A diadikus adatelemzés empíriával alátámasztott kritikája*

Dobos Imre,

a Budapesti Műszaki és Gazdaságtudományi Egyetem egyetemi tanára

E-mail: dobos@kgt.bme.hu

Gelei Andrea,

a Budapesti Corvinus Egyetem egyetemi tanára

E-mail: andrea.gelei@uni-corvinus.hu

A dolgozat célja, hogy empíriával „töltse fel" a diadikus adatelemzés néhány elméleti konstrukció kapcsán megfogalmazott matematikai-statisztikai kritikáját. A szerzők a felcserélhető eseteket tárgyalják először, majd a korrelációs mutatókra visszavezethető indikátorokat vizsgálják, végül pedig a diadikus adatelemzés ok-okozati modelljét tesztelik. Eredményeik szerint a kettős adatbevitel nem szolgáltat tényleges információs többletet a statisztikai elemzésben.

TÁRGYSZÓ:
Diadikus adatelemzés.
Kettős adatbevitel.
Diadikus korrelációk.

DOI: 10.20311/stat2018.01.hu0027

* A szerzők ezúton köszönik az OTKA K 115542 támogatását.

Az elmúlt évtizedekben a gazdasági jelenségek között különösen fontossá váltak a hálózati jelenségek. Ezek közé tartozik az üzleti hálózatok alapegysége, az üzleti kapcsolat, mely két vállalat kontextusában jön létre. Az üzleti kapcsolatok a vállalatok versenyképessége szempontjából kritikus jelentőségűek, hiszen azt jellemzőik közvetlenül is befolyásolják. Az üzleti kapcsolatok kutatása ugyanakkor módszertani kihívásokkal küzd. Ezért az elmúlt években részletesen foglalkoztunk a kérdőíves adatfelvételeken nyugvó statisztikai elemzések problémakörével. Munkánk során a nemzetközi szakirodalomban megjelent kritikai elemzések alapján abból indultunk ki, hogy a hagyományos adatfelvétel és a matematikai-statisztikai eszköztár, azaz az ún. egyvégű kutatások (*Brennan–Turnbull–Wilson* [2003]) számos kapcsolati jelenség vizsgálatára nem alkalmasak, az általuk kapott eredmények nem megbízhatók.

A szakirodalom egy része a hagyományos statisztikai elemzések korlátainak leküzdésére az ún. DA (dyadic data analysis – diadikus adatelemzés) használatát javasolja, melynek módszertanát, megközelítését, fogalmait, illetve elemzési eszközeit egy korábbi munkánkban (*Gelei–Dobos–Sugár* [2014]) már mi is ismertettük leíró módon, a 2016-ban megjelent tanulmányunkban (*Gelei–Dobos* [2016]) pedig bemutattuk konkrét gazdasági alkalmazását, illetve összehasonlítottuk eszköztárát a hagyományossal. Az utóbbi írásunk eredményei azt tükrözik, hogy az ajánlott új módszertan hozzáadott értéke nem nagy, annak néhány javasolt megoldását – ezen belül is elsősorban az ún. kettős adatbevitelt – érdemes újragondolni. A kettős adatbevitel matematikai-statisztikai kritikáját *Dobos* [2016] cikke fogalmazza meg. Erre építve jelen munkánkban azt mutatjuk be, hogy vajon az elméleti síkon megfogalmazott kritika és javaslatok az empirikus vizsgálatok tekintetében is megállják-e a helyüket.

Számításainkhoz a 2016-os írásunkban (*Gelei–Dobos* [2016]) bemutatott kérdőív néhány kérdését használtuk fel. A páros lekérdezés mintavételi eljárását alkalmazva, adatbázisunkat 89 adatközlő pár válaszai adták. Az akkori kutatásunk célja az volt, hogy teszteljük hipotézisünket, miszerint minél magasabb a bizalmi szint egy adott üzleti kapcsolatban, annál inkább jellemzik azt magas kockázati szintű cselekvések. A lekérdezéshez használt kérdőívet a Függelék tartalmazza. Az ebben szereplő kérdések közül jelen munkánkban véletlenszerűen választottunk ki kettőt (mint változókat, melyeket az elemzések leírásánál ismertetünk), hiszen célunk most nem a korábbi vizsgálat eredményeinek megismétlése, pusztán a diadikus adatfelvétel és a szakterület által javasolt páros minták alkalmazásával nyert korrelációk, illetve regressziós elemzések összevetése.

A következő fejezetben néhány elméleti alapkérdést tisztázunk, majd azokra építve végzünk homogenitáselemzést, és vizsgáljuk a DA korrelációs típusait, valamint az ok-okozati modelleket.

1. A diadikus adatelemzés kritikájának alapjai

Dobos [2016] kritikai munkájának alapgondolata a diadikus adatbevitelhez (double entry) kapcsolódik. A kettős adatbevitel lényege, hogy a páros lekérdezés révén nyert minden összetartozó adatpárból (diádból) két vektort képezünk úgy, hogy a diád elemeinek (azaz az összetartozó adatoknak) a sorrendjét megváltoztatjuk (*Gelei–Dobos–Sugár* [2014]). Az eljárás keretében két új változót definiálunk, ezeket X és X' szimbólumokkal jelöljük, amelyek az eredeti adatállomány n elemű vektorai helyett $2n$ elemű vektorok lesznek. Képzésüket az 1. táblázat szemlélteti, melyből kitűnik, hogy megfigyeléseik száma éppen a duplája a diádok, azaz a lekérdezett párok számának. E transzformációra azért van a diadikus adatelemzésben szükség, hogy táblázatok (mátrixok) helyett vektorokkal tudjuk az elemzéseket elvégezni.

1. táblázat

A kettős adatbevitel egy változójának (vektorának) képzése (double entry)

Megfigyelés	Változó	
	X	X'
1. számú pár (alapsorrend)	x_{11}	x_{12}
1. számú pár (felcserélt sorrend)	x_{12}	x_{11}
2. számú pár (alapsorrend)	x_{21}	x_{22}
2. számú pár (felcserélt sorrend)	x_{22}	x_{21}
3. számú pár (alapsorrend)	x_{31}	x_{32}
3. számú pár (felcserélt sorrend)	x_{32}	x_{31}
4. számú pár (alapsorrend)	x_{41}	x_{42}
4. számú pár (felcserélt sorrend)	x_{42}	x_{41}

Forrás: *Gelei–Dobos–Sugár* [2014] 426. old.

A diadikus adatelemzés bírálata megkérdőjelezi e transzformáció tényleges hasznát, a potenciális információtöbbletet. *Dobos* [2016] a téma matematikai hátterének kritikai vizsgálata során rámutatott arra, hogy a diadikus elemzések a fenti transzformáció nélkül, az alapadatok felhasználásával is elvégezhető. Jelen cikk ennek az elméleti kritikának az empirikus vizsgálatát tárgyalja.

Dobos [2016] elemzésének logikáját követve, – mint már említettük – egy korábbi páros lekérdezéssel kapott adatbázis (lásd *Gelei–Dobos–Sugár* [2014]) segítségével tárgyaljuk a témakört. Először tehát az összehasonlításhoz használt két adatbázist

mutatjuk be. Ezt követően azokon a diadikus adatelemzés ún. homogenitásvizsgálatával kapcsolatos elméleti javaslatokat teszteljük, majd a regressziókat vizsgáljuk empirikusan. Ezután a diadikus adatelemzés egyik regressziós modelljének, az ún. ICC-modellnek (intraclass correlation coefficient – osztályon belüli korrelációs koefficiens) a felhasználásával támasztjuk alá azt a kritikai megjegyzést, hogy a kettős adatbevitellel a regressziós modellek is rontják a becslést. Vizsgálataink során az SPSS 22 programcsomagjával és a Microsoft Excel statisztikáival dolgoztunk.

A diadikus adatelemzés két adatbázistípust különböztet meg, az ún. felcserélhetőt és a nem felcserélhetőt, azaz a megkülönböztethető megfigyelésekből álló adatpárokat tartalmazókat (*Gonzalez–Griffin* [2000]). A diadikus jelenségek esetén a páros minták alkalmazása akkor célravezető, ha az összetartozó párok (például ugyanannak a jelenségnek különböző időpontokban meghatározott jellemzői vagy bizonyos személyek [például egy orvos és betege, illetve házastársak]) között lényeges, aszimmetrikus kapcsolat van.

Más esetekben azonban előre nem határozható meg, hogy a vizsgált diádok tagjai között vajon aszimmetrikus viszony van-e. Erre példa a jelen írás adatfelvételében szereplő üzletemberek közötti kapcsolat és egy korábbi adatfelvétel hallgatói mintája is (*Gelei–Dobos–Sugár* [2014]).

A DA kritikáját *Dobos* [2016] a felcserélhető eset kapcsán fogalmazta meg, így megállapításainak tesztelésére szintén ez alapján kerül sor. A felcserélhető eset lényege, hogy az adatszolgáltatás során párokat alkotó két összetartozó válaszadó helyzete nem eltérő, közöttük (szemben a már említett ún. megkülönböztethető esettel) előzetesen semmilyen különbség nem állapítható meg.

Mint arra *Dobos* [2016] rámutatott, a felcserélhető esetben felcserélhetők a diádon belüli adatfelvételek is, így egy adott páros lekérdezéséből számos induló adatbázist képezhetünk. Alapadat-állománynak tekintjük a továbbiakban azt az adatbázist, amelyikben a páros lekérdezés során kapott diádokat a lekérdezés sorrendjében rögzítjük. Általánosságban igaz, hogy amennyiben a vizsgálatokat n darab diádon végezzük, úgy 2^n különböző induló adatbázis áll rendelkezésünkre, hiszen előre nem tudunk a diád elemei között különbséget tenni. Feltehető tehát a kérdés: melyik adatbázist válasszuk a további elemzésekhez?

A következő példa azt illusztrálja, hogy a diadikus adatelemzés ún. felcserélhető esetében egy adatfelvételből származó, de két, egymástól eltérő sorrendben rögzített adatbázis más-más eredményt adhat. Míg példánk első esetében az átlagok megegyeznek, addig a másodiknál szignifikánsan különböznek egymástól. (Összevetésüket a 2. táblázatban foglaljuk össze.)

2. táblázat

Két véletlenszerűen képzett adatbázis összehasonlítása a páros minták tesztjének felhasználásával

Adatbázis	A páros minták tesztjének eredményei							
	Páros különbség					*t*-teszt	Szabadságfok	Szignifikancia (kétoldalú)
	átlag	szórás	standard hiba	95 százalékos konfidenciaintervallum a különbségre				
				Alsó	Felső			
Első	0,079	1,798	0,191	–0,300	0,457	0,413	88	0,681
Második	1,135	1,391	0,147	0,842	1,428	7,694	88	0,000

Forrás: Itt és a továbbiakban saját számítás *Gelei–Dobos* [2016] adatbázisa alapján.

Az elemzéshez választott induló adatbázis kiválasztása tehát várhatóan befolyásolja az elemzés eredményét, ami probléma! Ennek kezelésére *Dobos* [2016] javaslata szerint olyan adatelemzési módszert szükséges alkalmazni, ami független az adatok felviteli sorrendjétől. Ilyen lehet a diádon belüli adatok összegének és/vagy különbségük abszolút értékének használata, hiszen az minden diád esetében állandó, függetlenül a rögzítés sorrendjétől. Ennek kapcsán két új változó (z_{i1} és z_{i2}) bevezetése javasolható:

$$z_{i1} = ½ (x_{i1} + x_{i2}), \text{ valamint } z_{i2} = ½ |x_{i1} - x_{i2}|,$$

ahol az *i*-edik diád első és második tagjának válasza ugyanazon kérdésre legyen x_{i1} és x_{i2}.

A két új változó közül z_{i1}-t az együttes hatást, míg z_{i2}-t a válaszok különbözőségét mérő új változóként értelmezhetjük. Mindkettő előnye, hogy érzéketlenek az adatrögzítés sorrendjére. E változókból könnyen visszaszámolhatók a felvett alapadatok:

a) Ha $x_{i1} \geq x_{i2}$, akkor $x_{i1} = z_{i1} + z_{i2}$, valamint $x_{i2} = z_{i1} - z_{i2}$.
b) Ha $x_{i1} < x_{i2}$, akkor $x_{i2} = z_{i1} + z_{i2}$, valamint $x_{i1} = z_{i1} - z_{i2}$.

A továbbiakban a fel nem cserélhető esetekkel foglalkozunk, de jelezzük, hogy az előbbi algoritmussal a felcserélhető esetek adatai is fel nem cserélhetővé tehetők!

A következőkben a kettős adatbevitellel nyert és az alapadatokat tartalmazó adatbázis felhasználásával kalkulált statisztikai mutatókat hasonlítjuk össze. Mint azt korábban említettük, célunk, hogy bemutassuk, a kettős adatbevitel nem feltétlenül jár többletinformációval.

2. Homogenitásvizsgálat és korrelációk

A fejezetben először a diadikus adatelemzés homogenitásvizsgálatát végezzük el a már korábban említett adatbázison. Ezt követően az alapadatok segítségével fejezzük ki a DA elméletében ismert korrelációs fogalmakat, elkerülve a kettős adatbevitelt és annak következményeit, majd az alapadatokkal közelítjük a kettős adatbevitellel meghatározott korrelációs együtthatókat.

2.1. Homogenitásvizsgálat kettős adatbevitellel és az alapadatokkal

A diadikus adatelemzés ún. páronkénti belső (csoporton belüli) individuális korrelációja (pairwise intraclass correlation) a diád tagjainak egy bizonyos kérdésre adott válaszainak hasonlóságát, azaz homogenitását méri. *Dobos* cikkében [2016] a kettős adatbevitel esetén előálló belső korrelációt alapadatokkal meghatározható korrelációra vezeti vissza. Az eredmények röviden a következőképpen összegezhetők:

$$r(X,X') = \frac{\operatorname{cov}(x_1,x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2}{\frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} = \frac{\sqrt{\operatorname{var}(x_1)} \cdot \sqrt{\operatorname{var}(x_2)} \cdot r(x_1,x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2}{\frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \leq$$

$$\leq \frac{\sqrt{\operatorname{var}(x_1)} \cdot \sqrt{\operatorname{var}(x_2)}}{\frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2}} \cdot r(x_1,x_2) \leq r(x_1,x_2) \ .$$

A képlet azon túl, hogy bemutatja, miképp lehet kiszámítani a kettős adatbevitellel kapott korrelációs értéket alapadatok felhasználásával, arra is rámutat, hogy az alapadatok közötti korreláció nagyobb, mint a kettős adatbevitelnél számított. Tehát az utóbbival információt vesztünk!

A következőkben a DA-val foglalkozó munkákban (például *Kenny–Kashy–Cook* [2006]) javasolt és az alapadatokra visszavezetett korrelációs képletek felhasználásával kapott eredményeket mutatjuk be. Számításainkhoz a már hivatkozott adatfelvétel kérdőívének első két kérdését használjuk, mely (egy –3-tól +3-ig terjedő skálán) azt méri, hogy a válaszadó párok mennyire ismerik egymást. (Lásd a Függeléket.) A belső korreláció értéke kettős adatbevitel esetén 0,4905, míg az alapadatokra visszavezetett képlettel számolva 0,4909 volt (mint arra már felhívtuk a figyelmet, az előbbi az alacsonyabb). Megítélésünk szerint ezért a kettős adatbevitellel kapott mutató pontossága nincs arányban a módszer okozta nehézségekkel.

2.2. Korrelációszámítás kettős adatbevitellel és az alapadatokkal

A korrelációs fogalmak tisztázásához egy újabb változót (kérdést) kell bevonni a vizsgálatba: ez az alapadatok szintjén az egyik változó esetén x_1, x_2, a másik változó esetén y_1, y_2 lesz. A kettős adatbevitellel az előbbi változókra történő transzformáció pedig (X, Y). A DA különböző korrelációs hányadosait egyenként vizsgáljuk. A DA-ban definiált és alkalmazásra javasolt korrelációs mutatók ismertetése *Giffin–Gonzalez* [1995] dolgozatában, a kiindulási adatállományra visszavezetett (azaz kettős adatbevitel nélkül kapott) korrelációs mutatók levezetése pedig *Dobos* [2016] cikkében olvasható.

Első lépésben röviden ismertetjük a bevezetőben bemutatott diadikus adatelemzéssel foglalkozó tanulmányokban tárgyalt korrelációs mutatókat, majd a kettős adatbevitellel nyert adatbázisunkra kiszámítjuk és az alapadatbázis statisztikai mutatóival (variancia-kovariancia) értelmezzük értékeiket. Végül e korrelációs mutatókat az alapadatbázisból számított korrelációs értékekkel közelítjük.

A válaszadó belső korrelációja

A válaszadó belső korrelációja (overall within-partner correlation) azt mutatja, hogy milyen lineáris kapcsolat mutatható ki a diád egyik szereplőjének két kérdésre (változóra) adott saját válaszai között. Képlete a következő:

$$r(X,Y)= \\ = \frac{\frac{1}{2}\left[\sqrt{\mathrm{var}(x_1)}\cdot\sqrt{\mathrm{var}(y_1)}\cdot r(x_1,y_1)+\sqrt{\mathrm{var}(x_2)}\cdot\sqrt{\mathrm{var}(y_2)}\cdot r(x_2,y_2)\right]+\frac{\left[E(x_1)-E(x_2)\right]\cdot\left[E(y_1)-E(y_2)\right]}{4}}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}.$$

Mint azt *Dobos* [2016] levezette, e korreláció az alapadatok korrelációival a következőképpen közelíthető:

$$r(X,Y)\sim\frac{\frac{1}{2}\left[\sqrt{\mathrm{var}(x_1)}\cdot\sqrt{\mathrm{var}(y_1)}\cdot r(x_1,y_1)+\sqrt{\mathrm{var}(x_2)}\cdot\sqrt{\mathrm{var}(y_2)}\cdot r(x_2,y_2)\right]}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}}}\leq \\ \leq\frac{1}{2}\cdot\left[r(x_1,y_1)+r(x_2,y_2)\right].$$

Egy válaszadó belső korrelációja tehát az alapadatbázis egy adott párját alkotó két válaszadó szóban forgó válaszai közötti korrelációk átlagával becsülhető.

A DA-n alapuló belső korreláció ($r(X,Y)$) értéke 0,588, az alapadatbázisból számított pedig 0,590, ami alátámasztja azt a megállapítást, hogy az alapadatokkal is jó közelítést adhatunk. A korrelációs együttható nagysága alapján a párt alkotó egyik személy vizsgált két (ismertségre és bizalomra irányuló) kérdésre adott válaszai között közepesen erős korreláció áll fenn.

A párt alkotó személyek közötti keresztkorreláció

A párt alkotó személyek közötti keresztkorreláció (cross-intraclass correlation) azt mutatja, hogy a diádot alkotó valamelyik személynek a vizsgált kérdések egyikére (például az ismertségre vonatkozóra) adott válasza milyen korrelációt mutat a párja által a másik (jelen esetben a bizalom szintje iránt tudakozó) kérdésre adott válasszal. Képlete:

$$r(X,Y') = \frac{\frac{1}{2}\left[\sqrt{\mathrm{var}(x_1)}\cdot\sqrt{\mathrm{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\mathrm{var}(x_2)}\cdot\sqrt{\mathrm{var}(y_1)}\cdot r(x_2,y_1)\right]+\frac{\left[E(x_1)-E(x_2)\right]\cdot\left[E(y_1)-E(y_2)\right]}{4}}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}.$$

E korrelációt a következőképpen közelíthetjük az alapadatok korrelációival:

$$r(X,Y') \sim \frac{\frac{1}{2}\left[\sqrt{\mathrm{var}(x_1)}\cdot\sqrt{\mathrm{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\mathrm{var}(x_2)}\cdot\sqrt{\mathrm{var}(y_1)}\cdot r(x_2,y_1)\right]}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}}} \leq$$

$$\leq \frac{1}{2}\cdot\left[r(x_1,y_2)+r(x_2,y_1)\right].$$

A párt alkotó személyek keresztkorrelációja tehát az alapadatbázis egy adott párjának két eltérő kérdésre adott válaszai közötti korrelációk átlagával becsülhető.

Az általunk vizsgált két kérdés esetén a párt alkotó személyek keresztkorrelációjának ($r(X,Y')$) értéke 0,291, az alapadatbázisból számított közelítés pedig 0,293, ami megfelelő. Ezért az eredmények szerint az a tény, hogy az egyik fél mennyire ismeri partnerét, gyenge korrelációban áll azzal, hogy benne e partner mennyire bízik meg.

Diádszintű korreláció

A diádszintű korrelációnak (mean-level correlation vagy correlation between dyad means) nevezett mutató, ami az összetartozó személyek két kérdésre adott vála-

szainak összege közötti lineáris összefüggést írja le, igazából nem feleltethető meg a klasszikus matematikai statisztikából ismert korrelációs hányados definíciójának. Sokkal inkább egy elméleti konstrukciónak tekinthető, melynek levezetését – ismereteink szerint – a szakirodalom nem teszi közzé. A diádszintű korreláció képlete:

$$r_m(X,X',Y,Y') = \frac{r(X,Y) + r(X,Y')}{\sqrt{1 + r(X,X')} \cdot \sqrt{1 + r(Y,Y')}} \ .$$

Az alapadatok variancia-kovariancia mutatóinak felhasználásával a következőképpen határozható meg:

$$r_m(X,X',Y,Y') = \frac{\frac{1}{2} \cdot \operatorname{cov}(x_1 + x_2, y_1 + y_2)}{\sqrt{\frac{1}{2} \cdot \operatorname{var}(x_1 + x_2)} \cdot \sqrt{\frac{1}{2} \cdot \operatorname{var}(y_1 + y_2)}} = r(x_1 + x_2, y_1 + y_2) \ .$$

Az előbbiek alapján nyilvánvaló, hogy az általunk javasolt mutató már megfelel a klasszikus korreláció fogalmának. A mutató alapadatokkal számolt értéke esetünkben 0,617. Következésképpen a két kérdésre adott válaszok diádszinten erősen közepesen korrelálnak.

Egyéni szintű korreláció

Az egyéni hatást mérő, más néven egyéni szintű korreláció (individual-level correlation) szintén egy elméleti konstrukció, amit nehezen lehetne egy matematikai-statisztikai korrelációs fogalommal leírni. Azt mutatja, hogy az összetartozó személyek két kérdésre adott válaszainak különbsége között milyen erős lineáris kapcsolat van. Definíciója a következő:

$$r_i(X,X',Y,Y') = \frac{r(X,Y) - r(X,Y')}{\sqrt{1 - r(X,X')} \cdot \sqrt{1 - r(Y,Y')}} \ .$$

Az alapadatok variancia-kovariancia hányadosai segítségével a következőképpen írható le:

$$r_i(X,X',Y,Y') = \frac{\sqrt{\operatorname{var}(x_1 - x_2) \cdot \operatorname{var}(y_1 - y_2)} \cdot r(x_1 - x_2, y_1 - y_2) + \left[E(x_1) - E(x_2)\right] \cdot \left[E(y_1) - E(y_2)\right]}{\sqrt{\operatorname{var}(x_1 - x_2) + \left[E(x_1) - E(x_2)\right]^2} \cdot \sqrt{\operatorname{var}(y_1 - y_2) + \left[E(y_1) - E(y_2)\right]^2}} ,$$

ami közelítve

$$r_i(X,X',Y,Y') \sim r(x_1 - x_2, y_1 - y_2) \ .$$

Az általunk vizsgált két kérdés esetén az egyéni szintű korreláció és az alapadatbázisból számított közelítés értéke egyaránt 0,522, ami erősen közepes korrelációra utal a párok tagjainak egymás általi ismerete és egymás iránti bizalmának szintjei között. Ez a korrelációtípus nem tekinthető hagyományos korrelációnak, ezért a közelítő értékét értelmezzük.

Páros szintű korreláció

A páros szintű korreláció (dyad-level correlation) elméletileg a szórásnégyzetnek azt a részét mutatja, ami a párok válaszai közötti eltérésekből adódik. Ez ugyancsak egy absztrakt konstrukció, mely definíció alapján a következő:

$$r_d(X,X',Y,Y') = \frac{\frac{1}{2}\left[\sqrt{\mathrm{var}(x_1)}\cdot\sqrt{\mathrm{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\mathrm{var}(x_2)}\cdot\sqrt{\mathrm{var}(y_1)}\cdot r(x_2,y_1)\right]-\frac{\left[E(x_1)-E(x_2)\right]\cdot\left[E(y_1)-E(y_2)\right]}{4}}{\sqrt{\mathrm{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\mathrm{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}.$$

Ez a típus nem létezik, ha

$$\mathrm{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2<0 \text{ és/vagy } \mathrm{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2<0.$$

A kifejezésünk számlálójában található kovarianciát elemezve látható, hogy a „helyes” korreláció ekkor a már korábban meghatározott, párt alkotó személyek közötti keresztkorrelációhoz ($r(X,Y')$) hasonlatos. E megközelítés arra épül, hogy a diád tagjai hasonló módon válaszolnak, ezért a páros szintű korreláció inkább a párt alkotó személyek közötti keresztkorrelációval mutat hasonlóságot. A kovariancia pedig közel varianciává válik abban az esetben, ha a két változó várható értékei és szórásai csaknem azonosak.

A kettős adatbevitellel nyert adatállományon számolva, a DA-n alapuló páros szintű korrelációs mutató értéke 0,689; az alapadatokból becsült értékre ugyanakkor 0,293 adható meg e korrelációtípus és a párt alkotó személyek közötti korreláció hasonlósága miatt. Így ez a fajta megközelítés a hasonlóság ellenére sem ad jó eredményt!

A korrelációs elemzések összefoglalása

A 3. táblázat a párok ismeretségi és bizalmi szintjei közötti lineáris kapcsolatokat mutatja be.

3. táblázat

A diadikus adatelemzés korrelációs mutatói és az alapadatokból becsült értékek korrelációtípusonként

A korreláció típusa	Kettős adatbevitellel kapott korrelációs mutató	Alapadatokból becsült korrelációs mutató
A válaszadó belső korrelációja	0,588	0,589
A párt alkotó személyek közötti keresztkorreláció	0,291	0,293
Diádszintű korreláció	0,617	0,617
Egyéni szintű korreláció	0,522	0,522
Páros szintű korreláció	0,689	0,293

Az eredmények alapján az alapadatokra visszavezetett becslés egyedül a páros szintű korreláció esetén nem adott jó közelítést a diadikus adatelemzési tanulmányokban javasolt korrelációs mutatóra. Minden más korrelációtípusnál közel azonosak voltak a kétféle módszerrel számított értékek. Ez szintén alátámasztja azt a megállapításunkat, hogy a kettős adatbevitelnek nincs valódi információtöbblet-értéke, így a vele járó nehézségek nem állnak arányban az általa elérhető pontosságjavulással.

3. Lineáris regresszió diadikus adatokkal: az ICC-modell

A lineáris kapcsolatok elemzése után az ok-okozati tényezők vizsgálatára térünk át. Ebben az esetben azt vizsgáljuk, hogy a magyarázóváltozók milyen hatással vannak az eredményváltozókra. A regressziós modell két változója szintén a párok közötti ismeretségi és bizalmi szinteket használja változóként. A magyarázóváltozó az ismertség, míg az eredményváltozó a párok között mért bizalom szintje.

A klasszikus statisztikában az előbbiek megválasztása egyszerűbbnek tűnik, mint a diadikus adatelemzésben. A diadikus jelenségek (például adott üzleti partnerek közötti bizalom) tanulmányozásának különlegessége ugyanis, hogy a vizsgált változók közötti összefüggéseket két hatás befolyásolja: egyrészt a kérdőívet kitöltők egyéni, személyes jellemzői, másrészt az, hogy éppen kivel kapcsolatosan, melyik konkrét párban kérik ki válaszukat. Ezeket általánosan egyéni (individual) és páros (dyadic) hatásnak nevezzük. Ezért akár egy független és függő változó esetén is a következő tényezőket lehet figyelembe venni a diadikus adatelemzés regressziós vizsgálatakor:

– a cselekvő hatása (actor effect),
– a partner hatása (partner effect) és
– a kölcsönös hatás (mutual effect).

E tényezők számának ismeretében felépíthetők a diadikus adatelemzés regressziós modelljei, melyek közül jelen cikkünkben az ICC-modellt tárgyaljuk (*Gonzalez* [2010], *Gelei–Dobos–Sugár* [2014]). Ez csak a cselekvő és a partner hatására épül. *Dobos* [2016] cikke már rámutatott arra, hogy a kettős adatbevitelen nyugvó diadikus regressziós modellek is visszavezethetők az alapadatokra, eredményeik azok felhasználásával jól becsülhetők. Célunk ezért most az, hogy ezt az elméleti levezetést adatbázisainkon teszteljük, azaz, hogy a regressziós paramétereket mindkét adatbázison „előállítsuk", és az eredményeket összevessük.

Kettős adatbevitel esetén az ICC-modellt alapadatokon a következő összefüggéssel becsülhetjük:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon_1 \text{ és } y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \varepsilon_2 \ .$$

A második egyenletben ugyanazok a regressziós együtthatók szerepelnek, mint az elsőben, ezért kettős adatbevitelnél pontatlan a becslés az y_2 adatokra a pár második tagjának válaszait tekintve.

A pontosabb becslés érdekében a következőt javasoljuk:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11} \text{ és } y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21},$$

ahol ε_{11} és ε_{21} a becslés hibái.

Ez esetben tehát az ICC-modell három együtthatója helyett hatot becslünk, mely által valóban pontosabbá tehetők eredményeink.

Először a kettős adatbevitellel kapott adatbázison végzünk számításokat az ICC-modell segítségével. A 4 *a*) táblázat szerint az *R* értéke 0,588, tehát az, hogy a pár adott tagja mennyire ismeri társát, csak gyengén közepes mértékben magyarázza a vele szemben mért bizalom szintjét. Az alapmodell és a modellben szereplő hatások közül az *X* magyarázóváltozó együtthatója szignifikáns, ám az X'-é nem az.

4. táblázat

Az X és az X', valamint az Y közötti ICC-modell eredményei a kettős adatbevitellel nyert adatbázis esetén

a) A korrelációs és a determinációs együttható

R	R^2	Korrigált R^2	Véletlen tényező becsült szórása
0,588	0,346	0,338	1,220

b) ANOVA-tábla

Megnevezés	Négyzetösszeg	d*f*	Átlagos négyzetösszeg	*F*	Szignifikancia
Regresszió	137,819	2	68,910	46,276	0,000
Maradványérték	260,591	175	1,489		
Összesen	*398,410*	*177*			

c) Együtthatók

Állandó/változó	Nem standardizált együttható		*t*	Szignifikancia
	B	Standard hiba		
Állandó	0,761	0,096	7,938	0,000
X	0,495	0,059	8,357	0,000
X′	0,004	0,059	0,063	0,950

Megjegyzés. Magyarázóváltozók: *X*, X', eredményváltozó: *Y*.

A következőkben a regressziót az alapadatokat tartalmazó adatbázison a korábban javasolt módon becsüljük, vagyis a korábbi három együttható helyett most hat paramétert kezelünk. Ebben az esetben a két becslőfüggvény két független egyenletre esik szét, azokat nem köti össze a közös együttható.

Dobos munkájában [2016] rámutatott, hogy az alapadatokra átírt lineáris modell pontosabb becslést nyújt, mint a kettős adatbevitellel nyert adatbázis alapján számított. Ez abból következik, hogy amennyiben a legkisebb négyzetek módszere alapján becslünk, akkor az utóbbi két egyenlet függetlenné válik. A legkisebb négyzetek módszerének becslőfüggvényei a paramétereket tekintve tulajdonképpen kvadratikus függvények, az első egyenlőségre $f_1(\beta_{01},\beta_{11},\beta_{21})$, míg a másodikra $f_2(\beta_{02},\beta_{12},\beta_{22})$. Mivel a paraméterek minimalizálják a becslőfüggvényeket, ezért a következő egyenlőtlenségek írhatóak fel:

$$f_1(\beta_{01},\beta_{11},\beta_{21}) \leq f_1(\beta_0,\beta_1,\beta_2) \text{ és } f_2(\beta_{02},\beta_{12},\beta_{22}) \leq f_2(\beta_0,\beta_2,\beta_1).$$

A matematikai statisztikából ismert, hogy az R^2-et maximalizálják a legkisebb négyzetek módszerével meghatározott paraméterek. Ebből pedig már következik, hogy az utóbbi két egyenlet jobb becslést ad. (Természetesen például a maximum likelihood becslésről is hasonlókat állapíthatunk meg, csak ott egy maximalizáló függvény áll elő becslőfüggvényként.) Az SPSS segítségével számított két regressziós modellt első és második modellnek nevezzük el. Számításaink eredményeit az 5. és a 6. táblázatokban mutatjuk be.

5. táblázat

Az x_1 és az x_2, valamint az y_1 közötti első regressziós modell eredményei az alapadatokból nyert adatbázis esetén

a) A korrelációs és a determinációs együttható

R	R^2	Korrigált R^2	Véletlen tényező becsült szórása
0,583	0,339	0,324	1,170

b) ANOVA-tábla

Megnevezés	Négyzetösszeg	d*f*	Átlagos négyzetösszeg	*F*	Szignifikancia
Regresszió	60,481	2	30,241	22,096	0,000
Maradványérték	117,699	86	1,369		
Összesen	*178,180*	*88*			

c) Együtthatók

Állandó/változó	Nem standardizált együttható		*t*	Szignifikancia
	B	Standard hiba		
Állandó	0,738	0,130	5,672	0,000
X	0,438	0,079	5,569	0,000
X′	0,035	0,082	0,429	0,669

Megjegyzés. Itt és a következő táblázatban magyarázóváltozók: x_1, x_2, eredményváltozó: *y*.

6. táblázat

Az x_1 és az x_2, valamint az y_1 közötti második regressziós modell eredményei az alapadatokból nyert adatbázis esetén

a) A korrelációs és a determinációs együttható

R	R^2	Korrigált R^2	Véletlen tényező becsült szórása
0,599	0,358	0,343	1,282

b) ANOVA-tábla

Megnevezés	Négyzetösszeg	d*f*	Átlagos négyzetösszeg	*F*	Szignifikancia
Regresszió	78,907	2	39,453	24,010	0,000
Maradványérték	141,318	86	1,643		
Összesen	*220,225*	*88*			

c) Együtthatók

Állandó/változó	Nem standardizált együttható		*t*	Szignifikancia
	B	Standard hiba		
Állandó	0,793	0,143	5,562	0,000
X	–0,028	0,086	–0,328	0,744
X′	0,558	0,090	6,192	0,000

Az alapadatokra, tehát a párok mindkét tagjára külön-külön számított regressziók – a diadikus ICC-modellhez hasonlóan – gyengén közepes magyarázóerővel rendelkeznek. A modellek eredményei alátámasztják korábbi megállapításunkat: gyakorlatilag a kettős adatbevitel nélkül pontosabb eredményeket nyerhetünk. Az APIM (actor-partner interdependence model – cselekvő-partner kölcsönös függőség modell) alapján is hasonló következtetésre juthatunk, ám annak bemutatásáról jelen munkánkban eltekintünk.

4. Összegzés

Dolgozatunkban a diadikus jelenségek statisztikai elemzésével foglalkoztunk, mellyel kettős célunk volt: egyrészt, hogy kritikát fogalmazzunk meg a felcserélhető esetekre (exchangeable cases) vonatkozóan, és megadjunk egy algoritmust az ezekből származó problémák kiküszöbölésére, másrészt, hogy a kettős adatbevitel módszerét állítsuk tanulmányunk középpontjába. Homogenitáselemzés mellett a diadikus adatelemzésben alkalmazott korrelációs fogalmakat is tárgyaltuk, és a diadikus jelenségeket ok-okozati szempontból regressziós modellekkel vizsgáltuk.

Eredményeink alapján arra jutottunk, hogy a felcserélhető esetek visszavezethetők fel nem cserélhető, ún. megkülönböztethető esetekre. Ez egy egyszerű algoritmussal valósítható meg, melyre írásunkban egy megoldást is bemutattunk.

A diadikus korrelációk típusait elemezve tanulmányoztuk, hogy a szakirodalomban javasolt kettős adatbevitellel valóban elérhető-e információs többlet. Ehhez a kettős adatbevitellel nyert korrelációs mutatókat először visszavezettük az alapadatokra, majd az így kapott indikátorokat az alapadatbázis korrelációs hányadosainak segítségével közelítettük. Empirikus eredményeink szerint az alapadatokra visszavezetett képletek többnyire jól becslik a kettős adatbevitellel nyert mutatókat. Regressziószámítással is kimutattuk, hogy a javasolt modellváltozatokkal nagyobb fokú illeszkedést érhetünk el.

Összességében ezért úgy gondoljuk, hogy a diadikus jelenségek vizsgálata során mindenképpen javasolt a páros adatfelvétel módszerének alkalmazása, de nem vagyunk teljes mértékben meggyőződve az ily módon nyert adatbázis megkettőzésének (double entry) indokoltságáról. Eredmények ugyanis azt mutatják, hogy e technikával statisztikai értelemben nem jutunk információtöbblethez.

Függelék

A páros lekérdezés során használt, a bizalom vizsgálatára kifejlesztett kérdőív

A kérdőív a BCE Logisztika és Ellátási Lánc Menedzsment Tanszékén végzett, az üzleti kapcsolatok vizsgálatát célzó kutatás kérdőíve. A kutatásban résztvevők a kérdőívet természetesen mindig anonim módon töltik ki. Az alábbiakban feltett kérdések az Önnel a kitöltés pillanatában párt alkotó személlyel és az őáltala képviselt vállalattal meglévő kapcsolatra és egy konkrét döntési szituációra vonatkoznak.

1. Kérjük x jel használatával –3-tól +3-ig terjedő skálán értékelje a kitöltés pillanatában éppen párját alkotó *személlyel* kapcsolatban a következő kapcsolati jellemzőket! (–3 = egyáltalán nem, 3 = teljes mértékben)

Értékelési szempont	–3	–2	–1	0	+1	+2	+3
Mennyire ismeri aktuális partnerét?							
Mennyire bízik meg aktuális partnerében?							

2. Kérjük x jel használatával –3-tól +3-ig terjedő skálán jelölje, hogy mennyire bízik meg abban a *vállalatban*, melynél aktuális párja most dolgozik? (–3 = egyáltalán nem; 3 = teljes mértékben)

Értékelési szempont	–3	–2	–1	0	+1	+2	+3
Mennyire bízik meg abban a vállalatban, melynél aktuális üzletfele dolgozik?							

3. Kérjük, jelölje, hogy a konkrét kapcsolatban megosztaná-e a jelzett információkat *konkrét párjával*! (*I* az Igen, *N* a Nem)

A megosztott információ jellege	Kérjük, jelölje, hogy igen, megosztaná (*I*-vel), vagy nem, nem osztaná meg (*N*-nel) a jelzett információtípusokat!
Operatív, a konkrét együttműködéshez szükséges információkat (például rendelési mennyiség vagy szállítási határidő)	
Operatív, de más együttműködő partnerrel folytatott kapcsolatot is befolyásoló információkat (például kapacitás- és készletadatok)	
Érzékeny pénzügyi információkat (például költségszint, profittartalom)	
Jövőbeli stratégiai tervekkel, innovációval kapcsolatos információkat (például új értékesítési útra vagy termékre vonatkozóan)	

4. Kérjük, jelölje, hogy ebben a konkrét vállalati együttműködésben megosztaná-e a jelzett információkat az adott vállalatnál dolgozó más partnereivel! *I*-vel (IGEN) jelölje, ha az alább felsorolt különböző együttműködési típusokban megosztaná a jelzett információkat partnerével annak érdekében, hogy az aktuális problémát megoldják! Amennyiben ezeket nem osztaná meg, azt kérjük, jelölje *N*-nel (NEM)!

A megosztott információ jellege	Kérjük, jelölje, hogy igen, megosztaná (*I*-vel), vagy nem, nem osztaná meg (*N*-nel) a jelzett információtípusokat!
Operatív, a konkrét együttműködéshez szükséges információkat (például rendelési mennyiség vagy szállítási határidő)	
Operatív, de más együttműködő partnerrel folytatott kapcsolatot is befolyásoló információkat (például kapacitás- és készletadatok)	
Érzékeny pénzügyi információkat (például költségszint, profittartalom)	
Jövőbeli stratégiai tervekkel, innovációval kapcsolatos információkat (például új értékesítési útra vagy termékre vonatkozóan)	

Irodalom

BRENNAN, R. – TURNBULL, P. W. – WILSON, D. T. [2003]: Dyadic adaptation in business-to-business markets. *European Journal of Marketing*. Vol. 37. Nos. 11–12. pp. 1636–1665. https://doi.org/10.1108/03090560310495393

DOBOS I. [2016]: A diadikus adatelemzés módszertanának egy kritikai vizsgálata: A kettős adatbevitel és felcserélhető eset. *Szigma*. XLVII. évf. 3–4. sz. 79–94. old.

GELEI A. – DOBOS I. – SUGÁR A. [2014]: Bevezetés a diadikus adatelemzésbe – elmélet és alkalmazás. *Statisztikai Szemle*. 92. évf. 5. sz. 417–446. old.

GELEI A. – SUGÁR A. [2016]: Diadikus jelenségek kutatási kihívása – a diadikus adatelemzés és a hagyományos statisztikai megoldások összehasonlítása. *Statisztikai Szemle*. 94. évf. 10. sz. 977–1003. old. https://doi.org/10.20311/stat2016.10.hu0977

GELEI A. – DOBOS I. [2016]: Bizalom az üzleti kapcsolatokban. *Közgazdasági Szemle*. LXIII. évf. Március. 330–349. old. https://doi.org/10.18414/KSZ.2016.3.330

GONZALEZ, R. – GRIFFIN, D. [2000]: On the statistics of interdependence: treating dyadic data with respect. In: *Ickes, W. – Duck, S.* (eds.): *The Social Psychology of Personal Relationship*. John Wiley and Sons Ltd. Chichester. pp. 181–213

GONZALEZ, R. [2010]: *Dyadic Data Analysis*. Paper presented at the „Research with Dyads and Families: Challenges and Solutions in Working with Interdependent Data" Conference. 19 May. West Lafayette. http://www-personal.umich.edu/~gonzo/purduedyad/capture-1.html

GRIFFIN, D. – GONZALEZ, R. [1995]: Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*. Vol. 118. No. 3. pp. 430–439. https://doi.org/10.1037/0033-2909.118.3.430

KENNY, D. A. – KASHY, D. A. – COOK, W. L. [2006]: *Dyadic Data Analysis*. The Guilford Press. New York, London.

Summary

The aim of the paper is to examine the mathematical statistics foundations of dyadic data analysis. In a former study, it was investigated whether the dyadic data analysis significantly contributes to traditional statistical analysis and provides surplus in understanding statistical phenomenon, or not. Further, the authors try to correct some of the mathematical structures of dyadic data analysis. The theoretical critics are supplemented by empirical tests.