Alkalmazott Matematikai Lapok **27** (2010), 157-174.

EGY NEM HAGYOMÁNYOS STATISZTIKAI ELJÁRÁS BEMUTATÁSA AZ OECD PISA ADATBÁZISON - ESETTANULMÁNY

TAKÁCS SZABOLCS

A hagyományos szimulációs technikák – mint amilyenek például a bootstrap és a jackknife eljárások – csak módosítások mellett alkalmazhatóak olyan esetekben, amikor nem egyszerű véletlen mintavétel történik. A soron következőkben bemutatunk egy olyan esetet, amikor a hagyományosan alkalmazott eljárások eredményeit egy szimulációs eljárás eredményeivel vetjük össze.

Az OECD PISA felmérések során a jackknife eljárás egy módosítását alkalmazzák [6] bizonyos hipotézisek kiértékelésére. A módszer elméleti hátterét [11] már vizsgálták. Az alábbiakban e módszer egy valódi nagymintás kutatás adatain történő bemutatására törekszünk – megmutatva a hagyományos és a szimulációs technikával számított eredmények közötti eltéréseket.

Az igazságot nem ismerve (lévén, nem generált adatokkal dolgozunk) próbáljuk értelmezni a különböző eljárásokból származó, eltérő eredményeket.

Megfogalmazunk továbbá néhány észrevételt, kritikát is, amelyek az alkalmazott módszer esetén felvetődnek és az elemzéshez kapcsolódó elérhető leírásokban nem található rájuk megnyugtató válasz.

1. Bevezető

Az OECD PISA felmérés-sorozat egy 3 évente megrendezett, OECD és OECD partner országok 15 éves diákjain elvégzett oktatáspolitikai felmérés. Ennek keretében 3 tudásterületet vizsgálnak, felmérésenként változó fókusszal. A felmérésbe kerülő iskolák száma, illetve az egy iskolából bekerülő diákok száma függ az ország nagyságától.

2000-ben a szövegértés, 2003-ban a matematikai, míg 2006-ban a természettudományos készség volt a felmérés fókuszában.

2009-ben újra a szövegértés került a figyelem középpontjába, ám a 2009-es adatok még nem nyilvánosak az elemzésünk pillanatában (vö: [18], [19], [20], [17]).

Az oktatáspolitikai felmérések során a mintaválasztás struktúrája következtében a hagyományos statisztikai eljárások helyett számítás- és így időigényes eljárásokat alkalmaznak. A mintaválasztás lényege egyfajta tömbösített rétegezésben rejlik, melynek során az adott országok nemi, települési, finanszírozási struktúráját is figyelembe veszik, és így nem egy egyszerű, véletlen mintavételezési eljárás történik.

Ráadásul a diákok súlyokat is kapnak annak elérése érdekében, hogy valóban, minden országon belül az adott ország specifikumait figyelembe véve reprezentatív mintát kapjunk.

Az elemzések során szimulációs technikákat vetnek be az országokban fellelhető, oktatás minőségét befolyásoló tényezők okozta különbségek megállapítására, azok statisztikailag szignifikáns voltának kiderítésére.

A hagyományos szimulációs technikák – mint amilyen a hagyományos bootstrap és a jackknife eljárások – csak módosítások mellett alkalmazhatóak (az alapeljárások a későbbiekben ismertetésre kerülnek). Az OECD PISA felmérések során – a reprodukálhatóság, az országonként elvégzett saját elemzések összehasonlíthatósága miatt – a jackknife eljárás egy módosítását alkalmazzák [6].

Ezt az eljárást fogjuk alkalmazni egy nem hagyományos elemzés elvégzése során [10] útmutatása alapján, a 2003-as, matematika fókuszú adatbázison.

Az adatok nyilvánosak, mind letölthetőek a [18], [19], [20] webhelyekről csakúgy, mint az elemzés elvégzéséhez szükséges eljárások és a [9] technikai segédlet.

Először a hagyományos, majd a nem hagyományos eljárások eredményeit mutatjuk be – szemléltetve azok eltérő mivoltát, ezzel némiképpen igazolva a számításigényesebb eljárás helyénvalóságát.

A két elemzésben a matematika teljesítményt fogjuk figyelemmel kísérni, és arra az egyszerű kérdésre keressük a választ, hogy a matematika teljesítmény különbözik-e a négy magyaroroszági képzési típusban.

Magyarországon az OECD PISA felmérésben résztvevő 15 éves diákok általános iskolába, gimnáziumba, szakközépiskolába vagy szakiskolába járnak.

A felsorolás sorrendje tudatos, az adatbázisban szereplő kódok sorrendjében történt. A kódolás sorrendje nem tükrözi a diákok a priori teljesítményét. Nem is szeretnénk, ha a teljesítmény alapján az iskolatípusok között valamifajta ordinalitás – és ezzel egyfajta megkülönböztetés – predesztinálva lenne.

Hasonló vizsgálatot végzett 2008-ban Slud és Thibaudeau [11], akik generált adatokon tesztelték a bemutatásra kerülő, nem hagyományos szimulációs eljárást.

1.1. ÁLLÍTÁS. *Slud és Thibaudeau azt tapasztalták, hogy rétegzett minták esetén, a [6] és [15] által leírt módszer eredményezi a valódi értékekhez leginkább közel álló becsléseket.*

A kísérletet generált adatokon végezték. A módszert ezen esettanulmányban egy valódi adatbázison alkalmazzuk a hozzá tartozó technikai leírások alapján.

2. A megválaszolandó kérdés, a probléma ismertetése

Az OECD PISA felmérés során hosszas elemzések és gondos előkészületek után olyan változókat hoznak létre, melyek az országonként és teljes OECD viszonylatban is jó közelítéssel normális eloszlásra vannak transzformálva.

Ez a teljesítményeket mérő változók esetén azt jelenti, hogy a teljes OECD átlag a fókusz évében olyan normális eloszlású valószínűségi változóként kezelendő, melynek várható értéke 500, szórása 100.

Megjegyzés. Az egyéb indexeket, mint amilyen pl. az ESCS vagy a SES index is, úgy transzformálják, hogy az OECD országokra vonatkoztatva standard normális eloszlást kövessenek – így a különböző országok diákjai az OECD átlagával összehasonlíthatóak lesznek ezen indexek mentén. (SES: szociokultúrális háttér index).

Ez egyben azt is jelenti, hogy a vendég – partner – országok szintén az OECD szintjéhez vannak viszonyítva, de a standardizálásban ők nem vesznek részt, csak ugyanazokat a teszteket írják és a többi országgal megegyező számítási elvek alapján határozzák meg a pontjaikat.

Egy egyszerű kérdést szeretnénk megvizsgálni, ám a választ három különböző módon fogjuk kiszámítani. A számításokhoz minden esetben az OECD PISA felmérés során elfogadhatónak ítélt és alkalmazott SPSS programcsomagot fogjuk használni.

A kérdés tehát az, hogy Magyarországon a 4 különböző képzési típusba járó 15 éves diákok átlagos teljesítménye között van-e szignifikáns különbség. Ezt az alábbi három különböző módon szeretnénk megvizsgálni.

1. Először az SPSS beépített rutinjait fogjuk alkalmazni. Értelemszerűen a fenti kérdés eldöntésére egyszempontos ANOVA-elemzést fogunk bevetni. Az első esetben az OECD PISA által használt súlyozást alkalmazzuk.
2. Második esetben módosítani fogjuk az OECD PISA felmérésben használt súlyozást és szintén az SPSS beépített eljárását alkalmazzuk a kérdés eldöntésére.
3. Harmadik esetben a felmérésben alkalmazott szimulációs eljárást fogjuk bevetni, ragaszkodva a [9]-ban megjelölt útmutatóhoz.

Ahhoz, hogy a különböző módszerek közötti különbségeket értelmezni tudjuk, illetve egyáltalán végre tudjuk hajtani az elemzéseinket, szükségünk van arra, hogy az adatbázisok felépítését megismerjük, vagy legalábbis az elemzéshez szükséges paramétereket rögzítsük.

1. Minden diák esetén a fent már említett matematika teljesítmény változó(ka)t fogjuk figyelemmel kísérni. Minden diák esetén úgynevezett plauzibilis értékeket (PV) határoztak meg, minden diákra 5 darabot.

   Ezt úgy kell értelmezni, mintha a diák egyetlen teljesítménye helyett 5 darab, azzal azonos eloszlású, egymástól független véletlen valószínűségi változót tekintetnénk.

   Azaz:

   $$PV_{M_j} \sim V(\nu_0, \sigma_0), \quad j = 1, \ldots, 5,$$

   és ν_0 az adott diák becsült teljesítménye, míg σ_0 az adott diák becsült teljesítményének szórása.

Ezt talán úgy lehet legkönnyebben interpretálni, hogy a tesztek során több, különböző nehézségű kérdést kapnak a diákok (mindegyik kérdés 1-1 véletlen változónak tekinthető), a teljesítmény várható értéke és szórása ezen véletlen változók együttesének figyelembevételével alakul ki [9].

A diákok teljesítményét Rasch-modell segítségével állítják elő, amely lehetővé teszi a plauzibilis értékek meghatározását [8]. E modell segítségével minden diák esetén megmondható, hogy bizonyos nehézségű feladatokat milyen valószínűséggel oldhat meg – így a diák képessége egy valószínűségi változó segítségével modellezhető. A teljesítmény meghatározásáról [9], illetve [8]-ben találhatunk információkat.

2. Minden diákot egy adott súllyal is ellátnak annak megfelelően, hogy az ország mely szempontok szerint végezte el a rétegzett mintavételt - és a diák milyen rétegnek megfelelő részpopulációt képvisel. Ennek segítségével teszik a mintát reprezentatívvá az egész országra vonatkoztatva (felsúlyozással). Itt gondolhatunk arra, hogy nemre, településtípusra, iskola típusára, méretére vonatkozó információkra, vagy a szülők iskolai végzettségére (melyet előre nem tudhatunk) kell reprezentatívvá tennünk a mintát.

   Ez természetesen országonként eltérő, hiszen vannak országonként specifikus tényezők is, melyeket figyelembe kívánnak venni az adott ország felmérésében résztvevő kutatócsoportok.

   Ez a felsúlyozás azt jelenti, hogy bár csak 4-5000 diák írja meg Magyarországon a tesztet, mégis úgy kezeljük a súlyozás segítségével, mintha mind a nagyjából 100000 diák megírta volna a tesztet. (Természetesen ilyenkor csak a reprezentativitás van helyreállítva, a mintánkban rejlő információ csak ilyen szempontból módosul).

   Megjegyzés. A felsúlyozást a későbbiekben jól megválasztott súlyokkal próbálják ellensúlyozni. Ezért teszünk próbát arra, hogy a felsúlyozás helyett a mintán belül állítsuk helyre az arányokat anélkül, hogy az esetszámot a sokszorosára növelnénk.

3. Az előző pontban bemutatott súlyozást a második módszerhez úgy transzformáljuk, hogy az arányokat a mintán belül állítjuk helyre, így nem fogjuk a minta nagyságát mesterségesen a sokszorosára növelni – ezáltal nem csökkentjük mesterségesen a standard hibát.

Szerencsére ez a struktúra a kezdeti felmérésektől kötött, így nem okoz gondot ugyanazon programok alkalmazása a felmérések különböző időpontjaiban.

Megjegyzés. Az is világos, hogy az így felvett minta nem kezelhető egyszerű, véletlen mintaként. Így bármilyen nagy is a számunkra rendelkezésre álló adathalmaz, annak rétegzett tulajdonságait mindenképpen figyelembe kell vennünk az elemzés során.

3. A részpopulációk átlagainak összehasonlítása, hagyományos eljárás

Világos, hogy a fenti kérdés megválaszolása nem más, mint különböző csoportok esetén az átlagok meghatározása, majd azok összehasonlítása.

Ezt hagyományos ANOVA-eljárás keretei között lehet vizsgálni. Ehhez nincs másra szükség – hagyományos esetben, – mint a részminták átlagára és szórására, illetve azok elemszámára [12].

Hagyományos esetben első fázisban csak azt tudjuk megmondani, hogy a képzés típusának van-e valamilyen hatása a vizsgált változónk csoportonkénti várható értékére. Minket azonban az is érdekel, hogy ha van hatás, akkor az miben nyilvánul meg. Valójában – leginkább – erre a második kérdésre keressük a választ.

Ebben az esetben az ANOVA-elemzés egyik utóelemzésére lesz szükségünk – nevezetesen, valamely páronként is alkalmazható összehasonlításra. Ismert tény, hogy pl. páronkénti t-próbák alkalmazása nem megfelelő az elsőfajú hibák kontrolljának elvesztése miatt.

Ezért páronkénti összehasonlításra pl. a Tukey, vagy annak általánosításaként felfogható, robusztus Tukey–Kramer-eljárásra van szükségünk (ezen utóbbi a mintaelemszámok eltérő volta esetén is alkalmazható). Ezek minden további nélkül alkalmazhatóak a fenti esetekre attól függően – az elméleti szórások inhomogén volta esetén szabadságfok korrekciót végrehajtva. [2], [3]. Ezen módosított eljárás a Games–Howell-féle robusztus eljárás.

A vizsgált változóink normalitása részcsoportonként is biztosított, így a hagyományos eljárások kiválasztása során csak a szórások egyezésére kell tekintettel lennünk [12].

A valódi gond azonban az, hogy a mintavételi eljárásnak köszönhetően az egyedek egyáltalán nem tekinthetőek függetlennek. Gondoljunk csak arra, hogy az egy iskolában (egy osztályban) tanuló diákok a hasonló (vagy azonos) tantervvel, tanárokkal tanulnak, középiskolás korban persze eltérő érdeklődéssel – bár gyakran nagyon hasonló otthoni háttérrel.

Ráadásul azonos hatások érik őket iskolai szinten, pl. az iskola felszereltsége, iskolai közérzet, a többi diák viselkedése, tanárokkal való viszony stb.

Ahhoz, hogy ennek hatását a páronkénti összehasonlítások esetén vizsgálni tudjuk, meg kell vizsgálnunk az eljárás számítási lépéseit – itt részint a Tukey–Kramer, részint a szóráshomogenitásra robusztus Games–Howell-tesztet is vizsgálni fogjuk, mert a szórások homogenitása nem feltétlenül biztosítható. Részletezettebb leírást lásd pl. [14].

1. Jelölések: Legyenek

n_i	az i-edik minta elemszáma, $i = 1, \ldots, k$
$x_{i,j}$	az i-edik minta j-edik eleme
$\overline{x}_i$	az i-edik minta átlaga
s_i	az i-edik minta korrigált tapasztalati szórása

Megjegyzés. Ezen fenti jelölések a mi példánkon az alábbiak szerint értelmezhetőek:

n_i	az i-edik iskolatípusba járó diákok létszáma, $i = 1, \ldots, 4$
$x_{i,j}$	az i-edik iskolatípus j-edik diákja
$\overline{x}_i$	az i-edik iskolatípus átlaga
s_i	az i-edik iskolatípus korrigált tapasztalati szórása

2. Számítsuk ki az alábbi mennyiségeket:

$$N = \sum_{i=1}^{k} n_i,$$

$$T_i = \sum_{j=1}^{n_i} x_{i,j} \quad \Rightarrow \quad \overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j},$$

$$G = \sum_{i=1}^{k} T_i, \quad \Rightarrow \quad \overline{X} = \frac{G}{N},$$

$$Q_i = \sum_{j=1}^{n_i} x_{i,j}^2 - \frac{T_i^2}{n_i}.$$

3. Legyen most

$$Q_B = \sum_{i=1}^{k} Q_i,$$

$$Q_K = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{X}\right)^2.$$

4. Legyenek továbbá

$$f_K = k - 1, \qquad f_B = N - k,$$

rendre a Q_K és Q_B mennyiségekhez tartozó szabadságfokok.

5. Végső lépésként kiszámítjuk az

$$s_K^2 = \frac{Q_K}{f_K}, \qquad s_B^2 = \frac{Q_B}{f_B}$$

mennyiségeket.

Megjegyzés. Az s_K^2 és s_B^2 mennyiségeket külső és belső variancia néven szokás nevezni.

3.1. KÖVETKEZMÉNY. *Amennyiben a részcsoportok várható értéke megegyezik (és az előzetesen feltett feltételek is teljesülnek), úgy az*

$$F = \frac{s_K^2}{s_B^2}$$

statisztika f_K, f_B szabadságfokú F-eloszlást követ.

Ezzel még csak azt határoztuk meg, hogy van-e az iskola típusának valami hatása a diák teljesítményének szintjére. Azt azonban, hogy mely iskolatípusok teljesítenek esetleg jobban vagy rosszabbul, még nem tudjuk. Erre való pl. a Scheffé-teszt, vagy a már említett Tukey–Kramer-próba, melynek feltétele a változók normalitásán túl a részpopulációkban mért szórások egyezése [16], [22]. A szórások inhomogenitása esetén pedig a Games–Howell-teszt alkalmazandó.

A páros összehasonlítások metodikája a következő:

1. Legyenek $\overline{x}_p$ és $\overline{x}_r$ a két vizsgált részpopuláció átlaga, ahol $1 \leq p, r \leq k$.
2. Számítsuk ki (a fenti jelölések használatával) az alábbi statisztikát:

$$F_{p,q} = \frac{(\overline{x}_p - \overline{x}_r)^2}{(k-1)\, s_B^2 \left(\frac{1}{n_p} + \frac{1}{n_r}\right)}.$$

3. A normalitási és szóráshomogenitási feltétel mellett, a részpopulációk várható értékének egyezése esetén a fenti $F_{p,q}$ statisztika $f_K = k - 1$ és $f_B = N - k$ szabadságfokú F eloszlást követ.

Megjegyzés. [22] szerint az eljárások egyik fontos feltétele az egyszerű, véletlen mintavételezés. Ez azt jelenti, hogy csak akkor alkalmazhatjuk ezeket az eljárásokat, ha ezen szempont szerint kellően robusztusak. Azonban, ahogy [5]-ben is láthatjuk, az eljárás csak a szórások egyezésének sérülésére robusztus, a többi feltételt nekünk kell garantálni.

A Tukey–Kramer-módszer nem sokban különbözik a fent ismeretett Scheffé-féle eljárástól.

1. Legyen továbbra is $\overline{x}_p$ és $\overline{x}_r$ a két vizsgált részcsoport átlaga.
2. A fenti jelölések mellett számítsuk ki a

$$D_{p,r} = q_\alpha \sqrt{\frac{s_B^2}{2}\left(\frac{1}{n_p} + \frac{1}{n_r}\right)}$$

mennyiséget, ahol q_α az α szignifikancia-szinttől és k, illetve $N-k$ szabadságfoktól függő, Tukey-táblázatban megtalálható érték, lásd pl. [13].

3. Ha bármely p és r sorszámú minta átlagára teljesül, hogy

$$|\overline{x}_p - \overline{x}_r| \geq D_{p,r},$$

illetve ezzel ekvivalensen:

$$\frac{|\overline{x}_p - \overline{x}_r|}{\sqrt{\frac{s_B^2}{2}\left(\frac{1}{n_p} + \frac{1}{n_r}\right)}} \geq q_\alpha,$$

úgy azt mondhatjuk, hogy a $H_0 : \mu_p = \mu_r$ nullhipotézis elutasítható.

4. A Games–Howell-eljárást a következőképpen nyerjük: a Tukey–Kramer-eljárásban használt $D_{p,r}$ mennyiséget a szóráshomogenitási feltétel sérülése esetén módosítsuk az alábbi módon: q_α helyett q_α^f érték használandó, ahol f nem más, mint az $N - k$ szabadságfok módosítása az alábbi képlet szerint, lásd pl. [3]:

$$a = \frac{s_p^2}{n_p},$$

$$b = \frac{s_r^2}{n_r},$$

$$f = \frac{(a+b)^2}{\frac{a^2}{n_p - 1} + \frac{b^2}{n_r - 1}}.$$

Ez esetén kiszámítandó:

$$T_{p,r} = \frac{\overline{x}_p - \overline{x}_r}{\sqrt{\frac{a+b}{2}}}.$$

Akkor tekintjük a két várható értéket különbözőnek, ha

$$|T_{p,r}| \geq q_\alpha^f.$$

Megjegyzés. A szabadságfok kiszámításának analógiája található pl. a kétmintás t-próba és annak Welch-féle robusztus változatában is [12].

Megjegyzés. Az SPSS programcsomag többfajta szóráshomogenitásra nézve robusztus páronkénti összehasonlítást tartalmaz, melyek mindegyike a különbség standard hibájának egyfajta fentihez hasonló, szabadságfok korrekcióján alapul [21]. Ezek az eljárások nem mutattak érdemi különbségeket a hibabecslés konfidencia-intervallumán.

4. Nem hagyományos eljárás ismertetése

A nem hagyományos eljárás során olyan szimulációs technikát alkalmaztunk, mely egyfajta házasítása a jackknife és a bootstrap algoritmusnak. E két módszer általában jól ismert eljárások, róluk bővebben lásd pl. [6, 4, 7].

5. A jackknife eljárás

A jackknife módszer alkalmazásának feltétele mindösszesen az, hogy a mintánk független, azonos eloszlású valószínűségi változókból álljon, illetve a közös eloszlásukhoz véges szórásnégyzet tartozzon. (Ez a centrális határeloszlás tételben foglaltak teljesüléséhez szükséges).

Azt is feltételezzük, hogy a statisztikánk a megfigyeléseinkben (a valószínűségi változóinkban) szimmetrikus, tehát az argumentumok sorrendje nem befolyásolja a statisztika értékét.

Tegyük fel, hogy adott egy n elemű minta: $X_1, \ldots, X_n \sim F$.

Megjegyzés. Olyan $\widehat{\theta}$ statisztikákat vizsgálunk, melyek bármely n-nél kisebb elemszám esetén is értelmezettek (szükséges feltétel a kiszámíthatóság érdekében).

Bármely általános, θ paramétert becslő $\widehat{\theta}$ statisztika esetén a jackknife módszerből származó becslés az alábbi alakot ölti. Jelölje $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$-et.

Tekintsük a következő n darab pszeudó-statisztikát:

$$\widehat{\theta}_{(i)} = \widehat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Ekkor

$$\widehat{\theta}_{(\bullet)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{(i)},$$

a keresett statisztika becslése, míg a becslés hibájának jackknife becslése:

$$\widehat{\sigma}_J\left(\widehat{\theta}\right) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(\bullet)}\right)^2}.$$

6. A bootstrap eljárás

A bootstrap algoritmus alapját képező approximáció konvergenciájához elégséges feltétel a mintánkat alkotó valószínűségi változók véges szórásnégyzete (közös eloszlást tételezzünk fel itt is az X_i változókra nézve). Ez a feltétel a centrális határeloszlás tétel feltételeinek teljesülése miatt szükséges is.

(1) $\widehat{\theta}(x_1, \ldots, x_n)$ a statisztika értéke. (Egy adott $X_1 = x_1, \ldots, X_n = x_n$ realizáció mellett).
Ekkor $\sigma(F) = \sqrt{Var_F(X_1, \ldots, X_n)}$ a statisztika valódi hibája.
Ezt többnyire nehézkes zárt formában felírni.

(2) Miután az F eloszlást nem ismerjük, ezért $\widehat{F}$-pal, a tapasztalati eloszlásfüggvénnyel becsüljük. Ekkor persze $\widehat{\sigma}_B = \sigma\left(\widehat{F}\right)$ becsüli $\sigma(F)$-et. (Mint azt majd az alábbiakban látni fogjuk).

Itt csak approximációról van szó, hiszen ezt sem tudjuk zárt alakban felírni. Így tehát egy approximációs eljárást kell végrehajtanunk, mely a következő lépésekből áll.

(i) $\widehat{F}$ meghatározása.

(ii) F-ból független mintavétel segítségével $X_1^*, \ldots, X_n^*$ úgynevezett bootstrap minta létrehozása. Itt be kell tartanunk, hogy

$$\forall i: \; P\left(X_i^* = x_j\right) = \frac{1}{n}.$$

(Minden mintaelem ugyanolyan valószínűséggel veheti fel a realizációban szereplő különböző értékeket).

(iii) $\widehat{\theta}^* = \theta\left(X_1^*, \ldots, X_n^*\right)$ bootstrap másolatból származó statisztika kiszámítása.

(iv) az (ii) és (iii) lépések B számú ismétlése. Így előállítunk egy $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ független boostrap másolatból származó, a becslés értékére vonatkozó mintát.

(v) $\widehat{\sigma}_B$ approximáció kiszámítása az alábbi formula segítségével:

$$\widehat{\sigma}_B = \sqrt{\sum_{b=1}^{B} \frac{\left(\widehat{\theta}_b^* - \widehat{\theta}_\bullet^*\right)^2}{B-1}}$$

(tapasztalati szórás), ahol

$$\widehat{\theta}_\bullet^* = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{\theta}_b^*\right).$$

Megjegyzés. Ekkor, ha $B \to \infty$, úgy a $\widehat{\sigma}_B$ közelíti $\sigma(F)$-et. B optimális megválasztásáról nincsenek különösebb viták: általában elegendő 100 és 500 közötti bootstrap minta kiszámítása.

Megjegyzés. Az eljárásokból származó, adott szintű konfidencia-intervallum könnyen meghatározható a pszeudó-statisztikákból számított, becsült paraméterek sorbarendezéséből és megfelelő méretű trimmeléséből.

Más filozófia alapján folytatható addig az (ii)–(v) lépések egymásutánja, ameddig a lépésenként kiszámított és korrigált $\widehat{\theta}_\bullet^*$ valamilyen, előre meghatározott, minőséget előíró korlátnál kevesebbet változik 1 lépés alatt.

Fontos megjegyezni, hogy az approximáció konvergenciájához elégséges feltétel a véges szórásnégyzet (közös eloszlást tételezzünk fel itt is az X_i változókra nézve), mely a centrális határeloszlás tétel feltételeinek teljesülése miatt szükséges.

7. Fay eljárása és a BRR-eljárás

A jackknife eljárás szisztematikussága mellett a bootstrap eljárás véletlenszerűségét is szimuláló eljárás elméleti hátteréről [4]-ban vagy [6]-ben olvashatunk bővebben. Generált (tehát nem egy éles felmérésben végzett) adatokon végzett számításokról, azok eredményeiről [11]-ben tájékozódhatunk. Az eltérést az itt leírtaktól az adja, hogy a mi esetünkben két helyen is véletlenítés szerepel. Részint a szimulációs technika is magában hordoz egyfajta véletlenítést, részint pedig a vizsgált egyedekhez sem egyetlen mért érték tartozik, hanem 5, egymástól független, azonos paraméterekkel rendelkező, normális eloszlásból származó véletlen valószínűségi változó.

A felmérés adatbázisában alkalmazott és rögzített változók megalkotásáról részint [10] ad képet, részint pedig [23] ad tájékozódási pontokat.

Vázlatosan ismertetem a plauzibilis értékek által generált átlagok közötti különbségek standard hibáját számító algoritmust, [10] és [11] alapján.

A jól ismert jackknife módszer interpretálható úgy is, hogy az eredeti megfigyeléseket súlyozzuk: minden lépésben az egyik mintaelem súlyát 0-ra módosítjuk, a többi mintaelem súlyát változatlanul hagyjuk [15].

Fay eredeti eljárása egyik változtatásként azt mondja, hogy vegyünk egy mintaelemet $\frac{1}{2}$, míg egy bizonyos másik mintaelemet $\frac{3}{2}$ súllyal. Így annyi történik, hogy egy mintaelemet kevésbé, míg helyette egy másik mintaelemet jobban veszünk figyelembe az egyes pszeudo-statisztikákban [6].

Az igazi különbség azonban abban rejlik, hogy bizonyos egyedek helyett csak meghatározott más egyedek léphetnek be a pszeudo-statisztikákba. Ezen csoportokat a mintavétel során feltételezett összefüggések határozzák meg (az általunk vizsgált felmérésben ezek a szempontok azok, amelyek alapján reprezentatív mintát készítettek az adott országok).

Az alapmódszert egy olyan példán ismertetem, amikor minden csoportban 2-2 egyed szerepel. Az általánosság megszorítása nélkül tehetünk ilyen egyszerűsítést. Így a jackknife módszernek egy olyan módosítását/általánosítását nyerjük, mely arányosan rétegzett mintavétel esetén is használható. (Az OECD PISA felmérésben ezeket a rétegeket és arányokat a mérést koordináló kutatók adják meg minden országra az adott sajátosságokhoz alkalmazkodva a mérést szervező OECD központnak).

Az alább ismertetésre kerülő leírás esetén a legfontosabb szempont – mely a hagyományos módszerek alkalmazását kizárja – a mintába kerülő egyedek függetlenségének hiánya.

Legyen adott H darab strátum vagy réteg. (Pl. $H = 10$ iskola). Ekkor az egy rétegbe tartozó változók helyére csak a nekik megfelelő rétegből választhatunk „helyettesítő" változót. (Azaz, csak az adott iskolák diákjai helyettesíthetőek egymással).

Az egyszerűség kedvéért tegyük tehát fel, hogy minden rétegbe 2-2 egyed tartozik. Ekkor a súlyok $\widetilde{w}$ vektora: $\widetilde{w} = (w_{11}, w_{12}, \ldots, w_{H1}, w_{H2})$.

7.1. Példa. Ha a fenti 10 iskola mindegyikébe jár 20 fiú és 30 lány, akkor $\widetilde{w} = (20, 30, \ldots, 20, 30)$, vagy a nekik megfelelő fiú-lány létszám.

Megjegyzés. Ez még módosulhat, ha valamely oknál fogva nem veszünk valakit figyelembe (pl. a tanulási teljesítmény mérésénél a Sajátos Nevelési Igényű diákokat kihagyjuk). A végső súlyokat az egyes strátumokban ezek meghatározása után nyerjük [10].

Így jelölje a h-adik strátum i-edik esetének végső w súlyát $w_{Vhi} = f_{hi}(w)$.

8. A replikánsok súlyainak meghatározása

A BRR-eljárás ezen annyiban módosít, hogy ezt az eljárást többször, egymás után végrehajtja, mind az ismétlések miatt, mind pedig a plauzibilisek miatt [15].

Az ismétléses eljárás során a t-edik replikáns (*ismételt minta*) súlyait jelölje $w_{Vhi}^{(t)} = f_{hi}\left(w^{(t)}\right)$, ahol

$$w^{(t)} = ((1 + d_{t1})w_{11}, (1 - d_{t1})w_{12}, \ldots, (1 + d_{tH})w_{H1}, (1 - d_{tH})w_{H2}),$$

és itt d_{th} egy csupa 1-ből és -1-ből álló D mátrix t-edik sorának és h-adik oszlopának eleme, ahol a mátrix sorai és oszlopai is ortogonálisak egymásra.

Megjegyzés. Az itt említésre kerülő, úgynevezett Hadamard-mátrix meghatározására [15], vagy egy program, [23] ad bővebb instrukciót. Az ortogonalitás biztosítja a minták függetlenségét, a ± 1 szorzók pedig megadják, hogy mely egyedek mely egyedek helyett kerülhetnek a replikánsokba.

Fay módosított eljárásának [6] súlyai egy korrigáló konstans szorzóban különböznek ettől, nevezetesen:

$$w^{(t)} = ((1 + d_{t1}(1 - k))\widetilde{w_{11}}, (1 - d_{t1}(1 - k))\widetilde{w_{12}},$$
$$\ldots, (1 + d_{tH}(1 - k))\widetilde{w_{H1}}, (1 - d_{tH}(1 - k))\widetilde{w_{H2}}).$$

Itt $k \in [0, 1)$. Látható, hogy a BRR-eljárás a módosított Fay-eljárás speciális esete $k = 0$ esetén.

Megjegyzés. Így egyes egyedek nem kikerülnek vagy bekerülnek egy-egy replikánsba, hanem kisebb vagy nagyobb súllyal vesznek részt a paraméterek becslésében.

Amennyiben a $\theta(w_V, x)$ statisztikára vagyunk kíváncsiak, akkor a varianciák becslése:

$$(\sigma_{BRR})^2 = \frac{1}{T}\sum_{t=1}^{T}\left[\theta\left(w_V^{(t)}, x\right) - \theta\left(w_V, x\right)\right]^2,$$

$$(\sigma_{FAY})^2 = \frac{1}{T}\frac{1}{(1-k)^2}\sum_{t=1}^{T}\left[\theta\left(w_V^{(t)}, x\right) - \theta\left(w_V, x\right)\right]^2,$$

ahol T a replikánsok (ismétlések) száma.

Megjegyzés. Hagyományos esetben a várható érték becslésének standard hibája csak a becsülni kívánt eloszlás varianciájától és az elemszámtól függ (amennyiben egyszerű véletlen mintavételezést alkalmazunk).

Az esetleges összefüggések kiküszöbölésére alkalmazható a bootstrap és a jackknife eljárás, melyek során a standard hiba lényegében a replikánsokból számított paraméterek varianciájától és a replikánsok számától függ.

Amennyiben rétegzett mintavételnél szimulációs technikát alkalmazunk, úgy a becsülni kívánt paraméterek becslésén lévő hiba most már a rétegek varianciájától és elemszámától, valamint a rétegzett mintákon külön-külön alkalmazott replikálásokban számított paraméterek varianciájától függ.

Ha minden egyed esetén még plauzibilis értékeket is számítunk, úgy a rétegzés mellett még az egyes egyedeken mért variancia is hozzáadódik a paraméter becslésének hibájához, így a teljes populációra vett becslés végső hibája három részből tevődik össze: egyik oldalról a rétegek egymáshoz viszonyított különbségeiből, másik oldalról a replikálánsok során számított paraméterek varianciájából, harmadrészt pedig az egyedeken lévő plauzibilisek varianciájának mértékéből.

9. A hagyományos eljárás alkalmazása

A matematika teljesítmény átlagának összehasonlítása két eltérő súlyozásra a különböző képzési típusok esetén

A hagyományos eljárás mellett számított értékek, kétfajta súlyozás mellett az alábbiak szerint alakultak:

Teljes populációra való felsúlyozás esetén:

Képzési típus	Esetszám	Átlag	Átlag hibája	Szórás	Szórás hibája
1 - Általános iskola	6510	391,911	0,93	74,93	NA
2 - Gimnázium	21027	551,907	0,37	72,86	NA
3 - Szakközépiskola	41668	491,541	0,33	66,59	NA
4 - Szakiskola	37837	405,981	0,43	62,91	NA

A mintán belül helyreállított arányokat alkalmazó súlyozás esetén:

Képzési típus	Esetszám	Átlag	Átlag hibája	Szórás	Szórás hibája
1 - Általános iskola	394	391,911	3,75	75,06	NA
2 - Gimnázium	1618	551,907	1,80	72,88	NA
3 - Szakközépiskola	1852	491,541	1,54	66,60	NA
4 - Szakiskola	901	405,981	2,09	62,95	NA

Megjegyzés. Megállapítható, hogy az átlagon lévő standard hiba - természetesen a teljes elemszám megváltozása miatt - drasztikusan megnövekedett a második súlyozás alkalmazásakor.

Megjegyzés. Az SPSS 15 nem tartalmaz olyan rutint, mely pl. a szórás konfidencia-intervallumát meghatározná.

Az átlagét az ismert $\frac{\sigma}{\sqrt{n}}$ képlet segítségével számított standard hibával meghatározhatjuk, míg a leíró statisztikák menüpontban opcionálisan a ferdeség és csúcsosság paraméterek mellett azok standard hibáját is kiszámítja.

Az is megfigyelhető, hogy a második súlyozás esetén a szórás néhány tizeddel eltérő értéket vett fel (ez teljesen elfogadható, ha figyelembe vesszük a korrigált tapasztalati szórás számítását).

Ezek után vizsgáljuk meg, hogy a hagyományos eljárás, illetve annak szórás-homogenitást nem igénylő, robusztus, páronkénti összehasonlítást elvégző számításai mit eredményeztek a kétféle súlyozás esetén. Azért a robusztus eljárást alkalmazzuk, mert a Levene-próba alapján sérül a szórások egyenlőségét igénylő szórás-homogenitási feltételünk.

Először ismételten a felsúlyozott, majd az arányaiban helyreállított súlyozás eredményeit ismertetem.

Eredmények felsúlyozás esetén:

Képzési típus	Átlagok különbsége	Különbség hibája	Konf. iv. alja	Konf. iv. teteje
1 - 2	−159,996	1,00	**−162,56**	**−157,42**
1 - 3	−99,63	0,98	**−102,16**	**−97,1**
1 - 4	−14,071	1,03	**−16,7**	**−11,44**
2 - 3	60,366	0,50	**59,09**	**61,64**
2 - 4	145,926	0,57	**144,45**	**147,4**
3 - 4	85,56	0,54	**84,16**	**86,95**

Eredmények arányosított, második számú súlyozás esetén:

Képzési típus	Átlagok különbsége	Különbség hibája	Konf. iv. alja	Konf. iv. teteje
1 - 2	−159,996	4,75	**−172,26**	**−147,73**
1 - 3	−99,63	4,67	**−111,69**	**−87,57**
1 - 4	−14,071	4,87	**−26,62**	**−1,52**
2 - 3	60,366	2,35	**54,31**	**66,42**
2 - 4	145,926	2,72	**138,94**	**152,91**
3 - 4	85,56	2,57	**78,94**	**92,18**

Jól látható, hogy mindkét esetben, 95%-os szignifikancia-szint mellett a különbségek mindenhol eltérnek 0-tól, azaz a vizsgált várható értékek egyenlősége elvethető. Még a leginkább közel lévő, általános iskola és szakiskola esetén is szignifikáns különbség adódott.

10. A nem hagyományos eljárás alkalmazása

A matematika teljesítmény átlagának összehasonlítása a különböző képzési típusok esetén

Ebben az esetben is a 4 különböző képzési típusban vizsgáltuk meg a matematika teljesítmény átlagát és szórását. Az átlag esetén képes az SPSS hibát számítani – ezt fel is tüntetjük – azonban egy lényeges különbség rögtön látható lesz. Az előzőekkel ellentétben a teljesítmény szórása esetén az SPSS nem képes standard hibát meghatározni (vagy akár konfidencia-intervallumot). A szimulációs eljárások ebben (is) segítségünkre lehetnek, hiszen a szimulációs technikákkal az eloszlás bármely paraméterének standard hibája becsülhető, illetve bármely paraméterre konfidencia-intervallum illeszthető.

Az átlagok esetén itt is páronkénti összehasonlítást alkalmazó ANOVA-elemzést alkalmaztunk. Így megállapíthatjuk, hogy a különböző képzési típusok esetén valóban vannak-e teljesítménybeli eltérések.

A Fay-féle módosított BRR-szimulációs eljárással számított értékek az alábbiak lettek (a súlyozás készítésekor figyelembe veszik tehát azt, hogy az egy iskolában tanuló diákok nem tekinthetőek egymástól függetlennek):

Képzési típus	Esetszám	Átlag	Átlag hibája	Szórás	Szórás hibája
1 - Általános iskola	6510	391,911	7,16	82,67	4,54
2 - Gimnázium	21027	551,907	5,00	77,19	1,91
3 - Szakközépiskola	41668	491,541	4,61	71,18	1,95
4 - Szakiskola	37837	405,981	6,41	68,50	2,2

Megjegyzés. A csoportonkénti szórások nem egyeznek meg azzal, amit a hagyományos eljárásokkal, a plauzibilis értékek helyett azok átlagával számítottunk. Ez abból fakad, hogy a diákonkénti átlagot (elvárt teljesítményt) még tudtuk garantálni, azonban a plauzibilis értékek átlagának alkalmazásával az egy diákon lévő bizonytalanságot megszüntettük – így ezt a varianciából is eltüntettük.

A csoportok közötti különbségek és azok standard hibája, illetve az ezekből számított 95%-os konfidencia-intervallumok az alábbi táblázatban láthatóak. A kódokat az előző táblázatokban már ismertettük.

Képzési típus	Átlagok különbsége	Különbség hibája	Konf. iv. alja	Konf. iv. teteje
1 - 2	−159,996	8,39	**−176,44**	**−143,55**
1 - 3	−99,63	8,45	**−116,2**	**−83,06**
1 - 4	−14,071	9,58	**−32,86**	**4,71**
2 - 3	60,366	6,84	**46,95**	**73,78**
2 - 4	145,926	8,17	**129,91**	**161,94**
3 - 4	85,56	7,8	**70,28**	**100,84**

Megfigyelhető a különbség a hagyományos eljárás és a nem hagyományos, szimulációs eljárás eredményei között.

Míg a hagyományos esetben, szóráshomogenitást nem igénylő robusztusabb változata esetén is, mindkét súlyozásra azt az eredményt kaptuk, hogy a különböző iskolatípusok esetén a matematika teljesítmény várható értéke szignifikánsan eltér, addig a szimulációs technika azt mutatja, hogy bár vannak szignifikáns eltérések, de kivétel mégis akad.

A 15 éves diákok közül azok, akik szakiskolákba vagy általános iskolákba járnak, a szimulációs módszer alapján nem különböztek szignifikáns módon az átlagos teljesítményüket tekintve. Megállapítható, hogy a gimnazisták messze a többi iskolatípus felett teljesítettek, míg a szakközépiskolások a két teljesítmény között helyezkednek el, valamivel az OECD országok átlaga alatt.

11. Összegés: a különböző eljárások eredényeinek összehasonlítása

Az általános formulában használt két eljárás (az egyik az eredeti súlyozással, felfújt minta esetén, a másik az eredeti súlyozás mértékét megtartva, eredeti mintanagyságon belül dolgozva) alul becsüli azt a hibát, amit szimulációs eljárás segítségével nyerünk.

Ez olyan szempontból mindenképpen eltérő eredményre vezet, hogy a kisebb standard hiba következtében a különbségek hamarabb válnak szignifikánssá.

A szimulációs eljárás [11] alapján elméletileg bizonyítottan jobb becslést nyújt számunkra a valós standard hibát illetően rétegzett mintavételezés esetén, hiszen figyelembe veszi azt, hogy a mintánk korántsem rendelkezik azzal a függetlenséggel, mely a megszokott, $\frac{\sigma}{\sqrt{n}}$ standard hiba képletének alkalmazásához szükséges.

Bár e módosított eljárások némiképpen számításigényesebbek, még ekkora minták esetén sem okoznak érdemi számításidő növekedést. Az alkalmazásukból származó esetlegesen pontosabb információ mindenképpen megéri a ráfordított időt. Azonban csak valószínűsíteni tudjuk azt, hogy a szimulált eredmények megbízhatóbbak erre a felmérésre, mint a hagyományos eljárások számításai.

Mindenképpen meg kell azonban fogalmazni legalább egy kritikát is ezzel a módszerrel kapcsolatban: az alkalmazásban található BRR-technika hivatott kiküszöbölni a mintába kerülő egyedek közötti összefüggéseket, melyek az egy strátumba való tartozás miatt fellépnek. Nem világos, hogy az adatbázisban található BRR-technika hogyan tudja kiküszöbölni azt a problémát, hogy az egy strátumba kerülő egyedek közötti összefüggések változónként eltérőek – az eljárásban azonban, változótól függetlenül mindig ugyanazokat replikáns variációkat alkalmazzuk.

Egyszerűsítési okokból nem a bootstrap eljárást alkalmazzák ebben az elemzésben, az azonban nem látszik tisztázottnak, hogy milyen módon lehet visszanyerni ebben a módszerben a bootstrap algoritmus empirikus eloszlást alkalmazó erősségét, mely strátumonként, minden változóra más és más. Kisebb elemszám

mellett a bootstrap algoritmus alkalmazhatatlan lenne, azonban az OECD PISA adatbázis kellő elemszámmal rendelkezik ahhoz, hogy ez a probléma ne merüljön fel. [1]-ben a jackknife algoritmus különböző variánsaival összehasonlítják, azonban a jackknife algoritmus semmilyen módon nem veszi figyelembe a vizsgált változók empirikus eloszlását.

Hivatkozások

[1] Brick, J. M., Morganstein, D. and Valliant, R.: *Analysis of Complex Sample Data Using Replication*, WESTAT, (2000).

[2] Dunnett, Ch. W.: *Pairwise Multiple Comparison in the Homogeneous Variance, Unequal Sample Size Case*, Journal of the American Statistical Association **75**, (1980) 789–795.

[3] Dunnett, Ch. W.: *Pairwise Multiple Comparison in the Unequal Variance Case*, Journal if the American Statistical Association **75**, (1980) 796–800.

[4] Efron, B. and Gong, G. (1983): *A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation*, The American Statistican **37**, 36–48.

[5] Huzsvai László: *Biometriai módszerek az SPSS-ben, SPSS alkalmazások*, Debreceni Egyetem, Mezőgazdaságtudományi Kar, Debrecen, (2004) 53–36.

[6] Judkins, D. R.: *Fay's method for Variance Estimation*, Journal of Official Statistics Vol. **6** No. **3**, (1990) 223–239.

[7] Kárász Judit (2004): *A bootstrap algoritmus és a jackknife módszer*, Szakdolgozat, ELTE

[8] Molnár Gyöngyvér (2005): *Az objektív mérés lehetősége: a Rasch-modell*, Iskolakultúra, 2005/3, 71–80.

[9] *PISA2006 Technical Report*, OECD, (2009).

[10] *PISA2006 Technical Report*, OECD, (2009) 188–192.

[11] Slud, E. V. and Thibaudeau, Y.: *BRR versus Inclusion-Probability Formulas for Variances of Nonresponse Adjusted Survey Estimates*, AMSTAT, Section on Survey Research Methods, JSM (2008) 2057–2064.

[12] Vargha András: *Matematikai Statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal (2. kiadás)*, Pólya Kiadó, Budapest, (2007).

[13] Vargha András: *Statisztikai táblázatok*, Tankönyvkiadó, (1983), XXXIII. táblázat.

[14] Vargha András: *Pszichológiai statisztika gyakorlat II.*, Tankönyvkiadó, (1981) 112–124.

[15] Wolter, K. M.: *Introduction to variance estimation*, Springer, New York, (1985).

[16] http://people.richland.edu/james/lecture/m170/ch13-dif.html

[17] www.pisa.oecd.org

[18] http://pisa2000.acer.edu.au

[19] http://pisa2003.acer.edu.au

[20] http://pisa2006.acer.edu.au

[21] http://www.psychology.nottingham.ac.uk/staff/pal/stats/C82MST/contrasts.pdf

[22] http://xenia.sote.hu/hu/biosci/docs/biometr/lecture/anova1.html

[23] http://www.westat.com

(Beérkezett: 2010. február 1.)

TAKÁCS SZABOLCS
Károli Gáspár Református Egyetem, Bölcsésztudományi Kar
Pszichológiai Intézet, Általános lélektani és módszertani tanszék
1037, Budapest, Bécsi út 324., 5. épület, fszt.
E-mail: tretarkhon@gmail.com

A NON-TRADITIONAL STATISTICAL METHOD IN THE OECD PISA DATABASE CASE-STUDY

SZABOLCS TAKÁCS

When we haven't got a simple random sample the original simulation technics, like the bootstrap and the jackknife method are usable only with changes . In the following article we show a case when we compare the results of the traditional method and the result of a simulation algorithm.

One of the jackknife's modification [6] is used for some hypothesis in the OECD PISA survey. The theorems of the applied method are introduced in [6], [11]. Now we would like to show these algorithms in a large sample survey and we show the difference between the empirical results of the traditional and the simulation method.

Because we don't use generated database we don't know the truth. Without this information we would like to understand the different results of the algorithms.

We put some questions and critics to these algorithms which are relevant and we can't find reassuring answers in the connected bibliography of the survey.