A SOKASÁGI ARÁNY MEGHATÁROZÁSÁRA IRÁNYULÓ STATISZTIKAI ELJÁRÁSOK VÉGES SOKASÁG ÉS KIS MINTÁK ESETÉN

LOLBERT TAMÁS[1]

A cikk fő célja annak vizsgálata, hogy az ellenőrzési gyakorlatban széles körben használt és oktatott nagymintás becslőfüggvény (M1) milyen feltételekkel alkalmazható a sokasági arány meghatározására (attribute sampling) kis minták esetén.

A téma általános megközelítése céljából több becslési eljárást is megvizsgáltunk, mind a hagyományos mintavételi megközelítés, mind a bayesi megközelítés területéről. A mintavételi statisztika megközelítését alkalmazva megszületett egy új becslőfüggvény (M3), ezen kívül megalkotásra került két konkrét bayesi intervallumbecslés, valamint egy vegyes szemléletű becslés is. Az egyes becslőfüggvények elemzése megmutatta, hogy a sokasági arány elvárt megbízhatóságú becsléséhez legalább 150-200 elemű minta szükséges a nagymintás (M1) becslőfüggvénnyel, ezzel szemben a többi becslés kielégítő módon működik kisebb mintákra is. Az informatív priort használó bayesi becslések nagyságrendekkel szűkebb intervallumot adnak, amennyiben helyes priort használunk.

A becslésekről szóló részt kiegészíti a hipotézisvizsgálat, továbbá az Állami Számvevőszéknél is használt IDEA nevű könyvvizsgálati szoftver megfelelő moduljainak rövid leírása.

TÁRGYSZÓ: Sokasági arány. Attribute sampling. Hipergeometriai eloszlás. Megbízhatósági szint.

A statisztikai eljárások egyre fontosabb szerepet töltenek be a pénzügyi-gazdasági ellenőrzés területén. Az ellenőrzés egyik legfontosabb funkciója az, hogy bizonyosságot szerezzen egy adott szervezet, szervezeti egység, vagy intézményrendszer megfelelő működésről, és gyakran felmerül az igény, hogy meghatározzuk a bizonyos szempontoknak (attribute, characteristic) megfelelő egyedek (például szabálytalan kifizetések vagy egyéb téves tranzakciók stb., amiket a későbbiekben *minősített egyed*eknek fogunk nevezni) sokasági arányát. Természetesen a legtöbb esetben az auditor eltűr egy minimális hibaarányt, de ha az átvilágítás során arra a következtetésre jut, hogy a valós hibaarány ezt meghaladja, akkor kénytelen elmarasztaló véleményt kiadni az adott szervezetről.

[1] A szerző köszönetet mond *dr. Hunyadi Lászlónak*, akinek javaslatára e tanulmány megszületett, és annak elkészítését kezdettől fogva felügyelte. Köszönet illeti *Kánnai Zoltánt*, alapos lektori munkájáért, továbbá az Állami Számvevőszék vezetőit és munkatársait, különösen *dr. Csapodi Pált* és *dr. Lóránt Zoltánt*, hogy lehetővé tették a cikk megírását, és tanácsaikkal javították annak minőségén.

Nyilvánvaló, hogy felesleges kapacitásokat vonna el az összes tranzakció tételes ellenőrzése, ezért bevett gyakorlat a mintavétel és a mintából való következtetés. Ezzel együtt azonban fontos elvárás, hogy ezek a következtetések megalapozottak (statisztikailag alátámaszthatóak) és összehasonlíthatóak legyenek, ezért szükséges, hogy a következtetéseknél feltüntessük a megbízhatósági szintet. A hazai és nemzetközi ellenőrzési gyakorlatban leginkább a 95 százalékos megbízhatósági szint használata honosodott meg.

A tanulmány az aránybecslési eljárásokat tekinti át. Alapozó statisztikai könyvekben többnyire a *visszatevéses* mintavétel esetére alkalmazható, normális eloszlással való közelítésen alapuló becslőfüggvényt mutatják be. Az ellenőrzési szakma azonban jellegéből adódóan kizárólag *visszatevés nélküli* mintavételi tervekkel foglalkozik, továbbá szintén az ellenőrzési szakma sajátosságai miatt szigorúan csak véges sokaságokat, és jellemzően kis (100 tételnél kisebb) egyedszámú mintákat vizsgálnak. Ezért indokolt megvizsgálnunk, milyen becslőfüggvények készíthetők a sokasági arányra ilyen keretek közt. Hangsúlyozni kell, hogy a sokaság végességéből következően a sokasági arány nem vehet fel bármilyen értéket; ennek ellenére az eljárásokat úgy vezetjük be, hogy figyelmen kívül hagyjuk az eloszlások, illetve a lehetséges értékek diszkrét jellegét, és a tárgyalás során a könnyebb érthetőség kedvéért több esetben folytonosnak feltételezzük őket.

A téma tárgyalása során hallgatólagosan többször ki fogunk használni egy olyan feltevést, ami az egész ellenőrzést szuperpopulációs kontextusba helyezi. Egy adott ellenőrzés során tipikusan *adott* szervezet(ek) *adott* időszakra vonatkozó tevékenységét vizsgáljuk, és a speciális esetektől eltekintve nincs okunk feltételezni, hogy a vizsgált időszak ügymenete érdemben eltért a nem vizsgált időszakok ügymenetétől, illetve – alternatív megközelítésben –, hogy a vizsgált szervezet ügymenete eltér a hozzá hasonló szervezetek ügymenetétől. Ebből következően a vizsgált sokaság maga is egy mintának tekinthető, konkrétan egy végtelen sokaságból vett meta-mintának, amelynek eloszlásáról a korábbi ellenőrzési tapasztalatok alapján már komoly előzetes ismeretekkel is rendelkezhetünk. Példaként említhetjük a számviteli szabálytalanságok előfordulási gyakoriságát, melynek nagysága több évre visszatekintve is stabil, rendszersajátosságokat tükröző paraméter.

1. BECSLÉS

Mielőtt rátérnénk a becslések részletes tárgyalására, vezessük be a következő jelöléseket:

Legyen N a sokaság egyedszáma, M a sokaság minősített egyedeinek száma, n és m pedig a minta hasonló értékei (adott esetben használni fogjuk a $P = \frac{M}{N}$ és a $p = \frac{m}{n}$ jelöléseket is). Legyen α a becslés során elkövethető elsőfajú (mintavételi) hiba rögzített valószínűsége (tehát $1-\alpha$ a megbízhatósági szint), tekintsük továbbá az N elemű sokaságból nyerhető, különböző kimenetelű minták $S:=\{S_i\}$ halmazát.[2] Első feladatunk a pontbecslés lesz, ami nem okoz különösebb nehézséget.

[2] Egy minta kimenetelén az n, m paramétereket értjük.

1.1. Pontbecslés

A sokasági arány pontbecslése első megközelítésben a megfelelő mintabeli értékkel ($\hat{P} = p = \frac{m}{n}$) történik. A mintabeli érték torzítatlanul becsüli a sokasági értéket:

$$E\left(\hat{P}\right) = E\left(\frac{m}{n}\right) = \frac{1}{n}E(m) = \frac{1}{n}\cdot n \cdot \frac{M}{N} = \frac{M}{N},$$

ugyanis m jelen feltételek mellett hipergeometriai eloszlást követ, és a hipergeometriai eloszlás várható értéke $\frac{nM}{N}$.

Alternatív megoldást kínál a pontbecslésre a bayesi megközelítés. Ehhez priorként szolgál, ha a sokasági hibaszámról feltesszük, hogy binomiális eloszlású $\left(N, P_0\right)$ paraméterekkel, ahol P_0 a korábbi ellenőrzési tapasztalatokból ismert, rendszersajátosságot tükröző stabil paraméter. A minta hibaszámáról tudjuk, hogy adott sokasági hibaarány mellett N, M, n paraméterű hipergeometriai eloszlást követ, tehát így már meghatározható $\frac{M}{N}$ (pontosabban M) posterior eloszlása is. A posterior eloszlás ismeretében többféle elven is (például a posterior eloszlás várható értéke, mediánja, módusza, stb.) készíthető pontbecslés, de ezek részletezését terjedelmi okok miatt mellőzzük. Céljainknak megfelel, ha becsült értékként a posterior eloszlás *móduszát* adjuk meg, tehát azt a sokasági hibaarányt, amely mellett legvalószínűbb a kapott minta előfordulása. Választásunkat egyrészt indokolja a módusz viszonylag könnyű meghatározhatósága, másrészt pedig vonzóvá teszi a maximum likelihood szerű interpretálhatóság is. Megjegyeznénk, hogy hátránya is van a módusz választásának, ugyanis az amúgy „jól viselkedő” (értsd: „haranggörbe-szerű”) eloszlásoknak diszkrét esetben két – amúgy szomszédos – módusza is lehet. Az egyértelműség kedvéért mi következetesen a nagyobb móduszt fogjuk használni.

Gyakorlatban a módusz meghatározásához minden $M \in \left[m; N-n+m\right]$ paraméterre ki kell számolni a posterior valószínűséget, és meg kell keresni azt az M-et, ahol ez felveszi a maximumát:

$$f_m(M) = \frac{\binom{N}{M} P^M (1-P)^{N-M} \binom{M}{m}\binom{N-M}{n-m}}{\sum\limits_{i\in\left[m;N-n+m\right]} \left\{\binom{N}{i} P^i (1-P)^{N-i} \binom{i}{m}\binom{N-i}{n-m}\right\}}.$$

Ismert N, n és m mellett ez a számítás nagyon könnyen elvégezhető, például az Excel felhasználásával. Ennek ellenére most levezetünk egy eljárást, amivel egyszerűbben is meghatározható a módusz helye.

Algebrai átalakításokkal igazolható, hogy a posterior valószínűségeloszlást a következő gyakoriságfüggvénnyel adhatjuk meg:

$$f_m(M) = \binom{N-n}{M-m} P_0^{M-m} (1-P_0)^{N-n-(M-m)}.$$

Ezt a formulát megvizsgálva látszik, hogy a gyakoriságfüggvények bármely m esetén egymásnak az „x-tengely” (a független változó, jelen esetben M) mentén való eltoltjai, ezért elégséges csak az $m=0$ esetre elvégezni a számítást, a többi értéket az $f_x(M) = f_0(M-x)$ azonosság felhasználásával kapjuk. Ha $m=0$, a következő, egyszerűsített formulát kapjuk:

$$f(M) = \binom{N-n}{M} P_0^{M} (1-P_0)^{N-n-M}.$$

M egységnyi növekedésekor ez a következőképpen változik meg:

$$\frac{f(M+1)}{f(M)} = \frac{N-n-M}{M+1} \cdot \frac{P_0}{1-P_0}.$$

Ez a függvény M-ben monoton csökken, tehát a (legnagyobb) módusz annál a legkisebb M-nél lesz, ahol a függvény értéke egynél kisebb (több módusz esetén a nem maximális móduszokra a monotonitás miatt a hányados értéke 1). Ezt a feltételt felírva az m=0-ra a módusz helye:

$$M_{m=0} = \lceil (N-n+1)P_0 - 1 \rceil,$$

ahol $\lceil \ \rceil$ jelöli a felfelé kerekítés műveletét.

Ez felhasználva kapjuk a végleges bayesi pontbecslést: $\hat{P} = \dfrac{M_{m=0}+m}{N}$.

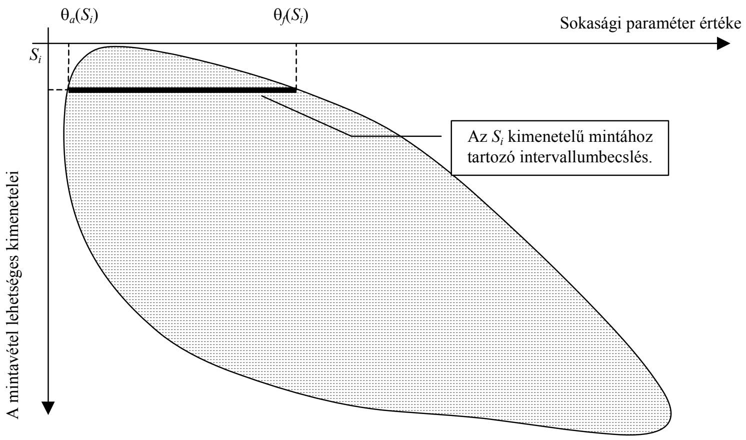
1.2. Intervallumbecslés

Az intervallumbecslés kérdései és tulajdonságai korántsem olyan kidolgozottak a szakirodalomban, mint a pontbecslésé. Elegendő arra utalni, hogy az intervallumbecslés tulajdonságait többnyire a neki megfeleltethető teszt tulajdonságaiból származtatják. Ezért indokolt egy kicsit részletesebben bemutatni az intervallumbecslés elméleti alapjait, különös tekintettel a hagyományos és a bayesi szemlélet egységes kezelésére, valamint a diszkrét eloszlásokból adódó és randomizálással megoldható problémákra.

1.2.1. Elméleti bevezetés

Formálisan az intervallumbecslés egy olyan halmazértékű leképezés, amely egy adott mintához és megbízhatósági szinthez hozzárendeli a valós számegyenes egy intervallumát úgy, hogy az megfeleljen bizonyos elvárásoknak. Azonban az, hogy mik ezek az elvárások, már korántsem ilyen magától értetődő.

1. ábra. Intervallumbecslés



A konfidenciaintervallum látszólag egyértelmű definíciója („az az intervallum, amely a becsülni kívánt sokasági jellemzőt adott valószínűséggel tartalmazza") pontatlan abban az értelemben, hogy nem határozza meg a valószínűségi mezőt, azaz azt, hogy milyen populáción kell kiszámítani, „mihez kell viszonyítani" a valószínűséget.[3]

A valószínűségi mezőt többféleképpen lehet meghatározni. A természetes definíció a „teljes" valószínűségi mező. Ez azt jelenti, hogy valószínűségi változónak tekintjük mind a becsülni kívánt sokasági paramétert, mind pedig a mintát.[4] Ilyenkor a konfidenciaintervallum definíciójában az összes lehetséges sokaságból vett összes elképzelhető mintán kell kiszámolni a lefedés valószínűségét, ami az 1. ábrán azt jelenti, hogy a besatírozott rész területének ki kell tennie a teljes terület megbízhatósági szintnek megfelelő százalékát.

Ennek az értelmezésnek egy speciális esete a hagyományos (mintavételi) statisztikához kötődik. A mintavételi statisztika egyik alapfeltevése, hogy a becsülni kívánt sokasági jellemző ismeretlen ugyan, de mégsem valószínűségi változó. A megbízhatósági szint ebben az esetben úgy is értelmezhető, mint egy feltételes valószínűség, ahol a feltételt a konkrét sokaság jelenti. Megemlítendő, hogy ezzel tökéletesen összhangban van az intervallumbecslés alapozó statisztika tankönyvekben található interpretációja: „ismételt mintavétel esetén az esetek átlagosan $(1-\alpha)\cdot 100$ százalékában igaz az, hogy az így számított intervallum lefedi a keresett sokasági jellemzőt". Nyilvánvaló, hogy ha egy becslés ebben az értelemben teljesíti a konfidenciaintervallum-becslés kritériumait, akkor a teljes valószínűségi mezőn nézve is teljesíti azt. (Lásd a 2. ábrát.)

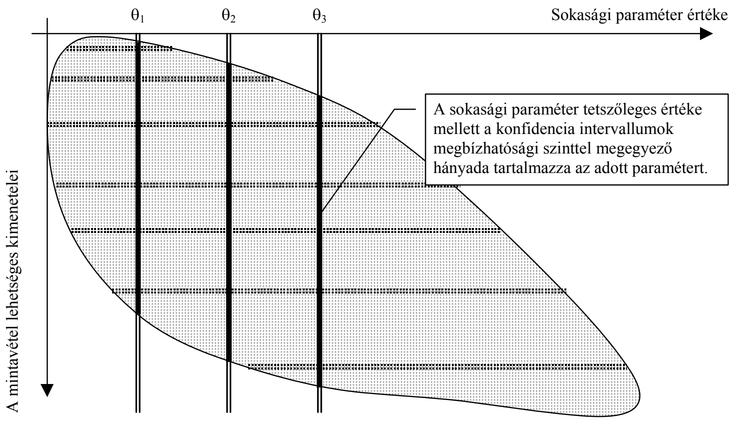
Egy másik speciális eseteként tárgyalható a bayesi felfogás, ahol a becsült intervallum a sokasági jellemző értékét az adott kimenetelű minták halmazán fedi le előre rögzített valószínűséggel. Az első speciális esethez hasonlóan most is minden, e definíció szerinti

[3] A következőkben felváltva, szinonimaként fogom használni a valószínűség, várható érték és a mérték fogalmakat. Ez megengedhető, hiszen egy esemény bekövetkezési valószínűsége megegyezik az esemény karakterisztikus függvényének várható értékével.

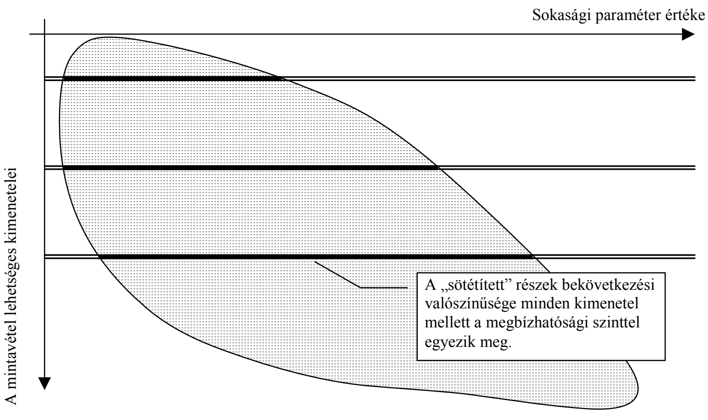
[4] A sokasági paraméter valószínűségi változóként való kezelése – mint arra korábban már utaltunk – legkönnyebben szuperpopulációs megközelítéssel ideologizálható, tehát azzal, hogy a vizsgált sokaság maga is egy nagyobb, adott esetben végtelen elemszámú sokaságból, a „szuperpopulációból" vett minta.

konfidenciaintervallum-becslés megmarad konfidenciaintervallum-becslésnek a teljes valószínűségi mezőn történő értelmezés alapján is.[5]

2. ábra. A mintavételi statisztika konfidenciaintervallumai



3. ábra. Bayesi konfidenciaintervallumok



A továbbiakban konfidenciaintervallum becslésén a következőt értjük: Az $(1-\alpha)\cdot 100$ százalékos megbízhatósági szinthez tartozó konfidenciaintervallum-becslésnek nevezzük azt az intervallumbecslést, amely a teljes valószínűségi mezőn számolva az esetek $(1-\alpha)\cdot 100$ százalékában tartalmazza a keresett sokasági jellemzőt. (Az ábrán a besatírozott rész területének kell kitennie a teljes terület $(1-\alpha)\cdot 100$ százalékát,

[5] Könnyen látható, hogy a megbízhatósági szint a bináris értékek (az adott mintához rendelt intervallum vagy tartalmazza, vagy nem tartalmazza a sokasági értéket) „átlagát” jelenti. A feltételes valószínűségekben szereplő feltételek teljes eseményrendszert alkotnak, tehát a teljes mezőn értelmezett várható érték ezeknek a feltételes várható értékeknek a súlyozott átlaga. Közismert, hogy az átlag mindig a legkisebb és a legnagyobb átlagolandó érték közé esik, és ebben az esetben az összes átlagolandó érték megegyezik.

egyéb feltevések nincsenek). A korábban leírtakból következik, hogy ez a definíció mind a bayesi statisztika, mind pedig a mintavételi statisztikai konfidenciaintervallum fogalmának általánosítása.

Az intervallumbecslés ilyen hosszú felvezetését az indokolta, hogy a két definiált speciális esetből teljesen más intervallumkészítési eljárás következik.

Az elemezni kívánt probléma (sokasági arány becslése) adott megbízhatósági szint mellett leegyszerűsíthető a következő problémára: Ismert N, n és m mellett készítsünk konfidenciaintervallum-becslést M-re.[6] Ennek a feladatnak a megoldása azért egyszerűbb, mivel az $S:=\{S_i\}$ halmaznak most csak az n mintaelemszámú elemeit kell megvizsgálni, melyeket a továbbiakban m-mel, a mintában lévő minősített egyedek számával reprezentálunk. Az intervallumkészítési módszerek bemutatásához tekintsük tehát az M és m összes lehetséges értékét, továbbá az adott m-hez tartozó intervallumbecslést tartalmazó, $n+1$ sorból ($m = 0\ldots n$) és $N+1$ oszlopból ($M = 0\ldots N$) álló mátrixot.

M

	0	1	2	3	4	5	6	7	8	9	.	.	.	.	.	.	.	.	N
0	X	X	X	X															
1	X	X	X	X	X														
2		X	X	X	X	X													
3			X	X	X	X	X	X											
4					X	X	X	X	X	X									
.						X	X	X	X	X	X	X							
.								X	X	X	X	X	X	X					
.										X	X	X	X	X	X	X			
n																			

m

A bayesi intervallumkészítési eljárás során először meghatározzuk minden m-hez M feltételes (posterior) eloszlásfüggvényét; az intervallumot úgy kapjuk, hogy meghatározzuk ezen eloszlás megbízhatósági szintnek megfelelő kvantiliseit. Nyilvánvaló, hogy az így kapott intervallumbecslés konfidenciaintervallum-becslés, hiszen minden m mellett a tartalmazás feltételes valószínűsége egyenlő a megbízhatósági szinttel.

Szót kell ejtenünk arról a gyakorta felmerülő nehézségről, hogy a probléma diszkrét jellege miatt nem található megfelelő kvantilis, előfordulhat, hogy egy adott M még alatta van a keresett értéknek, a szomszédja viszont már felette. Ebben az esetben a következő eljárást követjük. Első lépésként alapintervallumnak tekintjük azt a legbővebb intervallumot, amely az elméleti kvantiliseken belül esik, majd ehhez az alapintervallumhoz képezzük a randomizált intervallumot. A randomizált intervallum úgy keletkezik, hogy az alapintervallumot a következő szabály alapján kibővítjük a szomszédos elemmel (a példát az intervallum alsó végpontjára írjuk fel, a felső végpont esetén analóg módon járunk el).

[6] Könnyen látható, hogy az eredeti feladat megoldását az egyszerűsített feladat megoldásának N-nel való osztásával kapjuk.

Tekintsük az alapintervallum alsó végpontját, tehát azt az M_a értéket, amely minimalizálja az $F(M_a \mid m) \geq \frac{\alpha}{2}$ implicit egyenlet[7] bal oldalát (ekkor nyilván $F(M_a - 1 \mid m) < \frac{\alpha}{2}$). Legyen p olyan, hogy $p \cdot F(M_a - 1 \mid m) + (1 - p) \cdot F(M_a \mid m) = \frac{\alpha}{2}$. A randomizált intervallum alsó végpontja p valószínűséggel $M_a - 1$, $1 - p$ valószínűséggel pedig M_a.

A mintavételi statisztika filozófiájából következő intervallumkészítési eljárás ennél összetettebb: egy adott m_0-hoz úgy rendeljük az $M_a(m_0)$ és $M_f(m_0)$ intervallumvégpontokat, hogy azok minimalizálják a $Pr(m > m_0 \mid M_a) \geq \frac{\alpha}{2}$, illetve a $Pr(m < m_0 \mid M_f) \geq \frac{\alpha}{2}$ implicit egyenletek bal oldalát.

Mivel az $F(m \mid M)$ feltételes eloszlásfüggvény mindkét változójában monoton, ezért a fenti implicit egyenletek könnyen megoldhatók. Könnyen látható továbbá, hogy azon m-ek halmaza, melyekre egy adott M_i benne van az $M_a(m)$ és $M_f(m)$ alkotta intervallumban, szintén intervallum, éspedig olyan intervallum, aminek tetszőleges m elemére $Pr(x > m \mid M_i) \geq \frac{\alpha}{2}$ és $Pr(x < m \mid M_i) \geq \frac{\alpha}{2}$. Ebből következik, hogy tetszőleges M_i mellett a tartalmazás feltételes valószínűsége nem nagyobb, mint $(1 - \alpha) \cdot 100$ százalék.[8]

Az a feladatunk, hogy randomizálással kibővítsük ezt a becslést úgy, hogy az torzítatlan legyen (magyarul hogy a tartalmazás feltételes valószínűsége minden M_i mellett megegyezzen a megbízhatósági szinttel). Ezt a következő módon fogjuk megtenni.

M

	0	1	2	3	4	5	6	7	8	9	.	.	.	.	.	.	.	.	*N*
0	X	X	**X**	**X**	**X**														
1		**X**	X	X	X	**X**													
2			**X**	X	X	X	**X**	**X**											
3				**X**	X	X	X	X	**X**	**X**									
4					**X**	**X**	X	X	X	X	**X**	**X**							
.							**X**	X	X	X	X	X	**X**	**X**					
.								**X**	**X**	**X**	X	X	X	X	**X**	**X**			
.											**X**	X	X	X	X	X			
n																			

m

[7] A konvenciónak megfelelően $F(x){:=}Pr(\square < x)$ jelöli az eloszlásfüggvényt, azaz annak valószínűségét, hogy a valószínűségi változó az argumentumnál kisebb értéket vesz fel.

[8] A mintavételi statisztika intervallumkészítési eljárását szokás az úgynevezett pivot függvények definiálásával bevezetni. A pivot függvény olyan valós értékű függvény, amelynek két argumentuma a minta és a becsülni kívánt sokasági jellemző, továbbá a függvény a becsülni kívánt sokasági jellemzőben folytonos és monoton, valamint eloszlása független a becsülni kívánt sokasági jellemzőtől. Megmutatható, hogy az $F(m \mid M)$ feltételes eloszlásfüggvények eleget tesznek a pivot függvénnyel szemben támasztott követelményeknek, amiből pedig következik, hogy a fent leírt intervallumkészítési eljárás megfelel az $F(m \mid M)$ feltételes eloszlásfüggvény mint pivot függvény felhasználásával való intervallumkészítési eljárásnak.

Az ábrán az egyes *sorokban* látható, normál szedésű X-szel jelölt intervallumokat a $Pr(m > m_0 \mid M) \geq \frac{\alpha}{2}$, illetve a $Pr(m < m_0 \mid M) \geq \frac{\alpha}{2}$ implicit egyenletek *M*-ben közös megoldásai adják. A vastagított X-szel jelölt, randomizált értékek fogják biztosítani, hogy minden *M* esetén az X-szek által kijelölt (függőleges!!!) intervallum mértéke (az oszlopon belüli „súlya") megegyezzen a megbízhatósági szinttel. Ezen kívül azonban egy másik kritériumnak is eleget kell tennie a randomizált értékeknek: minden *m*-re a randomizálási folyamat végén a becslésnek intervallumnak kell maradnia (például a fenti táblázat $m = 4$, $M = 4$, 5 celláiban a 4;4 cella csak a 4;5 cellával együtt kerülhet kiválasztásra). Nyilvánvalóan ezeknek a kritériumoknak csak úgy lehet megfelelni, ha egyrészről egy adott „kivastagítás" kiválasztási valószínűségét az őt tartalmazó oszlop feltételes eloszlása határozza meg (például m, M alsó randomizált végpont esetén a kiválasztási valószínűség $p_{m,M} \cdot Pr(m \mid M) + F(m \mid M) = 1 - \frac{\alpha}{2}$ implicit egyenlet megoldása $p_{m,M}$-re), másrészről pedig a vízszintesen egymás mellett lévő kivastagítások kiválasztása függ egymástól. Szerencsére ez megtehető, ugyanis az $F(m \mid M)$ feltételes eloszlásfüggvény *M*-ben monoton csökkenő, azaz – ismét csak alsó randomizált végpontokra felírva – $p_{m,M} > p_{m,M-1}$. Ezt figyelembe véve m, $M - 1$ randomizált kiválasztására csak abban az esetben kerülhet sor, ha m, M már kiválasztásra került, és ilyenkor m, $M - 1$ kiválasztási valószínűsége $\frac{p_{m,M-1}}{p_{m,M}}$.

A randomizálás technikai kivitelezését a gyakorlatban a következő, ekvivalens módon tehetjük meg: Legyen *m* a mintabeli érték, és legyen *I*(*m*) a mintabeli értékhez a $Pr(m > m_0 \mid M_a) \geq \frac{\alpha}{2}$, illetve a $Pr(m < m_0 \mid M_f) \geq \frac{\alpha}{2}$ implicit egyenletek által adott alapintervallum. Tekintsük az *m* mintabeli értékhez az $I(m - 1)$, $I(m)$ és $I(m + 1)$ intervallumokat! Könnyen látható, hogy az alsó randomizálandó értékek $I(m - 1) \setminus I(m)$, a felsők pedig $I(m + 1) \setminus I(m)$. Az egyes randomizálandó értékekhez a fent leírt módon képezhetjük a $p_{m,M}$ valószínűségeket. Most az alsó értékekre bemutatjuk, hogyan lehet kiválasztani az intervallum randomizált alsó határát (a felső határ esetén analóg módon kell eljárni).

Ha $I(m - 1) \setminus I(m)$ elemei M, $M + 1$, … , $M + i$, akkor legyen $p'_M := p_{m,M}$, $p'_x := p_{m,x} - \sum_{j<x} p'_j$. Könnyen látható, hogy ha M, $M + 1$, … , $M + i$ potenciális alsó határokhoz ezeket a kiválasztási valószínűségeket rendeljük, akkor az ilyen módon kapott intervallumok mellett a becslésünk torzítatlan lesz.

Ennyi elméleti bevezetés után vizsgáljunk meg pár konkrét becslési eljárást.

1.2.2. A mintavételi statisztikához kapcsolódó becslések

A mintavételi statisztika ebben az esetben a bayesi szemlélettel szembenálló hagyományos (klasszikus, egyes bayesi megfogalmazás szerint ortodox) statisztikát jelenti. A feltételeket illetően ebben az alfejezetben tehát a vizsgálat tárgyát képező P, illetve M sokasági paramétert nem tekintjük valószínűségi változónak.

Nagymintás alapeset (M1). Kiinduló pontunk a tankönyvekben is részletesen tárgyalt becslés, amely esetén a visszatevéses nagy minta esetén a normális eloszlással való közelítés jogosnak tűnik. Ez az eset, mint a bevezetőben említettük, nem felel meg az ellenőrzési mintavétel követelményeinek, de itt, mint kiinduló pontot, etalont tekintjük.

Az eljárás a következő: *„végtelen"* (legalább több ezres nagyságrendű) sokaságból, vagy pedig kisebb sokaságból, de *visszatevéssel* történő mintavétel esetén a mintában található minősített elemek száma binomiális eloszlást követ. A mintabeli hibaarányról (p) bizonyítható, hogy szintén binomiális eloszlású, továbbá $E(p)=P$ és $Var(p)=\frac{P(1-P)}{n}$, ahol P jelöli a sokasági hibaarányt.

Ha $\min\{np, n(1-p)\} \geq 10$ (tehát legalább 10 hibás és 10 nem hibás elemet találtunk a mintában), akkor a minta hibaarányának transzformáltja közelítőleg standard normális eloszlást követ, amiből az intervallumbecslés:

$$p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}},$$

ahol $z_{1-\frac{\alpha}{2}}$ a standard normális eloszlás megfelelő kvantilise.

Az ellenőrzési gyakorlatban a gazdasági folyamatok jellegéből adódóan 5-10 százaléknál nagyobb arányú hibát már komolyabb kockázati tényezőként szokás figyelembe venni. Ha ezek alapján megvizsgáljuk ennek a módszernek az alkalmazási lehetőségeit, kiderül, hogy legalább többszáz elemű minta szükséges az $np \geq 10$ feltétel teljesítéséhez, ami a gyakorlati alkalmazások esetében általában nem valósul meg, így a közelítő eljárás alkalmazása torzítást visz az intervallumbecslésbe. További torzításra ad okot a modell kezdeti feltevése, azaz a végtelen sokaságból, vagy pedig visszatevéssel történő mintavétel. A visszatevés nélküli mintavétel esetén a mintabeli hibaarány varianciáját csökkenti egy $\frac{N-n}{N-1} < 1$ szorzó, ezért tehát ennek a feltevésnek a „megszegése" elvileg „jó" irányú torzítást okoz. A teljes torzítás mértékéről analitikusan nehéz pontosan nyilatkozni, de ha összevetjük ezt a becslést a később bemutatásra kerülő M3 becslőfüggvényünkkel, látható, hogy a mintaméret növekedésével (és a kiválasztási arány nullához tartásával) a két becslés is konvergál egymáshoz. Az egyes becslési eljárások pontos torzítási mértékét Excel segítségével meghatároztuk, ennek részleteiről bővebben a 4. rész számol be.

Visszatevés nélküli minta – egzakt hipergeometriai eloszlás (M2). A hipergeometriai eloszlás melletti intervallumbecslés (a továbbiakban: M2) esetén a feltételes eloszlás a következő gyakoriságfüggvény m szerinti („függőleges”) kumulálásával adódik:

$$Pr(m = x \mid M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

ha $\max(0, n - N + M) \le x \le \min(n, M)$, egyébként 0.

Miután meghatároztuk az m, $m - 1$ és $m + 1$ melletti alapintervallumokat és a p' értékeket, véletlenszám-generátorral kiválaszthatjuk a randomizált intervallumvégpontokat.

Az ellenőrzési szakma egyik „etalonnak” tekinthető szoftverének, az IDEÁ-nak a mintavételi modulja is az M2 eljárást használja azzal az eltéréssel, hogy randomizálás helyett mindig a legkonzervatívabb (a randomizálással kapható legbővebb) intervallumot adja meg.

Hipergeometriai eloszlás - normális közelítés (M3). A közelítő eljárások bevezetésére annak idején főként azért került sor, mivel sokáig nem álltak rendelkezésre táblázatok a hipergeometriai eloszláshoz, így valamilyen folytonos eloszlással helyettesítették a diszkrét eloszlást. Mivel a becslés végső célja a hibaarány meghatározása, a közelítő eljárásokban m és M szerepét általában $\frac{m}{n}$ és $\frac{M}{N}$ veszi át, és ezen kívül legtöbbször figyelmen kívül hagyják az előbbi hányadosok diszkrét jellegéből adódó sajátosságokat is, így nem kerül sor randomizálásra sem.

A korábban leírt intervallumkészítési eljárás figyelembevételével „elemi” módon könnyen készíthetünk közelítő módszerrel becslőfüggvényt. Jelen esetben közelítsük a feltételes eloszlást olyan normális eloszlással, melynek várható értéke és varianciája megegyezik a megfelelő hipergeometriai eloszláséval. A normális eloszlás kvantilis értékeit használva egy adott P sokasági hibaarány mellett azon mintabeli hibaarányok (x) halmaza, amelyekhez tartozó intervallumbecslések tartalmazzák P-t:

$$x \in \left[P - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P \cdot (1-P)}{n} \frac{N-n}{N-1}} ; P + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P \cdot (1-P)}{n} \frac{N-n}{N-1}} \right].$$

A konstrukció során azzal a feltevéssel élünk, hogy bármely x mintabeli hibaarány konfidencia intervallum végpontjához tartozó sokasági hibaarány fent definiált halmazának x a határpontja. Mivel folytonos közelítést alkalmazunk, ez a feltevés tartható, továbbá látszik, hogy ebben az esetben a konfidencia intervallum „felső” végpontja a halmaz alsó határa lesz, és fordítva. Írjuk fel ezt egyenletrendszer formájában:

$$x = P_f - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P_f \cdot (1-P_f)}{n} \frac{N-n}{N-1}},$$

$$x = P_a + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P_a \cdot (1-P_a)}{n} \frac{N-n}{N-1}}\,.$$

A feladatunk az, hogy ebből az egyenletrendszerből kifejezzük P_a -t és P_f -et. Könnyen látható, hogy átrendezés és négyzetre emelés után a két egyenlet ugyanúgy fog kinézni, és mivel szintén könnyen látható, hogy P_a -ban, illetve P_f -ben másodfokú egyenletet kapunk, P_a lesz a kisebbik, és P_f a nagyobbik gyök.

Az egyenlet megoldásához először is végezzük el a $c = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}}$ helyettesítést, ami után egyenletünk az $x = p \pm c \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$ formát ölti.

Átrendezve és négyzetre emelve:

$$nx^2 - 2nxp + np^2 = c^2 \cdot p \cdot (1-p)\,.$$

Ismét átrendezve:

$$(n+c^2)p^2 - (2nx+c^2)p + nx^2 = 0\,.$$

Felírva a megoldóképletet és tovább rendezve:

$$p_{1,2} = \frac{2nx + c^2 \pm \sqrt{(2nx+c^2)^2 - 4(n+c^2)nx^2}}{2(n+c^2)} =$$

$$= \frac{nx + c^2 \cdot 0{,}5}{n+c^2} \pm c\sqrt{\frac{n}{(n+c^2)}}\sqrt{\frac{n\frac{x(1-x)}{n} + c^2\frac{0{,}25}{n}}{(n+c^2)}}\,.$$

Bevezetve a $\lambda = \frac{n}{n+c^2}$ (pozitív, 1-nél kisebb számmal való) helyettesítést:

$$p_{1,2} = \left[\lambda x + (1-\lambda)\cdot 0{,}5\right] \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}}\sqrt{\lambda}\sqrt{\left[\lambda\frac{x(1-x)}{n} + (1-\lambda)\frac{0{,}5\cdot(1-0{,}5)}{n}\right]}\,.$$

Vegyük észre, hogy $n \to \infty$, és $\frac{n}{N} \to 0$ esetén $\lambda \to 1$ és $\frac{N-n}{N-1} \to 1$, így becslésünk határértékben az $x \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{x(1-x)}{n}}$ alakot ölti (ez megegyezik az M1 becsléssel).

Ezt a meglepően szép formájú, konvex kombinációt tartalmazó becslőfüggvényt nehéz volna intuitív módon előállítani, de a tesztekből ki fog derülni, hogy az elméleti konstrukcióval összhangban kis minták esetén is gyakorlatilag torzítatlan intervallumokat ad.

1.2.3. Bayesi szemléletű becslések (B)

A konfidenciaintervallum-becslés bayesi szemléletű definícióját, illetve a becslésre adott konstrukciós eljárást megvizsgálva látható, hogy a teljes valószínűségi mező intervallumokkal való lefedettségének mértéke általában nem egyenletes az egyes sokasági arányok mentén: ez amiatt van, hogy a priorban szereplő eloszlásnak megfelelően bizonyos sokasági arányok túlreprezentáltak a becslésben. Mindebből következik az is, hogy elvileg a bayesi szemléletű becslőfüggvények jóval pontosabb (szűkebb) intervallumokat eredményeznek; más kérdés, hogy ezen intervallumok ex-post megbízhatóságát hogyan befolyásolja, ha a prior jelentősen eltér a valóságtól.

Mivel az $F(m \mid M)$ feltételes eloszlás minden M-re olyan, hogy a hozzá tartozó feltételes sűrűségfüggvény n növekedésével egyre inkább egy pontra (éspedig $\frac{nM}{N}$-re) koncentrálódik, a priortól függetlenül minden intervallumbecslés $M = \frac{Nm}{n}$ értékre fog ráhúzódni, ha $n \to N$.

Kérdés azonban, hogy vajon ez a konvergencia milyen gyors, tehát hogy az ellenőrzésben használatos mintaméretek mellett érezteti-e a hatását. Fontos tudni, hogy a prior esetleges helytelen megválasztása milyen mértékben képes befolyásolni magát a becslést, és így az auditor által kialakított véleményt is.

Egy bayesi szemléletű intervallumbecslés elkészítéséhez két dolog ismerete szükséges: egyrészről ismernünk kell az $F(m \mid M)$ feltételes eloszlást, másrészről pedig az M prior eloszlását. Az ismert statisztikai összefüggések miatt $F(m \mid M)$-et a hipergeometriai eloszlás értékeit felhasználva kaphatjuk meg, ám M eloszlásával kapcsolatosan többféle feltevéssel is élhetünk:

(B1a.) Az első lehetséges feltevés, hogy *makro szemléletben* annak a valószínűsége (P_1), hogy egy adott egyed rendelkezik azzal a bizonyos jellegzetességgel, a korábbi ellenőrzési tapasztalatokból ismert, rendszersajátosságot tükröző, stabil paraméter. Egy ilyen – végtelen nagyságú – szuperpopulációt feltételezve a vizsgált sokaságra jellemző sokasági arány valószínűségi változó, amely (N, P_1) paraméterű binomiális eloszlást követ;

(B1b.) Egy másik lehetséges feltevés a vizsgált szervezetnél, vagy hasonló szervezeteknél lefolytatott korábbi ellenőrzések empirikus tapasztalatain alapul. Ezek alapján szintén megadható annak a valószínűsége (P_2), hogy egy adott egyed rendelkezik azzal a bizonyos jellegzetességgel, így a sokasági arány ismét csak binomiális eloszlású, (N, P_2) paraméterekkel;

(B2.) Ha egy adott jellemző sokasági arányáról hosszú időre visszamenően rendelkezünk megfigyelésekkel, és úgy találjuk, ez a sokasági arány eloszlásában stabil, akkor ezt az empirikus adatsort is felhasználhatjuk a becslésünkhöz;

(B3.) Végül használhatunk nem informatív priorként egyenletes eloszlást.

Az első két feltevés mellett a következő posterior gyakoriságot kapjuk:

$$f_m(M) = \frac{\binom{N}{M} P^M (1-P)^{N-M} \binom{M}{m} \binom{N-M}{n-m}}{\sum\limits_{i \in [m; N-n+m]} \left\{ \binom{N}{i} P^i (1-P)^{N-i} \binom{i}{m} \binom{N-i}{n-m} \right\}},$$

ha $M \in [m; N-n+m]$, egyébként 0.

Mivel a posterior valószínűség megegyezik annak valószínűségével, hogy a fennmaradó $N-n$ elemű sokaságban $M-m$ hibás elem található, ezért a fenti képlet a következő alakra egyszerűsíthető: $f_m(M) = \binom{N-n}{M-m} P^{M-m} (1-P)^{N-n-M+m}$, ami természetesen algebrai átalakításokkal is könnyen belátható.

Ebből a formulából kitűnik, hogy a posterior gyakoriság nem változik, ha M és m azonos módon változik, így fennáll az $M_a(m) = M_a(0) + m$ és $M_f(m) = M_f(0) + m$ összefüggés.

Az empirikus adatsor alapján minden sokasági hibaarányhoz hozzárendelhető egy $Pr_e(p)$ empirikus bekövetkezési valószínűség, aminek felhasználásával megadható a sokasági hibaszám prior eloszlása is.

A posterior valószínűségek:

$$f_m(M) = \frac{Pr_e\left(\frac{M}{N}\right) \binom{M}{m} \binom{N-M}{n-m}}{\sum\limits_{i \in [m; N-n+m]} \left\{ Pr_e\left(\frac{M}{N}\right) \binom{i}{m} \binom{N-i}{n-m} \right\}}.$$

Egyenletes eloszlású prior mellett a posterior gyakoriság a következő:

$$f_m(M) = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\sum\limits_{i \in [m; N-n+m]} \left\{ \binom{i}{m} \binom{N-i}{n-m} \right\}},$$

ha $M \in [m; N-n+m]$, egyébként 0.

A posterior eloszlásra a korábban leírt eljárást kell alkalmazni: először is meg kell határozni az alapintervallum végpontjait az $F(M_a) \geq \frac{\alpha}{2}$ implicit egyenlet minimalizálásával, illetve az $F(M_f) + f_m(M_f) = F(M_f + 1) \leq 1 - \frac{\alpha}{2}$ maximalizálásával, ahol $F(M)$ a fenti $f_m(M)$ függvények $M-1$-ig való kumulálásával adódik. Ezek után kell elvégezni a randomizálást $M_a - 1$, illetve $M_f + 1$ értékekre a $p_a = \dfrac{F(M_a) - \frac{\alpha}{2}}{f_m(M_a - 1)}$ és a $p_f = \dfrac{1 - \frac{\alpha}{2} - F(M_f + 1)}{f_m(M_f + 1)}$ kiválasztási valószínűségek felhasználásával.

A könnyebb kezelhetőség kedvéért bevezetünk egy közelítő módszert a B1 becsléssel kapható intervallum meghatározására. Mivel fennáll az $f_m(M) = f_{m+x}(M + x)$ összefüggés, és $M \; m = 0$ melletti posterior eloszlása $(N - n, P)$ paraméterű binomiális eloszlás, ezért M feltételes posterior várható értéke $(N - n)P + m$, feltételes posterior szórása pedig $\sqrt{(N - n)P(1 - P)}$. Ezt figyelembe véve átalakítások után a bayesi „gyorsbecslés" képlete:

$$\kappa \cdot \frac{m}{n} + (1 - \kappa) \cdot \left(P \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P \cdot (1 - P)}{N - n}} \right),$$

ahol $\kappa = \dfrac{n}{N}$, azaz a kiválasztási arány.

Ez a képlet a következő módon interpretálható: A sokaságból kiválasztott n elemű mintában $\dfrac{m}{n}$ a hibaarány. A mintavételkor „kihagyott" részben található hibák száma előzetes feltevésünk szerint $(N - n, P)$ paraméterű binomiális eloszlást követ. A becslést ennek a két résznek az átlagolásával kapjuk. Ez a becslés azért csak közelítés, mivel a binomiális rész kvantiliseit csak közelítve határoztuk meg.

1.2.4. Egy vegyes becslési módszer

A mintavételi statisztika keretében tárgyalt nagymintás becslési módszerből külső információ felhasználásával egy kis mintákra hatékonyan alkalmazható „hibrid" (MxB1) hozható létre. Emlékeztetőül: az M1 nagymintás becslési eljárás során p mintabeli hibaarányhoz hozzárendeltük a $p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\dfrac{p \cdot (1 - p)}{n}}$ intervallumot. Abban az esetben, ha előzetes tapasztalatokkal (P_0) rendelkezünk a hibaarányról, ezt a becslés során leginkább a standard hiba meghatározásánál használhatjuk fel.

Az új becslőfüggvényünk:

$$\hat{P} \in p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P_0 \cdot (1 - P_0)}{n}}.$$

Ez a becslés bayesi intervallumbecslésként is felfogható azzal a feltételezéssel, hogy a sokasági hibaarány prior eloszlása a P_0 pontra koncentrálódik.

Nyilvánvaló, hogy ez a becslőfüggvény torzított, és a torzítás aszimptotikusan sem szűnik meg. Azonban kis minták esetén, ha P valóban P_0 közelében szóródik, ez a becslés jóval megbízhatóbb, mintha a standard hibát is a mintából határoznánk meg.

2. HIPOTÉZISVIZSGÁLAT

Egy pénzügyi-gazdasági ellenőrzés lefolytatásakor a megvizsgált minta alapján több esetben nyilatkozni kell arról is, hogy a hibák gyakorisága nem lép túl egy előre rögzített szintet. A hipotézisvizsgálat hagyományos elmélete szerint ez történhet egyrészt úgy, hogy a minta eredménye alapján megmondjuk, mekkora bizonyossággal jelenthetjük ezt ki[9] (lásd H1 módszer), de történhet úgy is, hogy egy előre rögzített bizonyossági szint mellett kijelentjük, hogy elfogadható, vagy nem elfogadható ez a kijelentés[10] (lásd H2 módszer).

A hagyományos eljárással szemben a bayesi hipotézisvizsgálat alapvetően arról nyilatkozik, hogy a minta inkább a hipotézist, vagy annak tagadását támasztja-e alá.

2.1. Hagyományos hipotézisvizsgálat

A hipotézisvizsgálat hagyományos lefolytatásakor mintát veszünk a sokaságból, és a minta kimeneteléhez hozzárendeljük a döntésünket: vagy megmondjuk a bizonyosság mértékét, vagy elfogadjuk/elutasítjuk az állítást rögzített bizonyossággal. Az elfogadott kimeneteleket elfogadási tartománynak, az elutasított kimeneteleket elutasítási tartománynak nevezzük.

A továbbiakban nevezzük (null)hipotézisnek azt az állítást, hogy a sokaságbeli hibák gyakorisága meghaladja az előre rögzített szintet ($P > P_h$, ahol P_h az ellenőrzési szaknyelvben „tolerálható hibaarány” néven ismert mennyiség); ennek ellenhipotézise (vagy más néven az alternatív hipotézis) az, hogy a sokaság hibaszáma nem haladja meg ezt a szintet ($P \le P_h$).

Mindezek mellett döntésünkkel két fajta hibát követhetünk el:

– A minta alapján – tévesen – elutasítjuk a hipotézist, tehát annak ellenére, hogy a valós hibaszázalék meghaladja a rögzített szintet, mi ezt mégsem fogadjuk el. (elsőfajú hiba).

– A minta alapján – tévesen – elfogadjuk a hipotézist, tehát annak ellenére, hogy a valós hibaszázalék alatta marad a rögzített szintnek, nem vetjük el a hipotézist (másodfajú hiba).

A hipotézisvizsgálat (próba) megbízhatósági szintje annak valószínűségét mutatja meg, hogy a minta alapján helyesen fogadjuk-e el a nullhipotézist. Könnyen látható, hogy ez pont az elsőfajú hiba elkövetési valószínűségének komplementere, így ha az elsőfajú hiba elkövetési valószínűsége (más néven szignifikanciaszint) α , a megbízhatósági szint $1-\alpha$.

A próba erejének szokás nevezni annak a valószínűségét, hogy a helytelen hipotézist – helyesen – elutasítjuk. Adott megbízhatósági szint mellett nyilvánvaló cél, hogy a próba erejét maximalizáljuk.

[9] Ez vezet az ún. *p*-érték koncepcióhoz.
[10] Ez valójában a klasszikusnak számító Neyman–Pearson tesztelési elv, illetve stratégia.

Az eddigieket egy mátrixban szokás összefoglalni:

	Elfogadjuk (feltételes valószínűség)	Elutasítjuk (feltételes valószínűség)
Hipotézis igaz	Helyes döntés (1–□)	Elsőfajú hiba (□)
Hipotézis hamis	Másodfajú hiba (□)	Helyes döntés (1–□)

Mivel nem ismerjük a valós sokasági hibaarányt, ezért döntésünk megbízhatóságáról sem tudunk pontosan nyilatkozni; annak valószínűsége, hogy a mintavétel kimeneteléhez előre hozzárendelt döntésünk milyen valószínűséggel lesz helyes, függ a valós sokasági aránytól. Mivel azonban $F(p \mid P)$ P-ben monoton csökken, ezért a helyesen elfogadó döntés valószínűsége $P \geq P_h$ halmazon P_h fennállása esetén a legkisebb, és emiatt kijelenthető egy adott döntésről, hogy *legalább* mennyire megbízható. A továbbiakban ezt fogjuk megbízhatósági szintnek nevezni.

H1. Abban az esetben, ha a hipotézisvizsgálat célja a megvizsgált minta alapján megmondani, hogy mekkora legkisebb szignifikanciaszint mellett lehet elutasítani a hipotézist, formálisan a következő értéket kell kiszámolnunk.

$Pr(\xi > p \mid P_h)$, ahol $n \cdot \xi$ hipergeometriai eloszlású valószínűségi változó N, $N \cdot P_h$, és n paraméterekkel, továbbá p a mintából számított hibaarány; ez a valószínűség fogja megadni a megbízhatósági szintet. A képlet kiszámolásához egyszerűen a hipergeometriai eloszlás gyakoriságait kell $n \cdot p$-ig kumulálni.

H2. Terjedelmi okok miatt nem térünk a próbakészítési eljárásokra, ezért indoklás nélkül kijelentjük, hogy ez esetben a legjobb próbák elfogadási tartománya adott mintaméret mellett mindig intervallum. Mivel a minta hibaarányának nullhipotézis melletti eloszlása ismert, a feladatunk ezen eloszlás megbízhatósági szintnek megfelelő kvantilisét megkeresni. A kvantilis általában itt sem „lehetséges" érték, ezért szükséges lehet a kritikus érték randomizált meghatározása a becslési eljárásoknál már megismert módon.

A kritikus érték meghatározása után össze kell vetni a mintából kapott értéket a kritikus értékkel: amennyiben a kritikus érték meghaladja a mintabeli értéket, elutasítjuk a hipotézist, egyéb esetben viszont nem áll módunkban elvetni ezen a megbízhatósági szinten.

Az ellenőrzési szakmában, és ezzel összhangban az IDEA szoftverben a megbízhatósági szintet az eddig leírtakhoz képest némiképp eltérő módon értelmezik. Definiálják az alfa és béta kockázatot, amelyek értelmezése a következő:

alfa kockázat: annak a döntésnek a maximális bekövetkezési valószínűsége, hogy a valós hibaarány meghaladja a tolerálható hibaarányt, miközben valójában kisebb egy, úgynevezett alsó hibaaránynál. (Ez a kockázat a fenti nullhipotézis esetén hasonlít a másodfajú hiba definíciójához.)

béta kockázat: annak a döntésnek a maximális bekövetkezési valószínűsége, hogy a valós hibaarány nem haladja meg a tolerálható hibaarányt, miközben valójában mégis meghaladja.

Az alfa és béta kockázatok komplementereit nevezik alfa és béta megbízhatósági szintnek, tehát a korábban bevezetett megbízhatósági szint definíciónak a béta megbízhatósági szint a megfelelője abban az esetben, ha a nullhipotézis $P > P_h$.

2.2. Bayesi hipotézisvizsgálat

A bayesi hipotézisvizsgálat lényege, hogy összevetjük a hipotézis és az ellenhipotézis minta melletti bekövetkezésének valószínűségét. A valószínűségek meghatározásához ugyanazt az eljárást követjük, mint az intervallumbecslés esetén, nevezetesen első lépésként meghatározzuk a mintából kapott m értékhez M posterior valószínűségeloszlását. A döntési szabály ezek után az, hogy összevetjük a hipotézist a posterior eloszlás mediánjával: ha a medián a nagyobb, akkor elvetjük a hipotézist, ellenkező esetben elfogadjuk azt.

Az előbb leírtakat megfordítva és formalizálva, a bayesi hipotézisvizsgálat a következő módon történik.

1. Meghatározzuk az $F(P \mid p)$ posterior eloszlást minden p mellett;

2. A hipotézisként szereplő P_h-hoz hozzárendeljük azt a legnagyobb p_h-t, amely mellett még $F(P_h \mid p_h) \geq \frac{1}{2}$.

3. Ha a mintából származó $p \leq p_h$, elfogadjuk a hipotézist, ellenkező esetben elutasítjuk.[11]

3. A HIBAMENTES MINTÁBÓL LEVONHATÓ KÖVETKEZTETÉSEK

Ha a mintavételt követően a mintában nem találtunk hibát, a korábban ismertetett M2 (egzakt hipergeometriai eloszlást használó) becsléshez szükséges feltételes gyakoriságok a következő módon alakulnak.

$M = \ldots$	0	1	2	$\ldots$	i
$Pr(m = 0 \mid M)$	1	$\frac{N-n}{N}$	$\frac{(N-n)(N-n-1)}{N(N-1)}$		$\prod_i \frac{N-n-i+1}{N-i+1}$

Probléma lehet, ha a randomizált alsó végpont nagyobb, mint a randomizált felső végpont, ami természetesen ellentmondás. Ezt a jelenséget az teszi lehetővé, hogy $m = 0$-nál $F(m \mid M)$ minden M esetén konstans zérus, így nem létezik az „alapintervallum”. (Emlékeztetőül: egy adott m_0-hoz tartozó alapintervallumnak nevezzük az $M_a(m_0)$ és $M_f(m_0)$ végpontok által meghatározott intervallumot, ha azok minimalizálják a $Pr(m > m_0 \mid M_a) \geq \frac{\alpha}{2}$, illetve a $Pr(m < m_0 \mid M_f) \geq \frac{\alpha}{2}$ implicit egyenletek bal oldalát). Másként megfogalmazva: nincs olyan M, amely a randomizált intervallumok mindegyikében szerepelne.

A probléma megoldása lehet például, ha az alsó és felső végpontok randomizálással való meghatározása nem független egymástól. Ebben az esetben minden M értékhez hozzárendelünk p_M kiválasztási valószínűségeket (mindegyik szigorúan kisebb lesz egynél, hisz nincs alapintervallumunk). Rendezzük növekvő sorrendbe p_M-eket! Mivel $F(m \mid M)$ tulajdonságaiból következik, hogy a $\frac{p_{M+1}}{p_M}$ hányadosok monoton fogyó soro-

[11] Bár tartalmilag azonos ezzel az eredeti bayesi tesztelési stratégia, meg kell említeni, hogy az ottani keretek közt a döntést az ún. posterior esélyhányados (posterior odds) alapján hozzák meg. A posterior odds a nullhipotézis és az ellenhipotézis a posteriori bekövetkezési valószínűségeinek hányadosa; ha ez nagyobb 1-nél, akkor a nullhipotézis, ha kisebb 1-nél, az ellenhipotézis javára döntünk.

zatot alkotnak, ezért megmutatható, hogy léteznek olyan intervallumok, és az intervallumokhoz rendelt kiválasztási valószínűségek, melyekre bármely *M*-nek az előbbi intervallumokban való tartalmazási valószínűsége pontosan p_M.

Ha nagyon alacsony ($p << \alpha$) a korábbi tapasztalatok alapján várható hibaarány, megfigyelt hiba nélküli minta mellett felmerülhet olyan igény, hogy bizonyos megbízhatósági szinten kijelentsük: a vizsgált sokaságban nincs lényeges hiba. Adott szuperpopulációs hibaarány (p) mellett annak a valószínűsége, hogy a teljes sokaságban nincs hiba: $(1-p)^{N-n}$, ugyanis ez megegyezik annak valószínűségével, hogy a p hibaarányú végtelen sokaságból a véletlenszerűen kiválasztott *N–n* tétel egyike sem hibás. Ezért tehát ahhoz, hogy $1-\alpha$ megbízhatósági szinten – a minta alapján – kijelenthessük, hogy a sokaságban nincs hiba, a szükséges mintanagyság:

$$n = N - \frac{\ln(1-\alpha)}{\ln(1-p)},$$

ahol ln a természetes alapú logaritmust jelöli. A gyakorlati alkalmazás szempontjából ez a módszer csak akkor hasznos, ha *N* alacsony ugyan (<25), de az egy tranzakcióra eső ellenőrzési költség nagyon magas (például külső szakértőt kell igénybe venni, vagy túlzott időráfordítást jelentene a pótlólagos 5-10 tranzakció ellenőrzése).

4. A BECSLÉSI ELJÁRÁSOK ÉRTÉKELÉSE

A becslési eljárások torzítási mértékét a megbízhatósági szint – jelen tanulmány első felében adott – definíciója szerint a teljes valószínűségi mezőn kell vizsgálni.

A teljes valószínűségi mező a korábban leírtak alapján tartalmazza az összes lehetséges sokaság-minta forgatókönyvet. Egy konkrét sokaság-minta pároshoz tartozó valószínűség az ismert azonosság alapján felírható $Pr(m \cap M) = Pr(m \mid M) \cdot Pr(M)$ alakban, így – mivel a mintavételi terv ismeretében $Pr(m \mid M)$ is ismert – elegendő $Pr(M)$ meghatározása. Ehhez azonban feltevésekkel kell élnünk magáról a sokaságról.

A sokaságról való lehetséges feltevéseinket, továbbá azok indoklását részletesen kifejtettük a bayesi becslések bevezetésekor. Jelenlegi céljainkhoz ebből csak azt kell kiemelni, hogy a teljes valószínűségi mezőn értelmezett megbízhatósági szint tekinthető a mintavételi statisztika parciális (adott sokasági arány melletti ismételt mintavételt feltételező) megbízhatósági szintjeiből vett $Pr(M)$ súlyozású átlagnak.

A becslési eljárások értékelésekor első lépésben különböző sokaság- és mintaméretek mellett 15 százalék alatti sokasági hibaarányokra Excel segítségével meghatároztuk a parciális megbízhatósági szinteket. A B1 és MxB1 eljárások esetén a sokaság- és mintaméret mellett paraméterként szerepelt a feltételezett hibaarány is, természetesen más-más interpretációval.

A mintavételi statisztika becsléseinek konstrukciójukból adódóan minden sokasági arány mellett konstans 95 százalékos parciális megbízhatóságot kell(ett volna) mutatniuk, ezzel szemben a bayesi szemléletű becslésektől, különösen a B1 becsléstől, ez nem elvárható.

Az értékelés második szakaszára éppen azért volt szükség, mivel a B1 becslés megbízhatóságának priortól való függése az első szakasz eredményeiből közvetlenül nem volt megállapítható. A második szakaszban tehát az első szakasz eredményeit a szuperpopulációs hibaarány 0-15 százalék közötti értékeire kiszámított $Pr(M)$ súlyokkal átlagoltuk.

A bayesi szemléletű B2 becsléscsaládról nem tudunk univerzális érvénnyel nyilatkozni, de mivel ez a becslés átmenetet képez a „tiszta” B1 és B3 esetek közt, tulajdonságai is várhatóan valahol a kettő között találhatóak: minél kevesebb információval rendelkezünk a jellemző előfordulási valószínűségéről – minél kevésbé összpontosul az empirikus eloszlás egy értékre – annál inkább a B3 jellemzői érvényesülnek, ami egyenletesebb parciális megbízhatóságot de hosszabb (pontatlanabb) intervallumokat jelent.

4.1. Az első szakasz eredményei

A becslőfüggvények megbízhatóságának elemzéséhez explicit módon meghatároztuk adott N és n mellett minden lehetséges sokasági hibaarány mellett, hogy a becsült intervallumok m hipergeometriai eloszlását feltételezve az esetek hány százalékában tartalmazzák[12] M-et (illetve a valós sokasági hibaarányt). Ezzel párhuzamosan meghatároztunk egy várható intervallumhosszt is minden becsléshez minden sokasági hibaarányra. Az EXCEL-lel készült számítások néhány eredményét a Függelék 1-4. ábrái mutatják. Ezek példaként az $N = 600$ sokasági elemszám és a szokásos 95 százalékos megbízhatóság mellett az $n = 30$ (kisminta) és az $n = 150$ (nagyminta) esetekben mutatják a parciális (empirikus) megbízhatóságot, illetőleg a várható intervallumhosszt.

Az eredményekből kitűnik, hogy egyrészről az M3 becslés torzítása a többségében nem, vagy alig lépi túl a 1,5-2 százalékpontot, másrészről a torzítás hol pozitív, hol negatív irányú. Ezzel szemben az ellenőrzési szakirodalomban ajánlott M1 becslés akár 5-15 százalékpontnyi torzítást is tartalmaz. Az eredményeket tartalmazó grafikonon a valós sokasági hibaarány növekedésével meglepő periodicitás figyelhető meg az M1 és M3 becsléseknél rögzített N és n mellett. Ezen kívül megfigyelhető, hogy a becslések megbízhatósági görbéjének alakja gyakorlatilag érzéketlen a sokaság nagyságának változásaira, amiből az a következtetés vonható le, hogy a megbízhatóságot elsősorban a mintaméret befolyásolja (a kiválasztási arány növekedése a növekedéssel arányos megbízhatósági szint növekedést okoz az M1 becslésnél,[13] de a grafikon alakját érdemben sem az M1, sem az M3 becslésnél nem változtatja meg).

A mintavételi statisztika M2 becslőfüggvénye nem tartalmaz semmilyen közelítő eljárást és emiatt torzítást sem, amit a vizsgálat is alátámasztott.

Az informatív priorral való B1 becslés a szuperpopulációs hibaarány környezetében pozitív, azon kívül negatív irányban torzított. A pozitív torzítási tartomány nagysága fordítottan arányos a sokaság méretével, hozzávetőlegesen 2,5 (N=1500) és 5 (N=300) százalékpont között mozog. A mintaméret (kiválasztási arány) változása a pozitív torzítási tartomány méretét kevésbé, inkább a (parciális) megbízhatósági görbe alakját befolyásolja: minél magasabb a kiválasztási arány, a becslés (parciális) megbízhatósága priortól

[12] A randomizálást úgy vettük figyelembe, hogy az adott forgatókönyvre a tartalmazást kódoló 0-1 számok helyett a tartalmazás valószínűségét tüntettük fel.

[13] Az M1 becslésnél tapasztalt túlzott megbízhatósági szint növekedés például a statisztikai szakirodalomban ajánlott módon – a pontbecsléshez hozzáadott-kivont értékek (1-kiválasztási arány) tényezővel való szorzásával – korrigálható.

függetlenül annál jobban közelít a konstans 95 százalékhoz. Noha ez a konvergencia csak a 100 százalékhoz közeli kiválasztási arányoknál kezd igazán érezhető lenni, egy bayesi szemléletű becslésnél nem is elvárás, hogy a parciális megbízhatóság egyenletesen közel legyen az elvárt megbízhatósági szinthez.

A nem informatív prior használata melletti B3 becslés a mintavételi statisztika M3 becsléséhez nagyon hasonló eredményeket hozott.

Az M×B1 vegyes becslés torzítása a hipotézisben lévő tévedéssel ellentétes irányban változik: túlbecsült hibaarány esetén növekszik a megbízhatósági szint, alulbecslés esetén pedig 95 százalék alá csökken. Hüvelykujjszabályként megjegyezhető, hogy a hipotetikus hibaarány 1 százalékpontos csökkentése a megbízhatósági szintet 2-3 százalékponttal csökkenti.

Megvizsgáltuk azt is, hogy az adott forgatókönyv mellett az egyes becslőfüggvények milyen várható pontossággal (intervallumhosszal) dolgoznak. Az informatív priort tartalmazó B1 becslést leszámítva az egyes becslések intervallumhossza a 0-15 százalék közötti tartományon nagyságrendileg megegyezett, forgatókönyvtől függően 5 és 25 százalékpont között mozgott.[14]

4.2. A második szakasz eredményei

A B1 becslés megbízhatósági szintje a sokasági hibaarány binomiális eloszlása mellett akkor maximális, ha a feltételezett szuperpopulációs hibaarány megegyezik annak valós értékével. Ebben az esetben a megbízhatósági szint pontosan 95 százalék. A valós hibaaránytól távolodva csökken a becslés megbízhatósága is: minél nagyobb a kiválasztási arány, a csökkenés is annál lassabb ütemű. Ennek ellenére 200-300 elemű minták esetén is nagyon erős a prior jóságától való függés.

Itt kell szót ejtenünk arról, hogy ezt az eredményt hogyan is kell pontosan értelmezni. A bayesi becslés valójában nem csinál mást, mint a sokaságról szerzett többletinformációt a pontosság növelésére fordítja. Ez az átváltás mindaddig jól működik, amíg a becslésbe bevitt többletinformáció (megközelítőleg) helyes. Ha azonban az információ pontatlan, súlyos árat fizetünk: drasztikusan csökkeni fog becslésünk megbízhatósága (persze lehet olyan szerencsénk, hogy a konkrét sokasági realizáció mégis nekünk kedvez, de erre nem érdemes alapozni).

5. KÖVETKEZTETÉSEK

Az eddig leírtakból levonható fontos következtetések.

– A legfontosabb az M1 becslőfüggvény alkalmazhatóságáról levonható következtetés: a becslés 5-10 százalékos várható hibaaránya mellett az alkalmazásához legalább 150-200 elemű minta szükséges. Jól funkcionál az alapozó statisztika könyvekben alkalmazási feltételként javasolt $\min\{np, n(1-p)\} \geq 10$ hüvelykujjszabály, kiegészítve azzal, hogy a végesség miatt a varianciánál indokolt az $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$ korrekciós tényező használata.

[14] Elsőre meglepő lehet, hogy az elvileg fix pontosságú MxB1 becslésnek miért nem konstans a várható intervallumhossza. A jelenség magyarázata, hogy az intervallumoknak csak a nemnegatív szakaszát vesszük figyelembe.

– Az előbbi egyenes következménye, hogy a képlet „visszafelé történő alkalmazásával” számított szükséges mintaméretek sem megfelelőek, csak a nagyságrendek gyors kiszámítására alkalmasak. Ha tehát statisztikailag is megalapozott következtetéseket akarunk levonni a mintából, akkor ne használjuk az ellenőrzési szakirodalomban általánosan elterjedt $n = \frac{z^2 p(1-p)}{d^2}$ képletet a mintaméret meghatározására, ehelyett használjuk inkább az IDEA M2 eljáráson alapuló beépített mintavételi modulját.

– Kérdéses, hogy 10 százalékpont feletti becsült intervallumhossz mellett van-e egyáltalán értelme az intervallumbecslésnek, ha a hibaarány 5-10 százalék körül mozog.

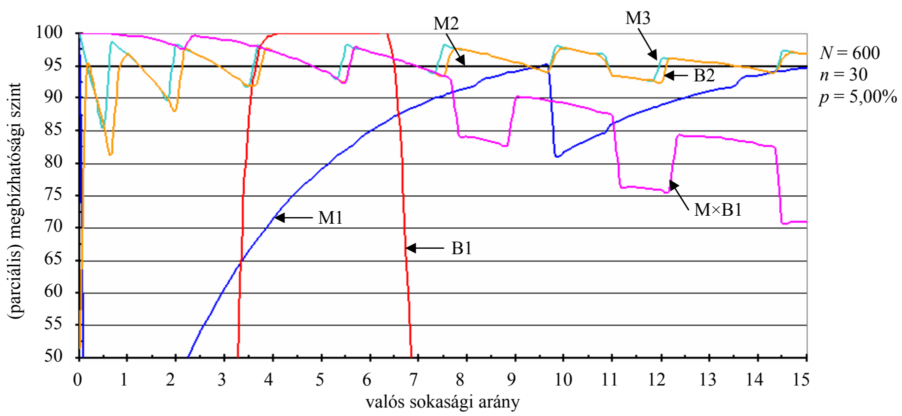
– Amennyiben többéves, részletes idősoraink vannak egy adott hiba, szabálytalanság előfordulási gyakoriságára vonatkozóan, és ennek a hibaaránynak az eloszlása időtől és/vagy vizsgált szervezettől függetlenül nagyfokú stabilitást mutat, érdemes megfontolni a bayesi intervallumbecslés használatát, ugyanis ez a becslés az elvárt megbízhatóság teljesítése mellett nagy pontosságú (rövidebb) intervallumokat szolgáltat.

*

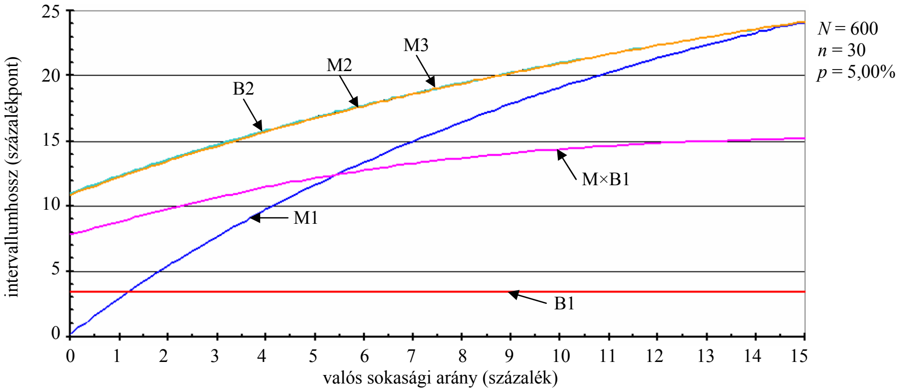
Az Állami Számvevőszék gyakorlatában a „financial audit” típusú vizsgálatoknál az IDEA szoftver támogatja a mintavételezést, így a következtetések megbízhatósága meghaladja az elvárt megbízhatósági szintet (emlékeztetőül: az IDEA nem használ randomizálást, hanem mindig a „legkonzervatívabb” becslést adja). További vizsgálat szükséges annak eldöntésére, hogy az önkormányzati szektor vizsgálatára alkalmazott kétlépcsős, tervezési-mintavételezési eljárások (1. lépcső: ellenőrizendő önkormányzat kiválasztása; 2. lépcső: az önkormányzatnál való mintavétel) milyen mértékben biztosítják, hogy az önkormányzati szektorról, mint aggregátumról levonható következtetések 95 százalékos bizonyosságúak.

FÜGGELÉK

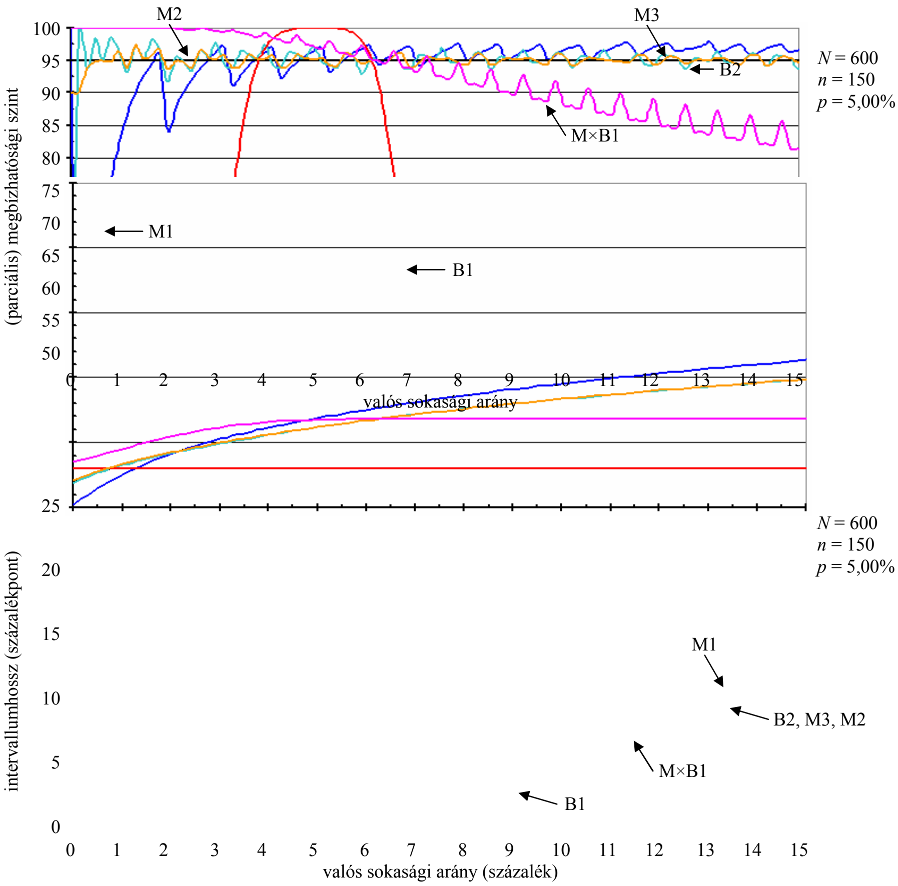
1. ábra. Az intervallumbecslési módszerek (parciális) megbízhatósága
(százalék)



2. ábra. Az intervallumbecslési módszerek várható intervallumhossza



3. ábra. Az intervallumbecslési módszerek (parciális) megbízhatósága
(százalék)



IRODALOM

CASEWARE IDEA RESEARCH DEPARTMENT [2003]: *White Papers on Attribute Sampling Technical Specification*. http://www.caseware-idea.com.

DENKINGER, G. [1990]: *Valószínűségszámítás*. Tankönyvkiadó. Budapest.

HUNYADI L. – VITA L. [2004]: *Statisztika közgazdászoknak* (harmadik kiadás). Központi Statisztikai Hivatal. Budapest.

HUNYADI L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak.* Központi Statisztikai Hivatal. Budapest.

KATZ, L. [1953]: Confidence intervals for the number showing a certain characteristic in a population when sampling is without replacement. *Journal of the American Statistical Association.* 48. évf. 256–261. old.

LEHMANN, E. L. [1959]: *Testing Statistical Hypotheses*. Wiley. New York.

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 Statistician Society.* 97. évf. 558–606. old.

WRIGHT, T. [1990]: When zero defectives appear in a sample: upper bounds on confidence coefficients of upper bounds. *The American Statistician*. 44. évf. 40–41. old.

WRIGHT, T. [1991]: *Exact confidence bounds when sampling from small finite universes*. Springer. New York.

WRIGHT, T. [1992]: A note on sampling to locate rare defectives with strong prior evidence. *Biometrika*. 79. évf. 685–691. old.

WRIGHT, T. [1997]: A simple algorithm for tighter exact upper confidence bounds with rare attributes in finite universes. *Statistics & Probability Letters*. 36. évf. 59–67. old.

SUMMARY

Since audit opinions can never be based on certainty but on “reasonable assurance”, sampling techniques play an important role in modern audit methodology. One of the audit’s objectives is to determine from small samples (<100) whether internal control systems in the audited organisations are operational. Theoretical controls (such as double signatures, sealing, etc.) have to be performed every time at a given stage of a procedure, and they are to ensure that the output is acceptable. Failing to perform a control can increase the risk of bad output, and “Attribute Sampling” is an audit technique that quantifies this risk by estimating the proportion of missing controls.

In this paper three major aspects of Attribute Sampling are addressed, however, interval estimation plays the largest role: we compare the effectiveness of six estimators. Methods for point estimation and hypothesis testing are also described, from mainstream and Bayesian view as well.