Szociológiai Szemle 27(1): 118.

SZAK-MA

Beszámoló a Módszeresen című rendezvénysorozatról

Bevezető

A *Szociológiai Szemle* örömmel vesz részt minden, a szakmai közéleti diskurzus formálására tett kezdeményezésben, és feladatunknak tartjuk, hogy ehhez nyílt fórumot teremtsünk. Ezúttal a Módszeresen előadás-sorozat első három alkalmáról szóló beszámolókat közöljük. Ezeket terveink szerint továbbiak követik majd, s reméljük, hogy mindez reflexióra, vitára ösztönzi a szélesebb szakmát.

A *Módszeresen* című szociológiai metodológiai előadás- és vitasorozat 2016 őszén indult.[1] A szervezők (Janky Béla, BME-MTA; Gárdos Judit, MTA; Németh Renáta, ELTE TáTK; Szakadát István, BME) tervei szerint az előadások és viták olyan módszertani kérdésekre koncentrálnak, amelyek alapvető fontosságúak, és melyekkel kapcsolatban új eredmények születtek a közelmúltban és/vagy élénk viták zajlanak a nemzetközi tudományos közösségben. A szervezők fel szeretnék hívni a gyakorló kutatók figyelmét azokra a területekre, amelyekben megfontolandó bizonyos korábbi kanonizált gyakorlatok felülvizsgálata – de nem szeretnék, hogy a sorozat régiek helyett új kánonokat jelöljön ki. Ahogy beharangozójukban írják: „Célunk, hogy a hallgatók kanonizált szokások követése helyett képesek legyenek az adatok, eszközök és következtetések kritikai megközelítésére, módszereik megújítására saját korlátaiknak és az általuk képviselt tudomány határainak szem előtt tartásával. Azt gondoljuk, hogy a módszerekről való párbeszéd a szociológia mint diszciplína intellektuális jövője szempontjából lényeges."

A szervezők várják a gyakorló kutatókat és a doktori iskolák, szakkollégiumok diákjait is. A *Módszeresen* sorozat honlapján (http://co.o-o-o.hu) az érdeklődők megtalálhatják az előadások felvételeit, diáit, az előadók által kijelölt felkészítő olvasmányokat, és a fórumokon lehetőségük van hozzászólásra is.

1 A sorozat támogatói: MTA TK „Lendület" Research Center for Educational and Network Studies (RECENS); BME GTK Szociológia és Kommunikáció Tanszék; ELTE Társadalomtudományi Kar; MTA TK Szociológiai Intézet Módszertan és Kutatástörténet Osztály; OTKA K 115644 (Kovács Éva).

Szociológiai Szemle 27(1): 119–125.

A statisztikai szignifikanciateszt rítusa – kortárs kritikák; a rítus a szociológiában

Bárdits Anna – Németh Renáta
barditsanna@gmail.com; nemethr@caesar.elte.hu

„Anyone knowing the problems, as described over one hundred years, who continues to teach, use or publish significance tests is acting unethically, and knowingly risking the damage that ensues."

„Bárki, aki a több mint száz éve ismert problémák tudatában továbbra is szignifikanciateszteket tanít, használ vagy publikál, etikátlanul viselkedik, tudatosan kockáztatva, hogy kárt okoz."

Stephen Gorard, Sociological Research Online, 2016. február

Bevezető

A rendezvénysorozat első alkalma *A statisztikai szignifikanciateszt rítusa – kortárs kritikák; a rítus a szociológiában* címet viselte. A vitaindítót Németh Renáta (ELTE TáTK) tartotta, akinek társszerzőivel nemrég jelent meg összefoglalója a témában (Bárdits–Németh–Terplán 2016). Ez a cikk áttekintést is közöl a *Szociológiai Szemle* cikkeinek szignifikanciateszttel kapcsolatos téves gyakorlatairól, a vitaindító ennek a review-nak az eredményeiből is válogatott. A felkért hozzászólók Bartus Tamás (BCE) és Ferenci Tamás biostatisztikus (Óbudai Egyetem) voltak, a vitát Janky Béla moderálta.

A vitaindító szerint évtizedek óta jelen vannak kritikák a nemzetközi tudományos diskurzusban a szignifikanciateszttel kapcsolatban, de az utóbbi évek sosem látott kiélesedést hoztak. Csak néhány példát kiemelve: 2014-ben jelent meg a *Nature*-ben R. Nuzzo *Scientific Method: Statistical Errors* című nagy figyelmet keltő cikke, amely a *p*-érték problémáit s az alkalmazásához kapcsolódó rossz beidegződéseket taglalja, és azóta is a folyóirat egyik legidézettebb cikke. 2015-ben a *Basic and Applied Social Psychology* folyóirat szerkesztőségi állásfoglalásában (Trafimow–Marks 2015) mindenfajta következtetési statisztika használatának közlését megtiltotta szerzőinek. 2016-ban az *American Statistical Association* történetében egyedülálló módon állásfoglalást adott ki a témában, a legnagyobb nemzetközi szaktekintélyeket sorakoztatva fel szerzőként (Wasserstein–Lazar 2016). A brit *Sociological Research Online* 2016-os első három száma vitasorozatot közölt a téma szociológiai vetületéről (Gorard

2016; Nicholson–McCusker 2016; Spreckelsen–van der Horst 2016). A vitaindító értékelése szerint ez a nemzetközi figyelem jó alkalom a kérdés hazai nyilvánosságba emeléséhez.

A szignifikanciateszt logikája

Azért, hogy megértsük a *p*-értéket körülvevő vitákat, Németh Renáta előadásában felelevenítette a teszt logikáját.[1] Képzeljük el tehát, hogy a mintánkba került nők átlagkeresete alacsonyabb a mintánkba került férfiak átlagkereseténél. Kérdés: a különbség létezik-e a valóságban a populációban is, vagy csupán a mintavétel véletlen volta okozta a mintabeli eltérést. A szignifikanciateszt egy valószínűségi eljárás ennek megválaszolására. A nullhipotézisünk az, hogy a valóságban nincs különbség, az ellenhipotézisünk, hogy a mintánkban tapasztalt keresetkülönbség a valóságban is létezik. Már itt érdemes megjegyezni, hogy a terminológia nem igazán megnyugtató, hiszen éppen az „ellenhipotézis" az, amit bizonyítani akarunk. Ennek oka, hogy – mint látni fogjuk – indirekt következtetési sémát alkalmazunk.

Az alapgondolat szerint ha a megfigyelt eltérés „túl nagy", azt nehéz a véletlennel magyarázni. Az ún. *p*-érték (vagy szignifikancia) a „túl nagy" eltérés megítélésére szolgál. A *p*-érték-definíció szerint annak a valószínűsége, hogy a kapott – vagy még annál is nagyobb – eltérést kapjuk abban a világban, amelyben a nullhipotézis fennáll. Tehát minél kisebb a *p*-érték, annál inkompatibilisebbek az adatok ezzel a világgal/annál erősebb a nullhipotézis elleni bizonyítékunk. Ha nagyon kicsi – a megegyezéses küszöb szerint 5% alatti – a *p*-érték, akkor elvetjük a nullhipotézist. Vagyis – a példánkhoz visszatérve – ilyenkor arra a következtetésre jutunk, hogy elég erős a bizonyítékunk a keresetegyenlőség ellen.

Érdemes megjegyezni, hogy R. A. Fisher, a teszt megalkotója nem törekedett ilyen küszöb meghatározására, sem arra, hogy elmélete bináris döntéseket alapozzon meg. A *p*-értéket ehelyett egyszerűen egyfajta informális eszközként tekintette, ami annak megítélésében támogatja a kutatót, hogy eldöntse, érdemes-e a mintában talált eltérés további vizsgálatra. Ahogy írta: *„No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas"* (Fisher 1956). [„Nincs olyan tudós, akinek egyetlen fix szignifikanciaszintje lenne, amelyhez évről évre, minden körülmények között ragaszkodna. Ehelyett ezt a bizonyítékai és elgondolásai fényében minden egyes esetben külön választja meg."]

1 Történetileg a teszt itt bemutatott változata a Fisher-féle szignifikanciateszt és a Neyman–Pearson-féle hipotézisteszt ötvözete; ezt találjuk a legelterjedtebb szociológiai alkalmazásokban is, lásd pl. az oktatásban gyakran használt Freedman–Pisani–Purves-féle Statisztika jegyzetet.

A *p* interpretációja

Bonyodalmasnak tűnhet a *p*-érték fenti definíciója, és valóban az. Az egyszerűbb értelmezések pedig sajnos nem jók. Az alábbiakban azokat az interpretációs típushibákat vesszük sorra, melyeket a teszt gyakorlati alkalmazásának kritikusai is gyakran citálnak.

Nem igaz, hogy a *p* a nullhipotézis valószínűsége. Sőt: e klasszikus keretben egyáltalán nincs rá mód, hogy meghatározzuk, milyen valószínűséggel igaz a nullhipotézis.

Láttuk: a teszt indirekt bizonyítási sémát követ, az ellenhipotézist akarjuk bizonyítani. Ezért a $p>5\%$ nem a nullhipotézis bizonyítéka! Ez sem jó: „elfogadtuk a nullhipotézist" – mert az eltérés bizonyítékának hiánya nem bizonyíték az eltérés hiányára. (Ha tényleg az azonosságot akarjuk bizonyítani, ellentétes logikájú ún. ekvivalenciatesztet kell alkalmazni, ahol a nullhipotézis az, hogy az eltérés meghalad egy releváns értéket.)

Továbbá, a fentebb jelzett szokásos 5%-os döntési küszöbnek elméleti alapja nincsen, ezért nem érdemes hozzá minden helyzetben ragaszkodni. Nyilván nem észszerű pl. $p=0{,}045$ és $p=0{,}055$ között különbséget tenni (ahogyan azt a *Szociológiai Szemle* egy szerzője impliciten tette: *„a vallásossághoz tartozó szignifikancia 0,49995"*). Éppen ezért nem elég a $p<>5\%$ közlése, a konkrét *p*-érték informatívabb.

A szignifikanciateszt az eltérést tehát a mérési hibához viszonyítja. Ennek meghaladása azonban nem implikál szakmai jelentőséget! A szakmai jelentőség megítéléséhez további fogódzó kell – például hatásnagyság-mutatók (egyszerűen az eltérés nagysága vagy a Cramer-féle V a kereszttábláknál, vagy a regressziós B együttható). Ha hatásnagyság-mutatóként az egyszerű eltérést vagy más értelmezhető mutatót tudunk használni, érdemes annak konfidencia-intervallumát is közölni. A konfidencia-intervallum ismertetése a rituális $p<>5\%$ döntések helyett a hatás tényleges megítélése, a szakmai mérlegelés felé irányítja a gondolkodást, vagyis kibillenti a kutatót a teszt rituáléjából.

Néhány, a statisztikai és szakmai jelentőség azonosításán alapuló típushiba a *Szociológiai Szemle*ből:

„A magyarországi modellben a független változók közül a legerősebb hatást az iskolázottság fejt ki a státusra, az egyetemi/főiskolai végzettség 0,582, a középiskolai 0,38, a szakmunkásképző/szakiskolai végzettség 0,179-es béta-értékkel." A változók egymáshoz képesti relatív erősségét megismerjük, de a hatásnagyságot nem, így az idézetből nem derül ki, hogy az iskolázottság hatása mennyire erős szakmailag.

Másutt: „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." Szintén nem kapunk a hatásnagyságról képet, annak ellenére, hogy könnyen értelmezhetően forintban kifejezhető lenne.

A szociológiában ugyanakkor gyakran használunk olyan kompozit indexeket, melyeknek skálája nincsen, ezért az eltérés nagysága sem ítélhető meg közvetlenül: „A két nyugat-európai országban szignifikánsan nagyobb az egyénenkénti posztmaterialis-materiális veszélyekért aggódás különbségét mérő bizonytalanság fókusza változó átlaga. Ez azt jelenti, hogy a franciák és a britek inkább fókuszálnak a posztmaterialis, globális ökológiai veszélyekre, mint a magyarok vagy a görögök (a materiális veszélyekkel összevetve)." A statisztikai szignifikancia önmagában ezekben az esetekben sem jelent szakmai fontosságot. A franciák és britek magyaroktól és görögöktől vett távolságának nagyságrendje nem ítélhető meg, ugyanakkor, mivel skála nélküli indexet használ a kutatás, a hatásnagyság az előző, forintosítható példával szemben nehezebben lenne megadható.

Ha azt a döntési sémát követjük, hogy 5%-nál kisebb *p*-értéket látva elvetjük a nullhipotézist, az eljárás megítéléséhez fontos tudatában lenni a döntéshez kapcsolható hibavalószínűségeknek. Az elsőfajú hibavalószínűség, vagyis annak a valószínűsége, hogy szignifikánsnak ítéljük az egyébként a populációban nem fennálló eltérést, 5%-on rögzített definíció szerint. A másodfajú hibavalószínűség, vagyis annak a valószínűsége, hogy nem találjuk szignifikánsnak a populációban egyébként meglevő eltérést, nincs rögzítve, értéke tesztről tesztre változik. Így a teszt ereje (ami a másodfajú hiba el nem követésének valószínűsége) is változik. Pedig (az indirekt sémából adódóan) épp a teszt ereje az, ami fontos nekünk! Hiszen kereseti diszkriminációra/közpolitikai programunk sikerességére/az idegenellenes attitűd erősödésére stb. gyanakszunk, és szeretnénk tudni azt kimutatni. Tehát az erő a szociológiai praxisban méltatlanul mellőzött mutató. A teszt ereje elsősorban a mintanagyságtól függ. Kis mintáknál számolnunk kell azzal, hogy az erő kicsi, így akár fontos eltérések sem kimutathatók. Vagyis a kis minta kevéssé informatív. Azaz nem szignifikáns eredmény kis minta mellett más interpretációt kíván, mit nagy minta mellett – utóbbi esetben tulajdonképpen kizárható az eltérés. De az az értelmezés sem helyes, amit a *Szociológiai Szemlé*ben olvashattunk egy helyen – *„A többváltozós elemzés a kilépőkről markáns, de az elemszámnak köszönhetően kevés szignifikáns összefüggést mutatott ki."* –, hiszen nem lehetünk biztosak benne, hogy nagyobb minta mellett megismételve a vizsgálatot továbbra is markáns összefüggést látnánk!

Ha pedig nagyon nagy a minta, hatalmas lesz az erő, így akár egy szociológiailag irreleváns kis eltérés is statisztikailag szignifikáns lesz. A társadalomtudományokban valamilyen szinten minden mindennel összefügg (működik egyfajta „zajfaktor"), ezért talált a *Szociológiai Szemle* egyik szerzője több tízezres mintanagyság mellett számtalan szignifikáns összefüggést: *„A politikai nézetek, az idegenellenesség, a tradicionális női szerepek elfogadása, valamint a demokrácia működésével és a saját élettel való elégedettség tekintetében szignifikáns összefüggéseket találtunk."* Tehát szignifikáns eredmény nagy minta mellett más interpretációt kíván, mit kis minta mellett – utóbbi esetben csak nagyon erős eltérés tud statisztikailag is szignifikáns lenni.

A szignifikanciatesztet érő kritikák

A kritikák első két csoportja módszertani (pl. ragaszkodás az 5%-os küszöbhöz), ill. interpretációs jellegű (pl. a p azonosítása a hatás erősségével), az idetartozó típushibákat fent felsoroltuk. További módszertani probléma, hogy a szignifikanciatesztnek túlreprezentált szerepe van a kutatásokban, miközben ez a kutatási területek feltáró szakaszában lenne csak indokolt. „Érettebb" szakaszban a hatásnagyságok fontosabbak lennének.

A kritikák másik fontos csoportja tudományszociológiai hátterű problémákat vet fel. Ilyen probléma a publikációs torzítás, vagyis hogy a folyóiratokban sokkal nagyobb valószínűséggel jelennek meg szignifikáns eredményt produkáló cikkek (egy kérdés vizsgálatánál felmerülhet, hogy adott témában talált nem szignifikáns eredmények valahol az asztalfiókokban hevernek, ezért sosem jutnak tudomásunkra). További, ettől nem független probléma az elemzési torzítás, melyet szignifikanciavadászat néven ismerhetünk. Erre példa, ha már az adatok ismeretében állítjuk fel a hipotézist, ha több teszt közül mazsolázzuk ki a szignifikánsat, vagy ha a változóinkat és modelljeinket újradefiniáljuk abból a célból, hogy szignifikáns eredményt kapjunk.[2] Utóbbira egy példa a *Szociológiai Szemle*ből: *„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)."*

A kritikák emellett azt is felvetik, hogy nem helyes, ha tudományos konklúziókat vagy közpolitikai döntéseket egyedül a $p<5\%$ kritériumhoz kötünk. Ehhez további kontextuális tényezők kellenek, mint a mérés minősége, a design megítélése, külső bizonyítékok, illetve általában a teljes kutatási jelentések elérhetővé tétele és a kutatások átláthatósága.

Az előadást követő vita fontosabb megállapításai

A szignifikanciateszt problémáinak számos és összetett oka van, így a megoldás sem egyértelmű. Az egyetemeken folyó statisztikai oktatás, a használt statisztikai tankönyvek, illetve szoftverek is újraerősíthetik a *p*-értékkel kapcsolatos rossz beidegződéseket. Ezenkívül pszichológiai oka is van annak, hogy sokszor rosszul használjuk a tesztet: erős a leegyszerűsítés vágya, a „szürkét" hajlamosak vagyunk feketébe vagy fehérbe transzformálni. A probléma tudományszociológiai okai, hogy a folyóiratok a „pozitív" eredményeket preferálják, és a kutatók is azt gondolják, hogy akkor tudnak érvényesülni, ha szignifikáns eredményeket produkálnak. Emellett a szignifi-

2 Ide kapcsolódó jó cikk egy antropológusnak a fejlődéspszichológiai laborban szerzett tapasztalatairól: Peterson (2016).

kanciateszt körül kialakult rituálé kemény fogódzót biztosít a puha tudományoknak, ezért is nehéz elengedni.

Ha túlélőcsomagot szeretnénk kapni a szignifikanciatesztekhez, azt mondhatjuk, hogy a $p< >5\%$ információ nem elég. Mindig közöljük mellé a *p* tényleges értékét valamely hatásnagyság-mutatóval, utóbbi konfidencia-intervallumát és a statisztikai erőt, vagy legalább a *p* megítélésénél vegyük figyelembe a mintanagyságot.

És ez sem elégséges, hanem csupán szükséges feltétele a helyes következtetésnek. Mit lehet tenni azért, hogy a *p*-érték körül csoportosuló összetett problémákra megoldást találjunk és érvényesebb kutatásokat végezzünk? Erre a kérdésre keresték a választ a beszélgetés meghívott résztvevői és a közönség tagjai az előadás után. Ferenci Tamás arról beszélt, hogy az orvostudomány területén (bár korántsem tökéletes a helyzet) bizonyos szigorú feltételeknek meg kell felelniük a kutatásoknak, ami elősegíti, hogy a kutatásokban kevésbé jelenjenek meg a fent felsorolt problémák. Például erőelemzés nélkül ma már elképzelhetetlen egy gyógyszerkísérlet. Az elemzési torzítás kiküszöböléséhez említette azt a gyakorlatot, hogy a kísérletek elvégzése előtt pontosan rögzíteni kell az adatelemzés menetét és a hipotéziseket. A résztvevők arra jutottak, hogy ez ilyen szigorú formában nem megvalósítható a társadalomtudományokban, hiszen az orvostudományhoz képest itt sokkal fontosabb szerepük van az exploratív kutatásoknak, ahol nem ilyen egyértelmű előzetesen, hogy pontosan milyen hipotéziseket szeretnénk majd tesztelni. Ennek ellenére a szociológiai kutatásoknál is lehetséges, hogy már a kutatás előtt gondolkodjunk a mintanagyságról, és arról, hogy például a modellekbe bevont változók száma hogyan befolyásolja a tesztjeink erejét – emlékeztetett rá Bartus Tamás.

Egyszerű megoldást nem találtak, és nem is találhattak a résztvevők, hiszen egyszerű megoldás nem létezik – ezt jól mutatja, hogy a teszt körül évtizedek óta jelenlevő elégedetlenség ellenére továbbra is léteznek félreértelmezések és rossz gyakorlatok. Az előadás egyik konklúziója volt, hogy a fő probléma nem magával a teszttel van, hanem, Stephen Senn (az *American Statistical Association* említett állásfoglalásának egyik felkért hozzászólója) megfogalmazásban, azzal, hogy bálványként tiszteljük azt – és erre nem lehet megoldás egy újabb hamis isten, vagyis egy újabb egyszerű alternatíva. Alátámasztva a Módszeresen sorozat szervezőinek azon figyelmeztetését, miszerint „mérési eszközeink nem a bizonytalan empíriát biztos következtetésekké alakító alkímia kémcsövei". Mindenestre a jobb kutatási gyakorlatok irányába tett fontos lépés lehet, hogy a mostanában újra fellángolt kritikák nagyobb nyilvánosságba emelték a kérdést, és ebben talán része volt a sorozat ezen epizódjának is.

Irodalom

Bárdits A. – Németh R. – Terplán Gy. (2016): Egy régi probléma újra előtérben: a nullhipotézis szignifikancia-teszt téves gyakorlata. *Statisztikai Szemle, 94*(1): 52–75.

Fisher, R. A. (1956): *Statistical Methods and Scientific Inferences*. New York, NY: Hafner.

Gorard, S. (2016): Damaging real lives through obstinacy: Re-emphasising why significance testing is wrong. *Sociological Research Online, 21*(1).

Nicholson, J. – McCusker, S. (2016): Damaging the case for improving social science methodology through misrepresentation: Re-asserting confidence in hypothesis testing as a valid scientific process. *Sociological Research Online, 21*(2): 11.

Nuzzo, R. (2014): Scientific method: Statistical errors. *Nature, 506*(7487): 150–152.

Peterson, D. (2016): The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius, 2*(2).

Spreckelsen, Th. F. – van der Horst, M. (2016): Is banning significance testing the best way to improve applied social science research? – Questions on Gorard. *Sociological Research Online, 21*(3): 13.

Trafimow, D. – Marks, M. (2015): Editorial. *Basic and Applied Social Psychology, 37*(1): 1–2.

Wasserstein, R. L. – Lazar, N. A. (2016): The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2): 129–133.