1

Mintavétel, adattípusok

(Ahol minden elkezdődik, ...és csaknem el is dől)

A biológia objektumai valós, szinte kézzel fogható dolgok: növény- vagy állategyedek, azok társulásai, egyes szervek, sejtek vagy más szerveződési egységek, stb. Ezen objektumok kiválasztását természetesen saját szándékaink, ízlésünk, a vizsgálat célja, a rendelkezésünkre álló pénz, az idő, és hasonló, inkább praktikus, mint tudományos szempontok nagymértékben meghatározzák. Amennyiben nem elégszünk meg puszta köznyelvi leírásukkal, hanem valamilyen szakmailag igényes értékelést is végre akarunk hajtani – vagyis olyasmit amiről e könyv egésze szól –, akkor az objektumok kiválasztását megelőzően még néhány fontos kérdésre választ kell adnunk. Olyanokra, mint például: Vajon eleget tesz-e az objektumok kiválasztási módja a később alkalmazandó módszerek szabta feltételeknek? A megfigyelt és feljegyzett adatok egyáltalán feldolgozhatók-e valamilyen módszerrel? Mikor nem szűkítjük le túlságosan a kutatás későbbi fázisaiban hozandó döntéseink körét? Összhangban van-e az objektumok kiválasztása a vizsgálat későbbi céljával? És így tovább.

Ha tehát a vizsgálat tárgyának leírásával nem tekintjük a munkát befejezettnek, akkor egy folyamatot indítunk el, melynek első lépése döntően befolyásolhatja a többit. Olyannyira, hogy egy rossz kezdet esetleg évek munkáját is tönkreteheti. Egy emlős-biogeográfiával foglalkozó értekezést például azért utasítottak el (Kanadában), mert a jól hangzó végső megállapítások igen gyenge alapon álltak: a szerző következtetései a mintavétel torzításai miatt nem voltak általánosíthatók. Más esetekben a kutató már régen befejezte az adatgyűjtést; vaskos jegyzetfüzetét teljesen teleírta valamilyen számokkal, s csak ezután próbált az adataihoz "illő" statisztikai eljárást keresni. Ez azonban gyakran sikertelen vállalkozás, s ilyenkor derül ki, hogy egészen másféleképpen kellett volna a munkát elkezdeni.

Ezzel már ki is mondtuk, hogy a *mintavételezés* alapjainak ismerete elsőrendű fontosságú. Sok felesleges munkától, elutasító bírálattól kíméljük meg magunkat, ha egy olyan kutatási tervet készítünk, amely jó előre tisztázza a mintavétellel, az adatok rögzítésével kapcsolatos teendőinket. Az alábbiakban ezt kívánjuk elősegíteni.

1.1 Mintavétel: alapfogalmak

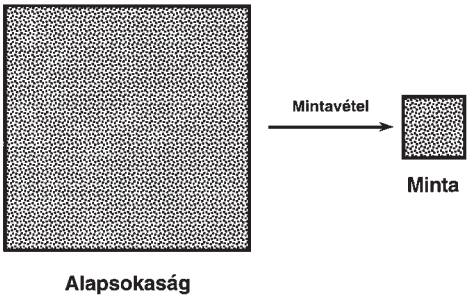
A biológus egyszerű megfigyelések révén is információhoz juthat az őt érdeklő objektumokról: nem vesz fel számszerű adatokat, vizsgálódásának eredményét az agyában összesíti. Elsősorban saját magának, bár esetleg másokkal is közli szóban. A növénycönológus például kimegy a terepre, alaposan bejárja azt, s – előzetes tapasztalataitól, ismereteitől függően – kialakít magában valamilyen képet a látott növénytársulásokról. Nem tagadva az ilyen előzetes tájékozódás fontosságát, ki kell mondanunk: adatok nélkül nincs lehetőség semmilyen elemzésre a későbbiek során.

Mintavételnek csak azt a műveletet tekintjük, melynek folyamán adatokat további feldolgozásra alkalmas formában rögzítünk.

Jogosan merül fel a kérdés: mi is valójában a minta? A vizsgálat során nyerhető összes lehetséges adat az ún. *statisztikai populáció.* Ez a szóhasználat, különösen a biológia területén, sok zavarnak lehet forrása, hiszen a populáció szó már régen foglalt. Ennek egyértelmű jelentése van a genetikában vagy a demográfiában. Nem lenne jó már a kezdet kezdetén kétértelműségekbe bocsátkozni. A lehetséges adatok összességét *alapsokaságnak* fogjuk nevezni (nevezhetjük univerzumhalmaznak is). Ilyen alapsokaságot alkotnak egy erdő összes fájának magassági adatai, egy faj összes példányának testsúly-értékei vagy, bár ez első pillantásra nem nyilvánvaló, egy tó összes halegyedének faji hovatartozásai is, stb. Az alapsokaság nem definiálható; mi magunk s a vizsgálat szempontjai döntik el, hogy mi tartozik bele, mi nem.

Elméletileg lehetséges, a gyakorlatban azonban ritkábban kivihető, hogy az alapsokaság összes értékét meghatározzuk. Ez nem mintavétel, hanem az ún. *teljes felmérés* (leszámlálás, enumeráció). Akiknek erre lehetőségük van, akár ki is hagyhatják az 1.1 részt. A mintavételezés során ugyanis az alapsokaság egy részét ismerjük csak meg, egy *részhalmazt*, azaz a *mintát* emeljük ki abból a mintavételezés folyamán (1.1 ábra). A minta lesz később az alapja a teljes alapsokaságra vonatkozó következtetéseinknek, s ezért is fontos idejekorán tisztázni a mintavétel lehetséges módozatait.

Ha a terepbiológus pH-mérés céljából kiemel öt kémcső vizet a Balatonból, s mindegyiket külön-külön is vízmintának nevezi, hogyan értse a fenti meghatározást? Való igaz, a szakzsargon és a köznyelv között itt ellentét feszül, ami csak úgy oldható fel, ha a fenti elméleti meghatározást határozottan elválasztjuk a mintavétel empirikus oldalától. Elméleti szempontból valóban arról van szó, hogy a kapott öt pH érték az összes, a Balatonban az adott pillanatban elvileg mérhető (voltaképpen végtelen számú, lásd 1.2.2) pH értékből származó egy



1.1 ábra. A mintavételezés egyszerű sémája.

lehetséges mintát alkot. Az ellentmondást feloldhatjuk azzal, hogy az egyes pH értékeket különböző *mintavételi egységekben* (s nem "minták"-ban) mértük meg, azaz egy kémcsövet tekintünk egy mintavételi egységnek. Nincs ilyen probléma ha fák magasságát vagy egy tó halainak faji hovatartozását határozzuk meg, hiszen itt a mintavételi egységek maguk az egyedek, s ezeket senki se nevezné mintának (ezt a témát az 1.2.2 rész fejti ki részletesebben). Elméleti értelemben tehát a mintavétel adatok egy részhalmazának előállítása, technikailag pedig mintavételi egységek kiválasztása vagy elhelyezése.

A minta kiválasztásának módját tekintve még egy nagyon fontos megkülönböztetést kell tennünk. A biológiában általános az olyan "mintavétel", amikor a kutató maga dönti el, hogy egy adott egyed belekerüljön-e a mintába vagy sem. A növénycönológus korábbi tapasztalatait figyelembe véve gyakran dönt például úgy, hogy egyes degradáltnak tűnő, "nem tipikus" részeket kevésbé vesz figyelembe vagy teljesen kihagy a vizsgálatból. Taxonómiai leírások esetében a "jó megtartású", szép példányok kiválasztása általános gyakorlat. Az ilyen típusú adatgyűjtést *preferenciális* mintavételnek nevezzük: a kutató az alapsokaság egyes részeit preferálja (előnyben részesíti) a többivel szemben. Az is előfordulhat, hogy – szándékunk ellenére – az alapsokaság egyes részei mégis kizáródnak a mintavételből, mert valamilyen okból nem férünk hozzájuk (a terület be van kerítve, nincs időnk a mintavételezést mindenre kiterjeszteni, stb). A közös mindezekben az, hogy a minta nem fogja statisztikailag reprezentálni a teljes alapsokaságot, ennélfogva a mintából levont következtetések *nem általánosíthatók az alapsokaságra!*

Mi tehát a *reprezentativitás* feltétele? Hogyan érhetjük el azt, hogy eredményeink és következtetéseink az egész alapsokaságra nézve elfogadhatóak legyenek? A válasz meglehetősen egyszerű: a mintavételezés folyamatában valahol egy véletlenszerű lépésnek kell szerepelnie. Ez biztosítja azt, hogy az alapsokaság minden eleme egyforma eséllyel kerülhessen bele a mintába. Amint a későbbiekben meglátjuk, ennek megvalósítása nem is olyan egyszerű feladat.

1.2 Mintavételezési alternatívák

A továbbiakban mintavételen kizárólag olyan eljárást értünk, amely reprezentatív minta előállítására alkalmas. A mintavételezés folyamatának megtervezése előtt három választási lehetőséget kell figyelembe vennünk (vö. Kenkel et al. 1989).

1.2.1 Becslés vagy mintázatelemzés

Az első választási lehetőség a vizsgálat végső céljára vonatkozik. Sok esetben a mintát azért választjuk ki, hogy az alapsokaság valamilyen *paraméterét* megbecsüljük. Ilyen paraméter egy mérhető tulajdonság (pl. testmagasság, testsúly, egyedszám, stb.) átlaga, helyesebben *várható értéke*, vagy egy növénytársulás faj/egyed *diverzitása* adott függvénnyel kifejezve. A mintavételezést részletesen taglaló szakirodalom szinte teljes egészében ilyen típusú problémákkal foglalkozik, hiszen a torzítatlan becslés az alapfeltétele minden közismert szignifikancia tesztnek, hagyományos biometriai elemzésnek. (A torzítatlanságot úgy határozhatjuk meg, hogy nagyon sok mintából vett becslések átlagai torzítatlanság esetén megegyeznek a keresett paraméter valódi értékével.) További kívánalom, hogy a mintavételi "hiba" is minél kisebb legyen. Ezért a mérések varianciáját is igyekszünk csökkenteni.

Le kell szögeznünk, hogy ebben a könyvben olyan típusú problémákkal foglalkozunk, amelyekben nem a becslés a végső cél, s az legfeljebb az adatgyűjtés első fázisában jelentkezik, ha jelentkezik. A többváltozós módszerek alkalmazásával ugyanis valamilyen biológiai *mintázat*ot igyekszünk feltárni. A mintázat a legtágabb értelemben véve lehet egy osztályozás, egy háttér grádiens, valamilyen folytonos trend, vagy valamilyen térbeli variáció. A mintázat legteljesebb igényű feltárása pedig éppen nem a "hiba" minimalizálására törekszik, hiszen egy homogenizált mintában nem sok feltárnivaló van, hanem ellenkezőleg: a mintavételezést úgy kell megválasztanunk, hogy a minta elemei minél sokfélébbek legyenek (például, a variancia maximalizálása lebeg a szemünk előtt). Ideálisan akkor tudunk meg többet egy faj populációjának a morfológiai mintázatáról, ha a mintában a lehető legteljesebb alaktani változatosságot reprezentáljuk. Egy cönológiai tanulmány is akkor mondja a legtöbbet a vizsgált társulásokról, ha a kapott minta a fajegyüttesek minél több megnyilvánulási formáját lefedi. Mondanunk sem kell, hogy a becslési célú ill. a mintázatelemző vizsgálatok eleve más mintavételezési stratégiát feltételeznek.

1.2.2 Diszkrét vagy folytonos alapsokaság

Az 1.1-ben említett példák (az erdő fáiról és a kémcsővel vett vízmintáról) illusztrálják a következő, rendkívül fontos szembeállítást. A fák egymástól jól elkülönülő, természetesen elhatárolódó egyedek (tudjuk, hogy ez nem mindig van így, de most ez nem lényeges a tárgyalás szempontjából), és mintavételi egységként közvetlenül alkalmazhatók. Az erdőben véges számú fa található (mondjuk N), így az erdőben vehető, legalább egyelemű minták száma 2^N-1 lesz, azaz ugyancsak véges számú. (Ez a szám úgy kapható meg, hogy minden egyes fa vagy bekerül a mintába vagy nem, N fára tehát $2\times2\times2\times \ldots \times2=2^N$ féle lehetőség van, de ezekből egyet, az egy fát sem tartalmazó "üres" mintát kizárjuk. Más kérdés, hogy az egy v. kevéselemű mintáknak sincs sok értelme.) A fák alkotta erdő jó példa tehát a *diszkrét* típusú alapsokaságra. A mintavételezésről szóló könyvek gyakorlatilag erre az esetre szorítkoznak, részletesen taglalva az egyes egyedek kiválasztásának módozatait.

Ha a Balaton vizének pH-ját akarjuk megmérni, nincs olyan természetes mintavételi egység, mint az előző esetben. Az alapsokaság, valójában a Balaton teljes víztömege, egy térbeli *folytonosságot, kontinuumot* képez, s ebből a kontinuumból egy mesterségesen elhatárolt darabot, a kémcsőbe jutó vizet veszünk ki mintavételi egységként. A mintavételi egység nagyságát magunk választjuk meg, de bármekkora is legyen az, végtelen sokféleképpen vehető ki a Balatonból. Következésképpen a megkapható minták száma is végtelen lesz. Hasonló a helyzet amikor egy növénytársulás fajainak borítási viszonyait elemezzük. Általunk önkényesen megadott méretű kvadrátokat kell elhelyezni a társulásban, s egy ilyen mintavételi egység bizony végtelenféleképpen helyezhető el. Ilyen jellegű a vérvétel is: a teljes vérmennyiség jelenti a folytonos alapsokaságot, s ebből veszünk ki néhány cm^3-t, hogy a vér bizonyos tulajdonságait becsülhessük.

1.2.3 Egyváltozós vagy többváltozós esetek

A legegyszerűbb esetben a mintavétel során egyetlen egy jellemzőre, például fák törzsének átmérőjére figyelünk. Ugyancsak egy jellemzőről van szó, ha valamely növényfaj egyedeinek térbeli elhelyezkedését elemezzük különféle kvadrátmódszerekkel. Az egyváltozós mintavétel

problémáit a szakirodalom részletesen tárgyalja, ezért erre nem kell részletesebben kitérnünk. Számunkra a többváltozós mintavétel az érdekes, amikor is minden mintavételi egységben egyidejűleg több jellemzőt (tulajdonságot, változót) figyelünk meg vagy mérünk. Ennek a témának azonban érthetetlen módon kevesebb figyelmet szenteltek eddig.

1.3 A mintavétel főbb jellemzői

A 1.2.1-3 részekben tárgyalt három választási lehetőség összesen nyolcféle kombinációt eredményez. A kötetünk szempontjából, de általában sem egyformán fontosak ezek a kombinációk. A további diszkusszió során a nyolcból csupán kettőre összpontosítunk. Tehát már a mintavételezés kritériumai szerint is behatárolhatjuk könyvünk témáját:

- *Mintázatelemzés, többváltozós eset, diszkrét alapsokaság.* A vizsgálat objektumai természetes egységek, pl. egy populáció egyedei, diszkrét élőhelyek (tavak, szigetek).
- *Mintázatelemzés, többváltozós eset, folytonos alapsokaság.* Az objektumok a mintavételi egységek, a folytonos alapsokaság általunk elhatárolt részei, pl. talaj, víz, levegőminták ("minta" a köznapi értelemben értve), élőlények társulási viszonyainak elemzésére használt mintavételi egységek (amelyek lehetnek pontszerűek, lineárisak, sík- vagy térbeli idomok).

Ezek ismeretében térünk most rá a mintavételezési stratégiák négy fő jellemzőjére. A diszkrét esetben csak az első kettőnek van értelme, a folytonos esetben mind a négyre ügyelnünk kell. Lényeges kérdés tehát az, hogy minden szituációban tudjuk, milyen mintavételi stratégiák jöhetnek egyáltalán számításba.

1.3.1 A minta nagysága

Az elméleti és az empirikus mintanagyság között kell elsősorban különbséget tennünk. Célszerű az *empirikus mintanagysággal* kezdeni. Ez a *mintavételi egységek száma*, melyet jelöljön mondjuk *m*. Miután már tisztáztuk, hogy a vízminta vagy a talajminta csupán egy-egy mintavételi egység, nem téveszthetjük össze a mintanagyságot a mintavételi egység nagyságával (lásd 1.3.3). Az ne zavarjon meg bennünket, hogy az angolszász szakirodalom sokszor következetlenül alkalmazza a mintanagyság ("*sample size*") fogalmát, hol a mintavételi egység nagyságát, hol pedig a mintavételi egységek számát értve alatta.

A mintát részhalmazként határoztuk meg (1.1), ennek nagysága tehát elméleti értelemben a mintában szereplő adatok száma lesz. Mivel vizsgálódásunk több, mondjuk *n* változóra is kiterjed, az *elméleti mintanagyság*, az adatok száma $n \times m$ lesz.

> Milyen szempontok vezéreljenek bennünket az empirikus és az elméleti mintanagyság megadásakor? Ami *m*-et, az empirikus mintanagyságot illeti: akkora legyen, amekkora csak lehet. Minél több mintavételi egységet veszünk be az elemzésbe, annál több információhoz jutunk az alapsokaságról. A rendelkezésünkre álló pénz, idő és egyéb tényezők természetesen korlátozzák *m* nagyságát. Érdemes esetleg a később használandó számítógépes programok maximális kapacitását is figyelembe venni, bár egy nagy mintát utólag akármikor a kívánt méretűre csökkenthetünk. Kicsi mintából viszont később már nem csinálhatunk nagyobbat!
>
> A változók száma (*n)* is legyen olyan sok, amennyit az objektumok leírására értelmesen használhatunk. A nagyon sok változó természetesen redundáns lesz a közöttük fennálló korrelációk miatt, más változókról pedig kiderülhet, hogy nem volt értelme bevonni őket az

elemzésbe. Sohasem tudhatjuk azonban előre, hogy mely változók bizonyulnak majd feleslegesnek! Elkerülendő azonban olyan változók alkalmazása, amelyek egymásnak függvényei. Például, ne szerepeljen egyidejűleg a testmagasság, a testsúly és a kettő hányadosa – csak kettőt tartsunk meg közülük. Hasonlóképpen, ne szerepeljen együtt pl. a levélnyél, a levéllemez és a teljes levél hossza (ugyanis az utóbbi az első kettő összege). Sok esetben a változók száma automatikusan adódik a mintavétel során. Ilyen pl. egy növénycönógiai elemzésben a fajok száma. Valamennyit vegyük figyelembe – a későbbiek során, ha okunk van rá, a változók száma csökkenthető. A változók számát érintő további megjegyzéseket találunk az 1.4.3-7 részben.

Az empirikus mintanagyság és a változók számának arányára is érdemes odafigyelnünk. Ha n jóval nagyobb m-nél, az bizonyosan azt jelenti, hogy a változók erősen korrelálni fognak (l. a 7. fejezetet). Azaz, ha magunk szabjuk meg a változók számát, akkor ennek nem érdemes sokkal túllépnie m-et, tehát *ne erőltessük n* növelését! A fordított esetben, ha $m>>n$, viszont érdemes újabb értelmes változókat keresni (ha nem megy, az se baj). Látjuk tehát, hogy általában elvi megkötés nincs, bár kivételek is akadnak. A diszkriminancia analízis során (7.5 alfejezet) a változók száma semmiképp sem haladhatja meg az objektumok számát, mert (szingularitási problémák miatt) a számítások nem végrehajthatók. Ugyanez a helyzet az általánosított távolság (3.94 egyenlet) esetében is.

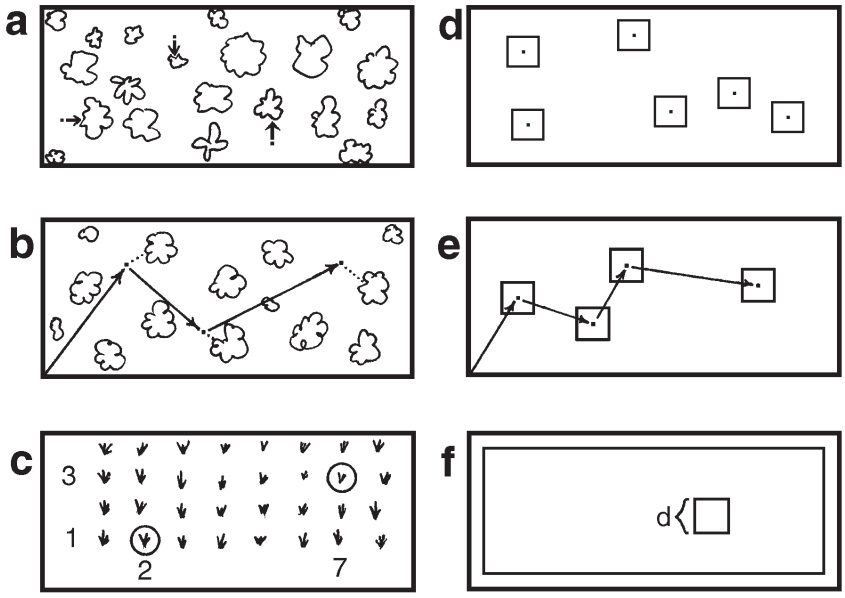
1.3.2 A minta származtatása az alapsokaságból

Diszkrét jellegű alapsokaság esetén az egyedek jelentik a mintavétel egységeit, a minta elemeit tehát *kiválasztjuk* az alapsokaságból. A folytonos esetben viszont a mintavételi egységek *elrendeződéséről* van szó a tér- v. időbeli kontinuumban. Most csak olyan módszereket tárgyalunk, amelyek a kiválasztással ill. elrendezéssel egyértelműen biztosítják a minta reprezentativitását az alapsokaságra nézve.

Az *egyszerű véletlen ("random") mintavétel* során az alapsokaság minden egyede, a kontinuum bármely pontja egyforma eséllyel kerül a mintába. A minta elemeit egymástól függetlenül választjuk ki.

Ezt a feltételt a gyakorlatban nem mindig egyszerű teljesíteni. A diszkrét esetben megtehetjük, hogy az alapsokaság összes elemét megszámozzuk (ez lesz a "mintavételi keret"), majd egy véletlenszám-generátor segítségével választjuk ki a mintát. Ilyen megszámozásra pl. terepi vizsgálatok esetén többnyire nincs lehetőség. Ekkor úgy is biztosíthatjuk a véletlenszerűséget, hogy a vizsgált terület térképén jelölünk ki véletlenszerűen elhelyezett pontokat, ezeket a terepen megkeressük, majd a minden egyes ponthoz legközelebb eső egyedet vesszük bele a mintába (1.2a ábra). Használható a "bolyongásos" módszer, amikor egy adott ponttól elindulva véletlen távolságokra és véletlen irányokban indulunk el (1.2b ábra), így jelölve ki a mintavételi pontokat, s a hozzájuk legközelebb eső egyedeket. Egy kukoricaföldön vagy egy szabályosan elrendeződő ültetvényben máshogy is eljárhatunk: sorokat és oszlopokat választhatunk ki véletlen számok alapján, és az így kapott sor- és oszlopindexek fogják azonosítani a minta elemeit (1.2c ábra).

A folytonos esetben eleve nem lehet szó mintavételi keretről. A mintavétel helyének kijelölésére azonban itt is használhatjuk a térképet ill. a bolyongásos módszert. A mintavételi egységet, pl. egy kvadrátot a növénytársulásban, ezután a véletlenszerűen kijelölt pontok körül helyezzük el (1.2d-e ábra). Egy, csak a folytonos esetre jellemző probléma merülhet itt fel, az úgynevezett *peremhatás*. Azt a kvadrátot, amely a mintavételi terület határával átfedésbe kerülne, azaz egy része "kilógna" a területről, nyilvánvalóan ki kell hagynunk. Ezáltal azonban egy, a kvadrát méretétől függő sávban a mintaterület szélén már nem biztosítjuk az egyenlő esélyt (1.2f ábra). Minél közelebb van egy pont ebben a sávban a terület határához, annál kisebb az esélye, hogy belekerülhessen egy mintavételi egységbe. Igazán egyenlő esélye csak a terület belsejében lévő pontoknak van. Ez a peremhatás annál jelentősebb, minél nagyobb a

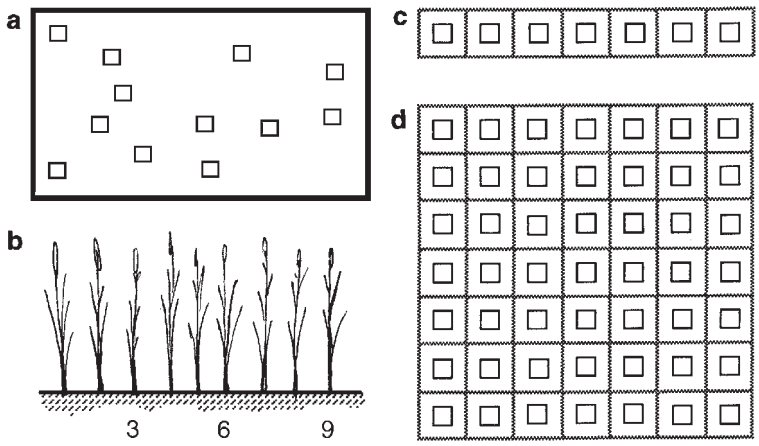


1.2 ábra. Az egyszerű random mintavétel megvalósítása a terepen. **a:** véletlen pont s a legközelebbi egyed, **b:** véletlen pontok s a legközelebbi egyed kiválasztása bolyongásos módszerrel, **c:** random indexek szabályos elrendeződésű alapsokaságból történő kiválasztásra, **d**: véletlen pont módszer kvadrátok kihelyezésére, **e:** kvadrátok elhelyezése a bolyongásos módszerrel. **f:** a peremhatás egy *d* oldalhosszúságú négyzetnél a mintaterület *d*/2 szélességű külső sávjában érvényesül.

mintavételi egység a mintaterülethez képest. Ezt figyelembe kell vennünk az eredmények kiértékelésében. A peremhatás teljes kiküszöbölésére ugyanis a többváltozós mintázatelemzés esetében nincs lehetőség. A peremhatás tipikusan növénycönológiai probléma, egy tóból származó vízminta véletlenszerűségét aligha befolyásolja.

A random mintavétel két vagy több lépcsőben is történhet ha az alapsokaság egységei eleve aggregátumokba tömörülnek. Az aggregátumok véletlenszerű kiválasztása az első lépcső (pl. sejtkolóniák kiválasztása sok közül), majd a második lépcsőben az imént kiválasztott aggregátumokon belül mintavételezünk (pl. sejtek kiválasztása a kolóniákból). Az ilyen mintavételezés valamilyen hierarchiát tételez fel: az alapsokaság kisebb halmazai benne vannak a nagyobb csoportokban, azok a még nagyobbakban, és így tovább. Innen ered a *beágyazásos* ("*nested*") mintavétel elnevezés. Az alárendeltségi viszonyokat tükrözi az ugyancsak gyakori "*subsampling*" kifejezés. A beágyazásos mintavételezés elsősorban becslési problémák esetén alkalmazható, bár többváltozós adatelemzést megelőzően is szóba jöhet (lásd pl. Green 1979, p. 36). Előfordulhat olyan eset is, hogy a változók egy részére, pl. egy növénytársulás fajaira, random mintavételt alkalmazunk kvadrát módszerrel, míg a környezeti változókra (pl. talajreakció, Ca-tartalom, stb.) beágyazásos mintavételt alkalmazunk, az egyes kvadrátokon belül sok ismétléssel.

A fenti eljárás egy változata a *rétegzett véletlen ("stratified random") mintavétel.* Erre akkor lehet szükség ha valamilyen *külső* szempont szerint az alapsokaság részhalmazokra ("rétegekre") osztható. Az egyes rétegekben külön-külön egyszerű random mintavételezést hajtunk végre olymódon, hogy az egyes rétegek arányosan szerepeljenek majd a mintában.



1.3 ábra. A véletlen mintavételezés egyenetlen lehet, nagy területek kimaradhatnak (**a**). A szisztematikus módszer egyenletes elrendeződést biztosít, pl. a diszkrét esetben minden *k*-adik egyed kiválasztásával (**b**). Szisztematikus elrendeződés a folytonos esetben a transzszekt (**c**) és a rács (**d**).

A rétegzett véletlen mintavétel alkalmazásakor figyelmünket a rétegeket elkülönítő kritériumokra és az arányokra kell fordítanunk. A rétegeknek valóban egy *külső szempont* szerint kell elkülönülnie, s nem pedig egy, a mintavételezésben is szereplő változó szerint. A vegetáció-kutatásban például érdemes lehet ilyen rétegeket, mondjuk, a terület mikrotopográfiája vagy talajtani sajátságai alapján elkülöníteni. Nem használható azonban egy adott növényfaj jelenléte vagy hiánya, ha az maga is szerepel a változóink között! Az *arányosság* pedig az egyes rétegekre jutó részminták nagyságával biztosítható. Legegyszerűbb esetben az egyes rétegekben alkalmazandó mintanagyság arányos magának a rétegnek a nagyságával. (Ez a vegetáció-kutatásban pl. területarányosságot jelent.) Más típusú arányosság is elképzelhető, de ezek elsősorban becslési problémák esetében jönnek számításba (pl. a rétegek varianciájával fordítottan arányos részminta-nagyságok).

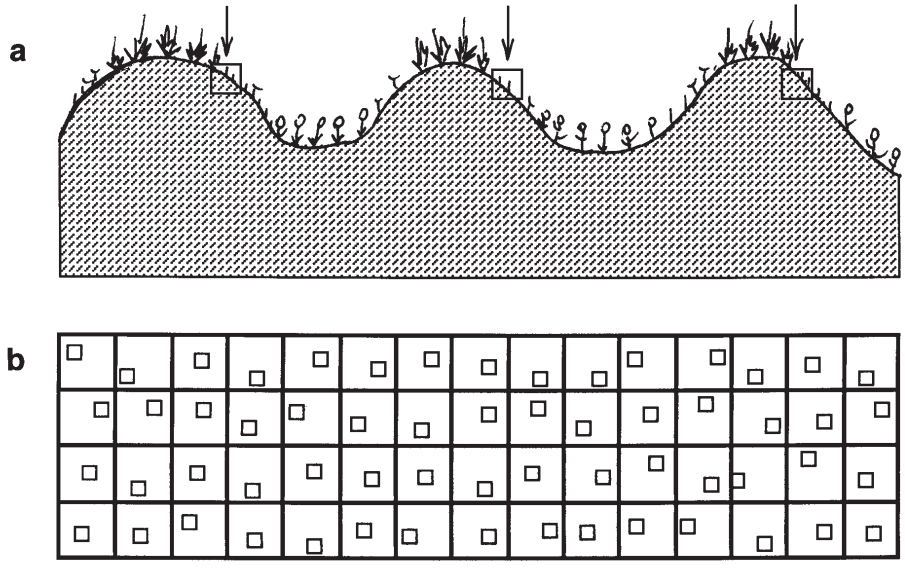
A véletlen mintavétel, bár elméleti szempontból a legjobb stratégia, a gyakorlatban nem mindig hajtható végre. A véletlenszerűség biztosítása eléggé nehézkes, a mintavételi keret kijelölése, a megszámozás pedig sokszor lehetetlen. További gondot jelent az, hogy a minta az alapsokaságot egyenetlenül képviselheti, amit az 1.3a ábra illusztrál. A térben véletlenszerűen elhelyezett kvadrátokra bizony előadódhat olyan eset is, hogy relatíve nagy területek teljesen kimaradnak a felvételezésből! A megoldás a *szabályos (szisztematikus)* módszer. Ekkor csupán egyetlen egy mintavételi egységet, az úgynevezett *kezdőelem*et választjuk ki véletlenszerűen, a többit szabályos szünetek, a *mintavételi intervallumok* kihagyásával kapjuk meg.

A diszkrét alapsokaságban a mintavételi intervallum nagysága egy *k* egész szám. Például egy kukoricaföldön előre eldöntjük, hogy csak minden 3. egyedet veszünk bele a mintába minden 3. sorban. Ekkor a kezdőelemet célszerűen a tábla egyik sarkában kijelölt 3x3 egyed közül választjuk ki teljesen véletlenszerűen. Ezután mindkét irányban minden 3. egyed belekerül a mintába, amíg el nem érjük az alapsokaság határát (1.3b ábra). A minta nagysága tehát az alapsokaság nagyságának és *k*-nak a függvénye. Folytonos esetben a mintavételezés interval-

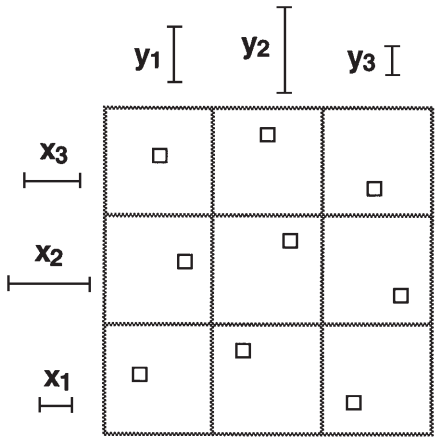
luma ("*spacing*") valamilyen távolság (térben v. időben). A kezdő mintavételi egységet véletlenszerűen kell kijelölni, majd ettől *k* távolságban vesszük fel a többit. Az egységek egy irányban sorakoznak a transzszektek esetében (1.3c ábra), két irányban pedig a rácsos mintavételkor ("gridek", 1.3d ábra). Transzszekteket használunk amikor valamilyen kitüntetett irányban ható tényező, gradiens hatását akarjuk kimutatni (pl. egy nedvesség-gradiens vízparti növényzet esetében). Időben is elképzelhető a transzszekt: a szabályos időközökben végzett fénycsapdázás is ennek tekinthető. A rácsos mintavétel biztosítja egy terület egyenletes lefedését, ezért előszeretettel alkalmazzák a vegetációtérképezésben. Speciális esetekben a mintavételi egységek összeérnek, és csak egy viszonylag kis területet borítanak be. Példaként a növényökológia egyes mintázatelemzési eljárásait említhetjük, melyek fő célja bizonyos strukturális jellemzők skálafüggésének az elemzése (egyváltozós esetre: Greig-Smith 1983, többváltozós esetre: Juhász-Nagy 1976, 1984, 1993). A rács ekkor csupán egy kiindulópont arra, hogy különféle méretű mintavételi egységeket állítsunk elő az alapegységek összevonásával ("*társorelemzés*", lásd 1.5.2).

A szisztematikus mintavételezés egyetlen, kivételes esetben lehet előnytelen: ha az alappopuláció térbeli elrendeződése eleve valamilyen szabályosságot követ és az egybeesik a mintavételi intervallummal. Például tételezzük fel, hogy egy viszonylag szabályosan váltakozó dünesor vegetációját vizsgáljuk transzszekttel, és *k* értéke éppen két düne távolságának felel meg (1.4a ábra). A szisztematikusan elrendezett mintavételi egységek mindegyike ekkor, a kezdőelem helyzetétől függően, azonos helyzetbe (pl. völgybe) kerül, s a kapott minta nem fogja hűen reprezentálni a teljes alapsokaságot, a dünesor növényzetét, hiszen az eltérő lehet a tetőkön és a völgyekben.

A teljesen szabályos elrendeződés esetleges torzító hatása kiküszöbölhető egy kevert stratégiával, a *félig szabályszerű (szemiszisztematikus)* mintavételezéssel. Ekkor az alapsokaságot egyenlő nagyságú blokkokra osztjuk pl. egy rács segítségével, majd minden egyes blokkon



1.4 ábra. A szisztematikus mintavételezés és egy természetes szabályosság esetleges egybeesése (**a**). Szemiszisztematikus stratégia kétdimenziós kontinuumra (**b**).



1.5 ábra. A félig szabályszerű elrendezés egy speciális esete, rögzített koordinátákkal.

belül egy (v. néhány, de blokkonként azonos számú) mintavételi egységet helyezünk el véletlenszerűen (1.4b ábra). A random és a szisztematikus elrendeződés előnyeit így egyesíthetjük.

A fenti stratégia elnevezésében a szakirodalom, mint oly sokszor, nem egységes. Többen (pl. Greig-Smith 1983, Southwood 1984) kifejezetten erre alkalmazzák a rétegzett mintavételezés megnevezést, mások (pl. Orlóci & Kenkel 1985, Green 1979) viszont a 17. oldalon leírtaknak megfelelően. Bár kétségtelenül van hasonlóság a rétegzett és a szemiszisztematikus mintavétel között (mindkét esetben a randomizáció egy részekre bontott alapsokaságban történik), célszerű megtartani az elnevezésbeli különbséget. A rétegzettség inkább tükrözi azt az esetet, amikor az alapsokaságot nem feltétlenül szabályos módon, hanem valamilyen külső tényező alapján osztjuk fel. A szemiszisztematikus elnevezés viszont jobban utal arra a tényre, hogy az alapsokaságot szabályosan, s ezáltal mesterségesen bontjuk részhalmazokra.

A félig szabályszerű elrendezés egy változatában a blokkokon belül nem teljes a randomizáció (Quenouille 1949). A mintavételi egységek blokkon belüli elrendeződését a blokkok egyes soraira és oszlopaira külön-külön megadott, s azokon belül egységesen alkalmazott random koordináták szabják meg (x_1, x_2, x_3 ill. y_1, y_2, és y_3 az 1.5 ábrán).

Smartt & Grainger (1974) azt találta, hogy ez az egészen speciális elrendezés vegetációtípusok arányainak becslésében még jobb eredményt adott, mint az előzőek. A módszer esetleges előnyei a többváltozós elemzésben még ismeretlenek.

1.3.3 A mintavételi egységek mérete

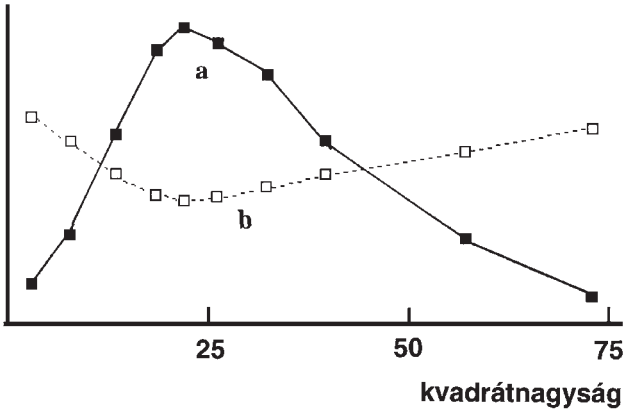
A folytonos alapsokaságban a mintavételi egység térbeli elhatárolása a kutató feladata. Rögtön adódik az első kérdés: vajon mekkora legyen ez az egység? Praktikus szempontok, mint például a könnyű kivitelezhetőség, sok mindent megszabnak. Az alapelvek ismertetésekor az a legfontosabb, hogy különbséget tegyünk az alapsokaság két típusa között. Különböztessük meg a *társulásokat* (élőlények valamilyen elhelyezkedése a térbeli kontinuumban) és *közegtípusú* alapsokaságokat (pl. víz, talaj, levegő).

Társulások esetében természetes követelmény, hogy a mintavételi egység ne legyen túl kicsi az élőlények méretéhez képest. Az is belátható, hogy túl nagy sem lehet, mert a rendelkezésünkre álló idő és pénz rendszerint nem korlátlan. E szempontokat figyelembe véve marad még egy rendszerint széles mérettartomány, melyen belül ránk van bízva a döntés. Ez a döntés

a mintavétel, s ezen keresztül az egész vizsgálat céljától függ elsődlegesen. Az összehasonlítás kedvéért megjegyezzük: *becslés* esetén az általános szabály a "minél kisebb, annál jobb" elve (Elliott 1977). Ha az idő és költségigény megszabja, hogy a kontinuum mekkora része vehető be a mintába (azaz a mintanagyság és a mintavételi nagyság szorzata eleve adott), akkor sok kisméretű egység az előnyösebb a kevés nagy egységgel szemben, mert ekkor kisebb a minta varianciája. (Ez ugyan az alapsokaság mintázatától is függ, de ennek részleteibe nem megyünk bele, hisz nem a becslés a célunk – de, lásd pl. Green [1979, p. 131-133].) *Mintázatelemzés* esetén azonban nem érdekünk a variancia csökkentése. A minta- nagyság és mintavételi egység nagyságának a szorzata itt már nemigen vehető figyelembe, és rendszerint valamilyen előmintavételt kell végeznünk a fő adatgyűjtést megelőzően. Ebben az előmintavételben állapítjuk meg azt a mérettartományt, melyen belül az adatok maximális "információt" adnak az alapsokaságról.

> Hogyan történhet ez az optimalizálás? Erre a kérdésre eleinte a növénycönológia/ökológia igyekezett gyors választ adni, a faj-area görbét, s ennek különböző módosításait alkalmazva. A növényzet osztályozásához – e javaslatok szerint – meg kell vizsgálni a fajszám változását a terület növelésének függvényében. Ahol a fajszám növekedése jelentéktelenné válik meg is állhatunk, mert megkapjuk az "optimális kvadrátnagyságot". Nos, ez a nagyság valóban optimális lehet, de csak arra, amit éppen vizsgálunk, azaz a fajszám, a legegyszerűbb diverzitási jellemző becslésére. A fajszámnak vajmi kevés köze van ahhoz a területnagysághoz, amelynél a lehető legtöbb információt tudjuk megállapítani a társulás szerkezetét, mintázatát illetően. Egyéb, a társulások textúráját leíró paraméterek alkalmazása is eleve kudarcra van ítélve.
>
> A megoldásban Juhász-Nagy (1967-1993) munkáira támaszkodhatunk. Ő kimutatta, hogy a faj/egyed diverzitás helyett a fajkombináció/kvadrát diverzitással és az ezzel rokon mennyiségekkel kell dolgoznunk. Miután azonban ezek az információelméleti mérőszámok viszonylag nagy mintát igényelnek, főleg ha sok faj szerepel a társulásban, egyszerűbb kritérium is alkalmazható. Ez pedig a várható (átlagos) hasonlóság függése a területtől, amely – távolsággal rokon index esetén (lásd 3. fejezet) – ugyanúgy szélső értéket vesz fel, mint az információelméleti mérőszámok (Podani 1984b). Az elmondottakat az 1.6 ábra illusztrálja. A mintavételezés fő stádiumában az a megfelelő méret, ahol e függvények maximumot ill. minimumot értek el. Mindezek azonban csak a bináris (prezencia/abszencia) esetben érvényesek, a "kvantitatív" esetre (pl. egyedszámok, borítás, biomassza, stb.) voltaképpen még nem ismerünk általánosan alkalmazható metódust az optimális kvadrátnagyság megkeresésére. Ez szükségképpen csak úgy kerülhető meg, hogy az előzetes mintavételt, majd a vizsgálat további lépéseit is megismételjük több méretet alkalmazva, s megvizsgáljuk ennek hatását az eredményekre. (Vagy a fő vizsgálatot hajtjuk végre több méretre, de nem vitatható: ez már nagyon költségigényes.) Így kiszűrhetővé válik, hogy a méretbeli változások milyen hatással vannak az eredményekre és következtetéseinkre. Azok, akik erre nem tudnak áldozni, kénytelenek beérni a kézikönyvekben táblázatosan összefoglalt, különféle társulástípusokra ajánlott "leginkább adekvát" méretekkel és mérethatárokkal (pl. Mueller-Dombois & Ellenberg 1974 p. 48, Westhoff & Maarel 1978, Gauch 1982, p. 55, Knapp 1984 p. 111, stb).

A fent leírtak rendszerint csak szesszilis élőlények (növények, bevonattársulások összetevői) esetében érvényesek. Állattársulások zöménél – éppen az egyedek nagy mozgékonysága miatt – speciális mintavételezési eljárásokra van szükség. Ilyen pl. a madármegfigyeléseknél alkalmazott sávmódszer, melyben a mintavételi egység szélessége, hossza, és az adatrögzítés időtartama a legfontosabb paraméterek. Vitathatatlan, hogy állattársulások esetében még inkább a tradíciók és praktikus szempontok döntik el a mintavételi egység nagyságát, hiszen igen nehéz az összhangot megtalálni a méret és az adatelemző eljárások között. Van persze



1.6 ábra. A mintázatelemzésben leginkább alkalmazható méretekről tájékoztat a fajkombináció/kvadrát (=florális) diverzitás (**a**) illetve a várható hasonlóság (**b**) függése a mintavételi egység nagyságától. A függőleges tengelyen felvett mértékegység itt önkényes skálájú, így nem tüntettük fel.

arra is példa, hogy állattársulások, nevezetesen planktonikus rákok esetében a Juhász-Nagy-féle módszerek is használhatók (Dévai et al. 1971).

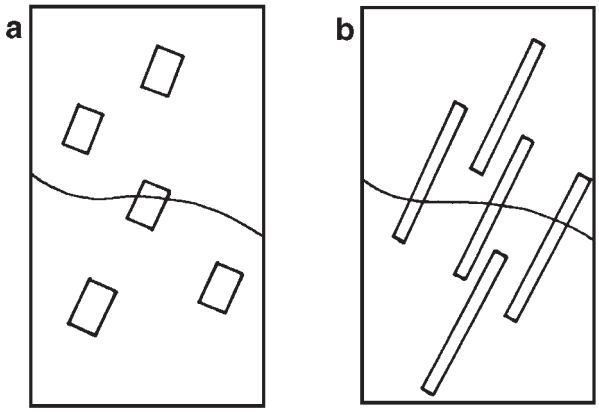
A vizsgálat célja persze nem feltétlenül társulástípusok elkülönítése, osztályozása és leírása. Ez csak a cönológiában van így. Green (1979) számos példát ad arra, hogy többváltozós módszerekkel elemezzük a társulásoknak a környezet leromlásával összefüggő változásait. Ez a monitoring a társulás strukturális megváltozásából von le következtetéseket, és ideálisan ugyanúgy optimális méretet kell alkalmazzon, mint az osztályozás. Ez a méret azonban változhat az idővel, különösen ha a változások erőteljesek. Joggal feltételezhető tehát: nem is létezik kitüntetett méret! Ugyanez mondható el a szukcessziós vizsgálatokról is, melyek során előszeretettel alkalmaznak permanens kvadrátokat. Ha ezen egységek mérete rögzített, akkor nem tudjuk az időbeli és térbeli változásokat elkülöníteni egymástól. Látjuk tehát, hogy – legalábbis elvileg – a többféle területnagyság alkalmazása elkerülhetetlen.

A *közeg-típusú* alapsokaságok témáját rövidebbre foghatjuk. A mintavételi egység mérete itt már inkább technikai kérdés, amely összefügg a rendelkezésre álló analitikai eszközökkel, azok pontosságával és méréshatárával (gondoljunk a pH mérésre, talajfúróra, légszennyezésmérőkre, hasonlókra). Ezen részletek ismertetése azonban nem lehet feladatunk.

1.3.4 A mintavételi egységek alakja

Az alak kiválasztásánál ismét a becslés–mintázatelemzés "ellentét" lebegjen a szemünk előtt. Becslési célból érdemes hosszabb, megnyúlt alakot használni, mert ez csökkenti a varianciát. Társulásokban azonban az ilyen alakú mintavételi egységeknek komoly hátránya van: a térben egymástól távol elhelyezkedő egyedeket tekintünk összetartozónak, s ez félrevezető értékeket ad az interspecifikus asszociációra (pl. Pielou 1977, Greig-Smith 1983) vagy sok faj egyidejű kapcsolatának kifejezésére (pl. Podani 1984a). Továbbá: megnyúlt egységek könnyebben átfednek a társuláson belüli v. azok közötti határvonalakkal, mint az izodiametrikus egységek (pl. négyzet vagy kör, 1.7 ábra). (Itt meg kell jegyezni, hogy a szakirodalom kvadrát néven nem feltétlenül négyzet alakra utal, olvashattunk már "kör alakú kvadrátról" is!) Ezért a többváltozós társulás-elemzésekre csak az izodiametrikus alak ajánlható.

A négyzet és főképpen a kör további előnye, hogy az egységen belüli *szegélyhatás* minimális (ez nem tévesztendő össze az 1.2f ábrán bemutatott peremhatással, célszerű tehát más



1.7 ábra. Négyzet alakú mintavételi egységek kisebb valószínűséggel fednek át az alapsokaságon belüli heterogenitásokkal (**a**). Nyújtott alakú egységek különböző jellegű részeket "mosnak össze"(**b**).

néven nevezni). Hosszú mintavételi egységeknél ugyanis nagyobb valószínűséggel jutnak a növény- (állat-) egyedek az egység szélére. Southwood (1984, p. 36) javaslata szerint e hatás úgy csökkenthető, hogy a mintavételi egység határának csak a felén vesszük figyelembe az egyedeket (pl. a négyzet baloldali és felső oldalán). Ezt a konvenciót mintázatelemzés során is betarthatjuk, ha elfogadjuk a kissé önkényes fele-fele megosztást.

Anizodiametrikus mintavételi egységeknél megemlíthetünk egy ötödik sajátságot is, a *térbeli irányultságot*. Nem mindegy ugyanis, hogy pl. a téglalap alakú egységek véletlenszerűen helyezkednek-e el vagy pedig egy irányba rendeződnek (mint az 1.7b ábrán). Az utóbbi esetben sok múlik azon, hogy ez az irányultság éppen egybeesik-e valamilyen természetes ható tényezővel, grádienssel. Ha tehát ragaszkodunk a téglalap alakú mintavételi egységhez, akkor az irányultságot mindenképpen véletlenszerűnek válasszuk.

1.3.5 A "plotless" mintavételről röviden

Szólnunk kell a növényökológia egyik kedvelt mintavételi eljárásáról, a *"plotless"* mintavételről is. A név azt fejezi ki, hogy itt már nem kétdimenziós egységeket jelölünk ki; a "kvadrátok" vonalra vagy pontra egyszerűsödnek. Vagy a véletlenszerűen elhelyezett pontok vagy vonalak által érintett egyedek faji hovatartozását rögzítjük (rendszerint becslési céllal), vagy a pontokhoz legközelebbi egyed távolságát mérjük meg (egy populáción belüli mintázatelemzési céllal; részletes áttekintést ad Mueller-Dombois & Ellenberg 1974, pp. 93-118, Greig-Smith 1983, pp. 47-53).

Többváltozós analízisre az ilyen típusú mintavételezés ritkán szolgáltat adatokat. Érdekes kivételek a Williams et al. (1969) által említett esetek. Vizsgálataikban azt rögzítették, hogy egy véletlenszerűen kijelölt pontnál milyen fajhoz tartozik a legközelebbi, a második, harmadik, ..., *n*-edik egyed. A ponthoz, mint mintavételi egységhez, tehát egy sorozat fajnév tartozik. Ezután, egy viszonylag kis terület mintázatának elemzésére a pontokat numerikus osztályozással csoportosították *n* különböző értékei mellett.

1.4 Adatok: mérési skálák és más jellemzők

A mintavételi egységeket azért választjuk ki, hogy az őket leíró változókat (tulajdonságokat, bélyegeket) adatok formájában rögzítsük, kódoljuk. Enélkül a mintavételezés voltaképpen

nem is mintavételezés, hiszen minta (adatok részhalmaza) sincs! A mintavételezési stratégiák után tehát tárgyalnunk kell a lehetséges adatformátumokat is. Az adatok típusának megválasztása nagy mértékben megszabja, hogy a későbbiek során milyen módszereket alkalmazhatunk.

Adatokat legtöbbször méréssel vagy leszámlálással kapunk. Megfigyeléseink azonban nem mindig eredményeznek közvetlenül adatokat. A mintavételi egységben megfigyelt tulajdonságokat sok esetben kódolnunk kell. (Más szóval, a tulajdonság lehetséges állapotaihoz egy-egy számot rendelünk, s így egy valószínűségi változót definiálunk.) Az így kapott adatok már számítógépbe vihetők és feldolgozhatók. A módszerek kiválasztását nagymértékben elősegíti az adattípusok ismerete. A köznyelvben, de a szakzsargonban is gyakran emlegetett "kvalitatív, kvantitatív" és "félkvantitatív", vagy pedig a "numerikus" megjelölések azonban pontatlanok és lehetőleg kerülendők. Sokkal egyértelműbb, ha az alábbi tipizálást követjük.

1.4.1 Skálatípusok

A változó lehetséges értékeit négyféle skálán mérhetjük (Anderberg 1973).

1) A *nominális* skálán a változó lehetséges állapotainak a megkülönböztethetősége az egyetlen feltétel. Vagyis, meg tudjuk mondani, hogy két állapot megegyezik-e vagy sem: csak az *azonos* (=) és a *nem azonos* (≠) operátoroknak van értelmük. Nominális változó például a levélalak (ovális, lándzsás, pajzsos, karéjos stb.). Ha az ovális állapotot 1, a lándzsást 2 és a karéjost mondjuk 3 jelöli, akkor nyilvánvalóan az értékek közötti műveleteknek (pl. különbség) nincs értelmük, hiszen a kódolás teljesen önkényes. A taxonómusok ezt a változótípust gyakran többállapotú ("*multistate*") karakternek nevezik. Ez a név kissé megtévesztő lehet, hiszen sok nominális változónak csak két lehetséges értéke van – de ezeket feltétlenül meg kell különböztetnünk más kétállapotú, azaz bináris változóktól (lásd később).

A 0, 1, 2, ..., egész, nem-negatív számokkal kódolt nominális változók közvetlenül feldolgozhatók két hasonlósági függvénnyel (3.103-104), valamint a blokk osztályozás módszerével is (8. fejezet). A módszerek jelentős része azonban nem tud mit kezdeni a nominális változókkal, azokat legfeljebb csak bináris formában tudja figyelembe venni: ha egy nominális változónak p lehetséges értéke van, akkor az behelyettesíthető p számú bináris változóval (Gordon 1981).

> A fenti példában minden egyes levélalak különálló változó lesz, s minden egyes alak hiányát 0, jelenlétét pedig 1 jelölheti. Ekkor azonban ügyelnünk kell arra, hogy az olyan hasonlósági koefficienseket ne használjuk, amelyek a prezenciát és abszenciát szimmetrikusan kezelik (3.2.1 rész), hiszen ekkor az ilyen dichotomizált változók túl nagy súllyal szerepelnek az elemzésben. Vizsgáljuk ezt meg két változóra és 10 egységre az alábbi hipotetikus értékek alapján:
>
> | | Mintavételi egységek |
> |---|---|
> | 1. változó | 1 2 1 3 4 3 2 5 1 2 |
> | 2. változó | 1 1 0 0 0 0 1 1 1 0 |
>
> Az első változó dichotomizálása után már 6 változónk lesz, öt új s egy, az utolsó, eredeti. Értékeink ekkor az alábbi táblázatban összesíthetők:

Mintavételi egységek
1 0 1 0 0 0 0 0 1 0
0 1 0 0 0 0 1 0 0 1
0 0 0 1 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0
- - - - - - - - - - - -
1 1 0 0 0 0 1 1 1 0

Ha most kiszámítjuk a Sokal - Michener féle egyezési koefficienst (3.6) azért, hogy az objektumok (oszlopok) közötti hasonlóságot kifejezzük, akkor azt kapjuk, hogy az első kettő hasonlósága 4/6 míg az első és a harmadik hasonlósága 5/6. Ez a nagyobb érték azért adódik, mert az első és a harmadik objektum egy olyan tulajdonságban egyezett meg, amelyet most öt változó ír le. Ha azonban a hasonlóságot a 3.23 egyenlettel számoljuk, akkor mindkét érték azonos (1/6) lesz, s az "egyensúly helyreáll". Látjuk tehát, hogy az adatok kódolása és a feldolgozó módszerek közötti összhang rendkívül fontos.

2) A következő skálatípusnál a megkülönböztethetőségen kívül a lehetséges értékek még egy logikus sorrendbe is rendezhetők. A < és > operátorok bevezetésével megkapjuk az *ordinális* skálát. Tipikus példa a szilárd anyagok Mohs-féle keménységi skálája. Itt a sorrendiségnél többet nem mondhatunk, különbségnek továbbra sincs értelme. Az első két anyag között (talkum és gipsz) keménységben korántsem biztosan olyan a különbség, mint az utolsó kettő (korund és gyémánt) között. Növénycönológiai példák a széles körben alkalmazott abundancia-dominancia (AD) skálák (Braun-Blanquet 1965, Soó 1964, van der Maarel 1979, Kent & Coker 1992, lásd az 1.1 táblázatot)

Ezt az adattípust nagyon nehéz feldolgozni; rendszerint vagy le kell egyszerűsíteni nominális típusúvá (amikor is információt veszítünk, hiszen a sorrendiség eltűnik), vagy pedig ki kell bővítenünk a következő, intervallum típusúra. Ez a "felbővítés" azonban csak valamilyen további információ figyelembevételével történhet (pl. az AD értékek behelyettesítése százalékokkal egy átlagolásos átszámítás alapján, lásd az 1.1 táblázatot), ami nem mentes az önkényességtől. Sneath & Sokal (1973) javaslata szerint egy p állapotú ordinális változó behelyettesíthető p–1 bináris változóval. Ha egy érték a k-adik a sorban, akkor az első k–1 bináris változó 1-es értéket, a többi 0-t vesz fel. Ha a fenti példa 5-állapotú 1. változóját most ordinálisnak tekintjük, akkor a Sneath & Sokal féle átalakítás után a következő 4 változót kapjuk:

0 1 0 1 1 1 1 1 0 1
0 0 0 1 1 1 0 1 0 0
0 0 0 0 1 0 0 1 0 0
0 0 0 0 0 0 0 1 0 0

Az ilyen átalakítás mindenképpen túlhangsúlyozza a kérdéses tulajdonságot, akármilyen koefficienst is alkalmazunk (ellentétben az előző példával, ahol az index megfelelő megválasztásával elkerültük ezt a veszélyt). További megoldást jelenthetnek a rangsoroláson alapuló együtthatók (3.4 rész).

3) Az *intervallum* skála "komoly előrelépést" jelent az előzőekhez képest. A megkülönböztethetőségen és a sorrendiségen kívül az értékek közötti különbségnek is van értelme. Tipikus példaként a hőmérséklet Celsius vagy Fahrenheit-féle skáláját említhetjük. (A 10 és 20 C^o közötti különbség ugyanakkora, mint a 20 és a 30 C^o közötti.) Azt azonban nem mondhatjuk, hogy a 30 C^o-os hőmérsékletű tárgy "háromszor olyan meleg", mint a 10 C^o-os, mert

1.1 táblázat. Ordinális skálák a növénycönológiából. Megjegyzendő, hogy a + "érték" nem dolgozható fel numerikusan, s ezt be kell helyettesíteni valamilyen kicsiny számmal, pl. 0,1. E skálák még abból az időből származnak, amikor nem állt számítógép rendelkezésre. Ma talán célszerűbb a borítási százalékokat közvetlenül megállapítani.

Érték	Braun-Blanquet	Domin
+	1 %-nál kisebb borítás	Egy egyed, mérhető borítás nélkül
1	1-5% borítás	1-2 egyed, nincs mérhető borítás.
2	6-25 % borítás	Több egyed, 1 %-nál kisebb borítás
3	26-50 % borítás	1-4 % borítás
4	51-75 % borítás	4-10 % borítás
5	76-100 % borítás	11-25 % borítás
6		26-33 % borítás
7		34 - 50 % borítás
8		51-75 % borítás
9		76-90 % borítás
10		91-100 % borítás

a skálának – matematikai értelemben – nincs természetes nullpontja. (A víz fagyáspontja egy teljesen önkényesen, bár célszerűen kiválasztott kezdőpont.)

> Az intervallumskálán kifejezett változók, éppen a különbség értelmezhetősége miatt, már szinte minden módszerrel elemezhetők, mégsem árt az óvatosság. Elsősorban az adatok transzformációjánál kell nagyon figyelnünk. A logaritmikus vagy a négyzetgyök transzformáció például az önkényes nullpont miatt értelmetlen.

4) Az *arányskálán* mért változók minden előző tulajdonsággal rendelkeznek, s a természetes nullpont meglétével már az értékek közötti arányoknak is van értelmük. Azaz, az osztás művelete is alkalmazható. A hőmérséklet mérése K^{o}-ban feloldja a más hőmérsékleti skálákkal kapcsolatos problémákat. De ilyen típusúak számlálással és a hossz-, tömeg-, terület- stb. mérésével kapott változók is. Gyakorlatilag bármilyen adattranszformációnak alávethetők.

1.4.2 Egy kiemelt típus, a bináris változó

A mérési skála mellett az is fontos, hogy a változó hány lehetséges értéket vehet fel. A legtöbb esetben végtelen számú lehetséges érték van, ezekről most többet nem is érdemes mondani. A fentiekben már sokat emlegettük viszont azt a változótípust, a *bináris* változót, amely csak két lehetséges értékkel rendelkezik, függetlenül a skálatípustól. A fajok prezenciája és abszenciája a mintavételi területeken az egyik leggyakoribb példa, de megemlíthetjük a van/nincs típusú taxonómiai karaktereket is.

A bináris változót általában a 0 és 1 értékek kódolják a numerikus feldolgozás előtt. Éppen erre a kódolásra kell ügyelnünk, amikor eldöntjük, hogy a változó melyik állapotát jelölje 0, melyiket 1. A kódolás ugyanis szoros összefüggésben van a később alkalmazandó hasonlósági függvényekkel.

> Voltaképpen teljesen mindegy, hogy mit jelöl a 0 és mit az 1, ha olyan függvényeket választunk, amelyek a két esetet szimmetrikusan kezelik. Ilyenek a 3.112 és 3.115 infor-

mációelméleti függvények, az euklidészi távolság és rokonai, a Sokal - Michener, *PHI*, Yule, Rogers - Tanimoto és Anderberg I-II indexek, azaz azok, amelyek a 2×2-es kontingencia tábla (l. a 3.2 alfejezet elejét) *a* és *d* értékét egyformán kezelik. Más szóval, az eredményeket a kódolás megfordítása egyáltalán nem befolyásolja (3.2.1 rész).

Sok más koefficiens esetében (pl. Sörensen, Jaccard, Baroni-Urbani - Buser I-II, azaz amelyek *a* és *d* értékét nem kezelik egyformán) a kódolás felcserélése rendszerint más eredményre vezet. Ekkor az a logikus, ha azt az állapotot, amely bizonyos értelemben "többet" jelent, mint a másik, 1 jelöli, a másikat pedig 0. Ez minden gond nélkül eldönthető az ordinális, intervallum és arányskálákon. Kétállapotú nominális változóknál azonban a kódolás teljesen önkényes lesz, ezért ilyen változókra ne alkalmazzuk ezeket a függvényeket. Ha bizonytalanok vagyunk egy, általunk nem említett módszert illetően, akkor ajánlatos egy rövid elemzést a kétféle kódolással külön-külön kipróbálni, s megvizsgálni az eredményeket.

1.4.3 Kevert adatok

A többváltozós módszerek jelentős része megköveteli, hogy az összes változó azonos vagy közel azonos típusú legyen (pl. intervallum és arány típusú változók szinte mindig szerepelhetnek együtt). Vannak azonban olyan esetek, elsősorban a taxonómiában, amikor többféle típussal van egyszerre dolgunk. A nominális és intervallum típusú változók, vagy a sokállapotú és a bináris változók együttes jelenléte viszont jelentősen leszűkíti az alkalmazható módszerek körét. A 3.103-104 egyenletek segítségével azonban sok osztályozó és ordinációs módszer kevert adattípusok esetében is alkalmazhatóvá válik. Ha viszont más módszerekhez ragaszkodunk, akkor vagy elhagyjuk a változók egy részét, vagy az Anderberg (1973) által ismertetett skálakonverziós eljárásokhoz folyamodunk. Ez utóbbiak részletezése nélkül megemlítjük, hogy a konverzió egyértelmű az arány → intervallum → ordinális → nominális irányban. Itt minden lépésben információt veszítünk, s magunknak kell eldönteni, hogy ez a veszteség elhanyagolható-e (pl. a növénycönológiában áttérés a fajok borítás értékeiről a prezencia-abszenciára). Fordított irányban viszont mindig szükség van valami külső információra.

A fentiekben a kevert típust úgy értelmeztük, hogy vagy a mérési skálában vagy a felvehető értékek számában van eltérés a mintát jellemző változók között. Ez összhangban van az általános terminológiával, de megjegyzendő, hogy az adatok "keveredése" másképpen is érthető. Az ökológiában például általános, hogy egy mintavételi helyről az ott előforduló fajok jellemzőit (pl. egyedszámát) és az ugyanott mért környezeti változókat is rögzítik. Nyilvánvaló: nem volna értelmes e két változócsoportot egy adathalmazba összevonni s a mintát ennek alapján – mondjuk – osztályozni. Vannak azonban olyan módszerek (pl. kanonikus korreláció elemzés, 7.2 alfejezet), amelyek a logikailag két csoportra osztható változókat külön kezelik, de emellett feltárják a közöttük lévő összefüggéseket is.

Van másféle keveredés is, például ha a változókat többféle mértékegységgel, jóllehet a fenti értelemben azonos típusú – mondjuk arány – skálán mérjük. Erre a problémára az 1.4.6 részben, az összemérhetőséggel kapcsolatosan visszatérünk.

1.4.4 Hiányzó adatok problémája

A többváltozós módszerek megkívánják, hogy az adatok táblázata hiánytalan legyen. Ez azt jelenti, hogy az elemzésbe bevett összes mintavételi egységre az összes változó értékét ismernünk kell. Sok esetben előfordul azonban, hogy néhány érték hiányzik. Rendszertani vizs-

gálatokban egyes egyedek sérültek lehetnek, s különösen igaz ez a paleontológiai leletanyagra. Máskor esetleg nincs mód mindent megfigyelni, és ez a későbbiekben már nem is lehetséges. Mondanunk sem kell, hogy ilyen esetben nem írhatunk be nullát a hiányzó adatok helyére, hiszen azt minden módszer létező értéknek fogja tekinteni.

Az egyik megoldás a 3.103-104 függvények használata, melyek tovább elemezhető távolságmátrixok kiszámítására alkalmasak. Ezen függvények alapján, ha az összehasonlított két objektum bármelyikére hiányzik valamely érték, az illető változó egyszerűen kimarad az elemzésből az adott párosításban. Túl sok ilyen lépés azonban csökkenti az eredmények megbízhatóságát, s a sok hiányzó adattal bíró objektumokat célszerű eleve kihagyni a vizsgálatból. Más lehetőség a hiányzó adatok becslése a meglevők alapján (Beale & Little 1975, Gordon 1981). Az alábbiak közül választhatunk:

1) Megkeressük, hogy ahhoz az objektumhoz (legyen ez *Q*), melyre hiányzó értéket találtunk, az ismert adatok alapján melyik a leghasonlóbb (a 3. fejezetben ismertetett valamely függvény szerint). Ezután a hiányzó értéket egyszerűen eme másik objektum ismert értékével becsüljük.

2) Végezzünk osztályozást valamely módszerrel (4-5. fejezet) az ismert értékek alapján. Ezután megkeressük, hogy *Q* melyik csoportba tartozik, s az e csoportba tartozó objektumok ismert értékeinek átlagával becsüljük a *Q*-nál hiányzó értéket.

3) Az ismert adatok alapján korrelációt (3.70 egyenlet) számítunk a változók között. Kiválasztjuk azt a változót, amely a *Q*-ra nézve nem ismert változóval maximálisan korrelál. Lineáris regressziót végzünk a két változó között, s a kapott egyenlet alapján becsüljük meg a *Q*-ból hiányzó értéket. (E módszer esetleg tovább "nehezíthető" parciális regressziós koefficiensek alkalmazásával).

Mint látjuk, még a legegyszerűbb eljárások is meglehetősen körülményesek, és semmi garancia sincs arra, hogy a hiányzó adat pótlása sikeres volt. Jobb tanács nem adható: ha lehetséges, a hiányzó adatokat mindenképpen kerüljük el.

Vigyáznunk kell arra is, hogy saját magunk se kreáljunk hiányzó adatokat. Ilyenre taxonómiában található könnyen példa, ha az egyes karakterek megléte függ egy másik jelenlététől. Gondoljunk például egy olyan rovarcsoportra, melyben egyes fajoknak van szárnyuk, másoknak pedig nincs. Ha több, a szárnyra utaló karaktert veszünk be az elemzésbe, akkor a szárnyatlanoknál ezen karakterekre természetesen nincs mit megadni; 0-t sem, mert ezen adatok egyszerűen "hiányoznak". Ez csak úgy kerülhető el, ha a szárny jellemzőit egyetlen egy nominális tulajdonság állapotaiként fogjuk fel, ahol 0 jelzi a szárnyatlanságot, 1 a szárny valamilyen tulajdonságkombinációját, és így tovább. Ez a megoldás – be kell ismernünk – nem mindig segít.

1.4.5 Negatív értékek és konstansok

Vannak esetek, amikor a mintavételezés negatív értékeket szolgáltat (pl. hőmérséklet mérése Celsius fokokban). Adatok standardizálása szórással, vagy az egyszerű centrálás (lásd 3.2.1. rész) is negatív számokat eredményez. Erre azért kell ügyelnünk, mert a negatív értékek az elemzés további lépéseiben súlyos problémákat okozhatnak. A számítógépes programok leállnak, ha negatív értékek logaritmusát akarjuk kiszámolni, ha egy objektumra nézve a pozitív és negatív értékek éppen 0 összeget adnak, s ezzel akarunk osztani bizonyos hasonlósági együtthatókban, és így tovább. Adatainkat célszerű tehát úgy átalakítani, hogy ne szere-

peljenek bennük negatív értékek (pl. egy konstans értéket hozzáadunk minden hőmérséklet-adathoz; az analízis eredményét ez nem befolyásolja). A szórással standardizált adatok további elemzésével legyünk nagyon elővigyázatosak: esetükben semmiképp se használjunk az értékek összegével operáló távolság- és hasonlóság-függvényeket.

Nincs értelme olyan változót bevenni az elemzésbe, amely minden mintavételi egységben azonos értéket vesz fel. Ezek a *konstans* vagy *invariáns* karakterek nem befolyásolják az eredményeket; 0 varianciájuknak köszönhetően még számítási problémákat is okozhatnak (pl. a főkomponens-elemzésben).

1.4.6 A változók súlyozása, összemérhetősége

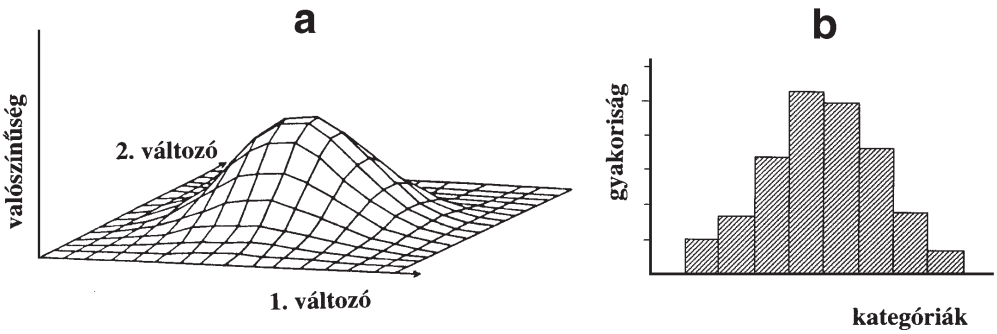
Előfordulhat, hogy véleményünk szerint bizonyos változók fontosabbak, mint mások, s ezt az elemzésben is érvényesíteni szeretnénk. Erre az önkényes lépésre, amit *külső súlyozásnak* nevezhetünk, a legtöbb módszer nem ad közvetlenül lehetőséget. Ha mindenképpen ragaszkodunk hozzá, akkor egy kis trükkel az egész számszoros súlyozást megoldhatjuk. Csupán az a teendő, hogy a kétszeresen (háromszorosan, ...) súlyozni kívánt változót kétszer (háromszor, ...) szerepeltetjük az adatokban (azaz az adatmátrix megfelelő sorát megismételjük)[1]. Hasonló jellegű súlyozásnak számít a nominális és ordinális változók fent említett binarizálása is.

> Míg a többváltozós elemzés általában nem alkalmaz külső súlyozást, a kladisztika területén már más a helyzet. Összhangban a kladisztika céljaival (6. fejezet) a karakterek nem tekinthetők egyformán fontosnak, egyesek sokkal inkább számításba jönnek a leszármazási viszonyok feltárásában, mint mások. Farris (1969), Fitch (1984, p. 238) és Maddison & Maddison (1992, pp. 197-198) tekintik át ezt a vitatott témát.

Az adatok magukban is rejtenek bizonyos *belső súlyozást.* Gondoljunk pl. egy erdőtársulásban felvett borításértékekre, amelyek várhatóan nagyon nagyok lesznek a fafajokra, gyepalkotó füvekre, de kicsik a szálanként növő orchideákra és egyebekre. Ezek az eleve meglévő, esetleg nagyságrendi különbségek a módszerek egy részénél változatlanok maradnak (pl. osztályozás v. ordináció az euklidészi távolságból, 3.47 egyenlet). Ennek következtében az eredményt a fafajok sokkal inkább befolyásolják, mint az orchideák. Az adatelemző módszerek megfelelő kiválasztásával, vagy az adatok előzetes átalakításával (2.3 rész) ez a belső súlyozás kiegyenlíthető (azaz minden faj egyformán fontos lesz), sőt fokozható is.

A belső súlyozástól nem választható el az *összemérhetőség* (Orlóci 1978) problémája. A fenti példát tekintve a fák illetve a szálanként nővő lágyszárúak borítása, akármekkora is az eltérés, összemérhető egymással, hiszen azonos dologról: növények által elfoglalt terület nagyságáról van szó. Egy fizikai-kémiai méréseket tartalmazó adathalmazban azonban sokféle változó szerepelhet, amelyek semmilyen értelemben sem összemérhetők. Ezt a különféle mértékegységek jelenléte okozza: egy ökológiai vizsgálatban pl. a pH értékek – mondjuk – a [4-8] tartományban mozognak, egy fém talajbeli koncentrációja pedig 100 és 200 ppm között. Azaz, egy kismértékű fémtartalomváltozás nagyobb súllyal szerepel az elemzésben, mint a pH maximális megváltozása, ami nyilván nemkívánatos. Ekkor adatainkat standardizálnunk kell (2.3 rész).

1 Ezek között persze 1-es korreláció adódik, és ez is komplikációkat okozhat sok módszernél.



1.8 ábra. A kétdimenziós normális eloszlás sűrűségfüggvényének diagramja (**a**) és egy empírikus sűrűséghisztogram (**b**).

1.4.7 A változók eloszlása

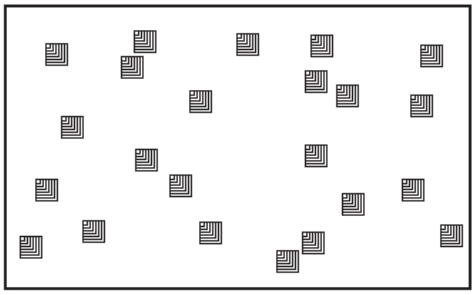
A többváltozós módszerek jelentős részét nem befolyásolja a változók eloszlása (egyszerűen mondva: az, hogy az alapsokaságból származó lehetséges értékek milyen valószínűek). Ide tartozik például az osztályozás (cluster analízis) majd minden módszere (4-5 fejezet), vagy a nem-metrikus többdimenziós skálázás (7.4.2 rész). Egyes hiedelmekkel ellentétben a főkomponens analízis (7.1 alfejezet) sem feltételez semmit a változók eloszlásáról (Chatfield & Collins 1980, p. 58, Rejtő 1986, p. 96), bár nem "hátrány" a normális eloszlás megléte. A diszkriminancia elemzés és a kanonikus korreláció esetében viszont alapfeltétel a *többdimenziós normális eloszlás.* Ezt két változóra az 1.8a ábra segítségével illusztráljuk. Ha nem teljesül e feltétel, attól az elemzés még végrehajtható, a számítógép kiad valamilyen eredményt, de azt rendkívül óvatosan kell kezelni.

> Az idézett módszerek erős non-normalitás esetén is jól értékelhető eredményt adhatnak: a kapott ordinációs diagram sikeresen szemléltetheti az objektumok csoportosulását két dimenzióban, az eredeti sok helyett (ezt úgy nevezzük, hogy a módszerek kellően *robusztusak* a feltételek megsértésével szemben). A szignifikancia próbáknak (7.2.1 és 7.5 részek) vagy a grafikus interpretációt elősegítő ellipsziseknek (9.5.2 rész) viszont már semmiképpen sincs értelmük. Ilyen esetekben mindenképpen meg kell vizsgálnunk az egyes változók eloszlását (pl. sűrűséghisztogramok segítségével, 1.8b ábra), mielőtt elhamarkodottan értékelnénk az ered- ményeket. Azt a változót, amely közelítőleg sem normális eloszlású, ki kell hagynunk vagy transzformálnunk kell (2.3.2 rész). A többváltozós normalitás azonban akkor sem biztos, hogy teljesül, ha az egyes változók külön-külön normális eloszlást követnek (l. Reyment 1991).

1.5 Speciális témák

1.5.1 Térsorelemzés

A mintavételi egység nagyságával kapcsolatosan már rámutattunk arra, hogy a mintavétel során (vagy az előzetes vagy pedig a fő adatgyűjtés alkalmával) többféle méretet célszerű kipróbálni. Annak érdekében, hogy csak a méret legyen a ható tényező, a mintavétel többi jellemzőjét (a mintanagyságot, az elrendezés módját és az alakot) változatlanul kell hagynunk (1.9 ábra). A növekvő kvadrátok sorozatát felhasználva ezután megvizsgálhatjuk a méret hatását magukra az adatokra, a hasonlóság- és távolságértékekre, osztályozásokra és ordinációkra. Más szóval, az eredmények skálafüggése elemezhetővé és értelmezhetővé válik. Az



1.9 ábra. Térsorelemzésre alkalmas mintavételi elrendezés, növekvő méretű, egymásba ágyazott kvadrátokkal.

ilyen mintavétellel egy, az idősorelemzéssel analóg műveletre nyílik lehetőség, amit *térsorelemzésnek* nevezhetünk (régebben "térfolyamat", vö. Podani 1984a, 1992). A növényökológia irodalmát áttanulmányozva megállapítható, hogy a térsorelemzés – kimondva – kimondatlanul – jelen van számos területen, pl. diverzitás becslésekben (Pielou 1975), és alapvető stratégia a populációk mintázatelemzésében (Greig-Smith 1983) és fajkombinációs diverzitás elemzésekben (Juhász-Nagy 1976 1984, Juhász-Nagy & Podani 1983). A térsorelemzés persze nem korlátozódik a terület nagyságának változtatására: a mintavételezés másik három jellemzőjével is végrehajtható, amint az alábbiakban bemutatjuk.

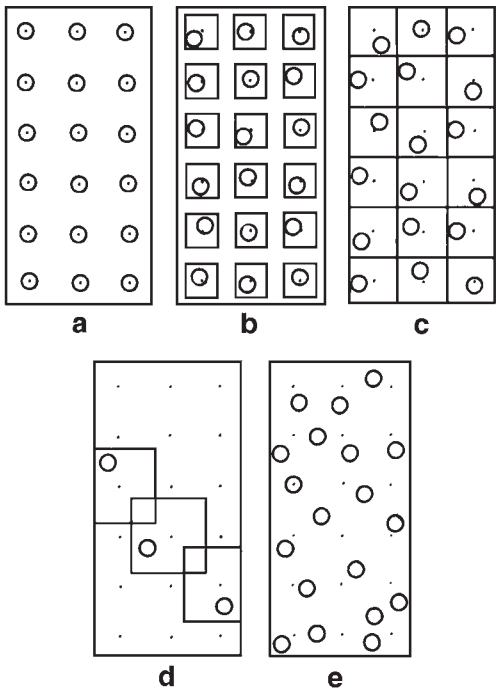
A *mintanagyság* növelése a legegyszerűbb térsor, melyet elsősorban egy kellő pontosságú becsléshez szükséges minta méretének előzetes megállapítására használhatunk. Ez a standard statisztika része, hiszen jól ismert a standard hiba összefüggése a mintanagysággal. Orlóci (1991) és Orlóci & Pillar (1989) ezenfelül javasolja a mintanagyság változtatását távolság- és sajátérték-struktúrák stabilizálására ökológiai vizsgálatokban. A mintanagyság csökkentése a kevésbé fontos változók fokozatos eliminálásával ugyancsak arra alkalmas, hogy többváltozós módszerek eredményeinek stabilitási tulajdonságait elemezzük (pl. Orlóci & Mukkattu 1973, Podani 1989d). Térsort alkothat a mintavételi egységek *elrendezése* folytonos alapsokaságban (Podani 1984a). Kiindulás a szisztematikus elrendezés, amelyből a szemiszisztematikus elrendezés különböző fokozatain keresztül a teljesen randomizált mintavételig jutunk (1.10 ábra). Ez hatékonyan csak számítógépes szimulációval végezhető el (1.5.2 rész). A mintavételi egység *alakjának* a fokozatos megnyújtása, konstans terület mellett, egy újabb lehetőség a térsorelemzésre (Nosek 1976, Podani 1984b, Bartha & Horváth 1987).

A későbbiek során látni fogjuk, hogy a térsorelemzés nem korlátozódik a valós térben definiált változásokra, és szinte mindenféle – az elemzés során szóba jövő – absztrakt térben is "lejátszható". Sőt mi több, végrehajtandó, ha saját, – a többváltozós elemzésben is elkerülhetetetlenül szubjektív – döntéseink hatását elemezni kívánjuk.

1.5.2 Számítógépes mintavételezés

A mintavételi körülmények térsorbeli változtatása rendkívül sok terepmunkát igényel, s erre nincs is minden esetben lehetőség. Ha sokféle kombinációt akarunk kipróbálni, akkor ez már szinte lehetetlen lenne a vizsgált terület alapos tönkretétele nélkül. Megoldást a számítógépes mintavételezés jelent. Palley & O'Regan (1961) és Arvanitis & O'Regan (1967), erdészeti becslésekre vonatkozó korai munkásságát követően Szőcs (1979) dolgozta ki a növénytársulások számítógépes mintavételezésének elvi alapjait.

A vizsgálandó terület növényzetéről fényképezéssel v. más módon ponttérképet kell készíteni. Ez bevihető a számítógép memóriájába. Másik lehetőség: nagyon finom felbontású négyzetrács egyes celláiban prezencia/abszencia adatokat veszünk fel, s ezeket tároljuk a



1.10 ábra. Mintavételi egységek elrendezési sora. **a**: szisztematikus mintavétel, **b**: szemiszisztematikus stratégia össze nem érő blokkokkal, **c**: szemiszisztematikus elrendezés összeérő blokkokkal, **d**: átmeneti állapot átfedő blokkokal (csak hármat mutat az ábra), **e**: teljesen random mintavétel, amikor is a blokkok mérete meghaladja a mintaterület nagyságát (Podani 1984a).

számítógép memóriájában. Megfelelő program segítségével a legkülönfélébb mintavételi stratégiák kipróbálhatók (pl. a **SYN-TAX**: Podani 1993, **MULTI-PATTERN**: Erdei & Tóthmérész 1993). Arvanitis & Reich (1989) programcsomagja elsősorban demonstrációs célokra, s nem konkrét adatok elemzésére való. A téma részletesebb áttekintését Podani (1987) adja meg. Megjegyzendő, hogy valós adatok számítógépes mintavételezése munka- és időigényes tevékenység s csak viszonylag kis alapsokaságra ajánlható.

1.5.3 Mintavételezés a mintából ("bootstrapping")

A név eredete az angol "pull yourself up by your own bootstraps", azaz kb. "segíts magadon" kifejezésben gyökerezik. Az eljárás a statisztikai becslések/hipotézisvizsgálatok témaköréből származik (Efron 1982), s egy ilyen jellegű példával mutatható be legkönnyebben. Először is veszünk egy *n* elemű véletlen mintát az alapsokaságból, s kiszámolunk valamilyen statisztikát (pl. átlag vagy variancia). Ezt a statisztikát nincs mivel összehasonlítani; nos, vegyünk ki nagyon sok véletlenszerű, ugyancsak *n*-elemű mintát a mintából, de már visszatevéssel! (A visszatevés azt implikálja, hogy az eredeti mintát most az alapsokaság olyan reprezentációjának tekintjük, amelyben minden egyes elem előfordulása egyformán valószínű.) Ez a mintavétel célszerűen számítógéppel történik, így a módszer a számítógépes mintavételezés egyik speciális esete. Minden ilyen mintából számoljuk ki ugyanazt a statisztikát, ez lesz az ún. bootstrap becslés. Több száz vagy ezer ilyen becslésből már egy empirikus eloszlás rajzolható fel, amelyben megvizsgálható, hogy az eredeti mintából kapott érték hol helyezkedik el. Ilymódon a statisztika torzítására, standard hibájára, megbízhatósági intervallumára sőt szignifikanciájára is következtethetünk, egyetlen mintából (Manly 1991).

A módszer többváltozós módszerek kiértékelésében, összehasonlításában és az adatok elemzésében is széles körben alkalmazható (pl. korreszpondencia-elemzés esetén Greenacre

1984, Knox 1989, Knox & Peet 1989, főkomponens-analízis ökológiai alkalmazásainál Stauffer et al. 1985). Kladisztikai elemzésekben Felsenstein (1985), Sanderson (1989), Hillis & Bull (1993) és mások a bootstrap technika segítségével azt vizsgálták, hogy mennyire befolyásolja a karakterek kiválasztása az eredményeket.

1.6 Irodalmi áttekintés

Többváltozós elemzéssel foglalkozó könyvek tucatjai nem is törődnek azzal, hogyan jutunk az adatokhoz. Azokat már adottnak veszik, s mintavételezésről sajnos egy szó sem esik. Az ökológusok számára írt ilyen művekre példa Williams (1976), Legendre & Legendre (1983), Pielou (1984), Digby & Kempton (1987).

Más források sokszor nem ismertetik részletesen, csak megemlítik és néhány irodalmi hivatkozással el is intézik az ügyet (pl. Ludvig & Reynolds 1988, Jongman et al. 1987), vagy rövid, velős összefoglalót adnak (Orlóci 1978). Nagyon rossz hatású lehet azonban a terjedelmesebb, de teljesen félrevezető prezentáció, amire több példa is akad, sajnos. Kershaw & Looney (1985) a véletlen elrendezést, a mintanagyságot, a mintavételi egység nagyságát és alakját becslési kontextusban tárgyalják. Ez a kötet populációbiológiai részét illetően úgy ahogy rendben is volna, de már teljesen irreleváns a többváltozós módszerek 65 oldalas leírására nézve. Mit tehetünk vajon olyan kijelentésekkel, hogy "elméleti alapon a legmegfelelőbb kvadrátméret a lehető legkisebb, amely a növényzet típusával ill. az adott méretű kvadrát praktikus voltával összhangban van" (Kershaw & Looney 1985, p. 27)? Greig-Smith (1983), egyébként kitűnő, több kiadást megért könyve is beleesik ebbe a csapdába, holott maga a szerző jegyzi meg a vonatkozó fejezet első sorában, hogy a "kvantitatív adatok értéke ... attól függ, hogy milyen mintavételi módszerrel jutottunk hozzájuk". Ahhoz képest, hogy a könyv 144 oldalt szentel a többváltozós módszereknek, a mintavételezésről szóló fejezet csak a becsléssel ill. a mérés pontosságával kapcsolatos szempontokat ismerteti. Mentségül felhozható, hogy a szerző mindezt tudatosan teszi, megemlítve, hogy [a növényzet] "általános összetételére ill. egy területen belüli variáció elemzésére nem biztos, hogy ugyanaz a legmegfelelőbb mintavételezési módszer". Greig-Smith egyébként az egyik első volt azok között, akik a mintavételezés és az adatelemzés közötti kapcsolat fontosságára rámutattak (Austin & Greig-Smith 1968).

Green (1979) ugyancsak becslési ill. tesztelési célú vizsgálatokra összpontosít, s nem foglalkozik a mintavételezés és a többváltozós módszerek kapcsolatával (pedig ő bőven szól e módszerekről). A tárgyalás folyamán viszont, szerencsére, teljesen nyilvánvaló, hogy mikor, milyen kontextusban értékeli a szerző az egyes mintavételi eljárásokat. Ennek ismeretében sok haszonnal forgathatjuk e könyvet (s némi plusz fáradtsággal, ui. a sajátos felépítésnek köszönhetően a mintavételezés témája eléggé elaprózódik).

Míg a Kershaw & Looney, a Greig-Smith- és a Green-féle kötetek mintavételi fejezeteinek egyoldalúsága legalább részben érthető, ez nem mondható el Gauch (1982) művéről. Gauch nemigen lép túl az általánosságokon, kritika nélkül átveszi az előtte leírtakat, függetlenül attól, hogy azok alkalmazhatók-e egyáltalán a többváltozós elemzésben, a kötet kizárólagos témájában. A 2. fejezet valóságos tárháza a teljesen használhatatlan kijelentéseknek. Ilyen pl. "általában az olyan téglalap, amely 2-4-szer hosszabb, mint amilyen széles, a legpontosabb" vagy "a mintanagyságot az egyes mintavételi egységek pontossága [=accuracy], az eredményektől elvárható pontosság ... befolyásolja". A faj-area görbéket ajánlani optimális kvadrátnagyság meghatározására, mint már utaltunk rá, egyenesen félrevezető.

Sok egyéb, elsősorban növénycönológiai-ökológiai indíttatású könyvet sem lehet megvádolni azzal, hogy a mintavételezés elméletét, többvátozós kontextusban használhatóan

tárgyalná. Knapp (1984), valamint Kent & Coker (1992) semmivel sem lép előbbre a Gauch-féle prezentációnál, holott az utóbbi mű több, mint 120 oldalt szentel a többváltozós módszereknek. Azokat a kézikönyveket pedig, amelyek csak a cönológia relevé módszerét találják egyedül üdvözítőnek, vagyis egy preferenciális jellegű mintavételt ajánlanak, ehelyütt meg sem kell említenünk.

Biostatisztikai szempontból teljesen megbízható kötetek, pl. Sampford (1962) és Cochran (1977), de kizárólag a becslési témában, így könyvünk szempontjából nem jöhetnek számításba. A Cormack et al. (1979) által szerkesztett kötetnek mind a 14 cikke különféle becslési célú mintavételezési módszerekről szól. Southwood (1984) is elsősorban csak azoknak ajánlható, akik populációs paraméterek becslésével foglalkoznak.

Kifejezetten az ökológiai mintavételezés a témája a Frontier (1983) szerkesztette könyvnek. Nagy figyelmet fordít a minta kiválasztásának módozataira, számos példát dolgoz ki, de célja ismét csak a becslés és statisztikai hipotézisvizsgálat. Egy fejezet röviden bemutatja a többváltozós módszereket is; a mintavételezés során a legfőbb kritériumnak a precizitás növelését tartja (azaz megint az adatok becslésénél tartunk). Mindenesetre a kötet sok-sok hasznos információval szolgál, és a mintavételezési technikák olyan részleteire is kitér, melyekre kötetünkben nem juthatott hely. Elsősorban hidrobiológusok forgathatják nagy haszonnal.

A rendszertanban láthatóan sokkal kevesebb figyelmet fordítanak a mintavételezésre. Cole (1969), Dunn & Everitt (1982) és Stuessy (1990) szinte meg sem említi a “minta” szót, ami egyértelműen arra utal, hogy a vizsgálatba bevont egyedek kiválasztása a kutató józan megítélésére van bízva, azaz preferenciális. Sneath & Sokal (1973) viszont már több helyen is foglalkozik a vizsgálatba bevont objektumok, az *OTU*-k[2] kiválasztásával. Számukra a leglényegesebb kérdések a következők: 1. miként befolyásolhatja a mintavétel a taxonómiai hasonlóság mértékét, 2. mennyiben tekinthető egy *OTU* reprezentatívnak az adott taxonra nézve? Központi jelentőségűnek tartják az *exemplar* módszert, amely feltételezi: elegendő minden egyes taxont egy példánnyal szerepeltetni a vizsgálatban, ha a taxonon belüli variabilitás kisebb, mint a taxonok közötti. (Az efajta ördögi körből persze nehezen mászunk ki, ha a kutatás célja éppen az, hogy a még nem ismert taxonokat elkülönítsük egymástól. A módszer viszont sok esetben bevált, amikor már leírt taxonok létét kellett megerősíteni.) Mindenesetre legalább egy tanulmány (Moss 1968) már részletesen foglalkozott azzal a kérdéssel, hogy mennyire befolyásolja a mintavétel az osztályozást. A konklúzió kedvező volt a “lustább” taxonómus számára: nem jelentősen.

Kladisztikai vizsgálatokban, különösen ha molekuláris alapon állanak (6.3-4 rész), korántsem hagyható figyelmen kívül a rendszertani csoporton belüli polimorfizmus kérdése, amely az alkalmazandó mintanagyságot nagymértékben befolyásolja. Ezt emeli ki Baverstock & Moritz (1990), a molekuláris szisztematikában alkalmazható mintavételi stratégiákat összegző áttekintésében. A fenti ördögi körből egy kétlépcsős vizsgálattal juthatunk ki: először a közelrokon taxonokat kell elemezni, majd földrajzilag távoleső populációkat kell minden egyes leszármazási vonalhoz adni. Így megállapítható, hogy a genetikai polimorfizmus vagy a taxonok eltérése-e a nagyobb. Az első esetben nagyobb mintanagyságra lesz szükség (Archie et al. 1989). Ha az elővizsgálat azt jelzi, hogy a variabilitás jelentős része a csoportok között mutatkozik, sokkal kisebb számú ismétléssel, vagy akár az exemplarral is beérhetjük. Baverstock & Moritz munkájával ellentétben más kladisztikai művekben szinte szó sem esik mintavételezésről (pl. Duncan & Stuessy 1984, Forey et al. 1992)

2 OTU="Operational Taxonomic Unit", a taxonómiai vizsgálat alapegysége, egy egyed vagy valamilyen taxon.

Fejezetünk másik fő témáját, az adattípusokat illetően sokkal kedvezőbb a helyzet, mint a mintavétel területén (maga a téma sem olyan "rázós"). A legfontosabb információkat szinte minden, többváltozós módszerekkel foglalkozó könyv összefoglalja. Mindenesetre vigyáznunk kell a terminológiai zűrzavarra a "kvantitatív, kvalitatív, numerikus, metrikus" és hasonló elnevezéseknél. Jobb, ha az 1.4 részben megadott skálatípus csoportosításhoz tartjuk magunkat. Ezekről és a skálák átalakításáról mind a mai napig a legrészletesebb leírást Anderberg (1973, pp. 26-69) adja. Orlóci (1978, pp. 6-13) részletesen vizsgálja az ökológiai változók kiválasztásának módozatait.

Míg az objektumok kiválasztásával a taxonómusok viszonylag keveset törődnek, sokkal jobban ügyelnek az objektumokat leíró bélyegek megfelelő kiválasztására. Sneath & Sokal műve (1973, pp. 90-109, 147-157) továbbra is az egyik legjobb áttekintés (taxonómiai karakterek főbb típusai, karakterek száma, súlyozás). Swofford & Olsen (1990, pp. 414-422) ajánlható a kladisztika speciális adattípusaival ismerkedőknek.

1.7 Kérdezz - válaszolok!

K: *A fentiekből kiderült, hogy nem vagy jó véleménnyel a preferenciális mintavételről, mondjuk a növénycönológia relevé módszeréről. Mit tegyen vajon az a kutató, aki már sok-sok év munkáját áldozta ilyen típusú terepmunkára? Alkalmazhat-e egyáltalán többváltozós módszereket az, aki nem tartja be a meglehetősen szigorúan megfogalmazott alapfeltételeket?*

V: A válasz egyértelműen az, hogy a preferenciális mintavétellel nyert adatok is rendkívül hasznosak a maguk helyén, hiszen ne felejtsük, biológiai tudásunk jelentős része a századok során végül is ilymódon halmozódott fel. Külön szerencse, hogy a többváltozós módszerek exploratív, adatfeltáró és adatösszesítő funkciója legalábbis részben független a mintavételezés körülményeitől. Egy osztályozás osztályozás marad akkor is, ha az objektumokat teljes mértékben a saját ízlésünk szerint válogattuk össze (más kérdés, hogy ez csak a kiválasztott objektumokra lesz érvényes). Teljesen használhatatlan és értelmetlen viszont a hagyományos, becslésekre és hipotézisvizsgálatokra orientált statisztika, ha a mintavételezés preferenciális.

K: *Ha ez így van, akkor mire jó az egész hercehurca ezzel a mintavételezéssel? Miért kellene nekünk annyira ügyelnünk a mintavételezés körülményeire, ha – a ritka szignifikancia tesztektől eltekintve – amúgy is használható a legtöbb többváltozós módszer?*

V: Annak elismerése, hogy biológiai tudásunk jelentős része preferenciális típusú adatnyerésből származik, és az a szerencsés körülmény, hogy a módszerek nem közvetlenül függenek a mintavételezéstől, még nem jelenti azt, hogy továbbra is figyelmen kívül hagyhatjuk ezt a témát. A biológusoknak két fontos kérdésre mindenképpen válaszolniuk kell magukban: 1) összhangban van-e a vizsgálat céljaival a mintavételezés stratégiája, és 2) általánosítani akarják-e következtetéseiket, vagy megelégednek azzal, hogy eredményeik csak a kiválasztott objektumok szűk körére lesznek érvényesek? Aki ezen csak egy kicsit is elgondolkodik, az nem fogja elsietni ezt az első, és nagyon jelentős munkafázist.

K: *A következő válaszod, már sejtem, összefügg az előzővel: összevonhatók-e egy mintába pl. a több személy által, esetleg különböző időpontokban felvett mintavételi egységek? Fontos-e az is, hogy egy mintán belül minden egység egyforma méretű és alakú legyen?*

V: Jól látod, a többváltozós módszereknek ilyen szempontból sincsenek kikötéseik, valamilyen eredmény mindenképpen kijön akkor is, ha a mintát nagyon sok ember, esetleg teljesen eltérő szempontok szerint gyűjtötte. Természetesen az sem kizárt, hogy értelmes eredményt kapjunk, de azt sohasem tudjuk meg, hogy a mintavételezés eltérései mennyiben befolyásolják az eredményeket. Amikor csak lehetséges, a mintavételezés körülményei legyenek egységesek az egész vizsgálatban. Még a növénycönológiai kvadrátok mérete is!

K: *Igen ám, de magad említetted, hogy az "optimális", azaz a mintázatot legnagyobb teljességben tükröző kvadrátnagyság változhat pl. az idővel a szukcesszió vagy degradáció során. Nem lehet ez az optimum különböző egy adott időpillanatban együtt elemzett társulásokra is?*

V: A kérdés jogos: bizony különböző lehet! A kérdést valójában a Poore-nak tulajdonítható, s Orlóci (1991) által feleleventett *szukcesszív approximációval* vizsgálhatjuk, melynek szerves része a térsorelemzés és a többváltozós adatfeldolgozás. A cönológiai kvadrátnagysággal kapcsolatosan ez azt jelenti, hogy a teljes mintára alkalmazott optimális méret csak a társulástípusok elválasztására alkalmas. Ha ez megvan, akkor a típusokon belül külön-külön kell optimumot keresnünk, majd ennek figyelembevételével revideálni az osztályozást mindaddig, amíg stabilis eredményt nem kapunk. A szukcesszív approximáció tehát voltaképpen egy többlépcsős műveletsorozat, melynek során akár a mintavételezés, akár az analitikai eszközök kismértékű módosítása vezet a végeredményre. Be kell ismerni, ez több erőfeszítést igényel a kutatótól, mint egy hagyományos "rajt-cél" vizsgálat.

K: *Sok szó volt a fentiekben a becslésről, de úgy tűnik, mintha az másodlagos lenne a többváltozós elemzés során? Biztosan másodlagos?*

V: Valóban, a becslés művelete a kutatás több állomásán is szerepel, s talán szólhattam volna róla előbb is. Az adatok megállapítása az első becslési tevékenység, gondoljunk pl. a cönológiai borításra. Becslés természetesen minden súly-, hossz-, koncentráció- stb. mérés is, az eszköztől függő pontossággal. A becslési célú vizsgálatot úgy értettem, hogy az a becsült adatok alapján az alapsokaság valamilyen paraméterét – mondjuk az adatok átlagolásával – megbecsüli s ezzel le is zárja az egészet, legfeljebb a paraméterek alapján valamilyen hipotézisvizsgálatot hajt még végre. A mintázatot feltáró többváltozós elemzések viszont csak most kezdődnek. Igaz ugyanakkor, hogy az adatokból számított hasonlóságok és távolságok is becslésnek számítanak. Sőt, a kapott ordinációk vagy osztályozások – a szó legeslegtágabb értelmében – akár maguk is "becslés"-nek tekinthetők. Hiszen a teljes vizsgálat is megismételhető, amely egy másik "becslést" adna a keresett osztályozásra vagy ordinációra.

K: *Bizonyos, hogy minden kutatási terv az adatok → hasonlóság v. távolság → osztályozás v. ordináció → eredmények értékelése sorrendet követi?*

V: Korántsem; de a könyvünk elsősorban olyan problémákra koncentrál, melyek ezzel a módszertani sorrenddel jellemezhetők. Nagyjából ez a főtengelye a 0.1 ábra sémájának is. Egyes lépéseket persze átugorhatunk, mint például a molekuláris szisztematikában, amikor adatok helyett közvetlenül távolságokat állapítunk meg (pl. DNS párosítási kísérletek alapján). Olyan eset is elképzelhető, amikor megfigyeléseink valamilyen egyszerű osztályozást v. ordinációt adnak, s ekkor az eredmények értékelése jelenti majd a számítógépes feldolgozás egyetlen állomását. A fenti sorrendtől leginkább eltérő a karakter alapon működő kladisztika (6. fejezet) stratégiája.