Vargha András,
a Károli Gáspár Református Egyetem és az Eötvös Loránd Tudományegyetem Pszichológiai Intézetének egyetemi tanára
E-mail: vargha.andras@kre.hu

Szignifikanciatesztek – negyven éve hibás elemzéseket végzek és téveszméket tanítok?

DOI: 10.20311/stat2016.04.hu445

Az empirikus vizsgálatokon, kutatásokon alapuló tudományos következtetéseknek csaknem száz éve fő módszere a biológiában, orvostudományban, pszichológiában és számos társadalomtudományban a statisztikai hipotézisvizsgálat, amelyet ún. *statisztikai próbák*, más elnevezéssel *szignifikanciatesztek* segítségével végzünk. Szokásosan statisztikai próbák segítségével döntünk arról, hogy egy szernek vagy terápiás technikának van-e gyógyító hatása, hogy szakmailag fontos csoportok (például férfiak és nők) között van-e bizonyos releváns szakmai változók tekintetében különbség, hogy ilyen-olyan változók függetlenek-e egymástól stb.

Ezt figyelembe véve szinte sokkszerűen robbant a statisztikai közéletbe a hír 2015 elején, hogy a neves amerikai folyóirat, a *Basic and Applied Social Psychology* (impakt faktora: 1,168) ezentúl nem fogad el szignifikanciateszteken alapuló kutatási eredményekről szóló cikkeket, mert ezt a statisztikai következtetési módszert érvénytelennek tartják (*Trafimow–Marks* [2015]). Persze nem előzmény nélküli a szignifikanciatesztek kritikája (például *Lykken* [1968], *Cohen* [1994], *Harlow* [1997]). Újabban *Cumming* [2014] foglalta össze igen logikusan és tetszetősen a szignifikanciateszteken alapuló következtetések hibáit és ezen irodalmak sorába illeszkedik, immár magyar nyelven *Bárdits–Németh–Terplán* [2016] cikke is a *Statisztikai Szemle* 2016. évi első számában. Egyébként meg kell jegyeznem, hogy a http://psych.colorado.edu/~willcutt/resmeth_signif.htm honlapról több mint húsz alapvető irodalom tölthető le a szignifikanciatesztekkel kapcsolatban.

Jelen rövid cikkemben nem szeretnék ezekből összefoglalót írni vagy felhasználásukkal egy (n + 1)-edik cikket gyártani. Inkább önvizsgálatot tartanék: több mint negyvenéves statisztikaoktatói, -kutatói és -alkalmazói tevékenységem hogyan értékelhető az újabb fejlemények tükrében? Én és statisztikus kollégáim mit csináltunk jól, illetve mit csinálnánk ma másképp?

1. A statisztikai hipotézisvizsgálat modellje

A statisztikai hipotézisvizsgálat elméleti modellje szilárd matematikai alapokon áll. E módszer szerint megfogalmazunk egy releváns matematikai modellt rögzített paraméterértékekkel (nullhipotézis), és empirikus adatok alapján valószínűségi döntést hozunk arról, hogy a modell megállja-e a helyét (nullhipotézis megtartása), vagy inkább egyéb paraméterértékekhez köthető, másik modellt (ellenhipotézis) preferáljunk (nullhipotézis elutasítása). Statisztikai hipotézisvizsgálatot rendszerint statisztikai próbák segítségével végzünk. Ha a vizsgált nullhipotézis a statisztikai próba során elutasítható, a próbát, illetve annak eredményét szignifikánsnak mondjuk. Tekintve, hogy a nullhipotézisekkel szemben megfogalmazott ellenhipotézisek releváns szakmai konzekvenciákat képviselnek (például szakmai hatások érvényesülése, változók összefüggése, fontos csoportok különbözősége), a statisztikai elemzések során a szignifikáns eredményeknek örülni szoktunk. Mi a rossz ebben? Az 1. táblázatban például néhány antropometriai változó Pearson-korrelációs mátrixa látható, amelyet egy $N = 500$ fős reprezentatív mintán számítottunk ki.

1. táblázat

Néhány antropometriai változó Pearson-korrelációs mátrixa (N = 500)

Változó	Születési súly	10 éves kori testsúly	Születési testhossz	10 éves kori testmagasság
Születési súly	1	0,164***	0,788***	0,236***
10 éves kori testsúly	0,164***	1	0,228***	0,662***
Születési testhossz	0,788***	0,228***	1	0,327***
10 éves kori testmagasság	0,236***	0,662***	0,327***	1

Megjegyzés. *** $p < 0{,}001$.

Az 1. táblázatban minden korreláció igen magas szinten ($p < 0{,}001$) korrelál, ami alapján szinte hiba nélkül ($p < 0{,}001$) kijelenthetjük: a születési és a tízéveskori testsúly, valamint a testhossz/testmagasság lineáris kapcsolatban van egymással. Természetesen elkerüljük azt a gyakorlatban nem ritkán elkövetett hibát, hogy a szignifikancia erősségét a kapcsolat szorosságaként interpretáljuk (vö. *Harlow* [1997], *Kline* [2004], illetve *Bárdits–Németh–Terplán* [2016] 1.3. alfejezet). Ezzel a következtetéssel hibát nem követtünk el, ám mégse mentünk vele sokra, mert szakmai trivialitást állapítottunk meg, vagyis alacsony információértékű kijelentést tettünk. Ez pedig főként abból fakad, hogy a statisztikai hipotézisvizsgálat, a szembeállított null- és ellenhipotézis dichotóm gondolkodásra, vagyis annak eldöntésére sarkall, hogy a két lehetséges modell közül melyik az igaz.

2. Hipotézisvizsgálat helyett becslések

A nullhipotézisek rendszerint nem igazak, ezért statisztikai elemzéseinkben nem ezek elutasítására kellene a hangsúlyt fektetnünk, hanem arra, hogy mekkora, milyen mértékű a nullhipotézisnek megfelelő modelltől való eltérés. Vagyis, ahogy *Cumming* ([2014] 14. old.) megfogalmazza: „Már magukat a tudományos kérdéseket is a becslés fogalomkörében fogalmazzuk meg. Ennek megfelelően így tegyük fel a kérdést: »Mekkora a hatás?« vagy »Milyen mértékű…?« Kerüljük az olyan dichotóm kifejezéseket, hogy »ellenőrizzük azt a hipotézist, hogy nincs különbség« vagy »Ez a kezelés jobb?«"

A szignifikanciatesztek helyett tehát inkább becslési eljárásokat, többnyire intervallumbecslést célszerű alkalmazni. Az 1. táblázat változói közötti korrelációk nagyság szerint sorba rendezett listája a 2. táblázatban látható, az elméleti értékek 99 százalékos intervallumbecslésével (C99) kiegészítve (N = 500).

2. táblázat

Az 1. táblázat változói közötti korrelációk az elméleti értékek 99 százalékos intervallumbecslésével kiegészítve (N = 500)

X változó	*Y* változó	Pearson-féle korrelációs együttható (*r*)	C99
Születési súly	Születési testhossz	0,788	(0,741; 0,828)
10 éves kori testsúly	10 éves kori testmagasság	0,662	(0,592; 0,722)
Születési testhossz	10 éves kori testmagasság	0,327	(0,220; 0,425)
Születési súly	10 éves kori testmagasság	0,236	(0,125; 0,341)
10 éves kori testsúly	Születési testhossz	0,228	(0,116; 0,334)
Születési súly	10 éves kori testsúly	0,164	(0,050; 0,273)

A 2. táblázatból nemcsak azt állapíthatjuk meg, hogy a vizsgált változók korrelációi nagy bizonyossággal mind pozitívak, hanem a nagyságrendjükről és a közöttük levő különbségekről is tájékozódhatunk. Például megállapíthatjuk, hogy a születési súly és a teszthossz közötti elméleti korreláció 99 százalékos megbízhatósággal a 0,741-től 0,828-ig terjedő, míg a tízéveskori súly és a testmagasság közötti a 0,592–0,722-es intervallumban helyezkedik el. Ezen elméleti értékek különbségeire is lehet intervallumbecslést készíteni (lásd *Zou* [2007], illetve https://seriousstats.wordpress.com/2012/02/05/comparing-correlations/). Az a statisztikai következtetés, hogy például az elméleti korreláció 99 százalékos megbízhatósággal 0,741 és 0,828 között keresendő, egyben azt is jelenti, hogy az elméleti korreláció 99 százalékos megbízhatósággal

nemcsak 0-nál, hanem 0,740-nél is nagyobb. Ez pedig lényegesen informatívabb állítás, mint a 0,788-as korreláció szignifikanciájának szimpla megállapítása bármilyen szinten.

Átlagok esetén a statisztikai hipotézisvizsgálatok helyett ugyancsak célszerűnek látom a becslésekre való áttérést, pontbecslés esetén persze kiszámítva (becsülve) a becslés pontosságát jelző standard hibát is. Ez a hangsúlyáthelyezés a statisztikai próbákról a becslésekre egyben jelzi az elemszám fontosságát is a statisztikai következtetésekben, hiszen más körülmények rögzítése mellett a mintanagyság növelése az egyetlen eszköz a standard hiba, illetve a konfidencia-intervallum szélességének csökkentésére.

3. A kutatás teljes vertikumát átfogó tanácsok

Az eddig mondottakból világosan kitűnik, hogy a legújabb irányzatok alapján a statisztikai elemzésekben el kell kerülnünk az automatikus döntéseket. A statisztikai elemzések eredményeit a statisztikai keretek és feltételek, valamint a teljes szakmai háttér figyelembevételével szabad csak értelmezni, ami a statisztikai mellett természetesen nagy szaktudást is igényel, és komoly felelősséget ró a szakmai következtetésekre jutó szakemberre. Ezzel a statisztikai adatfeldolgozás természetesen kilép abból a szokásos keretből, hogy a rutinszerűen alkalmazott statisztikai próbákból automatikus szakmai következtetéseket vonunk le. *Cumming* ([2014] 14. old.) például a következő nyolc pontban foglalja össze e kitágult kutatási koncepció főbb összetevőit.

– Kutatási kérdések megfogalmazása becslési terminusokban.

– A kutatási kérdésnek legmegfelelőbb hatásmértékek azonosítása.

– A kutatással kapcsolatos összes részlet előzetes megtervezése elemszámokkal és elemzési módszerekkel együtt.

– A kutatás elvégzése után a kiválasztott hatásmértékekre pontbecslések és konfidencia-intervallumok megadása.

– A konfidencia-intervallumok ábrázolása értelmes grafikonokon.

– A hatásmérték nagyságának és a konfidencia-intervallumok szélességének értelmezése szakmai szempontból is, és ezekből a kutatás céljával összhangban levő elméleti és gyakorlati következtetések levonása.

– Gondolkodás, ahol lehet, a metaanalízis modelljében. A kutatás és az elemzések végrehajtása úgy, hogy eredményeinket könnyű legyen integrálni a hasonló kutatások metaanalízisébe.

– A kutatás leírásakor (a nyers adatokkal együtt) minden releváns információ megadása más kutatók számára.

A metaanalízis előbbi hangsúlyozása azt a régebben is sokak által vallott nézetet képviseli, miszerint egy új kutatáson alapuló új tudományos eredményt addig nem szabad elfogadni, amíg azt független vizsgálatban mások meg nem erősítik.

4. Személyorientált módszerek

Ez az új adatelemzési szemlélet teljesen illeszkedik ahhoz a kutatási megközelítéshez, amely felhívja a figyelmet a hagyományos változóorientált kutatási paradigma gyengeségeire, és ehelyett, illetve mellette a személyorientált módszerek alkalmazását javasolja (*Bergman–Magnusson* [1997], *Bergman–Vargha* [2013]). E módszerek közé tartoznak olyan komplex elemzések is, mint a többváltozós klasszifikációk közé sorolható klaszteranalízis, konfiguráció- vagy sűrűsödéspont-elemzés stb. (lásd *Vargha–Torma–Bergman* [2015]), de a személyorientált módszerek egyszerűbb esetei nagyon jól illeszkednek a jelen cikk gondolatmenetébe, mint azt a következő példák illusztrálják.

1. Sima hatásvizsgálatot végzünk kvantitatív függő változóval (X). Két időpontban mérjük X értékét (X_1 és X_2) és arra vagyunk kíváncsiak, hogy van-e javulás, s ha van, akkor mekkora a második helyzetben (X_2) az elsőhöz (X_1) viszonyítva. A változóorientált szemlélet a $Z = X_2 - X_1$ különbségre koncentrál, s ha elvetjük is a Z szignifikanciájára vonatkozó (egymintás t-próbával végrehajtható) hagyományos elemzést, a Z elméleti átlagára vonatkozó konfidencia-intervallum szerkesztése is csak a két változó globális különbségének nagyságszintjéről fog tájékoztatni. Ha viszont a személyorientált szemlélet hívei vagyunk, az egymintás t-próba helyett (vagy mellett) inkább az előjelpróbát választjuk (lásd *Vargha* [2007] 8.4. alfejezet), amelynek során X_1 és X_2 különbségét nem az átlagaik különbségével jellemezzük, hanem az $X_1 < X_2$, $X_1 > X_2$ arányokkal. Ha például az jön ki, hogy az $X_1 < X_2$ (vagyis az X változó növekedése) domináns az $X_1 > X_2$ csökkenéssel szemben, mégis lesz információnk arról, hogy a globális tendenciának ellentmondó, ellentétes változás a vizsgált személyek mekkora arányára lesz igaz. Ezekre az arányokra természetesen konfidencia-intervallumok is szerkeszthetők hagyományos statisztikai módszerekkel.

2. Ugyanaz a szakmai probléma, mint az előbbi esetben, azzal a különbséggel, hogy nem összetartozó, hanem független minták állnak rendelkezésre. A változóorientált szemlélet itt is a $Z = X_2 - X_1$ különbségre koncentrál, s ha elvetjük is a *Z* szignifikanciájának (kétmintás *t*-próbával végrehajtható) hagyományos elemzését, a *Z* elméleti átlagára vonatkozó konfidencia-intervallum szerkesztése, valamint az ilyenkor szokásosan kiszámított Cohen-féle delta hatásmérték (lásd *Vargha* [2007] 247. old.) is csak a két változó globális különbségének nagyságszintjéről fog tájékoztatni. Ha viszont a személyorientált szemlélet hívei vagyunk, a kétmintás *t*-próba helyett (vagy mellett) inkább a rangsorolásos Mann–Whitney-próbát választjuk (lásd *Vargha* [2007] 10.1. alfejezet), amelynek során X_1 és X_2 különbségét nem az átlagaik különbségével jellemezzük, hanem az A_{12} valószínűségi fölény mutatóval, amely jelzi, hogy $X_1 > X_2$ milyen arányban fordul elő. Ha például $X_1 < X_2$ domináns az $X_1 > X_2$ relációval szemben (vagyis a második csoportbeliek többnyire nagyobb értékűek, mint az elsők), mégis lesz információnk arról, hogy a globális tendenciának ellentmondó, ellentétes állapot (amikor is egy első csoportbeli nagyobb értékű, mint egy második csoportbeli) milyen gyakran lesz igaz. Az A_{12} mutató elméleti nagyságára konfidencia-intervallum is szerkeszthető (lásd *Vargha* [2007] 10.4. alfejezet).

3. Két kvantitatív változó (*X* és *Y*) kapcsolatának szorossága érdekel bennünket. A változóorientált szemlélet szerint ilyenkor hagyományosan a Pearson-korrelációt szoktuk használni, amely jelzi a két változó közötti kapcsolat irányát és szorosságát. Ha a személyorientált szemlélet hívei vagyunk, a Pearson-korreláció helyett (vagy mellett) inkább a Kendall-féle tau rangkorrelációs együtthatót célszerű kiszámítani, amely a Pearson-korreláció által jelzett kapcsolati irány és szorosság mellett azt is jelzi, hogy *X* és *Y* értékeinek pozitív együttjárása (az ún. konkordancia) mennyivel gyakrabban fordul elő, mint a negatív együttjárás (az ún. diszkordancia; vö. *Vargha* [2007] 12.2. alfejezet). Például egy tau = 0,60 érték arról tájékoztat, hogy a pozitív kapcsolat (amikor is nagyobb *X*-érték nagyobb *Y*-értékkel jár együtt) 60 százalékkal gyakrabban lép fel *X* és *Y* viszonylatában, mint a negatív kapcsolat. Ha *X* és *Y* folytonos, akkor ebből már kikövetkeztethető, hogy a pozitív kapcsolat aránya 80 százalék, a negatívé pedig 20 százalék. Ha *X* és/vagy *Y* nem folytonos, a konkordancia és a diszkordancia aránya külön becsülhető.

Mindhárom esetben arról van szó, hogy ha a hagyományos paraméteres módszer helyett egy alkalmas nemparaméteres (többnyire rangsorolásos) statisztikai módszert

alkalmazunk, akkor a globális trenddel ellentétes, kisebbségben lévő, de mégsem elhanyagolható állapot (körülmény) gyakoriságáról is információt kapunk, amely a hagyományos változóorientált paraméteres módszerek esetében teljesen negligálódik. Ilyen nemparaméteres elemzések gazdag választéka áll rendelkezésre például a ROPstat statisztikai programcsomagban (lásd *Vargha* [2007], illetve www.ropstat.com).

Irodalom

BÁRDITS A. – NÉMETH R. – TERPLÁN GY. [2016]: Egy régi probléma újra előtérben: a nullhipotézis szignifikanciateszt téves gyakorlata. *Statisztikai Szemle*. 94. évf. 1. sz. 52–75. old. http://dx.doi.org/10.20311/stat2016.01.hu0052

BERGMAN, L. R. – MAGNUSSON, D. [1997]: A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*. Vol. 9. No. 2. pp. 291–319.

BERGMAN, L. R. – VARGHA, A. [2013]: Matching method to problem: A developmental science perspective. *European Journal of Developmental Psychology*. Vol. 10. No. 1. pp. 9–28. http://dx.doi.org/10.1080/17405629.2012.732920

CUMMING, G. [2014]: The new statistics: Why and how. *Psychological Science*. Vol. 25. No. 1. pp. 7–29. http://dx.doi.org/10.1177/0956797613504966

COHEN, J. [1994]: The earth is round ($p < .05$). *American Psychologist*. Vol. 49. No. 12. pp. 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

HARLOW, L. [1997]: *What If There Were No Significance Tests*? Lawrence Erlbaum Associates. Mahwah.

KLINE, R. B. [2004]: *Beyond Significance Testing*: *Reforming Data Analysis Methods in Behavioral Research*. American Psychological Association. Washington, D.C. http://dx.doi.org /10.1037/10693-000

LYKKEN, D. T. [1968]: Statistical significance in psychological research. *Psychological Bulletin*. Vol. 70. No. 3. pp. 151–159. http://dx.doi.org/10.1037/h0026141

TRAFIMOW, D. – MARKS, M. [2015]: Editorial. *Basic and Applied Social Psychology*. Vol. 37. Issue 1. pp. 1–2. http://dx.doi.org/10.1080/01973533.2015.1012991

VARGHA A. [2007]: *Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal*. *2. kiadás*. Pólya Kiadó. Budapest.

VARGHA A. [2008]: Új statisztikai módszerekkel új lehetőségek: a ROPstat a pszichológiai kutatások szolgálatában. *Pszichológia*. 28. évf. 1. sz. 81–103. old. http://dx.doi.org/10.1556/Pszi.28.2008.1.5

VARGHA, A. – TORMA, B. – BERGMAN, L. R. [2015]: ROPstat: A general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research*. Vol. 1. No. 1–2. pp. 87–98.

ZOU, G. Y. [2007]: Toward using confidence intervals to compare correlations. *Psychological Methods*. Vol. 12. No. 4. pp. 399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399