AZ INFORMÁCIÓ-AKKUMULÁCIÓ STATISZTIKUS TÖRVÉNYSZERŰSÉGEIRŐL

Írta: RÉNYI ALFRÉD

Bevezetés

A tudományos kutatás során és a gyakorlati életben egyaránt gyakran előfordul, hogy egy minket érdeklő tényállásra vonatkozó információ nem egyszerre jut teljes egészében birtokunkba, hanem azt részletekben szerezzük meg; az egyes, külön-külön keveset mondó adatokból — mintegy mozaikkövekből — lehet azután egy összképet kialakítani és a tényállást egészében felderíteni. *A teljes információt tehát részinformációk akkumulációja útján szerezzük meg.* Az információgyűjtés ezen folyamata során gyakran előfordul, hogy az egyes adatok részben átfedik egymást, és egy-egy új adat olyan információt is nyújt, amely az előzőleg szerzett adatokban már benne foglaltatik és csak részben ad valójában új felvilágosítást. Más szóval, az információ részletekben való gyűjtésénél általában bizonyos „redundancia" lép fel. Ezt a redundanciát csak úgy lehet kiküszöbölni, ha az információ gyűjtése igen jól átgondolt terv szerint történik; legtöbbször azonban ez olyan mértékben bonyolítja az információ összegyűjtését, hogy nem is célszerű erre törekedni.

Az információgyűjtés szóban forgó folyamata általában bizonyos véletlen elemeket is tartalmaz. Többnyire a véletlentől függ, hogy milyen részletadatok és milyen sorrendben jutnak az információt gyűjtő birtokába*.

Hogy az elmondottakat konkrétabbá tegyük, vizsgáljunk meg néhány példát. Első példaként tekintsük a kvalitatív kémiai analízis problémáját. Valamely anyagról meg akarjuk állapítani annak kémiai összetételét. E célból több vizsgálatnak, próbának vetjük alá az illető (homogénnek feltételezett) anyag egy-egy adagját. Minden egyes próba szűkíti a lehetőségek halmazát, míg végül elegendő számú próba után a lehetőségek halmaza egyetlenegy lehetőségre szűkül össze. Az egyes vizsgálatok részben egymást átfedő információt is nyújtanak. A vizsgálatok sorrendje befolyásolja azt, hogy milyen gyorsan sikerül a kérdést eldönteni. Persze, ha már eleve van bizonyos hipotézisünk, ez megkönnyíti az analízist, azonban minket az az eset érdekel, amikor a vizsgálat megkezdésekor még semmilyen támpontunk nincsen, és így a véletlennek is van bizonyos szerepe abban, hogy milyen sorrendben végezzük el az egyes vizsgálatokat.

Hasonló szituációval áll szemben az orvos is, amikor egy beteg diagnózisát kívánja megállapítani, e célból a beteget különböző vizsgálatoknak veti alá és ezek

* A valóságos helyzetet rendkívül bonyolítja, hogy az ember gyakran jut téves információkhoz is, vagyis a véletlen sokszor véletlen hibák (elírások, félreértések stb.) formájában is közrejátszik; ez a helyzet akkor, ha — az információelmélet szokásos terminológiájával élve — az egyes adatok „zajos" csatornán át jutnak el az információ gyűjtőjéhez. Bár az ilyen hamis adatok kiküszöbölése (pl. az adatok közötti ellentmondások felfedése útján) gyakran az információ feldolgozásának legnehezebb részét képezi, a kérdésnek ezzel az oldalával itt nem kívánunk foglalkozni; e kérdésre más alkalommal kívánunk visszatérni.

eredményeinek összevetése útján állapítja meg, hogy milyen betegségben szenved az illető beteg. Itt nagy szerepet játszik persze az orvos intuiciója és gyakorlata, amelyre támaszkodva már eleve csak kisszámú lehetőségre kell korlátoznia a figyelmét. Az orvosi tapasztalat és intuició szerepét azonban úgy foghatjuk fel, hogy az orvos a beteg panaszai, a betegről szerzett benyomásai alapján és a beteg szervezetének esetleges előzetes kezelésekből származó ismeretéből, még mielőtt a laboratóriumi vizsgálatokhoz folyamodna, máris nagy mennyiségű információ felett rendelkezik; vagyis tulajdonképpen az információgyűjtés nem a laboratóriumi vizsgálattal kezdődik, hanem abban a pillanatban, amikor az orvos az illető beteggel foglalkozni kezd (és ennek következtében gyakran laboratóriumi vizsgálatokra nincs is szükség). Minket e példából elsősorban az érdekel, hogy az orvos a végleges diagnózishoz nagyszámú részletinformáció összevetése révén jut el. Az, hogy milyen információk jutnak az orvos birtokába és milyen sorrendben, bizonyos mértékig a véletlentől is függ. Pl. az, hogy a beteg nem felejt-e ki valamit betegsége tüneteinek felsorolásából (esetleg azért, mert ő maga orvosilag tájékozatlan lévén, az illető, őt viszonylag kevéssé zavaró tünetnek nem tulajdonít jelentőséget), erősen a véletlentől függ.

Harmadik példaként tekintsük egy vizsgálóbíró munkáját, aki egy peres ügyben kívánja a tényállást felderíteni; e célból számos tanút hallgat ki, és azok vallomásaiból igyekszik a történteket rekonstruálni. Legtöbbször az a helyzet, hogy az egyes tanúk a szóban forgó eseményeknek csak egy-egy részletéről bírnak tudomással; továbbá az egyes tanúk által szolgáltatott felvilágosítások legtöbbször részben átfedik egymást*. A bíró feladata pl. állhat abban, hogy egyének egy bizonyos H halmazából kiválassza azt az egyént, aki egy bizonyos cselekményt elkövetett. Azt, hogy kiket és különösen, hogy milyen sorrendben hallgat ki a bíró, legtöbbször tartalmaz véletlen elemeket. (Pl. előfordulhat, hogy a megidézett tanúk közül egyesek akadályoztatás miatt csak későbbi időpontban vagy egyáltalán nem hallgathatók ki.)

További példaként vizsgáljuk meg az ún. „Bar—Kochba" játékot, amely a szóban forgó helyzetnek olyan messzemenően leegyszerűsített esete, hogy ennek alapján azonnal eljuthatunk a probléma matematikai modelljének felállításához. Az említett játékban ugyanis a minket érdeklő probléma szinte „laboratóriumi" tisztaságban áll előttünk. A Bar—Kochba játékban két játékos vesz részt; nevezzük őket A-nak és B-nek. Az A játékos gondol valamire, a B játékos pedig kérdéseket tesz fel A-nak és a kérdéseire kapott válaszok alapján igyekszik kitalálni, hogy A mire gondolt**. Azt, hogy mire gondolhat A, a játék szabályai nem korlátozzák lényegesen: A gondolhat egy meghatározott személyre, egy tárgyra, egy elvont fogalomra stb. B csak olyan kérdéseket tehet fel, amelyekre igennel vagy nemmel lehet válaszolni, A pedig köteles minden feltett és megválaszolható kérdésre legjobb tudása szerint helyesen válaszolni***. Csak ha a kérdés nem válaszolható meg, vagy A a helyes választ nem ismeri, áll A-nak jogában nem válaszolni egy feltett kérdésre.

* Mint már fentebb említettük, azt a (gyakorlatilag egyáltalán nem elhanyagolható) lehetőséget, hogy egyes tanúk (szándékosan vagy tévedésből) a valóságnak nem megfelelő vallomást tesznek, itt nem vesszük tekintetbe.

** A játékot szokták úgy is játszani, hogy többen együtt gondolnak valamire és felváltva válaszolnak a kérdezőnek; a mi szempontunkból ez nem jelent lényeges különbséget.

*** A téves információkat, amelyekkel a probléma matematikai tárgyalásának egyszerűsítése céljából itt amúgy sem kívánunk foglalkozni, a Bar—Kochba játék esetében tehát a játék szabályai eleve kizárják.

A Bar-Kochba játéknál a gyakorlott kérdező általában bizonyos rendszer szerint szokott kérdezni, a kérdések megválasztását és sorrendjét azonban még a gyakorlott játékosnál is kétségtelenül többé-kevésbé a véletlentől függő tényezők befolyásolják, míg olyan kérdező esetében, aki a játékban kevés tapasztalattal bír, a kérdések megválasztása teljesen véletlenszerűnek tekinthető.

A Bar—Kochba játék egy lehetséges módosításával is érdemes foglalkozni, amelyet „*többértékű Bar—Kochba játék*"-nak nevezhetünk. Ez abban különbözik a játék szokásos formájától, hogy a kérdezőnek jogában áll olyan kérdéseket is feltenni, amelyekre nem lehet igen-nel vagy nem-mel válaszolni, hanem a lehetséges válaszok száma kettőnél több. Ez esetben azonban a kérdezőnek fel kell sorolni a kérdésre adható összes lehetséges válaszokat, és az A játékos meg kell, hogy mondja, hogy az általa gondolt dolog esetében a kérdező által felsorolt válaszok közül melyik a helyes (másszóval, hogy az általa gondolt dolog a kérdező által felsorolt, egymást kizáró kategóriák közül melyikbe tartozik).

A „többértékű Bar—Kochba" esetében célszerű az alternatívák számát előre korlátozni (pl. úgy, hogy legfeljebb 3 vagy legfeljebb 4 alternatíva adható meg).

Ahhoz, hogy a Bar—Kochba játéktól eljussunk egy matematikailag kezelhető, leegyszerűsített modellhez, csak egyetlen egyszerűsítést kell bevezetni, mégpedig a gondolható dolgok halmazát kell egy — a játékosok által a játék megkezdése előtt elfogadott megállapítással előre lerögzített — *véges* H halmazra korlátozni. Az, hogy a gondolható dolgok halmaza véges, amennyiben az elemszámot nem korlátozzuk, nem jelent gyakorlatilag megszorítást. (Lehetséges megszorítás pl., hogy csak olyasmire lehet gondolni, ami, mint külön címszó, egy bizonyos lexikonban szerepel, vagy pl. ami egy meghatározott szótárban szereplő 5 szóval definiálható stb.).

Ebben az esetben minden egyes kérdésnek egyértelműen megfelel a H halmaz egy H' részhalmaza oly módon, hogy a kérdés azzal ekvivalens, hogy a gondolt dolog hozzátartozik-e a H' részhalmazhoz?

Könnyen megadhatunk egy elméletileg optimális, bár gyakorlatilag nemigen számításba jövő „stratégiát" a kérdező részére. Ha ugyanis a gondolható dolgok H halmaza n elemű, $2^{s-1} < n \leqq 2^s$ $(s = 1, 2, \ldots)$, és H elemeit a $0, 1, \ldots, n-1$ számokkal számozzuk meg; ez esetben (feltéve, hogy a számozásban a játékosok megállapodtak, vagy egy egyértelmű természetes számozás adva van) elegendő, ha a B játékos a gondolt dolog sorszámára kérdez. Más szóval azt is feltehetjük, hogy A valójában a $0, 1, \ldots, n-1$ számok valamelyikére gondolt, és ezt a *gondolt számot* akarja B kitalálni. Mármost a $0, 1, \ldots, n-1$ számokat a kettes számrendszerben felírva, azok mindegyike legfeljebb s diadikus jegyet tartalmaz, és így B minden esetben célhoz ér a következő s kérdéssel; „A gondolt dolog sorszámát a kettes számrendszerben felírva, abban a j-edik jegy zérus-e? $(j = 0, 1, \ldots, s-1)$".

Az n elemű H halmaz gondolt eleme tehát $s = \{\log_2 n\}$ kérdésre adott válaszból egyértelműen kitalálható; itt $\{x\}$ a legkisebb olyan egész számot jelöli, amely nem kisebb x-nél. Azt, hogy ennél kevesebb kérdéssel nem juthatunk célhoz, a következő egyszerű meggondolással láthatjuk be. Ha csak olyan kérdést tehet B fel, amelyre a válasz igen vagy nem, akkor minden elképzelhető kérdéshez hozzárendelhető a H halmaznak egy H' részhalmaza, amely H azon elemeiből áll, amelyek esetében a válasz a szóban forgó kérdésre igenlő. Ha mármost $f(n)$-nel jelöljük azon kérdések minimális számát, amelyek minden körülmények között elégségesek egy n elemű halmaz egy elemének kitalálásához, akkor először is nyilvánvaló, hogy $f(n)$ monoton

nem-csökkenő függvény, másrészt, ha az első kérdésnek megfelelő halmaz m-elemű, akkor még további $f(m)$ ill. $f(n-m)$ kérdésre van szükség aszerint, hogy az első kérdésre igenlő vagy tagadó választ kaptunk; ennélfogva

$$f(n) \geqq 1 + \operatorname*{Min}_{0 \leqq m \leqq n} [\operatorname{Max}(f(m), f(n-m))].$$

Tehát

$$f(n) \geqq 1 + f\left(\left\{\frac{n}{2}\right\}\right).$$

Ha azonban $2^{s-1} < n \leqq 2^s$, akkor $\left\{\frac{n}{2}\right\} > 2^{s-2}$. Ezt a meggondolást $(s-1)$-szer megismételve, azt kapjuk, hogy

$$f(n) \geqq s - 1 + f(2);$$

mármost nyilvánvaló, hogy $f(2) = 1$, tehát

$$f(n) \geqq s;$$

de s alkalmas kérdés bizonyosan célhoz vezet, tehát $f(n) = s$, ha $2^{s-1} < n \leqq 2^s$.

Hasonlóképpen látható be, hogy a többértékű Bar—Kochba játéknál, ha az egy kérdésre adandó válaszok megengedett száma r, a szükséges kérdések száma $\left\{\frac{\log_2 n}{\log_2 r}\right\}$, és egy optimális kérdésrendszer pl. az, hogy sorra megkérdezzük, hogy H gondolt elemének sorszámát az r-alapú számrendszerben felírva a 0-adik, első, második,... jegy a $0, 1, \ldots, r-1$ számok közül melyik?

A mondottakból látható, hogy ha egy-egy kérdésre r alternatív válasz adható, mindig feltehetjük, hogy a gondolható dolgok H halmazának elemszáma r-nek hatványa; ha ugyanis $r^{s-1} < n \leqq r^s$, semmit nem veszítünk, ha a H halmazt $r^s - n$ „fiktív" elemmel r^s eleművé egészítjük ki. Tegyük tehát fel, hogy $n = r^s$. Mármost könnyen belátható, hogy ez esetben a fentebb ismertetett optimális kérdésrendszerek a következő két okból optimálisak: a) először is minden válasz teljes egészében új információt nyújt, azaz redundancia nem lép fel; ugyanis egy számot az r-alapú számrendszerben felírva az egyes jegyek egymástól függetlenül vehetik fel a $0, 1, \ldots, r-1$ értékeket. b) Egy olyan kérdésre adott válasz, amely kérdésre elvileg r-féle válasz lehetséges, az információelmélet elemei szerint maximálisan $\log_2 r$ „bit" információt tartalmaz és az említett optimális kérdésrendszer esetében ezt a maximális információt minden egyes kérdésre ténylegesen meg is kapjuk. Ugyanis a $0, 1, \ldots, r^s - 1$ számokat az r alapú számrendszerben felírva, ezek közül pontosan r^{s-1} számnak lesz a j-edik jegye éppen i $(j = 0, 1, \ldots, s-1;\ i = 0, 1, \ldots, r-1)$. Ha tehát a $0, 1, \ldots, r^s - 1$ számok mindegyikére ugyanolyan valószínűséggel gondolt az A játékos (amit feltehetünk, hiszen ez éppen azt jelenti, hogy semmilyen előzetes tudomásunk sincs arról, hogy A melyik számra gondolt), akkor arra a kérdésre, hogy a gondolt dolog sorszámát az r-alapú számrendszerben felírva, mi a j-edik jegy, $\frac{1}{r}$ valószínűséggel kapjuk azt a választ, hogy ez a jegy i $(i = 0, 1, \ldots, r-1)$. Ez esetben tehát $\log_2 r$ információt kapunk minden egyes válaszból és így s válaszból összesen $s \log_2 r = \log_2 n$ információt kapunk, és éppen ennyi információra van szükség ahhoz, hogy egy n elemű halmaz egy elemét egyértelműen jellemezni tudjuk.

Mielőtt továbbmennénk, vegyük még észre az említett optimális kérdésrendszer következő sajátosságát: teljesen mindegy, hogy a kérdéseket egymás után egyenként tesszük-e fel vagy egyszerre. Míg a tényleges Bar—Kochba játéknál előbb megvárjuk az első kérdésre adott választ és csak azután tesszük fel a második kérdést, s. i. t., és első pillanatra úgy tűnhet, hogy ez előnyt jelent (vagyis, hogy érdemes a második kérdés megválasztását függővé tenni az első kérdésre kapott választól, s. i. t.), valójában ez az előny illuzórikus, hiszen az említett optimális kérdésrendszer esetében semmilyen hátrányt nem jelent, ha az összes kérdéseket egyszerre tesszük fel, hiszen akármi is az első kérdésre a válasz, úgyis ugyanazt a kérdést tesszük fel másodiknak. Pl. $r=2$ esetében az első kérdés úgy szól, hogy a gondolt szám páros-e, a második úgy, hogy a gondolt számból kivonva e szám mod 2 vett maradékát és a különbséget 2-vel osztva a hányados páros-e? Ha a második kérdés feltevése előtt már rendelkezésre áll az első kérdésre adott válasz, ez csak annyi különbséget jelent, hogy ha a válasz az volt, hogy a szám páros, a második kérdést úgy fogalmazhatjuk, hogy a szám fele páros-e, míg ha azt hallottuk, hogy a szám páratlan, másodszorra azt kérdezzük, hogy a gondolt számból egyet levonva a kapott páros szám fele páros-e? Ez azonban csak a második kérdés *fogalmazását* érinti, a szükséges kérdések *számát* azonban nem befolyásolja.

Más szavakkal kifejezve: teljesen mellékes, hogy a kapott válaszokból nyert információkat minden egyes válasz után „összegezzük"-e, vagy az „összegezést" csak akkor kezdjük meg, ha már mind az s válasz rendelkezésünkre áll.

Egy optimális kérdésrendszer az $r=2$, $n=2^s$ esetben a következőképpen jellemezhető. Ha $H_1, H_2, \ldots, H_s$ jelölik az egyes kérdéseknek megfelelő részhalmazokat (vagyis a j-edik kérdés azt kérdezi, hogy a gondolt dolog hozzátartozik-e a H halmaz H_j részhalmazához) és $\overline{H}_j$ jelöli a H_j halmaz kiegészítő halmazát H-ra nézve, akkor a H_j és $\overline{H}_j$ halmazok 2^{s-1} elemből állnak, a H_iH_j $H_i\overline{H}_j$, $\overline{H}_iH_j$ és $\overline{H}_i\overline{H}_j$ halmazok ($i\neq j$) mindegyike 2^{s-2} elemből áll és általában ha $\tilde{H}_j$ a H_j és $\overline{H}_j$ halmazok közül az egyiket jelöli, akkor az összes $\tilde{H}_{j_1}\tilde{H}_{j_2}\ldots\tilde{H}_{j_l}$ alakba írható halmazok ($j_1<j_2<\ldots<j_l$) mindegyike 2^{s-l} elemből áll. Speciálisan tehát a $\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_s$ halmazok mindegyike egyetlen elemet tartalmaz. Mivel ilyen alakú halmaz pontosan 2^s darab van, ezek tehát a H halmaz elemeihez kölcsönösen egyértelműen vannak hozzárendelve; könnyen belátható, hogy az összes lehetséges $2^s!$ ilyen hozzárendelésnek megfelel egy optimális kérdésrendszer, vagyis az optimális kérdésrendszerek száma éppen $2^s!$.

A tényleges Bar—Kochba játéknál az említett optimális kérdésrendszerek azonban nem alkalmazhatók, kivéve ha a gondolható dolgok halmazának elemszáma igen kicsiny, már pusztán azért sem, mert ha n nagy, a gondolható dolgok halmaza elemeinek megszámozása igen fáradságos volna és e számozás fejbentartása gyakorlatilag nem vihető keresztül. (Ha ez a rendszer ténylegesen alkalmazható volna, a játék elveszítené minden érdekességét.) Így tehát felmerül a kérdés, hogy mennyivel több kérdést kell feltenni akkor, ha nem optimális kérdésrendszert használunk. Ezt a problémát azon extrém feltevés mellett fogjuk megválaszolni, amikor nemcsak hogy nem optimális kérdésrendszer szerint kérdezünk, hanem kérdéseinkben semmiféle rendszert nem alkalmazunk, vagyis a kérdéseket teljesen találomra választjuk meg. Arra a meglepő eredményre jutottunk, hogy a szükséges kérdések száma általában nem sokkal lesz több, mint az optimális kérdésrendszer esetében, pl. n értékétől függetlenül az esetek kb. 99%-ában a kérdéseket találomra választva legfeljebb 7-tel több kérdésre van szükség, mint optimális kérdésrendszer esetében.

Az 1. §-ban a közönséges Bar—Kochba játék modelljével, a 2. §-ban a többértékű Bar—Kochba játék modelljével foglalkozunk.

A tárgyalás során azzal az általánosabb esettel is foglalkozunk, amikor a lehetséges válaszok egy-egy kérdésre eleve (a priori) nem egyformán valószínűek. Ha egy kérdésre a többértékű Bar—Kochbánál r válasz lehetséges és ezek valószínűségei a $\mathcal{S}=\{p_1, p_2,\ldots, p_r\}$ valószínűségeloszlás tagjai, akkor Shannon képlete szerint egy válasz $I_1(\mathcal{S})=\sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i}$ információt nyújt és így várható, hogy legalább $\left\{\frac{\log_2 n}{I_1(\mathcal{S})}\right\}$ kérdést kell feltenni. Ez esetben is kiszámítjuk, hogy ha a kérdéseket minden rendszer nélkül, találomra tesszük fel, mennyivel több kérdés szükséges.

A 3. és 4. §-okban egy, az előbbivel rokon másik problémát vizsgálunk. Ennek a problémának a modellje a Bar—Kochba játék egy másik általánosítása, amelyet „*szimultán Bar—Kochba játéknak*" nevezhetünk. E játékban nem egy, hanem több játékos — pl. az $A_1, A_2,\ldots, A_l$ játékosok mindegyike — gondol a H halmaz egy-egy általa választott elemére (feltesszük, hogy nincs két játékos, aki H ugyanazon elemére gondolna) és a kérdező B játékosnak ki kell találnia, hogy az $A_1, A_2,\ldots, A_l$ játékosok rendre mire gondoltak. E célból B kérdéseket tesz fel szimultán az $A_1, A_2,\ldots, A_l$ játékosoknak, amelyekre ezek mindegyike — az általa gondolt dologra vonatkozólag — a valóságnak megfelelően válaszol. Aszerint, hogy csak igennel vagy nemmel megválaszolható kérdések vannak-e megengedve, vagy olyanok, amelyekre $r>2$ válasz lehetséges, beszélünk *közönséges* ill. *többértékű szimultán Bar—Kochba* játékról. A matematikai tárgyalás szempontjából lényeges különbséget jelent, hogy l kis rögzített szám-e (pl. $l=2$ vagy $l=3$) vagy pedig l nagy szám, pl. $l=n$. Az $l=n$ esetben a H halmaz minden egyes elemére gondolt az $A_1, A_2,\ldots, A_l$ játékosok közül egy, tehát lényegében a H halmaz elemeinek egy ismeretlen permutációját kell B-nek kitalálnia. Az l szám nagyságától függetlenül B-nek a szimultán Bar—Kochba esetében valójában azt a *függvényt* kell meghatároznia, amely a $A_1, A_2,\ldots, A_l$ játékosok mindegyikéhez hozzárendeli a H halmaz egy elemét, vagyis egy véges halmaznak egy másik véges halmazra való *leképezését* kell „kitalálnia". Abban az esetben, amikor $l=n$, tehát a H halmaz minden egyes elemére gondolt az $A_1, A_2,\ldots, A_l$ játékosok közül legalább egy, B-nek olyan kérdéseket kell feltennie, amelyekre adott válaszok H minden egyes elemét egyidejűleg jellemzik.* Mivel egy olyan kérdésnek, amelyre r válasz lehetséges, egyértelműen megfeleltethető a H halmaz r részre való felosztása, tehát B akkor tudja kitalálni, hogy az $A_1, A_2,\ldots, A_l$ játékosok mire gondoltak, ha az általa feltett kérdéseknek megfelelő felosztások H elemeit teljesen *szeparálják*, azaz nincs két olyan eleme H-nak, amely minden egyes felosztásnál ugyanabba az osztályba tartozik.**

Ilyen módon a szimultán Bar—Kochba játék problémája lényegében azonos a

* A közönséges Bar—Kochbával kapcsolatban említett optimális kérdésrendszerek egyben a szimultán Bar—Kochbára nézve is optimálisak; ha azonban a kérdéseket véletlenszerűen választjuk, a közönséges és a szimultán Bar—Kochba között lényeges eltérés van: a szimultán Bar—Kochbánál az $l=n$ esetben kb. kétszerannyi kérdést kell feltenni, mint az $l=1$ esetben.

** A rövidség kedvéért a következő terminológiát használjuk: amennyiben az egyes részinformációk a H halmaz két részre vágásának felelnek meg („igen vagy nem" válasz), azt mondjuk, hogy az információt *dichotomiák*ból nyerjük (dichotomia = kettévágás), míg ha az egyes osztályozásoknál kettőnél több osztály is felléphet, azt mondjuk, hogy *általános osztályozás*okból nyerjük az információt.

következő problémával: Egy könyvtár könyveit különböző szempontok szerint osztályozzuk kategóriákba; milyen feltételek mellett fogják ezek az osztályozások együttvéve egyértelműen jellemezni a könyvtár minden egyes könyvét. Vagy tekintsük a kvalitatív kémiai analízis példáját. E példát, amelyet már fentebb említettünk, szintén lehet úgy módosítani, hogy az a „szimultán" Bar—Kochba esetének feleljen meg. Tegyük ugyanis fel, hogy nem egyetlen ismeretlen anyagot, hanem többféle anyagot kell egyidejűleg analizálnunk, mégpedig úgy, hogy ugyanazokkal a vizsgálatokkal állapítsuk meg ezek mindegyikének kémiai minőségét. Tehát a szóban forgó anyagokat többféle próbának vetjük alá és feljegyezzük, hogy melyik hogyan reagált azokra. Minden egyes vizsgálatnál a lehetséges reakciók az összes kémiai vegyületek egy-egy osztályozását adják meg. Kérdés, hogyan kell az elvégzendő vizsgálatok sorozatát összeállítani ahhoz, hogy a reakciókból egyértelműen meghatározható legyen, milyen anyagokkal állunk szemben. Hasonló probléma merül fel a képességvizsgálatoknál alkalmazott „teszt"-eknél is, stb.

A szóban forgó probléma érdekessége, hogy megoldásában a $\mathcal{S} = \{p_1, p_2, \ldots, p_r\}$ valószínűségeloszláshoz tartozó információnak nem a Shannon-féle $I_1(\mathcal{S}) = \sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i}$ mértékszáma, hanem egy másik információ-mérték, az $I_2(\mathcal{S}) = \log_2 \frac{1}{\sum_{i=1}^{r} p_i^2}$ ún. *másodrendű információ-mértékszám* játszik szerepet. Az $I_2(\mathcal{S})$ információ-mértékszám a szerző [1] és [2] dolgozataiban bevezetett információmértékszámok közé tartozik. Az említett dolgozatokban a $\mathcal{S} = \{p_1, p_2, \ldots, p_r\}$ valószínűségeloszláshoz tartozó információmennyiség α-adrendű $I_\alpha(\mathcal{S})$ mértékszámát $(\alpha > 0)$ az $\alpha \neq 1$ esetben a következőképpen definiáltuk:

$$I_\alpha(\mathcal{S}) = \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{r} p_i^\alpha \right),$$

megjegyezve, hogy ha $\alpha \to 1$, akkor $I_\alpha(\mathcal{S})$ az $I_1(\mathcal{S}) = \sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i}$ Shannon-féle mértékszámhoz konvergál, amely tehát úgy tekinthető, hogy szintén az α-adrendű információ-mértékszámok közé tartozik, mégpedig az $\alpha = 1$ esetnek felel meg.

Megjegyezzük még, hogy az $r = 2, p_1 = p_2 = \frac{1}{2}$ speciális esettel foglalkoznak a szerző [3] és [4] dolgozatai. [3]-ban a kérdést a Boole-algebra terminológiájával fogalmaztuk meg. A többértékű szimultán Bar—Kochba játék problémája a Boole-algebra nyelvén a következőképpen fogalmazható meg: a H halmaznak a B játékos által feltett kérdéseknek megfelelő részhalmazai milyen feltételek mellett generálják (a halmazelméleti műveletekre nézve) a H összes részhalmazaiból álló Boole-algebrát?

Az e dolgozatban tárgyalt eredményeket 1961 nyarán előadtam a Michigan State University-n tartott szemináriumomban. A szeminárium résztvevői közül elsősorban H. RUBIN professzor kapcsolódott be a szóban forgó témakör vizsgálatába. Értékes hozzászólásaiért, amelyekből sok ösztönzést merítettem, ezúton is köszönetet mondok. A szeminárium egy másik résztvevője, J. FOX, a következő módosítását vetette fel a közönséges szimultán Bar—Kochba játéknak: a kérdések feltevése során B mindig csak olyan kérdéseket tesz fel, amelyek biztosan nyújtanak új infor-

mációt az összes, az előző kérdések által még ki nem talált, gondolt dolgokra vonatkozólag. Más szóval, a j-edik kérdésnek olyannak kell lennie, hogy a H halmaznak az előző $j-1$ kérdésnek megfelelő két részre osztásai egyesítésével létrejött felbontásának minden egynél több elemet tartalmazó osztályát a j-edik kérdésnek megfelelő felosztás két valódi részre bontsa fel. A probléma ezen módosítását illetőleg H. RUBIN ért el érdekes eredményeket.

1. §. Információ-akkumuláció egy ismeretlenre vonatkozó dichotomiák esetében

E §-ban először a következő problémát oldjuk meg.

Legyen H egy n elemű halmaz. Válasszuk ki találomra H-nak k számú részhalmazát — ezeket jelöljék $H_1, H_2, \ldots, H_k$ —, oly módon, hogy az egyes részhalmazok kiválasztásai függetlenek és minden választásnál H minden egyes részhalmaza (beleértve az üres halmazt és magát H-t is) ugyanakkora, tehát $\frac{1}{2^n}$ valószínűséggel kerülhet kiválasztásra.* Kiszámítandó annak a valószínűsége — ezt P_{nk}-val jelöljük —, hogy kijelölve H egy x elemét, ez az elem egyértelműen meg legyen határozva annak megadása által, hogy a $H_1, H_2, \ldots, H_k$ halmazok közül melyeknek eleme és melyeknek nem eleme. Meg kívánjuk vizsgálni továbbá, hogy mekkorára kell k értékét választani, hogy a szóban forgó P_{nk} valószínűség egy megadott α $(0<\alpha<1)$ számot meghaladjon.

Jelöljék H elemeit $x_1, x_2, \ldots, x_n$ és legyen

$$\varepsilon_{jh}=\begin{cases}1 & \text{ha } x_h \in H_j \\ 0 & \text{ha } x_h \notin H_j\end{cases} \qquad (j=1,2,\ldots,k,\ h=1,2,\ldots,n). \tag{1.1}$$

Feltevésünk szerint a H_j halmaz $\frac{1}{2^n}$ valószínűséggel lesz azonos H bármely részhalmazával; mivel továbbá H-nak nyilván 2^{n-1} olyan részhalmaza van, amely x_h-t tartalmazza és ugyancsak 2^{n-1} olyan részhalmaza, amely x_h-t nem tartalmazza, tehát

$$\mathbf{P}(\varepsilon_{jh}=1)=\mathbf{P}(\varepsilon_{jh}=0)=\frac{2^{n-1}}{2^n}=\frac{1}{2}. \tag{1.2}$$

Hasonlóképpen látható be az is, hogy az ε_{jh} $(j=1,2,\ldots,k;\ h=1,2,\ldots,n)$ valószínűségi változók függetlenek; ugyanis először is az $(\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jn})$ vektorok függetlenek, mert feltevésünk szerint a $H_1, H_2, \ldots, H_k$ halmazokat egymástól függetlenül választjuk; másrészt viszont, ha $\delta_1, \delta_2, \ldots, \delta_n$ tetszőleges, a 0 és 1 számokból álló számsorozat, akkor

$$\mathbf{P}(\varepsilon_{j1}=\delta_1,\ \varepsilon_{j2}=\delta_2,\ \ldots,\ \varepsilon_{jn}=\delta_n)=\frac{1}{2^n}=\prod_{h=1}^{n}\mathbf{P}(\varepsilon_{jh}=\delta_h), \tag{1.3}$$

* Ily módon előfordulhat, hogy ugyanazt a részhalmazt többször is kiválasztjuk, ennek valószínűsége azonban rendkívül csekély lesz, ha n nagy szám, mivel vizsgálatainkban k nagyságrendje $\log n$, és ahhoz, hogy ugyanazon halmaz többszöri kiválasztásának a valószínűsége számottevő legyen, k-nak legalább $\sqrt{2^n}$ nagyságrendűnek kellene lennie.

ugyanis H-nak egyetlenegy olyan részhalmaza van, amely tartalmazza azon x_h-kat, amelyekre $\varepsilon_{jh}=1$ és nem tartalmazza azon x_h-kat, amelyekre $\varepsilon_{jh}=0$. Tehát az ε_{jh} $(j=1, 2,..., k;\ h=1, 2,..., n)$ valószínűségi változók függetlenek és mindegyik a 0 és 1 értékeket $\frac{1}{2}$ valószínűséggel veszi fel. E változókat rendezzük el egy $k\times n$ elemű mátrix alakjában, oly módon, hogy ε_{jh} a mátrix j-edik sorában a h-adik helyre kerüljön. Jelöljük e mátrixot E_{kn}-nel. Könnyen belátható, hogy az a követelmény, hogy az x elemet egyértelműen meghatározza az, hogy a $H_1, H_2,..., H_k$ halmazok közül melyeknek eleme, az $x=x_i$ esetben azt jelenti, hogy az E_{kn} mátrixnak az i-edik oszlopa különbözzék e mátrix összes többi oszlopától. Mivel az E_{kn} mátrix oszlopvektorai függetlenek és mindegyikük a k-dimenziós és a 0 és 1 komponensekből álló összes lehetséges 2^k vektorral $\frac{1}{2^k}$ valószínűséggel egyenlő, a keresett P_{nk} valószínűség nyilván

$$P_{nk}=\left(1-\frac{1}{2^k}\right)^{n-1}, \tag{1.4}$$

függetlenül attól, hogy mi i értéke és hogy milyen elemek állnak E_{kn} i-edik oszlopában. Másrészt az $e^{-\frac{x}{1-x}}<1-x<e^{-x}$ $(0<x<1)$ egyenlőtlenség szerint

$$e^{-\frac{n-1}{2^k-1}}<P_{nk}<e^{-\frac{n-1}{2^k}}.$$

Ha $2^k\geqq n$, akkor $\frac{n-1}{2^k-1}\leqq\frac{n}{2^k}$, ezért

$$e^{-\frac{n}{2^k}}<P_{nk}<e^{-\frac{n-1}{2^k}}, \quad \text{hacsak} \quad 2^k\geqq n. \tag{1.5}$$

Más szóval, ha $\frac{1}{e}<\alpha<1$, és $k\geqq\log_2 n+\log_2\frac{1}{\log\frac{1}{\alpha}}$, akkor $P_{nk}>\alpha$, míg ha $k\leqq\log_2(n-1)+\log_2\frac{1}{\log\frac{1}{\alpha}}$, akkor $P_{nk}<\alpha$. Ilyen módon bebizonyítottuk a következő tételt:

1a. TÉTEL. *Találomra válasszuk ki egy n elemű H halmaz k részhalmazát oly módon, hogy az egyes részhalmazokat egymástól függetlenül választjuk és minden részhalmaz kiválasztásakor H összes részhalmazai ugyanakkora valószínűséggel jönnek tekintetbe. Jelölje P_{nk} annak valószínűségét, hogy H egy tetszőleges rögzített x eleme egyértelműen meg lesz határozva azáltal, hogy megadjuk, hogy a kiválasztott k részhalmaz közül melyek tartalmazzák x-et. Ha $\frac{1}{e}<\alpha<1$ és $k\geqq\log_2 n+\log_2\frac{1}{\log\frac{1}{\alpha}}$, akkor $P_{nk}\geqq\alpha$, míg ha $k\leqq\log_2(n-1)+\log_2\frac{1}{\log\frac{1}{\alpha}}$, akkor $P_{nk}\leqq\alpha$.*

Megjegyzés. Ha $\alpha = 0{,}99$, akkor $\log_2 \frac{1}{\log \frac{1}{\alpha}} < \log_2 100 < 7$, tehát ha a Bar—Kochba játéknál kérdéseinket találomra választjuk meg, és n dolog valamelyikére lehet gondolni, a gondolt dolgot $\{\log_2 n\} + 7$ találomra megválasztott kérdés alapján 99%-ot meghaladó valószínűséggel ki fogjuk találni. Meglepő, hogy a szükséges kérdések száma mindig 7-tel lesz nagyobb, mint ha optimális módon tennénk fel a kérdéseket, függetlenül attól, hogy mekkora n, vagyis függetlenül attól, hogy mi a minimális kérdésszám. Ha például a játékosok abban állapodnak meg, hogy egy 6-jegyű telefonszámot kell kitalálni, akkor ez 27 találomra feltett kérdés alapján 99% valószínűséggel sikerül ($2^{20} > 10^6 > 2^{19}$).

Könnyen belátható, hogy a részhalmazok választására tett feltevésünkből következik, hogy általában a kiválasztott részhalmazok mindegyike a H halmaz elemeinek körülbelül a felét fogja tartalmazni.*

Felmerül a kérdés, hogy hogyan módosul a helyzet, ha a részhalmazok kiválasztási szabályát úgy választjuk, hogy a kiválasztott részhalmazok túlnyomórészt kb. pn elemből álljanak, ahol $0 < p < 1$ és $p \neq \frac{1}{2}$.

Egy ilyen kiválasztási szabályt könnyen nyerhetünk, ha továbbra is feltesszük, hogy az ε_{jh} valószínűségi változók függetlenek, azonban ezek eloszlását úgy választjuk, hogy

$$\mathbf{P}(\varepsilon_{jh} = 1) = p \quad \text{és így} \quad \mathbf{P}(\varepsilon_{jh} = 0) = q = 1 - p \tag{1.6}$$

legyen. Ez annak felel meg, hogy a H_j részhalmazt úgy választjuk meg, hogy H minden egyes x_h eleméről sorsolás útján döntjük el, hogy belekerüljön-e H_j-ba, és feltesszük, hogy a sorsolás úgy történik, hogy ennek valószínűsége j és h minden értékére p-vel egyenlő. Ez esetben is elkészítjük az ε_{jh} számokból az E_{kn} mátrixot, és az, hogy az ismeretlen x elem meg van határozva azáltal, hogy megadjuk, hogy a $H_1, H_2, \ldots, H_k$ halmazok közül melyek tartalmazzák x-et, az $x = x_i$ esetben továbbra is azzal az állítással ekvivalens, hogy az E_{kn} mátrix i-edik oszlopa különbözik-e a mátrix összes többi oszlopától. Most azonban ezen esemény P_{nk} valószínűsége már függ attól, hogy az i-edik oszlopban hány 1-es áll. A teljes valószínűség tételét alkalmazva a P_{nk} valószínűségre, a

$$P_{nk} = \sum_{v=0}^{k} \binom{k}{v} p^v q^{k-v} (1 - p^v q^{k-v})^{n-1}, \tag{1.7}$$

kifejezést nyerjük. Alkalmazva a Moivre—Laplace tételt, némi számolással adódik, hogy érvényes a következő tétel:

2a. TÉTEL. *Egy n elemű H halmaznak válasszuk ki találomra, egymástól függetlenül k részhalmazát, oly módon, hogy H minden egyes elemére vonatkozólag, egymástól függetlenül, sorsolással döntjük el, hogy az illető elem beletartozzék-e vagy nem a szóban forgó részhalmazba és ezen lehetőségek valószínűségei p és* $q = 1 - p$ *legyenek* ($0 < p < 1$; $p \neq \frac{1}{2}$). *Jelölje* P_{nk} *annak a valószínűségét, hogy H egy tetszőlegesen meg-*

* Pontosabban: mindegyik halmaz elemszáma nagy valószínűséggel az $\frac{n \pm \sqrt{(2+\delta) n \log \log n}}{2}$ határok közé fog esni, ahol $\delta > 0$ tetszőleges, és $n \geqq n_0(\delta)$, ahol $n_0(\delta)$ csak δ-tól függ.

választott rögzített x elemét egyértelműen jellemezze az, hogy a kiválasztott részhalmazok közül melyek tartalmazzák x-et. Ez esetben, bevezetve az $I_1(\{p, q\}) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}$ *jelölést, ha* $n \to +\infty$ *és*

$$k = k(n) = \frac{\log_2 n + y\sqrt{\log_2 n} + o(\log_2 n)}{I_1(\{p, q\})}, \tag{1. 8}$$

akkor

$$\lim_{n \to +\infty} P_{n,k(n)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y/\sigma} e^{-t^2/2} dt, \tag{1. 9}$$

ahol

$$\sigma = \sqrt{\frac{pq}{I_1(\{p, q\})}} \cdot |\log p/q|. \tag{1. 10}$$

Megjegyzés. Ha (1. 10)-ben $p = \frac{1}{2}$, akkor $\sigma = 0$; ez esetben a tétel állítása akkor marad érvényes, ha $y \neq 0$ és y/σ alatt $+\infty$ ill. $-\infty$ értendő, aszerint, hogy $y > 0$, ill. $y < 0$.

2. §. Információ-akkumuláció egy ismeretlenre vonatkozó általános osztályozások esetében

E §-ban az 1. §-ban tárgyalt probléma következő általánosításával foglalkozunk. Az n elemű H halmazt k-szor egymásután találomra r darab (számozott) részre bontjuk, $(r > 2)$ oly módon, hogy az egyes felbontások egymástól függetlenül történnek és minden felbontás ugyanakkora $\left(\text{tehát } \frac{1}{r^n}\right)$ valószínűséggel lesz egyenlő az összes lehetséges r-részre való felbontások mindegyikével. Kiszámítandó annak a valószínűsége — ezt P_{nk}-val jelöljük —, hogy kijelölve H egy tetszőleges rögzített x elemét, ez az elem egyértelműen meg legyen határozva annak megadása által, hogy a k felbontás mindegyikénél x hányadik részhalmazba tartozik.

Nyilván a szóban forgó felbontások mindegyike jellemezhető azon függvény által, amely H minden eleméhez hozzárendeli a felbontás azon részhalmazának sorszámát, amelyhez az illető elem tartozik. Ez a függvény (és ennek következtében a számbajövő felbontás) nyilván r^n-féleképpen adható meg. Jelölje ε_{jh} a j-edik felbontáshoz tartozó ilyen függvényt, tehát legyen

$$\varepsilon_{jh} = l, \tag{2. 1}$$

ha x_h a j-edik felbontásnál az l-edik osztályba tartozik $(l = 1, 2, \ldots, r;\ j = 1, 2, \ldots, k;\ h = 1, 2, \ldots, n)$. Ugyanúgy, mint az 1. §-ban, beláthatjuk, hogy az ε_{jh} változók függetlenek és

$$\mathbf{P}(\varepsilon_{jh} = l) = \frac{1}{r} \qquad (j = 1, 2, \ldots k;\ h = 1, 2, \ldots, n;\ l = 1, 2, \ldots, r). \tag{2. 2}$$

Elkészítve újból az ε_{jh} számokból az E_{kn} mátrixot, az $x = x_i$ esetben P_{nk} annak valószínűségével lesz egyenlő, hogy E_{kn} i-edik oszlopa az összes többi oszloptól különbözik. Így nyerjük, hogy

$$P_{nk} = \left(1 - \frac{1}{r^k}\right)^{n-1}, \tag{2. 3}$$

tehát, ha $r^k > n$, akkor

$$e^{-\frac{n}{r^k}} < P_{nk} < e^{-\frac{n-1}{r^k}}. \tag{2. 4}$$

Ily módon az 1. tétel következő általánosítását kaptuk:

1b. TÉTEL. *Ha az n elemű H halmaznak találomra kiválasztjuk k számú r (számozott) részre való felbontását ($r \geqq 2$) oly módon, hogy az egyes felbontásokat egymástól függetlenül választjuk ki úgy, hogy az r^n lehetséges felbontás mindegyike ugyanolyan valószínűséggel kerülhet kiválasztásra és P_{nk}-val jelöljük annak a valószínűségét, hogy H egy tetszőleges rögzített x eleme egyértelműen meg legyen határozva azáltal, hogy megadjuk, hogy a k felbontás mindegyikénél melyik osztályba tartozik, továbbá, ha* $\frac{1}{e} < \alpha < 1$ *és* $k \geqq \dfrac{\log_2 n + \log_2\left(\log \frac{1}{\alpha}\right)^{-1}}{\log_2 r}$, *akkor* $P_{nk} \geqq \alpha$, *míg ha* $k \leqq \leqq \dfrac{\log_2 (n-1) + \log_2\left(\log \frac{1}{\alpha}\right)^{-1}}{\log_2 r}$, *akkor* $P_{nk} \leqq \alpha$.

Megjegyzés. Ha $r = 2$, akkor az 1b. tételből speciális esetként az 1a. tételt kapjuk.

A mondott feltételek mellett minden egyes felbontás minden egyes részhalmaza 1-hez közeli valószínűséggel $\dfrac{n \pm O(\sqrt{n \log\log n})}{r}$ elemet fog tartalmazni. Ha azt akarjuk, hogy a felbontások részhalmazai ne álljanak körülbelül ugyanannyi elemből, hanem elemszámaik a $p_1 n, p_2 n, \ldots, p_r n$ számokhoz essenek közel, ahol $\mathfrak{P} = \{p_1, p_2, \ldots, p_r\}$ egy tetszőleges r elemű valószínűségeloszlás, azaz $p_l > 0$ $(l = 1, 2, \ldots, r)$ és $\sum_{l=1}^{r} p_l = 1$, akkor a felbontások véletlenszerű megválasztását úgy kell módosítani, hogy

$$\mathbf{P}(\varepsilon_{jh} = l) = p_l \quad (l = 1, 2, \ldots, r;\ j = 1, 2, \ldots, k;\ h = 1, 2, \ldots, n) \tag{2. 5}$$

legyen.

Ez esetben az 1. §-ban alkalmazott meggondoláshoz hasonló módon a

$$P_{nk} = \sum_{\substack{\sum_{l=1}^{r} k_l = k}} \frac{k!}{k_1!\, k_2! \cdots k_r!}\, p_1^{k_1} \cdot p_2^{k_2} \cdots p_r^{k_r} (1 - p_1^{k_1} \cdot p_2^{k_2} \cdots p_r^{k_r})^{n-1} \tag{2. 6}$$

képletet nyerjük. Ebből némi számolással adódik a következő

2b. TÉTEL. *Ha egy n elemű H halmaz elemeit k-szor találomra* $r \geqq 2$ *számozott osztályra bontjuk fel, oly módon, hogy az egyes felbontások egymástól függetlenek és H minden eleme minden egyes felbontás l-edik osztályába* p_l *valószínűséggel esik* $\left(p_l > 0,\ l = 1, 2, \ldots, r;\ \sum_{l=1}^{r} p_l = 1\right)$ *és* P_{nk} *jelöli annak a valószínűségét, hogy H egy tetszőleges rögzített x eleme egyértelműen meg legyen határozva azáltal, hogy megadjuk, hogy a k felbontás mindegyikénél x hányadik osztályba tartozik, továbbá, ha*

$$k = k(n) = \frac{\log_2 n + y\sqrt{\log_2 n} + o(\sqrt{\log_2 n})}{I_1(\mathfrak{S})}, \tag{2.7}$$

ahol $\mathfrak{S}$ *jelöli a* $(p_1, p_2, \ldots, p_r)$ *eloszlást, amelyről feltesszük, hogy elemei nem mind egyenlők, és*

$$I_1(\mathfrak{S}) = \sum_{l=1}^{r} p_l \log_2 \frac{1}{p_l}, \tag{2.8}$$

akkor

$$\lim_{n \to +\infty} P_{n,\,k(n)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y/\sigma} e^{-t^2/2}\, dt, \tag{2.9}$$

ahol

$$\sigma = \sqrt{\frac{\sum_{l=1}^{r} p_l \log_2^2 p_l - \left(\sum_{l=1}^{r} p_l \log_2 \frac{1}{p_l}\right)^2}{I_1(\mathfrak{S})}}. \tag{2.10}$$

Megjegyzés. Ha $r = 2, p_1 = p, p_2 = q$ $(p + q = 1, p \neq \frac{1}{2})$, akkor a (2. 10) által definiált σ értéke $\sqrt{\frac{pq}{I_1(\{p, q\})}}\,|\log p/q|$, és így a 2b. tétel az $r = 2$ speciális esetben a 2a. tételre redukálódik. Ha viszont r tetszőleges, viszont $p_1 = p_2 = \ldots = p_r = \frac{1}{r}$, akkor a (2. 10) által definiált σ értéke 0; a 2b. tétel állítása ez esetben is érvényes marad $y \neq 0$-ra, ha y/σ-t $+\infty$-nek ill. $-\infty$-nek értelmezzük, aszerint, hogy $y > 0$ vagy $y < 0$.

3. §. Szimultán információ-akkumuláció véletlen dichotomiák esetében

E §-ban először a következő problémával foglalkozunk: Egy n elemű H halmaznak egyidejűleg l különböző ismeretlen elemét kívánjuk meghatározni oly módon, hogy egymástól függetlenül találomra kiválasztjuk H k részhalmazát (úgy, hogy minden választásnál minden részhalmaz ugyanakkora valószínűséggel kerül kiválasztásra) és feltesszük, hogy az l ismeretlen elem mindegyikét illetőleg megállapítható, hogy azok a kiválasztott részhalmazok közül melyekhez tartoznak hozzá. Jelöljük P_{nkl}-lel annak a valószínűségét, hogy ezek az adatok egyértelműen meghatározzák a H halmaz szóban forgó l elemét. Jelöljék újból $x_1, x_2, \ldots, x_n$ a H halmaz elemeit, $H_1, H_2, \ldots, H_k$ a kiválasztott részhalmazokat és legyen $\varepsilon_{jh} = 1$ vagy 0 aszerint, hogy x_h eleme-e H_j-nak vagy nem $(j = 1, 2, \ldots, k;\ h = 1, 2, \ldots, n)$. Akkor P_{nkl} nyilván

egyenlő annak valószínűségével, hogy az ε_{jh} elemekből alkotott E_{kn} mátrixnak az l ismeretlen elemnek megfelelő oszlopa egymástól, valamint az összes többi oszloptól különbözzék. Ilyen módon azt kapjuk, hogy

$$P_{nkl} = \prod_{v=0}^{l-1}\left(1-\frac{v}{2^k}\right)\left(1-\frac{l}{2^k}\right)^{n-l}. \tag{3. 1}$$

Egyszerűen adódik (3. 1)-ből, hogy ha n, k és l úgy tartanak végtelenhez, hogy $l=o(n)$ és $k=\log_2 nl+\log_2 \frac{1}{\log\frac{1}{\alpha}}+o(1)$, ahol $0<\alpha<1$, akkor

$$\lim P_{nkl}=\alpha. \tag{3. 3}$$

Ezzel szemben, ha n és l úgy tartanak végtelenhez, hogy

$$l\sim\vartheta n \qquad (0<\vartheta\leqq 1), \tag{3. 4}$$

akkor (3. 3) abban az esetben teljesül, ha

$$k = \log_2 nl+\log_2\frac{1-\frac{\vartheta}{2}}{\log\frac{1}{\alpha}}+o(1), \tag{3. 5a}$$

vagyis, ha

$$k = 2\log_2 n+\log_2\frac{\vartheta(2-\vartheta)}{2\log\frac{1}{\alpha}}+o(1). \tag{3. 5b}$$

Speciálisan, ha

$$l=n, \tag{3. 6}$$

akkor (3. 3) fennállásához kell, hogy

$$k = 2\log_2 n+\log_2\frac{1}{2\log\frac{1}{\alpha}}+o(1) \tag{3. 7}$$

legyen.

Érvényes tehát a következő tétel:

3a. TÉTEL. *Ha az n elemű H halmaznak egymástól függetlenül találomra kiválasztjuk k részhalmazát, oly módon, hogy minden választásnál H 2^n részhalmaza közül mindegyik ugyanolyan valószínűséggel kerülhet kiválasztásra, és P_{nk} jelöli annak valószínűségét, hogy a kiválasztott k részhalmaz H bármely két elemét elválasztja egymástól, azaz, hogy nincs H-nak két olyan eleme, amelyek az összes kiválasztott részhalmazzal ugyanolyan relációban állnak (ugyanazoknak a részhalmazoknak*

elemei, ill. nem elemei) és n és k oly módon tartanak végtelenhez, hogy

$$k = 2\log_2 n + \log_2 \frac{1}{2\log\frac{1}{\alpha}} + o(1),$$

ahol $0<\alpha<1$, *akkor*

$$\lim_{n\to+\infty} P_{nk} = \alpha.$$

Összehasonlítva a 3a. tételt az 1a. tétellel, azt látjuk, hogy egy n elemű halmaz *összes* elemeinek szeparálásához kereken kétszer annyi dichotomia szükséges, mint e halmaz *egyetlen* elemének szeparálásához. E problémánál valójában egy ismeretlen n-edrendű permutációt kell meghatároznunk. Ehhez legalább $\log_2 n! \sim n\log_2 n$ „bit" információra van szükség; egy dichotomia a permutáció minden egyes elemére nézve maximálisan 1 bit, az egész permutációra nézve n bit információt adhat. Más szóval, ha véletlen dichotomiákból határozzuk meg az ismeretlen permutációt, kb. 100%-os redundancia lép fel.*

4. §. Szimultán információ-akkumuláció, általános osztályozások esetében

Ha az n elemű H halmaz l különböző elemét akarjuk meghatározni és minden egyes rész-információ abban áll, hogy találomra kiválasztjuk a H halmaz egy r részre való felbontását (úgy, hogy minden választásánál az összes ilyen felbontások egyformán valószínűek, és az egyes választások egymástól függetlenek), és ezután értesülünk arról, hogy az l ismeretlen elem mindegyike minden egyes felbontásnál melyik részhalmazhoz tartozik, akkor — P_{nkl}-lel jelölve annak a valószínűségét, hogy k ilyen felbontás elegendő az l ismeretlen elem meghatározására — az előző §-ban alkalmazott meggondoláshoz hasonlóan beláthatjuk, hogy

$$P_{nkl} = \prod_{\nu=0}^{l-1}\left(1-\frac{\nu}{r^k}\right)\left(1-\frac{l}{r^k}\right)^{n-l}. \tag{4. 1}$$

Ha tehát

$$l = o(n), \tag{4. 2}$$

akkor a

$$k = \frac{\log_2 nl + \log_2 \frac{1}{\log\frac{1}{\alpha}} + o(1)}{\log_2 r} \tag{4. 3}$$

* Ha $l>n$, viszont megengedett az, hogy — a szimultán Bar—Kochba terminológiáját használva — több játékos gondoljon ugyanarra a dologra, akkor sem szükséges több dichotomia mint az $l=n$ esetben, hiszen $2\log_2 n + \log_2 \frac{1}{2\log\frac{1}{\alpha}}$ dichotomia körülbelül α valószínűséggel elégséges a H halmaz minden elemének szeparálásához és teljesen mellékes, hogy hány játékos gondolt ugyanarra a dologra, hiszen akik ugyanarra gondoltak, mindig ugyanazokat a válaszokat adják és így, ha B kitalálta, hogy egyikük mire gondolt, akkor a többiről is tudja ugyanezt.

esetben lesz

$$\lim P_{nkl} = \alpha, \qquad (0 < \alpha < 1), \tag{4. 4}$$

míg ha

$$l \sim \vartheta n, \qquad (0 < \vartheta \leqq 1), \tag{4. 5}$$

akkor a

$$k = \frac{2 \log_2 n + \log_2 \dfrac{\vartheta(2-\vartheta)}{2 \log \dfrac{1}{\alpha}} + o(1)}{\log_2 r} \tag{4. 6}$$

esetben lesz (4. 4) érvényes.

Ezek után azt várnánk, hogy ha az r-részre osztásnál az egyes részek elemszámai körülbelül úgy viszonylanak egymáshoz, mint a $p_1, p_2, \ldots, p_r$ számok $\left(\sum_{i=1}^{r} p_i = 1\right)$, akkor a fenti eredmények csak annyiban módosulnak, hogy $\log_2 r$ helyébe $I_1(\mathcal{S})$ lép. A tüzetesebb vizsgálat azonban azt mutatja, hogy ez nem igaz, ugyanis $\log_2 r$ helyébe nem a Shannon-féle $I_1(\mathcal{S}) = \sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i}$ elsőrendű információ-mennyiség, hanem az $I_2(\mathcal{S}) = \log_2 \left(\frac{1}{\sum_{i=1}^{r} p_i^2} \right)$ másodrendű információ-mennyiség lép. (Persze a $p_1 = p_2 = \ldots = p_r = 1/r$ esetben $I_1(\mathcal{S}) = I_2(\mathcal{S}) = \log_2 r$.) Ezt a következőképpen láthatjuk be: Jelöljék $\pi_1, \pi_2, \ldots, \pi_{r^k}$ valamilyen sorrendben az olyan k-tényezős szorzatokat, amelyek minden tényezője a $p_1, p_2, \ldots, p_r$ számok egyike (sorrend is számít!). Nyilván a π_i számok mind $p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r}$ alakúak, ahol $\sum_{i=1}^{r} k_i = k$ és ez a szám a π_i számsorozatban $\frac{k!}{k_1!\, k_2! \ldots k_r!}$-szor fordul elő.

Az egyszerűség kedvéért csak a legérdekesebb esetet tekintjük, vagyis azt, amikor $l = n$.

Ez esetben P_{nk}-val jelölve annak valószínűségét, hogy a k felbontás H összes elemeit szeparálja,

$$P_{nk} = \sum{}' \pi_{i_1} \pi_{i_2} \ldots \pi_{i_n}, \tag{4. 7}$$

ahol az összegezés az összes *különböző* számokból álló $(i_1, i_2, \ldots, i_n)$-szám n-esekre terjesztendő ki.

A (4. 7) jobb oldalán álló összeg aszimptotikus viselkedésének megvizsgálásához éppen arra a szita-formulára van szükség, amelyet egy előző dolgozatunkban [5] bizonyítottunk be. Jelölje B_{ij} azt az eseményt, hogy az E_{kn} mátrix i-edik és j-edik oszlopa azonos. E jelölés mellett P_{nk} annak a valószínűségével egyenlő, hogy a B_{ij} $(1 \leqq i < j \leqq n)$ események egyike sem következik be. Jelölje G azt a gráfot, amelynek szögpontjai az (i, j) számpárok $(1 \leqq i < j \leqq n)$ és amelyben az (i, j) és (i', j') számpároknak megfelelő szögpontok akkor és csak akkor vannak összekötve egy éllel,

ha az i, j, i', j' számok nem mind különbözők (tehát, ha $i=i'$ vagy $i=j'$ vagy $j=i'$ vagy $j=j'$; nyilvánvaló, hogy e négy lehetőség közül $i<j$ és $i'<j'$ miatt legfeljebb egy állhat fenn). Jelölje Γ^* a G szögpontjai halmazának azon részhalmazai összességét, amelyek nem tartalmaznak egyetlen G-ben éllel összekötött szögpontpárt sem, és Γ^{**} a G szögpontjai halmazának azon részhalmazai összességét, amelyek legfeljebb egy G-ben éllel összekötött szögpontpárt tartalmaznak. Legyen

$$S_d^* = \Sigma^* P(B_{i_1 j_1} B_{i_2 j_2} \dots B_{i_d j_d}), \tag{4.8}$$

ill.

$$S_d^{**} = \Sigma^{**} P(B_{i_1 j_1} B_{i_2 j_2} \dots B_{i_d j_d}), \tag{4.9}$$

ahol Σ^* ill. Σ^{**} azt jelöli, hogy az összegezés csak az olyan $(i_1, j_1), \dots, (i_d, j_d)$ számpár d-esekre terjesztendők ki, amelyek Γ^*-hoz, ill. Γ^{**}-hoz tartoznak. Ez esetben az [5]-ben bebizonyított tételünk szerint s bármely nemnegatív egész értékére

$$1 - S_1^{**} + S_2^{**} - S_3^{**} + \dots - S_{2s+1}^{**} \leqq P_{nk} \leqq 1 - S_1^* + S_2^{**} - \dots + S_{2s}^{**}. \tag{4.10}$$

Azonban nyilvánvalóan

$$S_d^* = \left(\sum_{i=1}^{r^k} \pi_i^2 \right)^d \frac{n(n-1)\dots(n-2d+1)}{d!\, 2^k}, \tag{4.11}$$

továbbá

$$S_d^{**} = S_d^* \left(1 + \frac{8 \left(\sum_{i=1}^{r^k} \pi_i^3 \right) d(d-1)}{3 \left(\sum_{i=1}^{r^k} \pi_i^2 \right)^2 (n-2d+1)} \right). \tag{4.12}$$

Ha mármost

$$k = \frac{2 \log_2 n + \log \dfrac{1}{2 \log \dfrac{1}{\alpha}} + o(1)}{\log_2 \dfrac{1}{\sum\limits_{v=1}^{r} p_v^2}}, \tag{4.13}$$

akkor figyelembe véve, hogy

$$\sum_{i=1}^{r^k} \pi_i^2 = \left(\sum_{v=1}^{r} p_v^2 \right)^k, \tag{4.14}$$

azt kapjuk, hogy d minden rögzített értékére

$$S_d^* \sim \frac{\left(\log \dfrac{1}{\alpha} \right)^d}{d!}, \tag{4.15}$$

továbbá figyelembe véve, hogy

$$\sum_{i=1}^{r^k} \pi_i^3 = \left(\sum_{v=1}^{r} p_v^3 \right)^k, \tag{4.16}$$

és ezért, bevezetve a $\bar{p} = \max_{1 \leqq v \leqq r} p_v$ jelölést,

$$\sum_{i=1}^{r^k} \pi_i^3 < \bar{p}^k \left(\sum_{v=1}^{r} p_v^2 \right)^k, \tag{4.17}$$

azt kapjuk, hogy

$$S_d^{**} = S_d^{*}(1 + o(n^{-\varrho})), \tag{4.18}$$

ahol

$$\varrho = \frac{\log_2 \dfrac{\sum_{v=1}^{r} p_v^2}{\bar{p}^2}}{\log_2 \dfrac{1}{\sum_{v=1}^{r} p_v^2}} > 0. \tag{4.19}$$

Tehát s minden nemnegatív egész értékére

$$\sum_{d=0}^{2s+1} \frac{(-1)^d \left(\log \frac{1}{\alpha} \right)^d}{d!} \leqq \underline{\lim}\, P_{nk} \leqq \overline{\lim}\, P_{nk} \leqq \sum_{d=0}^{2s} \frac{(-1)^d \left(\log \frac{1}{\alpha} \right)^d}{d!}. \tag{4.20}$$

Ilyen módon, elvégezve az $s \to +\infty$ határátmenetet nyerjük, hogy

$$\lim P_{nk} = e^{-\log \frac{1}{\alpha}} = \alpha. \tag{4.21}$$

Ezzel tehát bebizonyítottuk a következő tételt:

3b. TÉTEL. *Ha egy n elemű H halmaz elemeit egymástól függetlenül k-szor r (számozott) osztályba soroljuk, oly módon, hogy H elemei minden egyes osztályba-sorolásnál egymástól függetlenül p_v valószínűséggel kerülnek a v-edik osztályba ($v = 1, 2, \ldots, r$), és P_{nk} jelöli annak a valószínűségét, hogy a szóban forgó k felosztás H bármely két elemét elválasztja, vagyis, hogy ne legyen H-nak két olyan különböző eleme, amelyek mindegyik választott felosztásnál ugyanabba az osztályba esnek, továbbá, ha n és k úgy tartanak végtelenhez, hogy*

$$k = \frac{2 \log_2 n + \log_2 \dfrac{1}{\log \frac{1}{\alpha}} + o(1)}{\log_2 \dfrac{1}{\sum_{v=1}^{r} p_v^2}} \tag{4.22}$$

ahol $0<\alpha<1$, *akkor*

$$\lim P_{nk}=\alpha. \tag{4.23}$$

Nyilvánvaló, hogy a 3b. tétel speciális esetként tartalmazza (az $r=2, p_1=p_2=\frac{1}{2}$ esetben) a 3a. tételt.

IRODALOMJEGYZÉK

[1] RÉNYI ALFRÉD: Az információelmélet néhány alapvető kérdése, *MTA III. Oszt. Közl.* **10** (1960) 251–282.

[2] „ On measures of entropy and information, Proceedings of the 4th Berkeley Symposium on *Mathematical Statistics and Probability*, University of California Press, 1961, Vol. I. 547–561.

[3] „ On random generating elements of a finite Boolean algebra, *Acta Sci. Math. Szeged*, **22** (1961) 75–81.

[4] „ Statistical laws of accumulation of information, *Bulletin of the International Statistical Institute, 33rd. Session of the ISI in Paris*, 1961. pp. 1–7.

[5] „ Egy általános módszer valószínűségszámítási tételek bizonyítására és annak néhány alkalmazása, *MTA III. Oszt. Közl.* **11** (1961) 79–105.

(Beérkezett: 1961. XI. 9.)