

A VALÓSZÍNŰÉGSZÁMÍTÁSI INFORMÁCIÓFOGALOM NÉHÁNY KÉRDÉSÉRŐL

Írta: VINCZE ISTVÁN

A $(p_1, p_2, \dots, p_n, \sum_1^n p_i = 1)$ diszkrét valószínűségeloszlás entrópiáját, azaz bizonytalanságának mértékét SHANNON [4] a

$$-\sum_1^n p_i \log p_i$$

formulával definiálta. Folytonos változó esetére az ún. Boltzmann-féle H -függvényt adta:

$$-\int_{-\infty}^{\infty} f(x) \log f(x) dx,$$

ahol $f(x)$ a valószínűségi változó sűrűségfüggvénye. SHANNON és más szerzők is kifejezésre juttatják, hogy a H -függvény nem tekinthető a diszkrét esetre adott fenti formula folytonos esetre való kiterjesztésének. [5, 6] dolgozatainkban rámutatunk arra, hogy a Shannon-féle diszkrét entrópia folytonos eloszlású változóra való egy lehető, közvetlen és természetes kiterjesztése az ún. I -divergencia kifejezéshez vezet, amelyben az eredeti változó sűrűségfüggvényén kívül egy további sűrűségfüggvény, tehát látszólag egy új elem lép fel.

Az alábbiakban a folytonos eloszlású változó ezen információfogalmának néhány elvi és gyakorlati kérdésével foglalkozunk (1. és 2. §), majd a 3. §-ban egy alkalmazásával a matematikai statisztikában, amely a konfidenciaintervallum fogalmához vezet. Az I -divergencia számos más matematikai statisztikai alkalmazását illetően utalunk S. KULLBACK könyvére [2].

1. §. A bizonytalanság mértéke folytonos eloszlás esetén

1. Tekintsük valamely kísérlet vagy jelenség esetében a bizonyos esemény egy felbontását egymást kizáró eseményekre, vagyis egy teljes eseményrendszert:

$$\mathfrak{A}: I = A_1 + A_2 + \dots + A_n.$$

Legyenek a megfelelő valószínűségek: $p_1, p_2, \dots, p_n, \sum_1^n p_i = 1$.

Ezen eseményrendszer SHANNON által definiált entrópiája, mint ismeretes:

$$(1) \quad E_n = -\sum_1^n p_i \log \frac{1}{p_i} \quad (0 \leq E_n \leq \log n).$$

Eseményrendszer helyett beszélhetünk arról a ξ valószínűségi változóról, amely valamely $k_1, k_2, \dots, k_n$ értékeket vehet fel és amelyre az $A_i = \{\xi = k_i\}$ reláció érvényes. Az (1) formula mindenesetre invariáns a változó minden olyan transzformációjával szemben, ami különböző értékeket különböző értékekbe visz át, vagyis nem függ a változó értékészletétől, hanem csakis a valószínűségek eloszlásától.

Ha ξ folytonos eloszlású változó $F(x) = \int_{-\infty}^x f(t)dt$ eloszlásfüggvénnyel, akkor az (1) formula közvetlen alkalmazása nem vezet eredményre, mert határátmenetben végtelen értéket ad. SHANNON folytonos változó esetére az

$$E(\xi) = - \int f(x) \log f(x) dx$$

formulát ajánlotta, amely már nem rendelkezik E_n néhány kívánatos tulajdonságával, mint a nemnegativitás, vagy az invariancia a változó értéktranszformációjával szemben.

Az (1) formulának folytonos esetre való kiterjesztéséhez néhány megjegyzést teszünk:

a) Több szerzővel egyetértésben kifejezésre juttatjuk, hogy az entrópia nem az eloszlás, hanem az eseményrendszer entrópiája:

$$(1') \quad E_n(\mathfrak{A}) = \sum_1^n p_i \log \frac{1}{p_i}.$$

Ugyanazon jelenségnél általában a bizonyos esemény többféle felbontásáról beszélhetünk: az $\mathfrak{A}$ mellett az

$$\mathfrak{A}': I = A'_1 + A'_2 + \dots + A'_n,$$

$$\mathfrak{A}'': I = A''_1 + A''_2 + \dots + A''_n,$$

stb. felbontásokról. Amikor tehát az entrópia kifejezéseként, illetve értékeként az (1)-ben megjelöltet választottuk, hallgatólagosan azzal a feltételezéssel éltünk, hogy bennünket az $\mathfrak{A}, \mathfrak{A}', \mathfrak{A}'', \dots$ lehető felbontások közül az $\mathfrak{A}$ felbontás érdekel. Más esetben az entrópia értékeként ettől különböző adódhatnak.

b) Térjünk át az entrópia kifejezése helyett a „kiegészítő” fogalomra, vagyis az

$$(2) \quad I_n(\mathfrak{A}) = \log n - E_n(\mathfrak{A}) = \sum_1^n p_i \log np_i \quad (0 \leq I_n \leq \log n)$$

kifejezésre, amit az $(A_1, A_2, \dots, A_n)$ eseményrendszer információjának (röviden: információnak) fogunk nevezni.

c) Amikor a folytonos eloszlású ξ valószínűségi változó entrópiáját, illetve információját a diszkrét esetre vonatkozó (1), ill. (2) formulára akarjuk visszavezetni, nem hagyhatjuk figyelmen kívül a szóban forgó eseményteret, ill. az azon értelmezett σ -algebrát. Ez a megjegyzés — az a) alatt mondottak figyelembevételével — azt jelenti, hogy folytonos eloszlás esetén a számegyeneset nem oszthatjuk — önkényesen — ekvidisztans részekre, hanem minden n -re olyan felosztást kell tekintenünk, amely az eseményternek „bennünket érdeklő”, ill. a problémának megfelelő felosztásának felel meg.

Fenti megfontolások alapján jutottunk említett dolgozatunkban a ξ valószínűségi változóval kapcsolatos „információnak”

$$(3) \quad I_{\Phi}(\xi) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{\varphi(x)} dx$$

kifejezéséhez, ahol $f(x)$ a ξ sűrűségfüggvénye, míg a $\Phi(x) = \int_{-\infty}^x \varphi(t) dt$ eloszlásfüggvényt „érdeklődésünk eloszlásának” neveztük. E fogalommal kapcsolatban a következőket jegyzem meg:

α) A (3) alatti információt ξ -nek a $\Phi(x)$ eloszlásra vonatkozó „relatív” információjának neveztük, vagyis látszólag a Shannon-féle „abszolútnak” tekinthető információval szemben relatív fogalom. Valójában nem kevésbé abszolút, mint a diszkrét esetre vonatkozó entrópia (1) alatti kifejezése, amely a választott $\mathfrak{A}$ felbontásra vonatkoztatott; sokkal inkább tekinthető mind (1), mind (3) relatív mérőszámnak a mondott értelemben. Számos gyakorlati példa adható arra, hogy ugyanazon jelenség bennünket különböző kimenetelei (felbontásai) szempontjából érdekelhet és ez más-más entrópia-értékhez vezet.

β) A (3) alatti információ mindig nemnegatív, ami külön bizonyításra nem szorul, ez a diszkrét esetre vonatkozó hasonló tulajdonság következménye; ugyanis (3) közvetlenül adódik (2)-ből és a határátmenetnél ez a tulajdonság változatlan marad.

γ) A (3) alatti információ a változó monoton transzformációjával szemben invariáns, vagyis ha $y = t(x)$ és $t(x) > t(x')$, ha $x > x'$, akkor az $\eta = t(\xi)$ változóra:

$$I_{\Phi}(\xi) = I_{\Phi'}(\eta),$$

ahol Φ' a transzformált érdeklődéseloszlás.

Ellentétben azonban a diszkrét esettel, ahol a felbontás általában a gyakorlati problémából közvetlenül adódik, a folytonos esetben az „érdeklődés eloszlásának” $\Phi(x)$ -nek megválasztása, illetve meghatározása problematikus lehet, és (A. PEREZ egy megjegyzése szerint) hasonlóságot mutat az a priori eloszlás kérdésével a Bayes-féle problémakör esetében. Erre a kérdésre térünk a következő pontban.

2. §. Az érdeklődés eloszlásáról

Legyen ismét a ξ folytonos valószínűségi változó sűrűségfüggvénye $f(x)$, amelyről feltesszük, hogy pozitív az egész x -tengelyen, ugyancsak legyen $\varphi(x) = \Phi'(x) > 0$ $-\infty < x < +\infty$ -re. Az

$$\frac{f(x)dx}{\varphi(x)dx}$$

információelem megadja, hogy az $\{x \leq \xi < x + dx\}$ esemény esetén egységnyi érdeklődésre eloszlásunkból mennyi jut. Ennek valamely függvényének várható értéke szolgálhat az információ mértékéül:

$$\mathbf{M}_{\xi} \left[\psi \left(\frac{f(x)}{\varphi(x)} \right) \right] = \int_{-\infty}^{\infty} \psi \left(\frac{f(x)}{\varphi(x)} \right) f(x) dx.$$

Az információelméletben, fizikában és matematikai statisztikában a $\psi(x) = -\log x$ kitüntetett szerepet játszik. Más függvényeket illetően RÉNYI ALFRÉD [3] dolgozatára utalunk.

A $\Phi(x) = \int_{-\infty}^x \varphi(t) dt$ eloszlásnak eleve valamely értelmezést adni, többet, mint

az előző paragrafusban tettünk, nem volna helyes. Más-más alkalmazási területen nyilván más-más értelmezése adódik. Az alábbiakban négyféle ilyen lehető értelmezésre adunk példát: a) gazdaságossági, b) valamely jelenség megfigyelésének módjából adódó, c) a természet által nyújtott értelmezésre; végül d) valamely a sztochasztikus rendszerre vonatkozó előzetes információ szolgáltatathatja az érdeklődés eloszlását, amire a 3. §-ban térünk ki.

a) Legyen az (x, y) síkban a $\mathcal{G}$ zárt tartomány egy mezőgazdasági terület. Jelentse $f(x, y)$ a jégeső sűrűségeloszlását ezen a területen pl. július hónapban. A következő évekre vonatkozó megfontolásainkhoz az információ mértékét nyilván igen egyoldalú volna csupán az $f(x, y)$ segítségével megadni. Közelfekvő annak a $q(x, y)$ függvénynek konstrukciója és az érdeklődés mértékéül való választása, amelyre $q(x, y) dx dy$ az (x, y) pont környezetében levő terméseredmény értékével arányos. Az $f(x, y)$ és $q(x, y)$ sűrűségfüggvények erős diszkrepanciája az információ nagy értékéhez vezet, ami arra utal, hogy a várható kár nagy valószínűséggel kicsiny.

b) Az (x, y) -sík x -tengelyén véletlenszerűen pontszerű jelek keletkeznek az $f(x)$ eloszlássűrűség szerint. A megfigyelő műszer az $(x=0, y=1)$ pontban helyezkedik el és erre nézve csak az bir jelentőséggel, hogy mely irányból kapja a jelet. Ha tehát az információt (entrópiát) a műszerre vonatkozólag akarjuk mérni, akkor figyelembe kell venni, hogy bármely dx -nyílásszög ugyanazon jelentőségű, ami az x -tengelyre átszámítva a

$$q(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

Cauchy-féle eloszláshoz vezet.

Más kérdés az, hogy az így számított információ milyen objektív jelentéssel bír a megfigyelő rendszerre. Ilyennek feltételezését bizonyos esetekben jogosulttá teszi, ha arra gondolunk, hogy ugyanaz a sztochasztikus rendszer különböző fizikai rendszerekre különböző hatással lehet.

c) INGARDEN és URBANIK [1] dolgozatukban figyelemre méltó kezdeményezést tesznek az információ és valószínűség fogalmainak viszonylatában: az információ mennyiség fogalmát axiomatizálják — ezt téve elsődlegesnek — majd ebből konstruálják a valószínűség fogalmát, ill. mérőszámát. Mint említik, vannak szituációk a fizikában, ahol az információ (pl. egy makroszkopikus rendszer entrópiája) ismert és nagy pontossággal mérhető, ugyanakkor a valószínűségeloszlás lényegében ismeretlen és gyakorlatilag nem mérhető. A szerzők említett dolgozatukban hívták fel a figyelmet arra, hogy a statisztikus termodinamikában az „érdeklődés eloszlása”, mint a modell által adott eloszlás, tehát *objektív módon* jelentkezik. Amikor ugyanis az energiaeloszlás entrópiáját számítják, az energiatengelyt nem egyenlő részekre osztják, hanem annak felosztását a fázistér egyenlő térfogatú részei determinálják.¹

¹ Megjegyezzük, hogy ebben az esetben az érdeklődés eloszlása nem a szokott értelemben vett eloszlás, ugyanis végtelen intervallumon nem normálható. Erre — a fizikában gyakran előforduló — esetre jelen dolgozatunkban nem térünk ki.

3. §. Konfidenciaintervallumok

A matematikai statisztikában az információ mennyiség fogalma általában két-féle jelentésben szerepel. Az egyik esetben azt kívánjuk mérni, hogy egy eloszlás valamely paraméterére nézve bizonyos számú megfigyelés mennyi információt nyújt, aminek egyik mértéke R. A. FISHER ismert kifejezése. Másik esetben valamely ismert eloszlású változó esetén mérőszámot kívánunk adni arra a bizonytalanságra, ami a jelenséggel szemben egy következő kimenetel szempontjából fennáll, aminek mértékéül az entrópia különböző kifejezései szolgálnak. Az alábbiakban rámutatunk arra, hogy ez utóbbi felfogásból kiindulva az érdeklődés eloszlásának alkalmas megválasztásával hogyan jutunk el a konfidenciaintervallumokhoz.

Az $f(x)$ sűrűségű ξ valószínűségi változó $\Phi(x) = \int_{-\infty}^x \varphi(t) dt$ érdeklődéseloszlásra vonatkozó információja, mint láttuk

$$I_{\Phi}(\xi) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{\varphi(x)} dx.$$

Ha az információ mértéke nagy, ez nagyjából azt jelenti, hogy ξ egy jövőben megfigyelt értéke nagy valószínűséggel esik olyan tartományra, amelyen érdeklődésünk mértéke kicsiny. Gondoljuk el, hogy $F(x)$ -et nem ismerjük, de $\Phi(x)$ -et és $I_{\Phi}(\xi)$ -t ismerjük. Ekkor az $I_{\Phi}(\xi) = 0$ vagy annak kicsiny értéke azt jelenti, hogy $f(x)$ és $\varphi(x)$ megegyeznek, ill. egymáshoz közel esnek.

Függjön most $F(x)$ egyetlen a ismeretlen paramétertől ($-\infty < a < \infty$); $F(x) = F(x; a)$ és tekintsünk ξ -re nézve n számú független megfigyelést: $\xi_1, \xi_2, \dots, \xi_n$. Válasszuk érdeklődésünk eloszlásaként a

$$\Phi^*(x) = \Phi(x; a) = F(x; a)$$

eloszlásfüggvényt, ahol a az a paraméter valamely becslése:

$$a = a(\xi_1, \dots, \xi_n), \text{ vagyis legyen}$$

$$I_a(\xi) = \int_{-\infty}^{\infty} f(x; a) \log \frac{f(x; a)}{f(x; \alpha)} dx.$$

A konfidenciaintervallumok módszerének analógiájára válasszuk meg most a -t oly módon, hogy ez az információ „ne legyen túl nagy”. Ezen a következőt értjük: $I_a(F)$ nemnegatív és $a = a(\xi_1, \dots, \xi_n)$ miatt valószínűségi változó; mondhatjuk tehát, hogy ne legyen nagyobb, mint várható értékének valamilyen λ -szorosa ($\lambda > 0$). Vagyis

$$(3.1) \quad I_a(\xi) < \lambda M_a(I_a(\xi)).$$

Ez a reláció kedvező esetben a -ra nézve konfidenciaintervallumhoz vezet.² Lássunk néhány példát.

² Megfigyelhetjük, hogy az alábbi esetekben az intervallumba esés valószínűsége adott λ esetén valóban nem függ a keresett paraméter értékétől.

1. A normális eloszlás várható értéke, adott szórás mellett.

' Legyen σ_0 az ismert szórás, μ az ismeretlen várható érték, $\alpha = \bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$,

$$F(x; \mu) = \frac{1}{\sqrt{2\pi} \sigma_0} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma_0^2}} dt,$$

$$\Phi^*(x) = F(x; \bar{\xi}) = \frac{1}{\sqrt{2\pi} \sigma_0} \int_{-\infty}^x e^{-\frac{(t-\bar{\xi})^2}{2\sigma_0^2}} dt,$$

$$I_{\Phi^*}(F) = \int_{-\infty}^{\infty} f(x; \mu) \log \frac{f(x; \mu)}{f(x; \bar{\xi})} dx = \frac{(\mu - \bar{\xi})^2}{2\sigma_0^2}.$$

Mint hogy $\bar{\xi}$ sűrűségfüggvénye

$$f_n(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi} \cdot \sigma_0} e^{-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}},$$

tehát

$$\mathbf{M}[I_{\Phi^*}(F)] = \frac{1}{2n}$$

vagyis a (3. 1) reláció a Neyman-féle

$$(\mu - \bar{\xi})^2 < \lambda \frac{\sigma_0^2}{n}$$

kofidenciaintervallumhoz vezet, ami esetünkben egyértelműen és szimmetrikusnak adódik.

2. A normális eloszlás szórása ismert várható érték esetén. Legyen

$$s = s_n = \sqrt{\frac{\sum_{i=1}^n (\xi_i - \bar{\xi})^2}{n}}.$$

Ekkor

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$\varphi^*(x) = f(x; \mu, s) = \frac{1}{\sqrt{2\pi} s} e^{-\frac{(x-\mu)^2}{2s^2}},$$

$$I_{\Phi^*}(F) = \int_{-\infty}^{\infty} f(x; \mu, \sigma) \log \frac{f(x; \mu, \sigma)}{f(x; \mu, s)} dx = \frac{1}{2} \left(\log \frac{s^2}{\sigma^2} + \frac{\sigma^2}{s^2} - 1 \right).$$

Ennek várható értékére a következőt nyerjük:

$$\mathbf{M} \left[\frac{1}{2} \left(\log \frac{s^2}{\sigma^2} + \frac{\sigma^2}{s^2} - 1 \right) \right] = \frac{1}{n-2} + \frac{1}{2} \left(\psi \left(\frac{n}{2} \right) - \log \frac{n}{2} \right) = c_n,$$

ahol $\psi(x)$ az ismert Euler-féle ψ -függvény,

$$\psi(x) = \frac{d}{dx} \log \Gamma(x), \quad \psi(n) = \sum_{k=1}^{n-1} \frac{1}{k} - c,$$

c az Euler-féle állandó.

A szórásra innen az

$$1 < \frac{\sigma^2}{s^2} e^{\frac{s^2}{\sigma^2}} < e^{2\lambda c_n + 1}$$

relációt kapjuk és ha az $y = g(x) = \frac{1}{x} e^x$ kétértékű inverz függvényét $x = g_1^{-1}(y)$, $x < 1$ és $x = g_2^{-1}(y)$, $x > 1$ jelöli, akkor σ^2 -re a nem szimmetrikus

$$s^2 g_1^{-1}(2\lambda c_n + 1) < \sigma^2 < s^2 g_2^{-1}(2\lambda c_n + 1)$$

konfidenciaintervallumot nyerjük. Minthogy c_n nagyságrendje $\frac{1}{n}$, ennél fogva a σ^2 -re kapott konfidenciaintervallum hossza $\frac{1}{\sqrt{n}}$ rendben tart adott λ mellett 0 felé, ha $n \rightarrow \infty$.

3. Az exponenciális eloszlás paramétere.

$$f(x; \lambda) = \lambda e^{-\lambda x},$$

$$\lambda \sim \frac{1}{\bar{\xi}} = \frac{n}{\xi_1 + \xi_2 + \dots + \xi_n},$$

$$\varphi^*(x) = \frac{1}{\bar{\xi}} e^{-\frac{x}{\bar{\xi}}},$$

$$I_{\Phi^*}(\xi) = \int_0^{\infty} f(x; \lambda) \log \frac{f(x; \lambda)}{f\left(x; \frac{1}{\bar{\xi}}\right)} dx = \log \lambda \bar{\xi} + \left(\frac{1}{\lambda \bar{\xi}} - 1 \right),$$

$$\mathbf{M}[I_{\Phi^*}(\xi)] = \frac{1}{n-1} + \psi(n) - \log n,$$

ahonnan $\frac{1}{\lambda}$ -ra a 2.-ben adott módon kapjuk az ismert konfidenciaintervallumot $\bar{\xi}$ segítségével.

IRODALOM

- [1] INGARDEN, R. S.—URBANIK, K.: Information without probability. *Colloquium Mathematicum* 9 (1962), 131—150 o.
- [2] KULLBACK, S.: *Information Theory and Statistics*. Wiley-New-York, 1959.
- [3] RÉNYI, A.: On measures of entropy and information. *Fourth Berkeley Symposium on Probability and Statistics*, 1960. Vol. I. pp. 547—561.
- [4] SHANNON, C. E.: A mathematical system of communication. *Bell System of Technical Journal*, 27 (1948), 379—423 o. és 623—656 o.
- [5] VINCZE, I.: Az információelmélet egy fogalmának értelmezéséről. *Matematikai Lapok* 10 (1959), 255—266 o.
- [6] VINCZE, I.: An interpretation of the I -divergence of information theory. *Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, from June 1 to 6. 1959. Prága 1960, 681—684. o.

(Beérkezett: 1961. X. 16.)