A RÉTEGZETT MINTAVÉTELRŐL

PETHŐ SZILVESZTER*

A MŰSZAKI TUDOMÁNYOK KANDIDÁTUSA

[Beérkezett: 1971. november 8-án]

A rétegzett mintavétel az egyszerű véletlen mintavételnél hatásosabb megfigyelési eljárás. Alkalmazása során a mintázandó sokaságot részsokaságokra, rétegekre kell bontani, vagy a sokaság amúgy is meglevő rétegződését kell kihasználni. Az egyes rétegekben nagyjából azonos tulajdonságú, kis szórású mintaelemek vannak, de a rétegátlagok jobban különböznek egymástól, vagyis a rétegátlagok közötti szórás nagy. Ilyen módon rendezett sokaság esetében a rögzített mintaelemszámot a rétegekre megfelelő módon (pl. az optimális elosztás segítségével) szét kell osztani és a rétegekből a mintákat véletlen módon kell kiemelni.

A tanulmány a rétegzett mintavétel paraméterei közötti legfontosabb összefüggéseket, az egyszerű torzítatlan becsléseket, valamint a különböző mintavételi módokat ismerteti. Az utolsó részben két speciális becslési eljárást közöl.

Bevezetés

A mintavételnek három fő formája ismeretes: az egyszerű véletlen, a szisztematikus és a rétegzett mintavétel.

a) *Az egyszerű véletlen mintavétel* esetében a diszkrét eloszlású sokaság paraméterei a binomiális, ill. a hipergeometrikus eloszlások segítségével adhatók meg attól függően, hogy a mintavétel visszatevéssel vagy visszatevés nélkül történt; folytonos eloszlású sokaság esetén a mintavétel megszervezése, ill. a paraméterek becslése rendszerint a normális eloszlás segítségével történik.

b) *A szisztematikus mintavétel* esetében a természet vagy az emberi beavatkozás által rendezett sokaság elemeit meghatározott terv szerint mintázzák. Szisztematikus mintavétel a rendszerint folytonos valószínűségi változójú természeti jelenségek — hőmérsékletmérés, folyó vízállásának megállapítása, hasznosítható ásványtelepülés szabályos háló szerinti megkutatása stb. — vagy olyan ipari termékek, mint a gumiszalagon szállított vagy vasúti kocsiban levő szemcsés anyagok, szabályos időközökben, ill. térbeli távolságokban való megfigyelése. Diszkrét valószínűségi változójú sokaság esetében a szisztematikus mintavétel érdekében az elemek sorszámmal látandók el. A mintából kivett elemek egymás után következő sorszámainak különbsége a mintavételi arány reciprokával (N/n) egyezik meg úgy, hogy a legkisebb sorszámú mintaelem megállapítása véletlen módon történik. Ha az időben vagy

* Prof. Dr. Pethő Szilveszter, Miskolc — Egyetemváros.

térben egymás után következő mintaelemek jellemző értékei között függőség van, akkor a megfigyelt paraméterek torzítatlan becslése csak ezzel a mintavételi móddal biztosítható.

c) *A rétegzett mintavételt* a statisztikai hivatalok, a közvéleménykutató intézetek alkalmazták először, mert segítségével az egyszerű véletlen mintavételnél hatásosabb becslés érhető el. Ha egy tőkés társadalomban pl. az átlagos jövedelmet és ennek szórását mikrocenzusok segítségével rendszeresen meg akarják állapítani, akkor a mikrocenzusok alkalmával a tőkés társadalom meglevő osztálytagozódását célszerű alapul venni. Az egyes társadalmi rétegekben nagyjából azonos jövedelmű egyedek vannak, de a rétegek átlagos jövedelmében igen nagyok a különbségek. Az összes megfigyelések száma, n, rögzítve van, amelyet az egyes társadalmi csoportokra a rétegzett mintavétel elvei szerint előre felosztanak és a rétegeken belül egyszerű véletlen mintavételt hajtanak végre. Belátható és matematikailag bizonyítható, hogy ilyen mintavétellel a megfigyelt paraméterek, a felhozott példa szerint az átlagos jövedelem, pontosabban figyelhetők meg, mint egyszerű véletlen mintavétel esetén. Ilyen természetes rétegződés a mintavétel során nem mindig figyelhető meg, ilyenkor mesterségesen törekszenek a sokaságon belül olyan rétegek kialakítására, melynek elemei a megfigyelt jellemző értékében nem nagyon különböznek egymástól. A rétegekre való bontás akkor a legtökéletesebb, ha az a becsléses illeszkedésvizsgálathoz szükséges értékközök szerinti rendezésnek felel meg, mert ilyen esetben lehet a leghatásosabb becslést elérni. Szemnagyságelemzési vizsgálatokban, szenek, ércek alapgörbéinek felvételéhez általában ilyen tökéletes rétegeket hoznak létre: a rétegen belül nem odavaló tulajdonságú szemcse nincsen. Ha a rétegeket véletlen módon alakítják ki, akkor a rétegzett mintavétel bevezetésével a véletlen mintavétellel szemben a becslés hatásossága nem fokozható.

A következőkben a rétegzett mintavétel legfontosabb tulajdonságai kerülnek megvizsgálásra, és megismerhetők lesznek a különböző mintavételi és becslési módszerek.

A rétegzett mintavétel paraméterei közötti összefüggések áttekintése

1. A sokaság paraméterei közötti összefüggések áttekintése

A sokaság elemeinek száma N — tehát véges sokaságról van szó — a rétegek száma L, a h-odik rétegben levő elemek száma N_h. A h-odik réteg i-edik elemének jellemző értéke X_{hi}: az első index (h) tehát a rétegszámot, a második (i) pedig az elemszámot jelenti, vagyis a mintaelemek, legalábbis gondolatban, sorszámozva vannak. A h-odik réteg totálja X_h, átlaga $\bar{X}_h$, az egész

sokaság totálja és átlaga X ill. $\bar{X}$. A jelölések áttekintése után a következő összefüggések azonnal beláthatók:

$$N = \sum_{h=1}^{L} N_h, \tag{1.1}$$

$$X_h = \sum_{i=1}^{N_h} X_{hi}, \qquad \bar{X}_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h}, \tag{1.2}$$

$$X = \sum_{h=1}^{L} X_h = \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi} = \sum_{i=1}^{N} X_i, \tag{1.3}$$

$$\bar{X} = \frac{\sum_{h=1}^{L} X_h}{L} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi}}{\sum_{h=1}^{L} N_h} = \frac{\sum_{i=1}^{N} N_i}{N}. \tag{1.4}$$

A h-odik réteg „korrigált empirikus szórásnégyzete" (S_h^2) és empirikus szórás négyzete (σ_h^2)

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2}{N_h - 1}, \qquad \sigma_h^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2}{N_h}. \tag{1.5}$$

A relatív szórás (V_h) és a számtani átlag szórásnégyzete ($D^2(\bar{X}_h)$):

$$V_h = \frac{S_h}{X_h}, \qquad D^2(\bar{X}_h) = \frac{S_h^2}{N_h}. \tag{1.6}$$

(1.5) és (1.6) alatti szórásnégyzetek az egész sokaságra vonatkozóan:

$$S^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{X})^2}{N - 1}, \qquad \sigma^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{X})^2}{N}. \tag{1.7}$$

$$V = \frac{S}{\bar{X}}, \qquad D^2(\bar{X}) = \frac{S^2}{N}. \tag{1.8}$$

Célszerű definiálni még a rétegátlagok szórásnégyzetét (σ_b^2) és a rétegátlagokra vonatkozó szórásnégyzetet is (σ_w^2):

$$\sigma_b^2 = \frac{\sum_{h=1}^{L} N_h (\bar{X}_h - \bar{X})^2}{N}, \qquad \sigma_w^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2}{N}. \tag{1.9}$$

Fontos összefüggés mutatható ki a sokaság elemei szórásnégyzete (σ^2), továbbá a rétegátlagok szórásnégyzete (σ_b^2) és a rétegátlagokra vonatkozó szórásnégyzet (σ_w^2) között. Az összefüggés a következő:

$$\sigma^2 = \sigma_b^2 + \sigma_w^2 . \tag{1.10}$$

1.10 bizonyítása σ^2-nek (1.7) alapján való felírásával történhetik:

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi} - \bar{X})^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi} - \bar{X}_h + \bar{X}_h - \bar{X})^2 +$$

$$+ \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi} - \bar{X}_h)^2 + \frac{2}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi} - \bar{X}_h)(\bar{X}_h - \bar{X}) +$$

$$+ \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\bar{X}_h - \bar{X})^2 .$$

A második tagban szereplő

$$\sum_{i=1}^{N_h}(X_{hi} - \bar{X}_h) = 0 ,$$

és a harmadik tag

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\bar{X}_h - \bar{X})^2 = \sum_{h=1}^{L} N_h(\bar{X}_h - \bar{X})^2 .$$

Ezek, valamint 1.9 figyelembevételével az 1.10. tétel igazolódott.

2. *A minta paraméterei közötti összefüggések és a sokaság paramétereinek egyszerű torzítatlan becslései*

A mintabeli paramétereket kisbetűkkel célszerű jelölni: a mintaelemszám n, a h-odik rétegből vett mintaelemszám n_h; a h-odik réteg i-edig mintaelemének jellemző értéke x_{hi}, a réteg totálja (n_h számú mintaelem alapján) x_h, átlaga $\bar{x}_h$; az egész sokaság totálja (n elemből kiszámítva) x és az átlag $\bar{x}$.

Az (1.1), (1.2), (1.3) és (1.4) összefüggések a mintabeli paraméterekre is érvényesek, de a mintabeli paraméterekre felírható összefüggésekben természetesen kisbetűk szerepelnek.

Az $\bar{X}_h$ átlag becslése $\bar{x}_h$-val történik, de az X_h és X totál, továbbá az $\bar{X}$ átlag becslése az egész sokaság és a rétegek elemszámának ismeretében

$$\left(N = \sum_{h=1}^{L} N_h\right),$$

— a becslési értékeket $\hat{x}_h$-val, $\hat{x}$-szel ill. $\hat{\bar{x}}$-sal jelölve — a mintabeli adatokkal a következőképpen hajtandó végre:

$$\hat{x}_h = N_h \bar{x}_h = N_h \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h}, \qquad \hat{\bar{x}}_h = \bar{x}_h, \tag{2.1}$$

$$\hat{x} = \sum_{h=1}^{L} \hat{x}_h = \sum_{h=1}^{L} N_h \bar{x}_h, \qquad \hat{\bar{x}} = \frac{\sum_{h=1}^{L} N_h \bar{x}_h}{N}. \tag{2.2}$$

$\hat{\bar{x}}$ természetesen különbözik $\bar{x}$-től, melynek kiszámítása

$$\frac{\sum_{h=1}^{L} n_h \bar{x}_h}{n}$$

szerint történik. A (2.1)-ben és (2.2)-ben szereplő becsléseket egyszerű *torzítatlan becsléseknek* szokás nevezni.

Az (1.5)-ben és az (1.7)-ben szereplő szórásnégyzetek torzítatlan becslései s_h^2 és s^2, amelyek kiszámítása az

$$s_h^2 = \frac{\sum_{h=1}^{n_h} (x_{hi} - \bar{x}_h)^2}{n_h - 1} \quad \text{és} \quad s^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} (x_{hi} - \bar{x})^2}{n - 1} \tag{2.3}$$

képletek szerint történik. A relatív szórások (v_h és v) és a számtani átlagok szórásnégyzetei ($D^2(\bar{x}_h)$ és $D^2(\bar{x})$) (1.6) és (1.8) alatti összefüggések alapján, a (2.3)-ban szereplő szórásnégyzetek figyelembevételével írhatók fel:

$$v_h = \frac{s_h}{\bar{x}_h} \quad \text{és} \quad v = \frac{s}{x}, \tag{2.4}$$

illetve

$$D^2(\bar{x}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} \quad \text{és} \quad D^2(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}. \tag{2.5}$$

2.5.-ben szerepelnek az n_h/N_h illetve n/N mintavételi arányok, amelyek f_h-val illetve f-el jelölhetők.

A (2.1)-ben és a (2.2)-ben szereplő egyszerű torzítatlan becslések szórásnégyzetei — $D^2(\hat{x}_h)$ és $D^2(\hat{x})$ a totálok, $D^2(\hat{\bar{x}}_h)$ és $D^2(\hat{\bar{x}})$ a számtani átlagok szórásnégyzetei — a következők:

$$D^2(\hat{x}_h) = N_h^2 (1 - f_h) \frac{s_h^2}{n_h}, \tag{2.6}$$

$$D^2(\hat{\bar{x}}_h) = D^2(\bar{x}_h) = (1 - f_h) \frac{s_h^2}{n_h}. \tag{2.7}$$

Az egész sokaságra vonatkozó szórásnégyzetek kiszámításaa (2.6) szerinti szórásnégyzet megfelelő összegezése és súlyozása útján történik:

$$D^2(\hat{x}) = \sum_{h=1}^{L} N_h^2(1 - f_h)\frac{s_h^2}{n_h}, \tag{2.8}$$

$$D^2(\hat{\bar{x}}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2(1 - f_h)\frac{s_h^2}{n_h}. \tag{2.9}$$

3. *A mintavételi módok ismertetése*

A rétegzett sokaságból történő mintavételi módok tárgyalásakor n, az összes mintaelemszám rögzítve van, az egyes mintavételi módok az n mintaelemszámnak az egyes rétegekre történő elosztásában különböznek egymástól. A következő mintavételi módok ismeretesek:

a) *A mintaelemszám véletlen elosztása.* Az n mintaelemszámnak a rétegekre való véletlen elosztásakor a becslési értékek — a totálok és átlagok — a (2.1) és (2.2) és ezek szórásnégyzetei a (2.6), (2.7), (2.8) és (2.9) egyenletekkel számíthatók. Az egész sokaság totáljának és átlagának szórásnégyzetei

$$D^2(\hat{x}) \text{ és } D^2(\hat{\bar{x}})$$

ennél a mintavételi módnál a legnagyobbak, illetve megegyeznek az egyszerű véletlen mintavétel megfelelő szórásnégyzeteivel.

b) *Az arányos elosztás.* Ebben a h-odik réteg n_{ha} mintaelemszáma,

$$n_{ha} = \frac{N_h}{N} n = f N_h \qquad (h = 1, 2, \ldots, L) \tag{3.1}$$

vagyis a mintaelemszám a rétegek N_h elemszámával arányos. A sokaság átlaga ennél a becslési módnál (2.2 alapján 3.1 felhasználásával és $\hat{\bar{x}}_a$-sal jelölve)

$$\hat{\bar{x}}_a = \frac{\sum_{h=1}^{L} N_h \frac{\sum_{i=1}^{n_{ha}} x_{hi}}{N_h}}{N} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_{ha}} x_{hi}}{n} = \bar{x}, \tag{3.2}$$

vagyis megegyezik $\bar{x}$-sal, amelynek kiszámításához a rétegek elemszámának (N_h illetve N) ismerete nem szükséges. A rétegátlag szórásnégyzete ($D^2(\hat{\bar{x}}_{ha})$, (2.7)-ből kiindulva):

$$D^2(\hat{\bar{x}}_{ha}) = \frac{1 - f}{f N_h} s_h^2. \tag{3.3}$$

A sokaság átlagának szórásnégyzete ($D^2(\bar{x}_a)$, (2.9)-ből kiindulva):

$$D^2(\bar{x}_a) = \frac{1 - f}{nN}\sum_{h=1}^{L} N_h s_h^2. \tag{3.4}$$

Mivel két torzítatlan becslés közül az a hatásosabb, amelyiknek a szórásnégyzete kisebb, az arányos elosztásnak a véletlen elosztással szembeni előnye (3.4) és (2.9) sokaságátlag szórásnégyzetek összehasonlításával ítélhető meg. A $D^2(\hat{\bar{x}}_a)$ és $D^2(\bar{x})$ szórásnégyzetek a következő formákban is felírhatók:

$$D^2(\hat{\bar{x}}_a) = \frac{1}{nN} \sum_{h=1}^{L} N_h s_h^2 - \frac{1}{N^2} \sum_{h=1}^{L} N_h s_h^2, \tag{3.5}$$

$$D^2(\bar{x}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2}{n_h} s_h^2 - \frac{1}{N^2} \sum_{h=1}^{L} N_h s_h^2. \tag{3.6}$$

Képezve a $D^2(\bar{x}) - D^2(\hat{\bar{x}}_a)$ különbséget,

$$D^2(\bar{x}) - D^2(\hat{\bar{x}}_a) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2}{n_h} s_h^2 - \frac{1}{nN} \sum_{h=1}^{L} N_h s_h^2 =$$

$$= \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2}{n_h} s_h^2 - \frac{1}{N^2} \sum_{h=1}^{L} \frac{NN_h}{n} s_h^2 \geq 0 \tag{3.7}$$

szerint, azonnal belátható, hogy az adott feltételekkel

$$N = \sum_{h=1}^{L} N_h, \qquad n = \sum_{h=1}^{L} n_h \quad \text{és} \quad N_h \geq n_h. \quad (h = 1, 2, \ldots, L),$$

a különbség általában 0-nál nagyobb, vagyis az arányos elosztással a véletlen elosztással szemben a sokaságról megbízhatóbb ismeretek szerezhetők.

c) *Az optimális elosztás:* Az n mintaelemszámnak a sokaság rétegeire való olyan elosztása is létezik, amelynél a sokaság totáljának, illetve átlaga szórásnégyzeteinek minimuma van — ez az optimális elosztás —, amellyel tehát a legmegbízhatóbb ismeretek szerezhetők a sokaság paramétereiről.

A mintaelemszám optimális elosztásához a (2.9) szerinti $D^2(\bar{x})$ szórásnégyzet szélső értékét kell megkeresni a

$$\sum_{h=1}^{L} n_h = n$$

feltétel megtartásával. Lagrange módszere szerint tehát írható:

$$F = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} + k \left(\sum_{h=1}^{L} n_h - n\right). \tag{3.8}$$

Az összefüggés jobb oldalának első tagja a $D^2(\bar{x})$ szórásnégyzet, mely minimumának megkeresése a cél; a második tag a mellékfeltételt fejezi ki; a mellékfeltételben szereplő k a Lagrange-féle szorzó. n_h optimális értékének meghatározásához F-nek n_h szerinti részleges deriváltját 0-val tesszük egyenlővé, és ebből:

$$n_h = \frac{N_h s_h}{N\sqrt{k}}, \qquad (h = 1, 2, \ldots, L), \tag{3.9}$$

továbbá

$$\sum_{h=1}^{n} n_h = n = \frac{\sum_{h=1}^{L} N_h s_h}{N\sqrt{k}} . \tag{3.10}$$

$\sqrt{k}$-t (3.10)-ből kifejezve és (3.9)-be behelyettesítve, az optimális elosztásra vonatkozó mintaelemszám megadható:

$$n_{h\,\mathrm{opt}} = \frac{N_h s_h}{\sum_{h=1}^{L} N_h s_h} , \qquad (h = 1, 2, \ldots, L), \tag{3.11}$$

vagyis az optimális elemszám a réteg elemszámának (N_h) és korrigált empirikus szórásának (s_h) szorzatával arányos. Ha az egyes rétegek elemszáma azonos — $N_h = N/L$ —, akkor az optimális elemszám értéke

$$n_{h\,\mathrm{opt}} = \frac{s_h}{\sum_{h=1}^{L} s_h} n , \qquad (h = 1, 2, \ldots, L), \tag{3.12}$$

vagyis az említett határesetben ($N_h = N/L$) az optimális elemszám a réteg korrigált empirikus szórásával arányos.

$n_{h\,\mathrm{opt}}$ figyelembevételével $\bar{\hat{x}}_{\mathrm{opt}}$, $D^2(\hat{\bar{x}}_{h\,\mathrm{opt}})$, $D^2(\hat{x}_{h\,\mathrm{opt}})$ és $D^2(\bar{\hat{x}}_{\mathrm{opt}})$ a következőképpen adható meg:

$$\bar{\hat{x}}_{\mathrm{opt}} = \frac{\sum_{h=1}^{L} N_h \frac{\sum_{i=1}^{n_{h\,\mathrm{opt}}} x_{hi}}{n_{h\,\mathrm{opt}}}}{N} , \tag{3.13}$$

$$D^2(\hat{\bar{x}}_{h\,\mathrm{opt}}) = \frac{N_h - n_{h\,\mathrm{opt}}}{N_h n_{h\,\mathrm{opt}}} s_h^2 = \frac{\sum_{h=1}^{L} N_h s_h - s_h n}{N_h n} s_h , \tag{3.14}$$

$$D^2(\hat{x}_{h\,\mathrm{opt}}) = N_h^2 \frac{N_h - n_{h\,\mathrm{opt}}}{N_h n_{h\,\mathrm{opt}}} s_h^2 . \tag{3.15}$$

(3.11)-et (3.15)-be is behelyettesítve, a réteg totális szórásnégyzete:

$$D^2(\hat{x}_{h\,\mathrm{opt}}) = \frac{N_h s_h \sum_{h=1}^{L} N_h s_h}{n} - N_h s_h^2 , \tag{3.16}$$

és

$$D^2(\bar{\hat{x}}_{\mathrm{opt}}) = \frac{1}{N^2} \sum_{h=1}^{L} D^2(\hat{x}_{h\,\mathrm{opt}}) = \frac{1}{N^2} \left[\frac{\left(\sum_{h=1}^{L} N_h s_h \right)^2}{n} - \sum_{h=1}^{L} N_h s_h^2 \right]. \tag{3.17}$$

Az optimális elosztásnak az arányos elosztással szembeni előnye a

$$D^2(\hat{\bar{x}}_a) - D^2(\hat{\bar{x}}_{\text{opt}})$$

különbség kiszámításából tűnik ki. (3.4) illetve (3.5) és (3.17) felhasználásával

$$D^2(\hat{\bar{x}}_a) - D^2(\hat{\bar{x}}_{\text{opt}}) = \frac{\sum_{h=1}^{L} N_h s_h^2}{N_h} - \frac{\left(\sum_{h=1}^{L} N_h s_h\right)^2}{N^2 n} =$$

$$= \frac{1}{n}\left[\frac{1}{N}\sum_{h=1}^{L} N_h s_h^2 - \frac{1}{N^2}\left(\sum_{h=1}^{L} N_h s_h\right)^2\right]. \tag{3.18}$$

(3.18)-at megvizsgálva, látható, hogy abban a szögletes zárójelen belül a rétegszórások szórásnégyzete szerepel. (A szögletes zárójelben az első tag a rétegelemszámokkal súlyozott szórásnégyzet, a negatív jel után következő második tag pedig a súlyozott szórásátlag négyzete.) (3.18) csak abban az esetben 0, ha az s_h $(h = 1, 2, \ldots, L)$ szórások egyenlők; ilyenkor a szóban forgó két becslési módszer azonos értékű, minden más esetben az optimális mintaelosztás az előnyösebb.

Ha nem az összes mintaelemszám, hanem a mintavétel összes C költsége van rögzítve és mindegyik rétegben egy-egy mintaelem vételének költsége, C_h a változó (C_h tehát a h-odik réteg fajlagos mintavételi költsége), a mintaelemszám optimális elosztása ebben az esetben is elvégezhető a

$$C = \sum_{h=1}^{L} C_h n_h \tag{3.19}$$

mellékfeltétel figyelembevételével. Az a függvény, amelynek Lagrange módszerével a szélső értékét megkeresve,

$$F_c = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} + k_c\left(\sum_{h=1}^{L} C_h n_h - C\right) \tag{3.20}$$

(3.20) függvény megoldása a (3.8) függvény megoldásához hasonlóan történik, és a rétegek mintaelemszámára az

$$n_{hc} = \frac{N_h s_h/\sqrt{C_h}}{\sum_{h=1}^{L} N_h s_h/\sqrt{C_h}}\, n_c, \qquad (h = 1, 2, \ldots, L) \tag{3.21}$$

összefüggés adható meg. (Az egyes rétegekből az összefüggéssel kiszámítható mintaelemszámot véve, az előre rögzített C költséghez tartozó minimális szórású sokaságátlag megadható). (3.21)-ben szerepel n_c, amelyet szintén meg kell adni, hogy a rétegek mintaelemszáma kiszámítható legyen. Ezért a (3.21) egyenletet be kell helyettesíteni (3.19)-be, amelyben C, az összes költség

ismert és így az egyenletben csak egy ismeretlen van, n_c, az összes mintaelemszám:

$$n_c = \frac{\sum_{h=1}^{L} N_h s_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h s_h / \sqrt{C_h}} C . \tag{3.22}$$

Numerikus feladat megoldásakor C ismeretében először n_c számítandó ki (3.22)-vel, majd n_{hc} (3.21)-gyel. Az egyes rétegekből a (3.22) ill. a (3.21) egyenletekkel kiszámítható mintaelemszámot véve, az előre rögzített C költséghez tartozó minimális szórású sokaságátlag megadható.

A mintavétel költségei minimálhatók olyan kikötéssel is, hogy a sokaságátlag szórása előre meghatározott E érték legyen, vagyis az

$$E^2 = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} \tag{3.23}$$

egyenlőség biztosításával az összes költség (C),

$$C = \sum_{h=1}^{L} C_h n_h \tag{3.24}$$

szélső értékének megkeresése a feladat. Mivel a mellékfeltételt a (3.23) sorszámú egyenlet tartalmazza, azért az a függvény, amelynek megoldása szintén Lagrange módszerével történik,

$$F_E = \sum_{h=1}^{L} C_h n_h + k_E \left[\frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} - E^2 \right]. \tag{3.25}$$

(3.25)-ből a rétegek mintaelemszámára (3.21)-gyel azonos alakú összefüggés adódik:

$$n_{hE} = \frac{N_h s_h / \sqrt{C_h}}{\sum_{h=1}^{L} (N_h s_h / C_h)} n_E . \tag{3.26}$$

A két képlet közötti különbség az összes mintaelemszám értékében van, általában $n_c \neq n_E$. Az n_E meghatározása érdekében n_{hE} a (3.24) egyenletbe helyettesítendő n_h helyére és a megfelelő műveletek elvégzése után az összes mintaelemszám

$$n_E = \frac{\sum_{h=1}^{L} \left(N_h s_h \sqrt{C_h} \sum_{h=1}^{L} N_h s_h / C_h \right)}{E^2 N^2 + \sum_{h=1}^{L} N_h s_h^2} . \tag{3.27}$$

Gyakorlati feladatok megoldásakor — E ismert — először n_E számítandó ki (3.27)-tel, majd n_{hE} (3.26)-tal. Az egyes rétegekből a (3.27)-tel ill. (3.26)-tal

számítható mintaelemszámot véve, az előre rögzített E pontossághoz (sokaságátlag szóráshoz) tartozó minimális költségű sokaságátlag megadható.

d) *Azonos megbízhatósági szinttel és megbízhatósági intervallummal történő mintaelosztás.* Az azonos megbízhatósági szinttel, illetve azonos megbízhatósági együtthatóval (u_p) és az azonos megbízhatósági intervallummal (2ε hosszúságú) történő mintanagyság meghatározásnak különösen a gazdasági és társadalmi jelenségek megfigyelésekor van jelentősége. Bérek, adók megállapításakor és általában a jövedelmek megállapításával kapcsolatos megfigyelések során jogos az az igény, hogy a megfelelő paramétereket azonos megbízhatósági szinttel, tehát azonos megbízhatósági együtthatóval (u_p) és megbízhatósági intervallummal (ε) állapítsák meg.

Ennél a mintaeloszlásnál célszerű előbb a végtelen sokaság esetét tárgyalni.

A végtelennek tekintett statisztikai sokaság h rétegéből vett mintaelemszám, mint ismeretes,

$$n_h \geq \frac{u_p^2}{\varepsilon^2} s_h^2 . \tag{3.28}$$

Mivel

$$\sum_{h=1}^{L} n_h = n$$

rögzített, ezért az összes mintaelemszám

$$n = \left(\frac{u_p}{\varepsilon}\right)^2 \sum_{h=1}^{L} s_h^2 . \tag{3.29}$$

(3.29)-ből $(u_p/\varepsilon)^2$-et kifejezve, majd (3.28)-ba behelyettesítve és (3.29)-hez hasonlóan ismét csak az egyenlőséget tartva meg, a rétegek mintaelemszáma

$$n_{hu_p} = \frac{s_h^2}{\sum_{h=1}^{L} s_h^2} n , \qquad (h = 1, 2, \ldots, L) . \tag{3.30}$$

Az összefüggésből látható, hogy a rétegek mintavételi száma s_h^2 korrigált empirikus szórásnégyzetükkel arányos.

A h-adik réteg és az egész sokaság átlagának szórásnégyzete (3.30) felhasználásával

$$D^2(x_{hu_p}) = \frac{s_h^2}{n_{hu_p}} = \frac{\sum_{h=1}^{L} s_h^2}{n} = \left(\frac{\varepsilon}{u_p}\right)^2 , \qquad (h = 1, 2, \ldots, L) \tag{3.31}$$

$$D^2(x_{u_p}) = \frac{1}{N^2} \left(\frac{\varepsilon}{u_p}\right)^2 \sum_{h=1}^{L} N_h^2 . \tag{3.32}$$

Véges sokaság esetén az n_{hup} mintanagyság meghatározása érdekében a következőképpen célszerű eljárni. A h-adik réteg mintaelemszáma, általános formában

$$n_h \geq \frac{ks_h^2}{1+\frac{ks_h^2}{N_h}}, \qquad k=\left(\frac{u_p}{\varepsilon}\right)^2, \tag{3.33}$$

amely csak az egyenlőség megtartásával

$$n_h + k\frac{n_h s_h^2}{N_h} = ks_h^2$$

szerint alakítható át. Áttérve az egyes rétegek mintaelemszámának összegezésére és figyelembevéve, hogy

$$\sum_{h=1}^{L} n_h = n \quad \text{és} \quad n_h/N_h = f_h,$$

$$k = \frac{n}{\sum_{h=1}^{L} s_h^2 - \sum_{h=1}^{L} f_h^2 s_h^2}, \tag{3.34}$$

amelyet (3.33)-ba helyettesítve, a megfelelő átalakítás után

$$n_h = \frac{N_h s_h^2}{N_h\left(\sum_{h=1}^{L} s_h^2 - \sum_{h=1}^{L} f_h s_h^2\right) + s_h^2 n}\, n. \qquad (h=1,2,\ldots,L) \tag{3.35}$$

Az egyenlet megoldása próbálgatással, pl. logarléc segítségével néhány lépésben elvégezhető. A problémát a nevezőben szereplő $f_h = n_h/N_h$ kifejezés okozza, amelyet első lépésben 0-nak célszerű felvenni. ([7] és [8] -ban más megoldások is megtalálhatók!) (3.35) alapján egyébként megállapítható, hogy a rétegek mintaelemszáma a réteg elemszámától és korrigált empirikus szórásnégyzetétől függ. A (3.31)-hez és (3.32)-höz hasonlóan az átlagok szórásnégyzete véges sokaság esetén is megadható.

4. Különleges becslési módok

Különleges becslési módokról a rétegzett mintavétellel kapcsolatban akkor van szó, ha az egész sokaság totálja (X) vagy átlaga ($\bar{X}$) ismert, vagy ezen értékek bizonyos hibával megmérhetők. A mérési eredmények jelölése ez utóbbi esetben O illetve O, szórásnégyzeteiké $N^2D^2(\bar{O})$ és $D^2(\bar{O})$. Ez az utóbbi eset fordul elő pl. mosási görbék felvétele esetén akkor, amikor nemcsak a frakciók (rétegek), hanem a kísérletbe bevont összes anyag tulajdonságát is meghatározzák. Ilyen esetben a rétegekből vett minták alapján számított és

a közvetlen méréssel kapott sokaságátlag vagy totális általában nem egyezik meg, de legvalószínűbb értéke, $\hat{\bar{x}}$, (vagy $\hat{\bar{x}}$) kiszámítható:

$$\hat{\bar{x}} = \frac{\frac{1}{\bar{D}^2(\hat{\bar{x}})}\hat{x} + \frac{1}{D^2(\bar{O})}\bar{O}}{\frac{1}{D^2(\hat{\bar{x}})} + \frac{1}{D^2(\bar{O})}}. \tag{4.1}$$

A különleges becslési módokban arról van szó, hogy azt az r értéket, amely

$$r = |X - x| \; ; \; r = |\hat{\bar{x}} - \hat{x}| \tag{4.2}$$

szerint van definiálva, és *zárlati hibának* nevezhető, miképpen kell az egyes rétegekre elosztani.

a) Azonos megbízhatósági határral való becslés

A rétegek egyszerű torzítatlan totálisai ($N_h\bar{x}_h$) összegének és az egész sokaság totálisának (X) ill. totálisa legvalószínűbb értékének ($N\hat{\bar{x}}$) különbsége adja r-t, a zárlati hibát:

$$\sum_{h=1}^{L} N_h\bar{x}_h - X' = r. \tag{4.3}$$

Ebben az összefüggésben, amelyet a kiegyenlítő számításban *ellentmondási egyenletnek* neveznek, X' jelenti X-et, illetve $N\bar{x}$-et. Ha $\bar{x}_h$ legvalószínűbb értéke $\bar{x}_h$, akkor a feltételi egyenlet a (4.3) ellentmondási egyenlet alapján a következőképpen alakul:

$$\sum_{h=1}^{L} N_h\bar{\bar{x}}_h - X' = 0. \tag{4.4}$$

A rétegátlagok legvalószínűbb értéke

$$\bar{\bar{x}}_h = \bar{x}_h \pm u_p D(\bar{x}_h) \tag{4.5}$$

szerint definiálható, ami annyit jelent, hogy a méréssel megállapított $\bar{x}_h$ átlag az r zárlati hiba eltüntetése érdekében az u_p megbízhatósági együttható és a $D(\bar{x}_h)$ szórás szorzatával arányosan változtatandó meg. Mivel az u_p szám mindegyik rétegre azonos, ezért az $\bar{\bar{x}}_h$ értékek ($h = 1, 2, \ldots, L$) azokat a megbízhatósági határokat jelentik, amely határok által kijelölt intervallumban való bekerülés valószínűsége egyforma. Ezért a szóban forgó becslési mód azonos megbízhatósági határokkal való becslésnek nevezhető.

(4.5)-nek (4.4)-be való helyettesítésével

$$\sum_{h=1}^{L} N_h\bar{x}_h - X' \pm u_p \sum_{h=1}^{L} N_h D(\bar{x}_h) = 0 \tag{4.6}$$

összefüggés adódik, amelyből az u_p megbízhatósági együttható

$$u_p = -\frac{\sum_{h=1}^{L} N_h \bar{x}_h - X'}{\sum_{h=1}^{L} N_h D(\bar{x}_h)} = -\frac{r}{\sum_{h=1}^{L} N_h D(\bar{x}_h)} . \tag{4.7}$$

Ha az u_p megbízhatósági együttható 3-nál nagyobb (sok szerző szerint 2-nél), akkor valószínű, hogy a méréssorozat során szisztematikus hiba is előfordult. A (4.7) egyenlettel meghatározott u_p tehát a mérési eredményekből számítható és a rétegátlag legvalószínűbb érték:

$$\bar{\bar{x}}_h = \bar{x}_h \mp \frac{\sum_{h=1}^{L} N_h \bar{x}_h - X'}{\sum_{h=1}^{L} N_h D(\bar{x}_h)} D(\bar{x}_h) . \tag{4.8}$$

Ennek alapján a többi paraméter, így pl. a $N_h \bar{\bar{x}}_h$ totális is, számítható.

Ha a rétegátlagok szórásai ($D(\bar{x}_h)$) egyformák, akkor a (4.8) egyenlet

$$\bar{\bar{x}}_h = \bar{x}_h \mp \frac{r}{N} \tag{4.9}$$

szerint, a rétegtotális pedig

$$N_h \bar{\bar{x}}_h = x_h \mp \frac{N_h}{N} r \tag{4.10}$$

szerint alakul. Ez utóbbi egyenlet szerint a rétegtotális legvalószínűbb értékének meghatározásakor a zárlati hiba a rétegek elemszámával arányosan osztandó szét.

b) A legkisebb négyzetek módszerével való becslés

Annak érdekében, hogy a legkisebb négyzetek módszerével a zárlati hiba az egyes rétegekre eloszthatóvá legyen, az x_h totálisok, illetve a totálisok $D^2(\hat{x}_h)$ szórásnégyzete veendő figyelembe. Az egyes rétegek totálisainak javítása $\Delta\hat{x}_h$; ezek összege ($h = 1, 2, \ldots, L$) a zárlati hibát adja. Mivel a legkisebb négyzetek módszerénél a javításoknak a szórással súlyozott négyzetösszege minimális, ezért az a függvény, amelynek szélső értéke Lagrange módszerével keresendő meg a következő:

$$F = \sum_{h=1}^{L} \left[\frac{\Delta\hat{x}_h}{D(\hat{x}_h)} \right]^2 - 2k \left[\sum_{h=1}^{L} \Delta\hat{x}_h - r \right] = 0 . \tag{4.11}$$

A függvényben a mellékfeltételt kifejező rész szorzója — a javítások összege a zárlati hibát adja — a könnyebb megoldás érdekében $-2k$. A megoldást a szokásos módon megkeresve, a totálisok javítására az írható, hogy

$$\Delta\hat{x}_h = \frac{D^2(\hat{x}_h)}{\sum_{h=1}^{L} D^2(\hat{x}_h)} r, \qquad (h = 1, 2, \ldots, L). \tag{4.12}$$

(4.12)-ből megállapíthatóan a javítások a rétegtotálisok szórásnégyzetével arányosak, az utóbbiak a (2.6) egyenlet szerint számíthatók ki.

A legkisebb négyzetek módszerével javított totális

$$\hat{\hat{x}}_h = \hat{x}_h + \Delta\hat{x}_h . \tag{4.13}$$

Ha a rétegtotálisok szórásnégyzetei egyformák, akkor a (4.12) összefüggés

$$\Delta\hat{x}_h = \frac{r}{L} \tag{4.14}$$

szerint alakul, vagyis ilyen esetekben a javítások azonosak.

*

Jelen tanulmány alapján a [3] alatti szakkönyv, továbbá a [4—8] és [10] alatti szakcikkek alkotják. A [11] és [12] alatti munkák speciális problémákat ölelnek fel, és nem is kerültek ezen tanulmányba bedolgozásra. Az idézett tanulmányoknak a jelenlegivel való összevetéséből kitűnik, hogy ebben a munkában tartalmilag az eddigiekhez képest új megállapítások nincsenek, csupán a tanulmány felépítése újszerű.

IRODALOM

1. Prékopa A.: Valószínűségelmélet műszaki alkalmazásokkal. Műszaki Könyvkiadó, Budapest 1962
2. Prékopa A.—Éltető Ö.: Matematikai jegyzetek IV. Matematikai statisztika. Kézirat, 1961
3. Moriss, H.—H.,—William, N.,—Hurwitz,—William, G.—Madow: Sample Survey Methods and Theory, Volume II. John Wiley Sons, Inc., New York—London 1960
4. Pethő Sz.: Meddőmennyiségek és széntermelés megállapítása munkahelyekre és bányaüzemekre, azonos megbízhatósági szinten. *Bányászati Lapok* (1966) 385
5. Pethő Sz.: Adatfelvételi igények meghatározása minőségelemzés céljára. *Bányászati Lapok* (1967). 100—104
6. Janositz J.—Pethő Sz.: A rétegzett mintavétel és a feltételes megfigyelések paramétereinek a legkisebb négyzetek módszerénél hatásosabb becslési módjáról. *Tatabányai Szénbányák Műszaki Közgazdasági Közleményei*, **10** (1970) ápr.—jún. 49—52
7. Pethő Sz.—Szarka Z.: Mintaelemszám egy meghatározási módja. *Minőség és megbízhatóság* (1969) 36—39
8. Pethő, Sz.—Szarka, Z.: Eine Art der Bestimmung des Stichprobenumfangs bei geschichteter Stichprobe. *Internationale Zeitschrift für Theorie und Praxis.* Statistische Hefte. **11** (1970), 234—236
9. Hazay I.: Kiegyenlítő számítások. Tankönyvkiadó, Budapest 1968

10. JANOSITZ J.—PETHŐ SZ.: Ásványelőkészítési kísérletek megbízhatóságáról és a kísérleti eredmények legvalószínűbb értékeiről. (Sajtó alatt)
11. JANOSITZ J.: Optimális mintaelhelyezés több halmazból történő mintavétel esetén. *Minőség és Megbízhatóság* (1969) 36—39
12. PIERRE, GY.: L'échantillonnage des minerais en vrac. Tome I. Théorie générale. 1967. *Mémoire du Bureau de Recherches Géologique et Minières.* 56. Éditions B. R. G. M.

Stratified Sampling. Stratified sampling is a better observation method than simple random sampling. When using it, the sampled set must be divided into subsets, strata, or an anyhow existing stratification of the set mut be used. In the different strata are members of approximately the same characteristics and low dispersion, but the strata means differ from each other the variation between strata means is large. With a set orderly arranged in this way the fixed number of samples must be divided among the strata in a suitable way (e.g. with the aid of the optimum distribution), and random samples must be taken from the subsets. The paper presents the most important relations between the parameters of stratified sampling, the simple unbiased estimations, and reviews the various sampling methods. The last part deals with two special estimating methods.

Die geschichtete Probeentnahme. Die geschichtete Probeentnahme ist eine wirkungsvollere Beobachtungsmethode als die einfache Probeentnahme. Die Menge aus der die Proben zu entnehmen sind wird in Teilmengen, Schichten aufgeteilt oder die vorhandene Schichtung der Menge wird ausgenützt. In den einzelnen Schichten befinden sich Elemente mit ungefähr gleichen Eigenschaften und kleiner Streuung, die Schichtmittelwerte unterscheiden sich jedoch bedeutender voneinander, d. h. die Streuung zwischen den Schichtmitteln ist groß. Bei einer derart geordneten Menge ist die fesgelegte Zahl der Proben entsprechend (z. B. mit Hilfe der optimalen Verteilung) auf die Schichten zu verteilen und aus den Schichten sind die Proben zufallsartig zu entnehmen. In der Arbeit werden die wichtigsten Zusammenhänge zwischen den Parametern, ferner die einfachen unverzerrten Schätzungen und die verschiedenen Verfahren für die Probeentnahme besprochen. Der letzte Teil der Arbeit bringt zwei spezielle Schätzungsverfahren