Hiányzó adatok és kezelésük a statisztikai elemzésekben

Oravecz Beatrix,

a Budapesti Corvinus Egyetem tanársegédje

E-mail: beatrix.oravecz@uni-corvinus.hu

Adathiánnyal szinte minden adatbázis esetén találkozunk. A hiányzó adatokat valamilyen módon kezelni kell, nem hagyhatjuk ki őket egyszerűen a mintából, mert a sokasági paraméterbecslések torzítottak lehetnek, hacsak az adathiány nem teljesen véletlenszerű. A hiányzó adatok kezelésének célja éppen ennek a torzításnak az eltüntetése. Ezt a célt a különböző módszerek annak függvényében érik el, hogy mennyire helyesen sikerül azonosítani és modellezni az adathiány sajátosságait. Ebben a tanulmányban áttekintjük a hiányzó adatok típusait és a kezelésük lehetséges módjait, kiemelve az egyes módszerek előnyeit, hátrányait és alkalmazásuk feltételeit. A hiányzó adatok kezelésére nem létezik egyetemesen legjobb megoldás. Lényeges szempont, hogy a választott eljárás összhangban legyen a később elvégzendő elemzésekkel, és az olyan adatbázisok esetében, ahol a hiányzó adatokat valamilyen módon pótolták, a felhasználók is láthassák az adatpótláshoz használt módszert.

TÁRGYSZÓ:
Statisztikai mintavétel.
Statisztikai módszer.
Statisztikai elemzés.

A hiányzó adatok sok kutatásnál okoznak problémát, mert a minta véletlenszerűségét rombolhatják le, pedig a legtöbb statisztikai módszer és következtetés alapja a véletlen minta. Ebben a tanulmányban röviden áttekintjük a hiányzó adatok típusait és a kezelésükre használt legelterjedtebb módszereket, kiemelve fő előnyeiket és hátrányaikat.

Egy általános adatmátrix sorai tartalmazzák a megfigyelési egységeket, vagy eseteket, az oszlopok pedig a változókat, amelyek értékét minden egység esetén ismerjük. Az adatmátrixban lévő adatok általában valós számok, amelyek vagy egy mennyiségi ismérv tényleges értékeit fejezik ki (például az életkor vagy a jövedelem), vagy egy minőségi ismérv kategóriáit reprezentálják (például az iskolai végzettség vagy a nem). A gyakorlatban azonban az a jellemző, hogy ez az adatmátrix nem teljes, bizonyos értékek hiányoznak.

Például egy háztartási bevételeket és kiadásokat vizsgáló kutatás során a megkérdezettek megtagadhatják a jövedelemre vonatkozó kérdés megválaszolását, vagy egy fogyasztói preferenciákat vizsgáló kutatás során előfordulhat, hogy a válaszadó nem tud választani két termék közül, egyiket sem preferálja a másikkal szemben. Az első esetben a jövedelem értékét tekinthetjük hiányzónak, hiszen van mögötte egy tényleges érték, csak mi nem ismerjük. A második esetben azonban nem tekinthetjük a termékpreferenciát hiányzónak, mert nincs mögötte valós érték, a válaszadó nem megtagadta a választ, hanem nem tudott válaszolni. Ebben az esetben a „nincs preferencia" vagy „nem tudom" is egy válaszadói réteget jelöl. A legtöbb statisztikai szoftver tartalmaz egy vagy több speciális kódot az adathiány bevitelére. Egynél több kód lehetővé teszi a különböző jellegű adathiányok beazonosítását, mint „nem tudja", „válaszmegtagadás", „értelmetlen adat". Ez utóbbi esetben van ugyan adatunk, tehát látszólag nincs adathiány, de tudjuk, hogy az nem megbízható, vélhetően hibás, így azt valójában nem használhatjuk az elemzésekben, hanem a hiányzó adatokhoz hasonlóan kell kezelnünk[1]. Felmerülhet a kérdés, miért kell egyáltalán a hiányzó adatokkal foglalkozni, ahelyett, hogy egyszerűen törölnénk őket a mintából. Válaszként álljon itt a következő példa.

1992. április 9-én a Konzervatív Párt megnyerte a brit választásokat, ami óriási bukást jelentett a közvélemény-kutatási iparágnak. A választások napján a négy legnagyobb közvélemény-kutató cég a Munkáspárt 0,9 százalék pontos győzelmét várta.

[1] Az outlierek esetében is van adatunk, de azt nem célszerű a többihez hasonló módon használni. Az outlierek meghatározásáról és kezeléséről olvashatunk például *Csereháti* [2004] cikkében. Ez utóbbi típusú „adathiányok" kezelésében az adatellenőrzésnek és korrekciónak nagy szerepe van, de ezekkel ebben a tanulmányban nem foglalkozunk.

Ezzel szemben a Konzervatív Párt győzött 7,6 százalékponttal. Ez 8,5 százalékpontos hiba, ami igen nagy. Egy utólagos vizsgálat megállapította, hogy a hiba fő oka az volt, hogy a kutatás során nem foglalkoztak a válaszmegtagadásokkal és a „még nem tudom" típusú válaszokkal, hanem egyszerűen törölték őket a mintából. Ez a gyakorlat végzetes volt az eredmények szempontjából, mert az utólagos kutatás megmutatta, hogy a konzervatív pártiak kevésbé tárták fel választási szándékukat. (Hasonló volt a helyzet a magyarországi 2002-es választások során is.)

Látható tehát, hogy a hiányos adatbázisokból való következtetések torz képet adhatnak. Törekedni kell tehát az adathiány természetének megismerésére, majd ezen információk figyelembevételével a hiányzó adatok valamilyen kezelésére.

A tanulmányban először áttekintjük az adathiányok jellemző mintázatait, majd megvizsgáljuk az ún. adathiány-mechanizmusokat, végül sorra vesszük azokat a lehetséges eljárásokat, amelyek adathiányos helyzetekben alkalmazhatók.

A hiányzó adatok kezelésével foglalkozott korábban a *Statisztikai Szemlében Máder Miklós Péter* [2005] „Az imputálási eljárások hatékonysága" című cikke. Ez a korábbi cikk nem foglalkozott az adathiány mintázatokkal, ezért ezeket ebben a tanulmányban ismertetjük. Máder cikke néhány eljárás hatékonyságát vizsgálta modellezéssel. Ez a tanulmány nem tartalmaz empirikus vizsgálatot, hanem az ott alkalmazott és egyéb alkalmazható módszerek elméleti hátterét és tulajdonságaikat tekinti át.

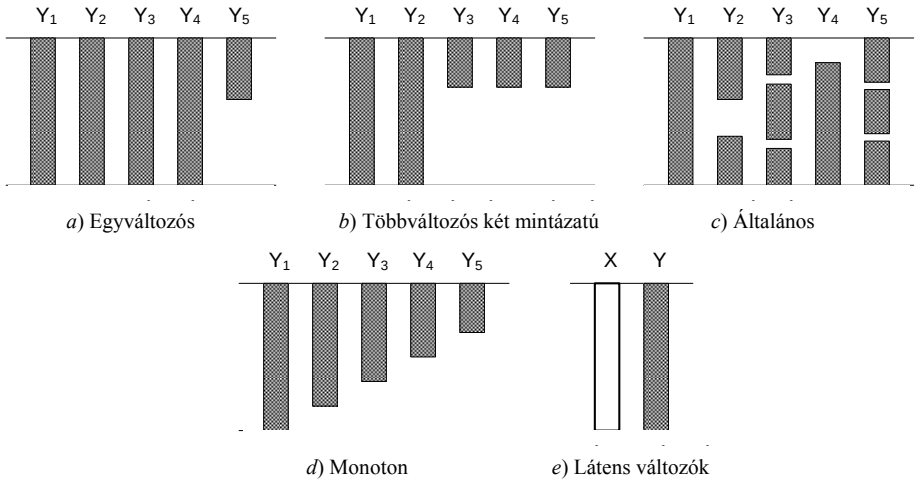
1. Hiányzó adatok típusai

A következőkben áttekintjük az adathiány típusait. A csoportosítás egyik szempontja az *adathiány mintázata*. A mintázat azt írja le, hogy mely adatok a megfigyeltek és mely adatok hiányoznak az adatmátrixban. A másik csoportosítási szempont az *adathiány-mechanizmus*, amely a hiányzás és az adatbázisban szereplő változók értékei közötti kapcsolatot veszi figyelembe.

1.1. Adathiány mintázat

Legyen $\mathbf{Y} = (y_{ij})$ egy $(n \times K)$ általános adatmátrix, hiányzó adatok nélkül, amelynek i-dik sora $y_i = (y_{i1}, \ldots, y_{iK})$, ahol y_{ij} az Y_j változó értéke az i-dik egységnél. Hiányzó adatok esetén legyen $\mathbf{M} = (m_{ij})$ az adathiány indikátor mátrix (*Little–Rubin* [2002]), ahol $m_{ij} = 1$, ha y_{ij} hiányzik és $m_{ij} = 0$, ha y_{ij} megfigyelt. Az $\mathbf{M}$ mátrix definiálja az adathiány mintázatot. Az 1. ábra mutat néhány példát az adathiány-mintázatokra. (A megfigyelt y-ok ($m = 0$) sötéttel jelölve.)

1. ábra. Adathiány mintázatok



a) Egyváltozós *b*) Többváltozós két mintázatú *c*) Általános

d) Monoton *e*) Látens változók

Az *egyváltozós adathiány* az 1. ábra *a*) esete, amikor csak egyetlen változóban van adathiány, a többi változó teljesen megfigyelt. Ilyen mintázata lehet például a mezőgazdasági kontrollált kísérletek eredményének, ahol azt vizsgálhatják, hogy milyen a kapcsolat az Y_K eredményváltozó (terméshozam) és az $Y_1,...,Y_{K-1}$ magyarázóváltozók (öntözővíz, hőmérséklet, műtrágya típusa, mennyisége) között. A magyarázóváltozók ekkor teljesen megfigyeltek, nincs hiányzó adat, a függő változóban viszont előfordulhat adathiány (például hibás vetőmag vagy rossz adatrögzítés miatt).

A többváltozós kétmintázatú adathiány egy másik általános mintázat, amikor az előző példában szereplő egyetlen adathiányos változó *(*Y_K) helyett több adathiányos változónk van ($Y_{j+1},...,Y_K$) , ahol mindegyik egyformán megfigyelt, vagy hiányzik ugyanazokra az esetekre. (Lásd az 1. ábra *b*) esetét, ahol $K = 5$ és $J = 2$.)

Erre a mintázatra lehet példa a kérdőíves felméréseknél az egységszintű nemválaszolás. (Amennyiben az adathalmazból egy-egy elem teljesen hiányzik teljes (vagy egységszintű) nemválaszolásról (unit nonresponse) beszélünk.) Ez az egységszintű nemválaszolás előfordulhat azért, mert a kiküldött kérdőívet meg sem kapta a címzett, vagy megkapta, de megtagadta a válaszadást. Ekkor a kérdőívben szereplő változók lesznek az adathiányos változók. A teljes, adathiányt nem tartalmazó változók a minta tervezéséhez használt változók lesznek, amelyek mind a válaszolók, mind a nemválaszolók esetében előzetesen ismertek egy listáról (például név→nem, lakcím).

Általános adathiány-mintázat úgy alakul ki, ha csak bizonyos kérdésekre adott válaszok hiányoznak, ekkor részleges (vagy tétel szintű) nemválaszolásról (item nonresponse) beszélünk. Ebben az esetben az adathiány mintázata általában semmiféle specialitással nem rendelkezik. (Lásd 1. ábra *c*) esetét.)

Monoton adathiány következik be például, ha a longitudinális felmérések időről időre gyűjtenek be adatokat ugyanazon megfigyelési egységekről. Ezekben a felmérésekben gyakori jelenség a lemorzsolódás, ami azt jelenti, hogy a megfigyelési egység kiesik a mintából, még a kutatás befejezése előtt. Például háztartás-panel esetén a család külföldre költözik, vagy klinikai kísérleteknél más gyógyszerek hatása, vagy egyéb betegség miatt a beteg nem tud tovább részt venni a kísérletekben. A lemorzsolódás egy példája a monoton mintázatú adathiányoknak. (Lásd 1. ábra *d*) esetét.) Ekkor a változókat lehet úgy sorba rendezni, hogy minden $Y_{j+1}, ..., Y_K$ hiányzik, ha Y_j hiányzik. Vannak olyan módszerek, amelyek csak az ilyen mintázatú adathiányt tudják kezelni. Az ilyen mintázat a gyakorlatban ritkán fordul elő, közel monoton mintázat azonban már gyakrabban.

A nem megfigyelhető *látens változókat* is felfoghatjuk adathiány problémaként, csak ezeknél a látens változóknál speciálisan minden megfigyelési érték hiányzik. Az 1. ábra *e*) esetében az *X* jelenti a látens változók csoportját, ahol minden érték hiányzik és *Y* pedig a teljesen megfigyelt változók csoportját. Ekkor természetesen bármiféle elemzéshez különböző feltételezésekkel kell élnünk. Látens változó lehet például a klinikai kísérleteknél a beteg gyógyulásba vetett hite, ha erre vonatkozóan nem szerepelnek adatok a mintában.

1.2. Adathiány-mechanizmus

A hiányzó adatok kezelésének legalkalmasabb módját akkor tudjuk megtalálni, ha ismerjük, hogy miként lettek hiányzóak. *Little* és *Rubin* [1987] az adathiány három alapvető esetét különbözteti meg, attól függően, hogy milyen a kapcsolat a hiányzás és az adatbázisban levő változók értékei között. Ezeket ők *adathiány-mechanizmusnak* nevezték el.

Intuitíve és formálisan is megadjuk az egyes csoportok definícióját. Legyen továbbra is az $\mathbf{Y} = (y_{ij})$ a teljes adatmátrix és az $\mathbf{M} = (m_{ij})$ az adathiány indikátor mátrix. Az adathiány mechanizmus jellemezhető az **M** adott **Y** melletti feltételes eloszlásával, az $f(M|Y,\theta)$-val, ahol θ ismeretlen paramétereket jelöl.

A teljesen véletlenszerű adathiány (Missing Completely at Random – MCAR) esetében a teljes adatállománnyal rendelkező egységek és a hiányzó adatokat tartalmazó egységek teljesen egyformák, ugyanazon eloszlásból származnak.

A hiányzás tehát nem függ az **Y** értékétől, sem a megfigyelt, sem a hiányzó adatokkal rendelkező változók értékétől, azaz:

$$f(M|Y, \theta) = f(M|\theta), \quad \text{minden } Y, \theta \text{ esetén.} \qquad /1/$$

Ez a mechanizmus például akkor fordulhat elő, ha minden válaszadó egy pénzérme feldobásával dönti el, hogy válaszol-e a kérdésre.

Véletlenszerű adathiány (Missing at Random – MAR) esetében a hiányzó adatokat tartalmazó egységek eltérnek a hiánytalan adatokkal bíró egységektől, de a hiány jellegzetességei nyomon követhetők, előre jelezhetők az adatbázis más változói segítségével. Az adathiány tehát más változókkal kapcsolatban van, de azzal a változóval, amelyikben a hiányzás felmerül nincs közvetlen kapcsolatban.

Legyen $Y_{megfigyelt}$ azon változók halmaza az Y-ból, amelyben nincs adathiány és $Y_{hiányzó}$ azon változók halmaza, amelyben van adathiány. A véletlenszerű adathiány tehát az jelenti, hogy:

$$f(M|Y, \theta) = f(M| Y_{megfigyelt}, \theta), \quad \text{minden } Y_{hiányzó}, \theta \text{ esetén.} \qquad /2/$$

Ez a mechanizmus fordul elő például, ha a magasabb jövedelemmel rendelkezők nagyobb valószínűséggel tagadják meg a jövedelemre vonatkozó kérdések megválaszolását, de a jövedelemre következtetni tudunk a felmérés más változói (például: fogyasztási szokások, fogyasztás és megtakarítás egymáshoz való viszonya) alapján.

A *nem véletlenszerű adathiány* (Not Missing at Random – NMAR vagy másként „nonignorable", nem elhanyagolható) esetében az adathiány nem véletlenszerű, és más változókkal sem becsülhető, mert közvetlenül az adathiányt tartalmazó változóval van kapcsolatban. Az M eloszlása tehát függ az Y hiányzó értékeitől (is). Ez az adathiány legveszélyesebb, legnehezebben kezelhető formája.

Ez a mechanizmus fordul elő például, ha a magasabb jövedelemmel rendelkezők nagyobb valószínűséggel tagadják meg a jövedelemre vonatkozó kérdések megválaszolását, és a jövedelemre nem tudunk következtetni a felmérés más változóiból.

A hiányzó adatok számos problémát okoznak. Ugyanazon az adatbázison különböző kutatók által végzett elemzések eredménye között inkonzisztenciát tapasztalhatunk, ha azok másképpen kezelték a hiányzó adatokat. A hiányzó adatok kezelésére pedig azért van szükség, mert a sokasági paraméterbecslések torzítottak lehetnek (mint ahogy az 1992-es brit választásoknál is történt), hacsak az adathiány nem teljesen véletlenszerű.

A hiányzó adatok kezelésének célja éppen ennek a torzításnak az eltüntetése. Ezt a célt a különböző módszerek annak függvényében érik el, hogy mennyire helyesen sikerül azonosítani és modellezni az adathiány sajátosságait.

2. Hiányzó adatok kezelésére szolgáló módszerek

A hiányzó adatokkal való elemzés irodalma nem túl hosszú múltra tekint vissza. A szakirodalomban ajánlott és alkalmazott módszereket a következőképpen csoportosíthatjuk (*Little–Rubin* [2002]).

1. Teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások

2. Átsúlyozás

3. Imputációalapú eljárások
4. Modellalapú eljárások

A csoportok nem átfedésmentesek, de ebben a csoportosításban tekintjük át az alábbiakban a nemválaszolások kezelésének legelterjedtebb módszereit. A felsorolás nem tartalmaz minden alkalmazható módszert, csak a széles körben használt megközelítéseket.

2.1. Teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások

Az *adathiányt tartalmazó esetek törlését* (listwise vagy casewise adat törlés) említjük elsőként. Ha egy megfigyelési egységnél akár csak egy változó tekintetében is hiányzik adat, az egész megfigyelést törlik az adatbázisból. Az eljárást számos statisztikai programcsomag tartalmazza alapmegoldásként. A megoldás előnye az egyszerűsége, és az hogy az egyváltozós statisztikák összehasonlíthatók, mert mindegyik ugyanazon adatokon alapulva lett számítva. Hátránya viszont, hogy a nem teljes megfigyelésekben meglevő információt egyáltalán nem hasznosítja. Csak teljesen véletlenszerű eredetű adathiány esetén alkalmazható, azaz ha a hiányzó adatokat tartalmazó esetek az összes eseten belüli véletlenszerű almintának tekinthetők. Ha az adathiány nem MCAR, akkor a módszer torzítást okoz. Relatíve alacsony nemválaszolási arány mellett ésszerű lehet az alkalmazása, mert ekkor az egyszerűségből fakadó előnyök ellensúlyozhatják a néhány hiányzó adat által okozott információveszteséget és minimális torzítást.

Az elérhető adatok elemzése (available case analysis) a második módszer. Az adatok törléséből származó információveszteség csökkenthető, ha minden változó elemzésekor az abban a változóban meglevő összes adatot használjuk. A módszer hátránya, hogy a változónkénti elemzések más-más adatbázison készülnek, így az eredmények összehasonlítása problémás lehet. E módszer alkalmazásakor kétváltozós korrelációvagy kovariancia-számításhoz mindig az adott két változó tekintetében elérhető adatpárokat használják (pairwise available case). Számos statisztikai programcsomag tartalmazza ezt a kezelési módot. Előnye, hogy jobban kihasználja a meglévő adatokat, de az eredményeként létrejövő korrelációs mátrix nem feltétlen lesz pozitív definit. (Ekkor pedig ez a mátrix már nem is nevezhető korrelációs mátrixnak.)

Nézzük a következő példát (*Little–Rubin* [2002]), ami három változóra vonatkozóan 12 megfigyelést tartalmaz. (A „?” hiányzó adatot jelent.)

Y_1	1	2	3	4	1	2	3	4	?	?	?	?
Y_2	1	2	3	4	?	?	?	?	1	2	3	4
Y_3	?	?	?	?	1	2	3	4	4	3	2	1

Ebben a mintában az elérhető adatpárokat használva a mintából számított korrelációs együtthatók $r_{12} = 1$, $r_{13} = 1$, $r_{23} = -1$. Ezek a becslések nem jók, mert a sokasági korrelációs együtthatóknál $\rho_{12} = \rho_{13} = 1$-ből az következik, hogy $\rho_{23} = 1$, nem lehet -1.

Mivel az elérhető adatokat használatával több információra támaszkodunk, azt várnánk, hogy ez a megoldás hatékonyabb, mintha csak a teljes adatokat használnánk. *Kim* és *Curry* [1977] is ezt találták MCAR és gyenge korreláció esetén. Erősebb korreláció esetén viszont a teljes adatok használata bizonyult jobbnak (*Azen–Van Guilder* [1981]).

2.2. Átsúlyozás

Az átsúlyozásos módszerek abból indulnak ki, hogy válaszmegtagadás esetén a válaszmegtagadó elemhez hasonló nem adathiányos esetek (vele azonos kategóriában vagy rétegben szereplő elemek) arányosan több sokasági elemet képviselnek, azaz nagyobb súlyt kell kapniuk. Általában, ha a j-dik alcsoportban (kategóriában) a válaszadók aránya p_j, akkor az itt szereplő elemek $1/p_j$ súlyt kapnak, azaz itt mindegyik elem ennyiszer több sokasági elemet képvisel. Véletlen mintákból való következtetésnél, amikor az elemek kiválasztása nem azonos valószínűséggel történik, gyakran súlyozzák a megfigyelési elemeket a tartalmazási valószínűségük (probability of inclusion, azaz a minták hány százaléka tartalmazza az adott elemet) inverzével (*Hunyadi* [2001]). Legyen például y_i az Y változó értéke az i-dik megfigyelési egységre. Ekkor, ha nincs hiányzó adat, a sokasági átlag Horvitz–Thompson becslőfüggvénye:

$$\hat{\bar{Y}}_{HT} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}}, \qquad /3/$$

ahol π_i az i-dik egység ismert tartalmazási valószínűsége, a szumma pedig a megkérdezettekre vonatkozik.

Hiányzó adatok esetén az átsúlyozás úgy módosítja a súlyokat, mintha a nemválaszolás is a mintavételi terv része lett volna, ekkor a fenti becslőfüggvény a következőképpen módosul:

$$\hat{\bar{Y}}_{HTm} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i \hat{p}_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i \hat{p}_i}} \qquad /4/$$

Itt a szumma nem a megkérdezettekre, hanem a ténylegesen válaszolókra vonatkozik, a $\hat{p}_i$ pedig az i-dik egység becsült válaszadási valószínűsége (általában a válaszadási arány a minta egy alcsoportjában).

A módszer alapelve tehát egyszerű, de többdimenziós feladatoknál már igen bonyolult lehet a kivitelezése. Ráadásul a túlságosan szóródó súlyok nagy korrekciót jelentenek, ami megnöveli a feltételezések szerepét a becslésekben (*Hunyadi* [2001]). Az átsúlyozás mögött az a feltételezés húzódik meg, hogy az adott rétegen belül a válaszadók a megkérdezettek véletlen almintájának tekinthetők, azaz a rétegen belül az adathiány MCAR-jellegű. Az átsúlyozott mintából sokszor relatíve egyszerű a sokasági paraméterek pontbecsléseit elkészíteni. Az intervallumbecslésekhez szükséges standard hibák számítása már korántsem ilyen egyszerű. A statisztikai programcsomagok lehetővé teszik aszimptotikus standard hibák számítását összetettebb mintavételi tervek esetén, beleértve az átsúlyozást, rétegzést is. Ezek a programok azonban tipikusan fixnek, ismertnek tartják a súlyokat, pedig adathiány esetén a válaszadási aránnyal arányos súlyok maguk is mintavételi ingadozásnak vannak kitéve.

Egyszerű véletlen mintára vannak képletek a hibaszámításhoz, komplexebb esetekhez azonban a minta mesterséges újrahasznosításán alapuló nagy számítógépigényű módszerek (jackknife, bootstrap, kiegyensúlyozott ismétlések) alkalmazására van szükség.

2.3. Imputációalapú eljárások

Az imputáció azt jelenti, hogy a hiányzó adatot utólag mesterségesen pótolják egy ahhoz vélhetően hasonló értékkel. Ezután az így létrejött „teljes" adatbázison elvégezhetők a standard statisztikai elemzések. A helyes következtetéshez azonban módosítani kell a standard elemzéseket, valahogyan meg kell különböztetni a valódi és az imputált értékeket, hiszen ez utóbbiak újabb bizonytalansági faktort képeznek. Ezt a bizonytalansági tényezőt építi be a modellbe például a többszörös imputáció (multiple imputation).

Logikai imputációról (data editing) akkor beszélünk, ha a hiányzó értékek más adatokból, vagy korábbi felvételekből logikailag következnek és azokkal pótolják őket. Az emberek neme például nem változik, és a hiányzó életkorra is következtethetünk, ha egy korábbi felmérésnél megadták. A módszer előnye, hogy nem csökkenti az adatokban levő tényleges változékonyságot.

Az átlaggal való pótlás esetében az adott változóban meglevő adatok átlagával (átlag helyett más középérték is használható (módusz, medián)) helyettesítik a hiányzó értékeket. Az átlaggal való imputálás előnye az egyszerűsége, és könnyű alkalmazhatósága. Hátránya viszont, hogy bár teljesen véletlenszerű adathiány esetén várható érték szempontjából nem torzít, az elemek változékonyságát alulbecsli. Ez

javítható, ha a megfigyeléseket homogénebb csoportokra bontjuk és csoportokon belüli részátlagokkal imputálunk, de a standard hibákat és a becslések konfidencia-intervallumát még így is alulbecsüljük. Ez a módszer tulajdonképpen az átsúlyozással azonos eredményt ad.

A regressziós módszerek esetében a teljes megfigyeléseken építenek egy regressziót, a hiányzó értéket tartalmazó változót eredményváltozóként, a többit magyarázóváltozóként kezelve. Aztán azokra az esetekre, ahol az eredményváltozó értéke hiányzik, a regresszió segítségével becslést készítenek. A módszer továbbfejlesztéseként a *sztochasztikus regressziós imputálások* esetén egy véletlen változót is adnak a becslésekhez, mert e nélkül a változók közötti kapcsolat a későbbi elemzésekben szorosabbnak mutatkozna, mint amilyen valójában lehet.

A *hot deck imputáció* esetében a hiányzó adatot tartalmazó megfigyeléshez leginkább hasonló hiánymentes esetet megkeresik és ennek Y értékével pótolják a hiányos eset hiányzó Y értékét. A hasonlóság mértékének megítélésére különböző módszerek használhatók. A hot deck módszer előnye a fogalmi egyszerűsége mellett, hogy megőrzi a változók eredeti mérési szintjét (a kategóriás kimenetelű változók kategóriások maradnak, a folytonosak pedig folytonosak). A módszer hátránya, hogy nehéz az esetek hasonlóságát definiálni és az elemzőnek esetleg saját programot kell készítenie a donor egységek kiválasztásához. Ezenkívül a standard hibák számítása is nehézségekbe ütközhet (*Roth–Switzer* [1995]). A nehézségek ellenére a hot deck imputáció igen népszerű technika, számos hivatalos statisztikai felmérésben is ezt a módszert alkalmazták. (Például: Statistics Canada (*Rubin* [1987]).)Vannak modellek, amelyek több hasonló esetet keresnek és azokból véletlenszerűen választják ki a donor megfigyelést, vagy ha az megfelelő, az átlagukat számítják az imputációhoz. A hot deck (belső) módszereken sokszor tágabb értelemben az olyan adatpótlást értik, amely csak az adott mintát használja az imputációhoz, cold deck (külső) módszerek esetén pedig más, külső forrásokat (az adott mintához képest külső, például múltbeli hasonló felmérések adatai) is felhasználnak.

A *közelítő bayesi bootstrap* (Approximate Bayesian Bootstrap – ABB) módszer logisztikus regressziót alkalmaz, hogy az Y függő változóban a válaszolás/nemválaszolás valószínűségét becsülje az X_i változók segítségével. (Ilyen logisztikus regressziós módszert alkalmaz *György* [2004] a munkaerő-felvételben szereplő nemválaszolás kezelésére.) A megfigyelési egységek az így kapott hiányzás hajlamossági score-ok alapján képzett kvantilisekbe csoportosíthatók. A csoportokon belül a nem hiányos esetekből visszatevéses mintavétel segítségével lehet imputálni a hiányzó értékeket. Az eljárás minden hiányzó adatot tartalmazó változóra megismétlődik. A módszer a hot deck imputáció egy formája, ahol a hasonlóságot a hiányzás hajlamossági score-ok határozzák meg.

Léteznek ún. *kompozit módszerek* (composite methods) is, amelyek különböző módszerek alapelemeit ötvözik. Például a hot deck és a regressziós imputáció keve-

réke, amely először regresszióval számítja a becsült átlagokat, majd ezekhez hozzáadja egy véletlenszerűen kiválasztott empirikus reziduum értékét.

A nemválaszolás miatti bizonytalanság pótlólagos varianciaforrást jelent, amit valahogyan be kell építeni a becslésekbe. Ez megoldható például a minták másodlagos hasznosításán alapuló számítógép intenzív módszerek alkalmazásával, amelyekkel bonyolult mintavételi terv és imputációs technika esetén is becsülhető a becslőfüggvények varianciája. Több imputált adatbázis létrehozásával és azok eredményeinek összesítésével szintén beépíthető az adathiány okozta pótlólagos bizonytalanság a rendszerbe.

A *többszörös imputáció* (Multiple Imputation – MI) esetén minden hiányzó elem helyére több lehetséges értéket imputálnak, ezáltal több (általában 3-10) „teljes" adatbázist készítenek az eredeti hiányos adatbázisból. Az elemző mindegyik adatbázison elvégzi a megfelelő statisztikai módszerekkel a kívánt elemzéseket, a kapott eredményeket összegyűjti és kombinálja egyetlen elemzésbe. Ez utóbbi sokszor nem egyszerű feladat. A többszörös imputáció egy lépéssel tovább megy azzal, hogy bevezeti a statisztikai bizonytalanságot a modellbe, azért, hogy egy teljes adatbázisban meglevő változékonyságot közelítse az imputációval teljessé tett adatbázis is.

A többszörös imputációt először *Rubin* [1987] javasolta a hiányzó adatok kezelésére.

Furcsának tűnhet, hogy viszonylag kevés (3–10) imputációval is érzékeltetni lehet a pótlások bizonytalanságát. *Rubin* megmutatta, hogy *m* imputáción alapuló becslés relatív hatékonysága végtelen számú imputáció hatékonyságához képest nagyjából $\left(1+\frac{\gamma}{m}\right)^{-1}$, ahol γ a hiányzó információk aránya (számítását lásd később). Az *m* és γ különböző értékei mellett elérhető hatékonyságokat mutatja az alábbi táblázat:

Többszörös imputációval elérhető relatív hatékonyság
(százalék)

m	γ				
	0,1	0,3	0,5	0,7	0,9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Ha a hiányzó információk aránya nem túl magas, akkor igen kevés javulást eredményez néhánynál több imputált adatbázis készítése és elemzése. Az *m* darab imputált adatbázison elvégzett elemzések eredményeinek összegzésére Rubin azt a

módszert ajánlotta, hogy minden elemzésből mentsük el a becsült paraméterek és a standard hibák értékét. Legyen $\hat{\theta}_j$ a becsülni kívánt paraméter értéke (például egy regressziós együttható) a *j*-edik adathalmazból (*j*=1,2,...,*m*). U_j pedig legyen a $\hat{\theta}_j$ varianciája. Az összesítés utáni becslés az egyedi becslések átlaga lesz:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j \qquad /5/$$

Ezen becslés standard hibájához először az átlagos imputáción belüli varianciát:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j \qquad /6/$$

és az imputációk közötti varianciát kell kiszámolni:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\theta}_j - \bar{\theta}\right)^2. \qquad /7/$$

A teljes variancia:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B. \qquad /8/$$

Ahol az $\left(1+\frac{1}{m}\right)$ a véges *m* miatti korrekciós tényező.

Az együttes standard hiba pedig $\sqrt{T}$ lesz.

A $\gamma = \frac{\left(1+m^{-1}\right)B}{T}$ a nemválaszolás miatt a θ-ról hiányzó információk becsült aránya.

Nagy minták esetén a θ-ra vonatkozó szignifikancia tesztelése a $t = \frac{\theta - \bar{\theta}}{\sqrt{T}}$ próbafüggvénnyel történhet, ami a nullhipotézis alatt Student-féle *t*-eloszlást követ a következő szabadságfokkal:

$$v = (m-1)\left(1 + \frac{1}{m+1}\frac{\bar{U}}{B}\right)^2 \qquad /9/$$

ami a Satterthwaite-közelítésen alapul (*Rubin–Schenker* [1986] és *Rubin* [1987]).

A szabadságfok javított értéke kis mintákra:

$$\nu' = (\nu^{-1} + \hat{\nu}_{megfigy}^{-1})^{-1}, \qquad /10/$$

ahol

$$\hat{\nu}_{megfigy} = (1-\gamma)\left(\frac{\nu_{telj}+1}{\nu_{telj}+3}\right)\nu_{telj}, \qquad /11/$$

és ν_{telj} az adathiányt nem tartalmazó adatbázis esetén alkalmazandó szabadságfok. (*Barnard–Rubin* [1999]).[2]

Intervallumbecslés a paraméterre szintén ezek felhasználásával készülhet. További módszereket ismertet az eredmények összesítésére többszörös imputáció esetén *Schafer* ([1997], 4. fejezet).

A többszörös imputáció különböző módszerekkel történhet attól függően, hogy milyen jellegzetességekkel bír az adathiány. A többszörös imputációt besorolhatnánk a modellalapú eljárások közé is, mert legtöbbször bayesi eljáráson alapul: szükség van egy parametrikus modellre a teljes adatokra vonatkozóan és prior eloszlásra az ismeretlen modell paraméterekre (ez esetlegesen lehet neminformatív), aztán a hiányzó adatokra készít *m* független szimulációt a hiányzó adatok feltételes eloszlását használva (Bayes-tétel). Bonyolultabb parametrikus modellek esetén speciális számítási technikákra is szükség lehet, ezek közül leggyakrabban a Markov-lánc Monte-Carlo-[3] (Markov chain Monte Carlo – MCMC) szimulációt használják. *Rubin* [2003] egy olyan MCMC-szimulációt és beágyazott többszörös imputációt alkalmazó modellt ír le, amelyet három változó esetén a következő módon lehet illusztrálni.

Legyen a három változónk *X*, *Y* és *Z*. Kezdjük azzal, hogy valahogyan kitöltjük az *Y* és *Z* hiányzó értékeit (ezek az induló értékek), majd a megfigyelt *X*-ekkel építünk egy *X*|*Y,Z* modellt és e modell segítségével imputáljuk a hiányzó *X*-eket. Ezek után dobjuk ki az imputált (induló) *Y* értékeket és illesszünk egy *Y*|*X,Z* modellt a megfigyelt *Y*-okra , majd ezzel a modellel imputáljuk a hiányzó *Y*-okat. Aztán dobjuk ki az imputált *Z* értékeket és illesszünk egy *Z*|*X,Y* modellt a megfigyelt *Z*-kre , majd ezzel a modellel imputáljuk a hiányzó *Z*-ket. Az iteratív eljárás mindaddig folytatja a fenti lépések ismétlését, míg a kapott paraméterek nem konvergálnak.

[2] A képletek pontos elméleti háttere megtalálható a hivatkozott művekben.

[3] A Markov-lánc véletlen változók sorozata, amelyben minden egyes elem eloszlása az előző értékétől függ. A módszert eredetileg a fizikában használták egymással kölcsönhatásba lépő molekulák egyensúlyi eloszlásának feltárására. A statisztikai alkalmazások során többdimenziós, más módszerekkel megfoghatatlan eloszlások generálására használják.

A többszörös imputáció előnye, hogy könnyen érthető és elég robusztus a változók normalitási feltételének sérülése esetén is. Még például a bináris vagy az ordinális skálán mérő kategóriás változók esetén is gyakran elfogadható a normalitási feltétel melletti imputáció, majd a kapott folytonos imputált érték kerekíthető a legközelebbi kategóriára. Az erőteljes aszimmetriával rendelkező eloszlások közel normálissá transzformálhatók (például logaritmizálással), majd imputáció után visszatranszformálhatók az eredeti skálára. Hátránya viszont, hogy időigényes a három-tíz adatbázis imputálása, majd külön-külön az elemzések elvégzése, végül ezek összegzése. Ráadásul az összegzés módszertana még nincs minden statisztikai modellre kidolgozva. A többszörös imputációt több statisztikai szoftverbe is beépítették (például: a SAS enterprise Miner-hez írt Intelligent Multiple Imputation Software System – IMISS) ezek használatával az eljárás időigénye csökkent és sok kutató számára vonzó megoldássá vált.

2.4. Modellalapú eljárások

A modellalapú eljárások egy modellt definiálnak a megfigyelt adatokra és a becsléseket a modell melletti posterior valószínűségekre, vagy likelihoodra alapozzák. A megközelítés előnye a rugalmasság, a modellnél alkalmazott feltételezések explicit volta és az adathiányt is beépítő varianciabecslések elérhetősége.

Ilyen modellalapú becslés a maximum likelihood (ML) becslés, ami kiváló nagymintás tulajdonságokkal rendelkezik (konzisztens, aszimptotikusan hatásos, határeloszlása normális) (*Hunyadi–Vita* [2002]). A hiányzó adatok mintázata azonban nem mindig teszi lehetővé az ML-becslések explicit számítását.

Tegyük fel, hogy van egy modellünk az Y-ra, melynek eloszlását az $f(Y|\theta)$ sűrűségfüggvénnyel írhatjuk le, ahol θ ismeretlen paraméter. Legyen $Y = (Y_{megfigyelt}, Y_{hiányzó})$, ekkor $f(Y|\theta) = f(Y_{megfigyelt}, Y_{hiányzó}|\theta)$ az $Y_{megfigyelt}$, és az $Y_{hiányzó}$ együttes eloszlását leíró sűrűségfüggvény, az $Y_{megfigyelt}$ peremeloszlása pedig :

$$f(Y_{megfigyelt}|\theta) = \int f(Y_{megfigyelt}, Y_{hiányzó}|\theta)\, dY_{hiányzó}$$

Ekkor MAR-adathiány esetén a likelihood:

$$L(\theta | Y_{megfigyelt}) = \int f(Y_{megfigyelt}, Y_{hiányzó}|\theta)\, dY_{hiányzó}$$

Ekkor a ML-becslés a következő egyenlet megoldásával kapható:

$$D_{\ell}\left(\theta \middle| Y_{megfigyelt}\right) = \frac{\partial \ln L\left(\theta \middle| Y_{megfigyelt}\right)}{\partial \theta} = 0$$

Ha ennek az egyenletnek nincs zárt alakú megoldása, akkor iteratív módszerek alkalmazására van szükség. Ilyen iteratív módszer például a Newton–Raphson-algoritmus. Egy alternatív módszer a hiányzó adatokkal való becslések készítéséhez a várakozás maximalizáció (expectation maximization – EM), ami nem igényli a második deriváltak számítását, így nincs szükség olyan komplex programozási megoldásra, mint a Newton–Raphson-algoritmust alkalmazó módszerek esetén.

A következőkben ezt a módszert mutatjuk be, mert a gyakorlatban nagyon elterjedt az alkalmazása.

A *várakozás maximalizáció* egy általános módszer maximum likelihood becslésre MAR-típusú adathiány esetén. A módszer egy iteratív eljárás, amely két lépésből áll. Először, a várakozási lépésben (*E*) kiszámítják a teljes adatokat tartalmazó állományra a loglikelihood várható értékét, azután a maximalizáló lépésben (*M*) a kapott várható értékeket behelyettesítik a hiányzó értékek helyére és maximalizálják a likelihood függvényt, mintha nem lett volna hiányzó adat. Így új paraméterbecsléseket kapnak. Ez az iteratív eljárás mindaddig folytatja a fenti két lépés ismétlését, míg a kapott paraméterek nem konvergálnak. Konvergenciáról akkor beszélhetünk, ha a paraméterbecslések változása lépésről lépésre egyre kisebb lesz mígnem teljesen elhanyagolhatóvá válik. A konvergenciához annál több iteráció szükséges, minél több a hiányzó adat.

Nézzük meg egy egyszerű példán, hogyan működik az EM-módszer. A becslés elvégzéséhez valójában nincs szükség az EM-algoritmusra, csak a szemléltetés kedvéért választottuk.

Tegyük fel, hogy négyszer egymás után feldobunk egy pénzérmét, aminek az eredménye: (fej, fej, írás, ?), ahol a ? azt jelenti, hogy a negyedik dobás eredményét valamilyen oknál fogva nem ismerjük. Legyen a becsülni kívánt sokasági paraméter a „fej-dobás” valószínűsége, π. A teljes Y adatállományt felbontjuk megfigyelt és hiányzó részre: $Y = (Y_{megfigyelt}, Y_{hiányzó})$, a megfigyelt adatok valószínűségét a következő módon kapjuk:

$$P(Y_{megfigyelt}|\pi) = \sum_{Yhiányzó} P(Y|\pi) =$$
$$= P((\mathrm{F},\mathrm{F},\acute{\mathrm{I}},\acute{\mathrm{I}})|\pi) + P((\mathrm{F},\mathrm{F},\acute{\mathrm{I}},\mathrm{F})|\pi) = \pi^2(1-\pi)^2 + \pi^3(1-\pi) = \pi^2(1-\pi)$$

A megfigyelt adatok valószínűsége tehát ugyanaz, mintha a negyedik dobást egyáltalán nem vennénk figyelembe. Ekkor a π maximum likelihood becslése:

$$L(\pi|Y_{megfigyelt}) = P(Y_{megfigyelt}|\pi) = \pi^2(1-\pi)$$

$$D_L\left(\pi\middle|Y_{megfigyelt}\right) = \frac{\partial L\left(\pi\middle|Y_{megfigyelt}\right)}{\partial\pi} = 2\pi(1-\pi) - \pi^2 = 2\pi - 3\pi^2 = 0 \rightarrow \hat{\pi}_{ML} = \frac{2}{3}$$

A szemléltetés kedvéért nézzük, hogyan kaptuk volna meg ezt az eredményt az EM-módszer segítségével! Az *E* várakozási lépésben felírjuk a teljes adatok loglikelihoodjának várható értékét a jelenlegi $\pi^{(t)}$ becslés mellett.

$$Q(\pi|\,\pi^{(t)}) = \pi^{(t)}(3\ln\pi + \ln(1-\pi)) + (1-\pi^{(t)})(2\ln\pi + 2\ln(1-\pi))$$

Az *M* maximalizálási lépésben keressük *Q* maximumát π szerint, hogy megkapjuk $\pi^{(t+1)}$–et.

$$D_Q(\pi|\,\pi^{(t)}) = \frac{\partial Q(\pi|\,\pi^{(t)})}{\partial\pi} = \pi^{(t)}\left(\frac{3}{\pi} - \frac{1}{1-\pi}\right) + (1-\pi^{(t)})\cdot\left(\frac{2}{\pi} - \frac{2}{1-\pi}\right) = 0$$

Ebben az egyszerű esetben zárt formát kapunk az iterációra: $\pi^{(t+1)} = 0{,}5+0{,}25\pi^{(t)}$

Ha a kiinduló becslésünk mondjuk $\pi^{(0)} = 0{,}25$, akkor az iterációk sorozata: 0,2500; 0,5625; 0,6406; 0,6602; 0,6650; 0,6663; …, ami konvergál a 2/3-hoz.

Az EM-megközelítés előnye, hogy jól ismert statisztikai tulajdonságai vannak és általában jobban működik, mint az egyszerűbb listwise és pairwise adattörlések, az átlaggal való helyettesítés, vagy a regressziós imputálás (*Little* [1979], *Donner–Rosner* [1982], *Lee–Chiu* [1990]). Monte-Carlo-szimulációk is hasonló eredményeket mutattak (*Malhotra* [1987], *Graham–Donaldson* [1993]). Ugyanakkor ez az előny sokszor igen kicsi lehet (*Donner–Rosner* [1982]). A módszer hátránya annak viszonylagos bonyolultsága, ami miatt inkább csak statisztikusok számára vonzó megoldás. A legfontosabb gyengéje a módszernek, hogy a becsült adathoz nem ad bizonytalansági komponenst. A gyakorlatban ez azt jelenti, hogy míg a paraméterbecslések torzítatlanok lesznek, addig a standard hibák és a kapcsolódó tesztek nem megbízhatók. Ez a hiányosság arra késztette a statisztikusokat, hogy újabb likelihood alapú módszereket fejlesszenek ki. Ilyenek a teljes információs maximum likelihood módszer vagy a fent már tárgyalt többszörös imputáció alkalmazása.

(A *teljes információs maximum likelihood* (Full Information Maximum Likelihood – FIML vagy Raw Maximum Likelihood) minden elérhető adatot használ, hogy maximum likelihood alapú becsléseket készítsen. A módszert részletesen ismerteti például *Wothke* [1998].)

A maximum likelihood módszer MAR-típusú adathiányt feltételez, de a listwise és pairwise törlésekhez képest még nem véletlenszerű adathiány esetében is jobb eredményeket ad (*Wothke* [1998]).

A korábban ismertetett eljárások alkalmazásának szükséges feltétele a véletlenszerű adathiány (MAR). Vannak azonban olyan körülmények, amelyek esetén ez a feltételezés nem tartható, mert az adathiány kapcsolatban van a hiányt tartalmazó változóval. Ekkor az adathiány jellegét figyelembe vevő, a *nem véletlenszerű adathiány kezelésére* szolgáló modellek alkalmazására van szükség.

A NMAR-adathiánnyal foglalkozó kutatások alapvetően eltérő megközelítésük alapján két csoportra bonthatók: *szelekciós modellek* és *mintázatkeverék- (pattern-mixture) modellek*. Ezek a modellek az együttes valószínűséget eltérő módon bontják fel. A *szelekciós modellek* a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{hiányzó}| y_{megfigyelt}) P(y_{megfigyelt})$ felbontást használják. A szelekciós modellek feltételezik, hogy az adathiányt tartalmazó változó akkor és csak akkor figyelhető meg, ha egy másik változó (ami nem megfigyelhető) átlép egy küszöbértéket. Ilyen módszert alkalmazott *Heckman* [1976] kétlépcsős probit modelljében. A szelekciós modellek esetén a likelihood szokatlan eloszlású lehet, mert a paraméterek becsléséhez sokszor kevés információ áll rendelkezésre (*Schafer–Graham* [2002]).

A megoldás alternatívájaként alkalmazhatók a *mintázatkeverék-modellek*, amelyek a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{megfigyelt}| y_{hiányzó}) P(y_{hiányzó})$ felbontást alkalmazzák.

A mintázatkeverék-modellekkel foglalkozó tanulmányok: *Hedeker–Gibbons* [1997], *Little–Schenker* [1994], *Little* [1993], és *Glynn–Laird–Rubin* [1986]. Ezek a modellek kategorizálják a hiányzó értékek különböző mintázatait egy magyarázó változóba és ezt a magyarázó változót beépítik az adott statisztikai modellbe. Ezek után meghatározható, hogy az adathiány jellegzetességének van-e prediktív ereje akár önállóan (közvetlen hatás), akár más változókkal együttesen (interakciós hatás). A módszer előnye, hogy nem feltételezi a véletlenszerű adathiányt és részben használhatók hozzá statisztikai szoftverek, például a SAS MIXED proc. (például *Hedeker–Gibbons* [1997]), hátránya viszont, hogy az elemzőnek magának kell bizonyos lépéseket leprogramozni. Ha a megfigyelések számához képest sok változó esetén van relatíve sokféle eredetű adathiány, akkor a módszer elegendő adat hiányában nem működik.

3. Összegzés

A hiányzó adatok kezelésére nem létezik tehát egyetemesen legjobb megoldás. Pontosabban a legjobb gyógymód itt is a megelőzés. Ez sajnos nem mindig lehetséges, így ha már van adathiány, és az nem teljesen véletlenszerű, akkor valamilyen módon kezelni kell.

Összességében elmondható, hogy az általánosan használt egyszerű adathiány kezelési eljárásoknál (listwise és pairwise törlés, átlag imputálás) a hot deck, a maximum likelihood alapú és a többszörös imputációs eljárások a legtöbb esetben jobban teljesítenek. Mivel egyre szélesebb körben elérhető és könnyen használható szoftverek is tartalmazzák ezeket az eljárásokat, így az elméleti szerepükön túl az alkalmazásuk is egyre gyakoribb. Ezen módszerek mindegyike feltételezi a véletlenszerű

adathiányt, vannak azonban újabb statisztikai modellek a nem véletlenszerű adathiány kezelésére is. Ezekhez is használhatók (részben) az ismert statisztikai programcsomagok.

Az eljárások közötti választásban fontos szerepe van annak, hogy a cél *paraméterbecslések és tesztstatisztikák készítése*, vagy *konkrét megfigyelések hiányzó adatának becslése*. Az első esetben az adatbázis felhasználója kezeli a hiányzó adatokat és választhatja a saját elemzéséhez leginkább megfelelő módszert. A második esetben, ha például statisztikai hivatalok, kormányzati szervek nyilvánosságnak szánt adatbázisairól van szó, vagy olyan vállalati adatbázisokról, amelyeket sokféle belső kutatáshoz használnak, akkor olyan megoldást kell választani, ami nem igényel túl komplex bánásmódot a végső elemzések elvégzésekor. Ekkor például nem nagyon alkalmazható a többszörös imputáció. Fontos, hogy a választott imputációs eljárás kompatibilis legyen az imputált adatbázison később elvégzendő elemzésekkel. Az imputációs modellel szembeni elvárás, hogy megőrizze a későbbi vizsgálat tárgyát képező változók közötti kapcsolatokat. Ha például az Y változót egy olyan modellel imputálták, amelyik csak az X_1 változót tartalmazta, majd imputáció után a kutató egy lineáris regressziós modellt illeszt Y-ra X_1 és X_2 változók felhasználásával, akkor az X_2 együtthatója torzított lesz 0 felé, a helytelen imputáció következtében. Hasonló okokból, panel felvételeknél, például a keresztmetszeti kapcsolatokon kívül az adott változó korábbi hullámbeli tényleges vagy imputált értékét is figyelembe kell venni. Az imputált adatbázisokhoz mellékelni kell az imputáló által alkalmazott modellt, mert így az elemző láthatja, hogy milyen változókat vontak be a modellbe és mely változók közötti kapcsolatokat tekintettek implicite 0-nak.

Az imputációt sokan egyfajta *statisztikai alkímiának* tartják, amelyben a semmiből valahogyan új információ keletkezik. Ez a felvetés helytálló lehet az olyan imputációs eljárásokkal kapcsolatban, amelyek az imputált értékeket ugyanúgy kezelik, mint a ténylegesen megfigyelteket. Ha viszont pontosan közlik az alkalmazott módszert és a hiányzó adatok bizonytalansága is megjelenik, akkor a hiányzó adatok megfelelő kezelésével eltüntethető, vagy legalábbis csökkenthető a nem teljesen véletlen adathiányból eredő torzítás.

Irodalom

AZEN, S – VAN GUILDER, M. [1981]: Conclusions regarding algorithms for handling incomplete data. *Proceedings of the Staistical Computing Section, American Statistical Association.* 53–56. old.

BARNARD, J. – RUBIN, D. B. [1999]: Small-sample degrees of freedom with multiple imputation. *Biometrica.* 86. évf. 4. sz. 949–955. old.

CSEREHÁTI Z. [2004]: Az outlierek meghatározása és kezelése a gazdaságstatisztikai felvételekben. *Statisztikai Szemle.* 82. évf. 8. sz. 728–746. old.

DEMPSTER, A. P. – LAIRD, N. M. – RUBIN, D. B. [1977]: Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*. 39. évf. 1. sz. 1–38. old.

DONNER, A. – ROSNER, B. [1982]: Missing value problems in multiple linear regression with two independent variables. *Communication in Statistics*. 11. évf. 2. sz. 127–140. old.

GLYNN, R. – LAIRD, N. M. – RUBIN, D. B. [1986]: Selection modeling versus mixture modeling with nonignorable nonresponse. In: *Wainer, H.* (szerk.): *Drawing Inferences from Self-Selected Samples*. Springer-Verlag. New York.

GRAHAM, J. W. – DONALDSON, S. I. [1993]: Evaluating interventions with differential attrition: the importance of nonresponse mand use of followup data. *Journal of Applied Psychology*. 78. évf. 1. sz. 119–128. old.

GYÖRGY E. [2004]: A nemválaszolás elemzése a munkaerő felvételben. *Statisztikai Szemle.* 82. évf. 8. sz. 747–772. old.

HECKMAN, J. J. [1979]: Sample selection bias as a specification error. *Econometrica*. 47. évf. 1. sz. 153– 161. old.

HEDEKER, D. – GIBBONS, R. D. [1997]: Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods.* 2. évf. 1. sz. 64–78. old.

HUNYADI L. [2001]: A mintavétel alapjai. *Egyetemi Jegyzet SZÁMALK*. Budapest.

HUNYADI L. – VITA L. [2002]: Statisztika közgazdászoknak. *Központi Statisztikai Hivatal.* Budapest.

KIM, J. O. – CURRY, J. [1977]: The treatment of missing data in multivariate analysis. *Sociological Methode Researche*. 6. évf. 2. sz. 215–240. old.

LEE, S. Y. – CHIU, Y. M. [1990]: Analysis of multivariate polychoric correlation models with incomplete data. *British Journal of Mathematical and Statistical Psychology*. 43. évf. 1. sz. 145–154. old.

LITTLE, R. J. A. [1979]: Maximum likelihood inference for multiple regression with missing values: a simulation study. *Journal of the Royal Statistical Society*. 41. évf. 1. sz. 76–87. old.

LITTLE, R. J. A. – RUBIN, D. B. [1987]: *Statistical analysis with missing data.* John Wiley & Sons. New York.

LITTLE, R. J. A. – RUBIN, D. B. [2002]: *Statistical analysis with missing data.* 2. szerk. John Wiley & Sons. New York.

LITTLE, R. J. A. [1993]: Pattern-mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*. 88. évf. 421. sz. 125–134. old.

LITTLE, R. J. A. – SCHENKER, N. [1994]: Missing Data. In: *Arminger, G. – Clogg, C. C. – Sobel, M. E.* (szerk.): *Handbook for Statistical Modeling in the Social and Behavioral Sciences.* Plenum. New York. 39–75. old.

MÁDER M. P. [2005]: Imputálási eljárások hatékonysága. *Statisztikai Szemle*. 83.évf. 7. sz. 628–644. old.

MALHOTRA, N. K. [1987]: Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research*. 24. évf. 1. sz. 74–84. old.

ROTH, P. L. – SWITZER, F. S. [1995]: A Monte Carlo analysis of missing data techniques in a HRM setting. *Journal of Management*. 21. évf. 5. sz. 1003–1023. old.

RUBIN, D. B. [1987]: Multiple imputation for nonresponse in surveys. *John Wiley & Sons*. New York.

RUBIN, D. B. [2003]: Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica.* 57. évf. 1. sz. 3–18. old.

RUDAS T. [1998]: *Hogyan olvassunk közvélemény-kutatásokat?* Új Mandátum Könyvkiadó. Budapest.

SCHLAFER, J. L. [1997]: *Analysis of incomplete multivariate data.* Chapman & Hall. London.

SCHAFER, J. L. – GRAHAM, J. W. [2002]: Missing data: our view of the state of the art. *Psychological Methods*. 7. évf. 2. sz. 147–177. old.

STATISTICAL SOLUTIONS, Inc. [1998]: SOLAS for missing data analysis. Version 1. Cork, Ireland: Statistical Solutions.

WOTHKE, W. [1998]: Longitudinal and multi-group modeling with missing data. *Mahwah. NJ: Lawrence Erlbaum Associates.*

Summary

Missing data cause several problems. Inconsistency can be experienced among the results of analyses done on the same database by different researchers, if they handled missing data in a different way. There is a need of handling missing data, because the populational parameter estimations may be biased, unless the missing of data is not completely at random.

The aim of handling missing data is exactly to make this bias disappear. Different methods reach this aim in the function of how correctly the features of missing data can be identified and constructed. In this article we look shortly over the types of missing data and the most often recommended methods used to handle them, highlighting their main advantages and disadvantages.

There is no universally best solution to handle missing data. But we can say, that methods, based on hot deck, maximum likelihood and multiple imputation methods usually perform better, than generally used methods, handling simple missing of data (listwise and pairwise deletion, mean imputation). As widely accessible and easily practicable software include these methods, their application over their theoretical role is more and more common. All these methods presume missing data at random, there are however newer statistical methods to handle data, not missing at random as well.