Statisztikai Szemle

Közzététel: 2026. március 30.

A tanulmány címe:

A mintavételből eredő torzítások: elméleti példák és korrekciós lehetőségek

Szerzők:
NAGY-GYÖRGY JUDIT
a Szegedi Tudományegyetem Bolyai Intézetének egyetemi docense
E-mail: Nagy-Gyorgy@math.u-szeged.hu
SZEITL BLANKA
a Szegedi Tudományegyetem Bolyai Intézetének adjunktusa
E-mail: Szeitl@math.u-szeged.hu

DOI: https://doi.org/10.20311/stat2026.03.hu0209

Nagy-György Judit – Szeitl Blanka

A mintavételből eredő torzítások: elméleti példák és korrekciós lehetőségek

Sampling biases: theoretical examples and possible corrections

Nagy-György Judit, a Szegedi Tudományegyetem Bolyai Intézetének egyetemi docense
E-mail: Nagy-Gyorgy@math.u-szeged.hu
Szeitl Blanka, a Szegedi Tudományegyetem Bolyai Intézetének adjunktusa
E-mail: Szeitl@math.u-szeged.hu

Tanulmányunk néhány különösen problémás mintavételi helyzetet mutat be, és az ezekből fakadó torzításokat Monte Carlo-szimuláció segítségével szemlélteti. Az elemzés középpontjában a statisztikai eljárások egyik kulcsfontosságú feltétele, a mintaelemek függetlensége áll, valamint az annak megsértéséből adódó következmények. A bemutatott példák homogén és rétegzett populációkra egyaránt kiterjednek, és elsősorban elméleti szemléltető célt szolgálnak. Bemutatunk ismert, de nem köztudott korrekciós módszereket is, továbbá néhány javaslatot is teszünk, hozzájárulva a statisztikai elmélet és az alkalmazó kutatások módszertani gyakorlatának összekapcsolásához.

Kulcsszavak: mintavétel, függetlenség, becslés, torzítás

During statistical analyses, the sample plays a central role in ensuring that the assumptions of the applied procedures are satisfied, a condition that is often not met in practice. This paper presents several particularly problematic sampling situations and illustrates the resulting biases using Monte Carlo simulations. The analysis focuses on one of the key assumptions of statistical procedures, namely the independence of sample elements, and on the consequences arising from its violation. The presented examples cover both homogeneous and stratified populations and primarily serve a theoretical illustrative purpose. In addition, we propose methods for correcting these biases, thereby contributing to the integration of statistical theory and the methodological practice of research that applies it.

Keywords: sampling, IID, bias

A statisztikai elemzés célja, hogy egy sokaság jellemzőiről információt nyerjünk, gyakran anélkül, hogy annak minden egyes elemét megfigyelnénk. A következtető statisztikai vizsgálatok esetében ez jellemzően egy megelőző lépést feltételez: a mintavételt. Ennek eredménye az az adathalmaz, amelyen elemzések történnek, és

alapvetően meghatározza az elemzések érvényességét. Megjegyezzük, hogy statisztikai elemzés teljes körű adatfelvételek esetén is végezhető, jelen tanulmány azonban azokra a helyzetekre koncentrál, amikor a következtetések mintán alapulnak. A statisztikai gondolkodás fejlődésében a mintavételi eljárások kidolgozása kiemelkedő jelentőségű. Már *Neyman (1934)* klasszikus tanulmánya is rámutat arra, hogy a jól megtervezett mintavétel nem csupán hatékonyabbá teszi az adatgyűjtést, hanem biztosítja a becslések torzítatlanságát és konzisztenciáját is. A mintavételen alapuló adatgyűjtésnek központi szerepe van az empirikus adatokat alkalmazó elemzésekben is. A mintavétel az ilyen típusú kutatások egyik legkritikusabb módszertani eleme, alapvetően meghatározza a következtetések megbízhatóságát *(Babbie, 2008)*. Cikkünk a mintavételi módból eredő torzításokról szól, és bemutat olyan lehetséges korrekciókat, amelyekkel a torzítások csökkenthetők. Eredményeink ezáltal kifejezetten a mintavételt alkalmazó kutatások esetében és az adatokból származó becslések készítésénél hasznosak. Fontos különbséget tenni a minta mint adathalmaz, valamint a mintavételi eljárás mint adatgyűjtési módszer között. Míg az utóbbi célja egy sokaság reprezentatív leképezése, addig a statisztikai elemzés szempontjából a minta mint megfigyeléshalmaz képezi az elemzés tárgyát. A továbbiakban a minta fogalmát tág értelemben használjuk: egy véges számú megfigyelést értünk alatta, amelyen statisztikai elemzéseket végzünk, függetlenül attól, hogy az adatok egy explicit mintavételi eljárás eredményeként vagy más adatgyűjtési folyamat melléktermékeként álltak-e elő.

A mintavételi torzítások kérdése rendre megjelenik a szakirodalomban. Ennek oka kettős. Egyrészt a matematikai statisztikai modellek gyakran ideális feltételezésekre épülnek, például arra, hogy a mintavétel visszatevéses, és ezáltal a mintaelemek egymástól függetlenek, vagy arra, hogy végtelen a populáció mérete, amelyből a mintaelemeket kiválasztjuk. Ezek az előfeltevések lehetővé teszik a klasszikus tételek (például a centrális határeloszlás tétele vagy a nagy számok törvénye) alkalmazását. Az adatfelvételek gyakorlatában viszont a visszatevéses mintavétel vagy megvalósíthatatlan, vagy redundáns információt eredményez, ezért a kutatók túlnyomórészt visszatevés nélküli mintavételt használnak. Ez azonban szükségszerűen sérti a függetlenség feltételét, ami a becslések varianciájára és konfidencia-intervallumainak megbízhatóságára is hatással van. *Kish (1965)* részletesen foglalkozik a *design effect* fogalmával, amely éppen arra hívja fel a figyelmet, hogy a mintavétel módja (és kifejezetten a mintaelemek csoportosítása és a mintavétel rétegezettsége) jelentősen befolyásolhatja a statisztikai következtetések megbízhatóságát.

A hazai szakirodalom is rámutat arra, hogy például a társadalomtudományokban alkalmazott mintavételi eljárások matematikai megközelítése nélkülözhetetlen. Korábbi munkákban megjelenik, hogy a ritka populációk mintavételi stratégiái esetében speciális eljárások nélkül a becslések bizonytalansága jelentősen

megnő *(Kapitány, 2010)*, továbbá rétegzett minták esetében a megfelelő mintavételi design képes mérsékelni a varianciát és javítani a reprezentativitást *(Galambosné Tiszberger, 2011)*. A függetlenség feltételének megsértéséhez kapcsolódó problémákra a válaszadó-vezérelt mintavétel adhat módszertani választ, ez viszont nem eredményez valóban véletlen mintát *(Simon, 2012)*. Újabb példa kifejezetten a mintavételezésből származó pénzügyi adatok vizsgálata, bemutatva, hogy a mintavételi design nemcsak a statisztikai tulajdonságok, hanem a kockázatkezelésre vonatkozó becslések szempontjából is meghatározó *(Varga, 2021)*. Cikkünk ebbe a gondolatmenetbe illeszkedik: célunk, hogy a függetlenség sérülésének hatásait matematikai modellekkel és szimulációkkal érzékeltessük, valamint javaslatokat fogalmazzunk meg a torzítások korrigálására.

A statisztikai mintavétel elméletének és gyakorlatának folyamatos párbeszédben kell állnia egymással. A matematikai statisztika ideális modelljei megadják a becslések tulajdonságainak vizsgálati keretét, míg a gyakorlati tapasztalatok rávilágítanak arra, hogy a valóságban ezek a feltételek maradéktalanul csak ritkán teljesülnek. A függetlenség hiányából fakadó torzítások és a mintavételi eljárásokból származó hibák korrekciója ezért alapvető fontosságú. E korrekciókhoz olyan módszerek szükségesek, amelyek egyszerre támaszkodnak a matematikai statisztika tételeire és a gyakorlati tapasztalatokra *(Särndal et al., 1992; Lohr, 2010)*.

Jelen cikk további célja, hogy bemutassa a függetlenség feltételének a mintavételből származó becslések tulajdonságaiban betöltött szerepét, és feltárja, milyen következményekkel jár, ha ez a feltétel sérül. A tanulmányban egyrészt ismertetjük a kapcsolódó matematikai statisztikai modellt, amely lehetővé teszi a különbségek formális leírását, másrészt szimulációkkal illusztráljuk a visszatevéses és a visszatevés nélküli mintavétel közötti eltéréseket, valamint a rétegzett mintavételből fakadó torzításokat. Szeretnénk rámutatni arra, hogy a klasszikus függetlenségi feltevés megsértése nem pusztán elméleti kérdés, hanem a gyakorlati adatfelvételek során is jelentős hatással van az eredményekre. Végül javaslatokat teszünk arra, hogy miként kezelhető a függetlenség sérülése, és hogyan kapcsolható össze a matematikai statisztika ideális kerete a valós kutatási helyzetekből fakadó kihívásokkal.

A fentiek által a tanulmány egyszerre kíván hozzájárulni a statisztikai elmélethez és az empirikus kutatások módszertani gyakorlatához. Az elméleti megalapozás és a szimulációs illusztrációk révén igyekszünk hidat képezni a matematikai statisztikai modellek és a valós adatgyűjtések problémái között. Úgy véljük, hogy a mintavételből eredő torzítások és korrekciójuk megértése nemcsak a statisztikusok, hanem minden empirikus kutató számára alapvető fontosságú.

Homogén populáció esetén alapvetően két mintavételi eljárást tekintünk át: a visszatevésest és a visszatevés nélkülit. Mindkét esetben uniform módon választjuk a mintaelemeket. Foglalkozunk rétegzett populációból történő mintavételezési

eljárásokkal is. Az empirikus kutatásokban gyakran alkalmazott mintavételi eljárás például a többlépcsős, csoportos mintavétel. Ennek során a populáció elemei első lépésben csoportokba (elsődleges mintavételi egységekbe) rendeződnek, majd a mintavétel második lépésében ezek közül választunk ki csoportokat, a továbbiakban pedig a kiválasztott csoportokon belül történik az adatgyűjtés. Ez a struktúra a gyakorlatban természetesen további lépésekkel bővíthető, és a kutatás céljai alapján kerül meghatározásra az egyes lépések során alkalmazandó konkrét mintavételi módszer. Klasszikus példa az ilyen típusú mintavételre, amikor iskolai osztályok elsődleges mintavételi egységként jelennek meg. Az elsődleges mintavételi egységek kiválasztása gyakran méretarányos bekerülési valószínűségekkel *(probability proportional to size*, PPS) történik. A csoportos mintavétel sajátossága, hogy a teljes populáció nem minden csoportja kerül kiválasztásra, ami a csoportokon belüli homogenitás és a csoportok közötti különbségek miatt torzításhoz vezethet. Az is előfordulhat, hogy a populáció olyan – a kutató számára nem ismert – strukturális tagoltsággal rendelkezik, amely a mintavétel során nem kerül expliciten kezelésre. Értelemszerűen ezáltal sérül a mintaelemek függetlenségére vonatkozó feltétel. Tipikus példa lehet még az a helyzet, amikor az adatgyűjtés szűk idősávban történik (pl. nappali órákban, hétköznap), ezért bizonyos csoportok rendszerszinten kimaradnak; közterületi kérdezésnél a helyszínek és az időpontok „ráhúzódnak" sajátos alcsoportokra; paneles/ismételt mérésben az elérhetőség és a kérdezői háló szerkezete rétegződést idéz elő. A továbbiakban a rétegzettség kifejezést nem mint rétegzett mintavételi eljárást, hanem mint a populáció belső heterogenitására utaló szerkezeti jellemzőt használjuk, amely a mintavétel módjától függetlenül befolyásolhatja a becslések tulajdonságait.

A tanulmány a következő szerkezetet követi: az elméleti részben (1. fejezet) elsőként a releváns matematikai statisztikai alapokat mutatjuk be az 1.1. fejezetben, ezután ismertetjük a korrekciós javaslatainkat, külön kitérve a homogén és a rétegzett populációkból választott mintákra (1.2. fejezet). A torzításokat és a korrekciós javaslatokat Monte Carlo-szimulációkkal is illusztráljuk a 2. fejezetben, külön a homogén (2.1. fejezet) és külön a rétegzett populációk esetére (2.2. fejezet). A tanulmányt a következtetések megfogalmazásával zárjuk a 3. fejezetben. A cikkben bemutatott eredmények részletesen a szerzők *Biases arising from sampling and their adjustments* című tanulmányában olvashatók, amelyben a korrekciós javaslatokhoz tartozó tételek bizonyítással együtt megtalálhatók *(Nagy-György–Szeitl, 2026)*.

1. Elméleti áttekintés

1.1. Matematikai statisztikai alapok

Ebben a részben áttekintjük a számunkra releváns matematikai statisztikai alapokat, kiegészítve néhány nem független mintára vonatkozó észrevétellel. A terminológia megválasztásánál tudatosan a matematikai statisztika nyelvhasználatát követjük, amely az elemzések elméleti tulajdonságainak vizsgálatához biztosít egységes keretet. A szakirodalomban többféle fogalom- és jelölésrendszer létezik, jelen tanulmányban ezek használatánál *Bolla és Krámli (2012)* meghatározásait alkalmazzuk.

A matematikai statisztikában a becslések és a hipotézisvizsgálatok elméleti tulajdonságait matematikai tételek írják le, amelyek meghatározott feltételek mellett érvényesek.[1] Amennyiben ezek a feltételek teljesülnek, a tételek klasszikus formájukban alkalmazhatók, a feltételek sérülése esetén pedig külön vizsgálni kell, hogy léteznek-e olyan módosított vagy közelítő eredmények, amelyek az adott helyzetben érvényesek. Például nagy minták esetében határeloszlás-tételek biztosítják, hogy számos statisztikai eljárás közelítő eloszlási eredményei akkor is alkalmazhatók, ha az eredeti háttéreloszlásra vonatkozó feltételek nem teljesülnek, ezek azonban matematikailag eltérő állításoknak tekintendők. Az egyik ilyen feltétel a mintaelemek függetlenségére vonatkozik. A következőkben azt mutatjuk be, hogy mi történik, amikor ez a feltétel a gyakorlatban nem teljesül.

A matematikai statisztikában a minta n darab független, azonos eloszlású változó. A mintaelemek közös eloszlása a háttéreloszlás, ennek a paramétereire szeretne becslést, hipotézisvizsgálatot adni a statisztika. A háttéreloszlás paramétereire mint elméleti értékekre fogunk hivatkozni. Ezek közül is kiemelt figyelmet kap a várható érték (populációból vett minta esetén ez a populációátlag), amelyet μ-vel jelölünk, és a variancia, más néven szórásnégyzet, ezt itt σ^2 jelöli.

Ha egy N elemű rögzített populációból vesszük a mintát, akkor visszatevéses mintavétel (azaz a populáció egy tagja újra kiválasztásra kerülhet) esetén a mintaelemek függetlenek lesznek. Ha visszatevés nélkül vesszük a mintát, akkor a mintaelemek nem függetlenek, ez esetben egyszerű számolással megadható az X_k és az X_l mintaelemek kovarianciája *(Nagy-György–Szeitl, 2026)*:

$$Cov(X_k, X_l) = -\frac{\sigma^2}{N-1}$$

[1] Megjegyezzük, hogy ezen feltételek pontos megértéséhez valószínűségszámítási ismeretek szükségesek.

Tehát visszatevés nélküli mintavétel esetében negatív korreláció van a mintaelemek között, amely abszolútértéke a populáció méretének növelésével csökken.

A populációátlagot (μ) a mintaátlaggal, a populáció varianciáját (σ^2) a korrigált empirikus varianciával szokás becsülni, utóbbi az alábbi formulával adható meg:

$$S_n^{*2} = \frac{n}{n-1} \cdot S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \underline{X})^2$$

Azért szokás korrigálni az S_n^2 empirikus varianciát, mert független minta esetén így kapunk torzítatlan becslést a σ^2 varianciára. Akkor nevezünk torzítatlannak egy becslést, ha a várható értéke éppen a becsült paraméter (a várható érték populációból vett minta esetén az összes lehetséges mintavétel esetén kiszámolt becslés átlaga, ha az egyes minták azonos valószínűséggel kerülhetnek kiválasztásra). A becslésektől a torzítatlanság mellett különböző konzisztenciatulajdonságokat is szokás elvárni. A statisztikai szakirodalomban több konzisztenciafogalom létezik, ezek közül leggyakrabban a gyenge és az erős konzisztenciát különböztetik meg. Az erős konzisztencia azt jelenti, hogy ha a mintaelemszám tart a végtelenbe, akkor a becslések sorozata 1 valószínűséggel a becsült paraméterhez tart, míg gyenge konzisztencia esetén a becslések a mintaelemszám növekedésével egyre nagyobb valószínűséggel kerülnek a becsült paraméter tetszőlegesen kis környezetébe (részletesebben lásd például *van der Vaart, 1998, 2. fejezet*). Független minta (visszatevéses mintavétel) esetén a nagy számok erős törvénye garantálja, hogy a fenti becslések erősen konzisztensek. Visszatevés nélküli mintavétel esetén a mintaelemszámot a populáció mérete felülről korlátozza, ezért klasszikus értelemben vett konzisztenciáról nem beszélhetünk; ilyen esetben azt vizsgáljuk, hogy a becslések hogyan viselkednek, amikor a mintaelemszám eléri a populációméretet.

A következőkben a populációátlag és a populációs variancia becslését vizsgáljuk meg a torzítatlanság és a konzisztencia szempontjából visszatevéses és nem visszatevéses mintavételek esetében. Elsőként homogén populációt feltételezünk (amikor a populációból nem rétegképzés után veszünk mintát), majd rétegzett populációval foglalkozunk.

1.2. Torzítások és korrekciós javaslatok

Első megállapításunk az, hogy homogén populációból vett független minta esetén a μ elméleti várható értékre akkor tudunk torzítatlan becslést adni, ha a populáció egyedeinek mintába kerülési valószínűsége ismert. Specifikusan, ha a kiválasztási valószínűségek azonosak, akkor a mintaátlag torzítatlan becslés lesz. Egyéb esetben korrekcióra van szükség: az ismert kiválasztási valószínűségek lehetővé teszik

a megfelelő súlyozást, amely még akkor is biztosítja a torzítatlanságot, ha a populáció egyedei eltérő eséllyel kerülnek a mintába. Megjegyezzük, hogy az ismert kiválasztási valószínűségek gyakorlati biztosítása mintavételi keret meglétét feltételezi. Ennek jól ismert bizonyítását megnézve láthatjuk, hogy a mintaelemek függetlensége egyik lépésben sem szükséges, ezért nem független minta esetén is teljesül, hogy a mintaátlag várható értéke. Ez azt jelenti, hogy a mintaátlag a visszatevés nélküli mintavétel esetén is torzítatlan becslést ad a populációátlagra. Megjegyezzük, hogy amennyiben nem uniform módon történik a mintavétel, azaz különböző valószínűséggel kerülnek be a populáció tagjai a mintába, akkor a mintaátlag nem lesz torzítatlan becslése a populációátlagnak. Ismert, hogy a korrigált empirikus kovariancia független minta esetén torzítatlan becslése a kovarianciának. Ennek bizonyítását megvizsgálva megkaphatjuk a torzítást nem független minta esetén. A korrigált empirikus variancia várható értéke a következő formulával adható meg:

$$E(S_n^{*2}) = \sigma^2 - Cov(X_1, X_2)$$

Ha a mintaelemek függetlenek, akkor kovarianciájuk 0, és azt kapjuk, hogy a korrigált empirikus variancia torzítatlan becslése σ^2-nek. Második megállapításunk azzal kapcsolatos, hogy mi történik, ha visszatevéses a minta. Ebben az esetben a mintaelemek kovarianciájának -1-szerese megjelenik torzításként, ezt behelyettesítve kapjuk, hogy visszatevés nélküli mintavétel esetén

$$E(S_n^{*2}) = \frac{N}{N-1} \cdot \sigma^2. \quad (1)$$

Ez a torzítás nem függ a minta nagyságától, tehát nem változik a minta méretének növelésével. Matematikailag ez azt jelenti, hogy aszimptotikusan sem torzítatlan a becslés.

Ha a populáció nagy, ez elhanyagolható, de kis populáció esetén érdemes korrigálni. Ez a korrekció az (1) formula alapján az

$$\frac{N-1}{N}$$

korrekciós tényező *(Nagy-György–Szeitl, 2026)*. Megjegyezzük, hogy a korrigált empirikus szórás független minta esetén sem torzítatlan becslés a szórásra, mivel a gyökfüggvény nem lineáris. Érdemes foglalkozni az átlag standard hibájával is, amellyel kapcsolatos a harmadik megállapításunk. Mivel független minta esetén is csak a varianciára tudunk torzítatlanságot igazolni, a standard hiba négyzetét írjuk fel. Független minta esetén ismert, hogy S_n^{*2}/n torzítatlan becslés a mintaátlag varianciájára:

$$D^2(\underline{X}) = \frac{n\sigma^2}{n^2} = E\left(\frac{S_n^{*2}}{n}\right)$$

Visszatevés nélküli minta esetén ez nem teljesül:

$$D^2(\underline{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}, \quad viszont \quad E\left(\frac{S_n^{*2}}{n}\right) = \frac{N}{N-1} \cdot \frac{\sigma^2}{n},$$

amely felülbecsüli a standard hibát, és meglepő módon a torzítás aránya az elméleti értékhez viszonyítva nő a mintaméret növelésével. Ha torzítatlan becslést szeretnénk a standard hiba négyzetére visszatevés nélküli mintavétel esetén, akkor pedig a(z) S_n^{*2}/n becslés esetében a korrekció a(z)

$$\frac{N-n}{N}$$

korrekciós tényező *(Nagy-György–Szeitl, 2026)*.

A konzisztencia szempontjából visszatevés nélküli mintavétel esetén nem tud a mintaelemszám végtelenbe tartani, hiszen a populációméret felső korlátot ad rá, így klasszikus értelemben nem beszélhetünk konzisztenciáról, de érdemes meggondolni, hogy mi történik, ha a mintaelemszámot addig növeljük, amíg lehet. A mintaátlagok sorozata a végén eléri a populációátlagot, a korrigált empirikus kovarianciák sorozatának utolsó tagja viszont

$$S_N^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2 = \frac{N}{N-1} \cdot \sigma^2,$$

amely nem a populáció varianciája, tehát konzisztencia szempontjából is szerencsés lenne a fent említett korrekció. Ez főleg kis populáció esetében lényeges. A standard hiba szokásos becslése erősen konzisztens független minta esetén a korrigált empirikus variancia erős konzisztenciája miatt, visszatevés nélküli mintavétel esetén viszont

$$\frac{S_N^*}{\sqrt{N}} = \frac{\sigma}{\sqrt{N-1}},$$

miközben az átlag szórása 0, ha a teljes populáció bekerül a mintába.

A továbbiakban rétegzett populációkkal foglalkozunk. Megfelelő eljárás esetén a mintaelemek egyforma valószínűséggel történő bekerülése biztosítható, de a teljes mintában a mintaelemek függetlensége nem. Megjegyezzük, hogy ha egy rétegből veszünk egy (rész)mintát, akkor a paraméterek becslésére a fentiek az adott rétegre és (rész)mintára vonatkoztatva érvényesek.

Tekintsünk egy rétegzett populációt r réteggel: az i-edik réteg átlaga μ_i, varianciája σ_i^2. Egyszerű számolással adódik, hogy a populáció elméleti értékei a következőképpen adhatók meg:

$$\mu = \frac{1}{N}\sum_{i=1}^{r} N_i \cdot \mu_i$$

és

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{r} N_i \cdot (\sigma_i^2 + (\mu_i - \mu)^2).$$

Most nézzük meg a rétegzett mintavétel egyik legegyszerűbb esetét: véletlenszerűen választunk egy R réteget úgy, hogy az i-edik réteg kiválasztásának valószínűsége p_i, majd a kiválasztott rétegből veszünk egy n elemű mintát. Azáltal, hogy egyetlen rétegből vesszük a teljes mintát, a mintaelemek függetlensége sérül. Ekkor

$$E(\underline{X}) = \sum_{i=1}^{r} \mu_i \cdot p_i.$$

A következő megállapításunk az, hogy ha $p_i = N_i/N$ minden i-re, akkor az átlag torzítatlan becslés, egyébként torzítás jelenhet meg, amelyet korrigálni kell, az i-edik réteg kiválasztása esetén $N_i/(Np_i)$ tényezővel. Megjegyezzük, hogy a torzítatlanság önmagában kevés ahhoz, hogy jó becslést kapjunk, hiszen ebben a modellben egy mintavétel átlaga valamelyik μ_i érték körül mozog, ha növeljük a mintaelemszámot, akkor ehhez egyre közelebbi értéket vesz fel. Könnyű meggondolni, hogy az átlag még gyengén sem lesz konzisztens becslése μ-nek, ha nem teljesül, hogy $\mu_1 = \cdots = \mu_r = \mu$.

Ha a mintát visszatevéssel vesszük,

$$E(S_n^{*2}) = \sum_{i=1}^{r} \sigma_i^2 \cdot p_i,$$

ha visszatevés nélkül, akkor

$$E(S_n^{*2}) = \sum_{i=1}^{r} \sigma_i^2 \cdot \frac{N_i}{N_i - 1} \cdot p_i,$$

azaz a korrigált empirikus variancia egyik esetben sem torzítatlan becslése σ^2-nek akkor sem, ha az i-edik réteget a méretével (vagy $N_i - 1$-gyel) arányos valószínűséggel választjuk. Ezt a becslést pedig akkor tudjuk korrigálni, ha (1) ismerjük vagy becsülni tudjuk a rétegek átlagait (μ_R); vagy (2) a populációátlag és a választott réteg átlagának különbségét ($\mu_R - \mu$). Utóbbihoz több rétegből vett mintákra van szükség.

Ha nem egy, hanem több, de nem az összes rétegből vesszük a mintát, hasonló problémák lépnek fel, viszont nehezebb kiszámolni a torzításokat. Ezzel a továbbiakban itt nem foglalkozunk, a több rétegből vett mintákkal kapcsolatos eredményeinket egy másik munkánkban *(Nagy-György–Szeitl, 2026)* tervezzük közölni.

2. Szimuláció

A következőkben bemutatott szimulációk célja nem tipikus társadalomtudományi mintavételi helyzetek empirikus modellezése, hanem bizonyos – a mintavételből fakadó – torzítások mechanizmusának szemléltetése olyan paraméterbeállítások mellett, ahol ezek az eltérések jól láthatók. Ebben a fejezetben a fentieket Monte Carlo-szimulációkkal is szemléltetjük, egy olyan módszerrel, amely bonyolult problémák közelítő megoldását keresi véletlen mintavételezés segítségével: sok lehetséges kimenetelt szimulálunk, majd ezek átlagából becsüljük meg a keresett mennyiséget. Matematikailag ez azon alapul, hogy egy valószínűségi változó várható értéke jól közelíthető nagyszámú független mintából számított mintátlaggal (lásd nagy számok törvényei). Itt sokszor szimulálunk egy mintavételt, és az egyes mintavételekből számolt becslések viselkedését vizsgáljuk. Elsőként homogén populációkat feltételezünk (2.1. fejezet), majd rétegzett populációkkal dolgozunk (2.2. fejezet).

2.1. Mintavétel homogén populációból

A vizsgált becslések közül a korrigált empirikus variancia és a standard hiba esetében is fontos tényező a populáció mérete. Ezért több, eltérő méretű populáción is bemutatjuk, hogy a különböző mintavételi módok milyen hatást gyakorolnak a becslésekre. Három kiinduló populációt használunk, amelyek $N = 20$, $N = 50$ és $N = 100$ elemből állnak. A kis elemszámú populációk választása tudatos: a cél, hogy a visszatevéses és a visszatevés nélküli mintavétel közötti különbségek, valamint a függetlenség sérüléséből fakadó torzítások hatása már kis minták esetén is jól elkülöníthetően jelenjen meg. Megjegyezzük, hogy kis elemszámú populációk valós kutatási helyzetekben is előfordulhatnak, például ritka betegségek vizsgálatánál, biológiai kutatásokban, illetve zárt vagy nehezen elérhető csoportok esetében. Mindhárom populációnál a tagjaik értékeit egy $\mu = 10$ és $\sigma = 5$ paraméterű normális eloszlásból generáltuk. A populációs értékek megoszlásai a Függelék F1–F3. ábráin láthatók. A generált populációkban a populációátlag rendre $10{,}71$, $10{,}17$ és $10{,}45$, a populációs varianciák pedig rendre $22{,}47$, $21{,}00$ és $20{,}62$ lettek. Ezek azok az elméleti értékek, amelyeket a következőkben a különböző mintavételek alapján becslünk.

A mintavételek során $n = 10$, $n = 40$ és $n = 50$ elemű mintákat veszünk. Mindegyik esetében elvégezzük a mintavételezést visszatevéses és visszatevés nélküli módon is. A Monte Carlo-ismétlések száma mindegyik mintavétel esetében egyenként $10\,000\,000$.

A becslésektől gyakran elvárt tulajdonság a torzítatlanság, amely azt jelenti, hogy a becslő statisztika várható értéke éppen a becsült érték. Ez a gyakorlatban azt jelenti, hogy ha nagyon sokszor végezzük el a vizsgálatot, akkor átlagosan körülbelül a becsült értéket kapjuk (a nagy számok törvényei matematikailag pontosan megfogalmazzák ezt). A torzítatlanság azonban nem az egyetlen becslési kritérium: A szakirodalomban a variancia, a konzisztencia és az átlagos négyzetes hiba is fontos szerepet játszik. A gyakorlatban nem feltétlenül tudjuk ezek mindegyikét biztosítani, ezért a becslések megválasztása gyakran kompromisszumot jelent a különböző kritériumok között. Az alábbiakban néhány elméleti értéket és ezeknek a kétféle minta alapján kapott becsléseinek várható értékének Monte Carlo-becsléseit vetjük össze. Először az elméleti várható érték (populációátlag) becsléseit nézzük meg. Ez mindkét mintavétel esetében az elvárásoknak megfelelően viselkedik. A torzítatlanságot a várható érték linearitás nevű tulajdonsága biztosítja, amihez nem kell a mintaelemek függetlensége, ezt a szimuláció is alátámasztja (1. táblázat). A konzisztencia is teljesülni fog (hétköznapi nyelven fogalmazva a mintaelemszám növelésével egyre pontosabb becslést kapunk).

1. táblázat

A mintaátlag várható értékének Monte Carlo-becslései
Monte Carlo estimates of the expected value of the sample mean

N: populációméret, n: mintaelemszám	Elméleti érték	Visszatevéses becslés	Visszatevés nélküli becslés
$N = 20$, $n = 10$	10,708	10,709	10,709
$N = 50$, $n = 10$	10,172	10,172	10,172
$N = 50$, $n = 40$	10,172	10,172	10,172
$N = 100$, $n = 10$	10,452	10,452	10,452
$N = 100$, $n = 50$	10,452	10,452	10,452

Forrás: saját szerkesztés.

A varianciánál már más a helyzet, hiszen elméleti úton igazoltuk, hogy nem független minta (azaz a visszatevés nélküli mintavétel) esetén a torzítatlanság nem is teljesül. A torzítás két mintaelem kovarianciája, amely a populációtól függ. A fenti populációk esetén a populációvariancia mellett a korrigált empirikus variancia várható értékének Monte Carlo-becsléseit a 2. táblázat tartalmazza. A táblázatba beillesztettük az általunk javasolt becslést is, amely az ismertetett korrekciós tagot használja. Ha nagyon nagy a populáció, a torzítás lényegében elhanyagolható. Megjegyezzük, hogy rétegzett mintavétel esetén a fentiek a rétegre vonatkoznak, tehát a réteg varianciájának becslése a réteg függvénye.

2. táblázat

A varianciabecslések várható értékének Monte Carlo-becslései
Monte Carlo estimates of the expected value of variance estimators

N: populációméret, *n*: mintaelemszám	Elméleti érték	Visszatevéses korrigált empirikus variancia	Visszatevés nélküli	
			korrigált empirikus variancia	javasolt becslés
$N = 20$, $n = 10$	22,469	22,476	23,655	22,472
$N = 50$, $n = 10$	21,002	21,004	21,436	21,007
$N = 50$, $n = 40$	21,002	21,005	21,430	21,001
$N = 100$, $n = 10$	20,622	20,619	20,834	20,625
$N = 100$, $n = 50$	20,622	20,624	20,831	20,622

Forrás: saját szerkesztés.

A standard hiba értékének, a szokásos becslése várható értékének, valamint az általunk javasolt korrigált becslés várható értékének Monte Carlo-becsléseit a 3. táblázat tartalmazza.

3. táblázat

A standard hiba értékének és a becslések várható értékének Monte Carlo becslései
Monte Carlo estimates of the standard error and the expected values of the estimators

N: populációméret, *n*: mintaelemszám	Visszatevéses		Visszatevés nélküli		
	elméleti érték	becslés várható értéke	elméleti érték	becslés várható értéke	javasolt becslés várható értéke
$N = 20$, $n = 10$	2,248	2,248	1,183	2,365	1,183
$N = 50$, $n = 10$	2,101	2,101	1,714	2,382	1,715
$N = 50$, $n = 40$	0,525	0,525	0,107	0,536	0,107
$N = 100$, $n = 10$	2,063	2,062	1,874	2,315	1,874
$N = 100$, $n = 50$	0,412	0,413	0,208	0,425	0,208

Forrás: saját szerkesztés.

A mintaátlag véletlen változó, mivel véletlen mintavétel történt. Az alábbi ábrák a mintaátlag sűrűségfüggvényének Monte Carlo-becslését mutatják kék színnel, a ráillesztett piros görbe a mintaátlag *Lukács Jenő* tétele *(Bolla–Krámli, 2012, 51. old.)* alapján feltételezett eloszlása, amely normális, paraméterei az elméleti értékek, tehát a populációátlagok és a populációs varianciák. Ha nem normális eloszlású a populáció, akkor a centrális határeloszlástétel állítása szerint a mintaátlag közel normális eloszlású, várható értéke a populációátlag, varianciája a populáció varianciájának n-edrésze független mintaelemek esetén, ahol n a mintaelemszám. Megjegyezzük, hogy az elméleti értékek a gyakorlatban nem

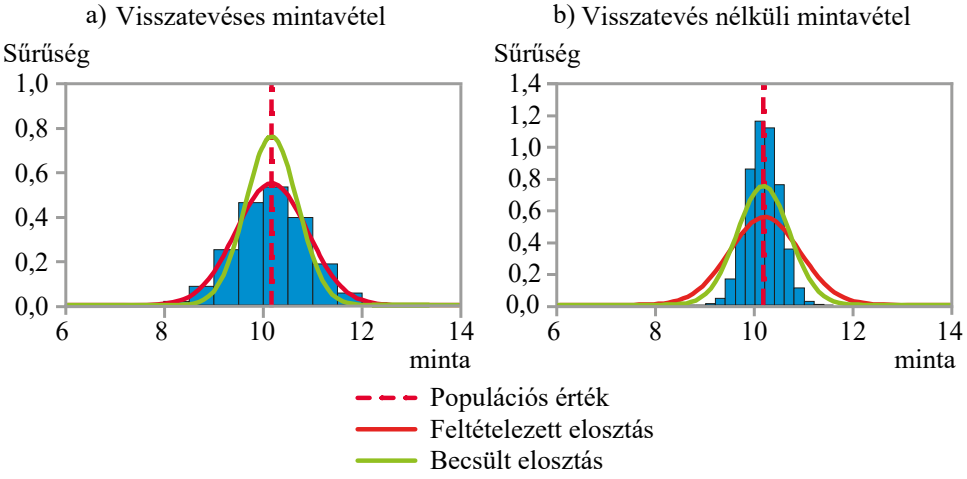
ismertek, tehát a mintaátlag eloszlásának ezen becslése nem adható meg, ezért egy másik, zöld színű görbét is illesztettünk minden ábrára, ahol a feltételezett eloszlás paraméterei a mintából becsült értékek.

A következő ábrák annak szemléltetései, hogy a mintaátlag valódi eloszlása (amelynek Monte Carlo-becslése kékkel látható) mennyire tér el a feltételezett (piros) eloszlástól, illetve annak becslésétől (zöld görbe) abban az esetben, ha a statisztikai mintára vonatkozó függetlenségfeltétel sérül a visszatevés nélküli mintavétel miatt. Az eltérés a visszatevéses és a nem visszatevéses mintavételek között minden ábrán jelen van, de abban az esetben látványosabb, amikor a mintaelemszám nagyobb. Ennek az az oka, hogy minél kisebb a populáció és minél nagyobb a minta, annál nagyobb súllyal jelenik meg a mintaelemek kovarianciája az átlag varianciájában (azaz a függetlenség sérülése esetén jobban torzul a becslés). A legnagyobb különbség abban az esetben látható a visszatevéses és nem visszatevéses mintavételekből származó becslések között, amikor az $N = 50$ elemű populációból $n = 40$ méretű mintákat veszünk (1. ábra).

1. ábra

A várható érték sűrűségének Monte Carlo-becslése 10 000 000 megismételt $n = 40$ elemű egyszerű véletlen minták (EVM) alapján az $N = 50$ elemű populációból

Monte Carlo estimate of the density of the sample mean (N = 50, n = 40)



Forrás: saját szerkesztés.

2.2. Egy módszertani ellenpélda: torzítás nem megfelelő mintavételi eljárás esetén

Ebben a fejezetben az olyan típusú mintavételek eredményeit mutatjuk be, amikor a populáció rétegzett. Az alábbiakban ismertetett helyzet nem tekinthető sem megfelelő, sem ajánlott mintavételi eljárásnak, célja kizárólag egy módszertani ellenpélda bemutatása, amely szemlélteti a mintavétel hiányosságaiból fakadó torzítások következményeit. A következőkben a populáció $r = 2$ rétegből áll. A populáció teljes mérete $N = 100$ fő, a rétegek mérete azonos, így $N_1 = N_2 = 50$. A populáció két rétegébe tartozó elemeket eltérő várható értékű ($-10 \leq \mu_1, \mu_2 \leq 10$), de azonos szórású ($\sigma_1 = \sigma_2 = 5$) normális eloszlásokból generáltuk. A populációs értékek megoszlásai a Függelék F4. ábráján találhatók. A generált populációban a rétegeken belüli populációátlagok rendre $\mu_1 = -1{,}65$ és $\mu_2 = 8{,}51$, a varianciák pedig rendre $\sigma_1 = 4{,}63$ és $\sigma_2 = 4{,}49$ lettek. A mintavételek során először egy R réteget választunk ki véletlenszerűen. A kiválasztás uniform, azaz mindkét réteg kiválasztási valószínűsége azonos. Ezután a kiválasztott rétegből $n = 10$ elemű mintákat veszünk, és mindegyik esetében elvégezzük a mintavételt visszatevéses és visszatevés nélküli módon is. Az ismétlések száma itt is egyenként 10 000 000, minden ismétlésnél újraválasztjuk a réteget.

Rétegzett mintavétel esetén a rétegek paramétereinek becslései során az előző fejezetben bemutatott jelenségek adódhatnak, ezen felül a populáció becsléseivel kapcsolatos további problémák is előfordulhatnak. Ebben a részben ez utóbbiakra fókuszálunk. A 4. táblázatban a várható érték (populációátlag) és becslései (mintaátlagok) várható értékének Monte Carlo-becslései találhatók. Ha megfelelő valószínűséggel választjuk ki a réteget, a becslés mindkét esetben torzítatlan, de könnyen belátható, hogy mivel nem teljesül, hogy $\mu_1 = \cdots = \mu_r = \mu$, az átlag még gyengén sem lesz konzisztens becslése μ-nek egyik esetben sem *(Nagy-György–Szeitl, 2026)*.

4. táblázat

A mintaátlagok várható értékének Monte Carlo-becslései

Monte Carlo estimates of the expected values of the sample means

N: populációméret, *n*: mintaelemszám, *r*: rétegek száma	Elméleti érték	Visszatevéses	Visszatevés nélküli
N = 100, *n* = 10, *r* = 2	3,38	3,429	3,432

Forrás: saját szerkesztés.

Az 5. táblázatban a populáció varianciája, valamint a korrigált empirikus variancia várható értékeinek Monte Carlo-becslései láthatók a két esetben. A szimulációs eredmények itt is alátámasztják az elméleti eredményeinket, azaz egyik esetben sem kapunk jó becslést a populáció varianciájára.

5. táblázat

Az empirikus variancia várható értékének Monte Carlo becslése

Monte Carlo estimates of the expected value of the empirical variance

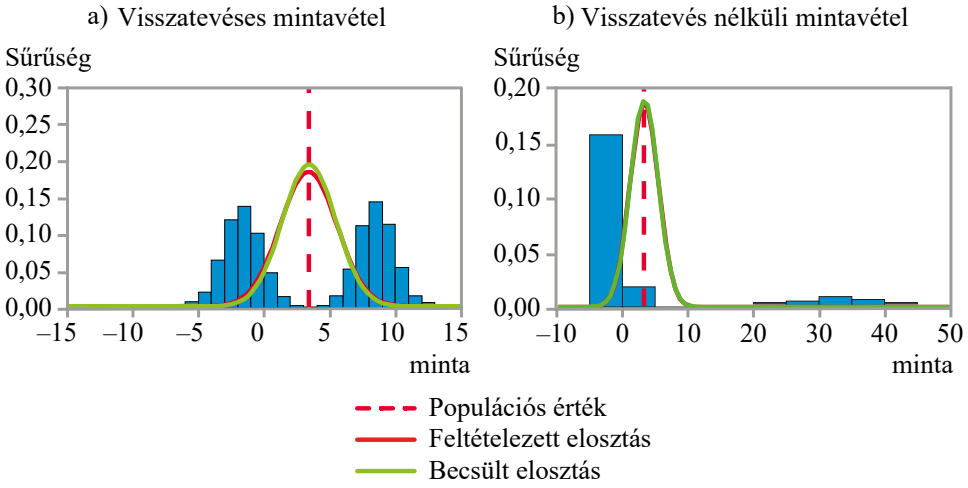
N: populációméret, n: mintaelemszám, r: rétegek száma	Elméleti érték	Visszatevéses	Visszatevés nélküli
$N = 100$, $n = 10$, $r = 2$	46,203	40,78	33,894

Forrás: saját szerkesztés.

2. ábra

Az átlagok sűrűségének Monte Carlo-becslései

Monte Carlo estimates of the density of the means of two strata using one stratum ($N_1 = 50$, $N_2 = 50$, $n = 10$)



Megjegyzés: 10 000 000 megismételt $n = 10$ elemű véletlen minták alapján az $N = 100$ elemű, $r = 2$ rétegből álló azonos szórású populációból ($N_1 = 50$, $N_2 = 50$), amely rétegek közül $k = 1$ került kiválasztásra
Forrás: saját szerkesztés.

A 2. ábrán látható eredmények annak szemléltetései, hogy a mintaátlag valódi eloszlása (amelynek Monte Carlo-becslése kék színnel jelölve) mennyire tér el a feltételezett (piros) eloszlástól, illetve annak becslésétől (zöld görbe) abban az esetben, ha a statisztikai mintára vonatkozó függetlenségfeltétel sérül a fent leírt mintavétel miatt. Az ábra alapján könnyen átlátható, hogy miért nem teljesül itt a konzisztencia, és hogy a mintaelemszám növelése csak akkor javítana a becslésen,

ha a rétegek várható értékei azonosak lennének, vagy legalábbis ismertek, ami alapján a korrekció megoldható lenne.

A bemutatott eredmények azt szemléltetik, hogy egy módszertanilag hibás mintavételi döntés milyen mértékű torzítást eredményezhet még egyszerű beállítások mellett is, ami nem is korrigálható.

3. Következtetések

Eredményeink gyakorlati üzenete az, hogy a mintavételezés módjára kiemelten érdemes figyelni, és a becslések értelmezésekor számításba kell venni a lehetséges torzításokat. Már a legegyszerűbb helyzetben (homogén populáció, egyszerű véletlen mintavétel visszatevés nélkül) is eltérnek a klasszikus, független mintára érvényes formulák. Ez számos esetben megjelenik a gyakorlatban, hiszen az adatfelvételek döntő többsége visszatevés nélküli mintavételt alkalmaz, és a populáció méretéhez képest nagy a minta aránya: (i) kis, zárt keretekben történő felmérések (vállalati dolgozói elégedettség, iskolai/osztályos vizsgálatok, kutatói hálózatok, kistelepülési lekérdezések), ahol n/N számottevő; (ii) választáskutatások, ahol egyes alminták (pl. kisebb megyék, fiatal férfiak alacsony válaszadási rátával) felül- vagy alulmintavétele miatt a valós kiválasztási valószínűségek és a rétegkülönbségek súlyos torzítást okozhatnak; (iii) piackutatási szegmensfelmérések, ahol szándékosan túlreprezentálunk egyes fogyasztói csoportokat; (iv) paneles/ismételt mérések, ahol a tényleges mintavétel a mintavételi keret véges voltából és az egységek ismételt megkereséséből adódóan sérti a függetlenséget; (v) terepi adatfelvételekben a címlisták csoportosulása és az interjúzói munkaszervezés földrajzilag szintén csoportosuló gyakorlata miatt a függetlenség megsértése elkerülhetetlen, és megjelenik a *design effect*[2], még akkor is, ha a tervezés elvben az egyszerű véletlen mintához közeli.

Rétegzett populáció esetében a torzítás kockázata jelentősen nő: ha a teljes mintát egyetlen rétegből nyerjük, az átlag még gyengén sem konzisztens a teljes populációs átlagra (kivéve azonos rétegátlagok esetén). Ilyenkor a kiválasztási valószínűségek figyelmen kívül hagyása a becslések torzításához vezet. A surveykutatások esetében leggyakrabban szándékosan nem veszünk mintát minden rétegből, és ezt a becslések korrekciójánál és az eredmények értelmezésénél figyelembe tudjuk venni. Fontos, hogy nem szándékosan hagyunk figyelmen kívül rétegeket a

[2] A *design effect* azt fejezi ki, hogy egy adott mintavételi eljárás mennyivel növeli vagy csökkenti egy becslés szórását ahhoz képest, mint ha azonos elemszámú, egyszerű véletlen mintát vettünk volna.

populáció nem megfelelő ismerete esetén vagy az adatgyűjtés pontatlan tervezése miatt, a kutató szempontjából rejtett rétegek lehetnek jelen. Ezek a helyzetek látszólag egyszerű mintavétel mellett is torzítást okozhatnak. Ennek mérsékléséhez célszerű előre azonosítani a potenciális rétegeket és a mintavételi tervbe beépíteni a megfelelő arányokat.

Eredményeink egyszerű mintavételi módszerekből származnak, a jövőbeli kutatásoknak érdemes kiterjeszteniük a vizsgálatot egyéb, a rétegzett eljárások torzításaira és a mintaátlagon, illetve a korrigált empirikus variancián túl más statisztikákra. Cikkünkkel ennek a vizsgálódási iránynak szándékoztunk kiindulási pontot adni.

Köszönetnyilvánítás

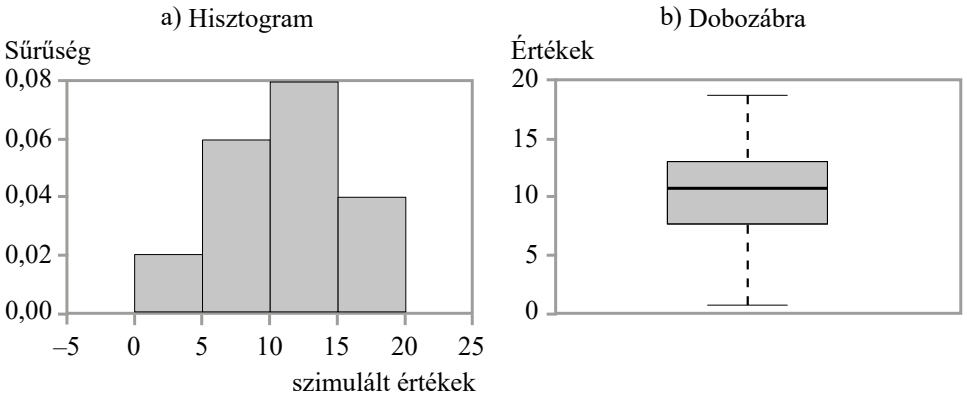
Ezt a tanulmányt a TKP2021-NVA-09 projekt támogatta. A TKP2021-NVA-09 projekt a TKP2021-NVA finanszírozási program keretében, a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból, a Magyar Kulturális és Innovációs Minisztérium támogatásával valósult meg.

Függelék

F1. ábra

A populációs értékek megoszlása $N = 20$ méretű populáció esetében

Distribution of the population values (N = 20)

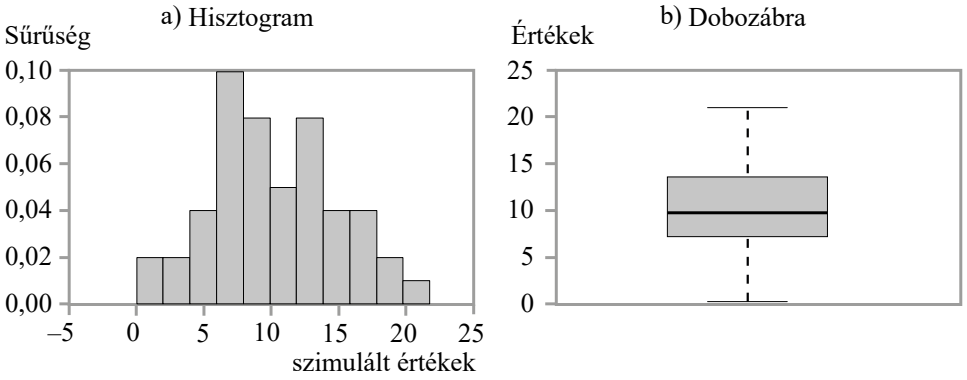


Forrás: saját szerkesztés.

F2. ábra

A populációs értékek megoszlása $N = 50$ méretű populáció esetében

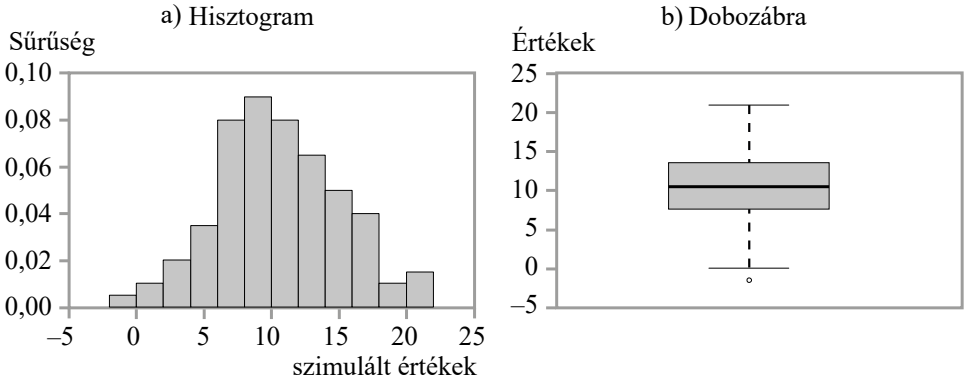
Distribution of the population values (N = 50)



Forrás: saját szerkesztés.

F3. ábra

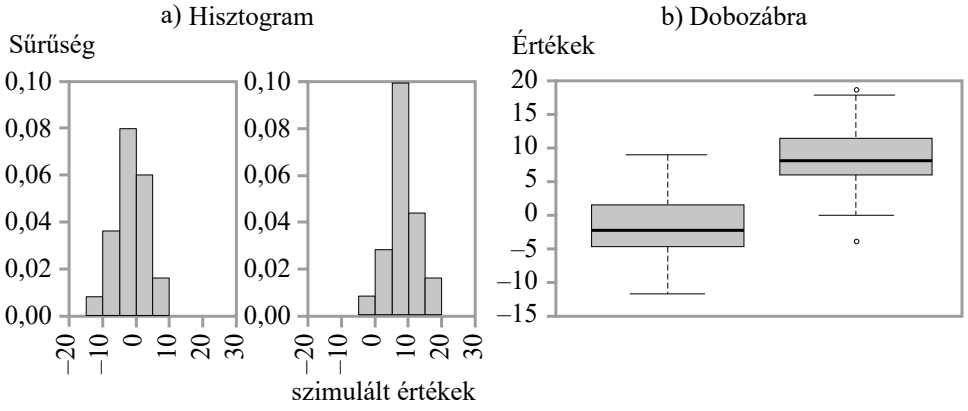
A populációs értékek megoszlása $N = 100$ méretű populáció esetében
Distribution of the population values (N = 100)



Forrás: saját szerkesztés.

F4. ábra

A populációs értékek megoszlása $N = 100$ és $r = 2$ azonos szórású populációk esetében
Distribution of the population values (N = 100) with r = 2 strata of equal variance



Irodalom

Babbie, E. (2008): *A társadalomkutatás módszertana.* Balassi Kiadó, Budapest.

Bolla M. – Krámli A. (2012): *Statisztikai következtetések elmélete*. Typotex Kiadó, Budapest.

Galambosné Tiszberger M. (2011): A rétegzett mintavételről. *Statisztikai Szemle,* 89(9), 909–929. https://www.ksh.hu/statszemle_archive/all/2011/2011_09/2011_09_909.pdf

Kapitány B. (2010): Mintavételi módszerek ritka populációk esetén. *Statisztikai Szemle,* 88(7-8), 739–754. https://www.ksh.hu/statszemle_archive/all/2010/2010_07-08/2010_07-08_739.pdf

Kish, L. (1965): *Survey Sampling.* John Wiley & Sons, New York.

Lohr, S. L. (2010): *Sampling: Design and Analysis.* Brooks/Cole, Boston.

Nagy-György, J. – Szeitl, B. (2026): *Biases arising from sampling and their adjustments. Manuscript.* https://www.math.u-szeged.hu/~ngyj/sampling/sample_bias.pdf

Neyman, J. (1934): On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society,* 97(4), 558–625. https://doi.org/10.2307/2342192

Simon D. (2012): Válaszadó-vezérelt mintavétel: ritka és rejtett csoportok kvantitatív vizsgálata. *Statisztikai Szemle,* 90(4), 249–275. https://www.ksh.hu/statszemle_archive/all/2012/2012_04/2012_04_249.pdf

Särndal, C-E. – Swensson, B. – Wretman, J. (1992): *Model Assisted Survey Sampling.* Springer, New York.

Van der Vaart, A. W. (1998): *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Varga J. Z. (2021): A mintavételezés hatása a pénzügyi adatok statisztikai tulajdonságaira és alkalmazása a kockázatkezelésben. *Statisztikai Szemle,* 99 (3), 233–252. https://doi.org/10.20311/stat2021.3.hu0233