Bibliometriai eloszlási függvények

Abraham BOOKSTEIN

A szerző The bibliometric distribution c. cikkét (Library, Quarterly, vol.46. 4.no. 416-423.p.) BOBOKNÉ BELÁNYI Beáta tömöritette.

A bibliometriában számos eloszlási törvény ismeretes és ezek egy része egymáshoz nagyon hasonló, más esetekben a hasonlóság matematikai átalakitások után mutatható ki. A különböző empirikus függvények gondos vizsgálatával megállapitható, hogy valamennyi eloszlási függvény egy és azonos általános eloszlási függvénynek adott konkrét feltételek közötti megoldása.

Az eloszlási függvények azért különböznek egymástól, mert a jelenséget más-más oldalról közelitik meg. Amikor pl. arról van szó, hogy egy témakör cikkei hogyan oszlanak meg a folyóiratokban, az eloszlási vizsgálat a megjelenő összes cikk ill. a sok cikket tartalmazó folyóiratok felől közelit. Ilyen esetekben kumulált függvényeket alkalmaznak. Ha pedig a szerzők publikációs tevékenységét vizsgáljuk, ott éppen a sok cikket publikálók összeszámlálása okoz problémát, ezért a nem-kumulativ formát szokás használni.

Célunk az alapvető fontosságu eloszlások kiválasztása és egymáshoz való viszonyuk vizsgálata.

Lotka az elsők között vizsgálta a kémikusok és az általuk publikált irodalom közti megoszlást és ugy találta, hogy ez minden esetben arányos $1/n^2$-tel. Törvénye - amely a nevét viseli -

$$N = A \frac{1}{n} 2 .$$

ahol N a kémikusok (vegyészek) száma, akik n számu cikket publikálnak, az A csupán egy arányossági tényező.

Később (1949-ben) Zipf arra a megállapitásra jutott, hogy ha a szavakat előfordulási gyakoriságuk szerint sorba rakjuk, akkor egy-egy szó előfordulásának gyakoriságát a következő képlet adja meg:

$$N = \frac{A}{r} ,$$

ahol r a gyakoriság szerint rendezett sorban elfoglalt helyet (rangot) jelenti, N pedig az r. szó előfordulása, az A itt is arányossági tényező. Ezt a törvényt gyakran kumulált formájában adja meg:

$$B = \sum_{n=1}^{r} 1/n$$

B az összes előfordulás az r rangig.

A kumulálást jól közeliti az integrál, amely analitikus formában adja meg a függvényt:

$$\int_{1}^{r} \frac{1}{x} \, dx = 1 \, nr$$

ez átalakitás után a következő alakra hozható:

$$\int_{1/2}^{r+1/2} \frac{1}{x} \, dx = 1 \, n \, (r+1/2) + \ln (2) = \ln (1+2r)$$

Ha ezt az utóbbi kifejezést kisérjük figyelemmel, igen érdekes megállapitásra jutunk a harmadik, a Bradford-féle eloszlási törvénnyel való összehasonlitásakor. Bradford azt állitja, hogy ha a folyóiratokat a bennük található egy adott témával foglalkozó cikkek száma szerint rendezzük, találunk egy magot (j számu folyóiratot), amely "a" számu cikket tartalmaz az adott témából, ahhoz, hogy további "a!" számu cikket kapjunk k x j folyóiratot kell átnézni, a harmadik harmad eléréséhez pedig k^2 x j. Vickery és később Leimkuhler kumulálta ezt az elosztást, és kimutatták, hogy a sorrendben az első folyóiratoknál (amelyeknek magas a rangjuk "r") ez megközelitőleg: A ln (1+Br), ahol A és B konstansok. A kumulálást az indokolja, hogy jobban követhető matematikailag a kumulált függvényforma. Bradford nem kumulált formában adta meg az eloszlást és nem határozta meg közelebbről, hogy hogyan kell megválasztani a nukleust (a magot). Ha a folyóiratok követik a Leimkuhler-féle eloszlást a nukleus tetszőlegesen választható meg.

Ezt a fent emlitett három eloszlást idézik legtöbbet a szakirodalomban. Megállapitották, hogy a legkülönbözőbb jelenségeket (városok lélekszámának eloszlása, könyvek könyvtári használata stb.) nagyon hasonló eloszlások irják le. Ezekről mondható, hogy

- először: igen egyszerüek, egymástól többé-kevésbé független, komplex helyzetekben alkalmazhatók és az eloszlásokat tükröző függvények nagyon egyszerü matematikai formában jelennek meg.
- Másodszor: az is világos, hogy ezek a megoszlások valóban léteznek és érvényesek. Például Lotka kisérletei azt igazolták, hogy ha öt éves időtartamra korlátozta a kémikusok (és vegyészek) és publikációik közötti megoszlás vizsgálatát, pontosan ugyanazt az eredményt kapta, mintha egy más generáció teljes életét tekintette volna a vizsgálat alapjának.
- Harmadsorban megállapitható, hogy majdnem teljesen azonosak a Lotka, Zipf, Bradford és Leimkuhler megoszlási függvények.

Kendall volt az első, aki bebizonyitotta, hogy Zipf eloszlása és Bradfordé kishiján azonos. A Zipf törvény kumulált formája (ln 1+2r) ugyanis a kumulált (egzakt) Bradford törvény (ln (1+Br) egy speciális esete.

Más szerzők kimutatták általánositva azt az észrevételt, hogy a Lotka és Bradford törvény ugyanannak az általános eloszlási függvénynek különböző közelitése és hogy legalábbis nagy "r" érték esetében mindkettő a Zipf egyenlettel azonos értéket ad.

Hogy mikor melyiket választják a szerzők eredményeik leirására, az attól is függ, hogy milyen jelenséget vizsgálnak. Ha például a folyóiratok-

ban megjelenő cikkeket vizsgáljuk (azok megoszlását), akkor a szakemberek elsősorban a sok témába vágó cikket tartalmazó folyóiratokat fogják ismerni, és kevesebb információjuk van az egy-két releváns cikket tartalmazó folyóiratok pontos számáról. Lotka esetében viszont könnyebb volt megállapitani, hogy hány szerző az, aki néhány cikket közöl, és egyre bizonytalanabb az eredmény, ha a nagyon sok cikket publikáló szerzőkről esik szó, ezért ő ezt az eloszlási függvény kis "r" értéktől (a kevésbé rangos szerzők felől) közelitette meg.

Sok erőfeszités történt ezeknek a tapasztalati megoszlási függvényeknek megmagyarázására. Sokkal ritkábban foglalkoztak ezek elméleti levezetésével, pedig csupán elméleti levezetéssel, információelméleti megfontolásokat figyelembe véve Mandelbrot a következő összefüggésig jutott:

$$N = \frac{A}{(1+ \beta \ r)^{\gamma}}$$

ahol β és γ állandó, "A" arányossági tényező, "n" az előfordulás gyakorisága, "r" pedig a rang (a sorrendbe sorolásnak megfelelően). Az eddig említett eloszlási függvények mind ennek az általános alaknak speciális esetei.

Problémát okozhat még, hogy az eddig tárgyalt eloszlási függvények matematikai megfogalmazása determinisztikus. A véletlen szerepére azonban jól rávilágit Simon modellje, amely szerint: ha egy szövegből véletlenszerüen választunk ki szavakat és egy-egy szó előfordulásának valószinüségét vizsgáljuk, az arányos az adott szó eddigi előfordulásai számának gyakoriságával. Ebből kiindulva kimutatható, hogy n-szer előforduló szavak száma arányos az

$$\frac{1}{n\ (n+1)\ldots,(n-k)\text{-val,}}$$

ahol a "k" konstans egészszám. "N" nagyértéke esetében megállapitható, hogy a fenti sorozat

$$\frac{1}{n^{\alpha}}\text{ -nal}$$

közelithető, ami nem más, mint a Mandelbrot-i eloszlás egy speciális esete.

Ezeket a determinisztrikus függvényeket ugy kell tehát értelmeznünk, hogy azok az eredmény legvalószinübb értékét adják. Kimutatható ugyanis, hogy ha az eredményként kapott legvalószinübb érték körül például Poisson eloszlást tételezünk fel, ez lényegesen nem befolyásolja a kapott végeredményt, viszont feloldja - a véletlenszerüség fogalmának bevezetése révén - az egyenletek determinisztikus értelmezéséből óhatatlanul fellépő olyan furcsaságokat, mint pl. két és fél szerzőnek összesen 80 cikket kellene publikálnia, vagy hogy két szerző nem publikálhat pontosan ugyanazon számu cikket.