MAGYAR PEDAGÓGIA
104. évf. 2. szám 183–201. (2004)

EMPIRIKUS PEDAGÓGIAI VIZSGÁLATOK OPTIMÁLIS MINTANAGYSÁGÁNAK MEGHATÁROZÁSA

Csíkos Csaba
MTA-SZTE, Képességkutató Csoport, Neveléstudományi Tanszék

A pedagógiai értékelés kvantitatív módszereinek elterjedésével mind gyakoribbá válnak a statisztikai mintavételt igénylő vizsgálatok. Az adatok gyűjtésének nagyon sokféle célja lehet: kutatás, fejlesztő kísérlet, tanulói teljesítmények nyomon követése stb. A tanulmány az empirikus pedagógiai vizsgálatok tervezésének kérdéseit elemzi abból a szempontból, mekkora létszámú tanulócsoport adatainak felhasználásával érdemes egy-egy adatgyűjtést lebonyolítani. A kizárólagosan leíró statisztikai jellegű pedagógiai felmérések mintanagyságának kérdését a tanulmány első részében érintjük. A statisztikai hipotézisvizsgálatokat felhasználó pedagógiai kísérletek esetén nem érvényes a fenti alapelv. Tézisünk, amit ebben a tanulmányban igazolni szeretnénk, hogy a fejlesztő kísérletek optimális mintanagyságának meghatározása alapvetően statisztikai alapokon nyugvó feladat.

Írásunk két nagyobb egységre tagolódik. Az első részben a nagymintás pedagógiai felmérések mintanagyságának meghatározása áll a középpontban. Azt igyekszünk megmutatni, hogy a minta reprezentativitásának szempontjai és a részminták leíró statisztikai mutatóinak felhasználása megtöbbszöröződheti a szükséges minta nagyságát. Ebben a részben a konfidencia-intervallum fogalma kapcsán a matematikai statisztika eszköztára is felvonul, de a statisztikai hipotézisvizsgálatok a második nagyobb részben kapnak majd főszerepet.

Mekkora legyen a minta a nagymintás vizsgálatokban?

Amennyiben egy pedagógiai jelenség vizsgálatakor a jelenség dokumentálása, leírása a cél, úgy nagymintás felmérést célszerű végezni. Különösen igaz ez a tétel, ha egy korábban még nem vizsgált személyiség-összetevő méréséről van szó. Ebben az esetben a pszichometria nagymintás mérésekre vonatkozó alapelvei szerint egyszerre történik az adott populáció minél jobb megismerése és a mérőeszköz fejlesztése, standardizálása. Az utóbbi évtizedekben számos példát láttunk arra Magyarországon is, hogy több ezres, sőt, több tízezres mintán történtek felmérések. (Egy közlemúltban lezajlott nagymintás tudományos felmérésről lásd *Józsa*, 2004). Az országos és nemzetközi rendszerszintű felmérésekhez (pl. a Monitor-vizsgálatok, a PISA-mérések, lásd OECD, 2005), a „szegedi

műhely" képességtesztjeinek beméréséhez nagy minták kiválasztására volt szükség. A minták kiválasztása során három fő tényezőt tartottak a kutatók szem előtt. A következőkben e három jelenséget mutatjuk be.

A reprezentativitási szempontok száma és a mintanagyság

Országos reprezentatív minták kiválasztása esetén a reprezentativitás szempontjainak száma szorzótényezőként jelenik meg a mintanagyság kiválasztásában. Ez azt jelenti, hogy ha van egy rögzített létszám, amelynek vizsgálatát finanszírozni lehet egy kutatásban, akkor az a létszám a reprezentativitás szempontjainak megfelelően osztódik kisebb, még önállóan vizsgálható egységekre. Így például bármilyen impozáns lehet egy tízezer fős felmérés, ha abban például az alapítványi iskolában tanuló, nem budapesti lányok eredményeinek átlagát szeretnénk meghatározni, akkor egy meglehetősen pici, esetleg 10 fő alatti részmintához juthatunk. Mélyebb statisztikai összefüggésekre hivatkozás nélkül is egyetérthetünk abban, hogy tíz fő alatti mintán az átlagolás, az egyik legalapvetőbb statisztikai eljárás, meglehetősen pontatlan eljárás.

Visszafelé gondolkodva: ha eldöntjük, hogy az önállóan vizsgálandó részminták létszáma körülbelül száz legyen, akkor három reprezentativitási szempontot, és az egyes szempontokon belül 3–5 lehetséges értéket feltételezve egy több ezer fős minta állhat elő. A reprezentativitás szempontjainak száma, és minden egyes reprezentativitási szempontón belül a változó értékeinek száma egyaránt szorzótényezőként jelentkezik. A jelenség szemléltetéséhez irányszámokat mutatunk meg egy elképzelt pedagógiai vizsgálathoz. Az elméleti arányszámok kiszámításához a „Jelentés a magyar közoktatásról 2003" könyv (*Halász* és *Lannert*, 2003) adatait használjuk föl. (Nem célunk a mintakiválasztás technikáinak áttekintése, hiszen az magyar nyelven több helyen is hozzáférhető, pl. *Babbie*, 2000).

Egy elképzelt kutatásban a magyarországi általános iskolai nappali tagozatos tanulók alapsokaságából szeretnénk mintát kiválasztani egy empirikus vizsgálathoz. Amennyiben nincs kiemelt szempontunk a reprezentativitáshoz, akkor a következő eljárás lenne célravezető: Az alapsokaság összes tanulójának névsorát egy adatbázisban szerepeltetjük, és a random számok táblázata segítségével, vagy a szisztematikus mintavétel technikájával kiválasztunk egy akkora mintát, amekkorát a kutatás finanszírozása elbír. Az egyszerű véletlen mintavétel technikáját alkalmaztuk így, amely a populáció minden tagjának egyforma esélyt biztosított a mintába kerüléshez, ebből adódóan a minta reprezentatív. Nagy valószínűséggel megfelelő számban szerepelnek az így kiválasztott mintában az ország különböző régiójából származó tanulók, akik megfelelő számban reprezentálják például az egyes iskolafenntartókat és a két nemet is.

Ha szeretnénk kiemelni, hogy a minta milyen szempontokból legyen reprezentatív, akkor a kutatás jellegétől függően például a régiók szerinti megoszlás, az iskolafenntartó típusa és a tanuló neme választható mintakiválasztási szempontként. Alapszabály, hogy egymástól független szempontokra van szükség. A függetlenség igazolása bonyolult procedúra, mert ebben az esetben azt kell bizonyítani, hogy minden régióban ugyanolyan a nemek és az egyes iskolafenntartók megoszlása is, valamint, hogy a fenntartó típusa és a nemek szerinti megoszlás is függetlenek. Egyelőre tételezzük fel, hogy van három,

egymástól független reprezentativitási szempontunk: (1) a régió, ahol a tanuló tanul, (2) az iskolafenntartó típusa, (3) a tanuló neme. A három szempontot tetszőlegesen egymásra vetítve minden egyes részmintának reprezentatívnak kell lennie a megfelelő részpopulációban ahhoz, hogy például az észak-magyarországi egyházi iskolákban tanuló lányok tudásszintjéről vagy a dél-dunántúli, megyei önkormányzati fenntartású iskolákban tanuló fiúk tantárgyi attitűdjeiről tegyünk megállapításokat.

A következő lépés az adott szempontok szerint lehetséges értékek populációbeli megoszlásának megállapítása. Ehhez nyújtanak segítséget különböző statisztikai kiadványok; példánkban az említett „Jelentés a magyar közoktatásról 2003" kötet táblázatait használjuk. A nappali tagozatos általános iskolai tanulók területi alapadatainak 4.28. számú táblázata szerint a 7 régió létszámadatai a következők (2001/2002-es tanévben): Közép-Magyarország 230 918, Közép-Dunántúl 108 109, Nyugat-Dunántúl 89 258, Dél-Dunántúl 92 457, Észak-Magyarország 128 799, Észak-Alföld 165 045, Dél-Alföld 129 658. A mintán belüli arányok kiszámításához az adatokat %-os értékre fogjuk átszámolni.

A „Jelentés a magyar közoktatásról 2003" könyv 4.27. számú táblázata az iskolafenntartók megoszlását %-ban kifejezve is mutatja: Települési önkormányzat 86,2%, megyei önkormányzat 4,9%, központi költségvetés 4,1%, egyház 3,8%, egyéb 1%. A kategóriák elaprózása nem célravezető, ezért dönthet úgy a kutató, hogy az iskolafenntartó szerint három kategóriát vesz figyelembe: települési önkormányzat 86,2%, megyei és központi költségvetéshez tartozó 9%, egyházi és egyéb 4,8%. Az eddigi két reprezentativitási szempont szerinti arányokat az 1. táblázat összegzi.

1. táblázat. Tanulói létszámarányok Magyarország régióiban az iskolafenntartó típusa szerint (%)

Fenntartó	*Közép-Magyar-ország*	*Közép-Dunántúl*	*Nyugat-Dunántúl*	*Dél-Dunántúl*	*Észak-Magyarország*	*Észak-Alföld*	*Dél-Alföld*
Települési	21,08	9,87	8,15	8,44	11,76	15,07	11,84
Megyei és központi	2,20	1,03	0,85	0,88	1,23	1,57	1,24
Egyházi és egyéb	1,17	0,55	0,45	0,47	0,65	0,84	0,66

A táblázatbeli %-os értékek megfelezésével kapjuk meg a nemek arányát is figyelembe vevő értékeket. A 2. táblázatban az 1. táblázatban megadott arányszámok felhasználásával azokat az egészre kerekített értékeket tüntetjük föl, amelyek 1000 fős minta esetén az egyes részminták celláira adódnak.

2. táblázat. Részminta-elemszámok egy 1000 fős minta esetén, a régió és az iskolafenntartó szempontja szerinti bontásban

Fenntartó	*Közép-Magyarország*	*Közép-Dunántúl*	*Nyugat-Dunántúl*	*Dél-Dunántúl*	*Észak-Magyarország*	*Észak-Alföld*	*Dél-Alföld*
Települési	211	99	82	84	118	150	118
Megyei és központi	22	10	9	9	12	16	12
Egyházi és egyéb	12	6	5	5	7	8	7

A kerekítéseknek köszönhetően végül 1002 fős minta adódott. Sok részmintába azonban így olyan kevesen kerülnének, hogy még az egyszerű átlagszámítás is csak bizonytalan becslést tudna adni a részpopulációk jellemző középértékére. Arra következtethetünk tehát, hogy ha a regionalitás és az iskolafenntartó szerepe is fontos szempont a reprezentativitáshoz, akkor az 1000 fő körüli minta nem teszi lehetővé a részminták releváns vizsgálatát.

Megfordíthatjuk a mintakiválasztás létszámadatainak keresését, és azt mondhatjuk meg, hogy legalább hány tanuló tartozzék az egyes kategóriákba. Legyen ez az érték olyan, hogy a nyugat-dunántúli régióban, egyházi vagy egyéb fenntartású iskolában tanulók közül is legalább 40-en szerepeljenek a vizsgálatban. Ebben az esetben 20 fiú és 20 lány tanulóra lehetne kettéosztani még ezt a legkisebb részmintát. A 3. táblázat megmutatja, hogy mekkora részmintákra lenne szükség, ha ezt a létszám-megállapító stratégiát követnénk.

3. táblázat. Részminta-elemszámok egy lehetséges nagy minta esetén, a régió és az iskolafenntartó szempontja szerinti bontásban

Fenntartó	*Közép-Magyarország*	*Közép-Dunántúl*	*Nyugat-Dunántúl*	*Dél-Dunántúl*	*Észak-Magyarország*	*Észak-Alföld*	*Dél-Alföld*
Települési	1858	870	718	744	1037	1328	1044
Megyei és központi	194	91	75	78	108	139	109
Egyházi és egyéb	103	48	40	41	58	74	58

A táblázatban szereplő, egészre kerekített értékek összegeként 8815-öt kaptunk. Ha ügyelünk arra, hogy az egyenlő fiú-lány arány minden cellában megvalósuljon, akkor a legközelebbi páros számokat szerepeltetve nagyságrendileg már nem módosul ez a szám. A mintavétel szabályaiból még az következik, hogy minden egyes részpopuláció teljes névsorát elő kellene állítanunk, és 7x3x2=42 esetben kell az egyszerű vagy a szisztema-

tikus véletlen mintavételt alkalmaznunk, hogy előálljon egy három szempontból reprezentatív országos minta. Mivel a pedagógiai felmérésekben többnyire az iskolai osztályok a mintavétel egységei, további korrekcióra lehet szükség a mintakiválasztásban, amely figyelembe veszi a régióktól kevésbé, az iskolafenntartóktól már sokkal inkább függő osztálylétszám-nagyságokat.

Az előzőekben részletesen végigvezetett példával az volt a célunk, hogy szemléltessük a mintakiválasztás szempontjainak száma és a mintanagyság közötti kapcsolatot. Ehhez mindössze a reprezentatív minták kiválasztásának szabályait követve azt vettük figyelembe, hogy minden vizsgált részmintában legalább 20 fő szerepeljen, akik reprezentálják a részminta tulajdonságainak megfelelő részpopulációt.

Összefüggés a mérés pontossága és a mintanagyság között

Ha eltekintünk a részminták vizsgálatának lehetőségétől, és homogénnek tekintjük a populációt, akkor a mért adatokból származó becslések pontosságára utalhatunk. Nagyobb minta esetén sokkal kisebb az az intervallum (a konfidencia-intervallum), amely a felmérésben kapott adatok átlaga köré olyan módon rajzolható, hogy 95%-os valószínűséggel beletartozzék a vizsgált, elvileg végtelen nagy populáció elméleti átlaga. Egy kis táblázatban bemutatom adott tapasztalati átlagot és szórást feltételezve, hogy a minta elemszámának növelésével hogyan szűkül az imént definiált konfidencia-intervallum (4. táblázat). Tegyük fel, hogy az alapsokaság elméleti szórása 1, a különböző nagyságú mintákon pedig rendre 3,0 adódott átlagként.

4. táblázat. A konfidencia-intervallum nagysága a mintanagyság függvényében 3,0-es mintaátlag és 1,0 elméleti szórás esetén, 95%-os szignifikancia-szint mellett

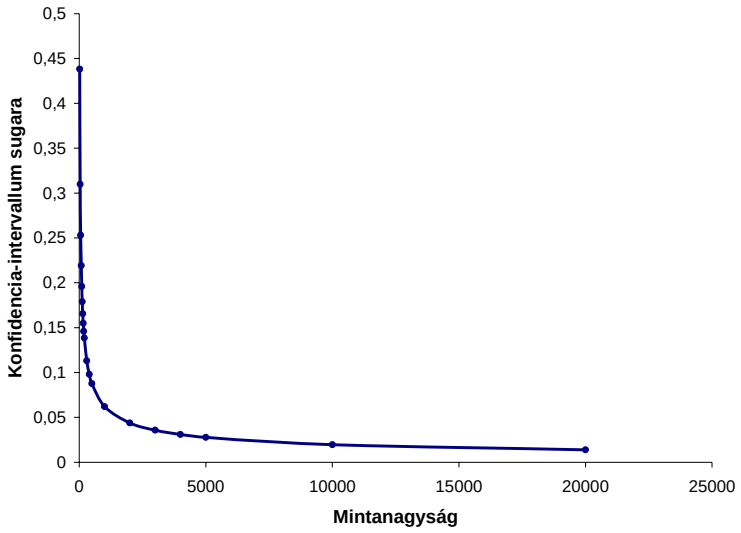
	Mintanagyság (N)		
	20	200	2000
95%-os szignifikancia-szinthez tartozó konfidencia-intervallum	[2,56 ; 3,44]	[2,86 ; 3,14]	[2,96 ; 3,04]

Megjegyzés:
A konfidencia-intervallum sugara, ismert elméleti szórás esetén, 95%-os bizonyossági szinten: $\frac{1{,}96 \cdot \sigma}{\sqrt{n}}$ ahol σ a populáció elméleti szórása, n pedig a minta elemszáma. A gyakorlatban szinte sosem ismerjük σ értékét, ezért helyette a minta korrigált tapasztalati szórását használjuk, és így az 1,96-os szorzó módosul [a minta elemszámától függően, a t-eloszlás táblázatából választva valamivel nagyobb lesz [(ld. részletesen *Vargha*, 2000)]. Abból következően, hogy a populáció elméleti szórása helyett „beérjük" a minta szórásával, csupán tágasabb intervallum-becslést tudunk adni a populáció elméleti átlagára.

Ha a 3,0-es átlag (például egy adott tantárgy osztályzatainak átlaga) egy 2000 fős mintán adódott ennyinek, akkor 95%-os bizonyossággal állítható, hogy a teljes, elvileg végtelen nagyságú alapsokaságban a valódi, elméleti átlag 2,96 és 3,04 közé esik. Nem kizárt, hogy 3,1 a populáció átlaga, de ennek esélye igen kicsi a 2000 fő adatai alapján számított értékek alapján. Ellenben ha „csupán" 200 tanuló vett részt a felmérésben, és az ő adataikból próbálunk következtetni az alapsokaság átlagára, akkor a 3,1-es érték

még beleesne a 95%-os valószínűséghez tartozó konfidencia-intervallumba. (Arra az ellenvetésre, hogy „hiszen 200 tanuló adataiból nem is akarunk soha a teljes alapsokaságra következtetni”, a statisztikai szempontú válasz az, hogy a 200 tanuló elvileg végtelen sokszor mérhető. Ebben az esetben az adott mintán elvileg végtelen sok méréssel kapható adatok populációjáról van szó.) Az eddigiek alapján nyilvánvaló, hogy egy végtelen alapsokaság statisztikai szempontú megismeréséhez mennyire tág becslés az, ami 20 fő eredményei alapján adódott.

Az 1. ábra a mintanagyság és a konfidencia-intervallum sugara közötti összefüggést grafikus formában is szemlélteti:



1. ábra
Összefüggés a mintanagyság és a konfidencia-intervallum sugara között 1,0 elméleti szórást feltételezve

Az ábráról szembetűnő, hogy amennyiben a vízszintes tengelyen több ezres mintanagyságokig történik az ábrázolás, akkor nagyon hirtelen ellaposodik a görbe 1000–2000 főnél nagyobb értékeknél. Ez másképpen fogalmazva azt jelenti, hogy a mérés pontossága alig-alig fokozható újabb ezrek felvételével a mintába. Az igazán jelentős különbségek a mérés pontosságának fokozásában az ezer fő alatti minták között lehetségesek.

Pedagógiai döntéshozatal és a mintanagyság

A szakmai közvélemény részéről érezhető az elvárás, hogy nagy mintás felmérések eredményei alapján hozzunk szakmai döntéseket. Aligha tévedek, ha azt állítom, hogy ha két felmérés eredményei ugyanarra a jelenségre hívják fel a figyelmet, de míg az egyik

száz, a másik tízezer tanulói adatai alapján, akkor a szakmai közvélemény a második felmérés eredményét jobban dokumentáltnak, meggyőzőbbnek fogja tartani.

Egyik célom a tanulmány megírásával annak megmutatása, hogy bár bizonyos empirikus felmérések esetén szükségszerű több száz vagy több ezer fős minták alkalmazása, nem minden jelenség megismeréséhez érdemes nagy mintát alkalmazni. Más tudományterületről vett példát hozva: több milliós alapsokaság főbb jellemzőinek pontos meghatározása lehetővé válik egy relatíve kis minta megvizsgálásával. Köztudomású, hogy a politikai közvélemény-kutatások 1000–1500 fő adatain nyugszanak, akik legalább 3–4 szempontból reprezentálják az ország választópolgárait. Ugyanakkor egy-egy országos pedagógiai felmérésben gyakran több ezer tanuló vesz részt, vagyis az alapsokaság létszámához képest relatíve sokkal nagyobb minták jellemzőek a pedagógiai kutatásokban. Mintha létezne egy kimondatlan alapelv, amely szerint lehetőség szerint minden tanulóról adatokat szeretnénk gyűjteni egy-egy felmérésben, és ennek csupán az anyagi lehetőségek szabnak gátat. Egészen szélsőséges manifesztálódása ennek a kimondatlan alapelvnek minden olyan vizsgálat, amely a teljes populáció felmérését célozza meg.

Az érem másik oldalát jelentik azok a pedagógiai kísérletek, amelyeket néhány tucat tanuló bevonásával végeznek el. Világszerte gyakori, hogy 50–100 fős mintán kipróbált pedagógiai kísérletek zajlanak, amelyekben nem cél a teljes alapsokaságról információt gyűjteni, viszont a konkrét mintát és a kísérlet körülményeit nagyon pontosan leírják. A nemzetközi tendenciákat látva a következő gondolatokat érdemes megfogalmazni. A tudománynak mint a társadalom egyik alrendszerének egyik feladata, hogy más alrendszereket (pl. az oktatási rendszert, a gazdaságot, a honvédelmet stb.) a hatékony döntésekhez szükséges információkkal ellásson. Hosszabb távon az a tudományág képes a maga műhelyeit működtetni, amely képes felhasználható eredményeket előállítani. A neveléstudománynak akkor van szüksége pedagógiai kísérleteket kidolgozó és azok eredményeit publikáló műhelyekre, ha az így kapott eredmények például az oktatási rendszert érintő döntésekhez támpontot adnak. Tegyük fel, hogy teljesül ez a feltétel. Ebben az esetben a pedagógiai kísérletek szakmai korrektségének kérdése központi jelentőségűvé válik. A tanulmány következő részében a kísérletek szakmai korrektségének problémái közül a mintanagyság kérdését emelem ki, és az ehhez szükséges statisztikai szemléletmódot igyekszem érvényesíteni. Elöljáróban kijelenthető, hogy bár nagyobb mintán könnyebb szignifikáns különbséget kimutatni egy pedagógiai kísérlet kulcsváltozóinak átlagai között, a „túlzottan nagy” minták alkalmazása csökkenti a kísérleti hatás nagyságát.

Mintanagyság és hipotézisvizsgálatok összefüggései

A hipotézisvizsgálatok jelentősége: az elsőfajú hiba szerepe

Az eddigiekben a leíró statisztikai vizsgálatok esetén a nagy (több száz fős) minták kiválasztása mellett szóló érveket tekintettük át. A továbbiakban a matematikai statisztika eszközeihez fordulunk, és ismertnek tételezzük fel a statisztikai hipotézisvizsgálatokkal kapcsolatos alapvető fogalmakat: populáció, minta, nullhipotézis, elsőfajú hiba, má-

sodfajú hiba. (Ezek leírását lásd pl.: *Vargha András* (2000) könyvében.) *Vargha* a statisztikai hipotézisvizsgálatokat nyomozáshoz hasonlítja, amelyben a valóságról (a populációról) igyekszünk minél biztosabbat megtudni egy korlátozott nyomozati anyag (a minta) alapján. Elképzelhető, hogy végül az ártatlant bűnösnek minősítjük (elvetjük a nullhipotézist, pedig az igaz), és ekkor az elsőfajú hibát követjük el. Az is lehetséges, hogy a bűnöst fölmenti a nyomozás (megtartjuk a nullhipotézist, pedig az nem igaz), és ekkor a másodfajú hibát követjük el.

Mielőtt az optimális mintanagyság statisztikai megközelítésével foglalkoznánk, áttekintjük azokat a tényezőket, amelyek alapvetően meghatározzák a statisztikai hipotézisvizsgálatokra törekvés célját és módját. Induljunk ki abból, hogy a körülöttünk lévő világ alapvetően nullhipotézisekkel jól leírható. Lényegében minden tagadószó nélküli mondat, amelyben valamilyen dolog valamilyen tulajdonságáról állítunk valamit, megfogalmazható nullhipotézis formájában. Ezzel szemben a pedagógiai kutató számára gyakran a változás, a változtathatóság problémája válik kulcsfontosságúvá. Azok az állítások, amelyek verbalizálva a „változás", „különbség", „összefüggés" szavakhoz kapcsolhatók, mindig tekinthetők valamilyen nullhipotézis tagadásának, vagyis egy konkrét vizsgálat ellenhipotézisének.

A pedagógiai kutató legtöbbször úgy jut releváns eredményekhez, ha megcáfol egy nullhipotézist. A nullhipotézis cáfolatára való törekvés közben ezért – az eddigiekből következően – az elsőfajú hiba elkövetését igyekszünk kontroll alatt tartani. Több évtizedes hagyomány a viselkedés- és társadalomtudományokban, hogy legföljebb 5% esélyt adunk annak, hogy úgy vessük el a nullhipotézist, hogy az valójában igaz. A kutatónak 5% alatt kell tartania annak valószínűségét, hogy elköveti az elsőfajú hibát.

Számos kutató hangsúlyozta ugyanakkor, hogy mennyire káros a tudomány számára az a gyakorlat, amely a nullhipotézisek tesztelését (NHST – null hypothesis significance test) állítja középpontba. A *Harlow, Mulaik* és *Steiger* (1997) által szerkesztett kötet szerzőinek többsége a nullhipotézisek tesztelését preferáló gyakorlat ellentmondásait tárta föl. Az ellentmondások egy része logikai természetű, más része technikai probléma, és ezen túl vannak még az eredmények interpretációjával kapcsolatos kommunikációs gondok.

Cohen (1997/1994) szerint a nullhipotézisek tesztelése látszólag követi az arisztotelészi kétértékű logikát. Tegyük fel, hogy fönnáll a következő két premissza:

1) Ha a nullhipotézis igaz, akkor ez a szignifikanciára utaló adat (D) nem fordulhat elő.
2) D mégis előfordult.

Ebben az esetben (modus tollens következtetési szabály): (konklúzió) A nullhipotézis hamis.

A nullhipotézisek tesztelése során az első kiindulási premissza valószínűségi természetű állítás, amelynek a kétértékű logikai szillogizmusban szerepeltetése megalapozatlan következtetésre vezethet.

Még nyilvánvalóbb logikai természetű problémát hoz felszínre annak felismerése, hogy vajon a következő két mondat közül melyik vonatkozik a nullhipotézis tesztelésére:

1) Bizonyos adatok előfordulása esetén mennyi a valószínűsége annak, hogy H_0 igaz?

2) Feltéve, hogy H_0 igaz, mennyi a valószínűsége bizonyos adatok előfordulásának?

A nullhipotézisre vonatkozó szignifikanciateszt a (2)-es számú állításra ad választ, pedig a kutató az (1)-esre sokkal inkább kíváncsi lenne.

A nullhipotézisek tesztelésével kapcsolatos furcsa jelenségek extrém nagy minták alkalmazása esetén kerülnek napvilágra. *Cohen* (1997/1994) idézi *Meehl* és *Lykken* vizsgálatát, akik 57 ezer Minnesota-i középiskolás személyi adatai közötti összefüggés vizsgálatát végezték el. A 15 felvett változó (közöttük pl. a tanuló neme, a továbbtanulásra vonatkozó terve, vallási hovatartozása, szabadidős elfoglaltságai) közötti 105 lehetséges $\chi 2$-próba mindegyike szignifikáns összefüggést mutatott ki $p<0{,}000001$ szinten, vagyis 99,9999%-os szignifikancia-szinten. Az említett technikai problémát jelzem ezekkel az adatokkal, vagyis azt, hogy a leghihetőbb nullhipotézisek is elvethetők elég nagy minták esetén.

A nullhipotézis tesztelésével kapcsolatos szemléleti problémára a következő mondat szolgál példaként (*Schmidt* és *Hunter*, 1997. 39. o.): „Képesnek kell lennünk arra, hogy elválasszuk egymástól azokat a jelenségeket, amelyek valóságosak, azoktól, amelyek a véletlennek köszönhetők". Ugye egyetértünk abban, hogy a pénzfeldobás eredménye a véletlennek köszönhető (50–50%, hogy fej vagy írás)? Ugyanakkor érdemes figyelembe venni, hogy a nullhipotézis tesztelésekor fellépő másodfajú hiba alapján meghatározható, mekkora az esélye annak, hogy adott kísérletben egy, a populációban ténylegesen létező hatást a kutató ki tudott mutatni. A pszichológiai szakirodalomban a legjellemzőbb valószínűség, ami adódik: 40–60% közötti. Ez körül-belül a pénzfeldobásnál megfigyelt valószínűség. Igazolható, hogy számos releváns kutatási kérdéshez nem lenne lehetséges akkora mintát összegyűjteni, hogy a ténylegesen létező hatás kimutatásának valószínűsége nagyobb legyen, mint 50%.

Másodfajú hiba a hipotézisvizsgálatokban

Fontos szerepéhez mérten elhanyagolhatóan kis figyelem jut a statisztikai hipotézisvizsgálatokban a másodfajú hibának. A másodfajú hiba elkövetése azt jelenti, hogy megtartjuk a nullhipotézist, pedig a valóságban az nem igaz. Pedagógiai példa: Ha nagy a másodfajú hiba elkövetésének valószínűsége, akkor könnyen előfordulhat, hogy látszólag eredménytelenül zárul egy fejlesztő kísérlet, pedig valójában különbséget, hatást lehetett volna kimutatni. Ugyanez egy gyógyszerészeti analógiával: a kutató hatástalannak ítélhet egy valójában hatásos szert. (Visszakanyarodva a nyomozás-analógiához: A másodfajú hiba azt jelenti, hogy felmentjük a bűnöst.)

Az első- és másodfajú hiba egymással kapcsolatban álló mennyiségek, azonban nem kapható meg egyszerű számítással egyikből a másik. *Keppel* (1991) az F-eloszlás kapcsán szemlélteti az összefüggést, amely szerint egymással ellentétes a viszonyuk: ha nagyon szigorúak vagyunk az elsőfajú hiba elkövetését illetően, akkor ezzel megnöveljük a másodfajú hiba elkövetésének valószínűségét. Ha tanulócsoportok közötti különbséget nagyon-nagyon precízen szeretnénk kimutatni, és például 99%-os bizonyossággal teszünk a csoportok közötti különbségre vonatkozó megállapításokat, akkor megnöveljük a másodfajú hiba elkövetésének valószínűségét, tehát megnöveljük az esélyét annak, hogy egy fejlesztő program ténylegesen létező hatását nem ismerjük föl.

A statisztikai hipotézisvizsgálatok során elkövethető következtetési hibák számbavételét tehát a másodfajú hibával, és az annak komplementer eseményeként szereplő próbaerő (power) fontosságának hangsúlyozásával folytatjuk. Az elsőfajú hibát akkor lehet elkövetni, ha a valóságban a nullhipotézis igaz, ám a konkrét mintaadatok alapján a kutató az ellenkező következtetésre jut. (A pedagógiai kísérletek nyelvén: valójában nincs kísérleti hatás, de a konkrét mintánk eredményei alapján arra a következtetésre jutunk, hogy van.) Ezzel szemben a másodfajú hiba akkor követhető el, ha a valóságban nem igaz egy nullhipotézis, ám a konkrét felmérés adatai alapján azt mégis megtartjuk. (A pedagógiai kísérletek nyelvén: ténylegesen létezik a kísérleti hatás, de eredményeinkkel nem sikerül azt kimutatni.) Ebből a leírásból is látható, hogy szükséges lenne a másodfajú hiba elkövetési valószínűségének tudatos kontrolljára, hiszen ha egy ígéretes pedagógiai kísérlet a másodfajú hiba elkövetése által eredménytelennek mutatkozik, kérdéses, mennyire lesz támogató a szakmai közeg egy újabb ismétléshez. A másodfajú hiba elkövetésének valószínűsége szoros kapcsolatban áll annak valószínűségével, hogy a szakmai közösség reprodukálni tudja egy kísérlet eredményét. *Keppel* (1991. 68–69. o.) több kutató egybehangzó megállapításait idézve megállapítja, hogy számos, szakmai folyóiratban közölt kísérlet esetén 50% körüli átlag adódott a másodfajú hiba elkövetésének valószínűségére. Ez azt jelenti, hogy számos esetben 50%-os esély van arra, hogy egy kísérlet reprodukálása ugyanazt a statisztikai következtetést teszi lehetővé.

A pedagógiai kísérletek mintanagyságának problémája

A pedagógiai kísérletekkel kapcsolatos megállapításaink általánosíthatók a nem kísérleti körülmények között gyűjtött adatokkal végzett statisztikai hipotézisvizsgálatokra is. Azt, hogy a pedagógiai kísérletek nyelvén fogalmazzuk meg téziseinket, elsősorban a fejlesztő kísérletek kiemelt szakmai jelentőségével indokolható. A pedagógiai kísérletek körében egy vagy több független változó esetére egyaránt érvényes mondanivalónk, hiszen az egyetlen független változó esetén adódó legalább két részminta összehasonlítására ugyanúgy használható a variancia-analízis és az annak részadataira építő statisztikai formulák, mint a többtényezős (multifactorial) kísérletek esetén.

Először az elsőfajú hiba szerepét bemutatandó azt a kérdést vizsgáljuk, hogy mekkora mintára van szükség ahhoz, hogy valamely átlagok között meglévő, szakmai szempontból jelentős különbség statisztikai szempontból is jelentősnek mutatkozzék. Leegyszerűsítve a szokásos kísérleti körülményeket: tételezzük fel, hogy azonos létszámú a kísérleti és a kontroll csoport, az összehasonlítandó eredmények százalékban vannak kifejezve, a szórás mindkét mintában 10%. Ebben az esetben a következő összefüggés adódik: minél nagyobb a mintánk, annál kisebb különbség már statisztikailag jelentősnek bizonyul.

Rövid számítást végzünk arra vonatkozóan, hogy a kétmintás t-próba alkalmazásával végzett hipotézisvizsgálat mekkora nagyságú különbségeket mutat szignifikánsnak. A próbastatisztika képlete:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

A képletben szereplő ismeretleneket – egy kivétellel – rögzítjük.

- A 95%-os szinthez tartozó t-táblázatbeli értékek függnek az elemszámtól, ezért érdemes azokra egy felső becslést választani: a jelenség illusztrálására felhasznált mintanagyságok esetén a t=2,1 érték mindig megfelelő felső becslés lehet.
- Feltételezzük, hogy a két minta egyenlő létszámú, vagyis $n_1=n_2$.
- Mindkét mintában 10-nek tételezzük fel a szórást.
- A számlálóban szereplő mennyiséget, az átlagok különbségét x-szel jelöljük.

Behelyettesítve a fenti képletbe a rögzített értékeket, a következő egyismeretlenes, egy paraméteres (n) egyenlőtlenséghez jutunk:

$$2{,}1 < \frac{x}{\sqrt{\frac{(n-1) \cdot 100 + (n-1) \cdot 100}{2 \cdot (n-1)}} \cdot \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

Az egyenlőtlenséget az átlagok különbségére mint ismeretlenre rendezve, és a szükséges számításokat elvégezve:

$$x > \frac{29{,}70}{\sqrt{n}}$$

A következő táblázatban kerekített adatokkal illusztráljuk, hogy az imént rögzített feltételek mellett adott mintanagyság esetén hozzávetőlegesen milyen különbség szignifikáns p=0,05 szinten. Mivel az egyenlőtlenségben szerepeltetett 2,1-es t-táblázatbeli érték helyett valójában kisebb küszöbértékek vannak, az 5. táblázat értékei felső becslésnek tekinthetők. A táblázat alapján megfogalmazható, hogy növekvő mintaelemszámok mellett egyre kisebb – átlagok közötti – különbségek válnak statisztikailag szignifikánssá.

5. táblázat. A részminták elemszáma és a statisztikailag szignifikáns különbségek összefüggései

	N=50	N=100	N=200	N=400
Szignifikáns különbség	4,2%	3,0%	2,1%	1,5%

Annak a kérdésnek a megválaszolására, hogy szakmai szempontból mekkora különbséget tartunk relevánsnak, két megközelítési lehetőség kínálkozik. Egy „objektív”, amely például a képességfejlődés tempója szerint releváns időközökhöz tartozó változás mérté-

két tekinti szakmai szempontból jelentősnek, és egy „szubjektív", amely arra vonatkozna, hogy a gyakorló tanárok és az oktatási rendszerrel kapcsolatos döntések más szereplői mekkora mértékű különbséget tartanak szakmai szempontból jelentősnek. A PREFER mérés adatbázisából (*Nagy*, 1980) például ki lehetett számítani – ha minden egyes hónapban születettek közül legalább 300 gyerek szerepel a mintában –, hogy kb. kéthónapnyi születési dátumbeli különbség már statisztikailag szignifikáns különbséget eredményez a teljesítményben. Ha tehát szakmai szempontból a két hónap alatt bekövetkező tudásváltozás nagyságát tekintenénk szignifikánsnak, akkor az a 300 fő körüli minta esetén a PREFER teszt alapján kb. 1,5%-nyi teljesítménykülönbséget jelent. Egy másik lehetőség a szakmai szempontból jelentős változás mértékének megállapítására *Csapó* (2003) eredményei alapján kínálkozik. A γ mutató mint a tanulói teljesítmény változásának standard mértéke lehetővé teszi, hogy különböző évfolyamokon, különböző mérőeszközökkel mért teljesítmények változását összehasonlítsuk. Az eddig elvégzett mérések alapján az körvonalazódik, hogy átlagosnak tekinthető a változás mértéke, ha egy év alatt az átlagok különbsége ötszöröse a szórások átlagának (azaz γ értéke 0,2 körüli).

Feltételezhető, hogy ha egy szubjektív küszöböt keresnénk (például szóbeli interjú vagy kérdőíves vizsgálat módszerével), akkor 1,5%-nál nagyobb teljesítménykülönbségről nyilatkozna úgy a többség, hogy az szakmai szempontból releváns. Ekkor viszont egy néhány száz fős minta már túlzottan nagynak bizonyulna, vagyis a szakmailag nem jelentősnek tartott különbségek is szignifikánsnak bizonyulnának – statisztikai értelemben.

Keppel (1991) szerint nem szerencsés az elsőfajú hiba elkövetésének valószínűségével jellemezni, hogy egy-egy kísérlet mennyire meggyőzően mutatja egy jelenség hatását. Szerinte „túlságosan gyakran" (64. o.) megtörténik, hogy a kutatók a $p<0{,}00001$-es valószínűségi szinten szignifikáns eredményt meggyőző erejűnek tartják, míg a $p<0{,}05$-ös csak futólagos figyelmet érdemel. Ezek a kutatók figyelmen kívül hagyják, hogy a minta nagyságának nagy szerepe van abban, hogy egy nullhipotézis elvetésekor milyen nagyságú az elsőfajú hiba elkövetésének valószínűsége.

A továbbikaban áttérünk a másodfajú hiba és a mintanagyság összefüggésének vizsgálatára. A másodfajú hibával együtt kezeljük a kísérlet érzékenységének mérőszámát, a próba erejét. A másodfajú hiba és a próba ereje akkor lép színre, amikor a valóságban létezik a kísérletnek tulajdonítható különbség. Annak valószínűsége, hogy ezt sikerül kimutatni, a próba ereje, és ennek komplementer eseménye, vagyis hogy a nullhipotézist megtartva nem mutatjuk ki a szóban forgó, ténylegesen létező különbséget, a másodfajú hiba. Amennyiben az elsőfajú hiba elkövetésének valószínűségét α-val jelöljük, a másodfajú hibáét β-val, akkor a kísérlet érzékenységét kifejező mérőszám: $1-\beta$. A másodfajú hiba elkövetésének valószínűsége (más tényezők változatlanul hagyása mellett) a minta elemszámának növelésével csökken.

A másodfajú hibát, és ezzel együtt a próba erejét meghatározó dolgok közé tartozik több tényező: az elsőfajú hiba elkövetésének valószínűsége (ezt p-vel jelöljük, de gyakori az α jelölés is), a minta mérete, a tanulók közötti és a mérési hibából fakadó variancia, a kísérleti hatás. Az itt említett négy tényező közül a kísérleti hatás meghatározása igényli a legtöbb magyarázatot. Az első- és másodfajú hiba mintanagysággal vett összefüggései mellett egy harmadik sarokpont a mintanagyság meghatározásához a kísérleti

hatás és a mintanagyság összefüggésének vizsgálata. A következőkben ezt vizsgáljuk meg kicsit részletesebben.

A kísérleti hatás mértéke – per definitionem – azt számszerűsíti, hogy a megfigyelt különbségek milyen mértékben tulajdoníthatók a kísérleti beavatkozásnak. Ebben a tanulmányban két, kiindulási ötletében egymástól jelentősen különböző megközelítésmódot használunk föl a kísérleti hatás értelmezésére. Közös a két megközelítésmódban, hogy variancia-analízissel vizsgálható kísérleti elrendezéseken alapulnak, vagyis kettő vagy több minta átlagai és varianciái szerepelnek bennük. Mindkét megközelítés mértékegység nélküli jelölőszámokat használ föl a kísérleti hatás becslésére. Az első megközelítésmódból származik az f, a másodikból az η^2-és az ω^2-mutató.

Elsőként az ötvenes-hatvanas években már részletesen tanulmányozott f-(„kis f”-) mutatót érintjük (*Cohen*, 1969). Az f kifejlesztése azt a célt igyekezett megvalósítani, hogy a két egyenlő létszámú minta esetén lehető legegyszerűbbnek számító d-mutató ($d = \frac{m_1 - m_2}{\sigma}$) általánosítható legyen több minta (pl. többtényezős kísérletek) esetére is. A d képletében egyszerűen a két mintaátlag különbsége szerepel, osztva a közös populáció szórásával. Az így kapott hányados hasonlít a t-statisztika képletére, ahol a számlálóban ugyancsak átlagok különbsége szerepelt, a nevezőben pedig a minták szabadsági fokaival súlyozott szórásértékek szerepeltek.

A d képletében a számláló fölfogható úgy, mint adott populációból származó mintaátlagok kételemű mintájának terjedelme. Az f képletben az általánosítás lényege, hogy a mintaátlagok szóródásának mérőszámaként nem a terjedelmet, hanem a szórást választjuk:

$f = \frac{\sigma_m}{\sigma}$, ahol σ a d képletében is szereplő, közös populációbeli szórás, míg σ_m az átlagok különbsége helyett a k db egyenlő létszámú minta esetén a mintaátlagok szórása:

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}$$

Nem egyenlő létszámú minták esetében az alapstruktúra ugyanaz marad, vagyis f ugyanúgy két érték hányadosaként áll elő, és σ értelemszerűen ugyanaz marad. p_i-vel jelölve az i-edik minta elemszámának és az összes elemnek hányadosát, a következő általánosabb képlet adódik, amelybe $p_i=1/k$ értékeket behelyettesítve előáll az egyenlő létszámú mintákra adott összefüggés.

$\sigma_m = \sqrt{\sum_{i=1}^{k} p_i \cdot (m_i - m)^2}$, ahol m a mintaátlagok súlyozott átlaga.

Cohen (1969) három f-értékhez fűz kvalitatív leírást: f=0,1 kicsi (small) kísérleti hatást jelez, f=0,25 közepes (medium), f=0,40 pedig jelentős (large) kísérleti hatásként interpretálható. Jelen esetben lényeges kérdés, hogy mi az összefüggés a kísérleti hatás nagysága és a kísérleti mintanagyság között. Egyetlen kísérleti és egy vele azonos létszámú kontrollcsoportot feltételezve, a próba erejét (és így a másodfajú hiba elkövetésé-

nek valószínűségét is) 50%-nak véve, a következő értékeket (l. 6. táblázat) találjuk p=0,05 szint mellett.

6. táblázat. Mintaelemszámok az f hatásméret függvényében p=0,05 és β=0,50 szinten [forrás: Cohen (1969. 377. o.)]

f	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,50	0,60	0,70	0,80
n	769	193	86	49	32	22	17	13	9	7	5	4

A táblázatban megfigyelhető tendencia először talán meglepő lehet: egyre nagyobb kísérleti hatáshoz egyre csökkenő mintaelemszám tartozik. Ha megfordítjuk a kapcsolat irányát, akkor könnyebb a magyarázatot megtalálni: minél kevesebb kísérleti személy adataiból születik meg a kísérleti eredmény, annál nagyobb hatással kellett rendelkeznie magának a kísérleti elrendezésnek (a kísérlet tényének) a résztvevők eredményei közötti különbségek kialakítására. Ha p=0,05 szinten mutatunk ki különbséget a kísérleti és kontroll csoport átlaga között, akkor nagy minta esetén kicsiny különbség is statisztikailag jelentőssé válik; azt a kicsiny különbséget pedig egy nagy minta esetén ezernyi tényező okozhatja, és magának a kísérleti elrendezésnek, a kísérlet tényének szerepe relatíve lecsökken.

A kísérlet relatív magyarázó erejének említésével eljutottunk a másik felfogáshoz, amely a kísérleti hatás számszerűsítését állítja középpontba. A megmagyarázott variancia fogalmáról van szó, amely egy korreláció-jellegű mutató. *Cohen* (1969) valamint *Steiger* és *Fouladi* (1997) is részletesen elemzi az úgynevezett η^2-mutatót. *Keppel*nél (1991) ω^2 néven szerepel az a mennyiség, amely azt mutatja meg, hogy az eredmények varianciájának hány százaléka tudható be a kísérleti elrendezésnek, a kísérlet tényének. A matematikai statisztika három alapfogalma, az átlag, a variancia és a korreláció „összeérnek”, amikor a kísérleti hatás definíciójáról van szó.

Ha a kísérleti hatás kiszámításához a megmagyarázott variancia fogalmán keresztül közelítünk, akkor két variancia hányadosaként írhatjuk föl a kísérleti hatást: a mintaátlagok közötti varianciát osztjuk a teljes varianciával. A hányadosban szereplő elméleti értékek becslésére a variancia-analízis lépéseiben kiszámolt mennyiségeket használhatjuk. Az η^2 és az ω^2 mutatók egyaránt arra valók, hogy segítségükkel a megmagyarázott variancia értékére becslést adjunk. Az η^2 az f-mutató függvényeként is megadható:

$$\eta^2 = \frac{f^2}{1+f^2}$$

ω^2 becslésére a legegyszerűbb út, ha a variancia-analízisben szereplő Fischer-féle F használásával írjuk föl (*Keppel*, 1991):

$\overline{\omega}^2 = \frac{(a-1)\cdot(F-1)}{(a-1)\cdot(F-1)+a\cdot n}$, ahol *a* a kísérletben résztvevő csoportok száma, n pedig az egy-egy kísérleti csoportban található mintaelemek száma. Észrevehető, hogy itt is azt a legegyszerűbb esetet dolgoztuk föl, amikor a kísérleti csoport(ok) és a kontroll csoport létszáma megegyezik. Amennyiben különböző létszámú csoportjaink vannak, úgy az ω^2 mutató *Keppel* szerint nem használható föl jól a kísérleti hatás becslésére. Ezekben az esetekben a fentebb definiált f mutató számítható ki, és felhasználhatjuk az η^2 és f^2 kö-

zötti összefüggést. A két becslés, η^2 és ω^2 eléggé hasonló értékeket ad. Elméleti szempontból az ω^2 mutató a jobb (kevésbé torzított becslést ad), és talán ezért nevezi *Keppel* (64. o.) a „legnépszerűbb mutatónak" – egyenlő létszámú minták esetén.

Saját kutatásunk példáján bemutatjuk az ω^2 becslésének eljárását. Fejlesztő kísérletünkben (*Csíkos*, 2005) négy utóteszt szerepelt, amelyek alapján összehasonlítható a kísérleti és kontroll csoportok teljesítménye. Mind a négy mérőeszköz esetén szignifikáns különbséget mutatott a kétmintás t-próba. A kísérleti hatás becsléséhez szükséges F értékeket és a kiszámított $\overline{\omega}^2$ mutatókat a 7. táblázat tartalmazza.

7. táblázat. „A metakogníció iskolai fejlesztésének alapjai" kísérletben szereplő utóteszteken mért kísérleti hatások

Mérőeszköz	*F-hányados*	$\overline{\omega}^2$	$\overline{\omega}^2$ *(%)*
Matematikai szöveges feladatok	43,92	0,200	20,0
Matematikai tudásszintmérő teszt	4,35	0,019	1,9
Szövegértés-teszt vegyes típusú szövegekkel	9,99	0,050	5,0
Szövegértés teszt dokumentum jellegű szövegekkel	5,38	0,025	2,5

Megjegyzés: A kísérleti hatás becslésére használt képletben szereplő többi mutató értékei: a=2, n=86. A táblázat F értékeit két tizedesre kerekítettük, de a kísérleti hatás kiszámításánál ennél több jeggyel számoltunk. Az F-hányados matematikai értelmezése meglehetősen absztrakt, ám *Csapó* (2004) tanulmányában szemléletes jelentést nyert.

A kísérleti hatás kiszámítási módjából látható, hogy az rendkívül érzékeny a mintanagyságra. Ha nagyobb minta felhasználása mellett adódik ugyanaz az F érték, akkor a kísérleti hatás mértéke kisebb lesz. A kísérleti módszereket alkalmazó társadalomtudományokban felhalmozódó tapasztalat szerint létezik egy optimális mintaelemszám, amelynek alkalmazása esetén a kísérleti hatás nagysága meggyőző, és ugyanakkor az első- és másodfajú hiba elkövetésének valószínűsége is megfelelő. E mintanagyság meghatározása matematikai alapokon nyugvó feladat. *Keppel* (1991) három tényezőt veszfigyelembe: (1) az elsőfajú hiba elkövetésének valószínűsége (p), (2) a kísérleti hatás (ω^2), (3) a próba ereje (1-β). Valójában ezen egymásra ható tényezők közül bármelyik kiemelhető, és vizsgálható a többiek függvényében.

1) Az elsőfajú hiba elkövetésének valószínűségét, avagy a szignifikancia-szintet a hagyomány okán 0,05-nak szokás választani. Ritkán célszerű ennél alacsonyabb értéket választani, mert megnőhet a másodfajú hiba elkövetésének valószínűsége.

2) A kísérleti hatás mérésének legnépszerűbb mutatója az úgynevezett ω^2-index. (omega-négyzet). Az ω^2 értékeit három kategóriába sorolják: 0,01 értéknél alacsony (small), 0,06 esetén közepes (medium), 0,15-nél és e fölött jelentős (large) kísérleti hatásról szokás beszélni. (Ezek a számadatok egybecsengenek az f értékekre megállapított jelzőkkel.)

3) A próba ereje nem mindig kapja meg a neki járó kitüntetett figyelmet. A már említett 0,5 körüli értékek, amelyek azt mutatják, hogy egy valójában létező hatás megismételt kísérleti kimutatására 50%-os esély van, nagyon alacsonynak nevezhetők. *Keppel* (1991. 73. o.) szavai szerint: „Ha a próba ereje alacsony, rossz minőségű a tudományos kutatás – időt, energiát és anyagi forrásokat fecsérlünk el".

A 8. táblázat mutatja az említett tényezők egymásra hatását. *Mulaik*, *Raju* és *Harshman* (1997) szerint négy, egymással kölcsönösen összefüggő mennyiség – a próba ereje, az elsőfajú hiba elkövetésének valószínűsége, a mintanagyság és a kísérleti hatás – közül bármelyik három megadása egyértelműen meghatározza a negyediket. Ezért egy kétdimenziós táblázatban, amely a négy mennyiség egymásra hatását illusztrálja, érdemes a négy tényező közül rögzíteni egyet: a pedagógiai vizsgálatokban legkevésbé az elsőfajú hibához kapcsolódó szignifikancia-szint változik.

A 8. táblázatban p értékének rögzítése mellett két tényező, a próba ereje és a kísérleti hatás adja az oszlopok és sorok fejlécét, a cellákba pedig a két változó eredőjenként a mintanagyságok kerülnek. A táblázat azt mutatja be, hogyan hat egymásra rögzített p=0,05 érték mellett a próba ereje, a kísérleti hatás és a mintanagyság. A táblázat értékei egy négyféle kísérleti körülményt (például négyféle olvasástanítási módszer) feltételező kísérlethez adottak, ahol mind a négy kísérleti csoportban ugyanannyi tanuló vesz részt.

8. táblázat. A próba ereje p=0,05 rögzített érték mellett, négyféle kísérleti csoportot feltételezve, a kísérleti hatásméret és a mintanagyság függvényében [forrás: Keppel (1991. 72. o.)]

ω^2	*A próba ereje (1-β)*								
	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,01	21	53	83	113	144	179	219	271	354
0,06	5	10	14	19	24	30	36	44	57
0,15	3	5	6	8	10	12	14	17	22

A fenti definíció alapján minél nagyobb a próba ereje, annál kisebb a másodfajú hiba (β) elkövetésének valószínűsége, ezt igazolja, hogy a táblázatban jobbra haladva, az elsőfajú hiba elkövetési valószínűségeként választott 5%-os szintet szigorúan megtartva adott hatásméret mellett egyre nagyobb mintára van szükség ahhoz, hogy a másodfajú hiba elkövetésének valószínűsége csökkenjen. A táblázat alapján a kísérleti hatás mértékének legelfogadottabb együtthatójával fordítottan függ össze az, hogy milyen mintanagyság mellett mutatható ki nagyobb kísérleti hatás. Ennek oka, hogy a kísérleti hatás a tanulók között egyébként is meglévő különbségek mellett számszerűsíti, hogy milyen mértékű különbséget okozott a kísérleti elrendezés. Minél több a mintaelem, annál nagyobbak és sokrétűbbek lehetnek a tanulók közötti és a mérési hibából adódó különbségek, és annál tompábban jelentkezhet egy adott kísérleti hatás.

A táblázat alapján akár olyan elhamarkodott következtetés is születhet, mely szerint a pedagógiai kísérleteket egy-egy osztálynyi méretű, 22 fős kísérleti és kontroll csoporttal

ideális megtervezni. Ez ellen szóló legerősebb statisztikai ellenérvünk, hogy így az átlagok változását nézve a szokásosan szakmailag relevánsnak számító különbségeknél nagyobb különbségek lennének statisztikailag szignifikánsak.

Amennyiben a közepes hatásméret elérésével megelégszünk, és a másodfajú hiba elkövetésének valószínűségét 10% alatt szeretnénk tartani, akkor statisztikai alapokon állva 50–100 fő közötti mintanagyság javasolható a pedagógiai kísérletek kísérleti és kontroll csoportjainak létszámaként. Ha fellapozunk tanítással-tanulással foglalkozó külföldi szaklapokat, azokban leggyakrabban ilyen mintanagysággal lebonyolított pedagógiai kísérleteket találunk.

Összegzés

A tanulmány az empirikus pedagógiai vizsgálatok mintanagyságát meghatározó tényezőkkel foglalkozott, és két nagy területet emeltünk ki: (1) nagymintás felmérések, (2) pedagógiai kísérletek.

A tanulmány első részében a nagymintás felmérések mintanagyságát meghatározó tényezők közül hármat emeltünk ki: (1) megmutattuk, hogy a minta reprezentativitásának szempontjai hogyan többszörözik meg a felmérésbe bevonandó minta nagyságát, (2) illusztráltuk a mintanagyság és a konfidencia-intervallum közötti összefüggést, (3) érintettük a kutatási eredmények relevanciájának, megalapozottságának kérdését – a mintanagyság szempontjából.

A pedagógiai kísérletek mintanagyságának meghatározásához elemeztük az első- és másodfajú hiba elkövetési valószínűségének szerepét, és bemutattunk néhány mérőszámot a kísérleti hatás meghatározására. Mindhárom tényező hatását megvizsgáltuk a kísérleti minta nagyságára, és néhány érték rögzítése mellett példákkal illusztráltuk a szükséges mintanagyság változását az említett tényezők függvényében.

Reményeink szerint a tanulmány az elméleti összefüggések szemléletet formáló erején túl a konkrét példákon keresztül is segítséget nyújt a hazai pedagógiai kutatóknak empirikus vizsgálatok tervezéséhez és azok mintanagyságának meghatározásához.

Köszönetnyilvánítás

Köszönettel tartozom *Csapó Benőnek*, *Józsa Krisztiánnak*, *Ollé Jánosnak* és *Kelemen Ritának* a tanulmány korábbi fogalmazványához fűzött értékes kritikai megjegyzéseikért.

A tanulmány „A metakogníció iskolai fejlesztésének alapjai” F038222 sz. OTKA kutatás támogatásával született.

Irodalom

Babbie, E. (2000): *A társadalomtudományi kutatás gyakorlata*. Ötödik kiadás. Balassi Kiadó, Budapest.

Cohen, J. (1969): *Statistical power analysis for the behavioral sciences*. Academic Press, New York, London.

Cohen, J. (1997/1994): The Earth is round (p<.05). In: Harlow, L. L., Mulaik, S. A. és Steiger, J. H. (szerk.): *What if there were no significance tests?* Lawrence Erlbaum Associates, Mahwah, NJ, London. 21–35.

Csapó Benő (2003): *A képességek fejlődése és iskolai fejlesztése*. Akadémiai Kiadó, Budapest.

Csapó Benő (2004): Az iskolai osztályok közötti különbségek és az oktatási rendszer demokratizálása. In: Csapó Benő: *Tudás és iskola*. Műszaki Könyvkiadó, Budapest. 225–241.

Csíkos Csaba (2005): A metacognition-based training program in grade 4 in the fields of mathematics and reading. Paper presented at the 11th Biennial Conference for Research on Learning and Instruction, Nicosia, Cyprus, 23–27 August.

Halász Gábor és Lannert Judit (2003, szerk.): *Jelentés a magyar közoktatásról 2003*. Országos Közoktatási Intézet, Budapest.

Harlow, L. L., Mulaik, S. A. és Steiger, J. H. (1997, szerk.): *What if there were no significance tests?* Lawrence Erlbaum Associates, Mahwah, NJ – London.

Józsa Krisztián (2004): Az első osztályos tanulók elemi alapkészségeinek fejlettsége – Egy longitudinális kutatás első mérési pontja. *Iskolakultúra*, 11. sz. 3–16.

Keppel, G. (1991): *Design and analysis. A researcher's handbook*. Prentice Hall, Englewood Cliffs, NJ.

Nagy József (1980): *5–6 éves gyermekeink iskolakészültsége*. Akadémiai Kiadó, Budapest.

Organisation for Economic Co-operation and Development (2005): *PISA 2003 Technical Report*. [on-line] Web: http://213.253.134.29/oecd/pdfs/browseit/9805011e.pdf

Schmidt, F. L és Hunter, J. E. (1997): Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow, L. L., Mulaik, S. A. és Steiger, J. H. (szerk.): *What if there were no significance tests?* Lawrence Erlbaum Associates, Mahwah, NJ – London. 37–64.

Steiger, J. H. és Fouladi, R. T. (1997): Noncentrality interval estimation and the evaluation of statistical models. In: Harlow, L. L., Mulaik, S. A. és Steiger, J. H. (szerk.): *What if there were no significance tests?* Lawrence Erlbaum Associates, Mahwah, NJ – London. 221–257.

Vargha András (2000): *Matematikai statisztika pszichológiai, biológiai és nyelvészeti alkalmazásokkal*. Pólya Kiadó, Budapest.

ABSTRACT

CSABA CSÍKOS: DETERMINING SAMPLE SIZE IN EDUCATIONAL INVESTIGATIONS

This study focuses on the factors that affect sample size in educational investigations. Two types of research are addressed: (1) large sample investigations and (2) educational experiments. In large sample investigations three factors are analyzed: i) the number of aspects from which a large sample can be considered representative; ii) the role of confidence intervals in the estimation of population mean; and iii) subjective factors concerning relevance and soundness of research results. Determining sample size in educational experiments requires taking account of i) the probability of type I error; ii) the power of the experiment; and iii) effect size. Beyond discussing theoretical relations between factors affecting sample size, detailed examples and illustrations are given in order to help researchers in designing investigations and determining sample size.

Magyar Pedagógia, **104**. Number 2. 183–201. (2004)

Levelezési cím / Address for correspondence: Csaba Csíkos, Department of Education
University of Szeged.