# KÉPVISELETI MINTA KORREKCIÓJA

TEKSE KÁLMÁN

Az 1960. évi magyar népszámlálás anyagából képviseleti minta készült. A képviseleti minta feldolgozásának célja, hogy az előzetes adatoknál jóval több demográfiai, családi és lakásismérvre vonatkozólag viszonylag rövid idő alatt részletesebb becslési eredményeket kapjunk. A népszámlálás teljes anyagából mechanikus eljárással, véletlenszerűen választottuk ki az ország lakosságának egy meghatározott százalékát.[1] A népszámlálás alkalmával az ország lakosságát háztartásonként írták össze (2). Ezért, továbbá egyéb elméleti megfontolások (1) eredményeképpen célszerűnek mutatkozott egyének helyett háztartások kiválasztása. Ilyen kiválasztási módszer mellett a mintába kerülő háztartások nagysága (tagjainak száma) valószínűségi változó, melynek szórása miatt a minta lélekszáma nem egyezik meg az ország lakossága adott százalékával, azonban annak torzítatlan becslése lesz. (Az intézeti háztartásokat a magánháztartásoktól elkülönítve kezeltük.)

A kiválasztás helyességének ellenőrzéséhez szükség van a minta nagyságára (elemszámára)[2] vonatkozó konfidencia-intervallum meghatározására. Ehhez a következő pontban meghatározzuk a minta nagyságának, mint valószínűségi változónak szórását. Miután e számításaink a kiválasztásra nézve igen megnyugtató eredményt adtak, elvégeztük a képviseleti minta korrekcióját. Eljárásunkat és indokolását a II. pontban ismertetjük.

## I. A MINTA NAGYSÁGA SZÓRÁSÁNAK MEGHATÁROZÁSA

Jelöljük $\xi_i$-vel ($i = 1, 2, \ldots 7$) az összesen $n$ háztartásból álló mintába került $i$ létszámú háztartások számát, $p_i$-vel pedig az $i$ létszámú háztartások ismert relatív gyakoriságát az alapsokaságban. Ismeretes (3), hogy a $(\xi_1, \xi_2, \ldots \xi_7)$ valószínűségi változók együttes eloszlása polihipergeometrikus eloszlás, melyet az alapsokaság nagy elemszáma miatt polinomiális eloszlással közelíthetünk. Ily módon:

$$P(\xi_i = k_1, \xi_2 = k_2, \ldots \xi_7 = k_7) = \frac{n!}{n!\ k_1!\ k_2! \ldots k_7!}\ p_1^{k_1}\ p_2^{k_2} \ldots p_7^{k_7}, \quad [1]$$

[1] A képviseleti minta nagyságával kapcsolatban vizsgálatok folytak a legalkalmasabb mintaelemszám (kiválasztási arány) meghatározására (1). Nagy pontosságot csak nagy mintaelemszámmal tudnánk biztosítani, ennek feldolgozása azonban rendkívül költséges, ugyanakkor a teljes népszámlálási anyag feldolgozását nagymértékben késleltetné. Kisebb mintából származó adatok céljainknak megfelelő pontosságúak és feldolgozásuk a teljes anyag feldolgozását alig késlelteti. E szempontok figyelembevételével 1%-os képviseleti mintát készítettünk.

[2] A minta nagysága alatt a háztartások — pontosan az alapsokaságbeli eloszlásnak megfelelő — adott %-a kiválasztása mellett adódó lélekszámot értjük.

ahol $$\sum_{i=1}^{7} k_i = n, \text{ és } \sum_{i=1}^{7} p_i = 1.$$

A minta $\eta$ terjedelmére fennáll:

$$\eta = \sum_{i=1}^{7} i\,\xi_i.$$

Mivel $$M(\xi_i) = n\,p_i,$$

ezért az $\eta$ valószínűségi változó várható értéke:

$$M(\eta) = M\left(\sum_{i=1}^{7} i\,\xi_i\right) = n \sum_{i=1}^{7} i\,p_i. \qquad [2]$$

A $\xi_i$ változók nem függetlenek és ezért a fenti egyenlőségek felhasználásával $\eta$ valószínűségi változó szórásnégyzetére kapjuk:

$$D^2(\eta) = \sum_{i=1}^{7} i^2\,M[(\xi_i - M(\xi_i))^2] +$$

$$+ 2 \sum_{1 \le j < k \le 7} \sum jk\,M[(\xi_j - M(\xi_j))(\xi_k - M(\xi_k))] \qquad [3]$$

ahol: $$M[(\xi_j - M(\xi_j))(\xi_k - M(\xi_k))] = R(\xi_j, \xi_k)\,D(\xi_j)\,D(\xi_k). \qquad [4]$$

Itt $R(\xi_j, \xi_k)$ $(1 \leqq j < k \leqq 7)$ $\xi_j$ és $\xi_k$ valószínűségi változók korrelációs együtthatóját jelenti.

Mivel a polinomiális eloszlás egyes komponensei külön-külön binomiális eloszlásúak, ezért [1]-ből:

$$M[(\xi_i - M(\xi_i))^2] = p_i\,(1 - p_i)\,n, \qquad [5]$$

és $$D(\xi_i) = \sqrt{n\,p_i\,(1 - p_i)}, \qquad (i = 1, 2, \ldots 7), \qquad [6]$$

$\xi_j$ és $\xi_k$ változók korrelációs együtthatója pedig:

$$R(\xi_j, \xi_k) = -\sqrt{\frac{p_j\,p_k}{(1 - p_j)(1 - p_k)}}. \qquad [7]$$

A $\xi_j$ és $\xi_k$ változók között tehát érthető módon mindig negatív korreláció áll fenn. Ily módon $\eta$ valószínűségi változó szórásnégyzete [3] és [4] egyenletekből [5], [6] és [7] felhasználásával:

$$D^2(\eta) = n\left[\sum_{i=1}^{7} i^2\,p_i\,(1 - p_i) - 2 \sum_{1 \le j < k \le 7} \sum jk \sqrt{\frac{p_j\,p_k}{(1 - p_j)(1 - p_k)}}\right.$$

$$\left. \sqrt{p_j\,(1 - p_j)}\ \sqrt{p_k\,(1 - p_k)}\right], \qquad [8]$$

azaz: $$D^2(\eta) = n\left[\sum_{i=1}^{7} i^2\,p_i\,(1 - p_i) - 2 \sum_{1 \le j < k \le 7} \sum jk\,p_j\,p_k\right].$$

[2] és [8] felhasználásával mintánk várható $\overline{m}$ terjedelmére az

$$M(\eta) - \lambda D(\eta) \leqq \overline{m} \leqq M(\eta) + \lambda D(\eta) \qquad [9]$$

konfidencia-intervallumot kapjuk, amelynek megbízhatóságát nagy $n$ esetén közelítőleg $[2\Phi(\lambda) - 1]$ szolgáltatja, ahol

$$\Phi(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{u^2}{2}} du .$$

Az 1949. évi magyar népszámlálás (4), valamint az 1959-ben végrehajtott magyar próbanépszámlálás eredményei (5) alapján lehetőség nyílt $p_i$-k közelítő meghatározására. Így Magyarországon számunkra szükséges pontossággal:

$$p_1 = 0{,}15\,;\ p_2 = 0{,}24\,;\ p_3 = 0{,}25\,;\ p_4 = 0{,}18\,;\ p_5 = 0{,}10\,;\ p_6 = 0{,}04\,;$$
$$p_7 = 0{,}04\,. \qquad [10]$$

Ezek felhasználásával [8]-ból kapjuk:

$$D(\eta) = 1{,}55\sqrt{n}\,.$$

Ily módon pl. a képviseleti minta Csongrád megyei anyagánál a minta nagyságára vonatkozó konfidencia-intervallum félhosszára — 95%-os valószínűségi szinten — 103 adódott,[3] a minta elemszámának a várható értékétől való eltérése viszont mindössze 41 volt. Teljesen hasonlóan, a képviseleti minta kiválasztása során a lélekszámokban mutatkozó eltérés nagysága (egy kivétellel) minden megyénél jóval a megfelelő hibahatárokon belül volt, 95%-os valószínűségi szinten. Ez azt mutatja, hogy a képviseleti minta háztartásnagyságra vonatkozó adatai elegendő pontossággal és megbízhatósággal közelítik az alapsokaság megfelelő adatait.

## II. A KÉPVISELETI MINTA KORRIGÁLÁSA

A minta elemszámának, mint valószínűségi változónak szórása miatt mintánk éppen a legfontosabb jellemző (a személyek száma) tekintetében nem adja a kívánt nagyságot. Ez a körülmény a minta adatainak az alapsokaságra történő kivetítésénél terjedelmes számolási munkát jelentene, amit az alábbi eljárásunkkal kiküszöbölünk, ugyanakkor az adatok közlésének egyszerűsítését is lehetővé tesszük.

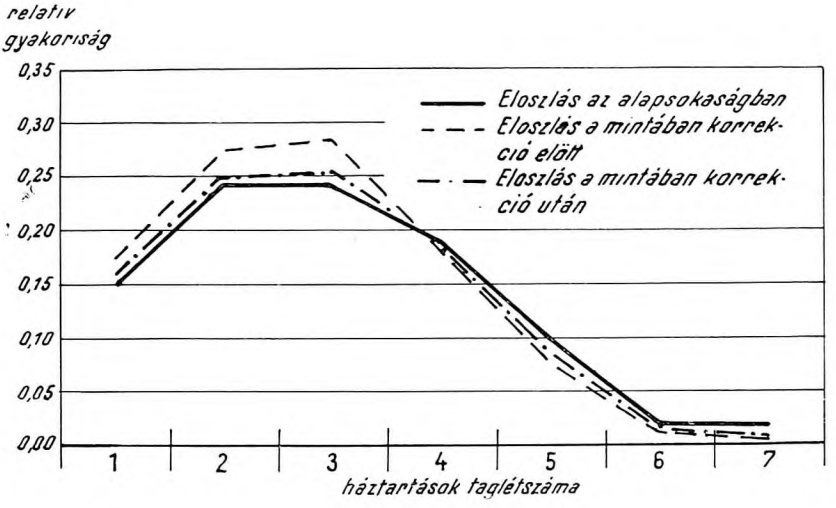
A nehézségek elkerülésének egyik módja a minta korrigálásának módszere.[4] Ennek lényege: bizonyos számú alkalmas nagyságú háztartásnak a mintából való kiemelése és bizonyos számú megfelelő nagyságú háztartás anyagának a mintába való helyezése oly módon, hogy a mintába került

[3] Mivel a háztartások nagyságszerinti eloszlása az egyes megyékre nem állott rendelkezésünkre, a megyékre vonatkozó számításainkban mindig a [10] eloszlást használtuk fel. A közelítésből származó hiba a konfidencia-intervallum hosszát lényegesen nem befolyásolja.

[4] Ezt a módszert alkalmazták pl. az 1951. évi angol népszámlálás anyagából készített 1%-os képviseleti minta kiválasztásánál is [6]. Egy másik lehetséges módszer a képviseleti minta adatait tartalmazó táblázatok korrigálásának módszere, ami két táblázat közlését teszi kívánatossá [7].

háztartások és személyek száma pontosan a kívánt mintanagyságot szolgáltassák. A megfelelő háztartások kiválasztása véletlen számtáblázat segítségével történik.

Ez a módszer teljesen megalapozott, ha a korrekció során felhasznált háztartások nagyság szerinti megoszlása megegyezik a háztartások alapsokaságbeli megoszlásával.[5] Könnyű belátni, hogy ebben az esetben a korrekció után a háztartások mintabeli megoszlása jobban megközelíti az alapsokaságbeli megoszlást, mint a korrekció előtt. (Lásd az ábrát.)




A minta korrigálásának módszere abban az esetben alkalmazható, ha ismeretes a háztartások nagyságszerinti megoszlása az alapsokaságban. Ez a mintavételek során általában ismeretlen, azonban gyakran lehetőség nyílik jó becslésére. E becsléshez Magyarországon is megfelelő adatok álltak rendelkezésünkre. Ilyen becslés útján kaptuk a [10] megoszlást, melynek felhasználásával végeztük számításainkat. Azok az egyezések, amelyekre az I. pont végén hivatkoztunk, bizonyos mértékig a [10] eloszlás használatának jogosságára is utalnak.

A korrekció végrehajtásához meg kell határozni a korrekció során a mintából kiemelésre kerülő és a mintába helyezendő különböző nagyságú háztartások számát.

Jelöljük ehhez $x_i$-vel a korrekció során felhasznált kiemelendő, vagy behelyezendő $i$-tagú háztartások számát, továbbá $p_i$-vel az $i$-tagú háztartások relatív gyakoriságát az alapsokaságban. Jelöljük $A$-val a minta terjedelmének az előírttól való előjeles eltérését, (mely pozitív vagy negatív, aszerint, hogy a kiválasztott minta terjedelme nagyobb, vagy kisebb a kívánt mintaterjedelemnél), $B$-vel a kiválasztott háztartások számának a kívánt háztartásszámtól való eltérése abszolút értékét.

[5] A módszer eddigi alkalmazásainál e fontos tényt nem vették kellően figyelembe, a korrekciónál felhasznált háztartások nagyságát csak hozzávetőlegesen, a lélekszámból kiindulva állapították meg. Ez a minta torzítását eredményezte.

A korrekció során a mintából kiemelésre kerülő, ill. pótlólag a mintába helyezendő háztartások számai közti különbségnek éppen $B$ eltéréssel kell egyenlőnek lennie és ezért

$$\sum_{i=1}^{3} x_i - \sum_{i=4}^{7} x_i = \pm B \text{ aszerint, hogy } A \gtrless 0. \qquad [11]$$

A korrekció során felhasznált $i$ tagú háztartások $x_i$ számát úgy kell megválasztanunk, hogy a korrekció eredményeképpen a minta terjedelme $A$-val változzon; növekedjék, vagy csökkenjen, aszerint, hogy eredetileg $A < 0$, vagy $A > 0$ volt. E feltételből kapjuk:

$$\sum_{i=1}^{3} i\,x_i - \sum_{i=4}^{7} i\,x_i = -A. \qquad [12]$$

Magyarországon az átlagos háztartásnagyság 3,14, ennek megfelelően [11] és [12]-ben az $x_i$ ($i = 1, 2, 3$; „kis háztartások" számai) és az $x_j$ ($j = 4, 5, 6, 7$; „nagy háztartások" számai) ismeretlenek ellenkező előjelűek.

A korrekció során felhasznált különböző nagyságú háztartások számainak (a minta torzítása elkerülése érdekében) követniök kell az alapsokaság háztartásai ismert nagyságszerinti eloszlását, vagyis a

$$p_1 : p_2 : p_3 = x_1 : x_2 : x_3 \text{ és a } p_4 : p_5 : p_6 : p_7 = x_4 : x_5 : x_6 : x_7$$

arányoknak kell teljesülniök. Ezek az arányok a következő öt független egyenlethez vezetnek:

$$p_i\,x_j = p_j\,x_i \quad (p \neq 0) \qquad \begin{matrix} i = 1;\ j = 2, 3; \\ i = 4;\ j = 5, 6, 7. \end{matrix} \qquad [13]$$

Tehát a korrekció során felhasználandó különböző nagyságú háztartások számainak ki kell elégíteniök a [11]—[13] lineáris egyenletrendszert. Feltehetjük, hogy ez a lineáris egyenletrendszer inhomogén, ellenkező esetben nem lenne szükség korrekcióra.

A korrekció abban az esetben hajtható végre, ha a [11]—[13] inhomogén lineáris egyenletrendszernek létezik a triviálistól különböző megoldása, azaz, ha e rendszer $D$ determinánsa nem nulla. A determináns:

$$D = \begin{vmatrix} 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 2 & 3 & -4 & -5 & -6 & -7 \\ p_2 & -p_1 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & -p_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_5 & -p_4 & 0 & 0 \\ 0 & 0 & 0 & p_6 & 0 & -p_4 & 0 \\ 0 & 0 & 0 & p_7 & 0 & 0 & -p_4 \end{vmatrix} \neq 0,$$

amelybe esetünkben a $p_i$-ket a [10] eloszlásból helyettesítve, valóban nullától különböző értéket nyerünk. Így a [11]—[13] inhomogén rendszernek

létezik a triviálistól különböző megoldása. A korrekciónál a kapott megoldásokat előjellel kell figyelembe venni.

A [11]—[13] rendszer ($x_1, x_2, \ldots x_7$) megoldásaként csak egész számok jöhetnek tekintetbe, amit megfelelő kerekítés után érünk el.

Általában a minta korrigálásának vizsgált módszere minden olyan esetben alkalmazható, amikor a képviseleti minta eleme és a minta kiválasztási egysége nem azonosak, és a fenti feltételek teljesülnek.

A bemutatott korrekciós módszert alkalmaztuk az 1960. évi magyar népszámlálás anyagából készült képviseleti minta korrigálásánál is. Mivel a minta anyagát megyei bontásban publikáljuk, a korrekciót is megyei egységekre kellett végrehajtani. Ehhez nem álltak rendelkezésünkre a háztartások megyékre vonatkozó taglétszám szerinti eloszlásai. Ezért az 1%-os minta egyes megyékre vonatkozó részének korrigálásánál is a [10] országos megoszlásokat vettük alapul. Az 1%-os mintában azonban $A$ és $B$ abszolút értékei olyan kicsinyek voltak, hogy a [11]—[13] rendszer megoldásainak e közelítésből származó hibái, a kerekítéseket is figyelembe véve, jelentéktelenek.

A módszer alkalmazásának illusztrálására nézzük meg pl. Csongrád megye anyagának korrekcióját. A képviseleti minta kiválasztása során Csongrád megye anyagából a szükségesnél két háztartással és 47 személlyel kevesebbet választottunk ki. ($A = -47$, $B = 2$.) E konstansokkal a [11]—[13] rendszer megoldásaira kapjuk:

$$x_1 = 3{,}3\,; \quad x_2 = 5{,}2\,; \quad x_3 = 5{,}5\,; \quad x_4 = 7{,}9\,; \quad x_5 = 4{,}4\,; \quad x_6 = 1{,}7\,; \quad x_7 = 1{,}7.$$

Kerekítés után a következő értékekkel hajtottuk végre a korrekciót:

$$x_1 = 3\,; \quad x_2 = 5\,; \quad x_3 = 6\,; \quad x_4 = 8\,; \quad x_5 = 4\,; \quad x_6 = 2\,; \quad x_7 = 2\,.$$

Ily módon a korrekció során 3 egytagú, 5 kéttagú, 6 háromtagú háztartást kellett kiemelni a mintából és 8 négytagú, 4 öttagú, 2 hattagú, 2 héttagú háztartást kellett pótlólag a mintába helyezni. E háztartásokat a mintából, ill. az alapsokaságból véletlen számtáblázat segítségével jelöltük ki.

*Megjegyzés:* Amennyiben nem írjuk elő a mintába kerülő háztartások számát, tehát bizonyos okokból kifolyólag az elsődleges kiválasztási egység száma nincs rögzítve (azaz a korrekció után $B$ nem feltétlenül 0), a fenti egyenletrendszer [11] egyenletét figyelmen kívül hagyhatjuk. Ily módon a [12]—[13] egyenletekből álló rendszer hét ismeretlent tartalmazó hat egyenletből álló rendszerre redukálódik. E rendszert

$$p'_4 x_1 \sum_{i=1}^{3} i\, p'_i - p'_1 x_4 \sum_{i=4}^{7} i\, p'_i = A\, p'_1 p'_4,$$

vagy

$$p'_4 \sum_{i=1}^{3} i\, p'_i = a\,; \quad p'_1 \sum_{i=4}^{7} i\, p'_i = b\,; \quad A\, p'_1 p'_4 = c \qquad [14]$$

jelölések bevezetésével

$$ax_1 - bx_4 = c \qquad [15]$$

alakra hozhatjuk, ahol

$$p_i' = 100\, p_i \quad (i = 1, 2, \ldots 7).$$

Korrekciónk végrehajtásához a [14] egész együtthatós lineáris diofantikus egyenlet egész megoldásait kell meghatároznunk. Ennek akkor van megoldása, ha $(a, b) = 1$, vagy $(a, b) = d > 1$ esetén $d/c$.[6] A [15] lineáris diofantikus egyenletnek az előbbi esetben pontosan egy, az utóbbi esetben éppen $d$ számú $c$-nél kisebb megoldása van (8).

Ha feltesszük, hogy képviseleti mintánk korrigálásánál fennállnak a megjegyzés követelményei, akkor [14]-ből [10] felhasználásával

$$a = 2484, \qquad b = 2610, \qquad c = 270\ \text{A}.$$

Így $(a, b) = d = 18$, azaz: $c = 15\, d \cdot A$.

Ekkor tehát a [15] lineáris diofantikus egyenletnek minden $A$ esetén létezik megoldása, éspedig pontosan $d = 18$ darab $c$-nél kisebb megoldás. A korrekció legegyszerűbb végrehajtásához ezen egyenlet legkisebb abszolút értékű megoldását kell felhasználni.

Kívánatos lenne a minta korrigálásának fenti módszerét kidolgozni többszörösen rétegezett mintavételek esetére is.

## IRODALOM


1. *Dr. Bene Lajos:* Előtanulmányok a népszámlálás képviseleti feldolgozásához. *Demográfia*, 1959. évi 4. sz. 501—519 p.
2. *Dr. Klinger András—dr Szabady Egon:* Az 1960. évi népszámlálás előkészítése, adatgyűjtési és feldolgozási programja. *Statisztikai Szemle*, 1959. évi 8—9. sz. 795—839 p.
3. *Rényi Alfréd:* Valószínűségszámítás. Tankönyvkiadó V., 322 p., Budapest 1954.
4. *Központi Statisztikai Hivatal:* 1949. évi népszámlálás 10. kötet. Családstatisztikai eredmények. Budapest 1951.
5. *Dr. Vukovich György:* Az 1959. évi népszámlálási próbafelvétel néhány módszertani kérdése. *Demográfia*, 1959. évi 1. sz. 101—109. p.
6. *General Register Office:* Census 1951. Great Britain one per cent sample tables. HMSO. London 1952.
7. *Badry, M. A. El.—Stephan, F. F.:* On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 1950. évf. 738—762. p.
8. *Vinogradov, I. M.:* A számelmélet alapjai. Tankönyvkiadó V., 42—46. p., Budapest 1951.


## О КОРРЕКЦИИ МАТЕМАТИЧЕСКОЙ ВЫБОРКИ


*Резюме*

Из материала общегосударственной переписи населения 1960 года приготовили выборку, где пропорция отбора — однопроцентная. Так как во время переписи населения перепись производилась по домашним хозяйствам (1), далее исходя из некоторых теоретических рассуждений (2), целесообразно было осуществить выборку путем отбора домашних хозяйств. При таком отборе размер домашних хозяйств — случайная величина, из-за дисперсии которой численность населения, входящая в выборку, не совпадает с одним процентом общей численности населения страны; однако она будет несмещенной оценкой последней. Чтобы избе-


[6] Itt $a$ és $b$ egész számok legnagyobb közös osztóját $(a, b)$ szimbólummal jelöltük, $d/c$ pedig azt jelenti, hogy $d$ osztója $c$-nek.

жать трудоемких вычислений, связанных с проекцией данных выборки на генеральную совокупность, желательна такая коррекция выборки, которая приведет точно к одному проценту и по численности населения.

1) Для проверки правильности отбора, необходимо определить доверительный интервал для объема выборки (численности населения, входящей в выборку). Совместная функция распределения чисел $\xi_i$ ($i = 1,2,...,7$) домашних хозяйств размера $i$ входящих в выборку, содержащую числа $n$ домашних хозяйств — полигипергеометрическое распределение. Из-за генеральной совокупности, содержащей большое число элементов, это можно оценить мультиноминальным распределением. Здесь обозначим через $p_i$ известную (по предыдущей переписи населения (4), а также по данным пробной переписи 1959 года (5)) относительную частоту домашних хозяйств размера $i$ в генеральной совокупности (см. (10), стр. . . . . .).

Таким образом (8) дает дисперсию объема

$$\eta = \sum_{i=1}^{7} i\,\xi_i$$

выборки, откуда легко получается доверительный интервал для математического ожидания $\bar{m}$ объема выборки, надежность которого при большом $n$ приблизительно равняется ($2\Phi(\lambda) - 1$), где через $M(\eta)$ обозначим математическое ожидание случайной величины $\eta$, и $\Phi(\lambda)$ — нормальная функция нормального распределения.

Эти доверительные интервалы всегда покрывали величины $m$ полученные при отборе выборки. Это показывает, что данные выборки, относящиеся к размеру домашних хозяйств, с достаточной точностью и надежностью дадут соответствующие данные генеральной совокупности, в то же время это свидетельствует о правильности употребления данных (10).

2) В следствии дисперсии объема (как случайной величины) выборки, наша выборка не будет иметь данного значения по важнейшему признаку (по числу лиц). Это затрудняет проекцию данных выборки на генеральную совокупность. Во избежание этих трудностей пользуемся одним из возможных методов (см. например (6)) : методом коррекции выборки. Этот метод в том случае применяемый, если задано распределение (или его приближение) домашних хозяйств по их размеру в генеральной совокупности.

Мы пользовались следующими обозначениями: через $A$ обозначали отклонение объема выборки от заранее заданного объема (положительное или отрицательное, если объем выборки больше или меньше заданного выборочного объема); через $B$ — отклонение по модулю числа отобранных домашних хозяйств от заданного числа хозяйств (известного из предварительных результатов переписи населения); и наконец через $x_i$ — число, используемое в ходе коррекции, домашних хозяйств размера $i$.

Суть метода коррекции выборки заключается в том, что в ходе коррекции путем выемки из выборки или восстановления в выборку числа $x_i$ домашних хозяйств размера $i$, достигнем того, чтобы число лиц и домашних хозяйств, входящих в выборку, доставило заданный объем выборки. (Так как в Венгрии средний размер домашних хозяйств — 3,14, поэтому $x_i$ обозначает при $i = 1,2,3$ — число «маленьких», при $i = 4,5,6,7$ — число «больших» домашних хозяйств). Для этого неизвестные величины $x_i$ удовлетворяют линейным уравнениям (11) и (12). Далее для чисел домашних хозяйств, использованных при коррекции, должны иметь место соотношения, вытекающие из известного распределения домашних хозяйств по размеру в генеральной совокупности. (Таким образом удастся избежать смещения выборки.) Поэтому $x_i$ удовлетворяют равенствам (13). После выполнения коррекции распределение домашних хозяйств в выборке точнее приближает, чем до коррекции, распределение хозяйств в генеральной совокупности (см. рис.).

Таким образом, числа домашних хозяйств разных размеров, используемых в ходе коррекции, удовлетворяют систему линейных неоднородных уравнений (11)—(13), которая имеет решения, отличных от тривиаль-

ных, если определитель системы отличен от нуля. Решения системы — после обычных округлений — можно употреблять при коррекции. Домашние хозяйства, использованные в ходе коррекции, выбираем при помощи таблицы случайных величин.

Вообще, исследованный метод коррекции выборки всегда можно употреблять, если элемент выборки и выборочная единица не совпадают, и удовлетворяются вышеуказанные требования.

Вышеуказанный метод коррекции был употреблен при отборе выборки, приготовленной из материала общегосударственной переписи населения 1960-го года. Коррекция была выполнена по комитатам на основании (как хорошей оценки) распределения (10). На примере показали метод коррекции выборки.

*Замечание:* В том случае, если число домашних хозяйств без особых условий, уравнение (11) нашей системы отпадает. Таким образом из системы уравнений получаем систему из шести уравнений, содержащих семь неизвестных, которую обозначениями (14) привели к виду (15). Теперь для выполнения коррекции надо отыскать минимальные, положительные целые решения линейного уравнения с целыми коэффициентами. Это уравнение: 1) имеет одно решение по модулю меньше $c$, если $a$ и $b$ не имеют общего делителя; 2) имеет $d$ решений по модулю меньше $c$, если наибольший общий делитель $d$ чисел $a$ и $b$ — делитель и $c$. Показали, что в случае венгерской выборки при любом $A$ существует $d = 18$ решений, по модулю меньше $c$.

## ON THE CORRECTION OF THE REPRESENTATIVE SAMPLE

*Summary*

From the material of the Hungarian census of the year 1960 a 1 % representative sample was taken. As the population was enlisted according to households [1], further on the base of other — theoretical — considerations [2], it seemed advisable to choose the sample not by individuals but by households. At such selection the size of the households (number of members) is a random variable, and in consequence of its variance the 1 % representation of the households does not involve the 1 % representation of the individuals, but is only an unbiassed estimate of it. In order to avoid in connection with the projection of the data extensive calculations, a correction of such a sample is desirable, leading to a 1 % representation of the population.

1. The control of the proper selection requires the determination of the confidence interval of the sample, with respect to the number of individuals. The joint distribution of the numbers $\xi_i$ $(i = 1, 2, \ldots, 7)$ of households with $i$ members in the sample of $n$ households is a hypergeometric one, which has been approximated — in consideration of the large sample size — by the polynomial distribution. Here the $p_i$-s denote the known (by the previous census [4] and the pilot census of 1959 [5]) relative frequency of the households with $i$ members in the parent distribution. The standard deviation of the sample size

$$\eta = \sum_{i=1}^{7} i\,\xi_i$$

is given by (8) so that the confidence interval of the expected size $\overline{m}$ of the sample can easily be determined for large $n$ on an approximate confidence level $(2\,\Phi\,(\lambda) - 1)$, where $M\,(\eta)$ denotes the expectation of the random variable $\eta$ and $\Phi\,(\lambda)$ the standard normal distribution function.

The values of $\overline{m}$ obtained during the selection of the sample were always within these confidence intervals, indicating that the data of the representative sample — in respect of the size of households — give a sufficiently close and reliable approximation to the data of the total population. At the same time this justifies the use of the values 10.

2. In consequence of the standard deviation of the sample size as a random variable our sample does not come up to the required size just in respect of the most important

characteristic (number of individuals). This causes difficulties in projecting the data fo the sample on the whole population. One way of obviating these difficulties is the correction of the sample. This method can be applied if the distribution of the size of households in the whole population is known, or at least a good approximation of it.

In the paper the following notations are used: $A$ denotes the signed deviation of the sample size from the prescripted value (which is positive resp. negative according to whether the sample size is larger or smaller than the prescription), $B$ denotes the absolute deviation of the number of selected households from the required (in the course of the technical realization of the selection this may occur too), $x_i$ the number of households consisting of $i$ members, to be used in the correction.

The principle of the correction is to attain by subsequent ommission resp. addition of $x_i$ households with $i$ members, that the number of households and individuals contained in the representative sample should be exactly as required. (As in Hungary the average size of households was equal to 3,14, hence the $x_i$-s give for $i = 1, 2, 3$, the number of "small" households and for $i = 4, 5, 6, 7$ that of the "large" ones.) For meeting this requirement the $x_i$-s must fulfill the linear equation system (11) resp. (12). The number of households of different sizes used in the correction must follow (in order to avoid the biassedness of the sample) the known distribution of the totality of households. (Till now this important circumstance has not been sufficiently considered.) Thus for the $x_i$-s the equalities (13) must hold. After carrying through the correction, the distribution of households in the sample comes nearer to that of the whole population. (See fig.)

I. e. the number of households of different sizes to be used in the course of the correction must fulfill the linear inhomogenous equation system (11)—(13) which has a non trivial solution if its determinant is not zero. The solutions of the system can be used — after the usual rounding off to integers — in carrying through the correction. The households to be used in the correction are selected by means of random number tables.

In general this method of correction is applicable in any problem where the elements of the representation do not coincide with the units of selection and above conditions are fulfilled.

Above method of correction was applied in the selection of the representative sample at the Hungarian census of 1960. The correction was carried through according to counties, making use of distribution (10) as an estimation. In the paper the method of correction is shown on an example.

*Remark:* If the number of households in the sample is not prefixed, equation (11) of above system can be disregarded. Thus the system consisting of equations (12) — (13) reduces to a system of six equations, containing seven unknowns, which can be written by introduction of notations (14) in the form (15). Now we ask for the smallest positive integer solutions of this linear (diophantic) equation with integer coefficients. This has one solution only if $a$ and $b$ are relative primes and $d$ solutions smaller than $c$, if the greatest common divisor of $a$ and $b$, $d$ is divisor of $c$ too. It is shown that in case of the Hungarian representative sample there exist for any $A$ exactly $d = 18$ solutions which are smaller than $c$.