Eötvös Loránd Tudományegyetem

Természettudományi Kar

Török Bálint

Véletlen permutációk modelljei

Szakdolgozat

Matematika Bsc, Matematikai elemző szakirány

Témavezető:

Csiszár Villő

Valószínűségelméleti és Statisztika Tanszék

Budapest, 2020

NYILATKOZAT

Név: Török Bálint

ELTE Természettudományi Kar, szak: Matematika Bsc

NEPTUN azonosító: RTG83S

Szakdolgozat címe:
Véletlen permutációk modelljei

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2020.05.25.

Török Bálint

a hallgató aláírása

Köszönetnyilvánítás

Ezúton szeretném megköszönni témavezetőmnek, Csiszár Villőnek a közreműködését, segítségét és iránymutatását, amikor elbizonytalanodtam. Külön szeretném megköszönni a türelmét és, hogy a kérdéseimre adott válaszai nagyon körültekintőek és alaposak voltak.
Továbbá hálás vagyok családomnak és barátaimnak, akik az egyetemi évek alatt végig támogattak, biztattak és nyugodt hátteret biztosítottak számomra.

Tartalomjegyzék

Bevezetés

Az emberi viselkedés egyik fontos eleme a legkülönfélébb dolgok sorbarendezése fontosságuk, vagy valamely más szempont alapján. Ez az egyén mindennapi életében ugyanúgy megvalósul például a napi teendők összeállításánál, mint az élet más területein. Általánosan elmondható, hogy szeretjük rangsorolni a minket körülvevő dolgokat, például, hogy kik a legjobb barátaink, kik a cégnél a legjobb dolgozók mind fontos információ számunkra. A rangsorolásnak rengeteg felhasználása van. Hogy néhányat említsek, a vállalatoknak tudniuk kell, hogy a vásárlóik mely termékeket szeretik jobban, a politikusoknak fontos tisztában lenniük a társadalmi értékekkel, a szavazói magatartással. Az adatok rangsorolása egyebek mellett gyakran előfordul pszhiológiában, szociológiában, az egészségügyben és a közgazdaságtanban is. A rangsorolási adatok elemzése és modellezése hatékonyan segít megérteni az emberek észleléseit és preferenciáit különböző dolgokkal kapcsolatban, melyek nem teljesen nyílvánvalóak első ránézésre. Az általános statisztikáknál sokszor az is előfordulhat, hogy a rendelkezésünkre álló minta között a nagyon kiugró értékek eltorzítják a becsléseinket, modelljeinket. Ilyen esetekben is hasznos lehet az adatok értékei helyett a rangjukkal számolni.

A rangsorolt adatok, a minta két alapvető része az n darab döntéshozó, akiket ezentúl bíráknak nevezek, valamint az m darab tárgy, amit rangsorolnak a bírák. Fontos megkülönböztetni, hogy rangsorolásról, vagy sorbarendezésről beszélünk. A rangvektorban az adott helyen lévő tárgy bíráktól kapott rangját, sorszámát tároljuk el, ahol az 1 a legjobbat az m pedig a legrossza-

bat jelöli. A sorrendvektorba a tárgyakat rendezzük sorba a bírák döntése alapján. A következőkben legtöbbször rangsorolásról fogunk beszélni, és ennek megfelelően rangvektorokat használunk. Teljes rangsorolásról akkor beszélünk, amikor az összes tárgyat sorba állítják a bírák, ezt $y \in \mathbb{R}^m$ vektorral jelölöm. Ezek az y vektorok az S_m térben vannak, ahol $S_m \equiv \{$az összes permutációja a rangoknak$\}$. Tehát az adott n bíró által létrehozott rangvektorok: $y^{(1)}, y^{(2)}, \ldots, y^{(n)} \in S_m$. A rangsorolt adatok elemzéséhez, majd a modellezéshez szükséges megértenünk az adataink matematikai hátterét. A szakdolgozat első fejezetében bevezetésre kerül a rangsorolás matematikai felírása, mint permutáció, és fontos tulajdonságai, melyek nélkülözhetetlenek a modellezés alapvető megértéséhez. Ezt követően a véletlen permutációk különböző modelljei kerülnek bemutatásra. Ezek bemutatásához a felhasznált fő irodalom az [1] könyv volt, mely segített megérteni a modellezés alapjait. A szakdolgozat lezárásaként a gyakorlatban is bemutatom az alkalmazásait a megismert modelleknek. Egy önkéntes kitöltésen alapuló internetes teszt formájában begyűjtött adathalmazt elemzek, és modellezek egy R-ben használható csomag segítségével.

1. fejezet

Matematikai háttér áttekintése

1.1. Permutációk

A bevezetőben megismert rangvektorokról elmondtuk, hogy az S_m halmaz elemei. A rangsorolást a matematikában permutációként értelmezzük.

1.1.1. Definíció. (Permutáció) Legyen X tetszőleges véges halmaz és π olyan függvény, melynek értelmezési tartománya és értékkészlete megegyezik, azaz $\pi\colon X \to X$. Ha π bijekció, akkor a π függvényt az X halmaz permutációjának nevezzük.

Jelöljük $[m]$-el, $m \in \mathbb{N}$ a legkisebb m pozitív egész számból álló halmazt. Ennek a halmaznak az összes permutációjának a halmazát a kompozíció műveletével ellátva egy szimmetrikus csoportot kapunk, melyet S_m-el jelölünk. Fontos megjegyezni, hogy $|S_m| = m!$. A $\pi \in S_m$ permutációt megadhatjuk kétsoros jelöléssel:

$$\pi = \begin{pmatrix} 1 & 2 & \dots & \mathrm{m} \\ \pi(1) & \pi(2) & \dots & \pi(m) \end{pmatrix}$$

Azonban ettől eltérően a következőkben az egyszerűség kedvéért a

$$\pi = (\pi(1), \dots, \pi(m))$$

jelölést fogjuk használni.

Ha π egy rangvektor, akkor $\pi(i) = j$ azt jelenti, hogy az i-edik tárgy a j-edik rangot kapta, vagyis a j-edik legkedveltebb. És értelmezhetjük hasonlóan a π^{-1} függvényt is, ahol $\pi^{-1}(j) = i$ jelentése az, hogy a j-edik legkedveltebb tárgy az i. Ez alapján a sorrendvektort is felírhatjuk:

$$(\pi^{-1}(1), \ldots, \pi^{-1}(m))$$

A rangvektor és sorrendvektor, mint permutációk tehát egymás inverzei. Az identitás permutációt, amely helyben hagyja az elemeket id_m-el jelöljük. Egy $\pi \in S_m$ permutációt transzpozíciónak nevezünk, ha kizárólag két elemét cseréli ki a halmaznak, a többit helyben hagyja.

1.1.2. Definíció. (Inverzió) Legyen $\pi \in S_m$ egy permutáció és $1 \leq i < j \leq m$. Ha $\pi(i) > \pi(j)$, akkor azt mondjuk, hogy ezek inverzióban állnak. Jelölje $inv(\pi)$ a π permutációban lévő inverziók számát, azaz

$$inv(\pi) =\mid \{(i,j) \in [m]^2 : i < j \text{ és } \pi(i) > \pi(j)\} \mid.$$

Definiálhatjuk a $\pi \in S_m$ permutáció ellentett permutációját is. Jelöljük ezt π^c-vel és legyen

$$\pi^c = (m+1)\big(1, 1, \ldots, 1\big) - \pi. \tag{1.1}$$

Tehát az ellentett permutációban a sorrend pontosan az ellenkezője az eredeti permutációnak.

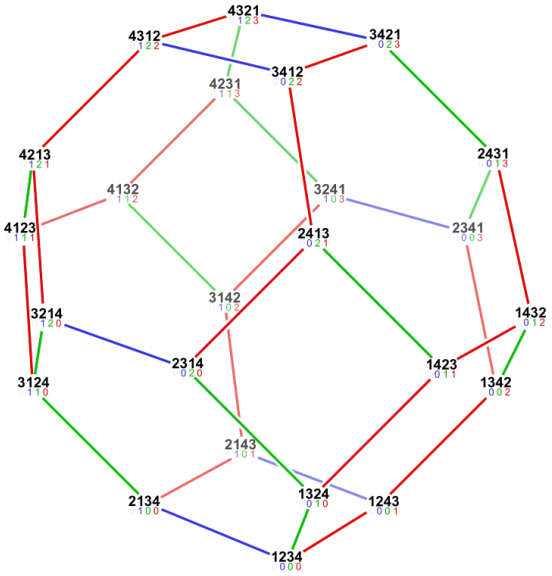
Egy permutációt megadhatunk az úgynevezett permutációs mátrix segítségével, melynek minden oszlopában és minden sorában egy darab 1-es van, a többi csupa nulla. Legyen az $m \times m$-es permutációs mátrixok halmaza $\mathcal{Q}_m$, ez kölcsönösen megfeleltethető S_m-el. Legyen $Q \in \mathcal{Q}_m$, ekkor $Q_{ij} = 1$, ha a hozzá tartozó $y \in S_m$ esetén $y_i = j$. Legyen $Q^{(y)} \in \mathcal{Q}_m$ az y-hoz tartozó permutációs mátrix, ezért

$$Q^{(y)} id_m = y, \tag{1.2}$$

ahol $id_m \in S_m$ az identitás permutációt jelöli. Jelöljük a sorbarendezések halmazát T_m-el. Az S_m és T_m halmazok kölcsönösen megfeleltethetőek egymásnak a már bevezetett inverzfüggvény segítségével. Azonban a permutációs mátrixok segítségével is megadható az $y \in S_m$-nek megfelelő $\omega \in T_m$. Legyen $Q^{(y)} \in \mathcal{Q}_m$ ahogy az előbb, ekkor $\omega = {Q^{(y)}}^T id_m$.

A permutációk ábrázolását az úgynevezett permutációs politóp segítségével tudjuk megoldani, ezt csak kis m-ekre tudjuk szépen ábrázolni.

1.1.3. Definíció. (Permutációs politóp) A permutációs politóp adott m-re a konvex burka az $y \in S_m \subset \mathbb{R}^m$ pontoknak.



1.1. ábra. S_4 politópja

Bevezetve a permutációs politóp fogalmát, már meg tudjuk határozni ennek a középpontját, ami egy konstans vektor. Ettől a ponttól a politóp minden egyes csúcsa egyenlő távolságra van. Tehát a bírók által megadott $y^{(i)}$ vektorok a politóp egy-egy csúcsát jelölik, de közel sem biztos, hogy a politóp összes csúcsa szerepel a rangvektorok között. Ha minden pontnak pontosan akkora súlyt adunk, ahányszor előfordul az adataink között, akkor

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \tag{1.3}$$

képlet segítségével egy új vektort definiálhatunk. Ezt nevezzük a rangvektorok átlagának. Így $\bar{y} \in \mathbb{R}^m$, vagyis a legtöbb esetben $\bar{y} \notin S_m$. Az $\bar{y}_i$ az i-edik tárgy átlagos rangját jelenti. Az $\bar{y}$ ponttól vett távolságok négyzetének összege minimális az euklideszi térben.

Egy cél lehet a központi rangsorolás megismerése, amit egy példán keresztül a legkönnyebb megérteni. Egy futóverseny végén ismerjük a versenyzők beérkezési sorrendjét, ezt nevezzük központi rangsornak, amit π_0-val jelölünk. A bírók ettől valószínűleg valamilyen mértékben eltérően állítják fel a saját sorrendjüket, hogy szerintük milyen sorrendben fognak befutni a versenyzők. Az esetek nagy részében azonban ez a központi rangsor számunkra ismeretlen. Ideális esetben az $y^{(i)}$-k közel helyezkednek el a π_0-hoz. De mit értünk az alatt, hogy közel helyezkednek el? Erre a kérdésre a választ a következő fejezetben kapjuk meg.

1.2. Távolság értelmezése permutációkon

A halmaz a matematika egy alapfogalma. A halmazokat különböző tulajdonságaik alapján szokták megkülönböztetni, az egyik legfontosabb, hogy lehet-e rajtuk valamilyen formában távolságot értelmezni. Ez nyilván azért fontos, mert ha értelmezünk rajta távolságot, akkor képesek vagyunk matematikailag összehasonlítani a halmaz elemeit, megmondani két elem mennyire van közel egymáshoz, milyen kapcsolatban állnak egymással. Ezen információk birtokában még több fontos következtetést lehet levonni a rendelkezésre álló adatokból.

A távolságot egy függvényként értelmezzük. Metrikák segítségével ki tudjuk számolni két permutáció egymástól való távolságát, valamint két véletlenszerűen választott permutáció egymástól való várható távolságát is, amiből már a szórást is meg tudjuk határozni a második momentum segítségével.

1.2.1. Definíció. (Metrikus tér) A metrikus tér egy olyan (X, d) pár, ahol

X tetszőleges nem üres halmaz, $d\colon X^2 \to \mathbb{R}_0^+$ pedig olyan nemnegatív valós értékű függvény, melyre tetszőleges $x, y, z \in X$ esetén:

1. $d(x, y) = 0 \Leftrightarrow x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

A permutációkon értelmezett távolság definíciója hasonló a metrikáéhoz, csak nem követeljük meg a háromszög egyenlőtlenség teljesülését. Ezeken kívül a permutációkon értelmezett távolságfüggvénynél megköveteljük, hogy jobbról invariáns legyen, azaz $\pi, \sigma, \gamma \in S_m$ esetén $d(\pi, \sigma) = d(\pi \circ \gamma, \sigma \circ \gamma)$, ahol $\pi \circ \gamma(i) = \pi(\gamma(i))$. Ez a feltélel azt biztosítja, hogy az elemek újracimkézése egy adott γ-val a távolságon ne változtasson.

Kétfajta megközelítése létezik a távolság értelmezésének két permutáció között. Az egyik a megszokott térbeli megközelítés, mennyi "időt" vesz igénybe az egyik pontból eljutni a másikba. A permutációkat a permutációs politópon tudjuk ábrázolni, ebben az esetben az ezen vett távolságokat értelmezzük. A másik fajta megközelítés a rendezettség, vagy éppen a rendezetlenség mértéke. Egy szemléletes példa erre, hogy a könyvek a polcon mennyire vannak az ábécé szerint sorbarendezve, mennyi időt vesz igénybe a helyes sorbarendezés.

Az előző fejezetben már beszéltünk a politóp középpontjának meghatározásáról, azonban ezek általában nem permutációk. Fontos lenne azonban tudni, hogy melyik az $\hat{\pi} \in S_m$, amely minimalizálja adott d mellett az átlagos távolságot

$$\bar{d}(\pi) \equiv \frac{1}{n} \sum_{i=1}^{n} d(y^{(i)}, \pi), \qquad \pi \in S_m. \tag{1.4}$$

Nézzünk néhány permutációkon értelmezett távolságot. Spearman a szokásos távolság megközelítést alkalmazta:

1.2.2. Definíció. (Spearman távolság) Legyen $y^*, y \in S_m$, ekkor

$$d_{Spear}(y^*, y) = \sum_{i=1}^{m} |y_i^* - y_i|^p, \tag{1.5}$$

ahol $0 < p < \infty$. Általában $p = 1$ és $p = 2$ értékeket szokták használni.

A $\bar{d}_{Spear}(\pi)$ kiszámolására $p = 2$ esetén létezik egy egyszerű formula:

$$\bar{d}_{Spear}(\pi) = \frac{m(m-1)(2m+1)}{3} - 2 \cdot \pi^T \bar{y}. \tag{1.6}$$

1.2.3. Definíció. (Diszharmonikus pár) Legyen $y^*, y \in S_m$. Diszharmonikus párnak nevezzük azokat az i, j tárgyakat, amelyeknek y^* és y rangvektor alapján különböznek a preferenciáik, azaz $i \neq j$ esetén

$$(y_i^* - y_j^*)(y_i - y_j) < 0. \tag{1.7}$$

Kendall két rangvektor távolságának a mérésére a diszharmonikus párok számát használja fel, tehát az egymáshoz mért rendezetlenséget veszi távolságnak.

1.2.4. Definíció. (Kendall távolság) Legyen $y^*, y \in S_m$, ekkor

$$d_{Ken}(y^*, y) = \sum \sum_{1 \leq i < j \leq m} I[(y_i^* - y_j^*)(y_i - y_j) < 0]. \tag{1.8}$$

Itt I az indikátor függvényt jelenti.

A Spearman távolsághoz hasonlóan ennek is létezik leegyszerűsített képlete az átlagos távolságra:

$$\bar{d}_{Ken}(\pi) = \sum \sum_{\{i,j|\pi_i > \pi_j\}} \hat{K}_{ij}, \tag{1.9}$$

ahol $\hat{K}$ az úgynevezett pármátrix, azaz $\hat{K}_{ij} = \frac{1}{n}|\{\ k|y_i^{(k)} < y_j^{(k)}\}|$. A $\hat{K}$ egy $m \times m$-es mátrix, ami az egyes tárgyak összehasonlításait tartalmazza. A mátrix ij-edik eleme azt mutatja, hogy az i-edik tárgyat a bírók hány százaléka preferálta a j-edik tárggyal szemben, azaz a minta hány százalékában kapott kisebb rangot az i-edik tárgy, mint a j-edik.

Ha az (1.5)-ben $p \to 0$ akkor a Hamming távolságot kapjuk. Ez a fajta távolságfügvény is a különbözőségeket vizsgálja két rangvektor között. Azonban itt nem számít, hogy egy hibás elhelyezés mennyivel rosszabb egy másik hibás elhelyezésnél, a lényeg, hogy rossz helyen van az adott objektum.

1.2.5. Definíció. (Hamming távolság) Legyen $y^*, y \in S_m$, ekkor

$$d_{Ham}(y^*, y) = \mid \{i \mid y_i^* \neq y_i\} \mid \tag{1.10}$$

1.3. Permutációk eloszlása

Már láttuk, hogy a rangvektorok az S_m elemei és $\mid S_m \mid = m!$. Az n bíró által megadott $y^{(i)}$ rangvektorokat megfeleltethetjük n darab független, azonos eloszlású valószínűségi változó sorozatának:

$$Y^{(1)}, Y^{(2)}, \dots Y^{(n)} \sim P, \tag{1.11}$$

ahol P egy valószínűségi eloszlás az S_m halmazon. Vagyis P egy $m!$ hosszúságú vektor, melynek az idexei az y permutációk, azaz az összes lehetséges rangvektorhoz tartozik egy index:

$$\mathbb{P}[Y = y] = P_y, \qquad y \in S_m. \tag{1.12}$$

Mivel P valószínűségekből álló vektor, ezért $P \subset \mathcal{P}_{m!} \subset \mathbb{R}^{m!}$, ahol

$$\mathcal{P}_k = \{x \in \mathbb{R}^k \mid x_i \geq 0 \text{ minden } i\text{-re, és} \sum_i x_i = 1\}. \tag{1.13}$$

Ezeket a fogalmakat bevezetve, már képesek vagyunk különböző eloszlásokat is definiálni. Ha minden egyes permutáció bekövetkezésére pontosan ugyanannyi az esély, akkor egyenletes eloszlásról beszélünk, ami precízen definiálva az előző jelöléseket használva:

$$Y \sim Egyenletes(S_m) \Leftrightarrow P = \left(\frac{1}{m!}, \frac{1}{m!}, \frac{1}{m!}, \dots \frac{1}{m!}\right) \in \mathbb{R}^{m!} \tag{1.14}$$

1.4. Maximum likelihood becslés

Tegyük fel, hogy van egy adathalmaz, melyet egy modell jól magyaráz. A modellnek van egy vagy több ismeretlen paramétere. Ezeket a paramétereket nem tudjuk mérni, azonban a rendelkezésre álló adatokból a lehető legjobban szeretnénk becsülni az értéküket. A paraméterbecslésnél a legáltalánosabb eljárás a maximum likelihood módszer.

1.4.1. Definíció. (Likelihood függvény) Legyen $X = (X_1, X_2 \dots X_n)$ független, azonos diszkrét eloszlású minta, θ az ismeretlen paraméter. A likelihood függvény:

- $L(\theta, x) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i)$

A likelihood függvény értékei azt mutatják, hogy adott θ mellett mennyire valószínű, hogy ez a minta jön ki. Mivel az a paraméterérték lesz a becslésünk, ami mellett a minta bekövetkezésének a valószínűsége a lehető legnagyobb, ezért azt a θ-t keressük, amelyik maximalizálja a függvényt. Előfordulhat, hogy nem létezik a függvénynek maximuma, vagy több maximumhelye is létezik. Ha több maximumhelye van, akkor mindegyiket maximum likelihood becslésnek tekintjük. Kihasználva a logaritmus függvény monotonitását sok esetben a loglikelihood függvényt használjuk, ami a logaritmusa a likelihood függvénynek, és megkönnyíti a számolást. A becsült paramétert $\hat{\theta}$-val jelöljük.

2. fejezet

Modellezés

A meglévő rangsorolt adatok modellezése segít mélyebben megérteni az adathalmaz felépítését az általános statisztikáknál. A 20. század elején kezdtek el először komolyabban foglalkozni a statisztika ezen területével. Az évek alatt két különböző megközelítése alakult ki a modellezésnek [1]:

- A rangsorolási eljárás modellezése
- A bírók összetételének a modellezése

Ez a két megközelítési mód nem választható szét teljesen egyértelműen egymástól, de mégis segít a modellalkotásban.

Az első megközelítés az úgynevezett "hard modeling", ami matematikailag próbálja meg leírni azt a folyamatot ami a bíróban lejátszódik a rangsorolás felállításakor. Általában akkor alkalmazzák, amikor viszonylag nagy a rangsorolandó tárgyak száma. Ez a fajta megközelítés feltételezi, hogy van valamilyen objektívan mérhető szempont, ami alapján a bírók alapvetően közel azonos sorrendet állítanak fel. Ilyen mérhető szempont lehet például visszatérve a versenyekhez az eddig elért eredmények.

A második megközelítés azt próbálja meg leírni, hogyan jellemezhető a rangsorolások eloszlása a bírók egy csoportjára nézve. A hangsúly a bírók

preferenciáin van, ami alapján csoportosítani lehet őket. Általában viszonylag kicsi a rangsorolandó dolgok száma, ilyen lehet például egy országgyűlési választás.

A első fajta megközelítés logikája segít fejleszteni a második megközelítés modelljeit, és ez fordítva is igaz, ezért nincsen nagy jelentősége ennek a kettősségnek.

A bevezetésben már leszögeztük, hogy rangsorolással és nem sorbarendezéssel fogunk foglalkozni. Vannak modellek melyeket eredetileg sorbarendezés szerint alkottak meg, de át lehet vezetni a sorbarendezést rangsorolássá, és fordítva is igaz ez a megállapítás.

2.1. Általános modell

Az előzőekben már definiáltuk a P eloszlást, mint egy vektor a $\mathcal{P}_{m!}$ térben. A valószínűségi modellek a valószínűség-eloszlások egy részhalmazát képezik. Általában ezt a részhalmazt θ-val szokták paraméterezni.

Legyen Θ olyan, hogy $\{P(\theta) \mid \theta \in \Theta\} \subset \mathcal{P}_{m!}$, ahol $P(\theta)$ egy függvény, amely Θ-ból $\mathcal{P}_{m!}$-be képez. Egy $y \in S_m$ bekövetkezésének valószínűsége a θ paraméter mellett:

$$P_y(\theta) = \mathbb{P}_\theta[Y = y] = \mathbb{P}_\theta(y), \qquad y \in S_m, \theta \in \Theta. \tag{2.1}$$

A gyakorlatban minden modell tartalmazza az egyenletes eloszlást, azaz megfelelő paraméter vagy paraméterek mellett egyenletes eloszlást produkál S_m-en, vagyis minden $y \in S_m$ permutáció bekövetkezésének a valószínűsége egyenlő. Legtöbbször a Θ egy alacsony dimenziószámú euklideszi tér részhalmaza. Két szélsőséges modell létezik, amit fontos megemlíteni. Az első modell, aminek csak egy eloszlása létezik, melyre $\mathbb{P}_\theta(y) = \frac{1}{m!}$ minden $y \in S_m$-re. A másik az úgynevezett telített modell, amikor az eloszlások halmaza az egész $\mathcal{P}_{m!}$. Ezek a modellek jó viszonyítási alapot szolgáltatnak amikor más modelleket tesztelünk. A nullmodell minden további modellnek a része, a telített modell

pedig az összes létező modellt tartalmazza. A cél az, hogy találjunk egy olyan modellt, ami jól illeszkedik a meglévő adatainkra, és hasznos tartalommal is bír. Annak ellenére hogy általában a mintát független, azonos eloszlásúnak tekintjük, nem feltétlenül szükséges, hogy a minta ennyire szigorú struktúrával rendelkezzen. Különböző statisztikákkal és ábrákkal is felfedezhetünk bizonyos mintázatokat, amik hasznos információval szolgálhatnak a modellezéshez. A minta alapján minden esetben meg tudjuk határozni a tapasztalati eloszlást, jelölje ezt $\hat{P}$. Ilyenkor a cél azt a $P(\theta)$-t meghatározni, ami a legjobban közelíti a $\hat{P}$-t. Az esetek döntő többségében a maximum likelihood módszer segítségével határozzuk meg a közelséget.

2.2. Távolság alapú modellek

A modellezni kívánt adatoknál feltehető, hogy létezik egy ismert központi rangsorolása a tárgyaknak, jelöljük ezt most $\pi_0 \in S_m$-mel. A bírók feltehetően ehhez a rangsorhoz viszonylag közeli rangsorokat állítanak fel, valamilyen távolság értelmezés alapján. A távolság alapú modellek a π_0-tól való távolság alapján adnak kisebb vagy nagyobb valószínűséget az adott rangsorolás bekövetkezésének, a θ paramétertől függően, ami a π_0 körüli kiterjedést paraméterezi. Ezek a távolság alapú modellek az exponenciális modellcsaládba taroznak. Adott π_0 központi rangsorolás, d távolságfüggvény, és egy valós θ paraméter mellett a távolságalapú modellek általános alakja:

$$\mathbb{P}_{\theta,\pi_0}(y) = exp(\theta d(\pi_0, y) - \psi(\theta)), \tag{2.2}$$

ahol ψ a "normáló konstans". Általában a θ paramétert becsüljük, a π_0 lehet ismert és ismeretlen is.

2.2.1. Definíció. (Elégséges statisztika) Legyen $X = (X_1, X_2, X_3 \ldots, X_n)$ diszkrét minta. A $T(X)$ statisztika elégséges, ha a $\mathbb{P}_\theta[X = x \mid T(X) = t]$ feltételes valószínűség független θ-tól. Másképp fogalmazva $T(X)$ elégséges statisztika, ha minden információt tartalmaz a θ paraméterre nézve.

2.2.2. Tétel. (Fisher–Neyman faktorizáció) *A $T(X)$ statisztika akkor és csak is akkor elégséges, ha létezik olyan h és g_θ függvények, melyekhez létezik a $\mathbb{P}_\theta(X = x)$ valószínűségnek $h(x)g_\theta(T(x))$ alakú faktorizációja.*

Abban az esetben, amikor ismert a központi rangsorolás, a θ paraméterre elégséges statisztika a $T(Y) = \sum_{i=1}^{n} d(y^{(i)}, \pi_0)$, ahol $Y = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})$, a rendelkezésre álló minta. Így a maximum likelihood becslést használva egyszerűen meghatározhatjuk $\hat{\theta}$-t. Probléma csupán akkor merülhet fel, amikor $T(Y)$ megegyezik a minimumával, 0-val, vagy a maximumával, ami $n \cdot Max(d)$. Ilyenkor nem létezik maximum likelihood becslése θ-nak, mivel $T(Y) = 0$ esetén az $L(\theta)$ likelihood függvény szigorúan csökken, $T(Y) = n \cdot Max(d)$ esetén szigorúan nő. Ezen esetek bekövektezésének a valószínűsége 0-hoz tart ahogy $n \to \infty$, azonban nagyon kis minta esetén előfordulhat.

2.2.1. Ismeretlen központi rangsor

Abban az esetben, amikor számunkra ismeretlen a központi rangsorolás, a (2.2) eloszlás a θ paraméter és az $\pi_0 \in S_m$ függvényeként írható fel. Legyen $T(Y) = \sum_{i=1}^{n} d(y^{(i)}, \pi_0)$. Ekkor a loglikelihood függvény:

$$l(\theta, \pi_0) = ln\big(\prod_{i=1}^{n} P_\theta(y^{(i)})\big) = \theta T(Y) - n\psi(\theta). \tag{2.3}$$

A normáló konstans nem függ π_0-tól, mivel $\psi(\theta) = ln \sum_{\pi \in S_m} exp(\theta d(\pi_0, \pi))$ és így bármely $\pi_0 \in S_m$ központi rangsor választással ez az összeg nem változik. Tehát a normáló konstans csak a θ paramétertől függ. Tekintsük $\theta > 0$-t rögzítettnek. Ebben az esetben a $L(\theta, \pi_0)$ függvény, csak a központi rangsortól függ, és azon π_0 mellett veszi fel a maximumát, amikor $T(Y)$ maximális. Tehát először azt a $\pi_0^{(1)}$-et kell meghatározni, amelynél $T(Y)$ a lehető legnagyobb. Ezután a maximum likelihood módszer segítségével könnyen kiszámolható $\theta^{(1)}$.

Ha $\theta < 0$ rögzített, akkor a likelihood fügvény maximuma a $T(Y)$ minimuma esetén következik be. Jelöljük $\pi_0^{(2)}$-val, ami esetén $T(Y)$ minimális. Így már $\theta^{(2)}$ is meghatározható.

A θ diszperziós paraméter, amely a permutációknak a központi permutációhoz viszonyított szétszóródását jellemzi. Ennek előjele az adott modelltől, és annak felírásától függ, általában a negatív előjel használata a jellemző. Előfordulhat, hogy $\hat{\theta} = 0$, ebben az esetben minden egyes $\pi \in S_m$-re ugyanazt a valószínűséget fogja adni a becslés. Tehát ha $\theta = 0$, akkor az eloszlás egyenletes.

Abban az esetben, ha a $d(y^*, y)$ távolságot a Spearman vagy a Kendall távolság szerint határozzuk meg, akkor $Max(d) = d(y^*, y) + d(y^*, y^c)$, minden $y^*, y \in S_m$-re, ezért tetszőleges permutációkra $d(y^*, y) = konst - d(y^*, y^c)$, és ebből látszik, hogy ha a távolságok összegét y maximalizálja, akkor y^c minimalizálja. Mindebből következik, hogy $(\pi_0^{(1)})^c = \pi_0^{(2)}$, $\theta^{(1)} = -\theta^{(2)}$, és $l(\pi_0^{(1)}, \theta^{(1)}) = l(\pi_0^{(2)}, \theta^{(2)})$, így elég megtalálni csupán az $\pi_0^{(1)}, \theta^{(1)}$ párt az előbb bemutatott eljárással. Mivel a távolság alapú modelleknek csak egy paramétere van, sok esetben az első lépésekben alkamazzák a modellezés során. Ritkán fordul elő, hogy csupán egyetlen paraméterrel tökéletesen jellemezni lehessen egy adathalmazt.

2.3. Összehasonlításon alapuló modellek

Eddig a bírók a tárgyak rangsorának felállítását rangvektorokkal, vagy sorrendvektorokkal adták meg. Egy másik lehetőség a tárgyak sorrendjének a felállítására, hogy a bírókat nem arra kérjük, hogy egy konkrét sorrendet állítsanak fel, hanem a tárgyakból alkotott párokról kell eldönteniük, hogy melyiket preferálják. Így m tárgy esetén $\binom{m}{2}$ összehasonlítást kell végezni. Egy konkrét ω sorbarendezésből nyílván egyértelműen meghatározhatóak a párösszehasonlítások, viszont ez fordítva nem feltétlenül igaz. Előfordulhat az úgynevezett "körbeverés", amikor nem lehet felállítani a megadott összehasonlításokból egy konkrét sorrendet. Megköveteljük tehát a bíróktól, hogy az összehasonlítások felállítása konzisztens legyen valamilyen sorbarendezéssel, vagyis, hogy egyértelműen meghatározzon egy rangsort.

2.3.1. Babington Smith modell

Mivel most sorbarendezésekkel foglalkozunk, ω_i az i-edik tárgynak az indexét jelenti, ahol $\omega \in T_m$ sorrendvektor, a sorbarendezendő tárgyak halmaza pedig $\{1, 2, \ldots, m\}$. Minden bírónak tehát el kell végeznie $\binom{m}{2}$ összehasonlítást. Feltehető, hogy a bírók homogén csoportot alkotnak, azaz minden egyes bíró az $\{ij\}$ párok összehasonlításánál ugyanolyan valószínűséggel preferálja az i-edik tárgyat a j-edikkel szemben. Ha ez a minta alapján nem mondható el, akkor a bírókat olyan csoportokra bonthatjuk, melyekre igaz ez az állítás. A bírók összetételének vizsgálatát nem részletezem. Tekintsük a bírók csoportját homogénnek. Legyen annak a valószínűsége, hogy a bírók az i tárgyat részesítik előnyben j-vel szemben p_{ij}, ahol $i < j$ az $\{i, j\}$ pár összehasonlításánál. Mivel a sorbarendezendő tárgyak halmaza $\{1, 2, \ldots, m\}$, ezt tekinthetjük egy rögzített sorrendnek. Legyen ez a sorrend $\omega_0 \in T_m$, ami alapján definiáljuk a p vektort. Mivel a párok összehasonlításának a sorrendjét rögzítettnek tekinthejük, p-nek az indexelése a következő felírásban egyértelműen megadható:

$$p = (p_{12}, p_{13}, \ldots, p_{1m}, p_{23}, \ldots, p_{2m}, \ldots, p_{m-1,m}) \tag{2.4}$$

Ennek megfelelően $p_{ji} = 1 - p_{ij}$. Fontos megjegyezni, hogy a p_{ij} paramétereknek nem feltétlenül kell konzisztensnek lenniük, csupán az összehasonlításoknak. Egy rangsor felállítása a következőképpen történik. Egymástól függetlenül elvégezzük az összehasonlításokat a p_{ij} valószínűségeket használva. Ha az összehasonlítások konzisztensek, akkor megkaptuk a rangsort, ha nem, akkor újra kell kezdeni az összehasonlításokat, amíg konzisztensek nem lesznek. Annak a valószínűsége, hogy az összehasonlítások konzisztensek adott p mellett:

$$C(p) = \sum_{\omega \in T_m} \prod_{i<j} p_{\omega_i \omega_j} \tag{2.5}$$

Szavakkal megfogalmazva, bármely ω sorrend bekövetkezésének a valószínűsége annak a valószínűsége, hogy a bíró azokat a párösszehasonlításokat állítja fel, ami megfelel ω-nak, feltéve, hogy olyan összahasonlításokat produkál, melyek konzisztensek valamilyen sorrendhez. Ez alapján a Babington Smith modell a

következő:

$$\mathbb{P}_p(\omega) = \frac{1}{C(p)} \prod_{i<j} p_{\omega_i \omega_j}. \tag{2.6}$$

A sorrendvektorokról könnyedén át lehet térni a rangvektorokra. Itt adott $y \in S_m$ rangvektorra definiáljuk az $I_{ij}(y)$ mennyiségeket. $I_{ij}(y) = 1$, ha $y_i < y_j$, ami pontosan azt jelenti, hogy i-t preferáljuk j-vel szemben, vagyis kisebb a rangja a rangvektorban. Egyébként pedig $I_{ij}(y) = 0$. Ezek alapján a Babington Smith modell felírható a következő alakban rangvektorokat használva:

$$\mathbb{P}_p(y) = \frac{1}{C(p)} \prod_{i<j} p_{ij}^{I_{ij}(y)} (1 - p_{ij})^{(1-I_{ij}(y))} = \frac{1}{C(p)} \prod_{\{i,j|y_i<y_j\}} p_{ij}. \tag{2.7}$$

Ebben az alakban a normáló konstans (2.5)-hoz hasonlóan meghatározható, $C(p) = \sum_{y \in S_m} \prod_{\{i,j|y(i)<y(j)\}} p_{ij}$.

Nagy m esetén viszonylag sok, $\binom{m}{2}$ paramétert használ ez a modell. A legnagyobb problémát mégis a normáló konstans kiszámítása okozza nagy m esetén, hiszen $m!$ elemet tartalmaz S_m, így $m!$ permutációnak kell összeadni a valószínűségeit. Ez $m \geq 10$ esetén már rendkívül időigényes. A Babington Smith modellnek különböző részmodelljei léteznek, melyek a paraméterek számát csökkentik, így egyszerűsítve a modellt.

Bradley és Terry a Babington Smith modell leegyszerűsítésére a p_{ij} értékek meghatározásához a $v = (v_1, v_2, \ldots, v_m)$ paraméterek bevezetését javasolták, ahol v_i az i-edik tárgy paramétere. Ezeket a v_i paramétereket használva:

$$p_{ij} = \frac{v_i}{v_i + v_j} \tag{2.8}$$

A paraméterek bevezetésének a koncepciója az volt, hogy minél inkább preferáltabb az i-edik tárgy, legyen annál nagyob a v_i értéke. Így egy adott $\omega \in T_m$-re (2.8)-at behelyettesítve (2.6)-ba megkapjuk a Mallows-Bradley-Terry modellt:

$$\mathbb{P}_v(\omega) = \frac{1}{C(v)} \prod_{i<j} \frac{v_{\omega_i}}{v_{\omega_i} + v_{\omega_j}} = \frac{1}{C^*(v)} \prod_{i=1}^{m} v_{\omega_i}^{m-i}, \tag{2.9}$$

ahol $C(v) = \sum_{\omega \in T_m} \prod_{i<j} (\frac{v_{\omega_i}}{v_{\omega_i} + v_{\omega_j}})$ behelyettesítve (2.5)-be a (2.8)-ban definiált értékeket. A (2.8) képlet nevezőjében lévő $(v_i + v_j)$ értékek minden

$\{i, j\}$ tárgy összehasonlításnál szerepelnek attól függetlenül, hogy melyik tárgyat preferálja a bíró, ezért ezen összegek szorzata is konstans, ezt felhasználva $C^*(v) = \left(\sum_{\omega \in T_m} \prod_{i<j} \frac{v_{\omega_i}}{v_{\omega_i}+v_{\omega_j}}\right) \prod_{i<j}(v_i + v_j) = \sum_{\omega \in T_m} \prod_{i<j} v_{\omega_i}$ a normáló konstans. Ez a modell már csak m paramétert tartalmaz. Sőt, mivel a paraméterek konstanssal való szorzása esetén nem változik a modell, valójában csak $m-1$ paramétert tartalmaz.

2.4. Mallows modellek

A Babington Smith modell paramétereinek további csökkentésével Mallows két darab egy-paraméteres modellt alkotott, melyeket rangvektorokkal definiálok.

2.4.1. Mallows ϕ modellje

Feltesszük, hogy létezik egy adott központi rangsorolás $\pi_0 \in S_m$. Ennél a modellnél a p_{ij} párösszehasonlítások valószínűségei csak attól függnek, hogy az π_0 központi rangsorban $\pi_0(i) < \pi_0(j)$ vagy $\pi_0(j) < \pi_0(i)$. Tehát minden $\pi_0(i) < \pi_0(j)$ esetén a p_{ij} értékek egyenlőek.A p_{ij} értékeket a következő módon definiáljuk:

$$p_{ij} = \frac{exp(\theta \cdot I[\pi_0(i) > \pi_0(j)])}{1 + exp(\theta)} \tag{2.10}$$

2.4.1. Állítás. *A Babington Smith modellbe behelyettesítve a (2.10)-ben definiált p_{ij} értékeket, egy távolság alapú modellt kapunk, melyben d a Kendall féle távolságfüggvény (1.8).*

Bizonyítás. Legyen $\pi_0 \in S_m$ a központi rangsor, így a p_{ij} paraméterek ez alapján a sorrend alapján definiálandók. A rangsorolandó tárgyak sorrendje legyen $\{1, 2, 3, \ldots, m\}$, és a központi rangsorolás egyezzen meg ezzel, vagyis a $\pi_0 \in S_m$ központi rangvektor egyenlő id_m-mel. Ennek köszönhetően $d_{Ken}(\pi_0, y) = inv(y)$.

Továbbá $i < j$ esetén $I[y_i > y_j] = I[(id_m)_i - (id_m)_j)(y_i - y_j) < 0]$. Ezért

$$\prod_{\{i,j|y_i<y_j\}} p_{ij} = \prod_{i<j} \frac{e^{\theta I[y_i>y_j]}}{1+e^{\theta}} = \frac{e^{\theta \cdot \sum_{i<j} I[y_i>y_j]}}{\left(1+e^{\theta}\right)^{\binom{m}{2}}} = \frac{e^{\theta \cdot inv(y)}}{\left(1+e^{\theta}\right)^{\binom{m}{2}}} =$$

$$\frac{e^{\theta \cdot d_{Ken}(id,y)}}{\left(1+e^{\theta}\right)^{\binom{m}{2}}} = (konst) \cdot e^{\theta \cdot d_{Ken}(\pi_0,y)}$$

□

Az eddig használt jelölésekkel a Mallows ϕ modelljének leggyakrabban használt alakja:

$$\mathbb{P}_{\theta,\pi_0}(y) = \frac{1}{\psi(\theta)} e^{\theta \cdot d_{Ken}(\pi_0,y)} \tag{2.11}$$

ahol $\psi(\theta) = \sum_{y \in S_m} e^{\theta \cdot d_{Ken}(\pi_0,y)}$ a normáló konstans, θ a π_0-tól való távolságok diszperziós paramétere. Ha $\theta = 0$, akkor teljesen egyenletes az eloszlás S_m-en. Mivel a p_{ij} értékek csak $\pi_0(i) - \pi_0(j)$ előjelétől függenek, bizonyos esetekben nem illeszkedik jól a mintára. Elég arra az esetre gondolni, hogy egy focibajnokság elején megtippeljük az egymás elleni mérkőzések győzteseit. Ekkor az első és második helyezett csapat egymás elleni mérkőzés győztesének p valószínűséggel az első csapatot választom, és ugyanúgy p valószínűséggel választom az első és utolsó helyezett mérkőzés győztesének az első csapatot. A valóságban az erőviszonyok figyelembevételével ez elég valószínűtlen.

2.4.2. Mallows θ modellje

Térjünk vissza Bradley és Terry formulához, ahol a p_{ij}-k meghatározásához a $v_1, v_2, \ldots, v_m$ paramétereket használjuk. A cél ismét a paraméterek számának a redukálása. Feltesszük, hogy létezik egy adott központi rangsorolás $\pi_0 \in S_m$. A p_{ij} paramétereket értelemszerűen π_0 alapján határozzuk meg. A Mallows-Bradley-Terry modellt felírhatjuk rangvektorokat használva. Legyen az $y \in S_m$ rangvektor, amiben az i-edik tárgy rangja y_i, ekkor az i-edik tárgy $(m - y_i)$ objektumot "előz meg" a rangosrolásban, tehát a hozzá tartozó v_i paraméter

$(m - y_i)$-szer szerepel (2.9)-ben látott produktumban. Így a modell:

$$\mathbb{P}_v(y) = \frac{1}{C^*(v)} \prod_{i=1}^{m} v_i^{(m-y_i)} \tag{2.12}$$

2.4.2. Állítás. *A (2.12) modellbe $v_i = exp(\pi_0(i) \cdot 2\theta)$ behelyettesítésével, ahol θ paraméter, egy távolság alapú modellt kapunk, melyben d_{Spear} a Spearmann féle távolság (1.5) $p = 2$ értékkel.*

Bizonyítás. A központi rangsor $\pi_0 \in S_m$ ismert. A tárgyak sorrendje egyezzen meg a központi rangsorolással, vagyis $\pi_0 = id_m$, mint az előbb, tehát $\pi_0(i) = i$. A p_{ij} értékek legyenek ugyanúgy definiálva mint eddig (2.8), melyeket a központi rangsorolás alapján meghatározhatunk, hiszen $v_i = exp(\pi_0(i) \cdot 2\theta)$ adott. Ez alapján annak a valószínűsége, hogy a bíró egy $y \in S_m$ rangsorolást állít fel a következő:

$$\prod_{i=1}^{m} v_i^{(m-y_i)} = \prod_{i=1}^{m} exp(2\theta \cdot i \cdot (m - y_i)) =$$

$$= exp(2\theta \cdot \sum_{i=1}^{m} i(m - y_i)) = exp(2\theta \cdot \sum_{i=1}^{m} im - (2\theta \cdot \sum_{i=1}^{m} iy_i)) =$$

$$= exp(2\theta \cdot (konst) - (2\theta \cdot \sum_{i=1}^{m} iy_i)) = (konst) \cdot exp(\theta \cdot d_{Spear}(\pi_0, y)),$$

mivel $d_{Spear}(\pi_0, y) = d_{Spear}(id, y) = (konst) - 2 \cdot \sum_{i=1}^{m} iy_i$, ahol d_{Spear} a Spearman féle távolságfüggvény (1.5) $p = 2$ mellett. □

Tehát ismét egy távolságalapú modellt kapunk, azonban Mallows ϕ modelljéhez képest, ahol csupán $\pi_0(i) - \pi_0(j)$ előjelétől függött a p_{ij} értéke, ennél a modellnél már az $\pi_0(i) - \pi_0(j)$ különbsége számít, ami sok esetben jobb illeszkedést biztosít a modellnek. Mallows θ modelljének általános alakja:

$$\mathbb{P}_{\theta,\pi_0}(y) = \frac{1}{\psi(\theta)} e^{\theta \cdot d_{Spear}(\pi_0, y)}, \tag{2.13}$$

ahol $\psi(\theta) = \sum_{y \in S_m} e^{\theta \cdot d_{Spear}(\pi_0, y)}$ a normáló konstans, θ a diszperziós paraméter ugyanazokkal a tulajdonságokkal, mint a Mallows féle ϕ modellnél.

2.5. Súlyozott távolság alapú modellek

A távolság alapú modelleknek az egyik bővítési lehetősége, hogy az eddigi egyenlő súlyokkal vett távolságok helyett a különböző rangokhoz, különböző súlyokat rendelünk, és így határozzuk meg a távolságokat. Ez azért is fontos, mert például egy versenyen általában sokkal fontosabb az első három helyezett rangsora, mint az utolsó háromé. Legyen a verseny végeredménye, azaz a központi rangsorolás $\pi_0 = (1, 2, 3, 4, 5)$, és a két bíró által megadott rangsor $\pi_1 = (1, 2, 3, 5, 4)$ és $\pi_2 = (2, 1, 3, 4, 5)$. Ekkor például a Kendall távolság esetén $d_{Ken}(\pi_0, \pi_1) = d_{Ken}(\pi_0, \pi_2)$, azonban a valóságban sokkal fontosabb, hogy az első bíró jól eltalálta az első helyezetteket. Tehát azt szeretnénk, hogy $d_{Ken}(\pi_0, \pi_1)$ kisebb legyen, mint $d_{Ken}(\pi_0, \pi_2)$. A rangokhoz a súlyokat a központi rangsorolás alapján rendeljük. A súlyok meghatározása általában azon az elven történik, hogy ha s_i nagy, akkor csak kevés bíró van azon az álláspontón, hogy a központi rangsorolásban az i-ediknek ítélt tárgynak nem az i-edik helyen kellene állnia. Tehát ha s_i nagy, akkor a bírók többsége egyetért azzal, hogy $\pi_0(k) = i$. Ha nem így lenne, akkor a távolság jelentősen nőne. Ebből adódóan, ha s_i nagyon közel van a 0-hoz, akkor a bírókat csak egy kicsit, vagy egyáltalán nem befolyásolja, hogy a központi rangsorolásban az i-ediknek rangsorolt elemet hányadiknak ítélik meg.

Lee és Yu a súlyozott Kendall távolságot felhasználva alkottak egy súlyozott távolságalapú modellt, amit súlyozott tau modellként ismerünk [6]. Ehhez a modellhez a távolságfüggvényt a következőképpen definiálták:

$$d_{Ken,s,\pi_0}(y^*, y) = \sum_{i<j} s_{\pi_0(i)} s_{\pi_0(j)} I[(y_i^* - y_j^*)(y_i - y_j) < 0] \tag{2.14}$$

Más súlyozott távolságokat is lehet generálni az előzőhöz hasonlóan, például a Spearman féle súlyozott távolságfüggvény a következőképpen írható fel:

$$d_{Spear,s,\pi_0}(y^*, y) = \sum_{i=1}^{m} s_{\pi_0(i)} |y_i^* - y_i|^p \tag{2.15}$$

3. fejezet

Társadalmi értékek rangsorolása

3.1. Adatok ismertetése

Az adataimat az ismerőseim köréből gyűjtöttem, a Google kérdőív anonim, önkéntességen alapuló kitöltésével. A kérdőívben arra kértem a válaszadókat, hogy rangsorolják a következő társadalmi értékeket fontosságuk szerint:

- Társadalmi egyenlőség
- Szabad véleménynyilvánítás
- Szolidaritás
- Hazaszeretet
- Család
- Önmegvalósítás
- Hit

Minden társadalmi érték mellé egy számot lehetett beírni 1-től 7-ig, úgy, hogy minden szám csak egyszer szerepelhet. Az 1-es az adott személy számára a legfontosabb értéket jelenti, a 7-es pedig értelemszerűen a legkevésbé fontosat.

Az adatokat .xlsx formátumban mentettem le, ahol a sorok egy személy válaszait tartalmazzák. Az adatok tisztításánál ellenőriztem, hogy a válaszadók helyesen értelmezték-e a feladatot, azaz permutációkat adtak-e meg. Sajnos a válaszadók kicsivel több, mint 30%-ának ez nem sikerült, ami jól mutatja, hogy egy teljes, 7 elemű rangsor felállítása is sok esetben problémát okoz. Abban az esetben amikor párok összehasonlításával szeretnénk mintához jutni, nyílvánvalóan még nagyobb problémát okoz a mintavétel, hiszen az összehasonlításoknak konzisztenseknek kell lenniük. Az adatok tisztítása után 387 helyes kitöltés maradt. A már megtisztított adatokat importáltam R-be, ahol többek között a *pmr* csomag felhasználásával elemeztem a mintát, amelyet kifejezetten rangsorolt adatok elemzésére és modellezésére fejlesztettek [4].

3.2. Elemzés

Első lépésben az adatokat átalakítom egy olyan formátumra, melyet a *pmr* csomag képes kezelni, és annak *destat* függvénye segítségével meghatároztam az alapvető statisztikákat:

```
$mean.rank
[1] 4.069948 3.981865 4.147668 4.518135 3.186528 3.917098 4.178756

$pair
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    0  189  199  225  145  166  207
[2,]  197    0  199  219  141  197  212
[3,]  187  187    0  222  131  180  194
[4,]  161  167  164    0  136  163  167
[5,]  241  245  255  250    0  245  236
[6,]  220  189  206  223  141    0  211
[7,]  179  174  192  219  150  175    0

$mar
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]   42   44   72   64   53   73   38
[2,]   31   55   65   83   72   51   29
[3,]    9   49   86   78   86   52   26
[4,]   31   66   41   33   43   93   79
[5,]  130   66   41   35   33   27   54
[6,]   43   68   55   69   62   41   48
[7,]  100   38   26   24   37   49  112
```

A minta átlagából megállapíthatjuk, hogy átlagosan az 5. elemet, azaz a Családot tartották határozottan a legfontosabb értéknek. Ez a megállapítás a párok mátrixát figyelembevéve is szépen igazolható, hiszen a párösszehasonlításoknál a Családot minden értékkel szemben többen részesítették előnyben. Itt a pármátrix elemei nem százalékban, hanem darabszámra vannak megadva. Eddig még nem volt szó a marginális mátrixról. A marginális mátrix ij-edik eleme:

$$\hat{M}_{ij} =\mid \{k \mid y_i^{(k)} = j\} \mid= \sum_{k=1}^{n} I[y_i^{(k)} = j] = \sum_{k=1}^{n} Q^{(y^{(k)})} \tag{3.1}$$

Az előző megállapítást, miszerint a Családot tartották átlagosan a legfontosabb értéknek, a marginális mátrix is igazolja. Az értékek ennél a mátrixnál is a darabszámot jelölik. Az marginális mátrix alapján a legtöbben, 130-an a Családot választották a legfontosabbnak, míg az átlagosan második legfontosabbnak tartott Önmegvalósítást csupán 43 válaszadó tekintette a legfontosabb értéknek. Különösen érdekes a Hit megítélése. Az átlag alapján nem mondható el, hogy a legfontosabb értékek közé tartozik. A marginális mátrix viszont egy fontos dologra hívja fel a figyelmet. Kiemelkedően sokan választották a legfontosabb, és ezzel együtt a legkevésbé fontos értéknek, míg a köztes értékek száma jóval alacsonyabb ezeknél. Ez alapján joggal feltételezhetjük, hogy a válaszadók csoportja nem homogén, és valószínűleg kettéválasztható egy vallásos, és egy nem vallásos csoportra. A bírók összehasonlításához majd később visszatérek.

Az eddigiek alapján azt sejthetjük, hogy nem egyenletes eloszlású a minta. Legyen a H_0: $P = Egyenletes(S_m)$. Többféle módon lehet ellenőrizni, hogy az $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$ minta alapján egyenletes-e az eloszlás. Az egyik általános eljárás, amikor az illeszkedésvizsgálatot a $\hat{P}$ tapasztalati eloszlást használva végezzük el. Ekkor a χ^2 próba esetén a próbastatisztika értéke

$$T_n = nm!\|\hat{P} - u_{m!}\|^2, \tag{3.2}$$

a szabadságfoka pedig $m!-1$. Itt $u_{m!} = (\frac{1}{m!}, \frac{1}{m!}, \ldots, \frac{1}{m!})$. Ez a próba csak akkor használható, ha a minta elemszáma jóval nagyobb, mint $m!$, ez a feltétel azonban most nem teljesül. Egy másik lehetőség az átlag rangvektor használata.

Ekkor a próbastatisztika értéke

$$T_n = \frac{12n}{m(m+1)} \|\bar{y} - c_m\|^2, \tag{3.3}$$

ahol $c_m = \frac{m+1}{2} 1_m$ a politóp középpontja, a szabadsági foka pedig $(m-1)$. A χ^2 próbát végrehajtva az átlag rangvektor segítségével, a próbastatisztika értéke a vizsgált adatok alapján 97.11522, a p érték pedig $4.810072\text{e} - 19$, így elutasítjuk a nullhipotézist, miszerint egyenletes eloszlású a minta.

3.3. Modellezés

A modellezés során a szakdolgazatban áttekintett modelleket alkalmaztam, melyeket a *pmr* csomag szintén tartalmaz. Annak eldöntésére, hogy melyik modell illik a legjobban az adatokra, azaz melyik magyarázza a legjobban a válaszadók preferenciáját több módszer is létezik, ezek közül kettőt vizsgáltam.

A BIC (Bayesian information criterion) érték meghatározásakor felhasználjuk a likelihood becslést, valamint a szabad paraméterek száma is szerepet játszik a modell jóságának meghatározásában.

$$\text{BIC} = t \cdot ln(n) - 2 \cdot ln(\hat{L}). \tag{3.4}$$

Itt t a szabad paraméterek számát jelöli, $\hat{L}$ pedig a likelihood függvény maximális értéke. Minél kisebb a modell BIC értéke, annál jobbnak tartjuk a modellt. Ebből a felírásból jól látszik, hogy ahogy nő a szabadsági fok, úgy a modell megbízhatósága is romlik.

A másik mérőszám a modellek vizsgálatakor a Pearson féle χ^2 teszt. Ez a teszt az illeszkedést a várható rangsorok és a megfigyelt rangsorok gyakorisága alapján határozza meg. Jelölje O_i az i-edik permutáció megfigyelt gyakoriságát, E_i pedig a modell alapján várható gyakoriságát. Ezeket a jelöléseket használva

$$\text{SSR} = \sum_{i=1}^{m!} \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

Itt is igaz, hogy minél kisebb az SSR értéke, annál jobbnak tekinthetjük a modellt, viszont ez az eljárás a pareméterek számát nem veszi figyelembe.

A modellek jelölésének megértéséhez szükséges megjegyezni, hogy a Spearman távolságokra $p = 1$ esetén Spearman Footrule távolságként szoktak hivatkozni, $p = 2$ esetén Spearman Rho-négyzet távolságnak hívjuk, melynek gyöke a Spearman Rho távolság. A különböző modellekre kapott eredményeket a következő táblázatban foglaltam össze:

Modellek	BIC	SSR	param.
Spearman rho	6470,596	9860,51	1
Footrule	6465,888	10250,56	1
Mallows ϕ	6511,307	9947,38	1
Mallows θ	6521,959	9743,12	1
Súlyozott Kendall távolságú	6528,322	9867,03	7
Súlyozott Spearman rho négyzet	6416,399	8717,95	7
Súlyozott Footrule	6451,799	9087.82	7

Az BIC értékek alapján az egyenlő súlyokkal vett távolságokon alapuló modellek közül Footrule modell illeszkedett a legjobban a mintára. A modellnek a becsült központi rangsorolása 3 4 5 6 1 2 7, tehát a Család és az Önmegvalósítás a két legfontosabb érték, míg a Hit kérdése a modell alapján a legkevésbé fontos érték. A becsült θ paraméter pedig 0.1260855. Az SSR értékek alapján az egyenlő súlyokkal vett távolságokon alapuló modellek közül Mallows θ modellje bizonyult a legjobbnak. A modellnek a becsült központi rangsorolása 4 3 5 7 1 2 6, tehát a Hit kérdése itt sem tartozik a legfontosabb értékek közé. A becsült paraméter pedig $0,01827257$. A súlyozott távolság alapú modellektől várható volt, hogy jobban teljesítenek az egyszerű távolság alapú modelleknél, hiszen magába foglalják azokat. Az összes kipróbált modell közül egyértelműen a Súlyozott Spearman Rho négyzet távolságú modell illeszkedik a legjobban az adatokra. A becsült központi rangsorolás ennél a modellnél 5 3 4 6 2 7 1, ami sokmindenben különbözik az előző modellekétől. A paraméterek 5 tizedesjegyre kerekített értékei:

```
[1] -0.01581  0.01989  0.05152  0.19220  0.04975  0.03138  0.00711
```

A paraméterek alapján a legnagyobb súlyt a Szolidaritás kapta, ezt követi nagyságredben a Társadalmi egyenlőség és a Szabad véleménynyílvánítás. Tehát a modell alapján a válaszadók többsége ezeknek a központi rangsorban elfoglalt helyével egyet ért. Láthatjuk, hogy a Hit a központi rangsorban az első helyen szerepel, azonban ehhez negatív súly tartozik, így itt annál valószínűbb lesz egy rangsorolás, minél jobban eltér a Hit rangja az 1-től. Tehát annak ellenére, hogy a központi rangsorolásban a Hit az első helyen szerepel, azok a permutációk a valószínűbbek, ahol az utolsó helyek között szerepel a rangsorolásban .

A marginális mátrix értékeiből már feltételezhető volt, hogy a válaszadók csoportja nem tekinthető homogénnak. Ennek vizsgálatára a *rankdist* R csomagot használtam, amelyben lehetőség van a bírók halmazát klaszterekre bontani. A bírók csoportját először két klaszterre bontottam. A két klaszterből álló modellek közül a Mallows ϕ modellje illeszkedett a legjobban a mintára, jobban mint az előző modellek.

```
Goodness of Fit
SSR:     8961.818
BIC:     6200.051
dof:     3
Parameter Estimation
Cluster A   B   C   D   E   F   G   p      Parameters
1       3   4   5   6   1   2   7   0.6    0.46
2       7   5   4   2   3   6   1   0.4    0.47
```

A két klaszter között a várakozásainknak megfelelően a Hit megítéléseben van drasztikus különbség. A modell alapján a második klaszter a vallásosnak feltételezett csoport. Ebben a klaszterban a Hit, Hazaszeretet és Család a három legfontosabb társadalmi érték a becsült központi rangsor alapján, ami talán nem meglepő. A másik csoportot nem ennyire egyszerű körülhatárolni, azonban fontosnak tartom megjegyezni, hogy a modell becsült központi rangsora alapján a Család az első helyen áll a társadalmi értékek között, míg a Hit és Hazaszeretet az utolsó helyeken. A két klaszter diszperziós paramétere csak

nagyon kis mértékben tér el egymástól, vagyis a központi rangsorolás körül közel azonos mértékben terjednek el a permutációk.

Abban az esetben amikor 3 klaszterre bontjuk a válaszadókat, szintén Mallows ϕ modelljét használva a következő eredményt kapjuk:

```
Goodness of Fit
SSR:      7829.551
BIC:      6112.262
dof:      5
Parameter Estimation
Cluster A   B   C   D   E   F   G   p      Parameters
1       7   5   4   2   3   6   1   0.41   0.47
2       4   3   5   7   1   2   6   0.37   0.63
3       1   2   3   6   4   5   7   0.22   0.58
```

Látható, hogy az előző, két klaszteres modellben lévő keresztény konzervatív tömb aránya megmaradt, és a központi rangsor is megegyezik az előző modellben látottal, vagyis joggal feltételezhetjük, hogy az új klaszterközéppont bevezetése ezt a csoportot nem érintette, vagyis a válaszadók 40%-a továbbra is ebbe a csoportba tartozik. Ezek alapján a másik klaszter vált ketté. Az egyikben a család és az önmegvalósítás kap nagyobb hangsúlyt, a másikban a társadalmi egyenlőség, szabad véleménynyívánítás, szoldidaritás, a nyugati világ sokat hangoztatott értékei. Látható, hogy a diszperzió paraméterek ebben a modellben már jelentősen eltérnek egymástól, vagyis a klasztereken belül már a permutációk szétszóródása különbözik.

A kérdőív kitöltői nem nevezhetőek reprezentatív mintának, ennek ellenére érdekes mintázatokat lehetett felfedezni a modelleket vizsgálva. Abban az esetben, ha több adattal is rendelkezünk a válaszadókról, például ismerjük a korukat, nemüket jóval pontosabb képet kaphatunk a társadalmunk társadalmi értékeiről, preferenciáiról. Összességében elmondható, hogy a válaszadók rangsorolásait vizsgálva a család egyértelműen az egyik legfontosabb érték.

Irodalomjegyzék

[1] John I. Marden: *Analyzing and Modeling Rank Data*, 1995, London: Chapman & Hall

[2] Michael A. Fligner, Joseph S. Verducci: *Probability Models and Statistical Analyses for Ranking Data*, 1993, New York: Springer-Verlag Berlin Heidelberg

[3] Douglas E. Critchlow: *Metric Methods for Analyzing Partially Ranked Data*, 1985, New York: Springer-Verlag Berlin Heidelberg

[4] R Documentation, PMR Package documentation
https://www.rdocumentation.org/packages/pmr/versions/1.2.5

[5] Zempléni András, Leíró és matematikai statisztika előadás fóliái
https://zempleni.elte.hu/stat_el19.html

[6] Zhaozhi Qian, Philip L. H. Yu: Weighted Distance-Based Models for Ranking Data Using the R Package rankdist, Journal of Statistical Software July 2019, Volume 90, Issue 5.
https://www.jstatsoft.org/article/view/v090i05

[7] https://machinelearningmastery.com/probabilistic-model-selection-measures/