Szegedy Miklós

Megoszlási strukturák vizsgálata információelméleti mérőszámokkal

A jövedelemegyenlőtlenség, az ipari koncentráció, bizonyos gazdasági mennyiségek területi vagy egyéb megoszlásai stb. elemzésének közös jellemzője, hogy egy egységnek tekinthető mennyiség részesedési arányai, illetve ezeknek az arányoknak a változásai, eltérései képezik a kvantitatív vizsgálat tárgyát. A sajátos gazdasági tartalomtól elvonatkoztatott, matematikai tárgyalás esetén az egységnek a részesedési arányai által meghatározott felbontását *megoszlási struktúrának* nevezzük. A gazdasági életben előforduló megoszlási struktúrákkal kapcsolatos mérésekhez számos eljárás áll rendelkezésünkre. Ezek között újabban egyre nagyobb teret nyer az információelméletből merített mutatószámok alkalmazása is, főleg azért, mert e mutatók, azon kívül, hogy csak egyszerű számítások elvégzését igénylik, visszavezethetők a struktúra részein belüli és a részek közötti viszonyokat kifejező mérőszámok összegére, ami a nem-információelméleti mérőszámokra általában nem érvényes. Ezért hasznos, ha betekintést nyerünk az információelmélet eredményeinek gazdasági alkalmazhatósága kérdéseibe azáltal, hogy megsimerkedünk néhány alapvető mérőszámmal és ezek tulajdonságaival a gazdasági értelmezhetőség szempontjából. Cikkünk éppen ezt a célt szeretné szolgálni.

Az információelmélet elnevezés, bár több értelemben használatos, általánosan elfogadottan a technikai hírközlés statisztikai elméletének egy területét jelöli. Az információelmélet eredete a statisztikus termodinamikába nyúlik vissza, ahol az entrópia alapvető jelentőségű fogalma kialakult.

Az elmélet úttörője R. V. L. Hartley volt az 1928-ban megjelent tanulmányával. Az elmélet tulajdonképpeni megalapítójának C. E. Shannon tekinthető, akinek 1948-ban jelent meg dolgozata erről a témáról. Shannon számos új fogalmat vezetett be és a tiszta matematikai kutatás számára is új lehetőségeket nyitott meg. Ma már az információelmélet ebben a tisztán matematikai értelemben a valószínűségelmélet egyik rohamosan fejlődő új ágának számít. Az elmélet bizonyos fogalmai és összefüggései annyira általánosak, hogy sokkal szélesebb körben értelmezhetők, mint a híradástechnika, így a gazdasági elemzésben is, amint azt igen kiterjedten és részletesen fejtegeti H. Theil Közgazdaságtan és információelmélet c. könyve.

Jelen cikk főleg Theil előbb említett művére támaszkodik, felhasználva egyéb forrásokat is (l. irodalomjegyzék). Annak a célkitűzésnek megfelelően, hogy figyelmünket csak néhány alapvető és egyszerű mérőszámra koncentrálva részletesen vizsgáljuk ezek fogalmát, tulajdonságait és egymással való kapcsolatukat, szükségesnek mutatkozott saját észrevételeinkkel kiegészíteni a tárgyalást, így az egyedi információ gazdasági értelmezhetőségét illetően, a mérőszámok dezaggregálhatóságára, ill. többszörös dezaggregálására vonatkozólag, de főképpen annak elemzésével, hogyan követik az ipari koncentráció mérőszámai a vállalatok egyesülésének folyamatát. Az ipari koncentráció egyik használatos mérőszámával, a vállalatok egyenletességi egyenértékszámával kapcsolatban egy, a koncentráció közvetlen kifejezésére alkalmas mutatót is bevezettünk, és ezt a vállalatok egyenlőtlenségi arányszámának neveztük.

A tárgyalásra kerülő fogalmak jobb megértése végett maradjunk egyelőre a híradástechnikai értelmezésnél. Tegyük fel, hogy véges sok különböző fajta hír vétele történik úgy, hogy minden egyes hírfajta vétele meghatározott valószínűséggel következik be. A hírfajták teljes sorozata tehát egy valószínűségeloszlást határoz meg, vagyis olyan $x_1, x_2, \ldots, x_n$ számok sorozatát, amelyekre

$$0 < x_i \leq 1 \quad \text{és} \quad \sum_{i=1}^{n} x_i = 1.$$

Minél nagyobb valószínűségű hír érkezik, annál kevésbé vagyunk meglepve, annál kisebb a hír *információtartalma*. Tehát egyetlen hír információtartalma vagy más szóval *egyedi formációj*a az illető hír valószínűségének csökkenő függvénye. Ha még megköveteljük az egyedi információ szemléletes fogalma alapján, hogy ez a mennyiség csakis a hír valószínűségétől függjön, hogy legyen folytonos a $0 < x \leq 1$ értelmezési tartományban, hogy ha az esemény valószínűsége 1, akkor a függvényérték legyen 0 (hiszen az ilyen hír vétele egyáltalán nem meglepő, vagyis semmiféle új információt nem tartalmaz), hogy 0-nál a függvény határértéke legyen ∞ (hiszen a 0 valószínűségű hír vétele „végtelenül" meglepő lenne), végül, hogy bármely két független hír vételének információtartalma legyen a két hír egyedi információjának összege (mivel minden további független információ additíve növeli az addigi információt), — akkor e követelmények egyértelműen meghatározzák az egyedi információ függvényének képletét:

$$h(x) = c \log \frac{1}{x}$$

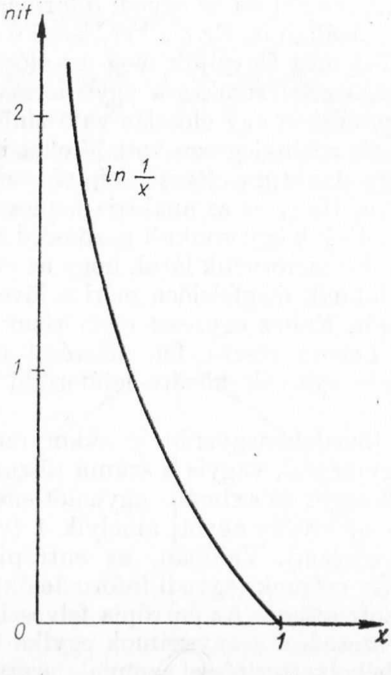
függvénytípust, ahol x független változó a hír vételének valószínűsége, a logaritmus alapja és c konstans értéke pedig tetszőleges. A gyakorlatban az egyszerűség kedvéért c értékét 1-nek szokták választani; ekkor, ha a logaritmus alapja 2, az egyedi információ egységét *bit*-nek nevezzük, természetes logaritmus esetén pedig az egység neve *nit*. Tehát például egy 0,5 valószínűségű hír információtartalma 1 bit, ami közelítőleg 1/1,433 azaz 0,693 nitnek felel meg (viszont 1 nit $\approx$ 1,443 bit). Matematikai levezetéseknél célszerűbb a természetes logaritmus használata, míg a gyakorlati számítások eredményeit inkább bit-ben szokták kifejezni.

Az egyedi információ, vagyis a $h(x) = \log 1/x$ függvény menetét az 1. ábra mutatja.

Ezek után képzeljük el azt, hogy nagyon sok esetben kapunk hírt, és a különböző fajta hírek előfordulási gyakoriságai jó közelítésben megegyeznek a hírek vételi valószínűségeivel. Ha kvantifikálni akarjuk a hírek okozta meglepődésünk átlagos szintjét, akkor az egyes meglepődéseink mértékeinek, az egyedi információknak az átlagát kell vennünk, ami a hírfajták információtartalmának a (gyakoriságnak megfelelő) valószínűséggel súlyozott átlaga:

$$H(x) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i} = - \sum_{i=1}^{n} x_i \log x_i$$

az $x_i (i = 1, 2, \ldots, n)$ valószínűségeloszlás esetén.



1. ábra. Az egyedi információ függvénye

A $H(x)$ függvényt értelmezhetjük abban az esetben is, amikor egy vagy több hír vételi valószínűsége 0. Figyelembe véve, hogy

$$\lim_{x \to 0} x \log \frac{1}{x} = 0,$$

megállapodunk abban, hogy a $H(x)$ függvényben azok a tagok, amelyekben az $x_i = 0$, 0-nak számítanak. A meghatározott valószínűséggel rendelkező hírfajták teljes rendszerének átlagos vételi bizonytalanságát jellemző $H(x)$ mennyiséget nevezzük entrópiának.

Az entrópiát mint az egyedi információk súlyozott átlagát szintén az egyedi információ egységében adjuk meg. Ha tehát az entrópia képletében szereplő logaritmus alapja 2, a kifejezés értékét bit-ben kapjuk meg.

Világos, hogy egy hírfajta-rendszer valószínűségeloszlása matematikailag pontosan azoknak a követelményeknek tesz eleget, amelyeknek egy megoszlási struktúra, hiszen ennek a részesedési arányszámai is 1-nél nem nagyobb nemnegatív számok és összegük 1. Persze, ha megoszlási struktúráról van szó, akkor az $x_i (i = 1, 2, \ldots, n)$ sorozat tagjait nem valószínűségeknek, hanem részesedési arányszámoknak, részarányoknak vagy megoszlási viszonyszámoknak nevezzük. Az $x_1, x_2, \ldots, x_n$ sorozat ilyen átértelmezése esetén az egyedi információ mint egy részarány nagyságának csökkenő függvénye csak azt fejezi ki, hogy az illető részarány „milyen nagyon" kicsi. Ez önmagában nem volna túlságosan érdekes, de ha figyelembe vesszük, hogy egy megoszlási struktúra egyenletességét a többi részarányhoz viszonyítva igen kis részarányok milyen nagyon csökkentik, akkor kezdjük megérteni az egyedi információ jelentőségét a megoszlási struktúra vonatkozásában is. Erre a kérdésre később még visszatérünk.

A részletes elemzés előtt még figyeljük meg a valószínűségeloszlás átlagos bizonytalansága és a megoszlási struktúra egyenletessége közötti analógiát. Minél közelebb esnek egymáshoz egy eloszlás valószínűségei, átlagosan annál bizonytalanabbak vagyunk a ténylegesen vett híreket illetően; és minél közelebb esnek egymáshoz egy struktúra részesedési arányai, annál egyenletesebbnek mondható a struktúra. Hogy ez az analógia nemcsak a felületes szemlélet eredménye, azt azonnal belátjuk egy konkrét gazdasági struktúra, a jövedelemeloszlás vizsgálata révén. Be szeretnénk látni, hogy az entrópia függvény valóban szemléletes fogalmainknak megfelelően méri a jövedelemeloszlás egyenletességét vagy egyenlőségét. Ehhez egyrészt ellenőrizni kell, hogy az entrópia szélsőértékeit megfelelő helyen veszi-e fel, másrészt megfelelően követi-e a függvény változása a részarányok között lejátszódó kiegyenlítődési folyamatokat.

Nyilvánvaló, hogy a jövedelemegyenlőség akkor maximális, ha az összes részarány megegyezik egymással, vagyis n számú részarány esetén mindegyik $1/n$; a minimális egyenlőséget (maximális egyenlőtlenséget) pedig az jelenti, ha mindegyik részarány 0, kivéve egyet, amelyik 1 (vagyis egyetlen kézben összpontosul az összjövedelem). Valóban, az entrópia értéke sosem lehet negatív, mert nem-negatív számok (egyedi információk) nem-negatív számokkal (részarányok) súlyozott átlaga. Az entrópia felveszi a 0 értéket, mégpedig pontosan akkor, ha a részesedési arányszámok egyike 1, a többi 0, amiről az entrópia képletébe való behelyettesítéssel azonnal meggyőződhetünk. Az entrópia minimuma tehát 0, és ez éppen a megkívánt esetben következik be. Az entrópia maximumát a Lagrange-féle multiplikátor-módszerrel határozhatjuk meg:

A feltételi egyenlet:

$$\sum_{i=1}^{n} x_i = 1.$$

Ezt 0-ra redukálva, az így nyert kifejezést λ-val, a Lagrange-féle multiplikátorral szorozva és az entrópia értékéből kivonva kapjuk a következő kifejezést:

$$-\sum_{i=1}^{n} x_i \log x_i - \lambda \left(\sum_{i=1}^{n} x_i - 1 \right).$$

Ennek $x_i (i = 1, 2, \ldots, n)$ szerinti deriváltját 0-val téve egyenlővé azt kapjuk, hogy

$$-1 - \log x_i - \lambda = 0 \qquad (i = 1, 2, \ldots, n)$$

ahol a logaritmus most természetes logaritmust jelent. (Ha más alapot használunk, akkor a $-1 - \log x_i$ egy konstanssal szorzódik, ami következtetésünkre nincs kihatással.) Tehát minden i-re

$$x_i = e^{-1-\lambda},$$

amiből az következik, hogy minden egyes x_i ugyanazzal a konstanssal, így egymással is egyenlő, vagyis

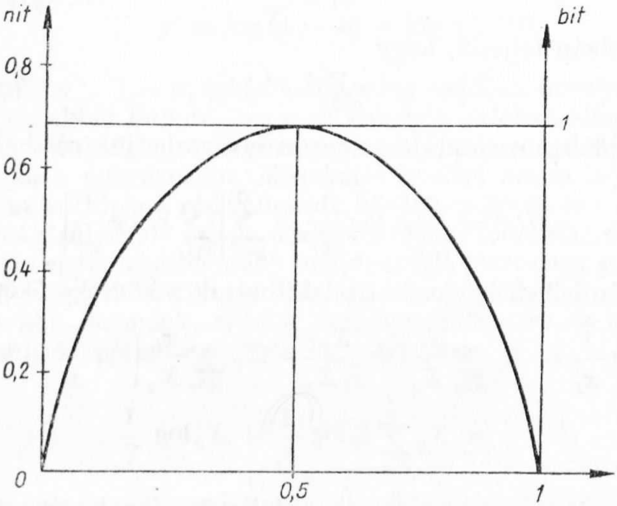
$$x_i = \frac{1}{n} \qquad (i = 1, 2, \ldots, n).$$

Ebben az esetben az entrópia értéke:

$$\sum_{i=1}^{n} x_i \log \frac{1}{x_i} = \sum_{i=1}^{n} \frac{1}{n} \log n = \log n,$$

ez a függvény maximuma, összhangban a szemlélet által előírt követelménnyel.

Az entrópia függvény menetét a 2. ábra mutatja abban a speciális esetben, amikor $n = 2$, $x_1 = x$ és $x_2 = 1 - x$.



2. ábra. Az entrópia függvénye ($n = 2$)

Problematikus lehet, hogy a függvény maximuma függ a struktúrát alkotó megoszlási viszonyszámok számától. Ha például a struktúrát csak két viszonyszám alkotja, akkor a teljes egyenlőségnek ($x_1 = x_2 = 1/2$) megfelelő entrópia 1 bit, ha viszont 1024 viszonyszámból áll a struktúra, akkor a maximális egyenlőség ($x_1 = x_2 = \ldots = x_{1024} = 1/1024$) entrópiája 10 bit. Mindkét esetben teljes egyenlőségről van ugyan szó, mégis, hogy 1024 jövedelem egyenlő egymással, az többet jelent, mint az, hogy csupán két jövedelem egyenlő egymással. (Úgy is mondhatnánk, hogy 1024 viszonyszám megegyezése több egyenlőségi relációt foglal magában, mint két viszonyszámé.) Ha tehát a jövedelemegyenlőség fogalmába beleértjük a viszonyszámok számából eredő hatást is, akkor az entrópia alkalmas mutatónak bizonyul.

Előfordulhat, hogy a mérésből ki akarjuk küszöbölni ezt a hatást; ekkor az entrópiát elosztjuk maximumával:

$$H_r(x) = \frac{H(x)}{\log n} \qquad (n > 1),$$

a hányadost *relatív entrópiá*nak nevezzük. Értéke a 0 és 1 határok között mozog. (Ha $n = 1$, akkor nincs értelmezve.)

Ahhoz, hogy megvizsgálhassuk, hogyan változik az entrópia értéke két részarány kiegyenlítődése közben, először látnunk kell az entrópia felbontását.

A teljes struktúrát alkotó $x_1, x_2, \ldots, x_n$ arányszámok (ill. ezek indexei) tartozzanak G számú diszjunkt halmazba; jelöljük ezeket $S_1, S_2, \ldots, S_G$-vel. Ekkor az S_g részhalmaz részesedési arányszáma az alaphalmazhoz viszonyítva:

$$X_g = \sum_{i \in S_g} x_i \qquad (g = 1, 2, \ldots, G).$$

Az S_g részhalmazon belüli részesedési arányszámok:

$$\xi_i = \frac{x_i}{X_g} \qquad (i \in S_g;\, g = 1, 2, \ldots, G)$$

amelyekre nyilván teljesül, hogy

$$\sum_{i \in S_g} \xi_i = 1 \qquad (g = 1, 2, \ldots, G).$$

A $H(x)$ entrópiában szereplő összegezést elvégezhetjük részhalmazok szerint csoportosítva:

$$H(x) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i} = \sum_{g=1}^{G} \left[\sum_{i \in S_g} x_i \log \frac{1}{x_i} \right].$$

A szögletes zárójelbeli kifejezést átalakíthatjuk a következőképpen:

$$\sum_{i \in S_g} x_i \log \frac{1}{x_i} = X_g \sum_{i \in S_g} \frac{x_i}{X_g} \log \frac{X_g}{x_i X_g} = X_g \sum_{i \in S_g} \frac{x_i}{X_g} \left(\log \frac{X_g}{x_i} + \log \frac{1}{X_g} \right) =$$

$$= X_g \sum_{i \in S_g} \xi_i \log \frac{1}{\xi_i} + X_g \log \frac{1}{X_g}.$$

A $\sum\limits_{i \in S_g} \xi_i \log \frac{1}{\xi_i}$ kifejezést részhalmazon belüli entrópiának nevezzük és $H_g(x)$-szel jelöljük, így

$$H(x) = \sum_{g=1}^{G} X_g \log \frac{1}{X_g} + \sum_{g=1}^{G} X_g H_g(x),$$

ahol a jobb oldal első tagja a *halmazközi entrópia*, második tagja a *részhalmazon belüli entrópiák súlyozott átlaga* (a súlyok az egyes részhalmazok részesedési arányszámai).

Most tegyük fel, hogy adott jövedelemeloszlás esetén két alany részesedése nem egyezik meg, és jövedelmük olyan módosuláson megy keresztül, a többi részesedést változatlanul hagyva, hogy a kisebb jövedelemmel rendelkező jövedelme javul a nagyobb jövedelmű társa rovására mindaddig, amíg kettőjük jövedelmének teljes kiegyenlítődése be nem következik. Vajon az entrópia függvény e kiegyenlítődést követi-e azáltal, hogy egyre nagyobb értéket vesz fel? Az entrópia felbontására támaszkodva igenlő választ adhatunk.

Álljon ugyanis S_1 részhalmaz abból a két részarányból, amelyeknek összege változatlan, de fokozatosan kiegyenlítődnek egymással, míg a többi részarány tartozzék S_2-be. Ekkor az entrópia:

$$H(x) = \sum_{g=1}^{2} X_g \log \frac{1}{X_g} + \sum_{g=1}^{2} X_g H_g(x).$$

Ebben a kifejezésben csak a kételemű S_1 részhalmaz entrópiája, $H_1(x)$ változik a részarányok eltolódása következtében. Vizsgáljuk meg a $H_1(x)$ függvény menetét, bevezetve a következő jelöléseket. Legyen a „szegényebb" alany részaránya $x_1 = x$, a „gazdagabbé" $x_2 = 1 - x$, az entrópia $H_1(x) = y$. Ekkor

$$y = x \log \frac{1}{x} + (1 - x) \log \frac{1}{1 - x}.$$

(Ennek a függvénynek a képét láthatjuk a 2. ábrán.) A kérdéses intervallum 0-tól 0,5-ig terjed. Ebben az intervallumban a függvény felveszi minimumát ($x = 0$-nál) és maximumát ($x = 0{,}5$-nél). Mivel (természetes logaritmus esetén)

$$y' = \log (1 - x) - \log x > 0$$

mindaddig, míg $x > 1 - x$, tehát a függvény valóban növekedést mutat.

Az elmondottakból látható, hogy az entrópia valóban elfogadható a jövedelemeloszlási struktúra — és persze bármely más megoszlási struktúra — egyenletességének mértékeként. Megokolás közben azt is láthattuk, hogyan bontható fel az entrópia a részhalmazok közötti entrópia és a részhalmazokon belüli entrópiák súlyozott átlaga összegére. Ez a felbonthatóság az entrópia mint mérőszám egyik legértékesebb tulajdonsága, mert nagy segítséget nyújt a struktúrában megnyilvánuló egyenletesség elemzéséhez.

Mielőtt tovább mennénk, álljunk meg egy pillanatra az entrópiát alkotó egyes tagoknál mint függvényeknél. Egy ilyen tag

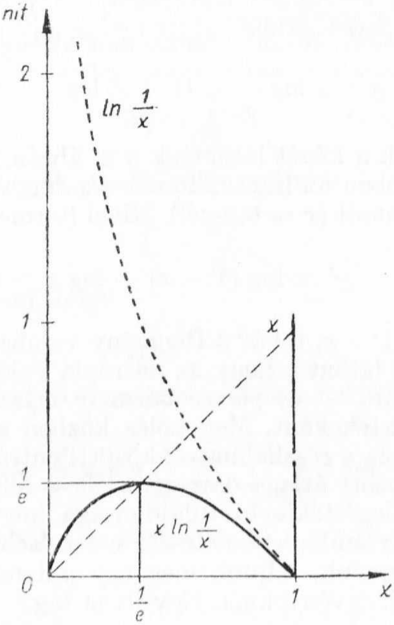
$$y = x \log \frac{1}{x}$$

egy részaránynak és a részarány egyedi információjának szorzata. Mint láttuk, az egyedi információ a részarány kicsiségének mértékéül is szolgálhat, maga a részarány pedig saját nagyságának mérőszáma. Ha szorzatukat mint függvényt diszkutáljuk a 0 és 1 határok között, megállapíthatjuk, hogy minimumát, 0-t a végpontoknál veszi fel, és a végpontoktól az intervallum belseje felé haladva növekedik az $x = 1/e$ pontig (a logaritmus alapjának megválasztásától függetlenül), ahol eléri maximumát ($y = 1/e$). A függvény menetét a 3. ábra mutatja (természetes logaritmus esetében).

Durván fogalmazva egy ilyen tag annál kevésbé növeli a struktúra entrópiájának értékét, minél jobban távolodik x a maximumhelytől, vagyis minél végletesebbé (vagy nagyon naggyá, vagy nagyon kicsivé) válik. Ezzel az egyedi információ átértelmezését próbáltuk egy kissé megvilágítani, bár ennek önmagában nincs jelentősége, csak az entrópia szempontjából.

Az entrópiát tehát a megoszlási struktúrán belül megnyilvánuló egyenletesség mérőszámának tekinthetjük. Felmerülhet azonban két struktúra összehasonlításának kérdése is, vagyis ha adva van két azonos számú részesedési arányszámból álló megoszlási struktúra: $x_1, x_2, \ldots, x_n$ és $y_1, y_2, \ldots, y_n$, amelyek között kölcsönösen egyértelmű megfeleltetés áll fenn úgy, hogy az azonos indexű arányszámok felelnek meg egymásnak, akkor kérdezhetjük, hogy mekkora a két struktúra közötti eltérés nagysága. Az információelmélet ehhez is kielégítő mérőszámot szolgáltat: az *információpontatlanságot* (vagy információ-divergenciát); képlete:

$$I(y{:}\, x) = \sum_{i=1}^{n} y_i \log \frac{y_i}{x_i}.$$



3. ábra. Az $x \log \frac{1}{x}$ *függvény és komponensei a (0,1) intervallumban*

Az összehasonlítás iránya nem közömbös, mert az x-ek és y-ok szerepe általában nem cserélhető fel a kifejezés értékének megváltozása nélkül. Az információelméletben az x értékek a hírek korábbi (a priori), az y értékek pedig későbbi (a posteriori) valószínűségeit jelölik. Ezt a gondolatot megtartva a gazdasági alkalmazásoknál is az x-ek időben vagy átvitt értelemben korábbi, az y-ok későbbi részesedési arányszámokat jelölnek.

Vizsgáljuk meg, hogy az információpontatlanság valóban alkalmas-e struktúrák egymástól való eltérésének mérésére. Ebből a szempontból az a követelmény látszik a legfontosabbnak, hogy ha két tetszőleges indexre $y_i > x_i$ és $y_j < x_j$, és a többi arányszám változatlansága mellett a két struktúra eltérése úgy nő, hogy x_i még kisebb lesz, x_j pedig (nyilván ugyanannyival) még nagyobb, akkor a mérőszám is nagyobb értéket vegyen fel. Az információpontatlanság valóban így viselkedik.

Legyen ugyanis x_i csökkenése (x_j növekedése) ε. Vizsgáljuk meg a következő függvény menetét:

$$f(\varepsilon) = y_i \log \frac{y_i}{x_i - \varepsilon} + y_j \log \frac{y_j}{x_j + \varepsilon}.$$

Ez a kifejezés az információpontatlanság változó része. Ha ez az $\varepsilon = 0$ helyen növekedést mutat, akkor a teljes összeg is növekedik, amit éppen ki szeretnénk mutatni. Mivel

$$\left.\frac{\mathrm{d}f(\varepsilon)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} = \left.\frac{y_i}{x_i - \varepsilon} - \frac{y_j}{x_j + \varepsilon}\right|_{\varepsilon=0} = \frac{y_i}{x_i} - \frac{y_j}{x_j},$$

és itt $y_i/x_i > 1$ és $y_j/x_j < 1$, tehát a két tört különbsége pozitív; ez éppen a függvény növekedését jelenti.

Az információpontatlanság értelmezhetőségét tekintve problémát jelent, ha valamelyik x vagy y értéke 0. Ha az egyik $y = 0$ és a megfelelő x is 0, akkor e számpár szempontjából a két struktúra egyáltalán nem tér el egymástól, tehát értelmes (és lehetséges) az

$$y \log \frac{y}{x}$$

tagot 0-nak értelmezni. Ha az egyik $y = 0$ és a megfelelő $x > 0$, akkor limesmeggondolással most is 0-nak vehető a tag értéke. Ha viszont $y > 0$, de $x = 0$, akkor a kifejezés még az $x \to 0$ határátmenettel sem értelmezhető, hiszen a kifejezés értéke ∞-hez tart.

Ez utóbbi meggondolásból az is következik, hogy az információpontatlanság tetszőleges nagy értéket is felvehet, hacsak valamelyik x érték a többi x-hez és y-hoz képest eléggé kicsi. Tehát az információpontatlanságnak nincs maximuma. Hogy minimuma éppen akkor van, amikor a megfelelő részarányok egymással páronként egyenlők, azt a következőképpen láthatjuk be.

Szorítkozzunk arra az esetre, mikor minden x_i és y_i pozitív és definiáljuk az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ mennyiségeket a következő egyenletekkel:

$$x_i = y_i (1 + \varepsilon_i) \qquad (i\ 1, 2, \ldots, n),$$

amiből következik, hogy

$$\sum_{i=1}^{n} y_i \varepsilon_i = 0,$$

mivel

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (y_i + y_i \varepsilon_i)$$

azaz

$$1 = 1 + \sum_{i=1}^{n} y_i \varepsilon_i.$$

Ezért az információpontatlanság így is felírható:

$$I(y:x) = -\sum_{i=1}^{n} y_i \log \frac{x_i}{y_i} = -\sum_{i=1}^{n} y_i \log (1 + \varepsilon_i) = \sum_{i=1}^{n} y_i [\varepsilon_i - \log (1 + \varepsilon_i)].$$

Az ε_i-ket definiáló egyenletekből világos, hogy $1 + \varepsilon_i > 0$, tehát a szögletes zárójelbeli kifejezést csak a $(-1, \infty)$ intervallumba vizsgáljuk. Ha $\varepsilon_i = 0$, akkor a kifejezés értéke is 0. Mivel a kifejezés ε_i szerinti deriváltja – természetes logaritmust használva –

$$\frac{\varepsilon_i}{1 + \varepsilon_i}$$

ez $-1 < \varepsilon_i < 0$ esetén negatív, $\varepsilon_i > 0$ esetén pedig pozitív, tehát a kifejezés $\varepsilon_i = 0$-nál veszi fel minimumát, a 0-t. Ebből már következik, hogy $I(y:x) \geq 0$, és 0 pontosan akkor, ha az összes ε_i eltűnik, vagyis $x_i = y_i$ minden i-re, tehát ha a két struktúra egymással azonos.

Ha egy vagy több y nulla, akkor limes-meggondolással ugyanerre az eredményre jutunk.

A fenti tulajdonságok alapján az információpontatlanságot joggal tekinthetjük két megfelelő struktúra eltérése mértékének és a megoszlási struktúrák összehasonlításakor nevezhetjük *struktúra-eltérésnek* is.

A struktúra-eltérésnek is megvan az az előnyös tulajdonsága, ami az entrópiának: szintén visszavezethető a részhalmazok mérőszámaira, mégpedig a következőképpen.

A részarányokat (ill. azok indexeit) osszuk be G számú részhalmazba, ezek: $S_g (g = 1, 2, \ldots, G)$. Legyen

$$X_g = \sum_{i \in S_g} x_i, \qquad Y_g = \sum_{i \in S_g} y_i,$$

$$\xi_i = \frac{x_i}{X_g} \; (i \in S_g), \qquad \eta_i = \frac{y_i}{Y_g} \; (i \in S_g).$$

Természetesen

$$\sum_{i \in S_g} \xi_i = \sum_{i \in S_g} \eta_i = 1.$$

Az $I(y : x)$ tagjait részhalmazok szerint csoportosítva:

$$\sum_{i=1}^{n} y_i \log \frac{y_i}{x_i} = \sum_{g=1}^{G} \left(\sum_{i \in S_g} y_i \log \frac{y_i}{x_i} \right),$$

ahol

$$\sum_{i \in S_g} y_i \log \frac{y_i}{x_i} = Y_g \sum_{i \in S_g} \frac{y_i}{Y_g} \log \frac{y_i / Y_g}{x_i / X_g} \cdot \frac{Y_g}{X_g} = Y_g \sum_{i \in S_g} \left(\eta_i \log \frac{\eta_i}{\xi_i} + \log \frac{Y_g}{X_g} \right) =$$

$$= Y_g \sum_{i \in S_g} \eta_i \log \frac{\eta_i}{\xi_i} + Y_g \log \frac{Y_g}{X_g}.$$

Végeredményben

$$I(y : x) = \sum_{g=1}^{G} Y_g \log \frac{Y_g}{X_g} + \sum_{g=1}^{G} \left(Y_g \sum_{i \in S_g} \eta_i \log \frac{\eta_i}{\xi_i} \right),$$

tehát a teljes struktúra-eltérés felírható a *halmazok közötti stuktúra-eltérés* és a *részhalmazokon belüli átlagos struktúraeltérés* összegeként.

Ezt a mennyiséget előnyösen használhatjuk fel gazdasági megoszlási struktúrák összehasonlítására, egy struktúra időbeli változásának mérésére, struktúra-előrejelzés pontosságának kvantifikálására.

A megoszlási struktúrán belüli egyenlőtlenséget, vagyis a részesedések koncentráltságát nyilvánvalóan egy olyan mérőszámmal fejezhetjük ki, amely az entrópiának inverze, amely annál nagyobb, minél jobban tér el a struktúra – adott elemszám mellett – a teljesen egyenletestől. Az eddig megtárgyalt mennyiségek és fogalmak alapján nem nehéz megfelelő mérőszámot találni egy struktúra koncentrációjára, vagyis a struktúrában fellépő egyenlőtlenségre.

Világos, hogy egy lehetséges mérőszám az entrópia tényleges értékeinek a maximálistól való eltérése:

$$\log n - H(y)\,.$$

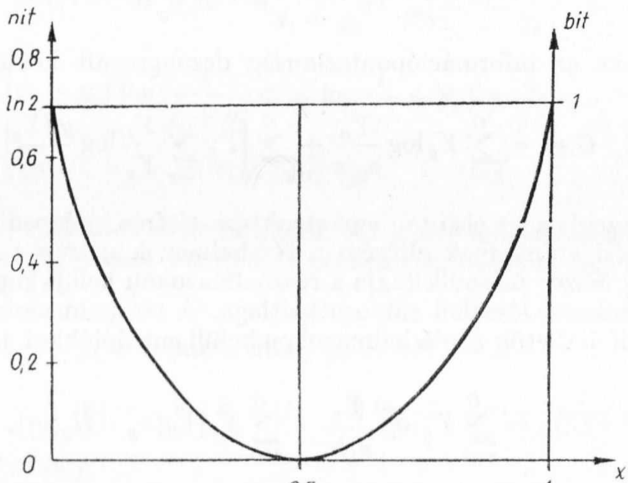
Ugyanezt a kifejezést kapjuk, ha kiszámítjuk a tényleges $(y_1, y_2, \ldots, y_n)$ struktúrának a teljesen egyenletestől $(x_1 = x_2 = \ldots = x_n = 1/n)$ való eltérését:

$$\sum_{i=1}^{n} y_i \log \frac{y_i}{x_i} = \sum_{i=1}^{n} y_i \log \frac{y_i}{1/n} = \log n - \sum_{i=1}^{n} y_i \log \frac{1}{y_i}\,.$$

Ezek szerint a

$$C(y) = \log n - H(y)$$

kifejezést a koncentráció információelméleti mérőszámának tekinthetjük és röviden *koncentrációnak* nevezzük. A koncentráció függvény menetét a **4.** ábra szemléltet abban a speciális esetben, amikor $n = 2$, $y_1 = x$, $y_2 = 1 - x$.



4. ábra. A koncentráció függvénye ($n = 2$)

A koncentráció szélsőértékei megegyeznek az entrópia szélsőértékeivel (tehát minimuma 0, maximuma $\log n$), csakhogy ahol az egyik minimumot vesz fel, ott veszi fel a másik a maximális értékét, és fordítva.

A koncentrációnál talán még érthetőbb, hogy a részesedések számának növekedésével a maximum növekedése miért vág egybe megszokott fogalmainkkal. Gondoljunk a jövedelemegyenlőtlenségre mint konkrét példára. $n = 2$ esetében egyiknek jut az összjövedelem, a másiknak pedig nem jut semmi, ekkor a koncentráció 1 bit. $n = 1024$ esetében a maximális egyenlőtlenség azt jelenti, hogy ismét egyetlen személynek jut az összjövedelem, viszont most 1023-nak a részesedése 0, ez pedig nyilvánvalóan nagyobb mértékű egyenlőtlenségnek fogható fel, ennek megfelelően a koncentráció számértéke is nagyobb: 10 bit.

Ha a részesedések számának hatását mégsem akarjuk figyelembe venni a koncentráció mérésénél, akkor ugyanúgy járhatunk el, mint az entrópiával, vagyis osztunk a maximummal:

$$C_r(y) = \frac{\log n - H(y)}{\log n}.$$

Ezt a mennyiséget *relatív koncentrációnak* nevezzük; értéke tetszőleges n esetén 0 és 1 között mozog.

A koncentráció mint az információpontatlanság speciális esete szintén felbontható, dezaggregálható.

Legyen $y_1, y_2, \ldots, y_n$ a vizsgált struktúra, $x_1 = x_2 = \ldots = x_n = 1/n$ pedig a neki megfelelő, teljesen egyenletes struktúra. Használjuk az eddigi jelöléseinket, kiegészítve azzal, hogy jelölje n_g a g-edik részhalmaz részesedéseinek számát. Ekkor

$$X_g = \frac{n_g}{n} \quad \text{és} \quad \xi_i = \frac{1}{n_g},$$

Behelyettesítve az információpontatlanság dezaggregált képletébe azt kapjuk, hogy

$$C(y) = \sum_{g=1}^{G} Y_g \log \frac{Y_n}{n_g/n} + \sum_{g=1}^{G} \left(Y_g \sum_{i \in S_g} \frac{Y_i}{Y_g} \log \frac{y_i/Y_g}{1/n_g} \right).$$

Ebben az összegben az első tag egy struktúra-eltérés, mégpedig a részhalmazok részesedési arányainak eltérése a részhalmazok nagyság szerinti arányszámaitól; az összeg második tagja a részhalmazokon belüli koncentrációknak a részhalmazrészesedésekkel súlyozott átlaga. A részhalmazokon belüli koncentrációk kifejezhetők a részhalmazokon belüli entrópiákkal, jelölésük $H_g(y)$. Ekkor

$$C(y) = \sum_{g=1}^{G} Y_g \log \frac{Y_g}{n_g/n} + \sum_{g=1}^{G} Y_g [\log n_g - H_g(y)].$$

Előfordul, hogy a koncentrációt olyan részletes elemzésnek akarjuk alávetni, amelyben kétszeres dezaggregálást alkalmazunk, például jövedelemegyenlőtlenség elemzése ágazati bontásban és azon belül nemek szerint. A kétszeres vagy általában többszörös dezaggregálás azért keresztülvihető, mert a részhalmazon belüli koncentráció mint koncentráció tovább bontható. (Egyébként hasonló érvényes az entrópiára is.)

Megjegyzendő, hogy míg az entrópia, információpontatlanság és koncentráció dezaggregálható, addig a relatív entrópia és relatív koncentráció nem rendelkezik ezzel a tulajdonsággal.

Bár a koncentráció mérőszáma általában alkalmas valamely struktúrában fellépő egyenlőtlenség mérésére, egyik legfontosabb alkalmazását az ipari koncentráció elemzése jelenti.

Könnyen belátható, hogy ha egy vállalat részaránya egy másik vállalatnál kisebb, és a nagyobb vállalat rovására növeli részarányát (úgy, hogy kettőjük részarányának összege állandó marad), akkor részarányaik kiegyenlítődéséig a koncentráció értéke csökken (vagyis éppen fordítva, mint az entrópiánál), megegyezésben a koncentrálódás fogalmával. Speciális problémát jelent azonban két vállalat egyesülése, ugyanis ekkor az eredeti struktúra részarányainak száma eggyel nagyobb, mint az egyesülés utánié; eddig viszont csak megfeleltethető struktúrák összehasonlításáról volt szó. Vizsgáljuk meg ezt a kérdést részletesebben.

Tegyük fel, hogy két vállalat úgy egyesül, hogy részarányaik összegeződnek és a többié változatlan marad. Az egyesülés következtében a struktúra entrópiája mindig csökken. Jelöljük ugyanis a szóban forgó két vállalat részarányát y_1-gyel, ill. y_2-vel ($y_1 > y_2 > 0$), a vállalatok számát (az egyesülés előtt) n-nel; ekkor az entrópia az egyesülés előtt:

$$H(y) = y_1 \log \frac{1}{y_1} + y_2 \log \frac{1}{y_2} + \sum_{i=3}^{n} y_i \log \frac{1}{y_i};$$

egyesülés után pedig:

$$H'(y) = (y_1 + y_2) \log \frac{1}{y_1 + y_2} + \sum_{i=3}^{n} y_i \log \frac{1}{y_2}.$$

Mivel

$$(y_1 + y_2) \log \frac{1}{y_1 + y_2} < y_1 \log \frac{1}{y_1} + y_2 \log \frac{1}{y_2},$$

tehát valóban

$$H(y) > H'(y).$$

A koncentráció változása szempontjából fontos az a megállapítás, hogy a két entrópia különbsége felvehet tetszőleges kis értéket. Ugyanis ha bevezetjük a következő jelöléseket:

$$y_1 + y_2 = k \text{ (konstants)}, \; y_2 = x, \; y_1 = k - x,$$

akkor

$$H(y) - H'(y) = k \log k - (k - x) \log (k - x) - x \log x.$$

Figyelembe véve, hogy

$$\lim_{x \to 0} x \log x = 0 \quad \text{és} \quad \lim_{x \to 0} (k - x) \log (k - x) = k \log k,$$

megállapíthatjuk, hogy

$$\lim_{x \to 0} [H(y) - H'(y)] = 0,$$

és ez éppen azt jelenti, hogy a különbség, amelyről már beláttuk, hogy mindig pozitív, tetszőlegesen megadott pozitív számnál kisebb értéket vesz fel, hacsak y_2 elég kicsi.

A koncentráció ugyan szoros kapcsolatban áll az entrópiával, azonban lényeges eltérés, hogy míg az entrópia növekedése csak a struktúra pozitív arányszámaitól függ, addig a koncentrációnál szerepet játszik a részesedési arányszámok száma is tekintet nélkül arra, hogy e részarányok nullától különbözők-e vagy sem. Nézzük meg, hogyan változik a koncentráció a két vállalat egyesülésekor.

A koncentráció az egyesülés előtt

$$C(y) = \log n - H(y),$$

egyesülés után

$$C'(y) = \log (n-1) - H'(y).$$

A koncentráció változása

$$C'(y) - C(y) = H(y) - H'(y) - \log \frac{n}{n-1}.$$

Tehát előfordulhat az is, hogy a koncentráció csökken, ha két vállalat egyesül, hiszen ennek feltétele

$$H(y) - H'(y) < \log \frac{n}{n-1},$$

ez pedig, mint az imént láttuk, lehetséges.

Hasonló érvényes a relatív koncentrációra is. Ez akkor csökken, ha

$$H(y) - H'(y) < H'(y) \cdot \frac{\log \frac{n}{n-1}}{\log (n-1)}$$

és ugyanez a feltétele a relatív entrópia növekedésének.

El kell-e vetnünk információelméleti mérőszámainkat e tulajdonságuk miatt? Ha a koncentrációt úgy fogjuk fel, hogy az elsősorban a részesedések egyenlőtlenségét jelenti, akkor egy aránytalanul kis vállalat beolvadása egy lényegesen nagyobb vállalatba valóban az (így értelmezett) koncentráció csökkenéséhez vezethet. Ha viszont megkívánjuk, hogy mérőszámunk a részesedések számának csökkenésére a hétköznapi értelemben reagáljon, akkor le kell mondanunk a közvetlen mérőszámról, helyette az inverz mutatót, az entrópiát használhatjuk.

Szemléletessé tehetjük a struktúra egyenletességének kifejezését azáltal, hogy az entrópia helyett az ún. *vállalatok egyenletességi egyenértékszámát* használjuk. Ez a szám arra a kérdésre ad választ, hogy hány egyenlő nagyságú vállalat egyenletessége egyezik meg a tényleges egyenletességgel (egyenletességen a struktúra entrópiáját értve). Képletben, ha az egyenértékszámot M-mel jelöljük:

$$\log M = \sum_{i=1}^{n} y_i \log \frac{1}{y_i},$$

tehát

$$M = e^{H(y)}$$

Ezt a számot úgy használhatjuk fel a koncentráció közvetlen kifejezésére, hogy kiszámítjuk, hányszorosa a vállalatok tényleges száma az egyenértékszámnak:

$$\frac{n}{M} = \frac{n}{e^{H(y)}}.$$

Ezt a hányadost *a vállalatok egyenlőtlenségi arányszámának* nevezzük. Ez tényleg követi a koncentráció változását, mert logaritmusa éppen $C(y)$.

Összefoglalásképpen álljon itt a következő táblázat a vizsgált mérőszámoknak és azok néhány jellegzetességeinek felsorolásával.

Megnevezés	Képlet	Értékkészlet határai	Mérték-egység	Dezaggregálható-e
Egyedi információ	$\log \frac{1}{x}$	$0, \log n$	bit	—
Entrópia vagy egyenletesség	$H(x) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i}$	$0, \log n$	bit	Igen
Relatív entrópia	$\frac{H(x)}{\log n}$	0,1	—	—
Információponatlanság vagy struktúraeltérés	$I(y:x) = \sum_{i=1}^{n} y_i \log \frac{y_i}{x_i}$	$0, \infty$	bit	Igen
Koncentráció	$C(x) = \log n - H(x)$	$0, \log n$	bit	Igen
Relatív koncentráció	$\frac{C(x)}{\log n}$	0,1	—	—
Vállalatok egyenletességi egyenértékszáma	$e^{H(x)}$	$1, \infty$	—	—
Vállalatok egyenlőtlenségi arányszáma	$\frac{n}{e^{H(x)}}$	$1, \infty$	—	—

Megjegyzések: (1) A fenti táblázatban n jelenti a struktúra részesedési arányszámainak számát, x_i és y_i az x ill. y struktúra i-edik részesedési arányszámát.

(2) A bit csak egyike a lehetséges egységeknek; a képletben szereplő logaritmus alapjától függően más és más elnevezést használunk.

IRODALOM

1. Corradi E.: Az információelmélet alapfogalmai és gazdasági alkalmazásának néhány kérdése. A KSH Statisztikai és Matematikai Módszerek Közgazdasági Alkalmazásának Laboratóriuma, 7. sz. munkaanyag.
2. Hall, M.—Tideman, N.: Measures of Concentration. Journal of the American Statistical Association, March 1967.
3. Horowitz, A.J.: Entropy, Markov Processes and Competition in the Brewing Industry. The Journal of Industrial Economics, 1968. No. 3.
4. Kotász Gy.-né—Szegedy, M.: Információelméleti mérőszámok alkalmazása a gazdasági elemzésben. KSH Ökonometriai Füzetek. 11. sz. 1970.
5. Meyer-Eppler, W.: Grundlagen und Anwendungen der Informationstheorie. Berlin, 1959. Springer-Verlag.
6. Reza, F. M.: Bevezetés az információelméletbe. Budapest, 1966. Műszaki Könyvkiadó.
7. Shannon, C. E.: A Mathematical Theory of Communication. Bell System Technical Journal. Vol. 27. 1948.
8. Shannon, C. E.—Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana, 111., 1949.
9. Theil, H.: Economics and Information Theory. Amsterdam, 1967. North-Holland Publishing Co., Magyar fordítás: Közgazdaságtan és információelmélet. Budapest, 1970. Közgazdasági és Jogi Könyvkiadó.