Egyszerűen végezhető statisztikai eljárások II.

Hajtman Béla

Az adatokkal kapcsolatban föltett leggyakoribb kérdések egyike, hogy az általuk képviselt *változók (1)* milyen szoros kapcsolatban állnak egymással. A kapcsolat *mérésére* a *korrelációs együttható (2)* alkalmas, ez azonban meglehetősen sok és kényelmetlen számolással jár. Nem sokkal egyszerűbb a rangszámokból készült *Spearman-féle együttható (3)* kiszámítása sem. A gyakran idézett cikksorozatban röviden leírtuk a *Kendall-féle rangkorrelációs együttható* számításmódját is *(4)*, de a pár soros leírás alapján aligha sikerül bárkinek is az együttható kiszámítása. Mivel fontos, jól használható és könnyen számolható eljárásról van szó, most részletesebben foglalkozunk vele, és számpéldát is készítünk rá.

A Kendall-féle rangkorrelációs együttható

Az alapelv igen egyszerű, de elég nehéz szavakba önteni; megértéséhez némi képzelőerő is szükséges. Segítségül bevezetünk néhány szimbolikus jelölést is, ezek azonban nem tartoznak a lényeghez. Mindez csak a *fogalom* megértésére vonatkozik: a számolás könnyű, mechanikus feladat.

Kapcsolat vizsgálatakor két változó szerepel; nevezzük őket *x*-nek és *y*-nak. Minden vizsgált személynek[1] két adatát ismerjük: egy *x*- és egy *y*-adatot. (Természetesen ennél csaknem mindig jóval több adatot ismerünk; a tárgyalt eljárás során azonban csak ezek jönnek szóba.) *Kapcsolatot* az jelent a két változó közt, ha az *egyik* ismeretében valami elképzelésünk lehet a *másik* nagyságáról is. Például „mennél nagyobb *x*, annál nagyobb *y* is". A dolog szimmetriájából következik, hogy *x* és *y* szerepe fölcserélhető, és az állítás megfogalmazható *fordítva* is: „mennél kisebb az egyik, annál kisebb a másik is".

Sietek megjegyezni, hogy ez nem jelenti azt, hogy *x* értéke *befolyásolja* *y* értékét (vagy *y* értéke *x*-ét): lehet, hogy mindkettő egy „külső", harmadik tényező befolyása alatt viselkedik így, vagy még ennél is általánosabb, lazább az az oksági lánc, amely a két változó „együttjárását" eredményezi. Az okság, a befolyásolás kimutatása, ha egyáltalán sikerül, mindig csak *szakmai* – tehát nem statisztikai – megfontolások eredménye lehet.

A fentebb említett „együttjárás" a kapcsolatnak csak egyik fajtája, az úgynevezett *pozitív kapcsolat;* a *negatív kapcsolat* így fogalmazható meg: „mennél nagyobb *x*, annál kisebb *y*" – hozzágondolva természetesen a megfelelő „szimmetrikus változatokat" is.

A Kendall-féle rangkorrelációs együttható értéke – akárcsak az 1999-es dolgozatban említett többi korrelációs mérőszámé – pozitív kapcsolat esetén pozitív, negatív kapcsolat fennállásakor negatív. Nulla a kapcsolat hiányát, szerencsés esetben a teljes függetlenséget jelenti. Ez a mérőszám is, akárcsak a többi, „normált": értéke nem lehet 1-nél nagyobb, és az 1 (akár plusz, akár mínusz) a „legszorosabb kapcsolatot" jelenti.[2]

Az előbbiekben vázolt „együttjárás" vizsgálatát nagyon egyszerűen végezhetjük, ha az egész minta egyidejű figyelembevétele helyett mindig csak *párokat* vizsgálunk: kiragadunk két személyt, és az ő adataikat – összesen négyet: két *x*- és két *y*-adatot – vesszük szemügyre. Az adatok számszerű értéke nem érdekel, mindössze annyit figyelünk meg, hogy a „pár" első egyedének *x*-adata nagyobb-e vagy kisebb, mint a másodiké. Jelöljük – ám csak ideiglenesen, amíg a gondolatmenet végére érünk – ↑-lal, ha az első, ↓-lal, ha a második személy *x*-adata nagyobb. Ugyanezt elvégezve *y*-ra, a párra vonatkozóan négyféle megállapítást tehetünk (az első nyíl az *x*-ek, a második az *y*-ok közti viszonyt mutatja): ↑↑, ↓↓, ↑↓, illetve ↓↑ típusú párról van szó. Az első két esetben „ugyanúgy rendezett", a második két esetben „fordítva rendezett" párról beszélünk. (Az „ugyanúgy rendezett" párok a pozitív, a „fordítva rendezettek" a negatív kapcsolatot támasztják alá.)

Levelezési cím: dr. Hajtman Béla, Klinikai Biostatisztikai Társaság, 1519 Budapest, Pf. 256

LAM 2001;11(6–7):466–473.

A Kendall-féle együttható kiszámításához csupán az kell, hogy összeszámláljuk az „ugyanúgy” és a „fordítva” rendezett párokat.[3]

Jelöljük az „ugyanúgy rendezett” párok számát U-val, a „fordítva rendezett” párokét F-fel. A Kendall-féle rangkorrelációs együttható[4] értéke:

A képlet valóban egyszerű, de ne hagyjuk becsapni magunkat: a különböző típusú „párok” összeszámlálása nem könnyű feladat! E dolgozat tulajdonképpeni célja megmutatni, hogyan lehet ezeket a párokat könnyen, szinte szórakoztató módon megszámlálni. Előbb azonban átalakítjuk a képletet úgy, hogy abban ne szerepeljen U; akkor elég lesz megszámolni a „fordított” párokat.

A képlet átalakítása

Az alábbi kis „levezetés” csak azoknak szól, akik nem szeretnek elfogadni be nem bizonyított állításokat. Kétségtelen, a következő bekezdések nem lesznek mentesek némi matematikától. Aki nem akarja, ne is olvassa el őket, és – a szakasz végére ugorva – folytassa mindjárt az új képlettel. Nem kell mást tennie, csak *elhinni* azt.

Az „érdeklődők” figyelmét először arra hívjuk fel, hogy a nevezőben álló $(U+F)$ nem más, mint az *összes párok* száma. Erre pedig nem nehéz általánosan érvényes, tehát a konkrét feladattól független képletet gyártani.

Kiválasztva egy személyt (mondjuk az „elsőt”, akinek az adata legfölül szerepel), vele $(n–1)$ „párt” képezhetünk: az *összes többit* párba állíthatjuk vele. (Mint általában mindig, n a minta elemszámát jelöli.) Más személyeket választva, azokkal is $(n–1)$ pár alakítható ki, az összes pár száma tehát $n(n–1)$. Csakhogy így minden pár *kétszer* szerepel! Nagybetűvel szimbolizálva az egyes személyeket, az AB páron kívül a BA is előfordul (amikor nem A, hanem B mellé válogatjuk a „párokat”), és ez valamennyi párral így lesz. A párok *helyes* száma tehát $½[n(n–1)]$.

Most arra törekszünk, hogy U ne szerepeljen önmagában, csak $(U+F)$ részeként:

Behelyettesítve ebbe az összes párok számára, $(U+F)$-re az imént kapott formulát, kapjuk a Kendall-féle rangkorrelációs együttható formailag bonyolultabb, de számolás szempontjából sokkal-sokkal kedvezőbb alakját:

$$\tau = 1 - \frac{4F}{n(n-1)}.$$

A fordított párok számának egyszerű kiszámítása

Már az is nagy előny, hogy nem kell az összes párt figyelembe venni, csak a legtöbbször kisebb számban szereplő „fordítottakat”. (Eszerint gyakoribb lenne a pozitív kapcsolat? Talán így van, de lehet, hogy nem; előző mondatunk tartalma akkor is igaz marad – látni fogjuk, hogy miért.) A következőkben megmutatjuk azokat a „trükköket”, ahogyan ezt a számlálást egyszerűbbé – végül szinte nevetségesen egyszerűvé – tehetjük.

A minta elemeinek sorrendje teljesen közömbös: így jöttek sorban a betegek vagy így vettük elő a kórlapokat, hogy az adatokat átmásoljuk róluk. Rendezzük hát át a mintát úgy, hogy az x-adatok nagyság szerint növekedő sorrendben szerepeljenek! Ha az x és y közötti korreláció 1, a „második oszlop” y-adatai ugyanúgy növekvő sorrendben fognak ott állni. (Erre rögtön mutatunk egy teljesen légből kapott példát.) De ha nem így van – mint ahogy általában nincs így –, már csak az y-oszlopban kell összeszámlálnunk a fordítottan elrendezett párokat. Minden fordítva elhelyezkedő y-pár az F számot gyarapítja, hiszen x-ben – a konstrukció következtében – minden párban a második elem a nagyobb. (Az első szakasz szimbolikája szerint, az x-párokra mindig $\downarrow$ teljesül, tehát minden $\uparrow$ elrendezésű y-pár fordított párt jelez.)

A fordított párok összeszámlálása eszerint egyszerű: megnézzük, hogy a legelső y-adatot hány olyan követi az oszlopban, amelyik nála kisebb, ehhez hozzáadjuk a második y-adatot követő, nála kisebb adatok számát, majd rendre elvégezzük a következő adatokkal ugyanezt. Az eredményül kapott szám F.

A legutóbbi gondolatmenetben nagyon sokszor kellett összehasonlítást végezni: el kellett dönteni, hogy egy érték kisebb-e vagy nagyobb valamelyik másik értéknél. Ez a látszólag igen egyszerű művelet is nagyon fárasztó azonban, ha sokszor meg kell ismételnünk. Ezért úgy járunk el, hogy az adatokat – külön az x-, külön az y-adatokat – nagyság szerint beszámozzuk: a legkisebb kapja az 1, a következő a 2, …, az utolsó az n értéket. Ezzel az úgynevezett *rangszámokat* osztottuk ki *(3)*; ezentúl ezek a rangszámok helyettesítik az adatokat.

A mintát is könnyebb így átrendezni, mint az eredeti adatok alapján végezni a rendezést. Az új

1. táblázat. A szisztolés vérnyomás (*x*) és a percenkénti pulzusszám (*y*) kapcsolata két szélsőséges esetben.

a) *x*	a) *y*	b) *x*	b) *y*
110	64	140	58
130	72	125	68
95	56	110	72
125	70	130	66
135	76	100	76
100	62	115	70

„táblázat” első, *x* oszlopában az 1, …, *n* számok állnak, ebben a sorrendben. A hozzájuk tartozó *y*-okat ugyanezek az egész számok alkotják, csak sorrendjük lesz más. Most azonban már sokkal könnyebb eldönteni, hogy egy adat (például 5) alatt hány nála kisebb szám fordul elő: csak azt kell megnéznünk, hogy az 1, 2, 3, 4 számokból hány áll „lejjebb” az oszlopban 5-nél. (Az eredeti adatokkal, esetleg sokjegyű számokkal ez lényegesen nehezebb lenne.)

Mielőtt az utolsó, leglényegesebb egyszerűsítést végrehajtanánk, nézzünk néhány példát! Valószínű, hogy a szisztolés vérnyomás és a percenkénti pulzusszám semmiféle – statisztikai értelemben vett – kapcsolatban nem áll egymással. De azért nézzünk előbb egy olyan példát, amelyikben a kettő közti Kendall-féle együttható 1, majd egy olyant, ahol –1 értéket ad. Az elemszám mindkét esetben 6 *(1. táblázat)*.

Az összes párok száma (az előző szakasz formulája alapján) 15. Az a) példarészben valamennyi pár ugyanúgy rendezett, a b) példarészben pedig valamennyi fordítva. Ezt azonban szinte lehetetlen ebben a formában észrevenni, és elég hosszú időbe telik mind a 15 párt megvizsgálni. Sokkal könnyebb a dolgunk, ha előbb kiosztjuk a rangszámokat *(2. táblázat)*, és a mintákat rendezzük az *x*-adatok növekvő sorrendjében (*3. táblázat)*. Utóbbiban az eredeti adatokat elhagytuk, és csak a rangszámokat tüntettük fel. Így rögtön látható, hogy az a) példarészben $F=0$: sehol nem fordul elő, hogy egy kisebb adat követne egy nagyobbat. A b) példarészben azonban csupa ilyenek vannak! Számláljuk össze őket (természetesen az *y*, a pulzusszám-oszlopban): a 6 alatt

2. táblázat. Az 1. táblázat adatai, a rangszámok kiosztása után. *(A kisebb számjegyek jelentik a rangszámokat.)*

a) *x*	a) *y*	b) *x*	b) *y*
110 3	64 3	140 6	58 1
130 5	72 5	125 4	68 3
95 1	56 1	110 2	72 5
125 4	70 4	130 5	66 2
135 6	76 6	100 1	76 6
100 2	62 2	115 3	70 4

3. táblázat. Az 1. táblázat adatai, rangszámokkal helyettesítve és átrendezve.

a) *x*	a) *y*	b) *x*	b) *y*
1	1	1	6
2	2	2	5
3	3	3	4
4	4	4	3
5	5	5	2
6	6	6	1

öt, nála kisebb adat található, az 5 alatt négy kisebb van, a 4-et három, a 3-at kettő ilyen követi, és a 2 alatt is nála kisebb szám áll. Összeadva ezeket: $5+4+3+2+1=15$; ennyi *F* értéke.

Ilyenkor az együtthatót számolni sem kell! Az első esetben $F=0$, a másodikban $U=0$, és ha a képlet eredeti, első formáját nézzük, azonnal látjuk, hogy

az első esetben a másodikban

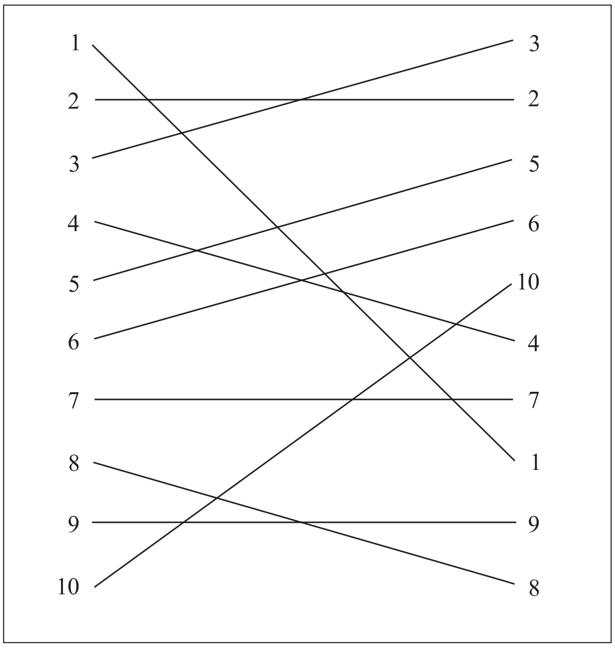
adódik, vagyis +1 és –1, ahogyan ígértük.

Lássunk ezek után egy „igazi” példát! Ez sem „valódi” abban az értelemben, hogy nem méréseken alapul, hanem kitalált adatokon. Mivel úgyis csak a módszert akarjuk illusztrálni, jelentse *x* most is a szisztolés vérnyomást és *y* a pulzusszámot. Ezúttal azonban legyen 10 adatunk *(4. táblázat)*.

Most pedig lássuk azt a trükköt, ahogyan *F*-et a legegyszerűbben lehet kiszámítani! Kis túlzással azt mondhatjuk, hogy az egész dolgozatot azért írtam, hogy ezt a módszert mennél több érdeklődő közkincsévé tegyem: ne hagyjuk azt statisztikai könyvek lapjain elfonnyadni! Maga a módszer (sajnos!) nem tőlem származik, több helyen megjelent már [először talán *Wilcoxon* és *Wilcox* kis könyvében *(6)*], de a számítógépek korában ezek az egyszerű, sőt merem így mondani: *kedves* eljárások kiesnek az érdeklődés köréből, és lassan feledésbe merülnek. Hiszen a GÉP, a nagy, szent masina úgyis kiszámít

4. táblázat. Újabb példa a szisztolés vérnyomás és a percenkénti pulzusszám kapcsolatának „vizsgálatára”. *(Föltüntettük a rangszámokat is, így könnyebb a példa és az 1. ábra kapcsolatát megtalálni.)*

x		*y*	
160	10	72	8
120	5	78	10
115	4	66	6
90	1	58	3
130	6	60	4
135	7	70	7
105	3	64	5
155	9	76	9
100	2	56	2
140	8	52	1



1. ábra. *F* meghatározása a 4. táblázat példájában. *(F a metszéspontok számával egyenlő; a példában F=15.)*

mindent – és olyan gyors, hogy tiszta vétek egyszerűsíteni, könnyebbé tenni valamit. Micsoda szegényes, ostoba ideológia!

Térjünk vissza utolsó példánkhoz! (A módszer természetesen használható az első két példa esetében is, csak ott nincs rá szükség. De tessék kipróbálni: ott is működik!) A fordítva rendezett párok számának egyszerű meghatározása érdekében az adatokat rangszámokkal helyettesítettük, és úgy írtuk föl a mintát, hogy az *x*-oszlopban a számok természetes sorrendben szerepeljenek *(1. ábra)*. Most pedig – és ez maga az új módszer – *kössük össze az azonos számokat a két oszlopban!* Ha megvan, számláljuk össze a metszéspontokat; ez a szám *F*-fel egyenlő! (Ilyenre szokták mondani, hogy aki ennél egyszerűbbet tud, az már csal.)

Az *1. ábráról* leolvashatjuk, hogy példánk esetében a metszéspontok száma 15, azaz $F=15$. Zsebszámológéppel mindig könnyen elvégezhető τ kiszámítása, de jelen esetben akár fejben is számolhatunk. Eszerint:

Ha sok a fordított elrendezésű pár, a vonalak „összegabalyodhatnak". (Erre már példánkban is vigyázni kellett.) Amennyiben több metszéspont torlódna valahol, „kanyarodjunk" kissé a vonalakkal[5]! Fontos, hogy a metszéspontokat jól össze lehessen számolni.

Alkalmazhatunk egy másik trükköt is. Ha meglehetősen negatív a kapcsolat (és ezért túlságosan sok a fordított pár), *megfordíthatjuk* az elrendezést: az *x*-adatokat továbbra is növekvő, az *y*-okat azonban csökkenő sorrendben látjuk el rangszámokkal. (Például az első példa b) részében célszerű így eljárni.) Az eredményt ez csak annyiban befolyásolja, hogy a Kendall-féle együttható ellenkező előjelűre fordul, abszolút értéke azonban változatlan marad. (Annyi történik, hogy *U* és *F* „helyet cserél"; az első képlet mutatja, hogy ez csupán előjelváltást jelent.) Megtehetjük ezt előre is, ha valamiért negatív kapcsolatot várunk, de ha a rangszámok összekötésekor látjuk úgy, hogy túl sok a metszéspont, fordítsuk meg az *y*-adatok számozását, és kezdjünk neki újból! Megéri ez a kis kényelmetlenség, hogy utána azután könnyen, gyorsan a végére érjünk.

A Kendall-féle rangkorrelácós együttható szignifikanciája

Azt már tudjuk, hogy mit jelent a pozitív, mit a negatív együttható, mi a jelentése a nulla körüli „kicsi" és az abszolút értékben 1-hez közeli „nagy" együtthatóknak. De ha a kapcsolat létének statisztikai alátámasztását is akarjuk, ha ki szeretnénk mondani, hogy az együttható „szignifikáns" (vagyis ha szükségünk van a mindnyájunk által oly jól imert *p*-értékre), akkor elő kell vennünk a Kendall-féle rangkorrelációs együttható táblázatát.

Valamire azonban mindenképpen vigyázzunk: ne hasonlítsuk össze τ-t más korrelációs együtthatókkal! A kapcsolatot „más skálán" méri a Kendall-féle együttható, mint az úgynevezett *lineáris korrelációs együttható (2)* és az ahhoz sokban hasonló *Spearman-féle rangkorrelációs együttható (3)*. (Jóllehet mindegyik mérőszám csak –1 és +1 közti értékeket vehet fel.)

Mielőtt megadnánk a τ együttható táblázatát, szeretnénk néhány szót szólni a táblázat jelentéséről és formájáról. A táblázat kifejezetten az együttható szignifikanciájának eldöntését segíti, ezért – mint minden hasonló statisztikai táblázat – a nullhipotézisnek megfelelő esetet képviseli. Ha a változók közt *semmiféle kapcsolat nincs* (nullhipotézis), akkor *x* és *y* sorrendje akárhogy alakulhat, és feltevésünk szerint minden sorrend egyformán valószínű (miért ne lenne az?): ha az *x*-adatokat sorba rendezzük (mint eddig is tettük), csupán a véletlen szabja meg, hogy milyen ehhez képest *y* sorrendje. Előállítva az *összes lehetséges* sorrendet és kiszámítva mindegyikből a τ együtthatót, meg lehet kapni τ eloszlástáblázatát. Ebben aztán kijelölhető az a τ-érték, amilyen nagy (és minden nála nagyobb) csak igen ritkán – az általunk előzetesen kiválasztott

5. táblázat. A Kendall-féle rangkorrelációs együttható (τ) táblázata. *(A táblázatban található küszöbértékek kiszámítása (7) alapján történt, ahol az U-értékek teljes eloszlástáblázata található meg n=2 és n=12 között.)*

	Kétoldali valószínűségek				
n	0,20	0,10	0,05	0,02	0,01
4	1,000	1,000	—	—	—
5	0,800	0,800	1,000	1,000	—
6	0,600	0,733	0,867	0,867	1,000
7	0,524	0,619	0,714	0,810	0,905
8	0,429	0,571	0,643	0,714	0,786
9	0,389	0,500	0,556	0,667	0,722
10	0,378	0,467	0,511	0,600	0,644
11	0,345	0,418	0,491	0,564	0,600
12	0,303	0,394	0,455	0,545	0,576

Egyoldali hipotézis vizsgálata esetén a feltüntetett valószínűségeket felezni kell.

szignifikancia-valószínűségnél ritkábban – fordul elő. Ezeket a τ-küszöbértékeket foglaltuk táblázatba különböző szignifikanciaszintek esetére *(5. táblázat)*.

A táblázatot a minta elemszámának megfelelő sorban kell mindig használni. A benne szereplő együtthatók *abszolút értékek*: akár pozitív, akár negatív a mi kiszámított együtthatónk, akkor kell azt szignifikánsnak tekinteni a fejlécben álló szinten, ha abszolút értéke *akkora vagy nagyobb,* mint a táblázatban álló szám.

A táblázatban kétoldali valószínűségek találhatók. Ha előre tudjuk, hogy milyen előjelű a kapcsolat – a dózis növelésével csak a hatás növekedésére számítunk: pozitív kapcsolat; azt várjuk, hogy idősebb gyermekek rövidebb idő alatt oldják meg ugyanazt a feladatot: negatív kapcsolat –, *egyoldali próbavégzéssel* is próbálkozhatunk. Tudjuk *(8)*, hogy a táblázatban álló szinteket ilyenkor *felezni* kell – de csak akkor végezhetjük el a próbát, ha az együttható előjele olyan, amilyennek vártuk.

Vegyük észre, hogy (kisebb elemszámok esetén) a táblázatban milyen sok az egyforma érték. Ennek az az oka, hogy egy-egy eloszlásban *nagyon kevés* különböző τ-érték fordul csak elő. (Pontosan annyi, ahány értéket F felvehet. A példák során láttuk, hogy $n=6$ esetén ez mindössze 16.) Így azután két lehetséges τ között esetleg nagyot „ugrik" a valószínűség.

Az *1.*, illetve *2.* és *3. táblázatban* álló két példa eredményét akár ki sem kell keresni: a +1 és –1 együtthatók „a lehető legszignifikánsabbak": nagyon valószínűtlen, hogy ez az (egyetlen) sorrend véletlenül forduljon elő[6]. Persze nem kizárt! Hiszen statisztikai próba esetén *mindig* számítani kell az (első fajta) hibára *(9)*.

A *4. táblázat* példájának szignifikanciáját azonban ellenőrizhetjük. Az $n=10$-hez tartozó sorban látjuk, hogy τ=0,333 semmilyen, a táblázatban szereplő szinten nem szignifikáns. (Írhatjuk, hogy $p>0{,}20$.) A nullhipotézist, amely a vérnyomás és a pulzusszám függetlenségét mondja ki, nem sikerült megcáfolni – de persze éppen erre számítottunk.

A τ együtthatót ugyan viszonylag könnyen ki tudjuk számítani nagyobb mintákból is – bár ha túl sok a metszéspont, meggyűlhet a bajunk az összeszámlálásukkal –, de a szignifikancia eldöntése az *5. táblázatból* nem megy, ha 12-nél nagyobb az elemszám. Nem ismerek olyan táblázatot, amelyik segítene ezen a problémán. (Sőt, ez a 12-ig elkészített táblázat is ritkaság: általában 10-nél vagy még hamarabb megállnak.) Biztosan van olyan gépi program, amelyik kiszámítja a τ-hoz tartozó p értéket, és jó közelítő formulák léteznek „házi használatra" is. Ezek azonban már nem illenek ebbe, az „egyszerű eljárásokat" tárgyaló dolgozatba; az érdeklődő olvasó megtalálja őket a statisztikai szakkönyvekben (például *5, 6*).

Az egyforma adatok problémája

Hallgatólagosan mindvégig feltételeztük azt, hogy az adatok egyértelműen rendezhetők nagyság szerinti sorba, azaz hogy két adatról mindig el lehet dönteni, hogy melyik a nagyobb. De mi van, ha egyforma adatok is előfordulnak?

Az egyszerű – de nagyon is elméleti ízű – válasz az, hogy ilyesmi pedig nem lehetséges, hiszen – folytonos változót feltételezve (márpedig a méréseredmények mind ilyenek) – egyformák előfordulásának nulla a valószínűsége. A gyakorlat azonban rácáfol erre. A korlátozott mérési pontosság még akkor is egyformának mutat néhány adatot, ha azok valójában mind különbözők.

A legtöbb szerző azt javasolja, hogy egyforma adatok előfordulásakor *ne* számítsunk Kendall-féle rangkorrelációs együtthatót. Ez elég érzékeny veszteség, mikor már kezdtünk örülni a fogalmi felépítését tekintve nagyon logikus és egyszerűen számítható mérőszámnak.

Pedig sajnos ebbe bele kell törődnünk. Vannak, akik a máskor bevált módszereket – a rangszámok „kapcsolását" *(3)* és az egyformák „elhagyását" *(10)* – javasolják, de ez nem látszik járható útnak. A sok javaslat közül, amit a probléma megoldására tettek, még az a legjobb, amelyik azt mondja: számítsuk ki az összes lehetséges együtthatót, amit akkor kapnánk, ha azok az egyforma adatok (így vagy úgy) mégis különböznének. De gondoljuk csak meg: ha x-ben van *három* egyforma adat, ez hatféle lehetséges sorrendet jelent. Ha ehhez járul még nem több, mint *két* egyforma az y-adatok közt, azokat kétféleképp rendezhetjük el „mesterségesen”. Mind a hat előbbi változathoz mindkettőt csatolnunk kell; ki kell tehát számítanunk 12 együtthatót. Ez nem valami vigasztaló perspektíva, de mondjuk, hogy megtesszük. És mit tudunk meg mindebből a kapcsolatról? A sok közül melyik számmal jellemezzük? Mit teszünk, ha a kapott együtthatók egyike-másika szignifikáns, a többi pedig nem?

Így hát jobb, ha az adatok különbözőségét szigorúan megköveteljük. Ha ez nem teljesül, számítsuk ki inkább az r, vagy az r_S együtthatót (a képletekre történő hivatkozásokat már korábban megadtuk): az elsőt egyáltalán nem zavarják az egyforma adatok, a másodikra vonatkozóan pedig jó korrekció ismeretes.

Az egyforma adatok problémája nem jelenti azt, hogy az *elvvel* lenne baj, ahogyan a Kendall-féle együttható méri a kapcsolatot; a *végrehajtás technikájába* nem férnek bele az egyformák. Más esetben ez semmi problémát nem okoz; erről lesz szó az utolsó szakaszban.

Kapcsolatmérés kontingenciatáblázatban

Ebben a részben feltételezzük, hogy az olvasó tisztában van a kontingenciatáblázat fogalmával, felépítésével és alkalmazási területeivel – legalább azon a szinten, ahogyan ez *(11)*-ben megtalálható. (Az egyes fogalmakat újabb hivatkozás nélkül fogjuk használni.) A táblázatból számolható együtthatók közül itt csak eggyel foglalkozunk,[7] amelyiknek a képlete *megegyezik* a Kendall-féle együttható első, definiáló képletével. Az együttható nevét tekintve nehéz helyzetben vagyok, mert az irodalomban nincs használható megállapodás: szinte ahány könyv, annyi elnevezés és annyiféle jelölés.

Azt hiszem, legjobb, ha azt választom, amit egyetemi előadásaimban használtam. (Nem mintha túlzottan egyetértenék vele…) Elképzelhető ugyanis, hogy olvasóim egy része hallotta ezeket az előadásokat, netán megvan még valahol a jegyzete; legalább őket ne zavarjuk össze. Eszerint az együttható jele a görög ψ (pszí) betű, elnevezése pedig: *asszociációs együttható.* Csak akkor alkalmazható, ha mind a soroknak, mind az oszlopoknak van *természetes sorrendje.* Ez nincs mindig így; az annak idején példaként emlegetett táblázat (a foglalkozások és betegségek összefüggéseinek vizsgálatára) *nem* ilyen: sem a foglalkozásoknak, sem a betegségeknek nincs egyértelmű sorrendje. (Hiába is definiálnánk: annak a sorrendnek nem lenne semmiféle jelentése.)

Sokszor azonban *van* természetes, magától értetődő elrendezés. Legegyszerűbb persze az az eset, ha ez valamilyen *nagyság szerinti* sorrend. Például olyankor, amikor mérhető változót osztunk be „kategóriákba”. (Példa: a magasságmérés centiméteradatait nem használjuk föl, hanem három csoportba – alacsonyak, középtermetűek, magasak – osztjuk a vizsgált személyeket.) Kézenfekvő ez a megközelítés, ha a változónak csak kevés különböző értéke fordul elő; rendszerint ez a helyzet például az alkalmazott *dózisok* esetében. De nemcsak a nagyság képezheti a kategóriák természetes sorrendjének alapját. Lehet például valamilyen *gyakoriság* (milyen gyakran vannak rohamai: naponta, hetente, havonta), *súlyosság, fontosság, színárnyalat* (világos–sötét) stb. (Az ezekhez rendelt mérőszámok persze ilyenkor is nagyság szerinti elrendezést határoznak meg.)

A jelölés és a szóhasználat igyekszik a dolgozat első részében használthoz alkalmazkodni; ezért beszélünk x-ről és y-ról, A és B változó helyett. Az x változó egyes kategóriáit az *oszlopok,* az y változó értékeit a *sorok* képviselik. (A koordinátarendszerben is *vízszintesen* mérjük föl mindig x-et és *függőlegesen* y-t!)

A továbbiakban egy kiragadott „példára” vonatkoztatva fogalmazzuk meg mondanivalónkat, mert az általános megfogalmazás nagyon nehézkes. A „példa” korántsem számpélda, még csak nem is „szöveges” példa: egyszerűen arról van szó, hogy egy olyan kontingenciatáblázatról fogunk beszélni, amelynek három sora és négy oszlopa van *(6. táblázat)*. A módszert erre vonatkozóan vezetjük be – de senkinek nem fog nehézséget okozni, hogy (ha szüksége van rá) tetszőleges méretű táblázatban alkalmazza az eljárást.

Az x változó kategóriáit (oszlopok) az **1, 2, 3, 4** jelekkel szimbolizáltuk, ezzel is mutatva, mit értünk azon, hogy egy érték „nagyobb”. Az y változó kategóriáit (sorok) az **I, II, III** római számok jelzik; ezekre az „inkább” határozót alkalmazzuk. (Lehetnek például a súlyosság mutatói.)

Pozitív kapcsolatra az utal, ha „mennél nagyobb, annál inkább”. (Összhangban a τ együttható beve-

6. táblázat. Egy háromsoros, négyoszlopos kontingenciatáblázat általános alakja. *(A gyakoriságokat a nehézkes, indexes betűk helyett egyszerűbben jelöltük. Az ábécé végére ugrás arra akar utalni, hogy amit erről a táblázatról mondunk, az sokkal nagyobb kontingenciatáblázatokra is érvényes.)*

***x* értékei**

***y* értékei**

zetésekor használt megfogalmazással; lásd az első szakaszt!) A negatív kapcsolat például így fogalmazható: „mennél nagyobb, annál kevésbé”. Az első tendenciát – ismét az első szakaszra rímelve – a ↓↓ párok, a másodikat a ↓↑ párok képviselik. Ezeket a párokat kell összeszámlálnunk, és az asszociációs együttható képletébe behelyettesítenünk.

Lássuk először az „ugyanúgy rendezett” párokat. Kiragadva egy személyt az első sor első cellájából, sem az első sorban, sem az első oszlopban nem kereshetünk hozzá „párt”: akik itt állnak, azoknak vagy x, vagy y „értéke” *ugyanakkora,* mint a kiragadott személyé, tehát nem lehet sem „nagyobb”, sem „inkább”. A nagyobbakat tőle *jobbra,* az „inkábbakat” *lefelé* kell keresnünk. De ezeken a helyeken mindenki ilyen! Így hát kiragadott személyünkkel „ugyanúgy rendezett” párt alkot (és ezzel U értékét gyarapítja) a *jobbra és lejjebb* álló cellákban állók összlétszáma: $(f+g+h+v+w+z)$. Az első sor első cellájához tartozó valamennyi személlyel ugyanennyi párt tudunk képezni, tehát ahhoz, hogy az első cellából kiindulva alkotott „ugyanúgy rendezett” párok számát megkapjuk, az előbbi zárójeles kifejezést a-val kell szorozni. Azután megnézzük, hogy a második cellában állókhoz képest hányat találunk „jobbra és lefelé”, és így tovább. A sor utolsó, d gyakoriságú celláját nem kell vizsgálnunk, hiszen attól „jobbra” senki sem található. Ugyanígy: az utolsó sort meg sem kell néznünk, hiszen attól „lefelé” már semmi sincs. Ezek szerint U-t így számítjuk ki:

$$U = a(f+g+h+v+w+z) + b(g+h+w+z) + \\ + c(h+z) + e(v+w+z) + f(w+z) + gz.$$

A „fordítva rendezett” párok megkeresését a *balra és lejjebb* elve alapján végezzük. (Ez annak felel meg, hogy „mennél kisebb, annál inkább”. De eljárhatunk az eredeti fogalmazás – „mennél nagyobb, annál kevésbé” – alapján is; ekkor a *jobbra és följebb* állókat kell összeszámlálnunk. Az eredmény mindkét esetben ugyanaz.) További magyarázkodásra nincs is szükség, rögtön fölírhatjuk F „formuláját”[8]:

$$F = b(e+u) + c(e+f+u+v) + d(e+f+g+u+v+w) + \\ + fu + g(u+v) + h(u+v+w).$$

Eddig csak szóban mondtuk, de most formálisan is fölírjuk az asszociációs együttható képletét:

Vagyis pontosan ugyanaz, mint a Kendall-féle rangkorrelációs együttható első szakaszban fölírt képlete.

Befejezésül érdemes megnézni, hogyan alakul ez a képlet a leggyakrabban előforduló kontingenciatáblázat, a *négymezős táblázat* esetén. A táblázat ilyen alakú:

x értékei

y értékei

A gyakoriságokat a szokásos módon jelöltük. Az előbb alkalmazott elvek alapján kapjuk, hogy $U = ad$ és $F = bc$.

Az asszociációs együttható eszerint:

.

* * *

Jó ez a II-es jelzés a dolgozat címében! Bizonyítja, hogy valóban *sorozatról* van szó: az első rész után elkészült a második is. De nem kötelezi semmire sem a szerzőt: ha van kedve és ideje – és főleg, ha megfelelő témát talál – folytathatja is a sorozatot (ebben, vagy legkésőbb a következő évben), de akár abba is hagyhatja. Én is így járok el: nem adok meg, nem ígérek semmiféle témát, mint a dolgozat I. részének végén, hanem azzal búcsúzom, hogy ha időm engedi – és főleg, ha valahonnan valami érdeklődésféle jutna el hozzám –, alkalomadtán folytatom.

JEGYZETEK

1. *Személy* helyett bármilyen más „vizsgálati egység” szerepelhet: a személytől levett vizsgálati minta, laboratóriumi anyag, sejttenyészet, kísérleti állat, vagy akár olyan általánosabb fogalom, mint egy kórházi osztály, egy szakma stb. A szöveg egyszerűsítése érdekében azonban mindvégig *személyekről,* és azok *két* (mérhető) *tulajdonságáról* fogunk beszélni.
2. Ilyenkor az x-adatok ismeretében meg tudjuk határozni az y-adatok nagyság szerinti sorrendjét. Az állítás természetesen fordítva is igaz: y-ból ugyanígy következtethetünk x-re.
3. Könnyen elképzelhető, hogy valakit csak zavar a megértés könnyítésére szánt, nyilakat felvonultató szimbolika. Ezért megkísérlem szavakkal is elmondani.
Kiválasztva a mintából két személyt, megvizsgáljuk, hogy x-adataik egymáshoz való viszonya és y-adataik egymáshoz való viszonya megegyezik-e. Ha mindkét változóban a pár első egyedének adata nagyobb, vagy ha mindkettőben az elsőé kisebb, akkor „ugyanúgy rendezett” párról beszélünk. Ha viszont az egyik

változóban a pár első, a másikban a pár második egyedének adata a nagyobb, a pár „fordítva rendezett”.

4. Nem tudni, miért, Kendall a görög τ-t használta együtthatója jelölésére *(5)*; ezt a hagyományos jelölést követjük mi is – ellentétben a *(4)*-ben mondottakkal.
5. Csak arra kell vigyázni, hogy „oda-vissza” ne kanyarodjon a vonal. Nézzük meg a *3. táblázat* a) részét. Ha ide berajzolnánk a vonalakat, csupa vízszintes vonalat kapnánk. Metszéspont nincs (F=0, ezt tudtuk is), de ha a vonalakat föl-le kanyarogva vezetnénk, jó néhány mesterséges metszéspontot sikerülne létrehozni.
6. Itt érzékelhető, hogy a statisztikai próba „ereje” mennyire függ az elemszámtól! Kis minta esetén olyan sokszor fordul elő „véletlenül” szélsőséges sorrend, hogy még a „legszorosabb kapcsolat” sem elég a szignifikancia kimondásához. Hiába 1 a Kendall-féle együttható abszolút értéke, ha n=5, ez sem elég az 1%-os szignifikanciához, ha pedig n=4, még az 5%-os szignifikanciához sem. (Lásd az *5. táblázatot!*)
7. Egy másik, ennél általánosabban alkalmazható együtthatót az idézett dolgozatban ismertettünk [*(11)*, 684. oldal].
8. Ez persze nem formula, hanem a kiszámítás szimbolikus jelölése. Azért használtuk a formula kifejezést, mert számok helyett betűket alkalmaztunk; így jobban látszik, mi történik, és általánosítani is könnyebb.

IRODALOM

1. *Hajtman Béla.* A biostatisztika alapjai II. *Lege Artis Medicinæ 1999;(9):121.*
2. *Hajtman Béla.* A biostatisztika alapjai III. *Lege Artis Medicinæ 1999;(9):210-1.*
3. *Hajtman Béla.* A biostatisztika alapjai VII. *Lege Artis Medicinæ 1999;(9):567.*
4. *Hajtman Béla.* A biostatisztika alapjai VII. *Lege Artis Medicinæ 1999;(9):568.*
5. *Kendall GM.* Rank correlation methods. *London: Griffin; 1948.*
6. *Wilcoxon F, Wilcox RA.* Some rapid approximate statistical procedures. *Pearl River: Lederle Laboratories; 1964. p. 13-4.*
7. *Owen DB.* Handbook of statistical tables. *Reading: Addison-Wesley; 1962. p. 396-9.*
8. *Hajtman Béla.* Egyszerűen végezhető statisztikai eljárások I. *Lege Artis Medicinae 2001;(11):380.*
9. *Hajtman Béla.* A biostatisztika alapjai V. *Lege Artis Medicinæ 1999;(9):397-8.*
10. *Hajtman Béla.* Egyszerűen végezhető statisztikai eljárások I. *Lege Artis Medicinæ 2001;(11):379.*
11. *Hajtman Béla.* A biostatisztika alapjai VIII. *Lege Artis Medicinæ 1999;(9):680-5.*