DIPLOMAMUNKA

Diverzitás, koncentráltság és Pareto-elv – epidemiológiai alkalmazásokkal

Bősze Beatrix

V. éves matematikus

Témavezető: dr. Izsák János egyetemi tanár

Berzsenyi Dániel Főiskola, Állattan Tanszék

Konzulens: dr. Tóth János egyetemi docens

BME TTK Matematika Intézet, Matematikai Analízis Tanszék

BME

2006

Tartalomjegyzék

1. fejezet

Bevezetés

A dolgozat felépítése a következő: pontos (az irodalomban megszokottnál néhány esetben körültekintőbb) definícióját adjuk a szereplő általános, illetve speciális mérőszámoknak [5]. Megvizsgáljuk az ezek között fenálló kapcsolatokat matematikai szempontból. Kiemeljük, hogy számos – jelen ismereteink mellett – megoldatlan problémát is megfogalmazunk. A következő fejezetben a Pareto-elv [10] matematikai leírásával foglalkozunk.

Végül pedig a KSH évekre és területi egységekre (megyék és Budapest) lebontott epidemiológiai adatait elemezzük a tanulmányozott eszközökkel.

2. fejezet

Biológiai diverzitás és koncentráltság

A címben szereplő diverzitás és koncentráltság kifejezéseket a közgazdaságtanban, az epidemiológiában, a biológiában, sőt a hétköznapi nyelvben is használjuk [5, 9, 14]. E két kifejezés jelentését tekintve ellentétes tartalmú. A diverzitást a szétosztottság mértékének maghatározásánál, míg a koncentráltságot a dominancia mértékének meghatározására használjuk. Ahhoz, hogy koncentráltsági és diverzitási szempontból össze tudjunk hasonlítani kettő vagy több mintát, számszerűsíteni kell ezeket a tulajdonságokat, tehát mérőszámok bevezetésére van szükség. Elsőre talán feleslegesnek érezhetjük mindkét szempont szerint vizsgálni egy mintát. Csakhogy különböző tudományokban eltérő időben és okból vált fontossá a koncentráltság, illetve a diverzitás mérése. Egyik feladatunknak éppen azt tűztük ki, hogy megvizsgáljuk, hogy szükség van-e ilyen sok mérőszámra, illetve egyáltalán a két fő csoportra. Az olyan tudományterületeken, ahol ismert a kategóriák száma (pl.: biológia, epidemiológia), ott a diverzitás mérésére vezettek be indexeket. Ezek általában szakterületenként különböző függvények. Ezzel szemben, ahol nincs nagy szerepe a kategóriák számának (pl. mert előre nem lehet meghatározni), ott koncentráltsági mérőszámokat használnak (pl.: közgazdaságtan:

jövedelem-eloszlás).

Annak ellenére, hogy számtalan tudományos területen használatosak, a következőkben a koncentráltsági és diverzitási mértékek bevezetésénél, illetve az indexekre kirótt követelményeknél biológiai kifejezéseket fogunk használni.

2.1. Diverzitás

Vizsgálódásunk tárgyai (leggyakrabban: biológiai) **populáció**knak nevezett (véges) halmazok, amelyek részhalmazai **faj**ok, a fajok elemei pedig **egyed**ek [6]. Az egyedeket $s \in \mathbb{N}$ számú faj esetén olyan X_s diszkrét valószínűségi változóval modellezzük, amelynek eloszlása az

$$S^s := \{(\pi_1, \ldots, \pi_s) \in (\mathbb{R}_0^+)^s; \sum_{i=1}^{s} \pi_i = 1\}$$

s-dimenziós szimplex valamely eleme, s amelynek értéke $i \in \{1, 2, \ldots, s\}$, ha a kiválasztott egyed az i-edik fajhoz tartozik, továbbá $P(X_s = i) = \pi_i$. Kiemelt szerepet fognak játszani az alábbiakban a $\boldsymbol{\pi}_0^s := (\frac{1}{s}, \frac{1}{s}, \ldots, \frac{1}{s}) = \frac{1}{s}\mathbf{1}^s$ **(diszkrét)egyenletes eloszlások** ($\mathbf{1}^s \in \mathbb{R}^s$ minden eleme 1), továbbá a **monopólium**ok, amely névvel itt $\mathbb{R}^s$ standard bázisának $\mathbf{e}_1^s, \mathbf{e}_2^s, \ldots, \mathbf{e}_s^s$ elemeire hivatkozunk. (Ez utóbbiaknál tehát az összes egyed egyetlen fajból származik.) Tetszőleges $s \in \mathbb{N}, \boldsymbol{\pi} \in S^s$ esetén jelölje $\boldsymbol{\pi}^\downarrow = (\pi_1^\downarrow, \pi_2^\downarrow, \ldots, \pi_s^\downarrow) \in S^s$ azt az eloszlást, amelynek tagjai azonosak a $\boldsymbol{\pi}$ eloszlás tagjaival, csak monoton csökkenő sorrendben követik egymást.

A populációk változatosságát olyan $\mathrm{Div} : \cup_{s\in\mathbb{N}}(\{s\} \times S^s) \longrightarrow \mathbb{R}$ **diverzitási függvény**ekkel (diverzitási indexekkel vagy diverzitásokkal) mérjük, amelyek eleget tesznek az alábbi, **kanonikus tulajdonság**oknak nevezett követelményeknek.

1. A második változójukban permutációra nézve invariánsak. Emiatt általában az eloszlás tagjait például nagyság szerint csökkenő sorrendben rendezzük, és így használjuk argumentumként.

2. Rögzített s mellett, minimumukat a monopóliumokon, maximumukat az egyenletes eloszlásokon felveszik: $\forall s \in \mathbb{N} \forall \boldsymbol{\pi} \in S^s \quad \mathrm{Div}(s, \mathbf{e}_1^s) \leq \mathrm{Div}(s, \boldsymbol{\pi}) \leq \mathrm{Div}(s, \boldsymbol{\pi}_0^s)$. (Megkövetelhetjük azt is, hogy az – argumentumok különbözősége esetén – szigorú egyenlőtlenség álljon fenn.)

3. A fajok számának növekedtével az egyenletes eloszlás diverzitása nem csökken, a monopóliumoké (amelyek közül ismét az 1. tulajdonság miatt nyilván elegendő csupán eggyel foglalkozni) nem nő:

   (a) $\forall s \in \mathbb{N} \quad \mathrm{Div}(s, \boldsymbol{\pi}_0^s) \leq \mathrm{Div}(s+1, \boldsymbol{\pi}_0^{s+1})$,

   (b) $\forall s \in \mathbb{N} \quad \mathrm{Div}(s, \mathbf{e}_1^s) \geq \mathrm{Div}(s+1, \mathbf{e}_1^{s+1})$.

4. Néha megköveteljük a következő tulajdonságot is. Osszuk fel az egyedek halmazát kétféle (A és B) osztályozás (pl. faj és élőhely típusa) szerint. Legyen $\pi_{1*}, \pi_{2*}, \ldots, \pi_{s*}$ az A-beli kategóriák előfordulási valószínűsége. $\pi_{*1}, \pi_{*2}, \ldots, \pi_{*t}$ a B-beli kategóriák előfordulási valószínűsége, végül pedig legyen $\pi_{ij} \quad (i = 1, 2, \ldots, s; j = 1, 2, \ldots, t)$ a szorzatosztályozás osztályainak valószínűsége. A diverzitásoknak az A osztályaira vonatkozó átlaga (előre rögzített diverzitási index-szel dolgozunk):

   $$\mathrm{Div}_A(B) = \sum_{i=1}^{s} \pi_{i*} \mathrm{Div}_i(B),$$

   ahol

   $$\mathrm{Div}_i(B) = \mathrm{Div}(\frac{\pi_{i1}}{\pi_{i*}}, \frac{\pi_{i2}}{\pi_{i*}}, \ldots, \frac{\pi_{ig}}{\pi_{i*}})$$

   Amennyiben teljesül, hogy

   $$\mathrm{Div}(A \times B) = \mathrm{Div}(\pi_{ij}) = \mathrm{Div}(A) + \mathrm{Div}_A(B),$$

   akkor azt mondjuk, hogy az adott diverzitási index teljesíti a harmadik kanonikus tulajdonságot.

Ezek a követelmények még az $s = 2$ esetben is sokféle diverzitási függvényt engednek meg, ugyanis például az átlagos rangszám és a Gini–Simpson-index

(a definíciókat lásd alább) teljesíti az első három tulajdonságot, és különbözik egymástól.

1. Állítás. *Ha* Div *diverzitási index,* $\varphi : \mathbb{R} \to \mathbb{R}$ *pedig olyan monoton növő függvény, amelyre* $\varphi(0) = 0$*, akkor* $\varphi \circ$ Div *is diverzitási index, ugyanis a diverzitási index mindhárom tulajdonsága egyenlőtlenséggel van definiálva.*

2.1.1. A diverzitás, mint átlagos ritkaság

Meglehetősen természetes módon származtathatunk diverzitási függvényeket az egyes fajok ritkaságát mérő $r : \{1, 2, \ldots, s\} \times S^s \longrightarrow \mathbb{R}$ **ritkasági függvény** segítségével. Megköveteljük, hogy nevüknek megfelelően, az adott populációban kisebb valószínűséggel előforduló fajhoz nagyobb értéket rendeljenek, vagyis, ha valamely $i, j \in \{1, 2, \ldots, s\}$ esetén $\pi_i \leq \pi_j$, akkor legyen $r(i, \boldsymbol{\pi}) \geq r(j, \boldsymbol{\pi})$.

Adott r ritkasági függvény esetén képezhető az $r(X_s, \boldsymbol{\pi})$ (a megszokottnál nagyobb minuciózitással: az $r(., \boldsymbol{\pi}) \circ X_s$) valószínűségi változó várható értéke, az **átlagos ritkaság**, sok esetben így kapunk diverzitási fügvényt.

Jöjjenek a példák.

2.1.2. Néhány ritkasági függvény

1. Legyen $\mathrm{rang}(i, \boldsymbol{\pi}) := \pi_i$ utolsó előfordulásának helye a $\boldsymbol{\pi}^{\downarrow}$ eloszlásban. Ez azt jelenti, hogy ha $\pi_i = \pi_j$, akkor $\mathrm{rang}(i, \boldsymbol{\pi}) = \mathrm{rang}(j, \boldsymbol{\pi})$.

   A rangot így számoljuk:

   ```
   Last[Position[Sort[p, #2<#1&], p[[i]]]]
   ```

   Ez nyilván az adott populációban kisebb valószínűséggel előforduló fajhoz nagyobb természetes számot rendel. Az ebből a ritkasági függvényből számolt diverzitás az R **átlagos rangszám**: $R(s, \boldsymbol{\pi}) = 1\pi_1^{\downarrow} + 2\pi_2^{\downarrow} + \cdots + s\pi_s^{\downarrow}$.

2. Állítás. *Az átlagos rangszám teljesíti a diverzitással szemben támasztott első három követelményt.*

1. Bizonyítás.

(a) Egy faj valószínűségének helye a csökkenő sorrendben rendezett eloszlás tagjai között független az fajok sorrendjétől.

(b) A függvény értéke a monopóliumokon 1*, az egyenletes eloszlásnál* s *számú faj esetén* $\frac{s+1}{2}$*. Teljes indukcióval belátjuk, hogy egy tetszőleges eloszlás esetén a diverzitás a két szélső érték közé esik. Az állítás* $s = 2$ *esetére nyilvánvaló. Föltéve, hogy tetszőleges* $\boldsymbol{\pi} \in S^s$ *esetén fennáll, hogy*

$$1 \leq 1\pi_1^{\downarrow} + 2\pi_2^{\downarrow} + \cdots + s\pi_s^{\downarrow} \leq \frac{s+1}{2} \tag{2.1}$$

bizonyítsuk be, hogy (2.1) *fennáll tetszőleges* $\boldsymbol{\varrho} \in S^{s+1}$ *mellett is. Az* (2.1) *indukciós feltevés miatt*

$$1 \leq 1(\varrho_1^{\downarrow} + \varrho_2^{\downarrow}) + 2\varrho_3^{\downarrow} + \cdots + s\varrho_{s+1}^{\downarrow} \leq \frac{s+1}{2}.$$

Mivel

$$\begin{aligned} &1(\varrho_1^{\downarrow} + \varrho_2^{\downarrow}) + 2\varrho_3^{\downarrow} + \cdots + s\varrho_{s+1}^{\downarrow} + (\varrho_2^{\downarrow} + \varrho_3^{\downarrow} + \cdots + \varrho_{s+1}^{\downarrow}) \\ &\qquad = 1\varrho_1^{\downarrow} + 2\varrho_2^{\downarrow} + 3\varrho_3^{\downarrow} + \cdots + (s+1)\varrho_{s+1}^{\downarrow}, \end{aligned} \tag{2.2}$$

és

$$0 \leq \varrho_2^{\downarrow} + \varrho_3^{\downarrow} + \cdots + \varrho_{s+1}^{\downarrow} \leq 1,$$

ezért (2.1) $s+1$ *esetén is fennáll.*

(c) Tetszőleges fajszám esetén a függvény értéke a monopóliumon 1*, tehát a fajszám növelésével nem változik, tehát speciálisan nem is csökken. Az egyenletes eloszlásnál a diverzitás értéke a fajszám monoton növő függvénye, vagyis ha növeljük a fajok számát, akkor nő a diverzitási függvény értéke is.*

2. Rögzítsünk egy q rangszámot ($q \in \{1, 2, \ldots, s-1\}$), majd képezzük a következő függvényt

$$r(i, \boldsymbol{\pi}) := \begin{cases} 0 & \text{ha rang}(i, \boldsymbol{\pi}) \leq q \\ 1 & \text{ha rang}(i, \boldsymbol{\pi}) > q \end{cases}$$

Ez a függvény nyilván az adott populációban kisebb valószínűséggel előforduló fajhoz nem kisebb számot rendel. Az ebből számolt diverzitási index pedig az $r(X_s, \boldsymbol{\pi})$ valószínűségi változó várható értéke:

$$T_q(s, \boldsymbol{\pi}) = \pi^{\downarrow}_{q+1} 1 + \pi^{\downarrow}_{q+2} 1 + \cdots + \pi^{\downarrow}_{s} 1,$$

amely mennyiséget **q-tól kumulált valószínűségi index**nek hívjuk.

3. Állítás. *A kumulált valószínűségi index teljesíti a diverzitással szemben támasztott első három követelményt.*

2. Bizonyítás.

(a) A permutációinvariancia az előzőhöz hasonlóan itt is nyilvánvaló.

(b) Tetszőleges $q \in \{1, 2, \ldots, s-1\}$ értékre a monopóliumon a függvény értéke 0*, az egyenletes eloszlásnál* 1 *(a valószínűségek összege). A többi eloszlásban T_q értéke éppen e kettő érték közé esik, mivel T_q egy eloszlás néhány tagjának összege.*

(c) Növelve a fajszámot a monopóliumon és az egyenletes eloszlásnál felvett függvényértékek nem változnak.

2.1.3. Dichotom ritkasági függvények

Dichotom ritkasági függvény értéke csak az i-edik faj előfordulási valószínűségétől függ, továbbá a függés módja fajonként azonos, azaz létezik olyan monoton csökkenő $\bar{r} : [0, 1] \longrightarrow \mathbb{R}$ függvény, amellyel $\forall i \in \{1, 2, \ldots, s\} \, \forall \boldsymbol{\pi} \in S^s \quad r(i, \boldsymbol{\pi}) = \bar{r}(\pi_i)$. (Ezzel a definícióval összhangban, az előző pontban bevezetett ritkasági függvényeket **nem-dichotom ritkasági függvények**nek nevezhetjük.)

4. Állítás. *Ha* $\mathrm{Div}(s, \boldsymbol{\pi}) := \mathbb{E}(r(X_s, \boldsymbol{\pi})) = \sum_{i=1}^{s} \pi_i \bar{r}(\pi_i)$*, ahol az* $\bar{r}$ *függvény monoton csökkenő, akkor a* Div *diverzitás teljesíti a szükséges három feltételt.*

3. Bizonyítás.

1. *Az index definíciójából nyilvánvaló a permutációinvariancia, hiszen az összeadás kommutatív.*

2. *A diverzitási függvény értéke s számú faj esetén a monopóliumon* $\bar{r}(1)$*, az egyenletes eloszásnál* $\bar{r}(\frac{1}{s})$*. Tetszőleges* $\boldsymbol{\pi}$ *eloszlás eseetén a következőket kapjuk:*

$$\bar{r}(1) \leq \bar{r}(\pi_1^{\downarrow}) = (\pi_1^{\downarrow}+\pi_2^{\downarrow}+\cdots+\pi_s^{\downarrow})\bar{r}(\pi_1^{\downarrow}) \leq \pi_1^{\downarrow}\bar{r}(\pi_1^{\downarrow})+\pi_2^{\downarrow}\bar{r}(\pi_2^{\downarrow})+\ldots \pi_s^{\downarrow}\bar{r}(\pi_s^{\downarrow}).$$

   A bizonyítandó egyenlőtlenség másik felét hasonlóan kaphatjuk meg.

3. Div *értéke a monopólimban s értékétől függetlenül mindig* $\bar{r}(1)$*. A diverzitás egyenletes eloszlásnál s faj esetén* $\bar{r}(\frac{1}{s})$*,* $s+1$ *faj esetén pedig* $\bar{r}(\frac{1}{s+1})$*. Mivel* $\frac{1}{s} \geq \frac{1}{s+1}$*, és az* $\bar{r}$ *függvény monoton csökken, így* $\bar{r}(\frac{1}{s}) \leq \bar{r}(\frac{1}{s+1})$*.*

A továbbiakban mutatunk néhány példát dichotóm ritkasági függvényekre.

$\bar{r}(\pi_i) :=$	$\mathrm{Div}(s, \boldsymbol{\pi})$	Név	Jel
$\frac{1}{\pi_i}$	s	fajok száma	
$\frac{1}{\pi_i - 1}$	$s-1$	redukált fajszám	
$\pi_i - 1$	$1 - \boldsymbol{\pi}^2$	Gini–Simpson-index	GS
$-\ln(\pi_i)$	$-\sum_{i=1}^{s} \pi_i \ln(\pi_i)$	Shannon-index	H
$\frac{1}{n} \ln \frac{n!}{\prod_{i=1}^{s} n_i!}$	$\frac{1}{n} \ln \frac{n!}{\prod_{i=1}^{s} n_i!}$	Brillouin-index	H_B

Az utolsó esetben az $\mathbf{n} = (n_1, n_2, \ldots, n_2)$ fajgyakoriság-vektorral és az $n := \sum_{i=1}^{s} n_i$ összegyedszámmal számoltunk.

2.2. Koncentráltság

A sokfajú populációk jellemzésére használt mérőszámok másik csoportja éppen azt méri, hogy milyen mértékben koncentrálódnak az egyedek a fajok körében. Ezeket a mérőszámokat hagyományosan nem az eloszlásokon, hanem az egyedszámvektoron értelmezik: $f : \cup_{s\in\mathbb{N}}(\{s\} \times (\mathbb{R}^+)^s) \longrightarrow \mathbb{R}$ és **koncentráltsági függvény**eknek vagy koncentráltságoknak nevezik. Velük szemben az alábbi követelményeket szokás előírni.

1. A második változójukban permutációkra nézve invariánsak, továbbá **skálainvariáns**ak: $\forall s \in \mathbb{N}\, \forall \mathbf{n} \in (\mathbb{R}^+)^s\, \forall c \in \mathbb{R}^+ \quad f(s, c\mathbf{n}) = f(s, \mathbf{n})$. Speciálisan ez azt is jelenti, hogy mégis csak értelmezhetők eloszlásokon és egyedszámvektorokon egyaránt.

2. A legkoncentráltabbak a monopóliumok, a legkevésbé koncentráltak pedig az egyenletes eloszlások. Az utóbbiakon a függvények értékét nullának szokás venni. $\forall s \in \mathbb{N}\, \forall\, \mathbf{n} \in (\mathbb{R}^+)^s \quad 0 = f(s, \mathbf{1}^s) \leq f(s, \mathbf{n}) \leq f(s, \mathbf{e}_1^s)$.

3. Legyen $s \in \mathbb{N}$, és tegyük fel, hogy $\mathbf{n} \in (\mathbb{R}^+)^s$ olyan vektor, hogy valamilyen $i, j \in \{1, 2, \ldots, s\}$ mellett $0 < n_i < n_j$. Ha mármost $0 < h < n_i$ tetszőleges, akkor az összes ilyen i, j indexre fennáll, hogy $f(s, \mathbf{n}) < f(s, \mathbf{n} + h(\mathbf{e}_j^s - \mathbf{e}_i^s))$. Ezeknek a feltételeknek az a jelentésük, hogy ha a kisebb létszámú faj egyedét egy nagyobb egyedszámú faj egyedével helyettesítjük (másképp: „szegény ad a gazdagnak”), akkor a koncentráltság nő.

Az alábbi példákból majd kitűnik, hogy ezek a követelmények még az $s = 2$ esetben is sokféle koncentráltsági függvényt engednek meg.

1. Példa. *Ha f koncentráltsági függvény, vagyis teljesülnek rá a vonatkozó feltételek, és $c \in \mathbb{R}^+$, akkor cf is koncentráltsági függvény.*

1. Megjegyzés. *Az alábbi példákból majd kiderül, hogy az viszont nem igaz, hogy ha f_1 és f_2 is koncentráltsági függvény, akkor $f_1 : f_2 =$ állandó.*

1. Tétel. *A fenti feltételeket kielégítő differenciálható f függvényre teljesülnek a következők.*

1. $\forall s \in \mathbb{N} \forall \mathbf{n} \in (\mathbb{R}^+)^s$

$$\exists i \; n_i = \max\{n_1, n_2, \ldots, n_s\} \Longrightarrow \forall\, h \in \mathbb{R}^+ \, f(s, \mathbf{n} + h\mathbf{e}_i^s) > f(s, \mathbf{n}).$$

2. *Ha $\mathbf{n} \in (\mathbb{R}^+)^s$ nem minden komponense egyenlő, akkor*

$$\forall\, h \in \mathbb{R}^+ \, f(s, \mathbf{n} + h\mathbf{1}^s) < f(s, \mathbf{n}).$$

3. *A maximális koordináták szerinti jobboldali parciális deriváltak nemnegatívak:*

$$\frac{\partial f(s+0, \mathbf{n})}{\partial n_i} \geq 0.$$

4. *Az $\mathbf{1}^s$ vektor irányában vett jobboldali iránymenti derivált nempozitív:*

$$\frac{\partial f(s+0, \mathbf{n})}{\partial \mathbf{1}^s} \leq 0.$$

5. $\lim_{h \to +\infty} f(s, \frac{\mathbf{n}+h\mathbf{1}^s}{n+hs}) = 0$. *Itt:* $n := \sum_{j=1}^{s} n_j$.

4. Bizonyítás.

1. *Legyen i olyan index, amire $n_i = \max\{n_1, n_2, \ldots, n_s\}$. A permutációinvariancia miatt feltehető, hogy $i = 1$. Ekkor a koncentráltsági függvény értékei közt a következő egyenlőtlenségnek kell teljesülnie:*

$$f(s, \frac{n_1 + h}{n+h}, \frac{n_2}{n+h}, \ldots, \frac{n_s}{n+h}) > f(s, \frac{n_1}{n}, \ldots, \frac{n_s}{n}).$$

Ez az egyenlőtlenség viszont következik a koncentráltsági indexek harmadik tulajdonságából (szegény ad a gazdagnak). Vagyis elég belátni, hogy

$$\frac{n_1 + h}{n+h} - \frac{n_1}{n} = \sum_{i=2}^{s} (\frac{n_i}{n} - \frac{n_i}{n+h}).$$

Ez gyakorlatilag azt jelenti, hogy a maximális relatív gyakoriságú faj relatív gyakorisága pontosan annyival nőtt, mint amennyivel a többi faj relatív gyakorsága összesen csökkent. Tehát a szegény adott a gazdagnak $s-1$ lépésben. Ez az állítás a bizonyítások után szereplő megjegyzésen alapul. Beszorozva az egyenlet mindkét oldalát $n(n+h)$-val, a következőt kapjuk:

$$nn_1 + hn - nn_1 - hn_1 = \sum_{i=2}^{s} nn_i + hn_i - nn_i$$

Kiemelve h-t:

$$h(n - n_1) = h(\sum_{i=2}^{s} n_i)$$

Minthogy $h \neq 0$, leosztunk h-val:

$$n = \sum_{i=1}^{s} n_i.$$

Minden lépés megfordítható, így az állítást bizonyítottuk.

2. *Lásd [5].*

3. *Az első állításban szereplő képletekből kapjuk az állítást.*

4. *A második állításban szereplő képletekből kapjuk az állítást.*

5. *A (folytonos) f függvény második argumentuma $\lim_{h\to+\infty}$ esetén az egyenletes eloszláshoz tart, melyre pedig a függvény értéke 0.*

2. Megjegyzés. *A harmadik tulajdonság úgy is teljesül, ha tetszőleges k $(k = 1, \ldots, s-1)$darab fajból veszünk elemeket, és a legnagyobb relatív gyakoriságú faj elemeivel helyettesítjük őket. Formálisan: legyen $h < n_i$, ahol $i = 2, \ldots, s$ és $n_1 = \max\{n_1, n_2, \ldots, n_s\}$ mint az előbb is! Ekkor:*

$$f(n_1 + kh, n_2 - h, \ldots, n_{k+1} - h, n_{k+2}, \ldots, n_s) > f(n_1, \ldots, n_s).$$

5. Bizonyítás. *Teljes indukció k-ra.*
$k = 1$-re igaz, mert ez volt a harmadik tulajdonság, amit teljesítenie kell az f koncentráltsági függvénynek.
Tegyük fel, hogy egy tetszőlegesen kiválasztott $k < s - 1$-re igaz. Lássuk be, hogy $k + 1$-re is igaz! Vagyis feltettük, hogy

$$f(n_1 + kh, n_2 - h, \ldots, n_{k+1} - h, n_{k+2}, \ldots, n_s) > f(n_1, \ldots, n_s).$$

Ha az egyenlőtlenség bal oldalán lévő kifejezés... tekintjuk egy kezdeti eloszlásnak, akkor ha erre alkalmazzuk a harmadik tulajdonságot, akkor a következőt kapjuk:

$$f(n_1+(k+1)h, n_2-h, \ldots, n_{k+2}-h, n_{k+3}, \ldots, n_s) > f(n_1+kh, n_2-h, \ldots, n_{k+1}-h, \ldots, n_s).$$

A tranzitivitás miatt pedig láthatjuk, hogy

$$f(n_1 + (k+1)h, n_2 - h, \ldots, n_{k+2} - h, n_{k+3}, \ldots, n_s) > f(n_1, \ldots, n_s).$$

Tehát ha feltesszük k-ra, akkor teljesül $k + 1$-re is.

2.2.1. Koncentráltsági mérőszámok

1. A korrigált **Berger–Parker-féle dominanciaindex**:

$$d(s, \mathbf{n}) = \frac{n_{\max}}{n} - \frac{1}{s},$$

ahol $n_{\max} := \max\{n_1, n_2, \ldots, n_s\}$. Ez gyakran használt index, annak ellenére, hogy egyedül a domináns gyakoriságra érzékeny, hiszen mindegy, hogy hány faj szerepel rajta kívül, csak az számít, hogy a domináns fajon kívüli egyedszám mennyi. Könnyű belátni, hogy erre az indexre teljesül a három alapkövetelmény.

   (a) A permutációinvariancia nyilvánvaló, hiszen az index csak a legnagyobb elemszámú fajtól függ. Mivel az indexet egy hányadosból

számoljuk, így a fajszámok konstansszorosára növelésével a konstans a számlálóban és a nevezőben egyaránt megjelenik, tehát az index skálainvariáns.

(b) A függvény értéke akkor lesz a legnagyobb ($\frac{s-1}{s}$), ha n_{max} értéke a lehető legnagyobb. Ez pedig akkor teljesül, ha egy kivételével az összes többi faj 0 egyedszámmal van jelen, ami éppen a monopóliumhelyzetet jelenti. Az egyenletes eloszlás esetében lesz rögzített n mellett n_{max} értéke a legkisebb. Ekkor a függvény értéke éppen 0. Az összes többi esetben a függvény értéke az előbbi két érték közt lesz.

(c) Ha a szegény ad a leggazdagabbnak h-t, akkor az index értéke $\frac{n_{\max}+h}{n}$, egyébként pedig nem változik.

2. A **Herfindahl-index**:

$$\mathrm{Herf}(s, \boldsymbol{\pi}) := \sum_{i=1}^{s} \left(\pi_i - \frac{1}{s}\right)^2 = \sum_{i=1}^{s} \pi_i^2 - \frac{1}{s} = \boldsymbol{\pi}^2 - \frac{1}{s}.$$

Ez a rögzített $s \in \mathbb{N}$ fajszámhoz tartozó $\boldsymbol{\pi} \in S^s$ eloszlásnak az $\frac{1}{s}\mathbf{1}^s$ egyenletes eloszlástól való eukleidészi távolsága. A második kifejezés – ami a formula eredeti alakja – egyezése az első formulával könnyen belátható, ha figyelembe vesszük, hogy $\sum_{i=1}^{s} \pi_i = 1$. Vizsgáljuk meg, hogy erre az indexre hogyan teljesülnek-e a feltételek.

(a) A permutációinvariancia itt is nyilvánvaló, mert az index a második változójában szimmetrikus. Mivel a függvény eloszlásra van definiálva, így a skálainvariancia itt úgy értendő, hogy tetszőleges $\mathbf{n} \in \mathbb{R}^{+s}$ esetére a következőképpen terjesztjük ki: $\mathrm{Herf}(s, \mathbf{n}) := \left(\frac{\mathbf{n}}{n}\right)^2 - \frac{1}{s}$. A képletből látszik, hogy a konstans szorzó egyaránt megjelenik a számlálóban és a nevezőben is, így lehet vele egyszerűsíteni.

(b) Látszik, hogy az egyenletes eloszlásnál a függvény értéke 0. (Az első formulában a szumma minden tagja 0.) Ez minimumhely, hiszen az index négyzetszámok összege, tehát nemnegatív értékű. A függvény értéke monopóliumon $1 - \frac{1}{s}$. Azt kell még belátnunk, hogy tetszőleges $\boldsymbol{\pi} \in S_s$ esetén

$$\sum_{i=1}^{s} \pi_i^2 - \frac{1}{s} \leqq 1 - \frac{1}{s}$$

$$\sum_{i=1}^{s} \pi_i^2 \leqq 1$$

$$\sum_{i=1}^{s} \pi_i^2 - \sum_{i=1}^{s} \pi_i \leqq 0$$

$$\sum_{i=1}^{s} \pi_i(\pi_i - 1) \leqq 0$$

Ez mindig igaz, mert az szummában lévő szorzat egyik tagja mindig negatív. Minden lépés megfordítható, így az állítást bizonyítottuk.

(c) Feltehető, hogy $n_1 = n_{\max}$, $n_1 \geqq n_2$. Ekkor a következő egyenlőtlenségnek kell teljesülnie a harmadik tulajdonság teljesüléséhez:

$$(\frac{n_1}{n})^2+(\frac{n_2}{n})^2+\sum_{i=3}^{s}(\frac{n_i}{n})^2-\frac{1}{s} \leqq (\frac{n_1+h}{n})^2+(\frac{n_2-h}{n})^2+\sum_{i=3}^{s}(\frac{n_i}{n})^2-\frac{1}{s}$$

A közös nevezővel való beszorzás, és az azonos tagok elhagyása után:

$$n_1^2 + n_2^2 \leqq (n_1 + h)^2 + (n_2 - h)^2$$

$$n_1^2 + n_2^2 \leqq n_1^2 + n_2^2 + 2h^2 + 2h(n_1 - n_2)$$

$$0 \leqq 2h^2 + 2h(n_1 - n_2).$$

Ez igaz, mert mivel feltettük, hogy $n_1 \geqq n_2$, így minden tag pozitív. Minden lépés megfordítható, így az állítást bizonyítottuk.

Nyilvánvaló az alábbi kijelentés.

5. Állítás. *Ha f koncentráltsági index, $\varphi : \mathbb{R} \to \mathbb{R}$ pedig olyan monoton növő függvény, amelyre $\varphi(0) = 0$, akkor $\varphi \circ f$ is koncentráltsági index, ugyanis a koncentráltsági index minden tulajdonsága egyenlőtlenséggel van definiálva.*

3. Megjegyzés. *Ha azonban nem csak egyenlőtlenségi feltételeket, hanem egyenlőségi feltétleket is kirovunk (függvényegyenletek formájában), akkor például egyértelműen megkaphatjuk az entrópiát.(lásd pl.[11])*

3. A **szóráshányados-index**:

$$\mathrm{V}(s, \boldsymbol{\pi}) := \sqrt{\frac{\sum_{i=1}^{s} \left(\pi_i - \frac{1}{s}\right)^2}{s}} \Big/ \frac{1}{s} = \sqrt{s \sum_{i=1}^{s} \pi_i^2 - 1}$$

Erre az indexre is teljesül mind a három feltétel, hiszen ez az index a **Herfindahl-index**nek monoton növő függvénye: $\varphi(z) = \frac{1}{s}\sqrt{\frac{z}{s}}$.

2.3. Kapcsolatok

A definiált mérőszámcsaládokkal és relációval kapcsolatban természetes módon merül föl egy sor kérdés. A kérdések egy részére még nincsenek válaszok, csupán azért vannak megemlítve itt, hogy egyrészről látható legyen a megismert indexcsoportok szerepe a matematikában, másrészről hátha valaki kedvet érez ezen elgondolkodni. Mindenekelőtt bevezetünk még egy további fogalmat, egy parciális rendezést az eloszlások halmazán ([1, 12]). Legyen $s \in \mathbb{N}; \boldsymbol{\pi}, \boldsymbol{\varrho} \in S^s$.

1. Definícó. *Azt mondjuk, hogy a $\boldsymbol{\pi}$ eloszlás **kevertebb** vagy **sztochasztikusan nagyobb**, mint a $\boldsymbol{\varrho}$ eloszlás, és azt írjuk, hogy $\boldsymbol{\pi} \prec_s \boldsymbol{\varrho}$, ha minden $k \in \{1, 2, \ldots, s\}$ esetén $\sum_{i=1}^{k} \pi_i^{\downarrow} \leq \sum_{i=1}^{k} \varrho_i^{\downarrow}$. Ez a reláció nyilván parciális rendezés az S^s halmazon, és $\forall s \in \mathbb{N} \, \forall \boldsymbol{\pi} \in S^s$ mellett $\frac{1}{s}\mathbf{1}^s \prec \boldsymbol{\pi} \prec \mathbf{e}_1^s$.*

Az alábbi programrészlet pontosan akkor ad `True` értéket, ha $\boldsymbol{\pi} \prec \boldsymbol{\varrho}$.

```
Apply[And,(FoldList[Plus, 0, Sort[pi, #2<#1&]]< FoldList[Plus,
0, Sort[rho, #2<#1&]])]
```

1. Igaz-e, hogy kevertebb eloszlás diverzitása nagyobb, koncentráltsága kisebb?

   Legyen $\boldsymbol{\pi}$ kevertebb eloszlás, mint $\boldsymbol{\varrho}$. A q-tól kumulált index azt jelenti, hogy egy bizonyos q indextől kezdve összeadjuk a valószínűségeket, tehát a $\boldsymbol{\pi}$ eloszlás q-tól kumulált indexe nagyobb, mint a mint $\boldsymbol{\varrho}$ eloszlásé, azaz ebben az értelemben $\boldsymbol{\pi}$ diverzitása nagyobb, mint $\boldsymbol{\varrho}$ diverzitása.

2. Mely esetben lesz egy diverzitási (koncentráltsági)index monoton csökkenő függvénye koncentráltsági (diverzitási) index?

3. Lehet-e egy diverzitási indexet Ljapunov-függvényként értelmezni, illetve Ljapunov-függvényből képezhető-e diverzitási index? Időtöl függő determinisztikus és sztochasztikus modellekben is fontos szerepet játszik az entrópia: a modellek egy részében csökken, minimumát az egyensúlyban veszi fel. Ez nyilván azt jelenti, hogy a modellek egy részében a diverzitás nő, az egyensúlyban veszi fel a maximumát. Fölmerül a kérdés, hogy az itt szereplő indexek is beválnak-e Ljapunov-függvényként, illetve van-e olyan Ljapunov-függvény, amelyikből további koncentráltsági és diverzitási index képezhető.

4. Mi a kapcsolat a diverzitási és koncentráltsági indexek, valamint a valószínűségeloszlások közötti távolságok között?

   Természetes módon merül fel a kérdés, hogy az eloszlások közötti távolságokból [Andai, 26.oldal] hasznos indexeket lehet-e definiálni a következő módon: valamely távolságban az egyik eloszlást rögzítjük; legyen az az egyenletes eloszlás vagy a monopólium. Ezek után vizsgáljuk az eloszlások ettől való távolságát! Definiálnak ezek a távolságok diverzitási, illetve koncentráltsági indexeket? A továbbiakban már csak a

számolás eredményeként kapott képleteket írjuk le, majd megpróbálunk párhuzamot vonni a már meglévő indexek és az itt adódott távolságok között.

(a) A D_{KL} **Kullback–Liebler-távolság** esetén a következőket kapjuk:

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}, \mathbf{1}^s) = \sum_{i=1}^{s} \pi_i \log(\pi_i) + \log(s) \tag{2.3}$$

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{1}^s, \boldsymbol{\pi}) = -\log(s) - \frac{1}{s} \sum_{i=1}^{s} \log(\pi_i) \tag{2.4}$$

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{e}_1, \boldsymbol{\pi}) = \boldsymbol{\pi}_1^{\downarrow} \log \boldsymbol{\pi}_1^{\downarrow} \tag{2.5}$$

Ezek közül az első a **Shannon-index** (-1)-szerese.

(b) A D_{H} **Hellinger-távolság** az alábbiakat szolgáltatja.

$$\mathrm{D}_{\mathrm{H}}(\boldsymbol{\pi}, \mathbf{1}^s) = 1 - \frac{1}{s} - \frac{2}{s} \sum_{i=1}^{s} \sqrt{\pi} \tag{2.6}$$

$$\mathrm{D}_{\mathrm{H}}(\boldsymbol{\pi}, \mathbf{e}_1) = (\sqrt{\boldsymbol{\pi}_1^{\downarrow}} - 1)^2 \tag{2.7}$$

(c) A D_{χ^2} χ^2-távolságból a következők adódnak:

$$\mathrm{D}_{\chi^2}(\boldsymbol{\pi}, \mathbf{1}^s) = s^2 \sum_{i=1}^{s} \pi_i^3 - 1 \tag{2.8}$$

$$\mathrm{D}_{\chi^2}(\mathbf{1}^s, \boldsymbol{\pi}) = \frac{1}{s^2} \sum_{i=1}^{s} \frac{1}{\pi_i^2} - 1 \tag{2.9}$$

$$\mathrm{D}_{\chi^2}(\mathbf{e}_1, \boldsymbol{\pi}) = \frac{1}{(\boldsymbol{\pi}_1^{\downarrow})^2} - 1 \tag{2.10}$$

A (2.10) távolság $\boldsymbol{\pi}_1^{\downarrow}$-nek monoton függvénye, így lényegében azonos a **Berger–Parker-index**szel. Míg (2.5) és (2.7) nem monoton függvénye $\boldsymbol{\pi}_1^{\downarrow}$-nek, így nem is hozható kapcsolatba a **Berger–Parker-index**szel.

5. Hogyan általánosíthatók a definiált fogalmak eloszlások helyett 1 nyomú önadjungált pozitív definit mátrixokra?

   Neumann János az ilyen $\mathbf{D}$ mátrixok entrópiájául a $\mathrm{Tr}(\mathbf{D}\log(\mathbf{D}))$ kifejezést javasolta. Ennek mintájára érdemes lehet bevezetni a **Gini–Simpson-index**et az $1-\mathrm{Tr}(\mathbf{D}^2)$ formulával, vagy a **Herfindahl-index**et a $\mathrm{Tr}(\mathbf{D}^2) - \frac{1}{s}$ képlettel. Formailag definiálható a **szóráshányados-index** is: $\sqrt{s\mathrm{Tr}(\mathbf{D}^2) - 1}$. Érdekes lenne azt is megvizsgálni, hogy hogyan kell a diverzitási és koncentráltsági indexekre vonatkozó általános kvalitatív kritériumokat megfogalmazni. Ezzel összefüggésben az is megvizsgálandó, hogy az így definiált indexek milyen $\mathbf{D}$ mátrix mellett veszik fel a szélsőértéküket.

2.3.1. Összefüggések a két indexcsoport között és azokon belül

Koncentráltsági indexek között:

	Herfindahl	Szóráshányados
Berger–Parker	1	2
Herfindahl	-	3

1. Rendezett minta esetén a Berger–Parker-index a Herfindahl-index első tagjának monoton növő függvénye.

2. Explicit összefüggés a Herfindahl-indexen keresztül van.

3. A két index közötti monoton explicit összefüggés:

$$\textbf{szóráshányados-index} = \sqrt{\textbf{Herfindahl-index} \times \textit{fajszám}}$$

Diverzitási indexek között:

	Redukált fajszám	GS	H	H_B
Fajok száma	1	2	2	2
Redukált fajszám	-	2	2	2
GS	-	-	3	4
H	-	-	-	5

1. **Redukált fajszám = Fajok száma** $- 1$

2. A Fajok száma a többi indexnél már csak a szummában jelenik meg, úgy mint az összeadandó tagok száma.

3. A Shannon-indexet első tagig sorbafejtve a Gini–Simpson-index mínusz egyszeresét kapjuk (egy additív konstanstól eltekintve).

4. Nincs közvetlen explicit összefüggés közöttük, csak a Shannon-indexen keresztül.

5. A Brillouin-index a Shannon-index „véges megfelelője”. Vagyis míg az előbbiben véges egyedszámokkal számolunk, addig az utóbbiban csak az eloszlást ismerjük, az egyedek számát nem [5].

Koncentráltsági és diverzitási indexek között:

	GS	H	H_B
Berger–Parker	1	2	3
Herfindahl	4	5	6
Szóráshányados	7	8	9

1. Összefüggés a Herfindahl-indexen keresztül.

2. Összefüggés a Herfindahl-indexen keresztül.

3. Összefüggés a Herfindahl- és Shannon-indexeken keresztül.

4. Explicit összefüggés:

$$\textbf{Herfindahl-index} = -\textbf{GS} + 1 - \frac{1}{s}$$

5. A Shannon-indexet első tagig sorbafejtve a Herfindahl-index mínusz egyszeresét kapjuk (egy additív konstanstól eltekintve).

6. Nincs explicit összefüggés.

7. Explicit összefüggés:

$$\textbf{szóráshányados-index} = \sqrt{-\textbf{GS} \times s + s - 1}$$

8. A Shannon-indexet első tagig sorbafejtve a szóráshányados monoton csökkenő függvényét kapjuk.

9. Nincs explicit összefüggés.

3. fejezet

A Pareto-elv

A Pareto-elv, vagy másnéven a 80-20-as szabály [2, 10] bizonyos értelemben egy monopóliumhoz közeli eloszlást ír le, vagyis a Pareto-elvnek eleget tevő eloszlás koncentráltsága közel van a monopóliuméhoz (ahol maximális). Ismert diszkrét és folytonos eloszlásokat fogunk vizsgálni, abból a szempontból, hogy milyen feltételek mellett, vagyis a milyen paraméter értékekkel teljesítik a 80-20-as szabályt. Továbbá megpróbáljuk a szabályt általánosítani a Pareto-elvben szereplő 80-20 helyett p-q önkényesen választott paraméterekre.

3.1. Pareto-elv

2. Definícó. *A Pareto-elv: gyakran bekövetkező eseményeknél fordul elő, hogy egy valószínűségi változó várható értékének 80%-a előáll a lehetséges értékeinek csak mintegy 20%-ából. Azaz,ha*

$$W(x) := \frac{\int_x^{+\infty} x' p(x') dx'}{\int_{x_{\min}}^{+\infty} x' p(x') dx'},$$

akkor $W(x_{0.2}) = 0.8$, *ahol* $x_{\min}$ *a vizsgált* $W(x)$ *minimális értéke,* $x_{0.2}$ *pedig a 20%-os alsó kvantilis:* $\int_{x_{0.2}}^{+\infty} p(x')dx' = 0.2$.

A Pareto-elvnek eleget tevő eloszlások bizonyos értelemben koncentráltabbak: bármelyik koncentráltsági mértékkel mérve olyan értéket kapunk, amely közelebb van a monopóliuméhoz.

Vizsgáljuk innentől, hogy milyen eloszláscsaládokra, és milyen paraméterekre igaz ez az elv!

Hatványeloszlás

Ennek az eloszlásnak leginkább a gyakorlati szerepe jelentős, hiszen a természetben előforduló jelenségeknek jelentős hányada követ hatványeloszlást. Ilyenek például: a földrengések nagysága, az internetes oldalak nézettsége, a városok lakosságának eloszlása, vezetéknevek eloszlása, a Hold-kráterek átmérőjeinek nagysága. Az, hogy például a vezetéknevek eloszlása hatványeloszlás, azt jelenti, hogy van néhány rendkívül gyakori vezetéknév, a legtöbb vezetéknév viszont elég ritka.

A hatványeloszlás sűrűségfüggvénye:

$$p(x) = Cx^{-\alpha},$$

ahol α az eloszlás paramétere, C pedig α-tól függő normáló tényező. Eloszlásfüggvényének komplementere pedig

$$P(x) = \int_x^{+\infty} p(x')dx' = \frac{C}{\alpha - 1} x^{-\alpha+1} = (\frac{x}{x_{\min}})^{-\alpha+1}, \tag{3.1}$$

ahol C értékét a következő módon kapjuk meg:

$$1 = \int_{x_{\min}}^{+\infty} p(x)dx = C \int_{x_{\min}}^{+\infty} x^{-\alpha}dx = -[\frac{Cx^{1-\alpha}}{1-\alpha}]_{-\infty}^{x}.$$

Ha $\alpha > 1$, akkor:

$$C = (\alpha - 1)x_{\min}^{\alpha-1}.$$

Ha $\alpha > 1$, akkor a medián, jelöljük ezentúl $x_{1/2}$-vel, egyértelműen meghatározható. A medián az értelmezési taromány azon pontja, amelyre igaz a

következő:

$$\int_{x_{1/2}}^{+\infty} p(x)dx = \frac{1}{2}\int_{x_{\min}}^{+\infty} p(x)dx,$$

vagyis

$$x_{1/2} = 2^{1/(\alpha-1)}x_{\min}.$$

Ha például azt vizsgáljuk, hogy hogyan oszlik meg a vagyon az emberek közt, akkor a medián elválasztja a társadalom gazdag rétegét a szegény rétegtől. Nézzük meg a gazdagabb réteg vagyona várható értékének arányát az összvagyon várható értékéhez képest:

$$\frac{\int_{x_{1/2}}^{+\infty} xp(x)dx}{\int_{x_{\min}}^{+\infty} xp(x)dx} = (\frac{x_{1/2}}{x_{\min}})^{-\alpha+2} = 2^{-(\alpha-2)/(\alpha-1)},$$

ha $\alpha > 2$ akkor mindkét integrál konvergál.

Általában, ha egy eloszlás eloszlásfüggvényének komplementere a (3.1)-ben meghatározott $P(x)$, akkor

$$W(x) = \frac{\int_{x}^{+\infty} x'p(x')dx'}{\int_{x_{\min}}^{+\infty} x'p(x')dx'} = (\frac{x}{x_{\min}})^{-\alpha+2}$$

és ha $\alpha > 2$, akkor

$$W = P^{(\alpha-2)/(\alpha-1)}$$

Lineáris sűrűségfüggvényű eloszlás

Ebben a részben olyan eloszlásokat vizsgálunk, amelyeknek lineáris a sűrűségfüggvénye. Nyilvánvalóan ezek olyan nemnegatív függvények lesznek, amik monoton növők vagy csökkenők. Egy lineáris függvény $f(x) = \alpha x + \beta$ alakban írható fel, ahol $\alpha, \beta \in \mathbb{R}$. Ha α pozitív, akkor a függvény szigorúan monoton nő, ha pedig negatív, akkor szigorúan monoton csökken. Legyen az f sűrűségfüggvényünk értelmezési tartománya az $[0, 1]$ intervallum, értékkészlete pedig $\mathbb{R}_0^+$. Keressük azt az $x_{0.8} \in [0, 1]$ (monoton csökkenő függvény esetén $x_{0.2} \in [0, 1]$) pontot, és azt az f sűrűségfüggvényt amire teljesülnek a következő feltételek:

1.

$$\int_{x_{0.8}}^{1} \alpha x + \beta \quad dx = 0.2, \tag{3.2}$$

(illetve monoton fogyó függvén esetén:

$$\int_{0}^{x_{0.2}} \alpha x + \beta \quad dx = 0.2,)$$

2. Továbbá az f sűrűségfüggvényre teljesülnek:

   (a) Az integrálja a $[0, 1]$ intervallumon 1

   $$\int_{0}^{1} f(x) \quad dx = \int_{0}^{1} \alpha x + \beta \quad dx = 1 \tag{3.3}$$

   (b) A keresett $[0.8, 1]$ (ill.$[0, 0.2]$) intervallumon a várható értékek aránya 0.8.

   $$\frac{\int_{x_{0.8}}^{1} x(\alpha x + \beta) \quad dx}{\int_{0}^{1} x(\alpha x + \beta) \quad dx} = 0.8 \tag{3.4}$$

   (illetve

   $$\frac{\int_{0}^{x_{0.2}} x(\alpha x + \beta) \quad dx}{\int_{0}^{1} x(\alpha x + \beta) \quad dx} = 0.8)$$

Ezek a Pareto-elv teljesülésének definíció szerinti feltételei. A függelékben található **Mathematica** programból [13] látszik, hogy egyetlen olyan $x_{0.8}$ pontot találtunk, amire igaz az (3.2) feltétel, a (3.3) kritériumnak eleget tevő α mellett, viszon erre az $x_{0.8}$ pontra nem teljesül (3.4). Konstans sűrűségfüggvény esetén az (3.4) egyenlet bal oldalán szereplő hányados negatív értéket vesz fel. Tehát lineáris sűrűségfüggvények nem teljesítik a Pareto-elvet.

4. fejezet

Alkalmazások

4.1. Az alapul vett epidemiológiai adatbázis leírása

A számításokban és a példákban konkrét, valóságos statisztikai adatokkal dolgoztunk (v.ö. [7]). 2005 tavaszán Kis Ildikó (e-mail: ildiko.kis@office.ksh.hu) rendelkezésemre bocsátott egy közel ötven táblázatból álló, a Központi Statisztikai Hivatal által készített adatállományt. A táblázatok epidemiológiai adatokat tartalmaznak különböző tulajdonságok szerint vizsgálva; alapvetően két fő szempont alapján: 2003-as évi adatok egy adott fertőző betegségben szenvedők számáról a 20 fő területi egység között elosztva (a 19 megye és Budapest), illetve egy adott betegségben szenvedők számának változása az idő függvényében (a szereplő évek 1970, 1980, 1990, 2000, 2002 és 2003). A programok, a részletes számolási eredmények és ábrák a Függelékben találhatók; itt csak néhány fontos, illetve jellegzetes eredményt és ábrát emelünk ki.

4.2. Diverzitás és koncentráltság epidemiológiai adatoknál

Először a Központ Statisztikai Hivatal által küldött *Bejelentett fertőző betegségek száma terület szerint* (2003) című táblázatot kellett a **Mathematica**nak úgy átadni, hogy egy olyan mátrix keletkezzen, amelynek az első oszlopa a területi egységek (megyék) nevét tartalmazza, az első oszlopa pedig a betegségek nevét, amik most csak számok 1-től 22-ig. A mátrix elemei számok, a következőképpen definiálva: $a_{i,j}$ az i-edik megyében a j-edik betegségben szenvedő regisztrált betegek száma.

4.2.1. Koncentráltság

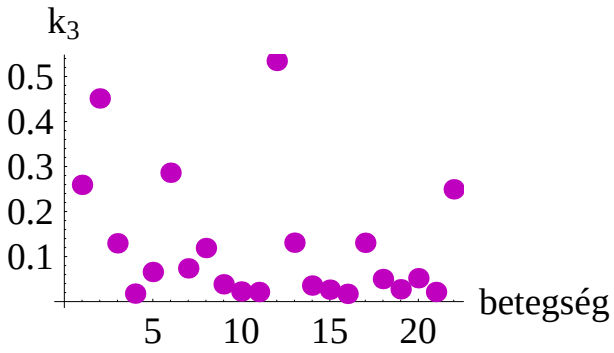
A betegségek területi eloszlásainak koncentráltság szerinti tulajdonságait fogjuk ebben a részben vizsgálni. A vizsgálatban használt koncentráltsági indexek a már korábban definiált függvények, a következők:

1. Korrigált **Berger–Parker-féle dominanciaindex**,
2. **Herfindahl-index**,
3. **szóráshányados-index**.

Még mielőtt lefuttatnánk a programot, vagyis alkalmaznánk ezeket a függvényeket a táblázatra, lehet következtetni az indexeknek az implicit alakjából is arra, hogy melyik milyen tulajdonságú. A **Berger–Parker-féle dominanciaindex**nél csak a legnagyobb relatív gyakoriság és a fajok száma számít. Ez az index nem különböztet meg két olyan populációt, amelyekben e kettő azonos, de a többi relatív gyakoriság különbözik bennük. A másik két index, ahogy azt már korábban láthattuk, egymásnak monoton függvénye. Így még a tényleges alkalmazás előtt gondolhatjuk, hogy a második két index értékeit érdemes jobban figyelni. Viszont az is igaz, hogy e két index

által felvett értékek valószínűleg közel ugyanúgy fognak viselkedni a táblázat oszlopain.

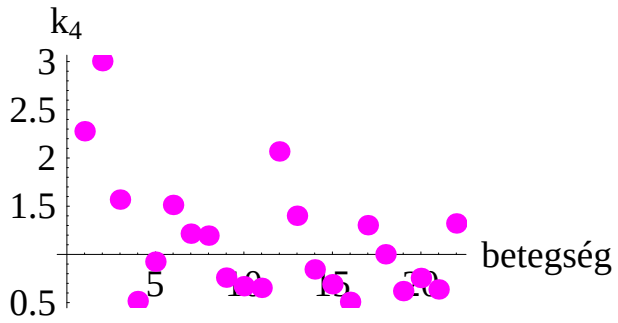
A program lefuttatása után a második és harmadik index szerint a betegségek két, egymástól elkülöníthető csoportra oszthatók. Néhány betegség koncentráltan jelenik meg a területeken, míg a többség inkább szétoszlik a megyékben. (Megjegyezzük, hogy az alábbi ábrák szemléltető jellegűek, mivel az ábrán szereplő formális betegség↦index függvényeknek nem értelmesek gyakorlati szempontból. Az EXCEL program szokásos oszlopdiagrammjaihoz hasonlítanak.) A **Herfindahl-index** értékei:



4.1. ábra. A Herfindahl-index értékei

A **szóráshányados-index** értékei:
A **Herfindahl-index** szerinti öt legkoncentráltabb betegség a következő:

- Hepatitis infectiosa,
- Hepatitis A,
- AIDS,
- Keratoconjunctivitis epidemica (Járványos kötőhártya-gyulladás),
- Halálos kimenetelű nosocomialis sepsis.



4.2. ábra. A szóráshányados-index értékei

A **szóráshányados-index** szerinti öt legkoncentráltabb betegség a következő:

- Hepatitis infectiosa,
- Hepatitis A,
- Hepatitis B,
- AIDS,
- Keratoconjunctivitis epidemica (Járványos kötőhártya-gyulladás).

Egy betegség eltérésével ugyanaz az eredmény mindkét index szerint.

4.2.2. Diverzitás

Most ugyanazt az adatsort fogjuk elemezni dichotom diverzitási indexekkel. Az indexek az alkalmazásuk sorrendjében a következők:

1. **fajok száma**,
2. **redukált fajszám**,

3. **Gini–Simpson-index**,

4. **Shannon-index**,

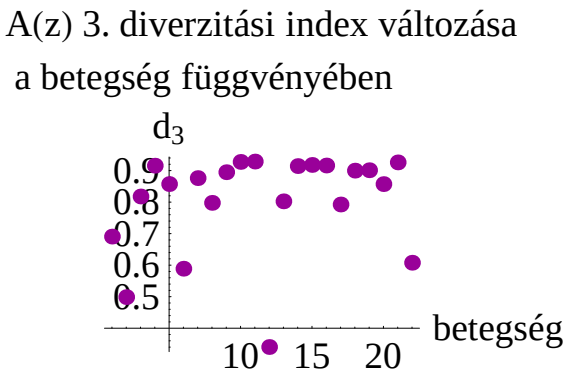
5. **Brillouin-index**.

A koncentráltsági indexekhez hasonlóan diverzitási indexeknél is már a képletből lehet látni, hogy melyik függvénynek mi a jelentősége. Mivel az első két index teljesen érzéketlen az eloszlásra, vagyis a relatív gyakoriságokra, így azok csak említés szintjén jelennek meg ebben a részben. A **Gini–Simpson-** és a **Shannon-index** közös tulajdonsága, hogy mindkettőben csupán a relatív gyakoriságokat kell ismerni, vagyis a populáció egyedszámára nincs szükség a mérőszám kiszámításához. Ezzel szemben a **Brillouin-index**ben relatív gyakoriságok helyett egyedszámokkal számolunk. Ezt a függvényt, ami az előzőhöz hasonlóan egy entrópia, a szakirodalom gyakran a **Shannon-index** véges megfelelőjének nevezi, még pedig éppen az előbb említett különbség miatt. Ha lefuttatjuk a programot a táblázat oszlopaira, akkor valóban láthatjuk, hogy az első két index nem mond el túl sokat a betegségek területi diverzitásáról, csupán azt, hogy hány megyében fordulnak elő. A többi függvénynél viszont, ahogy a koncentráltságoknál is, szét lehet bontani a betegségeket diverz és kevésbé diverz csoportra. Ha úgy gondoljuk, hogy kézzelfogható az ellentét a koncentráltság és a diverzitás jelentésének tartalma között, akkor igaznak kell lennie annak, hogy az öt legkoncentráltabb betegség egyben az öt legkevésbé diverz betegség is. Némely diverzitási indexnél ez teljesül is a táblázatra, viszont nem jellemző, hogy az indexek pontosan ugyanazt az öt betegséget választják ki.

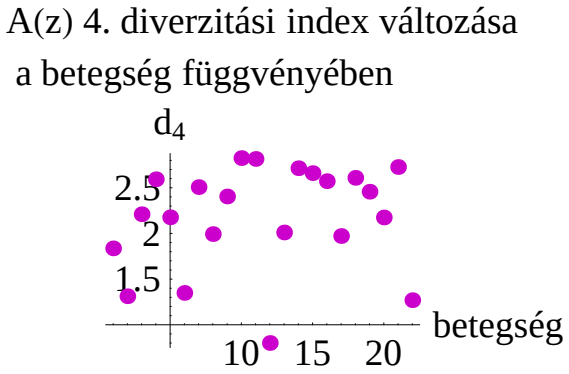
Nézzük meg e három index szerint legkevésbé diverz betegségeket, majd vessük össze a Herfindahl- és a szóráshányados-index szerint legkoncentráltabb betegségekkel!

A **Gini–Simpson-index** szerinti öt legkevésbé diverz betegség:

- Hepatitis infectiosa,



4.3. ábra. A Gini–Simpson-index értékei

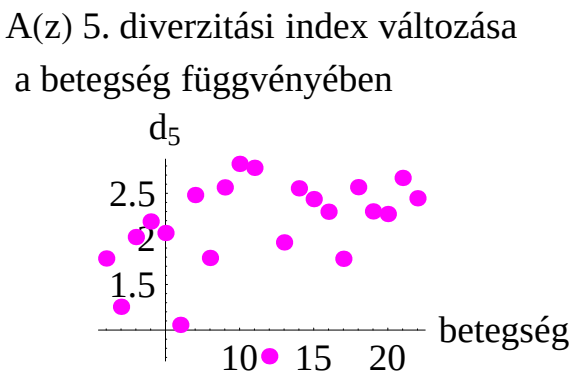


4.4. ábra. A Shannon-index értékei

- Hepatitis A,
- AIDS,
- Keratoconjunctivitis epidemica (Járványos kötőhártya-gyulladás),
- Halálos kimenetelű nosocomialis sepsis.

A **Shannon-index** szerinti öt legkevésbé diverz betegség:

- Hepatitis infectiosa,



4.5. ábra. A Brillouin-index értékei

- Hepatitis A,
- AIDS,
- Keratoconjunctivitis epidemica (Járványos kötőhártya-gyulladás),
- Halálos kimenetelű nosocomialis sepsis.

A **Brillouin-index** szerinti öt legkevésbé diverz betegség:

- Hepatitis infectiosa,
- Hepatitis A,
- AIDS,
- Keratoconjunctivitis epidemica (Járványos kötőhártya-gyulladás),
- Encephalitisinfectiosa.

A számítási eredményekből is jól látszik, hogy a legkoncentráltabb betegségek csoportja legfeljebb egy betegségben különbözik a legkevésbé diverz betegségek csoportjától. Ez is azt bizonyítja, hogy az az elméleti sejtés, miszerint a koncentráltsági és a diverzitási indexek közötti különbség nem több, mint azonos típusú indexek közti eltérés, a gyakorlatban is igazolódni látszik.

4.3. A koncentráltsági és diverzitási indexek időfüggése

A bejelentett fertőző megbetegedések száma és aránya című táblázat adataira alaklamaztuk a fent bevezetett koncentráltsági és diverzitási indexeket, így képet kaptunk azok időbeli változásáról. Eszerint megállapíthatjuk, hogy a különféle koncentráltsági indexek is hasonlóan változnak, és a különböző diverzitási indexek is hasonlóan változnak.

Ezek az adatok azt mutajták, hogy az utóbbi időben Magyarországon a fertőző betegségek koncentráltsága nő, ennek megfelelően diverzitásuk csökken, bármelyik mérőszámot használjuk is.

A részletes számolások és ábrák a 2. Mellékletben láthatók.

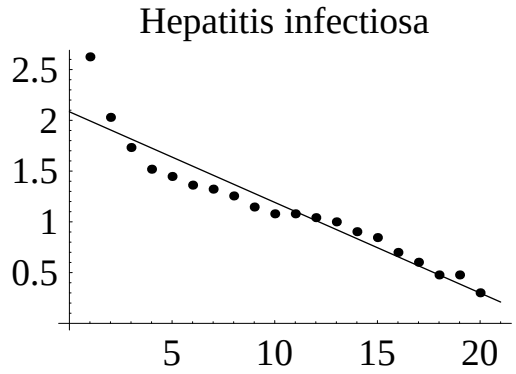
4.4. A Pareto-elv érvényesülése epidemiológiai adatoknál

A Pareto-elv teljesülésének a feltételeit ellenőrizni tudjuk egy adott mintán, ha meg tudjuk állapítani, hogy milyen eloszlásból származik a minta, és becslést tudunk adni az eloszlás paramétereire. Itt is a *Bejelentett fertőző betegségek száma terület szerint* (2003) táblázatot fogjuk elemezni az egyes betegségek területi eloszlása szerint. Először azt a sejtésünket igazoljuk, hogy a betegségek eloszlása hatványelszlás. Mint azt már láttuk az előző fejezetben, a hatványeloszlásokkal könnyű dolgozni, mert általánosságban kiszámítható, hogy milyen paraméterek mellett teljesítik a Pareto-elvet.

Az elemző program úgy működik, hogy veszünk egy betegséget, és az egyes területeken bejelentett fertőzöttek számát először csökkenő sorrendbe rendezzük, majd így ábrázoljuk egy logaritmikus koordináta-rendszerben. Erre a pontsorozatra illesztünk egy egyenest, ez lesz a regressziós egyenes. Majd egy beépített függvény segítségével megállapítjuk, hogy a minta illeszkedike az egyenesre. A pontokhoz illesztett regressziós egyenes illeszkedését a

varianciaanalízisen alapuló F-próbával vizsgálva azt találtuk, hogy 95%-os szinten az illeszkedés elfogadható. Ez minden esetben teljesül, vagyis mindegyik pontsorozat közelithető egyenessel. Ami azt jelenti, hogy eloszlásaik megfelelnek hatványeloszlásoknak [3, 4].

```
elemzes[2]
```



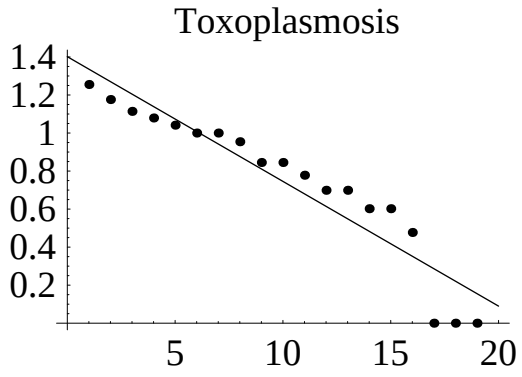

```
                          Estimate        SE             TStat         PValue
{ParameterTable → 1       2.0842          0.0869171      23.9792       0      ,
                  x       -0.0892459      0.0072557      -12.3001      0

 RSquared → 0.893675, AdjustedRSquared → 0.887768, EstimatedVariance → 0.0350091,

                          DF      SumOfSq       MeanSq         FRatio       PValue
 ANOVATable → Model       1       5.29661       5.29661        151.293      0
              Error       18      0.630163      0.0350091                          }
              Total       19      5.92677
```

4.6. ábra. Egy példa a program outputjára

Ezen az ábrán látható, hogy első közetítésként elfogadható a hatványfüggvény hipotézis, de valószínűnek látszik, hogy egy általánosabb függvény családdal való illesztés még pontosabbnak bizonyulhat. Ilyen általánosabb családot alkotnak a Zipf–Mandelbrot eloszlások [8]: $F(i) = p_i = \frac{a}{(b+i)^c}$, ahol a és c az eloszlás paraméterei, b pedig konstans.

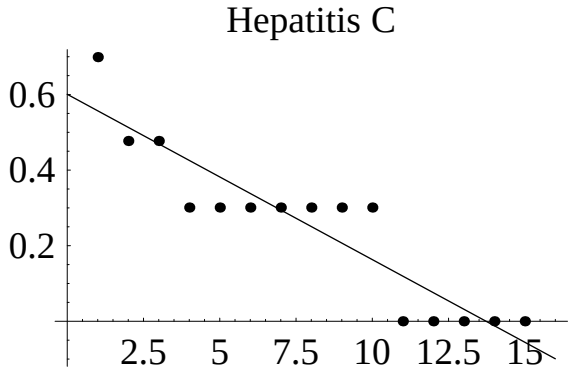
A következő ábrán egy láthatóan szép illeszkedés szerepel, ami főleg azért jött létre, mert elég sok adattal dolgozott a program. Ez jelen esetben azt jelenti, hogy minden megyében jelentős esetszámot regisztráltak. Ilyenek

például azok a fertőző gyerekbetegségek, amelyek ellen nem adnak védőoltást.



4.7. ábra. Látható az illeszkedés

Ezen az ábrán pedig olyan illeszkedés látható, ahol a betegség nem az egész országban elterjedt, sőt ahol előfordul, ott is viszonyleg kevés a fertőzöttek száma. Ilyenek például a vér útján terjedő betegségek.



4.8. ábra. Kevés adat miatt kevésbé illeszkednek a pontok

Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, Izsák János egyetemi tanárnak, konzulensemnek, Tóth János egyetemi docensnek folyamatos segítségükért és türelmükért. Továbbá köszönöm Kis Ildikónak az adatokhoz való hozzájutás lehetőségét. A dolgozat a T047132 számú OTKA részbeni támogatásával készült. [1]

[1] És köszönöm mindenkinek, akinek e és a betű szerepel a nevében.

Irodalomjegyzék

[1] Andai, A.: *Információgeometria a kvantummechanikában*, (Ph D értekezés), BME, Budapest, 2003.

[2] Arnold, B. C.: *Pareto Distributions*, International Co-operative Publishinghouse, USA, 1983.

[3] Bolla, M. – Krámli, A.: *Statisztikai következtetések elmélete*, Typotex Kiadó, Budapest, 2005.

[4] Ezekiel, M. – Fox, M. A.: *Korreláció- és regresszió-analízis*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1970.

[5] Izsák, J.: *Bevezetés a biológiai diverzitás mérésének módszertanába*, Scientia Kiadó, Budapest, 2001.

[6] Izsák, J.: Sensitivity Profiles of Diversity Indices, *Biometric J.* **38** (1996) 921–930.

[7] Izsák, J.: A pilot study on the frequency structure of histological neoplasm diagnosis in rats, *J. theor. Biol.* **236** (2005) 427–437.

[8] Izsák, J.: Some practical aspects of fitting and testing the Zipf–Mandelbrot model. A short essay, *Scientiometrics* **67** (2006) 107–120.

[9] Izsák, J. – Papp, L.: On diversity and concentration indices in ecology, *Coenoses* **13** (1) (1998) 29–32.

[10] Newman, M. E. J.: Power laws, Pareto distributions and Zipf's law, http://arXiv:cond-mat/0412004 v2 9 Jan 2005

[11] Ohya, M., Petz, D.: *Quantum Entropy and its Use*, Springer, Berlin, 2004

[12] Stoyan, D.: *Comparison methods for queues and other stochastic models*, John Wiley and Sons, Chicester, New York, Brisbane, Toronto, Singapore, 1983.

[13] Szili, L. – Tóth, J.: *Matematika és Mathematica*, ELTE Eötvös Kiadó, Budapest, 1996.

[14] Tóthmérész, B.: *Diverzitási rendezések*, Scientia Kiadó, Budapest, 1997.