Egy régi probléma újra előtérben: a nullhipotézis szignifikancia-teszt téves gyakorlata[1]

Bárdits Anna[2], Németh Renáta[3], Terplán Győző[4]

Kivonat

Tanulmányunkban az empirikus adatelemzés egyik sarokkövének, a nullhipotézis szignifikancia-tesztnek inherens problémáit és gyakorlati alkalmazási hibáit foglaljuk össze. A teszttel kapcsolatos régi vita áttekintésének apropóját a Basic and Applied Social Psychology (BASP) folyóirat 2015-ös szerkesztőségi állásfoglalása adja, melyben megtiltották szerzőik számára a szignifikancia-teszt alkalmazását. Összefoglalónkból kiderül, hogy a több, mint 50 éve jelen lévő kritikák ellenére a módszerrel kapcsolatos rossz gyakorlatok továbbra is fennmaradtak. Kitérünk az alkalmazási gyakorlat tudományterület-specifikusságára, és tárgyaljuk a téves alkalmazás tudományszociológiai gyökereit is. A Szociológia Szemle cikkeinek vizsgálatával megmutatjuk, hogy magyarországi empirikus szociológiában is azonosíthatók a téves gyakorlat egyes jegyei. Ezek elkerüléséhez több ajánlást gyűjtöttünk össze.

Title: An old problem in the spotlight again. The mistaken practice of the null-hypothesis significance test

Abstract: In our study we summarize the inherent problems and the practical mistakes of the null-hypothesis significance test, a cornerstone of empirical data analysis. The reason to review the old debate regarding the test is an editorial of Basic and Applied Social Psychology (BASP), published in 2015, which bans the use of null-hypothesis significance tests to the authors. Our summary reveals that despite the 50-year-long critics the bad practice continue to sustain. Discipline-specific practices and sociological roots of the problem are also discussed. By examining the papers of the Hungarian Sociological Review, we show that bad practices can be identified in Hungarian empirical sociological research as well. We collected several recommendations to avert these issues.

[1] A cikk Bárdits Anna survey statisztika mesterszakos szakdolgozatán alapszik, a dolgozat konzulense Németh Renáta volt.

[2] Az ELTE survey statisztika szakának végzős hallgatója. barditsanna@gmail.com
[3] Az ELTE Társadalomtudományi Karának docense. nemethr@tatk.elte.hu
[4] Az ELTE survey statisztika szakának végzős hallgatója. terplangyozo@caesar.elte.hu

Mottó:

„A Basic and Applied Social Psychology (BASP) 2014-es szerkesztőségi állásfoglalása hangsúlyozta, hogy a nullhipotézis szignifikancia-teszt eljárása (NHSzTE) érvénytelen, ezért szerzőinktől a továbbiakban nem követeltük meg annak használatát (Trafimow, 2014). Egyúttal türelmi időt jelöltünk ki szerzőinknek, mely idő alatt szerkesztőségünk nem tiltotta az NHSzTE-t. Szerkesztőségünk ezennel bejelenti, hogy a türelmi idő lejárt. Mától a BASP betiltja az NHSzTE használatát."

BASP, szerkesztői bevezető, 2015. február

1. Bevezető

Tegyük fel, kedves olvasó, hogy egy 20 elemű kísérleti és 20 elemű kontroll csoportban vizsgálja az átlagok egyenlőségére vonatkozó nullhipotézist, kétmintás t-próbával. Az eredmények: t=2,7; p=0,01. Melyik állítás igaz ekkor Ön szerint az alábbiak közül?

1. A nullhipotézist (vagyis hogy nincs különbség a populációs átlagok között) maradéktalanul cáfoltam.
2. Megtaláltam annak a valószínűségét, hogy a nullhipotézis igaz.
3. A kísérleti hipotézist (vagyis hogy van különbség a populációs átlagok között) maradéktalanul bizonyítottam.
4. Az eredmények alapján ki tudom számolni annak a valószínűségét, hogy a kísérleti hipotézisem igaz.
5. Ha úgy döntök, hogy elutasítom a nullhipotézist, tudom a valószínűségét annak, hogy rossz döntést hozok.
6. Megbízható kísérleti eredményem van abban az értelemben, hogy ha sokszor megismételnénk a kísérletet, akkor az esetek 99%-ában szignifikáns eredményt kapnánk.

A helyzet az, hogy egyik állítás sem igaz. De ha tévedett, ne higgye azt, hogy egyedül van ezzel. Haller és Krauss 2002-ben német egyetemeken tanító kutató-pszichológusoknak tette fel ugyanezt a kérdést. A megkérdezett 30, módszertani tárgyakat oktató tanár 80%-a jelölte meg valamelyik állítást igaznak. Ugyanez az arány a nem-módszertani jellegű tárgyat oktatók között 90% (!) volt. Jól mutatják ezek az eredmények, hogy még a szignifikancia-teszt gyakori használói, sőt oktatói között sem teljesen világos, hogy mit jelent pontosan egy alacsony p-érték. Mivel a p ténylegesen viszonylag keveset mond, erős az a vágy, hogy valami többet lássunk bele.

De nemcsak a gyakori félre-interpretálás jelent gondot: sokak szerint inherens problémákkal is küzd a teszt. 2014 februárjában az amerikai Basic and Applied Social Psychology (BASP) című tudományos folyóirat szerkesztőségi közleményt adott ki, melyben bejelentették, hogy szerzőiktől a továbbiakban nem követelik meg a nullhipotézis-tesztek alkalmazását Trafimow [2014] . Egy évvel később, 2015 februárjában a folyóirat újabb közleményt adott ki, eszerint a továbbiakban csak akkor közlik a publikációt, ha az írás nem tartalmaz utalást nullhipotézis-tesztre Trafimow és Marks [2015]. Mindezen felül a folyóirat szerkesztői semmilyen statisztikai következtetés megadását nem követelik meg szerzőiktől, mivel „a kifogástalan módszer továbbra sem ismert".

A BASP lépése nem előzmény nélküli a tudományos folyóiratok történetében. Az International Committee of Medical Journal Editors 1988-ban kiadott állásfoglalása szerint a kutatóknak el kell kerülni a hipotézistesztekre, különösen a p-értékekre való egyedüli

támaszkodást. Fidler és társai [2004] folyóirat-kutatása szerint az 1980-as évek végén Kenneth Rothman statisztikus előbb az American Journal of Public Health, később az Epidemiology című folyóirat szerkesztőjeként a hozzá kerülő cikkekből töröltette a szignifikancia-teszteket, és csak a konfidencia intervallumokat engedte közölni.

A nullhipotézis-teszt alkalmazási gyakorlata tehát több évtizede fel-fellángoló kritikák tárgya, a BASP tiltása ennek csak egy újabb, bár a korábbiaknál szélsőségesebb állomása. Írásunkban a módszer alkalmazási gyakorlatát övező kérdések szisztematikus áttekintését adjuk. Megvizsgáljuk, hogy a különböző társadalom- és viselkedéstudományi diszciplínákban illetve az orvostudományban milyen mértékben vannak jelen a módszerrel kapcsolatos hibák. Kitérünk a statisztikai tankönyvek tárgyalásmódjára és a hazai szociológiai publikációs gyakorlatra is. Végül az áttekintésből levonható tanulságként néhány ajánlást fogalmazunk meg. A cikk lezárásaként visszatérünk a BASP tiltására, tárgyalva annak következményeit is.

2. A nullhipotézis szignifikancia-teszt használatának kritikái

A felcserélt feltétel problémája

A nullhipotézis-teszt során arra a kérdésre kaphatunk választ, hogy ha a nullhipotézis (H_0) igaz, mennyire valószínű, hogy az adott adatokat (A) tapasztaljuk. Tehát a p-érték leolvasásával arra derül fény, hogy $P(A|H_0)$ mekkora, pedig természetes módon arra lennénk kíváncsiak, hogy mennyire valószínű, hogy a nullhipotézis igaz, feltéve, hogy az adott adatokat kaptuk ($P(H_0|A)$). Ezt támasztja alá, hogy Oakes [1986] szerint az általa vizsgált pszichológusok közel 40%-a élt abban a hitben, hogy a p-érték azt mutatja, hogy milyen valószínű, hogy a nullhipotézis igaz. Ziliak és McCloskey szerint szinte a közgazdászok 100 százaléka elköveti azt a hibát, hogy az alacsony p-értékből automatikusan arra következtet, hogy a nullhipotézis valószínűsége alacsony.

A nullhipotézis sosem igaz, így könnyen elutasítható

További kritika a nullhipotézis vizsgálattal szemben, hogy szigorúan véve a nullhipotézis sosem igaz. Meehl [1978] példája szerint gondoljunk például arra, hogy az IQ-t, az iskolai teljesítményt vagy bármilyen más pl. attitűdváltozót vizsgálunk a társadalom a különböző csoportjaiban: például nők és férfiak, városban és községben élők, stb. között. Nehéz elképzelni, hogy bármilyen egészen pontos állítás (pl. a nők és a férfiak átlagos IQ-ja közti különbség 2 pont) igaz lenne.

A társadalomtudományokban elméletünket legtöbbször úgy tudjuk tesztelni, hogy (lehetőleg valószínűségi) mintát veszünk a populációból, és a mintán megmért adatokból következtetünk arra, hogy a populációs paraméter megegyezik-e azzal, mint ami az elméletünkből következne. Tipikusan, a nullhipotézis (pl. nincs különbség a csoportok között) elutasítása, azzal jár, hogy az elméletünk valamilyen alátámasztást nyer, hiszen a hipotézisteszt logikája indirekt, az elméletet vagy annak következményeit az alternatív hipotézisben fogalmazzuk meg. A probléma, hogy az elméletünket ezzel a módszerrel a cáfolat valódi veszélyének nem tesszük ki, hiszen alátámasztásához csak annyi szükséges, hogy a könnyen elutasítható nullhipotézist utasítsuk el. Meehl szerint olyasmi ez, mintha egy klímára vonatkozó teóriát úgy próbálnánk alátámasztani, hogy azt az egyszerű állítást cáfoljuk, hogy áprilisban nem esik az eső. Ehelyett kívánatosabb gyakorlat lenne, ha az elméletünk alapján valamilyen szubsztantív, minél pontosabb előrejelzést tennénk (például április 4-én 0,66 cm csapadék fog esni), és ezt tesztelnénk utána.

Tehát a nullhipotézis könnyen megoldható elutasítása miatt fennáll a veszélye, hogy semmitmondó alátámasztását adjuk az elméleteknek. Meehl szerint a probléma igen súlyos:

„Úgy gondolom, hogy 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." Meehl [1986:817].

A tudományos fontosság összetévesztése a statisztikai szignifikanciával

Ez a kritika (a továbbiakhoz hasonlóan) nem a szignifiancia-teszt valamely inherens hátrányára, hanem kizárólag egy téves használati gyakorlatra utal, mégpedig a statisztikai szignifikancia és a tudományos fontosság azonosítására. Meehl már '67-es cikkében azt írta, hogy a nullhipotézis szignifikancia-teszttel kapcsolatos problémákra ellenszer lehetne, ha a kutatók gondosan megkülönböztetnék a statisztikai szignifikanciát a szubsztantív, elméleti fontosságtól. Egyrészt ha egy eredmény szignifikáns, még lehet tudományos szempontból érdektelen. Másrészt nem is feltétlenül szükséges a statisztikai szignifikancia ahhoz, hogy egy eredmény tudományos szempontból jelentős legyen – gondoljunk csak például arra, hogy akár egyetlen megfigyelés is cáfolhat egy elméletet. Ennek ellenére Ziliak és McCloskey [2008] számos példával támasztják alá, hogy bizonyos kutatók a statisztikai szignifikanciát, mint a tudományos fontosság (gyakran egyetlen) kritériumát használják, és hogy a probléma napjainkban is fennáll.

Szignifikancia-teszt használata kifejezetten nagy mintáknál

Egy további hiba, amit a kritikusok szerint a szignifikancia-teszt alkalmazásakor bizonyos kutatók elkövetnek, hogy olyan nagy mintákra használják, melyek mellett a statisztikai szignifikancia semmitmondó. Egy szélsőségesebb változata ennek, amikor a teljes populációról rendelkezésre állnak adatok, és mégis végeznek valamilyen szignifikancia vizsgálatot. Meehl és Lykken [1990] egy 57000-es mintában a rendelkezésre álló 15 kategoriális változót hasonlították össze az összes lehetséges módon kereszttáblával, hogy rámutassanak, hogy nagy mintákat használva semmitmondó a szignifikáns eredmény - és valóban, a khí-négyzet próba mind a 105 esetben szignifikáns eredményt adott. Meehlék azt a jelenséget, hogy a társadalomtudományokban valamilyen szinten minden mindennel összefügg, zaj-fatornak (*crud-factor*) nevezik. Aki nagyon nagy mintán végez szignifikancia vizsgálatot (mint amilyenek, magyar példával élve, a KSH több tízezres felmérései, pl. a munkaerő-felmérés), az a zaj-faktor miatt borítékolhatóan szignifikáns eredményt talál.

A p-érték azonosítása a nullhipotézis valószínűségével vagy a hatásnagysággal

Már említettük, hogy rendkívül elterjedt a p-értéknek a nullhipotézis valószínűségeként való értelmezése. A p-érték egy másik félreértelmezése, amikor a hatáserősséggel hozzák összefüggésbe. Duggan már a '60-as években felhívta rá a figyelmet, hogy a kicsi p-értéket a kapcsolat erősségével azonosítani téves, de a szociológiai folyóiratokban mégis létező gyakorlat. Duggan példaként rámutat, hogy a kereszttábla elemzéskor használt khí-négyzet próbához tartozó p-értéktől függetlenül lehet a (gammával mért) kapcsolat erős vagy gyenge a változók között. Egy nagy mintával nagyon gyenge kapcsolat esetén is kicsi p-értékeket kapunk, de ez nem jelenti azt, hogy erős lenne az összefüggés. A több évtizedes figyelmeztetés ellenére a félreértelmezés manapság is megtalálható – Ziliak és 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 ez hatás erősségét mutatná.

A tesztek erejének figyelmen kívül hagyása

A nullhipotézis vizsgálatkor fontos szempont lenne a teszt erejének vizsgálata, és a teszt használóit érő gyakori kritika, hogy erre egyáltalán nem fordítanak figyelmet. Az erő azt mutatja meg, hogy mekkora a valószínűsége, hogy elvetjük a nullhipotézist, amikor az ténylegesen hamis. Cohen [1962] a Journal of Abnormal and Social Psychology című folyóirat '60-ban és '61-ben megjelent cikkeiben vizsgálta meg a tesztek erejét *post hoc* erő-elemzéssel. Nem a cikkekben lévő mintákban tapasztalt 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%-os, közepes mellett 48%-os, míg nagy hatásnagyság mellett 83%-os erőt mért a vizsgált tanulmányokban. A kis hatások esetén ez igen alacsony érték – átlagosan az esetek 82%-ában (!) a használt vizsgálatok nem tudtak volna kimutatni a hatást, és tévesen elfogadták volna a nullhipotézist. Cohen mintájára később többen megvizsgálták a pszichológiai folyóiratok cikkeiben használt szignifikancia-tesztek erejét (pl. Rossi [1990]), és hasonló átlagos erőket mértek. Ezek az eredmények arra figyelmeztetnek, hogy a mintanagyságot érdemes úgy választani, hogy erő-elemzést is végzünk előtte. Érdemes 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.

Ragaszkodás az 5%-os küszöbhöz

Bár az elsőfajú hibavalószínűség megválasztása a kutató döntése lehetne – pl. a mintavételi módszer, a mintanagyság, és a teszt erejének függvényében – a legtöbb kutató rutinszerűen a 0,05-ös szintet használja Leahey [2005]. Ezen kívül még a 0,01-es és a 0,001-es szint jelzése a legelterjedtebb.

Lehet érvelni amellett, hogy a szokásos 5%-os szint használata bizonyos objektivitást ad a kutatási adatok elemzéséhez, és szükség van ilyen objektív procedúrákra ahhoz, hogy ne az egyes kutatók szubjektív döntésein múljon a kutatás kimenetele (Schmidt és Hunter [1997]). Amellett, hogy ez miért éppen 5%, már elég nehéz lenne érveket felhozni. A küszöb megválasztása valószínűleg Fisher 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." Fisher [1956:42]. Az a gyakorlat, hogy pl. egy 5,1%-os és egy 4,9%-os p-értékkel rendelkező eredmény egészen más következtetéshez vezet, nyilván nem üdvös, ha az 5%-os szint megválasztása csak rutinból, megszokásból történik, és nincs mögötte igazi érv.

A teszt feltételeinek figyelmen kívül hagyása

Ahhoz, hogy a szignifikancia-tesztet értelmezni tudjuk, bizonyos alkalmazási feltételeknek teljesülnie kell. Egy ilyen alapvető feltétel, hogy valamilyen valószínűségi mintával rendelkezzünk, és ebből próbáljunk következtetni a populációra. Kline [2004] megállapítása szerint a társadalom- és viselkedéstudományok esetében a legtöbb esetben kényelmi mintát használnak, ennek ellenére ezeknél a kutatásoknál is elterjedt a nullhipotézis szignifikancia-teszt alkalmazása. A Leahey [2005] által vizsgált empirikus szociológiai cikkek, több mint felénél nem-véletlen mintát használtak a szerzők, pedig (a pszichológiával szemben) ezen a területen elterjedtek a survey alapú kutatások. Összetettebb modelleknél, 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 [2002] pszichológiai folyóiratok cikkeit vizsgálva azt állapította meg, hogy ezek a

feltételek az esetek nagyon kis részében kerülnek ellenőrzésre. Kline hasonló következtetésekről számol be oktatáskutatatási illetve beszéd-, nyelv- és halláskutatással foglalkozó folyóiratokat kapcsán.

Szignifikanciavadászat

A Freedman által szignifikanciavadászatnak, Kline által horgászexpedíciónak nevezett rossz gyakorlat során a kutató csak az adatok megtekintése után dönti el, hogy mely adatokat ellenőrzi. Selvin [1957] megállapítja, hogy a szociológusok között talán a legelterjedtebb adatinterpretációs probléma, hogy azután fogalmazzák meg a hipotéziseiket, miután már megvizsgálták az adatokat, majd ezeket a hipotéziseket ugyanezeken az adatokon tesztelik. Feynman ugyanezt a problémát túlillesztésnek nevezi (Gigerenzer [2004]). Gigerenzer szerint a túlillesztés hibáját újra és újra elkövetik a kutatók a rutinszerű szignifikancia-tesztelés során, és a modern statisztikai programcsomagok lehetővé is teszik, hogy a változók közti összes lehetséges kapcsolatot teszteljék, addig „horgásszanak" a változók közti kapcsolatok között, amíg valami szignifikánsat nem találnak, majd ezt közöljék eredményként.

A „méret-nélküliség" kritikája

Az irodalomban talán a leggyakrabban említett kritika a Ziliak és McCloskey által „méret nélküli" (*sizeless*) tudománynak nevezett jelenséghez kapcsolódik. Ez összefoglalva annyit tesz, hogy a kutatás során a hatásnagyságot nem vizsgálják, és dichotóm döntéseket hoznak aszerint, hogy a nullhipotézis-teszt szignifikáns vagy nem. Ha a hatásnagyságot közlik is, akkor is előfordul, hogy nem interpretálják. Jellemzően például egy lineáris regresszió elemzést alkalmazva az értelmezésnél nem vizsgálják a regressziós együtthatók nagyságát, hanem csillagokkal, vagy félkövérrel jelzik, hogy melyik változónál észleltek szignifikáns hatást (Meehl [1978]). Ennek egy másik változata, amit Ziliak és McCloskey az „előjel-tudomány" kifejezéssel illetnek, amikor a kutatók a változók közötti kapcsolatnak csak az irányát interpretálják (szignifikáns pozitív vagy szignifikáns negatív kapcsolat), de a nagyságát nem. És ezután esetleg – egy korábban említett hibát elkövetve, a tudományos relevanciát a statisztikai szignifikanciával összetévesztve – a legkisebb p-értékekkel rendelkező változók fontosságát emelik ki. Kifejezetten gyakori ez az eljárás sok változót lefedő, feltáró jellegű kutatásoknál, amikor ezen az úton, egyfajta adatbányászati módszerként szűrik ki a szignifikáns kapcsolatokkal bíró változókat.

Számos szerző érvel amellett, hogy a dichotóm döntések (van összefüggés-nincs összefüggés, hipotézis elvetése-elfogadása) helyett hatásnagyságok becslése lenne a fontos, a hozzájuk tartozó konfidencia-intervallumok közlésével (Gardner [1986]). Annak, hogy a hatásnagyságot vizsgálni lehessen, előfeltétele, hogy a változóinkat általunk interpretálható egységekben mérjük, és hogy el tudjuk dönteni, hogy a tapasztalt hatásnagyság mennyire számít nagynak. Ez a döntés már alapvetően nem statisztikai, hanem szakmai kérdés.

3. Tudományterületek eltérései a szignifianciateszt téves gyakorlatában

Az egyes tudományterületek között nagy különbségek találhatók a szignifianciateszt téves gyakorlatának reflexiójával kapcsolatban.

Az **orvostudományok** területén, többek között a bevezetőben említett Ken Rothman szerkesztőnek köszönhetően a teszt túlértékelése a 20. század végére visszaszorult.

A **pszichológia** területén a kritikák már hamarabb megjelentek, mint az orvostudománynál (lásd pl. Meehl említett kritikáit). A '90-es években kezdték el bizonyos folyóiratok irányelveiket megváltoztatni annak érdekében, hogy a „méret nélküli" szignifikancia-

vizsgálatok helyett hatásnagyságokról és konfidencia intervallumokról is írjanak a publikálók. 2004-re 23 olyan pszichológiai folyóirat volt, ahol a szerkesztői irányelvek között kiemelték a nullhipotézis szignifikancia-teszttel kapcsolatos lehetséges buktatókat, és bátorították a hatásnagyságok és a konfidencia-intervallumok közlését. Az amerikai Pszichológiai Társaság publikációs kézikönyvében is megjelent '94-ben erre vonatkozó ajánlás. Ezeknek az intézkedéseknek azonban nem volt jelentős hatása a cikkekre: Cumming és társai 10 vezető pszichológiai folyóiratot vizsgálva 1998 és 2006 között megmutatták, hogy továbbra is jellemző, hogy a cikkek nullhipotézis szignifikancia-teszteket használnak (több, mint a cikkek 95%-a hagyatkozik erre), és a konfidencia-intervallumokat még 2006-ban is csak a cikkek 10%-ában közölték.

A **szociológiá**ban is népszerű a nullhipotézis szignifikancia-teszt használata. Leahey 1935 és 2000 közötti tanulmányokat vizsgált, 20 jelentős szociológia folyóirat cikkeinek 613 fős mintáján. Ebben az időszakban – az olyan tanulmányok között, ahol ez lehetséges volt – a cikkek 81%-ában használtak valamilyen szignifikancia-vizsgálatot. Ez az arány a '30-as évektől az '50-es évek végéig folyamatosan nőtt 30%-ról 80%-ra. Innentől kezdve a '70-es évekig – valószínűleg az akkoriban virágzó kritikák hatására (lásd pl. Duggan) – 60% körüli értékre esett vissza, majd újra növekedésnek indult és a '90-es években 90% körüli szinten mozgott. Leahey a tesztek előretörését a 70-es évektől alapvetően a statisztikai szoftverek elterjedésének tudja be.

Bár Leahey cikkének fókusza nem a nullhipotézis szignifikancia-teszttel kapcsolatos rossz gyakorlatok elterjedtségén van, kiderül, hogy a statisztikai szignifikancia-vizsgálatot alkalmazó cikkek 10%-ánál használtak egyszerű véletlen mintavételt, 30%-uknál más valószínűségi mintavételt, 6%-uknál a teljes populációról rendelkezésre álltak adatok, és a maradék 54%-nál valamilyen nem valószínűségi mintavételre hagyatkoztak. Vagyis a teszt is, és a véletlen mintára vonatkozó feltételeinek figyelmen kívül hagyása is igen elterjedt volt ebben az időszakban. Ezen kívül általánosnak volt mondható az a gyakorlat, hogy az 5, 1 vagy 0,1%-os küszöbhöz ragaszkodtak a cikkírók a nullhipotézis szignifikancia-teszteknél, más szempontokat, például a mintanagyságot figyelmen kívül hagyva.

A **közgazdaságtan** publikációs gyakorlatát vizsgálva Ziliak és McCloskey az American Economic Review (AER) '80-as illetve '90-es években megjelent cikkeit vizsgálták, azokra a publikációkra koncentrálva, amelyekben valamilyen regressziós modellt használtak a szerzők. A leggyakoribb hiba, melyet mindkét vizsgált időszakban az AER szerzőinek több, mint 90%-a elkövetett, hogy nem vették figyelembe a tesztek erejét. A Ziliakék által az egyik legsúlyosabbnak tartott problémát, a statisztikai szignifikanciának a tudományos fontosságként való értelmezését is elkövette a szerzők több, mint harmada.

4. A nullhipotézis szignifikancia-teszt téves használatának gyökerei

Tudományszociológiai okok

Ziliakék szerint a sziginifikancia-vizsgálat elterjedtségének, illetve annak, hogy tévesen tudományos fontosságként kezelik a statisztikai szignifikanciát a társadalom- és viselkedéstudományok képviselői, az is oka lehet, hogy ezeknek a „puha" tudományoknak a szignifikancia-teszt bizonyos biztonságot nyújt, csökkenti kisebbségi komplexusokat a „keményebb" tudományokkal szemben. A számolás és a behatárolt szabályok (például az 5%-os küszöb) objektívnek, „tudományosnak" mutatja az eredményeket.

Az is nyilvánvaló kiváltója lehet a nullhipotézis szignifikancia-teszt elterjedtségének, hogy a statisztikai szoftverek elterjedése és a számítógépes kapacitások intenzív növekedése óta ilyen teszteket rendkívül könnyű végezni.

Publikációs gyakorlat: mit érdemes közölni

Sterling, Rosenbaum és Weinkam 1995-ös tanulmányukban (Sterling és társai [1995]) 4 nagy pszichológiai folyóirat ’86-’87-es cikkeit vizsgálva azt találták, hogy azon cikkekben, melyek használnak valamilyen nullhipotézis szignifikancia-tesztet, a nullhipotézis az esetek 94%-ában elutasításra kerül. Matematikailag könnyen megmutatható, hogy ezek a publikált eredmények nem képezhetik reprezentatív mintáját az összes eredménynek. Ez arra a publikációs torzítás néven ismert jelenségre mutat, hogy a folyóirat-szerkesztők hajlamosabbak elfogadni a statisztikailag szignifikáns eredményeket bemutató cikkeket, illetve, hogy már eleve inkább az ilyen eredményt produkáló cikkeiket próbálják publikálni a kutatók. Sterlingék szerint a gyakorlat egyik fontos következménye, hogy azt a hamis látszatot kelti, hogy a tudományos fontosság szoros kapcsolatban áll a statisztikai szignifikanciával, hiszen arra ösztönzi a kutatókat, hogy végezzenek szignifikancia-vizsgálatot (esetleg akkor is, ha nincs jelentősége) és csak a statisztikailag szignifikáns eredményeiket tartsák tudományosan fontosnak.

Tankönyvek

A nullhipotézis szignifikancia-teszt vitájában 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 és McCloskey szerint rendkívül kevés ökonometria tankönyv különbözetheti meg a statisztikai szignfikanciát a tudományos fontosságtól, és úgy gondolják, hogy a statisztika tankönyvek többsége implicite elköveti a felcserélt feltétel hibáját. Thomas Scheff [2011] négy friss tankönyvet vizsgált, és úgy találta, hogy egyik sem hívja fel kellő mértékben a figyelmet a teszttel kapcsolatos problémákra és lehetséges félreértelmezésekre. Hasonló következtetésre jutunk, ha a Magyarországon használt tankönyveket (Fidy és Makara, 2005; Hunyadi és Vita, 2008; Lukács, 2002; Tómács, 2012) tekintjük. Ellenpélda is van ugyanakkor, lásd például a Freedman és társai [2005] jegyzetét. Csak néhány mondatot kiemelve: „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” (Freedman és társai [2005:623]).

5. A hazai gyakorlat: a Szociológia Szemle cikkeinek vizsgálata

A következőkben annak a kérdésnek a megválaszolására teszünk kísérletet, hogy a nullhipotézis szignifikancia-teszttel kapcsolatos rossz gyakorlatok a magyar empirikus szociológiában is jelen vannak-e. Ennek a kérdésnek azért is van jelentősége, mert a nemzetközi szakirodalomban sem tudunk olyan vizsgálatról, amikor szociológiai folyóiratot elemeztek volna ilyen szempontból. Ebben közrejátszhat az is, hogy a szociológiában a nullhipotézis szignifikancia-teszttel kapcsolatos vita a több, mint 50 éve nem lángolt fel igazán. Ekkor jelent meg a Morisson és Henkel kritikákat és ajánlásokat összegyűjtő kötete, amelyből például Duggan [1960] tanulmányát már idéztük korábban. Igaz, hogy a könyvet 2006-ban újra kiadták, de ezen kívül nem sok jele mutatkozik annak, hogy a szociológia területén olyan erősen jelen lenne a nullhipotézis szignifikancia-teszttel kapcsolatos diskurzus, mint a pszichológia vagy a közgazdaságtan esetében.

A Szociológia Szemle 2000 és 2014 közötti számaiban megjelenő tanulmányokat vizsgáltuk, azokra a cikkekre koncentrálva, amelyeknél valamilyen regressziós módszert használtak a szerzők. 38 ilyen cikk jelent meg a folyóiratban a 15 év alatt, és egyértelműen ez volt a legelterjedtebb módszer az empirikus szociológiai cikkek között. Ezen belül OLS regressziót,

logisztikus regressziót, OLS regresszión alapuló útmodellt illetve többszintű lineáris regressziót használtak a szerzők. Egy olyan cikk volt, ahol ugyan regressziót alkalmaztak, de szignifikanciát nem vizsgáltak, mert az adatok a teljes populációra vonatkoztak. Ezért ezt a cikket nem vettük bele a vizsgálatba, ami így végül 37 cikkre vonatkozik.

A tudományos fontosság összetévesztése a statisztikai szignifikanciával

Egy erre utaló jel, ha a nem szignifikáns eredményeket automatikusan ignorálják a szerzők. Csak két olyan szöveg volt a 37 közül, ahol ez nem történt meg, és a nem szignifikáns kapcsolatokkal is részletesebben foglalkoztak. Az egyik cikkben ezek közül a szerző figyelembe vette azt is, hogy kis minták nehezebben produkálnak statisztikailag szignifikáns eredményt. Más cikkekben gyakran találkozni olyan mondatokkal, amikből arra lehet következtetni, hogy valamennyire a szakmai fontosság („érdemi összefüggés", „megjegyzésre érdemes jelenség", „említésre méltó szerep") a statisztikai szignifikanciával definiálódik.

A szignifikanciának a fontosságként való értékelésére mutató további jel, amikor egy modellt úgy építenek fel a kutatók, hogy eleve csak a statisztikailag szignifikáns összefüggésekkel foglalkoznak. Ennek egy példája a stepwise, forward, vagy backward módszerrel felépített regresszió, ami automatikusan kidobja a nem szignifikáns változókat. Ilyen modellek használatával két cikkben találkoztunk.

A szignifikancia-teszt használata kifejezetten nagy mintáknál, esetleg a teljes populációra

Egyetlen tanulmányban követték el azt a hibát, hogy szignifikancia alapján döntöttek olyan esetben, mikor a teljes populációról rendelkezésre álltak adatok. A cikkben a mérési egységek országok voltak, és 68 országról rendelkezett adatokkal a szerző. Ez bár nem fedi le a Föld összes országát, semmi esetre sem tekinthető egy, az országokból vett véletlen mintának (mindössze arról volt szó, hogy bizonyos országokból nem álltak rendelkezésre adatok).

Ezen kívül három olyan tanulmánnyal találkoztunk, ahol 5000-nél magasabb esetszámmal dolgoztak a szerzők. Ezek közül egy volt olyan, ahol a nagy elemszám (ebben az esetben több, mint 40 ezer fős minta) ellenére hatásnagyságokat nem vizsgáltak, csak azt, hogy szignifikáns-e a regressziós együttható, és hogy milyen irányú. Nem meglepő módon szinte az összes általuk vizsgált kapcsolat szignifikánsnak mutatkozott, viszont ennek borítékolható volta miatt az eredmények viszonylag kevéssé voltak informatívak.

A p-érték azonosítása a nullhipotézis valószínűségével vagy a hatásnagysággal

Egyetlen cikkben sem azonosították a p-értéket a nullhipotézis valószínűségével. Az már inkább előfordult, hogy úgy interpretálták a p-értéket, mintha a hatásnagyság lenne: „A változó hatása erős (p=0,05 %-os szinten szignifikáns)". Ugyanakkor olyan cikk is volt, ahol a hatásnagyságot egyértelműen megkülönböztették a szignifikanciától: „A többváltozós elemzés a kilépőkről markáns, de az elemszámnak köszönhetően kevés szignifikáns összefüggést mutatott ki."

A tesztek erejének figyelmen kívül hagyása

Egy olyan cikk sem volt a 37 között, ahol foglalkoztak volna a tesztek erejével. Ez valószínűleg annak is betudható, hogy mivel jellemzően survey vizsgálatokról volt szó, és a minta elemszáma 1000 körüli vagy annál is magasabb volt, így a tesztek ereje is igen magas kellett, hogy legyen. Viszont abban a néhány cikkben, ahol alacsonyabb volt az esetszám, szintén nem merült fel a másodfajú hibavalószínűség kiszámítása. Valószínűleg a pszichológiában vagy más olyan diszciplínák esetében, ahol általában kisebb az elemszám, és

emiatt a tesztek ereje sem nagy, a statisztikaoktatás nagyobb hangsúlyt fektet erre, így a kutatók is jobban tudatában vannak a fontosságának.

Ragaszkodás az 5%-os küszöbhöz.

Az általunk vizsgált cikkek mindegyikében 5%-os szignifikanciaszintet használtak, vagy ezen kívül az 1%-os illetve 0,1%-os p-értékeket jelölték. Az a gyakorlat is jelen volt a cikkekben, hogy a pontos p-értékeket közölték, de szintén elterjedt, hogy félkövérrel, vagy (egy, kettő vagy három) csillaggal jelezték, hogy eléri e az 5, 1, vagy 0,1%-ot a p-érték. Az 5%-os szinttől való eltéréssel csak olyan cikkekben találkoztam, ahol bár alapvetően ezt a küszöböt tekintették irányadónak, a 10%-nál kisebb p-értékkel rendelkező kapcsolatokról is beszéltek. A következő példa szépen mutatja, hogy a magyar empirikus szociológiai gyakorlatban is kitüntetett szerepe van az 5%-os 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 kissé azt a benyomást is kelti, hogy a változók kategóriái nem valamilyen teoretikus szempontból kerülnek megváltoztatásra, hanem azért, hogy sikerüljön elérni a kívánt szignifikanciát (tehát a **szignifikanciavadászat** hibája is felmerül). Így az 5%-os küszöb már nem is tűnik annyira objektívnak, hiszen a változókat lehet úgy alakítani, hogy ezt elérjük. Ugyanebben a cikkben az szokásos küszöbhöz való ragaszkodást mutatja, hogy egy lábjegyzetben öt tizedes jegyig kényszerül a szerző kiírni a szignifikanciát, hogy bizonyítsa, az nem éri el az 5%-ot.

A teszt feltételeinek figyelmen kívül hagyása

A cikkek nagy részében valamilyen valószínűségi mintával dolgoztak a szerzők. Egy olyan (már említett) cikk volt, ahol a teljes populációról rendelkezésre álltak adatok. Egy másik cikkben a mérési egységek magyarországi városok voltak, ahol az önkormányzatok által kitöltött kérdőíveket elemezték. Az önkormányzatok 24%-ától érkeztek vissza a kérdőívek, így ezen a mintán elemezték az adatokat, ez azonban nem tekinthető valószínűségi mintának. További 4 cikknél dolgoztak nem valószínűségi mintán, de minden esetben felhívták a figyelmet arra, hogy emiatt korlátozottan általánosíthatóak az eredményeik, bár a szignifikanciát ennek ellenére a szokásos módon interpretálták. A 37 között két olyan cikket találtunk, ahol a szerző ellenőrizte a regressziós modell feltevéseit.

A „méret nélküliség” 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, irányukat. Például: „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 itt annak ellenére nem közölnek a szerzők hatásnagyságot, hogy a magyarázandó változók mérési egységei könnyen értelmezhetőek (a tőke ezer forintban volt mérve), és a módszer (OLS) sem nehezíti meg az interpretációt. A hatásnagyságok interpretálása már csak azért is fontos lenne, mert bár a regressziós együtthatókat tartalmazó táblázatból kiderül, hogy a becslés szerint a családi gazdaságok 4,8 millió forinttal kevesebb tőkével rendelkeznek, mint a nem családi gazdaságok, egy nem-szakmabeli számára nehéz értelmezni, hogy ez soknak vagy kevésnek számít.

Olyan esettel is találkoztunk, amikor az interpretáció vélhetően azért nem történt meg, mert a változók skálája eleve nehezen volt értelmezhető, és alapvetően nem volt világos, hogy mi számít nagy vagy kicsi eltérésnek ezen a skálán. „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 nem is lehetséges könnyen értelmezhető mértékegységekben mérni a változókat, akkor is hasznos lett volna, ha a szerzők közlik, hány szórásnyi a különbség az egyes országok között, és hogy (korábbi tapasztalatok szerint) jelentős mértékűek-e ezek az eltérések.

Olyan tanulmány is volt a vizsgáltak között, ahol a szerző indokolta, hogy miért nem közöl 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." A hatásnagyságok közlése a korábban írtak alapján nem csak akkor lehet fontos, ha jelenségek előrejelzéséről beszélünk. Másrészt, ha az egyszerű szignifikancia vizsgálatokkal ellenőrizni lehet az elméletet, akkor az a probléma jelentkezik, amit Meehl fejtett ki részletesen: 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 megtámogatnák azt.

A kutatási hipotézisek megfogalmazása is sok esetben arra engedett következtetni, hogy a hatásnagyságokra kevés hangsúlyt fektetnek a kutatók. Továbbá több esetben csak a standardizált regressziós együtthatókat közölték és interpretálták a szerzők. Így alapvetően a változók egymáshoz képesti erőségét tudjuk megítélni, de arról nem kapunk képet, hogy önmagukban mennyire számít nagynak egy-egy változó hatása.

Konfidencia-intervallumokat nem közöltek egyetlen tanulmányban sem. Hét olyan cikk volt, ahol a regressziós együtthatókat összefoglaló táblázatban bár a konfidencia-intervallumokat nem, de a standard hibákat feltüntették zárójelben, azonban ezeket egy esetben sem interpretálták semmilyen formában, a feltüntetés inkább csak formalitásnak tűnt.

Az alábbi táblázat foglalja össze azokat az eredményeket, amiket egyértelműen számszerűsíteni lehetett. Láthatjuk, hogy az összes tanulmányban közölték táblázatos formában a regressziós együtthatók nagyságát. Azon 21 tanulmány esetében, amely valamilyen lineáris regresszión alapuló modellt használt, 9 csak a standardizált regressziós együtthatókat tüntette fel a táblázatban. A 37 cikkből 20-ban interpretálták a szövegben is a regressziós együtthatók nagyságát, ebből 5 esetben csak a változók egymáshoz képesti nagyságát vizsgálták. A cikkek többségében (30 cikknél) valamilyen valószínűségi mintán dolgoztak. Mindössze 2 tanulmányban említették explicite, hogy ellenőrizték, hogy teljesülnek-e a használt modell feltevései.

1. Tábla – A Szociológiai Szemle 15 évének regressziós modellt alkalmazó írásainak vizsgálata - azon cikkek száma, amelyekre teljesülnek a felsorolt állítások.

Cikk jellemzője	Hány cikkre teljesül
Közli (táblázatban) a regressziós együtthatók nagyságát	37
Csak standardizált regressziós együtthatókat közöl	9
A szövegben elemzi a regressziós együtthatók nagyságát	20
Csak a változók egymáshoz képesti nagyságát interpretálja	5
Valószínűségi mintán dolgozik	30
Említi, hogy ellenőrizte a modell feltevéseit	2
Közli a teszt erejét	0
Közöl konfidencia-intervallumokat	0
N	37

Összességében tehát azt láttuk, hogy minden korábban felsorolt nullhipotézis szignifikancia-teszttel kapcsolatos rossz gyakorlatra van példa a Szociológiai Szemle elmúlt 15 évének cikkeiben. A nullhipotézis szignifikancia-teszttel kapcsolatos hibák elkövetése összekapcsolódik azzal, hogy magát a regresszió elemzést (a módszer megválasztását, az eredmények gyakran más interpretációt nélkülöző táblázatos közlését) is bizonyos szempontból ritualisztikusan használják a szerzők. A szubsztantív szignifikanciával kapcsolatos deficitet a hatásnagyságok fent említett elhagyásán kívül az is mutatta, hogy az olvasó számára nem derült ki minden esetben, hogy „mire jó” az, milyen jelentősége van annak, amit a kutatók találtak.

Mindezek arra mutatnak, hogy (legalábbis a magyar) szociológia területén sem lennének haszontalanok azok a törekvések, amik az adatok interpretálásának megreformálására irányulnak. Az, hogy a nemzetközi szociológia területén több, mint ötven éve lángoltak fel utoljára a nullhipotézis szignifikancia-teszttel kapcsolatos kritikák, akár előnyére is válhat a tudománynak, hiszen mostanra sikeres (orvostudomány) és sikertelen (pszichológia) példák is állnak előtte a tekintetben, hogy hogyan lehet megreformálni a hibásan elterjedt gyakorlatokat. A következő fejezetben erre vonatkozó ajánlásokat tekintünk át.

6. Ajánlások

Az alábbiakban Kline [2004] és Harlow [1997] összefoglaló munkái alapján néhány fontos ajánlást tekintünk át, melyet az utóbbi 50-60 évben fogalmaztak meg a nullhipotézis szignifikancia-teszt túlzott és egyes esetekben félreértett használatának visszaszorítása érdekében.

Az elméletek alapos és gondos értékelése, átgondolása

Bár triviálisnak tűnhet ez az ajánlás, ha minden kutató szem előtt tartaná, akkor valószínűleg el lehetne kerülni a nullhipotézis szignifikancia-teszt rituális és mechanikus használatát. Yates már 1951-ben úgy gondolta, hogy a túlzott hangsúly, amit szignifikancia-tesztek kapnak a tudományos következtetési folyamat során, és a tesztek mechanikus használata ahhoz vezetett, hogy olyan problémák kerülnek vizsgálatra, amiknek fontossága, gyakorlati haszna megkérdőjelezhető. Egészen az utóbbi évekig számos kutató (köztük Meehl, Rozeboom, Ziliak) fogalmazta meg különböző formákban, hogy a tudományos következtetés során nem tudjuk magunkat mechanikus procedúrákra bízni, és az emberi józan ész, kritikus gondolkodás, bölcsesség és intuíció része kell, hogy legyen a tudományos következtetési folyamatnak.

A nullhipotézis szignifikancia-teszt elsődlegessége csak felderítő kutatások esetében

Kline ajánlása szerint a nullhipotézis szignifikancia-tesztek csak olyan kutatások során kellene, hogy kiemelt szerepben legyenek, amikor a kutatás nagyon felderítő jellegű, s az adott témáról még nem áll rendelkezésre kellő mennyiségű információ. Ilyen estekben valóban hasznos lehet, ha alapvetően nullhipotézis szignifikancia-vizsgálatokra hagyatkozva próbáljuk megítélni, hogy egyáltalán létezik-e a kapcsolat bizonyos változók között. Ez a feltáró szakasz miden esetben átmeneti, a tájékozódást segíti. Mikor a kutatás (vagy egy kutatási terület) érettebb szakaszába ér, akkor a nullhipotézis szignifikancia-teszt dominánsabb szerepét átveszi hatásnagyságok becslése, illetve komplexebb, pontosabb modellek illesztése a rendelkezésre álló adatok alapján. Ha például egy szociológus korábbi kutatások alapján már tisztában van azzal, hogy a magasabb iskolázottságúak jobban keresnek, akkor a kapcsolat kimutatása önmagában nem cél. Kline úgy gondolja, hogy éppen a nullhipotézis szignifikancia-tesztekre való hagyatkozás hiánya lehetne a fémjelzője annak, hogy egy kutatási terület már a felderítő, puhatolózó állapotból tovább lépett egy érettebb szakaszba.

A statisztikai erő közlése nullhipotézis vizsgálat alkalmazásakor

Volt arról szó a kritikák felsorolásnál, hogy a rendkívül kiserejű tesztek alkalmazása elterjedt a társadalomtudományokban. Ennek ellenszere lehetne, ha az *a priori* erő számításokat minden statisztikai teszt esetében közölnék. Az erő közlése akkor különösen fontos, ha a tanulmányban vannak elfogadott nullhipotézisek, hiszen így képet kaphat az olvasó a másodfajú hibavalószínűségről is. Kline azt gondolja, hogy jelenleg azért is kevéssé elterjedt a tesztek erejének közlése, mert ritkán kerülnek olyan cikkek publikálásra, melyekben a nullhipotézis elfogadásra kerülne. Az empirikus szociológiai kutatásban talán azért is fordítanak erre kisebb figyelmet, mert jellemzően nagy elemszámú mintákkal dolgoznak, amelyeknél nagyobb a statisztikai erő, mint kis mintáknál. Azonban az erő kiszámítása nem csak önmagában lehet hasznos, hanem azért is, mert elengedhetetlen a kiszámolásához, hogy a hatásnagyságokkal is foglalkozzon a kutató. A legtöbb statisztikai szoftverben már lehetőség van erő-számításokat végezni, így technikai nehézségek nem állnak a kutatók útjába, ha a tesztek erejét ki szeretnék számolni.

A „szignifikáns" szó használatának megváltoztatása

Kline szerint a $p<\alpha$ jelenség leírására választott „szignifikáns" kifejezés utólag rossz döntésnek bizonyult. A szignifikáns a köznapi használatban a fontos, erős kifejezésekkel szinonim, és így egyrészt az olvasót félrevezetheti, másrészt táptalajt adhat annak a korábbiakban ismertetett kutatók között jelen lévő rossz gyakorlatnak, hogy statisztikai szignfikanciát a tudományos fontossággal, vagy a hatásnagysággal keverik. Ahelyett, hogy a változók közötti szignifikáns kapcsolatról beszélünk, a statisztikai kapcsolat kifejezést használhatjuk, amely talán kifejezőbben leírja, amit valóban jelent.

A dichotóm döntések helyett a becslésekre való koncentrálás, a hatásnagyságok és konfidenciaintervallumok közlése és interpretálása

Ennek a fontosságára talán korábban már rávilágítottunk. Azonban ennek is átgondoltan kell történnie. Fidler és társai [2004b] „A folyóirat-szerkesztők rá tudják venni a kutatókat, hogy konfidencia-intervallumokat használjanak, de arra nem, hogy gondolkodjanak" beszédes című cikkében rávilágít, hogy a konfidencia intervallumok közlése önmagában felületes, ha az eredmények interpretálásakor nem veszik figyelembe őket.

A kutatók ösztönzése arra, hogy az eredményeik szubsztantív szignifikanciájáról is beszéljenek

Az első ajánlással rokon javaslat a fenti: mivel a nullhipotézis szignifikancia-teszt erről semmit nem mond, a kutatóknak minden esetben részletesen ki kellene térniük arra, hogy az eredményeik miért fontosak.

A kutatások megismétlése és az eredmények értékelése meta-analízis segítségével

Lykken a kutatások megismétlésének három módját különbözteti meg, és a leghasznosabbnak a konstruktív replikációt tartja, amikor az empirikus tényt, amit egy kutató talált, egy másik kutató a saját maga által legjobbnak tartott módszerekkel és mérőeszközökkel próbálja megtalálni újra (Harlow [1997]). Ha több kutatás is rendelkezésre áll, akkor lehetőség van meta-analízist végezni, vagyis szintetizálni a különböző kutatások eredményeit és így pontosabb becslésekre, következtetésekre jutni. A meta-analízist nyilván megkönnyíti, ha a korábbi ajánlások szerint a kutatók mindig közlik a hatásnagyságot és a konfidencia-intervallumokat, illetve ha a statisztikailag nem szignifikáns eredményeiket is közlik.

A statisztikai módszerek kevésbé nullhipotézis szignifikancia-teszt centrikus oktatása

Kline szerint a bevezető statisztikai kurzusok során túlságosan nagy hangsúlyt fektetnek a nullhipotézisszignifikancia-tesztre. A korábbi fejezetekben láttuk, hogy ahhoz, hogy javuljon a helyzet ezen a téren, szükséges a megfelelő tankönyvek nagyobb elterjedtsége is. Ezen kívül Kline úgy gondolja a kutatás-módszertani és a statisztikai témájú kurzusokat jobban össze kellene hangolni, mert sokszor ezeket külön kurzusokon egymástól gyakorlatilag függetlenül tanítják.

Jobb statisztikai szoftverek

Kline véleménye szerint a statisztikai szoftverek nagy része túlságosan a szignifikanciák kiszámítására koncentrál, és az outputban nem közli minden esetben a hatásnagyságokat. Valóban, segíthet, ha a gyakran használt szoftverek beépített módszerei támogatják a méret-centrikus gondolkodást, de talán még jobb megoldás, ha ezeket a szoftvereket átgondoltan használják a kutatók: nem csak a beépített módszerek és alapbeállítások közül választanak, és nemcsak az alapbeállítások szerinti outputot vizsgálják. Ehhez persze a módszerek és a szoftverek valamivel mélyebb ismerete szükséges.

Bayesi módszerek alkalmazása

Az eddigi ajánlások mindegyike olyan volt, amely frekventista keretek között is megvalósítható. Több szerző amellett érvel, hogy inkább bayesi megközelítéssel kellene a hipotéziseinket tesztelni, bár Harlow megállapítja, hogy a nullhipotézis szignifikancia-teszttel foglalkozó szerzők között sincs ebben konszenzus. A bayesi keretek többek között azért lehetnek hasznosak, mert általuk lehetőség nyílik a nullhipotézis *a poszteriori* valószínűségének kiszámítására. Korábban láttuk, hogy többen érvelnek amellett, hogy a kutatókat valójában ez a valószínűség foglalkoztatja, és ezért értelmezik félre vágyaiktól vezérleve a p-értéket.

7. A BASP tiltásának utóélete

Írásunk befejezéseként térjünk még vissza kiindulásunkhoz, a BASP tiltásához. Láthattuk, hogy a több, mint 50 éve jelenlévő kritikák ellenére a módszerrel kapcsolatos rossz gyakorlatok nem szorultak vissza, így a BASP-hoz hasonlóan valóban hasznos lehet aktívan tenni a nullhipotézis szignifikancia-teszt túlzott, illetve téves használata ellen. A BASP tiltásának utóéletét vizsgálva több reakciót is találhatunk a tágabb tudományos közösségben. Az Amerikai Statisztikai Szövetség (American Statistical Association) rövid közleményt adott közre, amiben elismerik a „következtetési statisztikai eljárások használata és interpretálása

körül kialakult problémákat” (Wasserstein [2015]) . Ugyanakkor a szövetség szerint a tiltásnak negatív konzekvenciái lesznek és a megfelelő következtetési statisztikai eljárásokról szélesebb vitát kell folytatni a tudományos közösségben.

A tiltással a brit Királyi Statisztikai Társaság is foglalkozott (Flanagan [2015]). A társaság elnöke, Peter Diggle üdvözli és osztja a BASP szerkesztőinek aggodalmait a következtetési statisztikákat illetően, viszont nem tartja konstruktívnak a teljes tiltást. Diggle rövid kritikájában kiemeli, hogy a szerkesztőségi állásfoglalás adós maradt annak magyarázatával, hogy a leíró statisztikák alapján a szerzők és olvasók hogyan vonjanak le következtetéseket.

A BASP ajánlásokat is tett a megfelelő minőségű kutatások lefolytatására. A szerkesztőség véleménye szerint a szerzőknek a szociálpszichológiában megszokott mintáknál nagyobb mintákon kell végezniük kutatásaikat, csökkentve a mintavételi hibából fakadó bizonytalanságot, ezzel elősegítve a robosztusabb eredmények megtalálását. A szerkesztőség szerint fontos, hogy a szerzők részletes leíró statisztikákat, gyakoriságokat közöljenek kutatásuk után. A szerkesztőség szerint a tiltás hatására a szerzők felszabadulnak a nullhipotézis szignifikancia-tesztek által kényszerített gondolkodási sémából és nagyobb tere lehet a kreatív gondolkodásnak. A szerkesztőség szerint a nullhipotézis szignifikancia-tesztek mellőzése nem fog a publikált írások színvonalán rontani, épp ellenkezőleg, mivel számos esetben a null-hipotézis teszt alkalmazása igazolt rossz minőségű kutatásokat.

Kérdés, hogyan implementálódnak ezek az ajánlások. Bár a tiltás óta viszonylag rövid idő telt el, a fél év alatt megjelent 13 tanulmány által betekintést nyerhetünk abba, hogy milyen közvetlen hatásai vannak a tiltásnak. A tanulmányok többsége klasszikus kísérleteket tartalmaz, elvétve találunk csak többváltozós elemzéseket. Mindegyik empirikus munkában találhatunk az ajánlásoknak megfelelő leíró statisztikákat, legtöbbször az átlagokat és szórásokat közlik. Ugyanakkor hiába a leíró statisztikák részletes közlése, azok magyarázat nélkül maradnak, csak formalitásnak tűnik szerepeltetésük.

Van olyan cikk, amely nem közöl szignifikancia-teszteket, de a biztonság kedvéért a szerzők hozzáteszik, hogy a sokat kárhoztatott $p<0,05$-ös szignifikancia szinten szignfikáns az összes eredményük és egy olyan cikk is van, amiben egy teljes t-teszten alapuló szignifikancia-teszt eredményéről olvashatunk. A szignifikancia-tesztek helyett a klasszikus kísérletek eredményeit a hatásnagyság különböző mérőszámaival (Cohen-féle d, Glass-féle delta) értékelték ki. Az interpretációjuk a tanulmányokban hasonló a nullhipotézis szignifikancia-tesztek esetében elterjedt interpretációs hagyományához, bizonyos hüvelykujj szabályok szerint történik.

Hiába kerüli a cikkek többsége a nullhipotézis szignifikancia-tesztet, a szerkesztőség által korábban nehezményezett módszertani problémák továbbra is megtalálhatók. Például a statisztikailag szignifikáns eredmények továbbra is összemosódnak a szakmailag szignifikáns eredményekkel. A különbség csupán abban figyelhető meg, hogy a szerzők nem a p-érték küszöbértékei alapján ítélkeznek, hanem például a Cohen-féle d alapján. Tehát a tudományos eredmények mechanikus előállítása továbbra is detektálható. A keményhangú szerkesztőségi állásfoglalás ellenére nemcsak, hogy megjelennek többváltozós következtetési módszerek, de egyik cikkben sem olvashatunk arról, hogy az elemzési eszközök használatához szükséges vizsgálatokat elvégezték-e volna.

Az év eleji szerkesztői állásfoglalásban arra is kitértek a szerkesztők, hogy az elterjedt gyakorlattal ellentétben szeretnék, ha a publikáló kutatók nagyobb minta elemszámmal dolgoznának kutatásaik során. Az azóta megjelent tanulmányok többsége legfeljebb 150 fős mintával dolgozik, de gyakoriak a 60-80 fős minták is, ráadásul több esetben nem-valószínűségi „kényelmi mintát” vettek a kutatók.

Elmondhatjuk tehát, hogy az állásfoglalás tiltása és ajánlásai nem valósultak meg széleskörűen a folyóirat hasábjain. Ennek vagy az eltelt idő rövidsége lehet az oka, vagy az, hogy valamilyen okból – talán a teljes tiltás radikalizmusa miatt - meghátráltak a szerkesztők.

8. Hivatkozott irodalom

Cohen, J. [1962]: The Statistical Power of Abnormal-Social Psychological Research: A Review. Journal of Abnormal and Social Psychology 65.évf. 3. sz. 145–153. old.

Cumming, G. et al. [2007]: Statistical reform in psychology. Is anything changing? Psychological Science 18. évf. 3. sz. 220-232. old.

Duggan, T. J. [2006]: Common misinterpretations of significance leveles in sociological journals. In The Significance Test Controversy: A Reader. Második kiadás. Aldine ,Chicago..

Fidler, F. et al. [2004]: Editors Can Lead Researchers to Confidence Intervals, but They Can't Make Them Think: Statistical Reform Lessons from Medicine. Psychological Science 15. évf. 2.sz. 119–126.

Fidler, F. [2005]: From Statistical Reform to Effect Size Estimation: Statistical Reform in Psychology, Medicine and Ecology. Doktori disszertáció, The University of Melbourne. http://www.cultureofdoubt.net/download/docs_cod/Book%20statistical%20significance,%20epidemiology%20%28med%29.pdf. : 2015.04.13.

Fisher, R. A. [1956]: Statistical methods and scientific inference. Oliver & Boyd, Edinburgh.

Flanagan, Oz. [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 2015. július 6-án .

Freedman, D. et al ([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, 292 évf. 6522 sz. 746-750 old.

Gigerenzer, G. [2004]: Mindless Statistics. Journal of Socio-Economics 33. évf. 5. sz. 587-606. old.

Haller, H. – Krauss, S. [2002]. Misinterpretations of Significance: A Problem Students Share with Their Teachers? Methods of Psychological Research Online 7. évf. [1. sz.].

Harlow, L. [1997]. What if there were no significance tests? Mahwah, N.J.: Lawrence Erlbaum Associates.

Kline, R. B. [2004]. Beyond significance testing: reforming data analysis methods in behavioral research. Washington DC, USA: American Psychological Association.

Leahey, M. [2005]. Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology. Social Forces 84. *évf. 1. sz.: 1-24. old.*

Meehl, P. E. [1967]: Theory-Testing in Psychology and Physics: A Methodological Paradox. Philosophy of Science 34 évf. 2. sz. 103–115. old.

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 46 évf. : 806–34. old.

Meehl, P. E. [1990]: Why Summaries of Research on Psychological Theories Are Often Uninterpretable. *Psychological Reports* 66. évf.144sz.: 195–244.

Oakes, M.W. [1986]: Statistical inference: a commentary for the social and behavioural sciences. J. Wiley & Sons, Inc, Chichester.

Osborne, J. W. – Waters, E. [2002]: Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, 8. évf. 2. sz. http://www-psychology.concordia.ca/fac/kline/601/osborne.pdf [2015.04.13.].

Rossi, J. [1990]: Statistical Power of Psychological Research: What Have We Gained in 20 Years? Journal of Consulting and Clinical Psychology 58. évf. 646–56. old.

Rozeboom, W. W. [1960]: The fallacy of the null-hypothesis significance test. Psychological Bulletin, 57 évf. 416-428. old.

Schmidt, F. – Hunter, J. [1997]: Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: *What if there were no significance tests*? Mahwah, N.J.: Lawrence Erlbaum Associates.

Scheff, T. [2011]: The Catastrophe of Scientism in Social/Behavioral Science. Contemporary Sociology: A Journal of Reviews: 264-268. old.

Selvin, H. C. [1957]: A Critique of Tests of Significance in Survey Research. American Sociological Review 22. évf. 5. sz. 519-527. old.

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 49. évf. 1. sz. 108-108. old.

Trafimow, David [2014]: "Editorial." Basic and Applied Social Psychology 36. évf.1. sz.1–2. old. [].

Trafimow, David, – Michael Marks [2015]: "Editorial." Basic and Applied Social Psychology 37. évf. [1.sz.1–2. old. http://www.tandfonline.com/doi/pdf/10.1080/01973533.2015.1012991 [http://dx.doi.org/10.1080/01973533.2015.1012991] 2015. július 6..

Wasserstein, Ronald [2015]: "ASA Comment on a Journal's Ban on Null Hypothesis Statistical Testing - American Statistical Association." http://community.amstat.org/blogs/ronald-wasserstein/2015/02/26/asa-comment-on-a-journals-ban-on-null-hypothesis-statistical-testing 2015. július 6. .

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 46. évf. 253. sz.19–34. old.

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.