Bárdits Anna,
az Eötvös Loránd Tudományegyetem hallgatója
E-mail: barditsanna@gmail.com

Németh Renáta,
az Eötvös Loránd Tudományegyetem docense
E-mail: nemethr@tatk.elte.hu

Terplán Győző,
az Eötvös Loránd Tudományegyetem hallgatója
E-mail: terplangyozo@caesar.elte.hu

Egy régi probléma újra előtérben: a nullhipotézis szignifikanciateszt téves gyakorlata*

DOI: 10.20311/stat2016.01.hu0052

Tanulmányunkban az empirikus adatelemzés egyik sarokkövének, a nullhipotézis szignifikanciateszt inherens problémáit és alkalmazásának hibáit foglaljuk össze. A teszttel kapcsolatos régi vita áttekintésének apropóját a *Basic and Applied Social Psychology* (Alap- és Alkalmazott Szociálpszichológia) folyóirat 2015. évi szerkesztőségi állásfoglalása adja, mely megtiltja a szignifikanciatesztek használatát a publikálni vágyók számára. Jelen összefoglalóból kiderül, hogy a probléma nem újkeletű, több mint 40 éve kritika tárgyát képezi. Megközelítésünk inkább tudományelméleti, nem alakítunk ki a vita minden kérdésével kapcsolatban állásfoglalást, de ismertetünk néhány általunk is tévesnek ítélt gyakorlatot. Emellett kitérünk a szignifikanciatesztek használatának tudományterület-specifikusságára és a helytelen alkalmazások tudományszociológiai gyökereire is. A *Szociológia Szemle* cikkeinek vizsgálatával rámutatunk arra, hogy a magyarországi empirikus szociológiában is azonosíthatók a téves gyakorlat egyes jegyei, melyek elkerülésére több ajánlást teszünk.

Tegyük fel, hogy egy húszelemű kísérleti és egy húszelemű kontrollcsoportban vizsgáljuk az átlagok egyenlőségére vonatkozó nullhipotézist (nincs különbség a

* A cikk *Bárdits Anna* survey statisztika mesterszakos hallgató szakdolgozatán alapul; konzulense *Németh Renáta*.

populációs átlagok között), kétmintás *t*-próbával. A *t*-próbastatisztika eredménye $t = 2{,}7$, az empirikus szignifikancia $p = 0{,}01$. Melyik állítás igaz ekkor a következők közül?

1. A nullhipotézist maradéktalanul cáfoltuk.

2. Megtaláltuk annak a valószínűségét, hogy a nullhipotézis igaz.

3. A kísérleti hipotézist (vagyis hogy van különbség a populációs átlagok között) maradéktalanul bizonyítottuk.

4. Az eredmények alapján ki tudjuk számolni annak a valószínűségét, hogy a kísérleti hipotézisünk igaz.

5. Ha elutasítjuk a nullhipotézist, ismerjük a valószínűségét annak, hogy rossz döntést hozunk.

6. Megbízható kísérleti eredményünk van abban az értelemben, hogy ha sokszor megismételnénk a kísérletet, az esetek 99 százalékában szignifikáns eredményt kapnánk.

A felsorolásból egyik állítás sem igaz. Ám, ha tévedtünk, ne higgyük azt, hogy egyedül vagyunk ezzel. *Haller–Krauss* [2002] német egyetemeken tanító kutatópszichológusoknak tette fel ugyanezt a kérdést. A megkérdezett 30, módszertani tárgyat oktató tanár 80 százaléka jelölte valamelyik állítást igaznak. Ugyanez az arány a nem módszertani jellegű tárgyat oktatók között 90 százalék (!) volt. Ezek az eredmények azt mutatják, hogy még a szignifikanciateszt gyakori használói, sőt oktatói számára sem teljesen világos, mit jelent pontosan az alacsony *p*-érték. Mivel a *p* viszonylag keveset mond, erős az a vágy, hogy többet lássunk bele. De nemcsak a gyakori félreértelmezés jelent gondot: sokak szerint inherens problémákkal is küzd a teszt.

Az amerikai „*Basic and Applied Social Psychology*" (Alap- és alkalmazott szociálpszichológia) című tudományos folyóirat 2015. évi 1. számának szerkesztői bevezetője (*Trafimow–Marks* [2015]) a következővel kezdődött:

„A *Basic and Applied Social Psychology* (*BASP*) 2014. évi szerkesztőségi állásfoglalása hangsúlyozta, hogy a nullhipotézis szignifikanciateszt (nhszte) nem alátámasztott, ezért szerzőinktől a továbbiakban nem követeljük meg annak használatát (*Trafimow* [2014]). Azonban, türelmi időt adva szerzőinknek, szerkesztőségünk nem tiltotta be azonnal az nhszte-t. Jelen bevezető célja viszont, hogy bejelentse, a türelmi idő lejárt. A továbbiakban a *BASP* betiltja az nhszte használatát." 1. old.

A *Basic and Applied Social Psychology* emellett tiltja a konfidencia intervallumok használatát is. A szerkesztők semmilyen statisztikai következtetési módszer

használatát nem követelik meg a szerzőktől, mivel „a kifogástalan módszer továbbra sem ismert". Bár a többváltozós módszerek használatával külön nem foglalkoznak, feltételezhető, hogy a többi iránymutatás figyelembe vételével lehet ezeket használni. A folyóiratnak ez a lépése nem előzmény nélküli a tudományos folyóiratok történetében. Az Egészségügyi Folyóiratok Szerkesztőinek Nemzetközi Bizottsága által 1988-ban kiadott állásfoglalás szerint a kutatók nem támaszkodhatnak kizárólag a hipotézistesztekre és különösen nem a *p*-értékekre. *Fidler et al*. [2004] folyóirat-kutatása szerint az 1980-as évek végén *Kenneth Rothman* orvos előbb az „*American Journal of Public Health*" (Amerikai Közegészségügyi Folyóirat), később az „*Epidemiology*" (Epidemiológia) című folyóirat szerkesztőjeként a hozzá kerülő cikkekből töröltette a szignifikanciateszteket, és csak a konfidencia intervallumokat engedte közölni. A nullhipotézisteszt alkalmazása tehát több évtizede fel-fellángoló kritikák tárgya, a *Basic and Applied Social Psychology* lap tiltása ennek csak egy újabb, bár a korábbiaknál szélsőségesebb állomása.

Írásunkban először áttekintjük a módszer használatát övező kérdéseket, bemutatva több eltérő álláspontot is. Majd megvizsgáljuk, hogy a különböző társadalom- és viselkedéstudományi diszciplínákban, illetve az orvostudományban milyen mértékben vannak jelen a módszerrel kapcsolatos hibák. Kitérünk a hazai szociológiai publikációs gyakorlatra is. Később az áttekintésből levonható tanulságként néhány ajánlást fogalmazunk meg, végül pedig visszatérünk a *Basic and Applied Social Psychology* folyóirat tiltására, tárgyalva annak következményeit.

1. A nullhipotézis szignifikanciateszt használatának kritikái

A következőkben a nullhipotézis szignifikanciateszt alkalmazói által elkövetett leggyakoribb 10 hibát vesszük sorba. A kritikák legtöbbje már évtizedekkel ezelőtt írott cikkekben megjelenik, de emellett az újabb forrásokban is megtalálható. Jól mutatja, hogy milyen széles körben elterjedtek a kritikák, hogy több cikk szerzője már a bevezetőjében megemlíti, észrevételeivel nincs egyedül, hiszen már számos kutató leírta a nullhipotézis szignifikanciavizsgálattal kapcsolatos problémáit (például *Ziliak–McCloskey* [2008]).

1.1. A felcserélt feltétel problémája

A nullhipotézisteszt során arra a kérdésre kaphatunk választ, hogy igaz nullhipotézis (H_0) esetén mennyire valószínű, hogy az adott adatokat (A) tapasztal-

juk. Tehát a *p*-érték segítségével a $P(A|H_0)$ nagyságára derül fény, pedig annak valószínűségére lennénk kíváncsiak, hogy a nullhipotézis igaz-e (feltéve, hogy az adott adatokat kaptuk ($P(H_0|A)$). *Oakes* [1986] eredményei alapján a vizsgált pszichológusok közel 40 százaléka élt abban a hitben, a *p*-érték annak a valószínűségét mutatja, hogy a nullhipotézis igaz. *Ziliak–McCloskey* [2008] szerint pedig a közgazdászok szinte 100 százaléka elköveti azt a hibát, hogy a kis *p*-értékből automatikusan a nullhipotézis alacsony valószínűségére következtet.

1.2. A nullhipotézis sosem igaz, így könnyen elutasítható

További kritika a nullhipotézis-vizsgálattal szemben, hogy a nullhipotézis szigorúan véve sosem igaz. Gondoljunk többek között az IQ[1], az iskolai teljesítmény vagy más (például attitűd-) változók vizsgálatára a társadalom különböző csoportjaiban (nők és férfiak, városban és községben élők stb. között) (*Meehl* [1978]). Nehéz elképzelni, hogy bármilyen pontos állítás (például „A nők és a férfiak átlagos IQ-ja közötti különbség 2 pont.") igaz lenne.

A társadalomtudományokban elméletünket legtöbbször a következőképpen tesztelük: (lehetőleg valószínűségi) mintát veszünk a populációból, és az azon mért adatokból következtetünk arra, hogy a populációs paraméter megegyezik-e azzal, mint ami az elméletből következne. A nullhipotézis (például „Nincs különbség a csoportok között.") elutasítása tipikusan azzal jár, hogy feltevésünk valamilyen alátámasztást nyer, hiszen a hipotézisteszt logikája indirekt; teóriánkat vagy annak következményeit pedig egy alternatív hipotézisben fogalmazzuk meg. A probléma az, hogy elgondolásunkat ezzel a módszerrel nem tesszük ki a cáfolat valódi veszélyének, hiszen igazolásához csak a könnyen elutasítható nullhipotézis elvetése szükséges. *Meehl* [1978] szerint olyasmi ez, mintha egy klímára vonatkozó elképzelést annak az egyszerű állításnak a tagadásával próbálnánk bizonyítani, hogy áprilisban nem esik az eső. Ehelyett kívánatosabb lenne, ha a teóriánk alapján lehetőleg minél pontosabb, a lényeget szem előtt tartó előrejelzést adnánk (például április 4-én 0,66 cm csapadék fog esni), és utána azt tesztelnénk.

A nullhipotézis könnyű elutasíthatósága miatt tehát fennáll a veszélye, hogy elméletünk semmitmondó igazolását adjuk. *Meehl* [1978] szerint a probléma igen súlyos: „Úgy gondolom, hogy a szubsztantív elméletek alátámasztásának standard módszereként pusztán a nullhipotézis elutasítására támaszkodni […] egy szörnyű tévedés, alapvetően hibás és szegényes tudományos stratégia, és az egyik legrosszabb dolog, ami valaha a pszichológia történetében előfordult." (817 old.).

[1] IQ (intelligence quotient): intelligenciahányados.

1.3. A szubsztantív szakmai fontosság összetévesztése a statisztikai szignifikanciával

Egy másik kritika (a továbbiakhoz hasonlóan) nem a szignifikanciatesztben rejlő valamely hátránnyal, hanem annak téves használatával, mégpedig a statisztikai szignifikancia és a szubsztantív szakmai fontosság azonosításával szemben fogalmazódott meg. A szubsztantív szakmai fontosság alatt azt értjük, hogy az adott eredménynek a felhasználási terület szempontjából van-e fontossága, elég nagy-e a mért hatásnagyság ahhoz, hogy szakmai szempontból érdemes legyen foglalkozni vele. *Meehl* már az 1967-es cikkében azt írta, hogy a nullhipotézis szignifikanciateszttel kapcsolatos problémákra ellenszer lehetne, ha a kutatók gondosan megkülönböztetnék e két fogalmat egymástól. Ugyanis amennyiben egy eredmény szignifikáns, még lehet tudományos szempontból érdektelen, másrészt nem feltétlenül szükséges statisztikai szignifikancia ahhoz, hogy tudományos szempontból jelentős legyen – gondoljunk csak például arra, hogy akár egyetlen megfigyelés is cáfolhat egy elméletet. *Ziliak–McCloskey* [2008] számos példával alátámasztották, hogy e probléma napjainkban is fennáll, bizonyos kutatók a statisztikai szignifikanciát mint a szakmai fontosság (gyakran egyetlen) kritériumát használják.

1.4. A szignifikanciateszt használata kifejezetten nagy mintáknál

Egy további hiba, amit a kritikusok szerint a szignifikanciateszt alkalmazásakor bizonyos kutatók elkövetnek, hogy olyan nagy mintákra használják, melyek esetén a statisztikai szignifikancia semmitmondó. Ennek szélsőséges változata, amikor a teljes populációról rendelkezésre állnak adatok, de mégis végeznek szignifikanciavizsgálatot. *Meehl* [1990] *Lykken*nel végzett kutatásában egy 57 000-es mintában 15 kategoriális változót hasonlított össze minden lehetséges módon kereszttáblával, hogy rámutasson, nagy mintákat használva nem sokat mond a szignifikáns eredmény – és valóban, a khinégyzet próba mind a 105 esetben szignifikáns eredményt adott. A szerzők azt a jelenséget, hogy a társadalomtudományokban valamilyen szinten minden mindennel összefügg, zajfaktornak (crud factor) nevezték el. Aki tehát nagy mintán végez szignifikanciavizsgálatot (mint amilyenek a KSH több tízezres felmérései, többek között a munkaerő-felmérés is), az a zajfaktor miatt szignifikáns eredményt talál.

1.5. A p-érték azonosítása a nullhipotézis valószínűségével vagy a hatásnagysággal

A korábbiakban már említettük, rendkívül elterjedt, hogy a *p*-értéket a nullhipotézis valószínűségének tekintik. Sokan pedig a hatáserősséggel hozzák azt

összefüggésbe. *Duggan* már az 1970-es években felhívta arra a figyelmet, hogy téves a kis *p*-értéket a kapcsolat erősségével azonosítani, de a szociológiai folyóiratokban ez mégis létező gyakorlat. Duggan példaként rámutatott, hogy a kereszttábla-elemzéskor használt khi-négyzet próbához tartozó *p*-értéktől függetlenül lehet a (gammával mért) változók közötti kapcsolat erős vagy gyenge. Egy nagy mintával nagyon gyenge kapcsolat esetén is kis *p*-értékeket kapunk, de ez nem jelenti azt, hogy erős lenne az összefüggés. A több évtizedes figyelmeztetés ellenére a félreértelmezés manapság is számos munkában megtalálható – *Ziliak–McCloskey* [2008] és *Fidler* [2005] például azt a gyakorlatot említik, amikor regressziós modelleket interpretálva a változókat aszerint rangsorolják, hogy mekkora a *t*-statisztika abszolút értéke, azt a látszatot keltve, mintha az a hatás erősségét mutatná.

1.6. A tesztek erejének figyelmen kívül hagyása

Nullhipotézis-vizsgálatkor fontos szempont lenne a teszt erejének vizsgálata, ám gyakran a szerzők erre nem fordítanak figyelmet. Az erő azt mutatja, hogy mekkora a valószínűsége a nullhipotézis elvetésének, amikor az hamis. Függ a mintanagyságtól, a szignifikanciaszinttől (vagyis a még elfogadott elsőfajú hibavalószínűségtől) és attól, hogy a nullhipotézistől mekkora eltérést, más szóval mekkora hatásnagyságot szeretnénk kimutatni. A nagyobb minta és a magasabb szignifikanciaszint növeli a teszt erejét, illetve minél nagyobb a kimutatni kívánt hatás, annál erősebb a teszt. *Cohen* [1962] a *Journal of Abnormal and Social Psychology* (Abnormális és Szociálpszichológiai Folyóirat) 1960-ban és 1961-ben megjelent cikkeiben vizsgálta a tesztek erejét post hoc erőelemzéssel. Nem a cikkekben közölt hatásnagyságokat használta az erő meghatározására, hanem arra volt kíváncsi, hogy a használt tesztek mekkora erővel tudnak kimutatni kicsi, közepes vagy nagy eltéréseket. Kis hatásnagyság mellett átlagosan 18, közepesnél 48, míg nagynál 83 százalékos erőt mért. A kis hatások esetén ez igen alacsony érték – így az esetek átlagosan 82 százalékában (!) a szerzők nem tudtak volna kimutatni hatást vizsgálataikkal, és tévesen fogadták volna el a nullhipotézist. Cohen mintájára később még többen vizsgálták a pszichológiai folyóiratok cikkeiben bemutatott szignifikanciatesztek erejét (például *Rossi* [1990]), és hasonló átlagos erőket mértek. Ezek az eredmények arra figyelmeztetnek, hogy mintanagyságot úgy érdemes választani, hogy erőelemzést is végzünk előtte. Fontos azt is megjegyezni, hogy egy szociológiai vagy közgazdaságtani cikkeket vizsgáló tanulmány ennél várhatóan nagyobb erejű teszteket találna, mert ezekben a tudományokban jellemzően nagyobb esetszámmal dolgoznak, mint a pszichológiában.

1.7. Ragaszkodás az ötszázalékos küszöbhöz

Bár az elsőfajú hibavalószínűség megválasztása a kutató döntése – például a mintavételi módszer, a mintanagyság és a teszt erejének függvényében –, a legtöbb kutató rutinszerűen a 0,05-os szintet használja (*Leahey* [2005]). Ugyancsak elterjedt a 0,01-os és a 0,001-es szint is.

Természetesen lehet amellett érvelni, hogy a szokásos 5 százalékos szint alkalmazása bizonyos objektivitást ad a kutatási adatok elemzésének, és azért van rá szükség, hogy ne a kutatók szubjektív döntésein múljon a kutatás kimenetele (*Schmidt–Hunter* [1997]), de amellett, hogy az miért éppen 5 százalék, már elég nehéz indokot felhozni. A küszöb megválasztása valószínűleg *Fisher* [1956] nevéhez köthető, aki azonban nem ragaszkodott az érték rögzítéséhez: „[…] egyetlen tudósnak sincs rögzített szignifikanciaszintje, amelyhez évről évre, minden körülmények között ragaszkodna, ehelyett a bizonyítékai és elgondolásai fényében minden egyes esetben ezt külön választja meg.” (42. old.). Nyilvánvalóan problémás, ha egy 5,1-os és egy 4,9 százalékos *p*-értéket adó szignifikanciateszt egészen más következtetéshez vezet, főként, ha az 5 százalékos szint megválasztása csak rutinból, megszokásból történik, és nincs mögötte igazi érv.

1.8. A teszt feltételeinek figyelmen kívül hagyása

Ahhoz, hogy a szignifikanciatesztet értelmezni tudjuk, bizonyos alkalmazási feltételeknek teljesülnie kell. Ilyen alapvető feltétel az, hogy rendelkezzünk valamilyen valószínűségi mintával, és ebből próbáljunk következtetni a populációra. *Kline* [2004] megállapítása szerint a társadalom- és viselkedéstudományokhoz tartozó kutatásokban a legtöbb esetben kényelmi mintát használnak, ennek ellenére ezeknél is elterjedt a nullhipotézis szignifikanciateszt alkalmazása. A *Leahey* [2005] által vizsgált empirikus szociológiai cikkek több mint felében szerepelt nem véletlen minta, pedig (a pszichológiával szemben) ezen a területen elterjedtek a survey-alapú kutatások. Összetettebb (például regressziós) modelleknél számos más feltételnek is teljesülnie kell, hogy a modell paraméterbecslései, a hozzájuk tartozó konfidencia intervallumok és *p*-értékek érvényesek legyenek. Ennek ellenére *Osborne–Waters* [2002] pszichológiai folyóiratok cikkeit vizsgálva azt állapította meg, hogy ezeket a feltételeket a szerzők csak az esetek nagyon kis részében ellenőrzik. *Kline* [2004] szintén hasonló következtetésekre jutott oktatás-, illetve beszéd-, nyelv- és hallásku-tatással foglalkozó folyóiratok kapcsán.

1.9. Szignifikanciavadászat

A *Freedman–Pisani–Purves* [2005] által szignifikanciavadászatnak, *Kline* [2004] által pedig horgászexpedíciónak nevezett rossz gyakorlat során a kutató csak az adatok megtekintése után dönti el, hogy melyek szignifikanciáját ellenőrzi. *Selvin* [1957] megállapítása szerint talán a szociológusok között a legelterjedtebb az az interpretációs probléma, hogy csak az adatok vizsgálata után fogalmazzák meg hipotéziseiket, melyeket aztán ugyanazokon az adatokon tesztelnek. *Feynman* [1998] ugyanezt a problémát túlillesztésnek nevezi. *Gigerenzer* [2004] szerint ezt a hibát újra és újra elkövetik a kutatók a rutinszerű szignifikanciatesztelés során, hiszen a modern statisztikai programcsomagok lehetővé is teszik, hogy a változók közötti összes lehetséges kapcsolatot vizsgálják, illetve addig „horgásszanak" közöttük, míg valami szignifikánsat nem találnak, amit aztán közölni lehet.

1.10. A „méretnélküliség" kritikája

Az irodalomban talán a leggyakrabban említett kritika a *Ziliak–McCloskey* [2008] által „méretnélküli" (sizeless) tudománynak nevezett jelenséghez kapcsolódik. Ez összefoglalva annyit tesz, hogy a kutatók a hatásnagyságot nem vizsgálják, hanem dichotóm döntéseket hoznak aszerint, hogy a nullhipotézisteszt szignifikáns-e vagy sem. Ha a hatásnagyságot mégis közlik, akkor is előfordul, hogy nem magyarázzák. Jellemző például, hogy a tanulmányokban lineáris regressziószámítás alkalmazásakor az együtthatók mellett csillagokkal jelzik, melyik változónál észleltek szignifikáns hatást, majd az eredmények értelmezésekor csak ezeknek az együtthatóknak a szignifikáns vagy nem szignifikáns voltát veszik figyelembe, nagyságukat nem (*Meehl* [1978]). Ennek egy másik változata, amit *Ziliak–McCloskey* [2008] „előjeltudomány" kifejezéssel illetnek, hogy a kutatók a változók közötti kapcsolatnak csak az irányát interpretálják (szignifikáns pozitív/negatív kapcsolat), de a nagyságát nem, majd esetleg – egy korábban említett hibát, a szubsztantív szakmai relevancia és a statisztikai szignifikancia összetévesztését is elkövetve – a legkisebb *p*-értékekkel rendelkező változók fontosságát emelik ki. Kifejezetten gyakori ez az eljárás sok változót lefedő, feltáró jellegű kutatásoknál, amikor egyfajta adatbányászati módszerként szűrik ki a szignifikáns kapcsolatokkal bíró változókat.

Számos szerző érvel amellett, hogy a dichotóm döntések (van/nincs összefüggés, hipotézis elvetése/elfogadása) helyett a hatásnagyságok becslése lenne fontos a hozzájuk tartozó konfidencia intervallumok közlése mellett (*Gardner–Altman* [1986]). A konfidencia intervallumok megadása nem a nullhipotézis szignifikanciateszt valamely inherens hátrányát orvosolná. Jelentősége abban lehet, hogy a teszt szokásos rituáléjából kibillentve az azt alkalmazó kutatót, mind saját maga, mind pedig az

olvasó számára elősegítheti az eredmények helyes értelmezését. Nézzünk erre egy egyszerű példát. Legyen két, nőkből és férfiakból álló mintánk, A országban 40-40 fős mintanagysággal, 20 000Ft-os keresetkülönbséggel, B országban 4 000-4 000 fős mintanagysággal, 2 000Ft-os keresetkülönbséggel, C országban pedig 5 000-5 000 fős mintanagysággal, 5 000 Ft-os keresetkülönbséggel a férfiak javára. A kétmintás *t*-próba által adott három *p*-érték 0,54, 0,54, 0,012; tehát a nullhipotézis szignifikanciatesztet rituálisan alkalmazva a kutató arra jut, hogy a próba szerint csupán a harmadik országban keresnek statisztikailag szignifikánsan jobban a férfiak. A férfiak keresetelőnyére felírt konfidencia intervallumot (esetünkben 45 000–85 000 Ft; 5 000–9 000 Ft; valamint 1 000–9 000 Ft) és a keresetkülönbséget mint hatásnagyságot is figyelembe véve azonban látszik, hogy A ország populációjában a férfiak nagy keresetelőnye és keresethátránya is elképzelhető, vagyis bizonytalan a minta nyújtotta információ. Szemben a B országgal, ahol a szűk konfidencia intervallum kismértékű hatásra utal, csakúgy, mint a C ország esetében, melynél a nagy mintanagyság miatt ez a hatás meghaladja a statisztikai hibát, vagyis 5 százaléknál kisebb *p*-t eredményez. A konfidencia intervallum ismertetése tehát a hatásnagyság felé irányítja a gondolkodást dichotóm döntések esetén.

A hatásnagyság vizsgálatának előfeltétele, hogy változóinkat értelmezhető egységekben mérjük, és el tudjuk dönteni, mennyire számít az nagynak. Ez a döntés azonban már alapvetően nem statisztikai, hanem szakmai kérdés.

2. Tudományterületek közötti eltérések a szignifianciateszt téves gyakorlatában

Az egyes tudományterületek között nagy különbségek találhatók aszerint, hogy miként viszonyulnak a szignifikanciateszt helytelen gyakorlatához.

Az *orvostudományokban*, többek között a bevezetőben említett Kenneth Rothman szerkesztőnek köszönhetően, a teszt túlértékelése a XX. század végére visszaszorult.

A *pszichológia* területén hamarabb megjelentek a kritikák, mint az orvostudományban (lásd például Meehl már bemutatott bírálatát). Az 1990-es években bizonyos folyóiratok megváltoztatták irányelveiket annak érdekében, hogy a „méret nélküli" szignifikanciavizsgálatok helyett a szerzők hatásnagyságokról és konfidencia intervallumokról is írjanak. 2004-re 23 olyan pszichológiai folyóirat volt, melynek szerkesztősége útmutatóban emelte ki a nullhipotézis szignifikanciateszttel kapcsolatos lehetséges buktatókat, és bátorította a hatásnagyságok, illetve a konfidencia intervallumok közlését. Például az Amerikai Pszichológiai Társaság 1994. évi publikációs kézikönyvében (*American Psychological Association* [1994]) megjelent egy erre vonatkozó ajánlás.

Ezeknek az intézkedéseknek azonban nem volt jelentős hatása: *Cumming et al*. [2007] tíz vezető pszichológiai folyóiratban 1998 és 2006 között megjelent tanulmányokat vizsgálva rámutattak, hogy azokban továbbra is nagy számban szerepelnek nullhipotézis szignifikanciatesztek (az írások több mint 95 százaléka hagyatkozik erre), és konfidencia intervallumokat még 2006-ban is csak 10 százalékuk közölt.

A *szociológiában* is népszerű a nullhipotézis szignifikanciateszt. *Leahey* [2005] húsz jelentős szociológiai folyóirat 1935-től 2000-ig kiadott tanulmányait vette górcső alá. A 613 írásból álló mintában a cikkek 81 százaléka (ahol az lehetséges volt) közölt szignifikanciavizsgálatot. Ez az arány az 1930-as és az 1950-es évek vége között 30-ról 80 százalékra nőtt. Majd az 1970-es évekig – valószínűleg az akkoriban virágzó kritikák hatására (lásd például *Duggan* [1970]) – csaknem 60 százalékra esett vissza. Később újra növekedésnek indult, és az 1990-es években 90 százalék körüli szinten mozgott. *Leahey* alapvetően a statisztikai szoftverek elterjedésének tudja be a tesztek 1970-től kezdődő előretörését.

Bár Leahey cikke nem a nullhipotézis szignifikanciateszttel kapcsolatos rossz gyakorlatok elterjedtségére fókuszál, kiderül belőle, hogy a statisztikai szignifikanciavizsgálatot alkalmazó cikkek 10 százalékánál használtak egyszerű véletlen, 30 százalékuknál más valószínűségi mintavételt, 6 százalékuknál rendelkezésre álltak adatok a teljes populációról, a maradék 54 százaléknál pedig valamilyen nem valószínűségi mintavételre hagyatkoztak. Vagyis a teszt is és a véletlen mintára vonatkozó feltételek figyelmen kívül hagyása is elterjedt volt az adott időszakban. Ugyancsak általános volt, hogy a cikkírók az 5, 1 vagy 0,1 százalékos küszöbhöz ragaszkodtak a nullhipotézis szignifikanciatesztek végzésekor, más szempontokat (például a mintanagyságot) figyelmen kívül hagyva.

A *közgazdaságtan* publikációs gyakorlatát tekintve *Ziliak–McCloskey* [2008] az *American Economic Review* (Amerikai Közgazdasági Szemle) című folyóirat 1980-as, illetve 1990-es években megjelent cikkeit vizsgálták. Azokra a publikációkra koncentráltak, amelyekben a szerzők regressziós modellt használtak. Mindkét időszakban a leggyakoribb, a tanulmányok íróinak több mint 90 százaléka által elkövetett hiba volt az, hogy figyelmen kívül hagyták a tesztek erejét. De a Ziliakék által egyik legsúlyosabbnak tartott tévedés (a statisztikai szignifikanciának szubsztantív szakmai fontosságként való értelmezése) is a publikálók mintegy harmadánál előfordult.

3. A nullhipotézis szignifikanciateszt téves használatának gyökerei

Tudományszociológiai okok. *Ziliak–McCloskey* [2008] szerint az is oka lehet a szignifikanciavizsgálat túlzott elterjedésének, illetve a társadalom- és viselkedéstu-

dományok képviselői által a statisztikai szignifikancia szubsztantív szakmai fontosságként való téves kezelésének, hogy ezeknek a „puha" tudományoknak a szignifikanciateszt bizonyos biztonságot nyújt, csökkenti „kisebbrendűségi érzésüket" a „keményebb" szakterületekkel szemben. A számolás és a behatárolt szabályok (például az 5 százalékos küszöb) objektívnek, „tudományosnak" mutatja eredményeiket.

Az is kiváltója a nullhipotézis szignifikanciateszt gyakori alkalmazásának, hogy azt rendkívül könnyű végezni a statisztikai szoftvereknek és a számítógépes kapacitások intenzív növekedésének köszönhetően.

Publikációs gyakorlat: *mit érdemes közölni*. *Sterling–Rosenbaum–Weinkam* [1995] tanulmányukban négy nagy pszichológiai folyóirat 1986. és 1987. évi cikkeit vizsgálva azt találták, hogy azokban, melyekben a szerzők nullhipotézis szignifikanciatesztet végeztek, a nullhipotézis az esetek 94 százalékában elutasításra került. Matematikailag tehát könnyen bizonyítható, hogy a publikált eredmények nem képezhetik reprezentatív mintáját az összesnek. Ezzel arra a publikációs torzítás néven ismert jelenségre utalunk, hogy a folyóirat-szerkesztők főleg a statisztikailag szignifikáns eredményeket bemutató cikkeket fogadják el, illetve a kutatók már eleve csak ezeket teszik közzé. *Sterling* és szerzőtársai szerint ennek a gyakorlatnak az a fontos következménye, hogy azt a hamis látszatot kelti, a szubsztantív szakmai fontosság szoros kapcsolatban áll a statisztikai szignifikanciával, hiszen (esetleg akkor is, ha azoknak nincs jelentőségük) szignifikanciavizsgálatok végzésére és arra ösztönzi a kutatókat, hogy csak a statisztikailag szignifikáns eredményeiket tartsák tudományosan fontosnak.

Tankönyvek. A nullhipotézis szignifikanciatesztről szóló vitában több hozzászóló is amellett érvelt, hogy a publikációs szabályok helyett vagy mellett az oktatás módszerein is változtatni kellene annak érdekében, hogy a hipotézisvizsgálatokat ne övezze annyi félreértés, és a publikációk minősége javuljon. *Ziliak–McCloskey* [2008] szerint csak rendkívül kevés ökonometria-tankönyv különböztetheti meg a statisztikai szignifikanciát a szubsztantív szakmai fontosságtól, és még a statisztikatankönyvek többsége is implicit módon elköveti a felcserélt feltétel hibáját. *Scheff* [2011] négy új tankönyvet vizsgált, és úgy találta, hogy egyik sem hívja fel kellő mértékben a figyelmet a teszttel kapcsolatos problémákra, illetve a lehetséges félreértelmezésekre. Ugyanakkor létezik ellenpélda is, lásd például *Freedman–Pisani–Purves* [2005] jegyzetét. Csak néhány mondatot kiemelve ebből: „A próba nem ellenőrzi, hogy a modell illeszkedik-e a vizsgált kérdéshez, és ésszerű-e. Az eltérés fontosságát sem méri. Az eltérés okát sem állapítja meg. A próba tehát csak egyetlen, nagyon speciális kérdésre tud felelni. Márpedig mi sokszor nem erre a kérdésre volnánk kíváncsiak." (623. old.).

4. A hazai gyakorlat: a *Szociológia Szemle* cikkeinek vizsgálata

A következőkben annak a kérdésnek a megválaszolására teszünk kísérletet, hogy a nullhipotézis szignifikanciateszttel kapcsolatos rossz gyakorlat a magyar empirikus szociológiában is jelen van-e. Ennek azért van jelentősége, mert a nemzetközi szakirodalomban nem tudunk olyan vizsgálatról, amelyben szociológiai folyóirato(ka)t elemeztek volna e szempontból. Ebben közrejátszhat az is, hogy a szociológiában több mint negyven éve nem lángolt fel vita a nullhipotézis szignifikanciateszttel kapcsolatosan. Ekkor jelent meg ugyanis *Morisson–Henkel* [1970] kritikákat és ajánlásokat összegyűjtő kötete, amelyből *Duggan* [1970] tanulmányára korábban már hivatkoztunk. Bár a könyvet 2006-ban újra kiadták, más jele nem mutatkozik annak, hogy (hasonlóan a pszichológiához vagy a közgazdaságtanhoz) a szociológia területén is komoly diskurzus folyna a témával kapcsolatosan.

Jelen munkánkban a *Szociológia Szemle* 2000 és 2014 közötti számaiban megjelent tanulmányokat vizsgáljuk, azokra koncentrálva, amelyekben a szerzők valamilyen regressziós módszert használtak. Az empirikus szociológiai írásokban egyértelműen ez volt a legelterjedtebb eljárás, harmincnyolc ilyen cikk jelent meg a folyóiratban tizenöt év alatt. A regressziós technikák közül a legkisebb négyzetek módszerét, a logisztikus regressziót, a legkisebb négyzetek módszerén alapuló útmodellt, illetve a többszintű lineáris regressziót alkalmazták a szerzők. Csak egy olyan cikk volt, amelyben ugyan regressziót számítottak, de szignifikanciát nem vizsgáltak, mert az adatok a teljes populációra vonatkoztak. Ezért ezt nem vettük be az elemzésbe, ami így végül 37 cikkre vonatkozik.

4.1. A szubsztantív szakmai fontosság összetévesztése a statisztikai szignifikanciával

E két fogalom összekeverésére utaló jel, ha a nem szignifikáns eredményeket automatikusan figyelmen kívül hagyják a szerzők. Csak két olyan munka volt a 37 közül, ahol ez nem történt meg, és a nem szignifikáns kapcsolatokkal is részletesebben foglalkoztak. Ezek egyikében a szerző tekintetbe vette azt is, hogy a kis minták nehezebben „produkálnak” statisztikailag szignifikáns eredményt. Más cikkekben azonban gyakran találhatók olyan mondatok, amelyekből arra lehet következtetni, hogy a szakmai fontosság („érdemi összefüggés”, „megjegyzésre érdemes jelenség”, „említésre méltó szerep”) a statisztikai szignifikanciával definiálódik.

További jel, ha egy modellt úgy építenek fel a kutatók, hogy az eleve csak a statisztikailag szignifikáns összefüggésekkel foglalkozik. Erre jó példa a stepwise, a forward vagy a backward regresszió, ami automatikusan „kidobja” a nem szignifikáns változókat. Ilyen modellek használatával két cikkben találkoztunk.

4.2. A szignifikanciateszt használata kifejezetten nagy minták vagy a teljes populáció esetén

Csak egyetlen tanulmány szerzője követte el azt a hibát, hogy szignifikancia alapján döntött olyan esetben, mikor a teljes populációról rendelkezésre álltak adatok. A cikkben a mérési egységeket 68 ország alkotta. Ezek bár nem fedik le a Föld minden országát, semmi esetre sem tekinthetők egy, az összes országból vett véletlen mintának (mindössze arról van szó, hogy bizonyos adatok nem álltak rendelkezésre).

Ezen kívül három olyan tanulmánnyal találkoztunk, ahol 5 000-nél nagyobb esetszámmal dolgoztak a szerzők. Ezek közül egy volt olyan, ahol a nagy elemszám (több mint 40 ezer fős minta) ellenére sem vizsgáltak hatásnagyságokat, csak azt, hogy szignifikáns-e a regressziós együttható, és milyen irányú kapcsolatra utal. Nem meglepő módon szinte az összes általuk vizsgált kapcsolat szignifikánsnak mutatkozott, viszont eredményeik viszonylag kevéssé voltak informatívak. (A másik két tanulmányban a regressziós együtthatóknak nemcsak az irányát illetve szignifikanciáját vizsgálták, hanem interpretálták azok nagyságát is.)

4.3. A *p*-érték azonosítása a nullhipotézis valószínűségével vagy a hatásnagysággal

Egyetlen általunk vizsgált cikkben sem azonosították a *p*-értéket a nullhipotézis valószínűségével. Viszont az előfordult, hogy egyes szerzők úgy interpretálták azt, mintha az a hatásnagyság lenne: „A változó hatása erős ($p = 0{,}05$ százalékos szinten szignifikáns).” Volt olyan cikk is, ahol a hatásnagyságot egyértelműen megkülönböztették a szignifikanciától: „A többváltozós elemzés a kilépőkről markáns, de az elemszámnak köszönhetően kevés szignifikáns összefüggést mutatott ki.”

4.4. A tesztek erejének figyelmen kívül hagyása

Egy olyan cikk sem volt a 37 között, ahol foglalkoztak volna a tesztek erejével. Jellemzően survey vizsgálatokról és 1 000 körüli vagy annál is nagyobb mintaelemszámokról lévén szó, ez valószínűleg annak tudható be, hogy a tesztek erejének is igen nagynak kellett lennie. Viszont abban a néhány cikkben, ahol alacsonyabb volt az esetszám, szintén nem merült fel a másodfajú hibavalószínűség kiszámítása.

Feltételezhetően a pszichológiában vagy más olyan diszciplínák esetében, ahol a kutatók általában kisebb elemszámokkal dolgoznak, és emiatt a tesztek ereje sem nagy, a statisztikaoktatás nagyobb hangsúlyt fektet a téma ismertetésére, így a szerzők is jobban ismerik annak fontosságát.

4.5. Ragaszkodás az ötszázalékos küszöbhöz

Az általunk vizsgált cikkek mindegyikében 5 százalékos szignifikanciaszintet használtak, de némely esetben emellett még az 1, illetve 0,1 százalékos szignifikanciaszint alatti *p*-értékeket is jelölték. Egyes cikkekben pontos *p*-értékeket közöltek, de sokban csak félkövér szedéssel vagy (egy, kettő vagy három) csillaggal jelezték, hogy eléri-e az 5, 1, vagy 0,1 százalékot. Az 5 százalékos szinttől való eltéréssel csak olyan cikkekben találkoztunk, ahol bár alapvetően ezt a küszöböt tekintették irányadónak, 10 százaléknál kisebb *p*-értékkel rendelkező kapcsolatokról is beszéltek. A következő példa jól mutatja, hogy a magyar empirikus szociológiai gyakorlatban is kitüntetett szerepe van az 5 százalékos küszöbnek: „2002-re eltűnik a régió szignifikáns hatása ($p = 0{,}07$) abban az esetben, ha a régióváltozó a szokásos három értéket (fejlett európai régió, posztszocialista országok, USA) veszi fel. Ha viszont létrehozunk egy dummy változót, amely azt méri, hogy a kérdezett a posztszocialista országok régiójába tartozik-e vagy sem, akkor ez a régióváltozó már szignifikáns hatást fejt ki a frusztrációt mérő változóra ($p = 0{,}034$).” Az idézet azt a benyomást kelti, hogy a változók kategóriái nem teoretikus szempontból kerültek megváltoztatásra, hanem a kívánt szignifikancia elérése érdekében (tehát felmerül a „szignifikanciavadászat” hibája is). Így az 5 százalékos küszöb már nem tekinthető objektívnek. Az is a szokásos küszöbhöz való ragaszkodást mutatja, hogy ugyanennek a cikknek egy lábjegyzetében a szerző öt tizedesjegyig kényszerül kiírni a szignifikanciát, ezáltal bizonyítva, az nem éri el az 5 százalékot.

4.6. A teszt feltételeinek figyelmen kívül hagyása

A tanulmányok nagy részében valószínűségi mintával dolgoztak a szerzők. Egy olyan (már korábban is említett) cikket találtunk, amelyben a teljes populációról rendelkezésre álltak adatok. Egy másikban a mérési egységek magyarországi városok voltak, és a cikk írói az ezek önkormányzata által kitöltött kérdőíveket elemezték. Az összes önkormányzatnak csak 24 százalékától érkeztek vissza válaszok, így a szerzők ezen a (nem valószínűségi) mintán elemezték az adatokat. Még további négy írásban szerepelt nem valószínűségi minta, de ezek írói minden esetben felhívták a figyelmet arra, hogy emiatt az eredmények csak korlátozottan általánosíthatók (bár ennek ellenére a szignifikanciát a szokásos módon értelmezték). A 37 tanulmány között csak kettő olyat találtunk, ahol a szerző említette, hogy ellenőrizte a regressziós modell feltevéseit.

4.7. A „méretnélküliség” kritikája

Bár az összes tanulmányban közölték a regressziós együtthatókat, a 37-ből 17 esetben fordult elő, hogy a hatások nagyságát nem interpretálták, csak az előjelüket és az irányukat. Például a következőképpen: „Az eredmények azt mutatják, hogy az együtthatók előjele megfelel az elméleti előrejelzéseknek és minden specifikációra szignifikánsak (5. táblázat). Másképpen fogalmazva, a családi gazdaságok kevesebb tőkét használnak, mint a nem családi gazdaságok. A becslések azt is mutatják, hogy az idősebb farmerek magasabb tőkeállománnyal rendelkeznek.”

Azért emeltük ki ezt az idézetet, mert ebben annak ellenére nem ismertették a szerzők a hatásnagyságot, hogy a magyarázóváltozók mérési egységei könnyen értelmezhetők voltak (a tőke ezer forintban volt mérve), és a (legkisebb négyzetek) módszer sem nehezítette az interpretációt. A hatásnagyságok vizsgálata azért lett volna fontos, mert – bár a cikk regressziós együtthatókat tartalmazó táblázatából kiderült, hogy a becslés szerint a családi gazdaságok 4,8 millió forinttal kevesebb tőkével rendelkeztek, mint a nem családiak – egy nem szakmabeli számára nem világos, hogy az soknak vagy kevésnek számít.

Olyan esettel is találkoztunk, amikor a regresszió eredményeinek magyarázata vélhetően azért nem történt meg, mert nem volt kézenfekvő, mi számít nagy vagy kis eltérésnek a változók nehezen értelmezhető skáláján: „A két nyugat-európai országban szignifikánsan nagyobb az egyénenkénti posztmateriális-materiális veszélyekért aggódás különbségét mérő bizonytalanság fókusza változó átlaga. Ez azt jelenti, hogy a franciák és a britek inkább fókuszálnak a posztmateriális, globális ökológiai veszélyekre, mint a magyarok vagy a görögök.” Ha e cikkben nem is lehetett számszerű mértékegységekben mérni a változókat, hasznos lett volna, ha a szerzők közlik, hány szórásnyi a különbség az egyes országok között, és vajon (korábbi tapasztalatok szerint) jelentős mértékű-e ez az eltérés.

Volt olyan tanulmány is a vizsgáltak között, melynek szerzője indokolta, miért nem adja meg hatásnagyságokat: „Mivel a kutatás célja elméleti magyarázatok ellenőrzése és nem egy jelenség előrejelzése volt, az elemzés során nem tértem ki a szignifikáns hatással bíró változók magyarázó erejének összehasonlítására vagy a magyarázó modell erejének elemzésére, csupán a hatások és irányuk regisztrálására.” Véleményünk szerint egyrészt a hatásnagyságok közlése a korábban írtak alapján nem csak akkor lehet fontos, ha jelenségek előrejelzéséről beszélünk. Másrészt, ha egyszerű szignifikanciavizsgálatokkal ellenőrizni lehet az elméletet, akkor a *Meehl* [1978] által részletesen kifejtett probléma jelentkezik: az elméletet nem tesszük ki valódi kockázatnak, ha olyan nullhipotéziseket fogalmazunk meg, amelyeket könnyen el tudunk vetni. Hasznosabb lenne, ha az elmélet alátámasztása nemcsak ilyen nullhipotézisek elvetéséből állna, hanem további érvekkel is támogatnák azt.

A *Szociológiai Szemle*ben a kutatási hipotézisek megfogalmazása sok esetben arra enged következtetni, hogy a hatásnagyságokra kevés hangsúlyt fektettek a kutatók. Továbbá, több esetben csak a standardizált regressziós együtthatókat közölték és interpretálták. Így alapvetően csak a változók egymáshoz képesti erősségét tudták megítélni, de arról nem kaptak képet, hogy önmagukban mennyire számított nagynak egy-egy változó hatása.

Konfidencia intervallumot nem tartalmazott egyetlen tanulmány sem. Ugyan volt hét olyan cikk, ahol a standard hibák a regressziós együtthatókat összefoglaló táblázatban, zárójelben szerepeltek, ezeket egy esetben sem magyarázták, így feltüntetésük inkább csak formalitás volt.

Táblázatunk összefoglalja a számszerűsíthető eredményeket. A szerzők az összes tanulmányban közölték a regressziós együtthatók nagyságát. Huszonegy cikkben használtak lineáris regresszión alapuló modellt, de ezek közül kilencben mindössze a standardizált regressziós együtthatókat tüntették fel, a standardizálatlanokat nem. A regressziós együtthatók nagyságát a harminchét írásból húsznak a szövegében is értelmezték, ezek közül öt esetben viszont csak a változók egymáshoz képesti nagyságát vizsgálták. A tanulmányok többségében (30-ban) valószínűségi mintán dolgoztak; és mindössze kettőben említették egyértelműen, hogy ellenőrizték az alkalmazott modell feltevéseinek teljesülését.

A Szociológiai Szemle regressziós modellt alkalmazó írásainak jellemzői, 2000–2014

Jellemző	Cikkek száma
A szerző(k) (táblázatban) közli(k) a regressziós együtthatók nagyságát	3
Csak standardizált regressziós együtthatókat ismertet(nek)	9
A szövegben elemzi(k) a regressziós együtthatók nagyságát	20
Csak a változók egymáshoz képesti nagyságát interpretálja/interpretálják	5
Valószínűségi mintán dolgozik/dolgoznak	30
Említi(k), hogy ellenőrizte/ellenőrizték a modell feltevéseit	2
Közöl(nek) erőszámításokat	0
Bemutat(nak) konfidencia intervallumokat	0
N	37

Tehát minden felsorolt, nullhipotézis szignifikanciateszttel kapcsolatos rossz gyakorlatra volt példa a *Szociológiai Szemle* cikkeiben a vizsgált 15 év során. E hibák arra utalnak, hogy magát a regresszióelemzést (a módszer megválasztását, az eredmények gyakran más értelmezést nélkülöző táblázatos közlését) is bizonyos szempontból „ritualisztikusan” használják a szerzők. A szubsztantív szignifikanciával kapcsolatos deficitet a hatásnagyságok már említett elhagyásán kívül az is

mutatta, hogy az olvasók számára nem derült ki minden esetben, „mire jó”, milyen jelentősége van annak, amit a kutatók találtak.

Mindezek arra világítanak rá, hogy a (magyar) szociológia területén sem lennének haszontalanok az adatok értelmezésének megreformálására irányuló törekvések. Az pedig, hogy a nemzetközi szociológiában több mint negyven éve lángoltak fel utoljára a nullhipotézis szignifikanciateszttel kapcsolatban viták, akár előnyére is válhat a tudománynak, hiszen a hibás gyakorlat megváltoztatására vannak sikeres (orvostudomány) és sikertelen (pszichológia) példák is.

5. Ajánlások

Jelen fejezetben *Harlow* [1997] és *Kline* [2004] összefoglaló munkái alapján ismertetünk néhány fontos ajánlást, melyeket a szakemberek az utóbbi 50-60 évben fogalmaztak meg a nullhipotézis szignifikanciateszt túlzott és egyes esetekben helytelen használatának visszaszorítása érdekében.

5.1. Az elméletek alapos és gondos értékelése, átgondolása

Bár triviálisnak tűnhet ez a javaslat, mégis, ha minden kutató szem előtt tartaná, valószínűleg el lehetne kerülni a nullhipotézis szignifikanciateszt rituális és mechanikus használatát. *Yates* már 1951-ben úgy gondolta, hogy a túlzott hangsúly, amit a szignifikanciatesztek kapnak a tudományos következtetési folyamatban, valamint e tesztek mechanikus alkalmazása olyan problémák vizsgálatához vezetett, amelyek fontossága, gyakorlati haszna megkérdőjelezhető. Azóta számos kutató (köztük *Meehl* [1967], [1978]; *Rozeboom* [1960]; *Ziliak–McCloskey* [2008] is) megfogalmazta, hogy tudományos következtetések levonásához nem mechanikus procedúrákra, hanem a józan észre, kritikus gondolkodásra, bölcsességre és intuícióra kell bíznunk magunkat.

5.2. A nullhipotézis szignifikanciateszt elsődlegessége „felderítő” kutatások esetében

Kline [2004] ajánlása szerint a nullhipotézis szignifikanciateszt „felderítő” jellegű kutatásokban és akkor kell, hogy kiemelt szerepet kapjon, mikor az adott témáról nem áll rendelkezésre kellő mennyiségű információ. Ilyen esetekben valóban hasz-

nos lehet, ha erre hagyatkozva próbáljuk megítélni, hogy létezik-e a kapcsolat bizonyos változók között. Ez a feltáró szakasz azonban minden esetben átmeneti, a tájékozódást segíti. Mikor a kutatás (vagy egy kutatási terület) már „érettebb" szakaszba ér, a nullhipotézis szignifikanciateszt domináns szerepét átveszi a hatásnagyságok becslése, illetve a rendelkezésre álló adatokra építő, komplexebb, pontosabb modellek illesztése. Ha például egy szociológus korábbi vizsgálatok alapján már tisztában van azzal, hogy a magasabb iskolázottságúak jobban keresnek, akkor a kapcsolat kimutatása önmagában nem cél. Kline úgy gondolja, éppen az, hogy nem hagyatkozunk kizárólag a nullhipotézis szignifikanciatesztre, lehetne a fémjelzője annak, hogy egy kutatási terület a felderítő, puhatolózó állapotból már továbblépett az érettebb szakaszba.

5.3. A statisztikai erő számítása nullhipotézis-vizsgálat esetén

A kritikák felsorolásánál már említettük, hogy a társadalomtudományokban a kutatók elterjedten alkalmaznak rendkívül kis erejű teszteket. Ennek ellenszere lehetne, ha az *a priori* erőszámításokat minden statisztikai teszt esetében elvégeznék, tehát már a kutatás kezdeti szakaszában a minimálisan kimutatandó hatásnagyság, az erő és az elsőfajú hibavalószínűség függvényében választanák meg a mintanagyságot. Egy példa erre: mekkora mintát kell vennünk, ha 5 százalékos elsőfajú hibavalószínűség mellett 80 százalékos erővel szeretnénk kimutatni a férfiak és a nők átlagjövedelme közötti legalább 10 ezer Ft-os különbséget? (Tegyük fel, hogy a jövedelem szórása mindkét csoportban 20 ezer Ft). Könnyen kiszámítható, hogy a kívánt mintanagyság 64 férfi és 64 nő (amennyiben azonos számú személyt választanánk a mintába a két csoportból). A kutatónak ekkor szakmai ismeretei alapján kell megbecsülnie a populációs szórásokat, illetve megválasztania a minimálisan kimutatandó hatásnagyságot. Az *a priori* erőszámítás bizonyos alkalmazási területeken, például a biostatisztikában bevett gyakorlatnak számít, lásd például *Kraemer–Blasey* [2015]. De ha nem is végzünk ilyet, akkor is lehetséges az adatok ismeretében a tesztekhez tartozó erő bemutatása. A hatásnagyság függvényében felrajzolható az erőfüggvény, ám még egyszerűbb, ha a minimálisan kimutatandó hatásnagysághoz tartozó erőt közöljük. Ez akkor különösen fontos, ha a tanulmányban vannak elfogadott nullhipotézisek, hiszen így képet kaphat az olvasó a másodfajú hibavalószínűségről. *Kline* [2004] azt gondolja, jelenleg azért is kevéssé elterjedt a tesztek erejének ismertetése, mert a szerzők ritkán publikálnak olyan cikkeket, melyekben a nullhipotézist elfogadják. Az empirikus szociológiai kutatásokban talán azért fordítanak erre kisebb figyelmet, mert jellemzően nagy elemszámú mintákkal dolgoznak, amelyeknél nagyobb a statisztikai erő, mint kis mintáknál. Az erő kiszámítása azonban azért is hasznos lehet, mert közben a hatás-

nagyságokkal is foglalkozik a kutató. A legtöbb statisztikai szoftverben már van lehetőség erőszámításokat végezni, így nem okoz technikai nehézséget a tesztek erejének kiszámítása.

5.4. A „szignifikáns" szó használatának megváltoztatása

Kline [2004] szerint a $p < \alpha$ jelenség leírására választott „szignifikáns" kifejezés rossz döntésnek bizonyult. Ez a szó a köznapi használatban a fontos, erős szavakkal szinonim, és így egyrészt félrevezetheti az olvasókat, másrészt táptalajt adhat annak a korábbiakban ismertetett rossz gyakorlatnak, hogy a kutatók a statisztikai szignifikanciát összekeverik a szubsztantív szakmai fontossággal vagy a hatásnagysággal. Ezért nem a változók közötti szignifikáns, hanem a szemléletesebb statisztikai kapcsolat kifejezést kellene használni.

5.5. A dichotóm döntések helyett becslések, a hatásnagyságok és a konfidencia intervallumok közlése, értelmezése

Ezt az ajánlást, melynek fontosságára korábban már rávilágítottunk, átgondoltan kell megvalósítani. *Fidler et al.* [2004] „A folyóirat-szerkesztők rá tudják venni a kutatókat, hogy konfidencia intervallumokat használjanak, de arra nem, hogy gondolkodjanak" beszédes című cikkükben rámutattak arra, hogy a konfidencia intervallumok közlése önmagában felületes eljárás, ha az eredmények interpretálásakor azokat nem veszik figyelembe.

5.6. A kutatók ösztönzése arra, hogy eredményeik szubsztantív szignifikanciáját is ismertessék

Ez az első ajánlással rokon javaslat. A kutatóknak, mivel a nullhipotézis szignifikanciateszt erről semmit sem mond, részletesen ki kellene térniük minden esetben eredményeik fontosságára.

5.7. A kutatások ismétlése és az eredmények értékelése metaanalízis segítségével

Lykken [1968] a kutatások ismétlésének három módját különbözteti meg: a szó szerinti, az operatív és a konstruktív replikációt. Az első esetben az eredeti kutatást a

lehető legpontosabban megismétlik, minden kutatási körülményt és módszert tökéletesen reprodukálva. A valóságban ehhez a legközelebb úgy kerülhetünk, ha az eredeti cikk szerzőit megkérjük, hogy ismételjék meg kutatásukat új alanyokon. Az operatív replikáció esetében a cikkben közölt mintavételi és kísérleti módszertant pontosan követve (de ezen kívül eső faktorokat nem kontrollálva) ismétlik meg a vizsgálatot. Lykken a leghasznosabbnak a konstruktív replikációt tartja, amikor a kutatók egy korábban már megállapított empirikus tényt a saját maguk által legjobbnak tartott módszerekkel és mérőeszközökkel próbálnak újra feltárni. Ha több vizsgálat is rendelkezésre áll, akkor lehetőség van metaanalízist végezni, vagyis szintetizálni azok eredményeit, mellyel pontosabb becslésekre, következtetésekre juthatunk. A metaanalízist nyilván megkönnyíti, ha a korábbi ajánlások szerint a publikálók közlik a hatásnagyságot, a konfidencia intervallumokat, illetve a statisztikailag nem szignifikáns eredményeket is.

5.8. A statisztikai módszerek kevésbé „nullhipotézis szignifikanciateszt központú” oktatása

Kline [2004] szerint a bevezető statisztikai kurzusok túlságosan nagy hangsúlyt fektetnek a nullhipotézis szignifikanciatesztre. A korábbi fejezetekben bemutattuk, hogy a helyzet javítása érdekében a helyes gyakorlatot bemutató tankönyvek szélesebb körű használatára van szükség. Kline emellett úgy gondolja, hogy a kutatásmódszertani és a statisztikai témájú kurzusokat jobban össze kellene hangolni. Ugyanis, ha ezeket külön tárgyként, egymással nem összeegyeztetve tanítják, a diákok nem kapnak tiszta képet arról, hogy a statisztikai módszereket milyen módon alkalmazzák a kutatási gyakorlatban.

5.9. Jobb statisztikai szoftverek

Kline [2004] meggyőződése, hogy a statisztikai szoftverek nagy része túlságosan is a szignifikanciák kiszámítására koncentrál, és az outputban nem közli minden esetben a hatásnagyságokat. Véleményünk szerint e probléma megoldásában segítene, ha a gyakran használt szoftverek beépített módszerei támogatnák a méretcentrikus gondolkodást, de talán még jobb lenne, ha ezeket a kutatók inkább átgondoltan használnák, és nem csak a beépített módszerek, illetve az alapbeállítások közül választanának, valamint nem kizárólag az alapbeállítások szerinti outputot vizsgálnák. Ehhez persze a módszerek és a szoftverek valamivel mélyebb ismerete szükséges.

6. A Basic and Applied Social Psychology folyóirat tiltásának utóélete

Írásunk befejezéseként térjünk vissza a kiinduláshoz, a *Basic and Applied Social Psychology* folyóirat tiltásához. Mint azt már többször említettük, a több mint negyven éve megfogalmazódó kritikák ellenére a módszerrel kapcsolatos rossz gyakorlat nem szorult vissza, így a folyóirat szerkesztőinek aggodalmai jogosak lehetnek. Ezért aktívan kell tenni a nullhipotézis szignifikanciateszt túlzott, illetve téves használata ellen. A folyóirat által közzétett tiltásra a tágabb tudományos közösség többféleképpen reagált. Az Amerikai Statisztikai Szövetség például egy rövid közleményt adott közre, amiben elismerte a „következtetési statisztikai eljárások használata és interpretálása körül kialakult problémákat" (*Wasserstein* [2015]). Ám szerintük a tiltásnak negatív következményei lesznek, és a tudományos közösségnek szélesebb vitát kell folytatnia a statisztikai következtetési eljárásokról.

A korlátozással a brit Királyi Statisztikai Társaság is foglalkozott (*Flanagan* [2015]). Elnökük, *Peter Diggle* üdvözölte és osztotta a folyóirat szerkesztőinek aggodalmait a statisztikai következtetéseket illetően, viszont nem tartotta konstruktívnak a teljes tiltást. Rövid kritikájában kiemelte, a szerkesztőségi állásfoglalás adós maradt annak magyarázatával, hogy a szerzők és az olvasók miként vonjanak le következtetéseket a leíró statisztikák alapján.

A *Basic and Applied Social Psychology* folyóirat ajánlásokat is tett e témában. A szerkesztőség véleménye szerint a szerzőknek a szociálpszichológiában megszokottnál nagyobb mintákon kell végezniük kutatásaikat, csökkentve a mintavételi hibából fakadó bizonytalanságot, és elősegítve a robusztusabb eredmények elérését. Fontos, hogy a tanulmányok írói részletes leíró statisztikákat, gyakoriságokat is közöljenek kutatásukat követően. A szerkesztőség meggyőződése, hogy a tiltás hatására a szerzők felszabadulnak a nullhipotézis szignifikanciateszt által kikényszerített gondolkodási séma alól, és így nagyobb teret kaphat a kreatív gondolkodás. A nullhipotézis szignifikanciateszt mellőzése ezáltal nem rontja a publikált írások színvonalát, sőt, épp ellenkezőleg, javítja azt (mivel korábban számos esetben ezek alkalmazásával igazoltak rossz minőségű kutatásokat).

Kérdés, hogy miképp valósulnak meg ezek az ajánlások. Bár a tiltás óta viszonylag rövid idő telt el, közvetlen hatásaiba a fél év alatt megjelent tizenhárom tanulmányon keresztül nyerhetünk betekintést. Az utóbbiak többsége klasszikus kísérleteket tartalmaz, elvétve találunk csak bennük többváltozós elemzéseket. Mindegyik empirikus munka közöl az ajánlásoknak megfelelő leíró statisztikákat, legtöbbször átlagokat és szórásokat. Ugyanakkor hiába ezek részletes ismertetése, ha mégis magyarázat nélkül maradnak, mivel így bemutatásuk csak puszta formaságnak tűnik.

Van olyan cikk, amelyben a szerzők ugyan nem közölnek szignifikanciateszteket, de a biztonság kedvéért leírják, hogy eredményeik a sokat kárhoztatott $p < 0{,}05$-ös szignifikanciaszinten szignifikánsak. Egy másik munkában pedig egy teljes *t*-statisztikán alapuló szignifikanciavizsgálat értékeiről olvashatunk. Más tanulmányokban, ahol klasszikus kísérleteket alkalmaztak, az eredményeket a hatásnagyság különböző mérőszámaival (a Cohen-féle *d*-értékkel vagy a Glass-féle Δ-val) értékelték ki hüvelykujjszabály szerint, hasonlóan a nullhipotézis szignifikanciateszt esetében elterjedt értelmezéshez.

Hiába kerüli tehát a szerzők többsége a nullhipotézis szignifikanciateszt alkalmazását, a szerkesztőség által korábban nehezményezett módszertani problémák továbbra is fennállnak. Többek között azóta is összemosódnak a statisztikailag szignifikáns és a szubsztantív szakmai fontosságú eredmények. Különbség csupán abban figyelhető meg, hogy a szerzők nem a *p*-érték küszöbértékei, hanem például a Cohen-féle *d*-érték alapján „ítélkeznek". Így a tudományos eredmények mechanikus előállítása továbbra is folyik. A cikkekben megjelennek többváltozós következtetési módszerek is, de sajnos, egyik munkában sem olvashatunk arról, hogy az elemzési eszközök használatához szükséges előzetes vizsgálatokat elvégezték volna.

A 2015. évi állásfoglalásukban arra is kitértek a szerkesztők, hogy a gyakorlattal ellentétben szeretnék, ha a kutatók nagyobb mintaelemszámmal dolgoznának. Az azóta megjelent tanulmányok többsége legfeljebb 150 fős mintán alapul, de gyakori a 60-80-as elemszám is, ráadásul ezek több esetben nem valószínűségi, „kényelmi mintára" épülnek.

Összegzésképpen megállapíthatjuk, hogy az állásfoglalás ajánlásai nem valósultak meg széleskörűen a folyóirat hasábjain. Ennek vagy az eltelt idő rövidsége vagy az lehet az oka, hogy valamilyen okból – talán a radikális tiltás miatt – meghátráltak a szerkesztők.

Irodalom

AMERICAN PSYCHOLOGICAL ASSOCIATION [1994]: *Publication Manual of the American Psychological Association. 4th edition.* American Psychological Association. Washington, D.C.

COHEN, J. [1962]: The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*. Vol. 65. No. 3. pp. 145–153. http://dx.doi.org/10.1037/h0045186

CUMMING, G. – FIDLER, F. – LEONARD, M. – KALINOWSKI, P. – CHRISTIANSEN, A. – KLEINIG, A. – LO, J. – MCMENAMIN, N. – WILSON, S. [2007]: Statistical reform in psychology. Is anything changing? *Psychological Science*. Vol. 18. No. 3. pp. 220–232. http://dx.doi.org/10.1111/j.1467-9280.2007.01881.x

DUGGAN, T. J. – DEAN, C. D. [1970]: Common misinterpretations of significance levels in sociological journals. In: *Morrison, D. E. – Henkel, R. E.* (eds.): *The Significance Test Controversy: A Reader*. Aldine Publication Company. Chicago. pp. 161–165.

FEYNMAN, R. [1998]: *The Meaning of It All: Thoughts of a Citizen-Scientist*. Perseus Books. Reading.

FIDLER, F. – THOMASON, N. – CUMMING, G. – FINCH, S. – LEEMAN, J. [2004]: Editors can lead researchers to confidence intervals, but they can't make them think: Statistical reform lessons from medicine. *Psychological Science*. Vol. 15. No. 2. pp. 119–126.

FIDLER, F. [2005]: *From Statistical Reform to Effect Size Estimation*: *Statistical Reform in Psychology, Medicine and Ecology*. Ph.D. thesis. The University of Melbourne. http://www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf?origin=publication_detail

FISHER, R. A. [1956]: *Statistical Methods and Scientific Inference*. Oliver & Boyd. Edinburgh.

FLANAGAN, O. [2015]: *Journal's ban on null hypothesis significance testing: Reactions from the statistical arena*. StatsLife. http://www.statslife.org.uk/opinion/2114-journal-s-ban-on-null-hypothesis-significance-testing-reactions-from-the-statistical-arena

FREEDMAN, D. – PISANI, R. – PURVES, R. [2005]: *Statisztika*. Typotex. Budapest.

GARDNER, M. J. – ALTMAN, D. G. [1986]: Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*. Vol. 292. No. 6522. pp. 746–750. http://dx.doi.org/10.1136/bmj.292.6522.746

GIGERENZER, G. [2004]: Mindless Statistics. *Journal of Socio-Economics*. Vol. 33. No. 5. pp. 587–606. http://dx.doi.org/10.1016/j.socec.2004.09.033

HALLER, H. – KRAUSS, S. [2002]: Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*. Vol. 7. No. 1. http://www.dgps.de/fachgruppen/methoden/mpr-online/issue16/art1/article.html

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

KRAEMER, C. H. – BLASEY, C. M [2015]: *How Many Subjects? Statistical Power Analysis in Research*. 2nd edition. SAGE Publications. Thousand Oaks.

LEAHEY, M. [2005]: Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*. Vol. 84. No. 1. pp. 1–24. http://dx.doi.org/10.1353/sof.2005.0108

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

MEEHL, P. E. [1967]: Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*. Vol. 34. No. 2. pp. 103–115. http://dx.doi.org/10.1086/288135

MEEHL, P. E. [1978]: Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*. Vol. 46. No. 4. pp. 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

MEEHL, P. E. [1990]: Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*. Vol. 66. No. 1. pp. 195–244. http://dx.doi.org/10.2466/pr0.1990.66.1.195

MORRISON, D. E. – HENKEL, R. E. (eds.) [2006]: *The Significance Test Controversy: A Reader*. Aldine Transaction Publishers. New Brunswick.

OAKES, M. W. [1986]: *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley & Sons. Chichester.

OSBORNE, J. W. – WATERS, E. [2002]: Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*. Vol. 8. No. 2. http://www-psychology.concordia.ca/fac/kline/601/osborne.pdf

ROSSI, J. [1990]: Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*. Vol. 58. No. 5. pp. 646–656. http://dx.doi.org/10.1037/0022-006X.58.5.646

ROZEBOOM, W. W. [1960]: The fallacy of the null-hypothesis significance test. *Psychological Bulletin*. Vol. 57. No. 5. pp. 416–428. http://dx.doi.org/10.1037/h0042040

SCHMIDT, F. – HUNTER, J. [1997]: Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: *Harlow, L. L. – Mulaik, S. A. – Steiger, J. H.* (eds.): *What If There Were No Significance Tests?* Lawrence Erlbaum Associates Publishers. Mahwah. pp. 3-1–3-28. http://www.phil.vt.edu/dmayo/personal_website/Schmidt_Hunter_Eight_Common_But_False_Objections.pdf

SCHEFF, T. [2011]: The catastrophe of scientism in social/behavioral science. *Contemporary Sociology*: *A Journal of Reviews*. Vol. 40. No. 3. pp. 264–268. http://dx.doi.org/10.1177/0094306110404513

SELVIN, H. C. [1957]: A critique of tests of significance in survey research. *American Sociological Review*. Vol. 22. No. 5. pp. 519–527. http://dx.doi.org/10.2307/2089475

STERLING, T. – ROSENBAUM, W. – WEINKAM, J. [1995]: Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*. Vol. 49. No. 1. pp. 108–108. http://dx.doi.org/10.2307/2684823

TRAFIMOW, D. [2014]: Editorial. *Basic and Applied Social Psychology*. Vol. 36. No. 1. pp. 1–2. http://dx.doi.org/10.1080/01973533.2014.865505

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

WASSERSTEIN, R. [2015]: *ASA Comment on a Journal's ban on null hypothesis statistical testing*. American Statistical Association Community. http://community.amstat.org/blogs/ronald-wasserstein/2015/02/26/asa-comment-on-a-journals-ban-on-null-hypothesis-statistical-testing

YATES, F. [1951]: The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*. Vol. 46. No. 253. pp. 19–34. http://dx.doi.org/10.2307/2280090

ZILIAK, S. – MCCLOSKEY, D. [2008]: *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press. Ann Arbor.

A nullhipotézis szignifikanciateszt alapvető bizonyítási eszköz a statisztikában.
A *Statisztikai Szemle* teret kínál az alkalmazásával kapcsolatos
véleményeknek és tapasztalatoknak.
Ezért várjuk Olvasóink hozzászólását a vitaindító cikkhez.