

ELEMİ BIZONYÍTÁSOK A RENDEZETT MINTÁK ELMÉLETÉNEK NÉHÁNY ALAPVETŐ ÖSSZEFÜGGÉSÉRE

HAJÓS GYÖRGY r. tag és RÉNYI ALFRÉD lev. tag

Legyen $\xi_1, \xi_2, \dots, \xi_n$ egy $F(x)$ folytonos eloszlásfüggvényű statisztikai sokaságból vett n -elemű minta, más szóval legyenek $\xi_1, \xi_2, \dots, \xi_n$ teljesen független és közös $F(x)$ folytonos eloszlásfüggvénnyel bíró valószínűségi változók. Legyen $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ a fenti változók nagyság szerint növekedőleg elrendezett sorozata, azaz legyen

$$\xi_k^* = R_k(\xi_1, \xi_2, \dots, \xi_n) \quad (k = 1, 2, \dots, n)$$

ahol $R_k(x_1, x_2, \dots, x_n)$ az $x_1, x_2, \dots, x_n$ számok nagyság szerint növekedőleg elrendezett sorozatában a k -adikat jelenti.*

Ez a dolgozat a $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ valószínűségi változókkal, tehát a *rendezett minták elméletével* foglalkozik. Egyszerű módszerekkel fogunk bebizonyítani bizonyos alapvető összefüggéseket. Kifejezetten arra törekszünk, hogy kerüljük az analízis eszközeinek használatát és minimálisra redukáljunk minden számítást. Eredményeinket sokféle más úton is lehet számolással igazolni. Erre itt nem fogunk kitérni.

A dolgozatban szereplő tételek jórészt ismertek, bár néhányat közülük nem találtunk meg explicit alakban az irodalomban. Cikkünkben főleg arra törekedtünk, hogy a rendezett minták elméletének elemi és rendszeres tárgyalását adjuk. Dolgozatunknak tehát főleg módszertani érdekessége van. Az irodalomra vonatkozóan S. S. WILKS [3] cikkének bibliográfiájára és RÉNYI ALFRÉD [4] dolgozatára utalunk.

1. Abból a célból, hogy eloszlásmentes eredményekre jussunk, azaz olyan eredményekre, melyek nem függenek az $F(x)$ eloszlásfüggvénytől, bevezetjük az $\eta_k = F(\xi_k)$ ($k = 1, 2, \dots, n$) valószínűségi változókat. Ha feltesszük, hogy az $y = F(x)$ függvény szigorúan monoton növekvő x azon értékeire, amelyekre $0 < F(x) < 1$ és folytonos, az $x = F^{-1}(y)$ inverz függvény is ilyen és így**

$$P(\eta_k < x) = P(\xi_k < F^{-1}(x)) = F(F^{-1}(x)) = x \quad (0 \leq x \leq 1)$$

Ebből kitűnik, hogy az $\eta_1, \eta_2, \dots, \eta_n$ változók egyenletes eloszlásúak a $(0, 1)$ intervallumban. Az $\eta_k^* = F(\xi_k^*)$ ($k = 1, 2, \dots, n$) jelöléssel:

$$\eta_k^* = F(\xi_k^*) = F(R_k(\xi_1, \xi_2, \dots, \xi_n)) = R_k(F(\xi_1), F(\xi_2), \dots, F(\xi_n)) = R_k(\eta_1, \eta_2, \dots, \eta_n).$$

* Amennyiben az $x_1, x_2, \dots, x_n$ számok között egyenlők is vannak, minden szám annyiszor számlálандó, ahányszor előfordul.

** $P(A)$ az A esemény valószínűségét jelenti.

Az $\nu_1^*, \nu_2^*, \dots, \nu_n^*$ változókat tehát úgy tekinthetjük, mint a $(0, 1)$ intervallumban egyenletes eloszlású statisztikai sokaságból vett n -elemű rendezett minta elemeit.

A következőkben az $\nu_1^*, \nu_2^*, \dots, \nu_n^*$ rendezett minta vizsgálatára szorítkozhatunk: eredményeink könnyen átfogalmazhatók az eredeti $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ rendezett mintára vonatkozó összefüggésekké.

2. Az ν_k^* változók nem függetlenek, mint azt az $\nu_j^* < \nu_k^*$, ha $j < k$, relációkból láthatjuk.

Az $\nu_1^*, \nu_2^*, \dots, \nu_n^*$ változók együttes sűrűségfüggvénye:

$$(1) \quad f(x_1, x_2, \dots, x_n) = n! \quad (0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1).$$

Valóban, ha E a $0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$ egyenlőtlenségekkel definiált n -dimenziós szimplex tetszőleges mérhető részhalmaza, akkor

$$(2) \quad P((\nu_1^*, \nu_2^*, \dots, \nu_n^*) \in E) = \sum P((i_{i_1}, i_{i_2}, \dots, i_{i_n}) \in E),$$

ahol az összegezést az $1, 2, \dots, n$ számok összes $i_1, i_2, \dots, i_n$ permutációira kell kiterjeszteni. Mivel az $i_{i_1}, i_{i_2}, \dots, i_{i_n}$ változók együttes sűrűségfüggvénye a $0 \leq x_k \leq 1$ ($k = 1, 2, \dots, n$) egyenlőtlenségekkel definiált egységkocka tetszőleges $(x_1, x_2, \dots, x_n)$ pontjában 1-gyel egyenlő, (2)-ből (1) egyszerűen következik.

Vizsgáljuk most azt az esetet, amikor az $\nu_k^* = c_k, \nu_{k+1}^* = c_{k+1}, \dots, \nu_n^* = c_n$ értékek rögzítve vannak ($0 \leq c_k \leq c_{k+1} \leq \dots \leq c_n \leq 1$). Ekkor az $\nu_1^*, \nu_2^*, \dots, \nu_{k-1}^*$ változókat a $(0, c_k)$ intervallumban egyenletes eloszlású statisztikai sokaságból vett $k-1$ -elemű rendezett minta elemeiként foghatjuk fel. Az $\nu_k^* = c_k, \dots, \nu_n^* = c_n$ esemény úgy valósulhat meg, hogy az $\nu_1, \nu_2, \dots, \nu_n$ változók közül tetszőleges $n-k+1$ változót kiválasztunk és azoknak a $c_k, \dots, c_n$ értékeket adjuk, a többi $k-1$ változóra pedig csak azt a megkötést tesszük, hogy egyikük se legyen nagyobb c_k -nél. Így a kiválasztott $k-1$ változó mindegyike egyenletes eloszlású a $(0, c_k)$ intervallumban.

Ennélfogva (1) szerint az $\nu_1^*, \dots, \nu_{k-1}^*$ változók együttes feltételes sűrűségfüggvénye az $\nu_k^* = c_k, \dots, \nu_n^* = c_n$ feltételek mellett:

$$(3a) \quad f(x_1, x_2, \dots, x_{k-1} | c_k, \dots, c_n) = \frac{(k-1)!}{c_k^{k-1}} \quad (0 \leq x_1 \leq \dots \leq x_{k-1} \leq c_k).$$

Hasonlóképpen látható be, hogy ha az $\nu_1^* = c_1, \dots, \nu_k^* = c_k$ értékeket rögzítjük, akkor az $\nu_{k+1}^*, \dots, \nu_n^*$ változókat a $(c_k, 1)$ intervallumban egyenletes eloszlású statisztikai sokaságból vett $(n-k)$ -elemű rendezett minta elemeiként foghatjuk fel és az $\nu_{k+1}^*, \dots, \nu_n^*$ változók együttes feltételes sűrűségfüggvénye az $\nu_1^* = c_1, \dots, \nu_k^* = c_k$ feltételek mellett:

$$(3b) \quad f(x_{k+1}, \dots, x_n | c_1, \dots, c_k) = \frac{(n-k)!}{(1-c_k)^{n-k}} \quad (c_k \leq x_{k+1} \leq \dots \leq x_n \leq 1).$$

Minthogy a (3a) és (3b) sűrűségfüggvények csak c_k értékétől függnak, a fenti két állítás akkor is igaz, ha csak az $\nu_k^* = c_k$ feltételt tesszük fel, azaz a

(3a) és (3b) egyenletek az $f(x_1, \dots, x_{k-1}|c_k)$ és $f(x_{k+1}, \dots, x_n|c_k)$ feltételes sűrűségfüggvényeket is megadják, továbbá ugyanezeket a feltételes sűrűségfüggvényeket nyerjük akkor is, ha a bennük nem szereplő változókra vonatkozólag tetszőleges más feltevést teszünk. Ebből következik, hogy az $\eta_{ik}^* = c_k$ feltétel mellett az $(\eta_{i1}^*, \dots, \eta_{ik-1}^*)$ változók függetlenek a $(\eta_{k+1}^*, \dots, \eta_n^*)$ változóktól. Más szóval: *a rendezett minta elemei Markov-láncot alkotnak.**

3. Az $\eta_{i+1}^*, \dots, \eta_k^*$ ($1 \leq i < k \leq n$) változók együttes eloszlásának sűrűségfüggvénye:

$$(4) \quad f_{ik}(x_{i+1}, \dots, x_k) = \frac{n!}{i!(n-k)!} x_{i+1}^i (1-x_k)^{n-k} \quad (0 \leq x_{i+1} \leq \dots \leq x_k \leq 1).$$

Valóban, az $\eta_{i1}^*, \dots, \eta_{ik}^*, \eta_{k+1}^*, \dots, \eta_n^*$ változók együttes feltételes sűrűségfüggvénye az

$$(C_1) \quad \eta_{i+1}^* = x_{i+1}, \dots, \eta_k^* = x_k$$

feltételek mellett nyilván:

$$f(x_1, \dots, x_i, x_{k+1}, \dots, x_n | x_{i+1}, \dots, x_k) = \frac{f(x_1, \dots, x_n)}{f_{ik}(x_{i+1}, \dots, x_k)}.$$

Másrészt a Markov-tulajdonság miatt az $(\eta_{i1}^*, \dots, \eta_i^*)$ változók függetlenek az $(\eta_{k+1}^*, \dots, \eta_n^*)$ változóktól a (C_1) feltétel mellett, tehát

$$f(x_1, \dots, x_i, x_{k+1}, \dots, x_n | x_{i+1}, \dots, x_k) = f(x_1, \dots, x_i | x_{i+1}, \dots, x_k) f(x_{k+1}, \dots, x_n | x_{i+1}, \dots, x_k).$$

Az (1), (3a) és (3b) egyenleteket figyelembe véve megállapításaink (4)-et adják.

Az $i+1 = k$ speciális esetben (4)-ből η_k^* sűrűségfüggvényét kapjuk:

$$(5) \quad f_k(x_k) = \frac{n!}{(k-1)!(n-k)!} x_k^{k-1} (1-x_k)^{n-k} = \beta_{n,k}(x_k)$$

A rendezett minta elemei tehát béta-eloszlásúak.

Az $\eta_{k_1}^*, \dots, \eta_{k_r}^*$ változók együttes sűrűségfüggvénye

$$f_{k_1, k_2, \dots, k_r}(x_{k_1}, \dots, x_{k_r}) = \frac{n!}{(k_1-1)!(k_2-k_1-1)!\dots(k_r-k_{r-1}-1)!(n-k_r)!} x_{k_1}^{k_1-1} (x_{k_2}-x_{k_1})^{k_2-k_1-1} \dots \dots (x_{k_r}-x_{k_{r-1}})^{k_r-k_{r-1}-1} (1-x_{k_r})^{n-k_r}.$$

A bizonyítás a fentihez teljesen hasonló; a következő módosításokkal: (C_1) helyett a

$$(C_2) \quad \eta_{k_1}^* = x_{k_1}, \dots, \eta_{k_r}^* = x_{k_r}$$

feltételt tekintjük, a $(0, 1)$ intervallumot az $x_{k_1}, \dots, x_{k_r}$ osztópontokkal $r+1$ részintervallumra osztjuk, és figyelembe vesszük, hogy a Markov-tulajdonság miatt a különböző részintervallumokba eső η_i^* változók a (C_2) feltétel mellett függetlenek.

* Ezt elsőnek A. N. KOLMOGOROV jegyezte meg [1] dolgozatában.

4. Vezessük be egyszerűség kedvéért az $\nu_0^* \equiv 0$ és $\nu_{n+1}^* \equiv 1$ jelöléseket és definiáljuk a

$$\delta_k = \nu_k^* - \nu_{k-1}^* \quad (k = 1, 2, \dots, n+1)$$

valószínűségi változókat. Minthogy a $\delta_1, \delta_2, \dots, \delta_n$ változókat az $\nu_1^*, \dots, \nu_n^*$ változókból mértéktartó lineáris transzformációval nyertük, együttes sűrűségfüggvényeik a megfelelő pontokban megegyeznek egymással. A $\delta_1, \delta_2, \dots, \delta_n$ valószínűségi változók $g(y_1, \dots, y_n)$ együttes sűrűségfüggvényét tehát (1)-ből az

$$y_1 = x_1, y_k = x_k - x_{k-1} \quad (k = 2, \dots, n)$$

transzformációval nyerjük:

$$(6) \quad g(y_1, y_2, \dots, y_n) = n! \quad (y_k \geq 0; k = 1, 2, \dots, n; y_1 + y_2 + \dots + y_n \leq 1).$$

A (6) relációból következik, hogy a $\delta_1, \dots, \delta_n$ változók azonos eloszlásúak. Minthogy az egyes ν_k változók eloszlása szimmetrikus az $\frac{1}{2}$ pontra vonatkozólag, az $(\nu_1^*, \dots, \nu_n^*)$ változók együttes eloszlása is ilyen. Speciálisan $\delta_1 = \nu_1^*$ és $\delta_{n+1} = 1 - \nu_n^*$ azonos eloszlásúak. Végül is arra a következtetésre jutunk, hogy a $\delta_1, \delta_2, \dots, \delta_n, \delta_{n+1}$ változók valamennyien azonos eloszlásúak és közös sűrűségfüggvényük ν_1^* sűrűségfüggvényével azonos (lásd (5)-öt a $k = 1$ esetben). Közös várható értékük a $\delta_1 + \dots + \delta_{n+1} = 1$ reláció miatt:

$$(7) \quad M(\delta_k) = \frac{1}{n+1} \quad (k = 1, 2, \dots, n+1)$$

5. $\delta_1, \dots, \delta_{n+1}$ nemcsak azonos eloszlású, hanem egyben ekvivalens valószínűségi változók is, azaz együttes eloszlásuk invariáns a változók tetszőleges permutálásával szemben. (6)-ból ugyanis láthatjuk, hogy a $\delta_1, \dots, \delta_n$ változók együttes sűrűségfüggvénye a változók szimmetrikus függvénye, a $\delta_1, \dots, \delta_{n+1}$ változók együttes eloszlása pedig invariáns a $(\delta_1, \dots, \delta_{n+1}) \rightarrow (\delta_{n+1}, \dots, \delta_1)$ permutációval szemben: ezt a két tényt szukcesszive alkalmazva azt kapjuk, hogy a $(\delta_1, \dots, \delta_{n+1})$ változók együttes eloszlása invariáns változónak tetszőleges permutációjával szemben.

Következésképpen az

$$\nu_{i+k}^* - \nu_i^* = \delta_{i+1} + \dots + \delta_{i+k} \quad (0 \leq i < i+k \leq n)$$

különbség eloszlása csak k -tól függ, tehát megegyezik pl. ν_k^* (5) alatti eloszlásával és várható értéke (7) szerint $\frac{k}{n+1}$ -gyel egyenlő.

Speciálisan a minta $J = \nu_n^* - \nu_1^*$ terjedelmének sűrűségfüggvénye

$$\beta_{n, n-1}(x) = n(n-1)x^{n-2}(1-x) \quad (0 \leq x \leq 1),$$

és várható értéke

$$M(J) = \frac{n-1}{n+1}.$$

6. Mint láttuk, az $\nu_k^* = c_k, \dots, \nu_n^* = c_n$ feltételek mellett az ν_i^* ($1 \leq i < k \leq n$) változót úgy tekinthetjük, mint a $(0, c_k)$ intervallumból vett $(k-1)$ -elemű rende-

zett minta i -edik elemét. Így az $\frac{\nu_i^*}{\nu_k^*}$ hányados feltételes eloszlása c_k -tól sem függ, tehát ugyanaz marad, bármilyen értéket is vesznek fel a $\nu_k^*, \dots, \nu_n^*$ változók, s így sűrűségfüggvénye (5) szerint $\beta_{k-1, i}(x)$. Tehát a rendezett minta elemeiből képzett különbségek és hányadosaik is béta-eloszlásúak.

Az $\frac{\nu_k^*}{\nu_{k+1}^*}$ ($k = 1, 2, \dots, n$) változók teljesen függetlenek. Valóban, a fentiek szerint az ν_k^*/ν_{k+1}^* változó eloszlása nem függ az $\nu_{k+1}^*, \dots, \nu_n^*$ változók felvett értékeitől, tehát $\frac{\nu_k^*}{\nu_{k+1}^*}$ független az $\frac{\nu_{k+1}^*}{\nu_{k+2}^*}, \dots, \frac{\nu_n^*}{\nu_{n+1}^*}$ változóktól.

A

$$\zeta_k = \left(\frac{\nu_k^*}{\nu_{k+1}^*} \right)^k \quad (k = 1, 2, \dots, n)$$

változók egyenletes eloszlásúak a $(0, 1)$ intervallumban.* Valóban, ν_n^* eloszlásfüggvénye:

$$P(\nu_n^* < x) = \prod_{k=1}^n P(\nu_k^* < x) = x^n \quad (0 \leq x \leq 1).$$

Minthogy az $\frac{\nu_k^*}{\nu_{k+1}^*}$ változót úgy tekinthetjük, mint a $(0, 1)$ intervallumban egyenletes eloszlású statisztikai sokaságból vett k -elemű rendezett minta k -adik elemét, kapjuk, hogy

$$P\left(\frac{\nu_k^*}{\nu_{k+1}^*} < x\right) = x^k \quad (0 \leq x \leq 1),$$

amiből

$$P(\zeta_k < x) = P\left(\frac{\nu_k^*}{\nu_{k+1}^*} < \sqrt[k]{x}\right) = x \quad (0 \leq x \leq 1),$$

és éppen ezt akartuk bebizonyítani.

7. Bevezetjük a

$$(8) \quad \vartheta_k = -\log \zeta_k = -k \log \frac{\nu_k^*}{\nu_{k+1}^*} \quad (k = 1, \dots, n)$$

valószínűségi változókat. Előző eredményeink szerint ezek teljesen függetlenek és azonos eloszlásúak, mégpedig közös eloszlásfüggvényük:

$$P(\vartheta_k < x) = P(\zeta_k > e^{-x}) = 1 - e^{-x} \quad (x \geq 0)$$

tehát ϑ_k exponenciális eloszlású valószínűségi változó ($k = 1, 2, \dots, n$) és várható értéke 1. A (8) egyenletből

$$(9) \quad \ln \nu_k^* = - \sum_{j=k}^n \frac{\vartheta_j}{j} \quad (k = 1, \dots, n)$$

A rendezett minta elemeinek logaritmusai tehát nemcsak Markov-láncot, hanem *additív* Markov-láncot alkotnak, azaz teljesen független valószínűségi változók sorozatának szükcesszív részletösszegeiként állíthatók elő.

* Ez D. VAN DANTZIG és S. MALMQUIST tétele [2].

A (9) egyenlet segítségével a rendezett minta elemeit független és azonos eloszlású valószínűségi változók összegeinek egyszerű függvényeiként lehet előállítani. Ebből a tényből kiindulva a rendezett minták elméletének *határeloszlástételeit* a valószínűségszámítás centrális határeloszlástételei segítségével egyszerű és közvetlen úton lehet bebizonyítani. Ezzel foglalkozik RÉNYI ALFRÉD [4] cikke.

IRODALOM

[1] A. N. KOLMOGOROFF: Sulla determinazione empirica di una legge di distribuzione. Giorn. Istit. Ital. Att. 4 (1933), 83—91.

[2] S. MALMQUIST: On a property of order statistics from a rectangular distribution. Skand. Aktuarietidskrift, 33 (1950), 214—222. A tétel szerepel már D. VAN DANTZIG 1948-ban kiadott előadásainak sokszorosított szövegében is, amint arra D. VAN DANTZIG volt szives a figyelmünket felhívni.

[3] S. S. WILKS: Order statistics. Bull. Amer. Math. Soc. 54 (1948), 6—50.

[4] A. RÉNYI: On the theory of order statistics. Acta Math. Acad. Sci. Hung. 4 (1953), 191—231.