A harmadik probléma tisztázásával szeretnénk befejezni a cikksorozatot. Emlékeztetőül ismét idézzük fel, miről is van szó.

**3. Probléma.** (Hogyan lehet egy közvéleménykutatási eredményből, tehát egy minimodellből következtetni az egész sokaságra?)

A Wash&Go cég szeretne képet kapni, hogy legújabb termékük milyen fogadtatásban részesült a magyar piacon. Ezért valahogyan kiválasztanak 1000 családot, s megkérdezik a véleményüket. Ha 342 család nem észlelt javulást az új termék esetén, sőt 47-en még rosszabbnek is tartják, míg 611 család véli jobbnak, akkor vajon jobbnak tartsuk-e az új terméket? Milyen előfeltételeket használunk az okoskodásban?

A harmadik probléma, ha figyelmesen elolvassuk, valahogyan „inverze" a másodiknak. Ezúttal a sokaságról nincs információnk, ismerünk viszont egy mintát, s ebből szeretnénk valamilyen módon arra következtetni, mit lehet nagy valószínűséggel mondani az egész halmazról. Biztosan természetesen most sem lehet semmit. De ha a mintáról tudjuk, hogy ténylegesen véletlenszerűen választott, azaz nincs manipulálva, akkor abból következtethetünk a sokaságra.

A manipuláció valóban gyakori a valóságban, s ez vezet oda, hogy az emberek gyakran mondják, nem hiszünk el semmilyen következtetést, mert az úgyis csalás. Ismert vélemény: A statisztikával mindent lehet bizonyítani, s mindennek az ellenkezőjét is. Ezért gyakori az az álláspont, hogy ez nem matematika, nem is illik bele a matematika tanításába, hiszen ott az ellentmondásmentes logika a fő vezérelv. Ez a vélemény azonban alapvetően hamis. A statisztika semmit sem bizonyít a matematikai bizonyosság klasszikus értelmében. Pontosan akkor kap szerepet — de akkor nincs is jobb nála — ha bizonyosan semmit nem lehet mondani, de szeretnénk mégis valahogyan mérni, hogy egyes állítások mennyire hihetőek. Tehát azért, mert nincs abszolút bizonyosság, még nem kell teljesen lemondani egy rangsor felállításáról aszerint, hogy a biztosan meg nem válaszolható kérdésekre melyik válasz a hihetőbb, vagy a matematika precízebb megfogalmazása szerint, melyik a valószínűbb.

Ha valami nagyon valószínűtlen, akkor „praktikusan" nem fogjuk azt várni, hogy gyakori jelenség legyen, ezért az ellenkezőjét tekintük majdnem biztosnak.

Ezen kis kitérő után fogalmazzuk meg, mit értünk korrekt választáson.

A második cikk végén leírtak itt is érvényesek, hiszen csak a kérdésfeltevés s az ismert információ más. Tehát feltehetjük, hogy a minta véletlenszerűen választott, méghozzá úgy, hogy egyik elem sincs kitüntetve, azaz a választás egyenletes eloszlás szerint történt. Erről írtunk [1]-ben az utolsó részben.

A kérdés az, hogy vajon mit, s hogyan lehet a mintából következtetni a teljes sokaságra? Eljárásunkat egy ábrán fogjuk szemléltetni, amelynek lényege, hogy az „inverz", már megoldott feladatra próbáljuk meg visszavezetni a problémát.

A felső szakaszon szemléltetjük az alapul vett sokaságban egy adott tulajdonság előfordulási gyakoriságát (relatív gyakoriságát) ami 0 és 1 vagy százalékban kifejezve 0 és 100 közé esik. Ezt a statisztikában a becsülni kívánt paraméternek szokás nevezni, jele általában $q$.

Az alsó szakasz reprezentálja a mintát, ahol szintén van egy relatív gyakoriságunk, ezt jelöljük $X$-szel. *(1. ábra).*

Második feladatunk eszerint az volt, hogy ha ismerjük $\Theta$ értékét a sokaságban, akkor egy véletlenül választott mintában előírt velószínűséggel milyen intervallumba esik $X$, a várt relatív gyakoriság. Lásd a *2. ábrát*, ahol $\alpha$ jelöli a rizikó faktort, tehát az $X$ $1-\alpha$ eséllyel az alsó szakaszon berajzolt intervallumba esik. A cikksorozat második részében mutattuk meg, hogyan lehet ezt kiszámolni. Az itt bemutatott 2. ábra csak szemléltetés.

A mostani harmadik feladat tényleg az inverz, itt az alsó szakaszon ismerünk egy $X$ pontot, s a felső szakaszon keresünk egy intervallumot, amibe előírt eséllyel kell beleesnie $\Theta$-nak, amely most ismeretlen, s éppen ezt kívánjuk becsülni. Lásd a *3. ábrát.*

A kérdés az, hogyan lehetne a felső intervallumra valamilyen becslést adni. Itt segít a második megoldott probléma. Azt fogjuk mondani, hogy azok a pontok jönnek szóba fent, ahonnan a 2. ábrának megfelelő alsó intervallumba beleesik az ismert $X = X_0$. Ezt szemléltetjük a *4. ábrán.* $\Theta_1$ és $\Theta_2$ biztosan bele fog tartozni a keresett intervallumba, mert ezekből indított $\alpha$ rizikójú alsó intervallumba beleesik $X$. A mi feladatunk a két szélső $\Theta_b$ és $\Theta_j$-vel jelölt pont meghatározása lesz.

Írjuk fel, mit jelent a $\Theta_b$, illetve a $\Theta_j$ pont feltételeink szerint. Ha a sokaság paramétere $\Theta_b$, akkor annak az esélye, hogy $X \leq X_0$ egyenlő kell legyen $(1-\alpha)$-val. Tehát, ha a minta éppen $n$ elemű, akkor:

$$\sum_{k=0}^{nX_0} \binom{n}{k} \cdot \Theta_b^k \cdot (1-\Theta_b)^{n-k} = 1-\alpha.$$

Feltételezve, hogy ismét lehet a normális eloszlással becsülni, most az $n \cdot \Theta_b \cdot (1-\Theta_b) > 9$ kell, hogy teljesüljön. A várható érték természetesen $n\Theta_b$, tehát a standardizálás esetünkben azt jelenti, hogy $n\Theta_b$ értékét kell levonni, s osztani a szórással, ami $\sqrt{n \cdot \Theta_b \cdot (1-\Theta_b)}$.

Azaz a $\Phi\left(\dfrac{nX_0 - x\Theta_b}{\sqrt{x\Theta_b(1-\Theta_b)}}\right) = 1-\alpha$ egyenlőséget kell megoldani. Mivel $\alpha$ mint rizikófaktor általában előre adott, innen a $\Phi$ táblázatok segítségével meghatázotható az a $z_\alpha$ érték, amire a két oldal egyenlő.

Innen továbbléphetünk a következő egyenlet megoldásával:

(I)
$$\frac{nX_0 - n\Theta_b}{\sqrt{n\Theta_b(1-\Theta_b)}} = z_\alpha.$$

Mivel $X_0$ és $n$ (a minta elemszáma) adott, azért csak $\Theta_b$ az ismeretlen, amely ebből az egyenletből kiszámítható. Előbb irjuk fel a hasonló egyenletet $\Theta_j$-re is, s azután együtt oldjuk meg a két egyenletet.

Ezúttal a következő egyenlőségnek ($\Theta_j$ is határeset) kell teljesülni:

$$\sum_{k=nX_0}^{n} \binom{n}{k} \cdot \Theta_j^k \cdot (1-\Theta_j)^{n-k} = 1-\alpha.$$

Ismét feltételezve, hogy $n \cdot \Theta_j \cdot (1-\Theta_j) > 9$ közelíthetünk a normális eloszlással. A normálást hasonlóképpen végrehajtva, felhasználva, hogy most az ábra szerint $n\Theta_j > nX_0$:

$$1 - \Phi\left(\frac{nX_0 - n\Theta_j}{\sqrt{n\Theta_j(1-\Theta_j)}}\right) = 1-\alpha,$$

ekkor a $\Phi(-z) = 1-\Phi(z)$ tulajdonságot használva:

$$\Phi\left(\frac{n\Theta_j - nX_0}{\sqrt{n\Theta_j(1-\Theta_j)}}\right) = 1-\alpha,$$

azaz, ha $z_\alpha$ jelöli ismét azt a számot, amelyre $\Phi(z) = 1-\alpha$, akkor most a következő egyenlőséget kapjuk:

(II)
$$\frac{n\Theta_j - nX_0}{\sqrt{n\Theta_j(1-\Theta_j}} = z_\alpha.$$

Látható, mennyire hasonlít (I) és (II).

Emeljünk négyzetre mindkét esetben. Ekkor az alábbi másodfokú egyenlet két gyöke közül értelemszerűen a kisebb lesz $\Theta_b$, míg a nagyobb $\Theta_j$:

$$n^2(X_0-\Theta)^2 = z_\alpha^2 \cdot n \cdot \Theta(1-\Theta).$$

Ha ezt rendezzük, a következő másodfokú egyeneletet kapjuk $\Theta$-ra:

$$\left(x + z_\alpha^2\right)\Theta^2 - \left(2nX_0 + z_\alpha^2\right)\Theta + nX_0^2 = 0$$

Ennek megoldásakor, mint minden másofokú egyenletnél, a paraméterektől függően különböző lehetőségek lépnek föl. A diszkrimináns:

$$\begin{aligned}\left(2nX_0 + z_\alpha^2\right) - 4nX_0^2\left(n + z_\alpha^2\right) = z_\alpha^2 - 4nX_0^2 z_\alpha^2 + 4nX_0^2 z_\alpha^2 = \\ = z_\alpha^2\left(z_\alpha^2 + 4nX_0[1-X_0]\right) > 0.\end{aligned}$$

Tehát esetünkben mindig lesz valós gyök, mivel $(1-X_0)$ mindig nem negatív, hiszen $X_0$ egy relatív gyakoriság, tehát 0 és 1 közé esik.

A két gyök tehát a következő:

$$\Theta_b = \frac{\left(2nX_0 + z - \alpha^2\right) - z_\alpha \cdot \sqrt{\left(z_\alpha^2 + 4nX_0[1-X_0]\right)}}{2\left(n + z_\alpha^2\right)}$$
$$\Theta_j = \frac{\left(2nX_0 + z - \alpha^2\right) + z_\alpha \cdot \sqrt{\left(z_\alpha^2 + 4nX_0[1-X_0]\right)}}{2\left(n + z_\alpha^2\right)}$$

Természetesen semmi lényegeset nem mondtunk, ha vagy $\Theta_b$ kisebb, mint 0, vagy ha $\Theta_j$ nagyobb, mint 1.

(Az olvasóra bízzuk, milyen feltételek esetén lép ez föl.)

Befejezésül nézzük meg, a fenti eredményekből milyen válasz adható a mi kérdésünkre. Tegyük fel, hogy $\alpha = 0,025$, ekkor $z_\alpha = 1,96$.

Ha azokat nézzük, akik szerint javult a minőség, akkor $X_0 = \frac{611}{1000}$. Mivel $n = 1000$, azért ki lehet számolni a két $\Theta$ értéket: $\Theta_b = 0,58$; $\Theta_j = 0,64$. Tehát 97,5% eséllyel mondhatjuk, hogy a lakosság 58–64%-a elégedett, míg

kifejezetten a változás ellen szavazókra adódik: $X_0 = \frac{47}{1000}$, s így $\Theta_b = 0,035$, $\Theta_j = 0,062$, azaz a lakosságnak 97,5%-os biztonsággal legfeljebb 3,5–6,4%-a tartja rosszabbnak az új terméket.

Annak megítélése, hogy ez a cégnek jó vagy rossz eredmény, már nem tartozik feladataink közé.

A legrosszabb eseteket véve: 6,4%-nál közel 9-szer több az 58%, ezért azt mondhatjuk, hogy a lakosság nem elégedetlen az új terméket illetően. Gyakorlásul érdemes az olvasónak kiszámolni, milyen intervallumba esik azok száma, akik szerint változatlan a termék. Vajon hogyan változnak ezek az értékek, ha növeljük vagy csökkentjük a rizikót? Ezen is érdemes elgondolkozni, s esetleg számolni is. Ezzel jó gyakorlatot lehet szerezni az ilyen típusú becslésekben.

Befejezésül ennek a klasszikus módszernek egy meglevő hiányosságára szeretnénk rámutatni, amelynek lehetséges javítása már túlmutat ezen cikk keretein. Mi tulajdonképpen annak az esélyét vettük alapul, hogy ha $\Theta = \Theta_0$, akkor adott eséllyel lehet-e a minta gyakorisága az éppen megfigyelt $X = X_0$ érték. Tehát olyan $\Theta$ értékeket engedtünk meg, amelyekre teljesül: $P(X_0 \in (x_1, x_2) \mid \Theta = \Theta_m) \geq 1-\alpha$, ahol $\mathrm{P}(\ldots \mid \ldots)$ jelöli a feltételes valószínűséget, míg $\Theta_m$ egy megengedett érték