin, 1993.

[2] Sachs, L.: Angewandte Statistik. Anwendung statisstischetischer Methoden. 7. Auflage. Springer, Berlin, 1991.

[3] Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., De Yong, S. P., Levi, J. and Smeyers-Verbeke, J.:Handbook of Chemometrics and Qualimetrics. Elsevier, Amsterdam, 1998.

[4] Frank, I.E., Todeschini, R.: The data analysis handbook. Elsevier, Amsterdam, 1994.

[5] Podani, J.: Bevezetés a többváltozós biológiai adatfeltárás rejtelmeibe. Scientia Kiadó, Budapest, 1997