Permutációk és metrikák alkalmazása a statisztikában

Szakdolgozat

Írta: Drahos Csaba

Matematika BSc

Alkalmazott matematikus szakirány

Témavezető:

Csiszár Villő

adjunktus

Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem

Természettudományi Kar

2012

Köszönetnyilvánítás

Hálával tartozom témavezetőmnek, Csiszár Villőnek, aki precíz figyelmével, hasznos ötleteivel és tanácsaival segítette munkámat e dolgozat megírásában. Köszönettel tartozom családomnak és barátaimnak támogatásukért és bátorításukért.

Budapest, 2012. május 31.

Drahos Csaba

Tartalomjegyzék

1. fejezet

Bevezető

A statisztikában jelentős szerepe van a permutációk alkalmazásának. Gyakran a mintát néhány kiugró érték annyira eltorzítja, hogy a matematikai módszerek nem a várakozásainknak megfelelő eredményeket adják. Ezért felmerült egy olyan ötlet, hogy a konkrét mintaelemek helyett a rangjukkal számoljunk. Ebben a szakdolgozatban a célom olyan matematikai összefüggések áttekintése, amelyek lehetővé teszik rangsorok alkalmazásával nyert statisztikai eredmények születését. Ötleteimet főként a [1] könyvből merítettem. A legtöbb metrikának kiszámítjuk a várható értékét, ezt próbáltam minél különbözőbb módszerek útján bemutatni, megmutatva, hogy a permutációkkal való számolásokban milyen lehetőségek rejlenek, egy-egy új ötlet felhasználása során. Látni fogjuk, hogy egyes esetekben a várható érték és a metrika maximuma milyen szoros kapcsolatban állnak, illetve, hogy milyen összefüggések kötik össze a metrikák világát a korrelációval, végül az alkalmazások között speciális függetlenség-vizsgálati és homogenitás-vizsgálati módszereket mutatunk.

2. fejezet

Permutációk és tulajdonságaik

2.1. Definíció. *(Permutáció) Legyen X tetszőleges halmaz és $s : X \to X$ függvény, ha $D(s) = R(s) = X$ és s bijekció, akkor az s függvényt az X halmaz permutációjának nevezzük.*

2.2. Definíció. *(Szimmetrikus csoport) Legyen X tetszőleges halmaz, ekkor X összes permutációjának halmazát a kompozíció műveletével ellátva S_X szimmetrikus csoportnak nevezzük.*

Legyen $n \in \mathbb{N}$ és jelöljük az n legkisebb pozitív egész számból álló halmazt $[n] = \{1, 2, \ldots, n\}$, ekkor az $S_{[n]}$ szimmetrikus csoportot csak S_n-nel jelöljük.

2.3. Definíció. *(Transzpozíció) Legyen $s \in S_X$ permutáció, ha létezik olyan $x, y \in X$, amelyre $x \neq y$ esetén $s(x) = y$, $s(y) = x$ és minden $z \in X \setminus \{x, y\}$ esetén, $s(z) = z$, akkor az s permutációt transzpozíciónak nevezzük.*

Tehát egy $s \in S_X$ permutáció pontosan akkor transzpozíció, ha az X halmaz két különböző elemét kicseréli és a többi elemét helyben hagyja.

2.1. Állítás. *Legyen X véges halmaz és $s \in S_X$ permutáció, ekkor létezik olyan $t_1, t_2, \ldots, t_k \in S_X$ transzpozíció, amelyre $s = \prod_{i=1}^{k} t_i$.*

2.4. Definíció. *(Ciklus) Legyen $s \in S_X$ permutáció, ha létezik olyan $x_1, x_2, \ldots, x_k \in X$, amelyre minden $i \in \{1, 2, \ldots, k-1\}$ esetén $s(x_i) = x_{i+1}$, $s(x_k) = x_1$ és minden $x \in X \setminus \{x_1, x_2, \ldots, x_k\}$ esetén $s(x) = x$, akkor az s permutációt ciklusnak nevezzük.*

2.5. Definíció. *(Diszjunkt ciklusok) Legyen $c_1, c_2 \in S_X$ ciklus, ha $\{x \in X \mid c_1(x) \neq x\} \cap \{y \in X \mid c_2(y) \neq y\} = \emptyset$, akkor c_1 és c_2 diszjunkt ciklusok.*

2.2. Állítás. *Legyen X véges halmaz és $s \in S_X$ permutáció, ekkor léteznek olyan $c_1, c_2, \ldots, c_k \in S_X$ diszjunkt ciklusok, amelyekre $s = \prod_{i=1}^{k} c_i$.*

A ciklus definíciója szerint minden $c = (x_1 x_2 \ldots x_l) \in S_X$ ciklus esetén $c = (x_1 x_2 \ldots x_l) = (x_2 x_3 \ldots x_l x_1) = \cdots = (x_l x_1 \ldots x_{l-1})$ és ha $c_1, c_2 \in S_X$ diszjunkt ciklusok, akkor $c_1 c_2 = c_2 c_1$, mert különböző elemeket mozgató ciklusok esetén az eredmény nem függ a szorzás sorrendjétől. Legyen $s \in S_X$ tetszőleges permutáció, $s = \prod_{i=1}^{k} c_i$ az s diszjunkt ciklusfelbontása és minden $c_i = (x_{i1} x_{i2} \ldots x_{il_i})$, ekkor minden i esetén létezik olyan j_i amelyre $x_{ij_i} = \min\{x_{ij} \mid j \in \{1, 2, \ldots, l_i\}\}$. Írjunk fel minden c_i ciklust $c_i = (x_{ij_i} x_{ij_i+1} \ldots x_{il_i} x_{i1} \ldots x_{ij_i-1})$ alakban, ezzel elértük, hogy minden ciklus a legkisebb általa mozgatott elemmel kezdődik. Most tekintsük a ciklusok kezdő elemeit, legyen $x_1^* < x_2^* < \cdots < x_k^*$ az $x_{1j_1}, x_{2j_2}, \ldots, x_{kj_k}$ elemek növekvő sorrendje, ekkor egyértelműen létezik olyan $c_i^* \in \{c_1, c_2, \ldots c_k\}$ ciklus, amely az x_i^* elemet mozgatja, ha minden c_i ciklust $c_i = (x_{ij_i} x_{ij_i+1} \ldots x_{il_i} x_{i1} \ldots x_{ij_i-1})$ alakban írunk fel és a ciklusok sorrendjét is a kezdő elemük szerinti növekvő sorrendnek választjuk, akkor az s permutáció így meghatározott $s = \prod_{i=1}^{k} c_i^*$ ciklusfelbontása egyértelmű.

2.3. Állítás. *Legyen X véges halmaz és $n \in \mathbb{N}$, ha $|X| = n$, akkor $S_X \cong S_n$.*

3. fejezet

Metrika értelmezése szimmetrikus csoportokon

Legyen $(\Omega, \mathcal{A}, P)$ valószínűségi mező, (S_n, d) metrikus tér, $S, T : \Omega \to S_n$ egyenletes eloszlású permutáció értékű valószínűségi változó és $X_d : \Omega \to \mathbb{R}$ olyan valószínűségi változó, amelyre minden $\omega \in \Omega$ esetén
$X_d(\omega) = d\left(S\left(\omega\right), T\left(\omega\right)\right)$, ekkor az X_d valószínűségi változót a d metrika által generált valószínűségi változónak nevezzük és minden $x \in \mathbb{R}$ esetén

$$P(X_d = x) = \frac{\left|\left\{(s,t) \in S_n^2 \mid d\left(s,t\right) = x\right\}\right|}{|S_n^2|}.$$

A relatív gyakoriság szerint felírt képletből az eloszlás meghatározása a legtöbb esetben még nem túl nagy n esetén is elég sok számolással jár. Minden n esetén $\frac{n!^2}{2}$ számítást kell elvégeznünk, ha belegondolunk, hogy a metrika szimmetrikus tulajdonsága miatt az összes lehetőség fele épp ugyanazt az eredményt adja, mint a másik fele. Ez $n = 10$ esetén is már 6.584.094.720.000, viszont a következő tulajdonság teljesülése drasztikusan csökkenti ezt az értéket.

3.1. Definíció. *(Invariáns metrika) Legyen (S_n, d) metrikus tér és $s, t \in S_n$ tetszőleges permutáció, ha minden $r \in S_n$ permutáció esetén $d(s,t) = d(sr, tr)$, akkor a d metrikát jobbról invariáns metrikának nevezzük és hasonlóan, ha $d(s,t) = d(rs, rt)$, akkor a d metrikát balról invariáns metrikának nevezzük.*

Legyen minden $s \in S_n$ permutáció esetén $D(s) = d(s, id)$ és az $S, T : \Omega \to S_n$ egyenletes eloszlású valószínűségi változók esetén $X_D(S(\omega)) = X_d(S(\omega), id)$, ha d jobbról invariáns metrika, akkor minden $t \in S_n$ permutáció esetén $d(s,t) = d(st^{-1}, id) = D(st^{-1})$, így $X_d(S(\omega), T(\omega)) = X_D(S(\omega)T^{-1}(\omega))$. Ekkor meggondolható, hogy ST^{-1} is egyenletes eloszlású. Ha d balról invariáns metrika, akkor minden $s \in S_n$ permutáció esetén $d(s,t) = d(id, s^{-1}t) = D(s^{-1}t)$, így $X_d(S(\omega), T(\omega)) = X_D(S^{-1}(\omega)T(\omega))$. Ekkor meggondolható, hogy $S^{-1}T$ is egyenletes eloszlású. Ez az ötlet segítségünkre lesz abban, hogy metrikákat valószínűségi változóként tekintve kiszámoljuk a várható értékét és a szórását.

3.2. Definíció. *(Metrika várható értéke és szórásnégyzete) Legyen (S_n, d) metrikus tér, ekkor a d metrika várható értéke $E(d) = E(X_d)$ és szórásnégyzete $D^2(d) = E(X_d^2) - E^2(X_d)$.*

3.1. Diszkrét metrika

3.3. Definíció. *(Diszkrét metrika) Legyen $s, t \in S_n$ permutáció és*

$$d_D(s,t) = \begin{cases} 1 & s \neq t \\ 0 & s = t \end{cases},$$

ekkor a $d_D : S_n^2 \to \mathbb{R}$ függvényt diszkrét metrikának nevezzük.

A diszkrét metrika konstrukciója elég egyszerű ahhoz, hogy példát mutassunk arra, hogy egy metrika várható értéke és szórása hogyan számolható ki az általa generált valószínűségi változó eloszlásának segítségével. Legyen minden $s \in S_n$ permutáció esetén $D(s) = d(id, s)$, a diszkrét metrika invariáns metrika ezért, ha egyenletes eloszlás szerint véletlenszerűen választunk egy permutációt, akkor $P(X_D = 0) = \frac{1}{n!}$, mert $D(s) = 0$ pontosan akkor, ha $d(id, s) = 0$, ami csak úgy teljesülhet, ha $s = id$. Ezért $P(X_D = 1) = \frac{n! - 1}{n!}$. Definíció szerint kiszámíthatjuk a várható értéket:

$$E(X_D) = \sum_{k=0}^{1} kP(X_D = k) = P(X_D = 1) = \frac{n! - 1}{n!}.$$

Tehát két permutáció átlagos távolsága a diszkrét metrikában $\frac{n! - 1}{n!}$.
A szórásnégyzet meghatározásához szükségünk van a valószínűségi változó második momentumára, viszont ez éppen a várható értéke. Ezért
$D^2(X_D) = \sum_{k=0}^{1} k^2 P(X_D = k) - \left(\frac{n! - 1}{n!}\right)^2 = \frac{n! - 1}{n!} - \left(\frac{n! - 1}{n!}\right)^2 = \frac{n! - 1}{n!^2}$.
Tehát a diszkrét metrikában a permutációk távolságának szórása $\frac{\sqrt{n! - 1}}{n!}$.

3.2. Hamming metrika

3.4. Definíció. *(Hamming távolság) Legyen $s, t \in S_n$ permutáció, ekkor a*

$$d_H(s, t) = \sum_{i=1}^{n} \delta(s_i, t_i)$$

kifejezést az s és t permutációk Hamming távolságának nevezzük (ahol δ a Kronecker-féle δ függvényt jelöli).

3.1. Állítás. *(S_n, d_H) metrikus tér.*

Bizonyítás: Legyen $s,t \in S_n$ tetszőleges permutáció, ekkor $d_H(s,t) = 0$ pontosan akkor, ha $s = t$ és $d_H(s,t) = d_H(t,s)$ a metrika definiciójából következik. Minden $s,t \in S_n$ esetén $d_H(s,t)$ felírható

$$d_H(s,t) = \sum_{i=1}^{n} \mathrm{sgn}|s_i - t_i|$$

alakban. Mivel minden $a,b \in \mathbb{R}$ esetén $\mathrm{sgn}\left(|a+b|\right) \leq \mathrm{sgn}|a| + \mathrm{sgn}|b|$ és $\mathrm{sgn}\left(|a|+|b|\right) \leq \mathrm{sgn}|a|+\mathrm{sgn}|b|$, ezért ezt felhasználva beláthatjuk a háromszög-egyenlőtlenséget. Legyen $s,t,r \in S_n$ tetszőleges permutáció, ekkor

$$\begin{aligned} d_H(s,t) = \sum_{i=1}^{n} \mathrm{sgn}|s_i - t_i| = \sum_{i=1}^{n} \mathrm{sgn}|s_i - r_i + r_i - t_i| \leq \\ \leq \sum_{i=1}^{n} \mathrm{sgn}\left(|s_i - r_i| + |r_i - t_i|\right) \leq \sum_{i=1}^{n} \mathrm{sgn}|s_i - r_i| + \\ + \sum_{i=1}^{n} \mathrm{sgn}|r_i - t_i| = d_H(s,r) + d_H(r,t) \end{aligned}$$

Tehát a $d_H : S_n^2 \to \mathbb{R}$ függvény metrika, tehát (S_n, d_H) metrikus tér. □

Számítsuk ki a Hamming távolság várható értékét és szórását. Legyen minden $s \in S_n$ permutáció esetén $\mathrm{Fix}(s) = \left|\{x \in [n] \mid s(x) = x\}\right|$ az s fixpontjainak száma és $H(s) = d_H(id, s)$, ekkor $H(s) = n - \mathrm{Fix}(s)$, mivel $H(s)$ éppen annak a száma, hogy hány pont nem fix az s permutációban. Jelölje az X valószínűségi változó egy egyenletes eloszlás szerint véletlenszerűen választott permutáció fixpontjainak számát és X_i az i-edik pozícióban lévő fixpontok számát. Ekkor

$$X_i = \begin{cases} 1 & s_i = i \\ 0 & s_i \neq i \end{cases}$$

és $X = \sum_{i=1}^{n} X_i$. Mivel $(n-1)!$ féleképpen fordulhat elő, hogy az i-edik pozícióban fixpont van, ezért

$$E(X_i) = \sum_{k=0}^{1} kP(X_i = k) = P(X_i = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Így $E(d_H) = E(n) - E(X) = n - \sum_{i=1}^{n} \frac{1}{n} = n - 1$. Tehát két véletlenszerűen választott permutáció átlagos Hamming távolsága $n-1$.

A szórás meghatározásához szükségünk lesz a Hamming távolság második momentumára, ehhez kiszámoljuk az X valószínűségi változó második momentumát.

$$E\left(X^2\right) = E\left(\sum_{i=1}^{n} X_i\right)^2 = E\left(\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right) =$$
$$\sum_{i=1}^{n} E\left(X_i^2\right) + 2\sum_{i<j} E\left(X_i X_j\right) = \sum_{i=1}^{n}\sum_{t=0}^{1} t^2 P(X_i = t) +$$
$$2\sum_{i<j}\sum_{t=0}^{1} tP(X_i X_j = t) = \sum_{i=1}^{n} P(X_i = 1) + 2\sum_{i<j} P(X_i X_j = 1)$$

Minden $i \neq j$ esetén a $P(X_i X_j = 1)$ éppen annak a valószínűsége, hogy két különböző pozícióban fixpont van. Az i és j pozíciókban egyféleképpen lehet fixpont, a többi pozícióban szereplő értéket a megmaradt $n-2$ elem elhelyezésével tetszőlegesen választhatjuk így $(n-2)!$ féle lehetőség képzelhető el, ezért $P(X_i X_j = 1) = \frac{(n-2)!}{n!}$, tehát

$$\sum_{i<j} P(X_i X_j = 1) = \binom{n}{2}\frac{(n-2)!}{n!} = \frac{1}{2},$$

eszerint $E(X^2) = 1 + 2 \cdot \frac{1}{2} = 2$. A második momentum felhasználásával

$D^2(X) = E(X^2) - E^2(X) = 2 - 1 = 1$, így $D(X) = 1$. Persze a metrika definíciója szerint $H = n - X$, ezért $D(H) = 1$. Tehát két véletlenszerűen választott permutáció távolságának szórása a Hamming metrikában 1.

Tekintsük az $n - H$ valószínűségi változót, megmutatható, hogy a $\lim_{n\to\infty}(n - H)$ határeloszlás létezik és Poisson(1) eloszlású.

A $H(s)$ kifejezés megegyezik az s permutáció diszjunkt ciklusfelbontásában szereplő ciklusok összhosszával. Például $s = (1, 3, 4, 2, 5)$, ekkor $s = (234)$, tehát pontosan 3 olyan elem van, amely nem önmagába megy, így $H(s) = 3$.

3.3. Spearman metrika

3.5. Definíció. *(Spearman távolság) Legyen $s, t \in S_n$ permutáció, ekkor a*

$$d_S(s,t) = \sum_{i=1}^{n} |s_i - t_i|$$

kifejezést az s és t permutációk Spearman távolságának nevezzük.

3.2. Állítás. *(S_n, d_S) metrikus tér.*

Bizonyítás: Legyen $s, t \in S_n$ tetszőleges permutáció, ekkor $d_S(s,t) = 0$ pontosan akkor, ha $s = t$ és $d_S(s,t) = d_S(t,s)$ a metrika definíciójából következik. Legyen $s, t, r \in S_n$, ekkor

$$d_H(s,t) = \sum_{i=1}^{n} |s_i - t_i| = \sum_{i=1}^{n} |s_i - r_i + r_i - t_i| \leq \sum_{i=1}^{n} (|s_i - r_i| + |r_i - t_i|) = \sum_{i=1}^{n} |s_i - r_i| + \sum_{i=1}^{n} |r_i - t_i| = d_S(s,r) + d_S(r,t),$$

tehát a $d_S : S_n^2 \to \mathbb{R}$ függvény metrika, ezért (S_n, d_H) metrikus tér. $\square$

A Spearman metrika azonos az L^1 metrikával, a szimmetrikus csoport elemein.

A Spearman metrika két permutáció távolságát pozíciónként kiszámolt értékek összegeként számítja ki, ezért ha az X valószínűségi változó jelöli egy egyenletes eloszlás szerint véletlenszerűen választott permutáció Spearman távolságát az identitástól, és X_i jelöli az i-edik pozícióban számolt különbséget, akkor X felírható $X = \sum_{i=1}^{n} X_i$ alakban. Tehát X várható értékének kiszámításához határozzuk meg $E(X_i)$ értékét minden $i \in \{1, 2, \ldots, n\}$ esetén. A várható érték definíciója szerint

$$E(X_i) = \sum_{k=0}^{n} kP(X_i = k).$$

Tegyük fel, hogy $i \leq \frac{n+1}{2}$ Ha $k = 0$, akkor $P(X_i = k) = \frac{1}{n}$, mert csak akkor lesz $|i - s_i| = 0$, ha $s_i = i$. Ha $k \in \{1, 2, \ldots, i-1\}$, akkor s_i kétféleképpen lehet pontosan k távolságra i-től, tehát $P(X_i = k) = \frac{2}{n}$. Ha $k \in \{i, i+1, \ldots, n-i\}$, akkor az $|i - s_i| = 0$ egyenlet egyik megoldása negatív lenne, tehát $P(X_i = k) = \frac{1}{n}$. Végül, ha $k > n - i$ akkor az egyik megoldás szintén negatív, a másik viszont n-nél nagyobb, így $P(X_i = k) = 0$. Összefoglalva tehát

$$P(X_i = k) = \begin{cases} 0 & k > n - i \\ \frac{1}{n} & k \in \{0, i, i+1, \ldots, n-i\} \\ \frac{2}{n} & k \in \{1, 2, \ldots, i-1\} \end{cases}.$$

Így már könnyen kiszámítható a várható érték, ugyanis

$$E(X_i) = \sum_{k=0}^{n} kP(X_i = k) = \sum_{k=1}^{i-1} \frac{2k}{n} + \sum_{k=i}^{n-i} \frac{k}{n} = \frac{i(i-1)}{n} + \frac{n+1-2i}{2}$$
$$= \frac{i(i-1) + (n+1-i)(n-i)}{2n}.$$

Nyilvánvaló, hogy X_{n+1-i} ugyanolyan eloszlású, mint X_i és mivel $E(X_i)$ szimmetrikus kifejezés i-ben és $n+1-i$-ben, így ez lesz a várható érték minden $i > \frac{n+1}{2}$ esetén is. Tehát

$$E(X) = \sum_{i=1}^{n} \frac{i(i-1)}{n} + \sum_{i=1}^{n} \frac{n+1-2i}{2} = \frac{(n-1)(n+1)}{3}.$$

Ezért két permutáció átlagos távolsága a Spearman metrikában $\frac{(n-1)(n+1)}{3}$. Levezethető, hogy $D^2(d_S) = \frac{(n+1)(2n^2+7)}{45}$. Megmutatható, hogy X aszimptotikusan normális eloszlású, a közelítés $n \geq 10$ esetén már elég jó.

3.4. Cayley metrika

3.6. Definíció. *(Cayley távolság) Legyen $s, t \in S_n$ permutáció és minden t_i transzpozíció, ekkor a*

$$d_C(s,t) = \min_k \{s \prod_{i=1}^{k} t_i = t\}$$

kifejezést az s és t permutációk Cayley távolságának nevezzük.

3.3. Állítás. *(S_n, d_C) metrika.*

Bizonyítás: Természetesen a triviális tulajdonságok könnyen ellenőrizhetők. A háromszög egyenlőtlenség is könnyen megmutatható. Legyen

$s, t, r \in S_n$ tetszőleges permutáció, a Cayley távolság megadja a minimális transzpozíciók számát, amelyekkel az s permutáció áttranszformálható a t permutációba. Ezért $d_C(s,t)$ transzpozíció szükséges ahhoz, hogy s-et áttranszformáljuk t-be és $d_C(t,r)$ transzpozíció szükséges ahhoz, hogy t-t áttranszformáljuk r-be, ezért legfeljebb $d_C(s,t) + d_C(t,r)$ transzpozícióval s átvihető r-be, vagyis $d_C(s,t) + d_C(t,r) \geq d_C(s,r)$. $\square$

Ha az egyik változót az identitásnak választjuk, akkor legyen $C(s) = d_C(id, s)$. A $C(s)$ kifejezés megadja, hogy legalább hány transzpozíció szükséges az s permutáció felírásához, ezért $C(id) = 0$. Ha $K(s)$ jelöli a körök minimális számát, azaz a diszjunkt ciklusfelbontásban szereplő körök számát, $(K(id) = 0)$, akkor az alábbi összefüggés teljesül:

3.4. Állítás. *Legyen $s \in S_n$ tetszőleges permutáció, ekkor*

$$H(s) = C(s) + K(s).$$

.

Bizonyítás: Legyen $s = \prod_{i=1}^{k} c_i$ és legyen $l_i = |\{x \in [n] \mid c_i(x) \neq x\}|$ a c_i ciklus hossza. Tegyük fel, hogy s felbontásában a ciklusok hosszuk szerinti sorrendben szerepelnek, azaz $l_i \leq l_{i+1}$. Számoljuk meg, hogy hány azonos hosszú ciklus szerepel a felbontásban, tehát legyen $\alpha_m = |\{c_i \in S_n \mid l_i = m\}|$, azaz hogy hány ciklus pontosan m hosszú a c_i ciklusok között. Legyen $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, ahol $\alpha_1 = 0$, mert minden 1-hosszú ciklust az identitásnak tekintjük, amit ebben a reprezentációban nem számoltunk a ciklusok közé. Az α vektor meghatározza a körök számát, így $\sum_{i=1}^{n} \alpha_i = K(s)$, mert a körök száma nem függ a körök hosszától, így csak össze kell adni, hogy hány 1-hosszú, 2-hosszú, ..., n-hosszú ciklus szerepel

benne. Úgy kapjuk meg, hogy hány transzpozíció szükséges s felbontásához, ha a diszjunkt ciklusfelbontásban szereplő ciklusokat bontjuk fel transzpozíciók szorzatára, ekkor a

$c_i = (x_{i1}x_{i2}\dots x_{il_i}) = (x_{i1}x_{i2})(x_{i2}x_{i3})\dots(x_{il_i-2}x_{il_i-1})(x_{il_i-1}x_{il_i})$, tehát

$c_i = \prod_{j=1}^{l_i-1}(x_{ij}x_{ij+1})$. Írjuk fel α segítségével a ciklusok számából a transzpozíciók számát, α szerint s-ben α_2 2-hosszú ciklus, azaz transzpozíció szerepelt, ezek szerepelni fognak a transzpozíciókkal felírt szorzatban is, egy 3-hosszú ciklus az előző felbontás szerint előáll 2 transzpozíció szorzataként, de α_3 darab ilyen volt s-ben, így ezek a ciklusok $3\alpha_3$ darab transzpozícióval növelik az eddig felfedezett transzpozíciók számát. Hasonlóan, egy i-hosszú ciklus $i-1$ transzpozíció szorzatára bontható, így α_i darab i-hosszú ciklus $(i-1)\alpha_i$ darab transzpozíció szorzatára bomlik, ezért összesen

$0\alpha_1+1\alpha_2+2\alpha_3+\dots+(n-1)\alpha_n = \sum_{i=1}^{n}(i-1)\alpha_i$ darab transzpozíció szorzatára bontottuk s-et. Könnyen látszik, hogy ennél kevesebb transzpozíció szorzatára már nem tudjuk felbontani, ezért $C(s) = \sum_{i=1}^{n}(i-1)\alpha_i$. Az $i\alpha_i$ szorzat éppen az i-hosszú ciklusok számának i-szeresét határozza meg, tehát egyenlő az összes i-hosszú ciklus által mozgatott elemek számával, ezt minden i-re összeadva az eredmény éppen a ciklusok összhossza, tehát $\sum_{i=1}^{n} i\alpha_i = H(s)$. Mivel a $\sum_{i=1}^{n}\alpha_i$ kifejezésben összeadjuk az 1-hosszú, 2-hosszú, n-hosszú ciklusok számát, ezért ez az összeg éppen az s-ben lévő ciklusok számát adja meg, vagyis $\sum_{i=1}^{n}\alpha_i = K(s)$. Ezért adódik, hogy

$$C(s) = \sum_{i=1}^{n}(i-1)\alpha_i = \sum_{i=1}^{n} i\alpha_i - \sum_{i=1}^{n}\alpha_i = H(s) - K(s),$$

tehát $H(s) = C(s) + K(s)$. □

A Cayley metrika várható értékét kiszámolhatjuk a következő módon: Annak a valószínűsége, hogy egy véletlenszerűen választott $S = (S_1, S_2, \ldots, S_n)$ permutációban $S_1 = 1$, az $P(S_1 = 1) = \frac{1}{n}$. Ha $S_1 = 1$, akkor annak a valószínűsége, hogy $S_2 = 2$, az $P(S_2 = 2 \mid S_1 = 1) = \frac{1}{n-1}$ és hasonlóan minden $P(S_k = k \mid S_1 = 1, S_2 = 2, \ldots S_{k-1} = k-1) = \dfrac{1}{n-k+1}$. Mivel nem tudjuk, hogy az egyik pozícióban lévő érték a helyén van vagy cserével érhetjük el, hogy a helyére kerüljön, ezért jelölje X_i azt, hogy az i-edik pozícióban kell-e cserélnünk. Így $P(X_{n+1-i} = 1) = 1 - \dfrac{1}{i}$ és $P(X_{n+1-i} = 0) = \dfrac{1}{i}$. Ezért
$X = \sum_{i=1}^{n} X_i$, így $E(X) = \sum_{i=1}^{n} \left(1 - \frac{1}{i}\right) = n - \sum_{i=1}^{n} \frac{1}{i}$. Tehát
$E(C) = n - \sum_{i=1}^{n} \dfrac{1}{i}$ és $D^2(C) = \sum_{i=1}^{n} \dfrac{1}{i}\left(1 - \dfrac{1}{i}\right)$.

3.5. Kendall metrika

3.7. Definíció. *(Kendall metrika) Legyen $s, t \in S_n$ tetszőleges permutáció, ekkor a*

$$d_\tau(s,t) = \sum_{i,j} \chi(s_i < s_j, t_i > t_j)$$

kifejezést az s és t permutációk Kendall távolságának nevezzük.

3.5. Állítás. *(S_n, d_τ) metrikus tér.*

Bizonyítás: Legyen $s, t \in S_n$ tetszőleges permutáció, ekkor
$d_\tau(s,t) = 0$ pontosan akkor, ha $s = t$ és $d_\tau(s,t) = d_\tau(t,s)$ triviálisan következik a metrika definíciójából. A háromszög egyenlőtlenség igazolásához használjuk a definícióban szereplő kifejezés egy ekvivalens alakját. Legyen

$s, t, r \in S_n$ tetszőleges permutáció, ekkor

$$d_\tau(s,t) = \sum_{i,j} \chi(s_i < s_j, t_i > t_j) =$$
$$\sum_{i<j} \chi(s_i - s_j < 0, t_i - t_j > 0) =$$
$$\sum_{i<j} \mathrm{sgn}|\mathrm{sgn}(s_i - s_j) - \mathrm{sgn}(t_i - t_j)|.$$

Ezt felhasználva beláthatjuk a háromszög egyenlőtlenséget

$$d_\tau(s,t) = \sum_{i<j} \mathrm{sgn}\big|\mathrm{sgn}(s_i - s_j) - \mathrm{sgn}(t_i - t_j)\big| =$$
$$\sum_{i<j} \mathrm{sgn}\big|\mathrm{sgn}(s_i - s_j) - \mathrm{sgn}(r_i - r_j) + \mathrm{sgn}(r_i - r_j) - \mathrm{sgn}(t_i - t_j)\big|$$
$$\leq \sum_{i<j} \mathrm{sgn}\Big(\big|\mathrm{sgn}(s_i - s_j) - \mathrm{sgn}(r_i - r_j)\big| + \big|\mathrm{sgn}(r_i - r_j) - \mathrm{sgn}(t_i - t_j)\big|\Big)$$
$$\leq \sum_{i<j} \mathrm{sgn}\big|\mathrm{sgn}(s_i - s_j) - \mathrm{sgn}(r_i - r_j)\big| + \sum_{i<j} \mathrm{sgn}\big|\mathrm{sgn}(r_i - r_j) - \mathrm{sgn}(t_i - t_j)\big|$$
$$= d_\tau(s,r) + d_\tau(r,t).$$

□

Számítsuk ki a Kendall-féle távolság várható értékét és szórásnégyzetét. Legyen $\tau(s) = d_\tau(id, s)$ és jelölje $\mathrm{Inv}(s)$ az s permutáció inverziószámát, ekkor $\tau(s) = \mathrm{Inv}(s)$. A metrika invariáns tulajdonságát kihasználva számítsuk ki a metrika várható értékét és szórását. A Kendall-féle távolság annak a számát adja meg, hogy hányféleképpen választhatunk ki olyan $(i, j) \in [n] \times ([n] \setminus \{i\})$ párokat, melyekben az első változóban szereplő permutációban a megadott pozícióban lévő elemek ellentétes relációban állnak a második változóban

szereplő permutációban megfelelő pozícióiban található értékekkel.

Jelölje az X valószínűségi változó egy egyenletes eloszlás szerint véletlenszerűen választott permutáció inverziószámát és legyen minden $i < j$ esetén

$$X_{ij} = \begin{cases} 1 & s_i > s_j \\ 0 & s_i < s_j \end{cases}.$$

Ekkor $X = \sum_{i<j} X_{ij}$ mert az inverziószám éppen azt adja meg, hogy hány esetben fordul elő hogy kisebb indexű pozícióban nagyobb érték található a nagyobb indexű pozícióban található értéknél.

$$E(X) = E\left(\sum_{i<j} X_{ij}\right) = \sum_{i<j} E(X_{ij}) = \sum_{i<j} \sum_{t=0}^{1} tP(X_{ij} = t) = \sum_{i<j} P(X_{ij} = 1)$$

A $P(X_{ij} = 1)$ valószínűség kiszámításához ki kell választanunk azt a két elemet, amelyek inverzióban állnak, ezt $\binom{n}{2}$ féleképpen tehetjük meg és a maradék $n-2$ elemet $(n-2)!$ féleképpen választhatjuk meg, így $P(X_{ij} = 1) = \frac{\binom{n}{2}(n-2)!}{n!} = \frac{1}{2}$, ezért $E(X) = \sum_{i<j} \frac{1}{2} = \frac{1}{2}\binom{n}{2}$.

A szórás kiszámításához szükségünk lesz X második momentumára.

$$E\left(X^2\right) = \sum_{i<j} E\left(X_{ij}^2\right) + \sum_{\substack{i<j \\ k<l \\ (i,j)\neq(k,l)}} E\left(X_{ij}X_{kl}\right) = \sum_{i<j} \sum_{t=0}^{1} t^2 P(X_{ij} = t)$$

$$+ \sum_{\substack{i<j \\ k<l \\ (i,j)\neq(k,l)}} \sum_{t=0}^{1} tP(X_{ij}X_{kl} = t) = \sum_{i<j} P(X_{ij} = 1) + \sum_{\substack{i<j \\ k<l \\ (i,j)\neq(k,l)}} P(X_{ij}X_{kl} = 1).$$

A $P(X_{ij}X_{kl} = 1)$ valószínűség kiszámításához esetszétválasztást kell végezni aszerint, hogy az i, j, k és l indexek közül melyek egyenlők, mivel $i < j$,

$k < l$ és $(i,j) \neq (k,l)$ ezért, ha $i < j = k < l$ vagy $k < l = i < j$, akkor $P(X_{ij}X_{kl} = 1) = \frac{\binom{n}{3}\cdot(n-3)!}{n!} = \frac{1}{6}$, ha $i = k$ és $j \neq l$ vagy $i \neq k$ és $j = l$, akkor $P(X_{ij}X_{kl} = 1) = \frac{2\cdot\binom{n}{3}\cdot(n-3)!}{n!} = \frac{1}{3}$ és ha $|\{i,j,k,l\}| = 4$, akkor $P(X_{ij}X_{kl} = 1) = \frac{\binom{n}{2}\cdot\binom{n-2}{2}\cdot(n-4)!}{n!} = \frac{1}{4}$. Így:

$$\sum_{\substack{i<j \\ k<l \\ (i,j)\neq(k,l)}} P(X_{ij}X_{kl} = 1) = \frac{1}{6}\big|\{(i,j,k,l) \in [n]^4 \mid i < j = k < l\}\big| +$$

$$\frac{1}{6}\big|\{(i,j,k,l) \in [n]^4 \mid k < l = i < j\}\big|$$

$$\frac{1}{3}\big|\{(i,j,k,l) \in [n]^4 \mid i = k < j \neq l, k < l\}\big| +$$

$$\frac{1}{3}\big|\{(i,j,k,l) \in [n]^4 \mid i \neq k < j = l, i < j\}\big| +$$

$$\frac{1}{4}\big|\{(i,j,k,l) \in [n]^4 \mid |\{i,j,k,l\}| = 4\}\big| = \frac{1}{3}\binom{n}{3} + \frac{2}{3}\binom{n}{3} + \frac{2}{3}\binom{n}{3} + \frac{6}{4}\binom{n}{4}$$

Ekkor $E\left(X^2\right) = \frac{1}{2}\binom{n}{2} + \frac{5}{3}\binom{n}{3} + \frac{3}{2}\binom{n}{4}$.

Ezt felhasználva a szórásnégyzet már könnyen kiszámolható, mivel $D^2(X) = E(X^2) - E^2(X)$, így egyszerűsítés után adódik, hogy $D^2(X) = \frac{n(n-1)(2n+5)}{72}$. A Cayley távolság azt méri, hogy két permutáció hány transzpozícióval transzformálható egymásba, viszont a Kendall távolság a két permutáció közt lévő inverziót számítja, ami éppen az egymásba transzformáláshoz szükséges szomszédos transzpozíciók száma, ezért $C \leq \tau$. Megmutatható, hogy minden $s,t \in S_n$ permutáció esetén $d_\tau(s,t) + d_C(s,t) \leq d_S(s,t)$ és $d_S(s,t) \leq 2d_\tau(s,t)$.

3.6. Ulam metrika

3.8. Definíció. *(Ulam távolság) Legyen $s, t \in S_n$ tetszőleges permutáció és $L(s)$ a leghosszabb növekvő részsorozat hossza s-ben,ekkor a*

$$d_U(s,t) = n - L(ts^{-1})$$

kifejezést az s és t permutációk Ulam távolságának nevezzük.

A szokásos módon rögzítsük a metrika egyik változóját, legyen $U(s) = d_U(s, id)$ és tekintsük a következő problémát: legyen egy polcon n darab könyv valamilyen sorrendben, ekkor hány lépésben tudjuk elérni azt, hogy a könyvek, egy előre megadott sorrendben kövessék egymást? Legyen a könyvek eredeti sorrendje az identitás és az elérni kívánt sorrend s, ekkor minimálisan $U(s)$ lépésben tudjuk a kívánt sorrendbe rakni a könyveket.

Ennek a metrikának az eloszlása, várható értékének és szórásának kiszámítása máig megoldatlan probléma. Az várható értékre felírható, hogy

$$\lim_{n\to\infty} \frac{n - E(d_U)}{\sqrt{n}} = 2.$$

4. fejezet

Rangkorreláció

Legyen $d : S_n^2 \to \mathbb{R}$ metrika az S_n szimmetrikus csoport elemein, mivel S_n véges halmaz, ezért $\max_{s,t} d(s,t)$ létezik és véges. Legyen $\widehat{d} = 1 - \frac{2d}{\max_{s,t} d}$, ekkor a d metrikából nyert $\widehat{d} : S_n^2 \to [-1,1]$ függvény korrelációs együtthatóként használható és, mivel nem a mért adatokat csak az azok rangjának megfelelő permutációkat használja, ezért nevezzük rangkorrelációnak. Az eddig említett metrikák esetén megadható belőlük nyert rangkorreláció, de erre nem térünk ki, inkább két gyakran használt rangkorrelációs eszközt említünk.

4.1. Spearman-féle rangkorreláció

A statisztikában egy adott $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ minta esetén a korrelációs együtthatót az

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

kifejezés határozza meg. Bizonyos esetekben, néhány kiugró értéknek köszön-

hetően, ez a képlet nem megfelelően fejezi ki a két mennyiség közti korrelációt. Annak megelőzése érdekében, hogy a kiugró értékek eltorzíthassák az eredményt, a korrelációs együttható helyett a mennyiségek rangkorrelációs együtthatóját számoljuk ki. Ez azt jelenti, hogy a mintában szereplő értékek helyett, azok rangját helyettesítjük a képletbe.

Legyen $s_i = |\{x_j \mid x_j \leq x_i\}|$ az x_i rangja és $t_i = |\{y_j \mid y_j \leq y_i\}|$ az y_i rangja, ekkor a

$$\varrho = \frac{\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2 \sum_{i=1}^{n}(t_i - \bar{t})^2}}$$

kifejezés meghatározza a rangok korrelációját, ezért nevezzük rangkorrelációnak. A képlet egyszerűbb alakra hozható. Mivel
$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$ és ez minden rangsor esetén teljesül, ezért
$\bar{t} = \bar{s} = \frac{n+1}{2}$. A számláló a következő alakban írható:

$$\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t}) = \sum_{i=1}^{n}\left(s_i - \frac{n+1}{2}\right)\left(t_i - \frac{n+1}{2}\right) = \sum_{i=1}^{n} s_i t_i - \frac{n+1}{2}\sum_{i=1}^{n} s_i$$
$$- \frac{n+1}{2}\sum_{i=1}^{n} t_i + \sum_{i=1}^{n}\left(\frac{n+1}{2}\right)^2 = \sum_{i=1}^{n} s_i t_i - \frac{2n(n+1)^2}{4} + \frac{n(n+1)^2}{4} =$$
$$\sum_{i=1}^{n} s_i t_i - \frac{n(n+1)^2}{4}$$

A nevező a következő módon egyszerűsíthető:

$$\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2 \sum_{i=1}^{n}(t_i - \bar{t})^2} = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 =$$

$$\sum_{i=1}^{n}\left(i^2 - i(n+1) + \frac{(n+1)^2}{4}\right) = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} = \frac{(n-1)n(n+1)}{12}$$

A képletbe a rangok esetén kiszámolt összefüggéseket behelyettesítve

$$\varrho = \frac{\sum_{i=1}^{n} s_i t_i - \frac{n(n+1)^2}{4}}{\frac{(n-1)n(n+1)}{12}}.$$

Ezt a képletet átalakíthatjuk a következő módon:

$$\varrho = \frac{\sum_{i=1}^{n} s_i t_i - \frac{n(n+1)^2}{4}}{\frac{(n-1)n(n+1)}{12}} = \frac{12}{(n-1)n(n+1)}\sum_{i=1}^{n} s_i t_i - 3\frac{n+1}{n-1} =$$

$$= \frac{12}{(n-1)n(n+1)}\sum_{i=1}^{n} s_i t_i + \frac{(n-1)-(4n+2)}{n-1} = 1 - \frac{2(2n+1)}{n-1}$$

$$+ \frac{12}{(n-1)n(n+1)}\sum_{i=1}^{n} s_i t_i = 1 - \frac{2(2n+1)n(n+1)}{(n-1)n(n+1)} +$$

$$\frac{12}{(n-1)n(n+1)}\sum_{i=1}^{n} s_i t_i =$$

$$1 - \frac{6}{(n-1)n(n+1)}\left(\frac{2n(n+1)(2n+1)}{6} - 2\sum_{i=1}^{n} s_i t_i\right) =$$

$$1 - \frac{1}{\binom{n+1}{3}}\left(2\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} s_i t_i\right) =$$

$$1-\frac{1}{\binom{n+1}{3}}\left(\sum_{i=1}^{n} i^2-2\sum_{i=1}^{n} s_i t_i+\sum_{i=1}^{n} i^2\right)=$$

$$1-\frac{1}{\binom{n+1}{3}}\left(\sum_{i=1}^{n} s_i^2-2\sum_{i=1}^{n} s_i t_i+\sum_{i=1}^{n} t_i^2\right)=$$

$$1-\frac{1}{\binom{n+1}{3}}\sum_{i=1}^{n}(s_i^2-2s_i t_i+t_i^2)=1-\frac{1}{\binom{n+1}{3}}\sum_{i=1}^{n}(s_i-t_i)^2=$$

$$1-\frac{6\sum_{i=1}^{n}(s_i-t_i)^2}{n\left(n^2-1\right)}$$

4.1. Definíció. *(Spearman-féle rangkorreláció) Legyen* $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ *minta,* s_i *az* x_i *rangja,* t_i *az* y_i *rangja, ekkor a*

$$\varrho(x,y)=1-\frac{6\sum_{i=1}^{n}(s_i-t_i)^2}{n\left(n^2-1\right)}$$

kifejezést az (x,y) *minta Spearman-féle rangkorrelációjának nevezzük.*

Legyen minden $s,t\in S_n$ permutáció esetén

$$f_\varrho(s,t)=\sum_{i=1}^{n}\left(s_i-t_i\right)^2$$

Eddig konzekvens módon d-vel jelöltük a metrikákat, viszont az f_ϱ függvény nem metrika az S_n szimmetrikus csoport elemein, viszont a $\sqrt{f_\varrho}$ függvény igen, mert megegyezik az euklideszi metrikával. Számítsuk ki a várható értékét. Az invariancia kihasználásával rögzítsük az egyik változót az identitásnak és legyen $F_\varrho(s)=f_\varrho(id,s)$, ekkor

$$E(f_\varrho) = \frac{1}{n!}\sum_s F_\varrho(s) = \frac{1}{n!}\sum_s\sum_{i=1}^{n}(i-s_i)^2 = \frac{1}{n!}\sum_{i=1}^{n}\sum_s\left(i^2 - 2is_i + s_i^2\right) =$$
$$\frac{1}{n!}\sum_{i=1}^{n}\left(i^2\sum_s 1 - 2i\sum_s s_i + \sum_s s_i^2\right) =$$
$$\frac{1}{n!}\sum_{i=1}^{n}\left(i^2 n! - 2i(n-1)!\sum_{j=1}^{n} j + (n-1)!\sum_{j=1}^{n} j^2\right) =$$
$$\frac{1}{n!}\sum_{i=1}^{n}\left(i^2 n! - 2i(n-1)!\frac{n(n+1)}{2} + (n-1)!\frac{n(n+1)(2n+1)}{6}\right) =$$
$$\frac{(n-1)n(n+1)}{6}$$

Így f_ϱ várható értéke $\frac{(n-1)n(n+1)}{6}$. Az tudjuk hogy az f_ϱ függvényből megfelelő transzformációval rangkorreláció alkotható, ekkor a transzformáltra az teljesül, hogy $\widehat{f_\varrho} = 1 - \frac{2f_\varrho}{\max_{s,t} f_\varrho}$ és tudjuk, hogy $\varrho = 1 - \frac{6f_\varrho}{n(n^2-1)}$, tehát $\varrho = \widehat{f_\varrho}$, ezért $\max_{s,t} f_\varrho = 2E(f_\varrho)$.

4.2. Kendall-féle rangkorreláció

A Kendall metrikánál megismert függvényből kiindulva, rangkorrelációs együtthatót állíthatunk elő a $\widehat{d_\tau} = 1 - \frac{2d_\tau}{\max_{s,t} d_\tau}$ transzformációval. Mivel $\max_{s,t} d_\tau = \frac{n(n-1)}{2} = \binom{n}{2}$, ezért $\widehat{d_\tau} = 1 - \frac{2d_\tau}{\binom{n}{2}} = \frac{\binom{n}{2} - 2d_\tau}{\binom{n}{2}}$. Legyen az inverzióban lévő párok száma n_d és a nem inverzióban lévő párok száma n_c, ami éppen $n_c = \binom{n}{2} - n_d$. Ezt behelyettesítve adódik, hogy $\widehat{d_\tau} = \frac{2(n_c - n_d)}{n(n-1)}$.

4.2. Definíció. *(Kendall-féle rangkorreláció) Legyen* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *minta,* s_i *az* x_i *rangja,* t_i *az* y_i *rangja, ekkor a*

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}$$

kifejezést az (x, y) minta Kendall-féle rangkorrelációjának nevezzük.

4.2.1. Kendall próba

Egy adott $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ minta esetén fontos kérdés annak eldöntése, hogy a mintában szereplő változók függetlenek-e. Legyen az együttes eloszlás $H(x, y)$, tegyük fel, hogy ez folytonos. Ekkor logikus azt a hipotézist vizsgálni, hogy ez felbomlik-e $H(x, y) = F(x)G(y)$ alakban. Legyen minden $1 \leq i < j \leq n$ esetén

$$Z(x_i, x_j, y_i, y_j) = \begin{cases} 1 & (x_i - x_j)(y_i - y_j) > 0 \\ -1 & (x_i - x_j)(y_i - y_j) \leq 0 \end{cases}.$$

A $\tau = 2P((x_i - x_j)(y_i - y_j) > 0) - 1$ a Kendall-féle τ, könnyen látszik, hogy ha x és y függetlenek, akkor $\tau = 0$, viszont ez fordítva nem mindig teljesül. Legyen $T = \sum_{i<j} Z(x_i, x_j, y_i, y_j)$, ekkor megmutatható, hogy a $\widehat{\tau} = \frac{2T}{n(n-1)}$ kifejezés torzítatlan becslése τ-nak, ahol $\widehat{\tau}$ éppen a Kendall-féle rangkorreláció, így $E(\widehat{\tau}) = \tau$. Legyen a nullhipotézis $H_0 : \tau = 0$, ekkor tetszőleges α terjedelem esetén a $H_1 : \tau > 0$ ellenhipotézis mellett a kritikus tartomány a $K = \{T \geq k_\alpha\}$ halmaz, ahol k_α-ra teljesül, hogy
$P(T \geq k_\alpha = \alpha$. Ha $H_1 : \tau < 0$ az ellenhipotézis, akkor $K = \{T \leq -k_\alpha\}$ a kritikus tartomány. Kétoldali ellenhipotézis esetén
$K = \{T \geq k_{\alpha_1}$ vagy $T \leq -k_{\alpha_1}\}$, ahol $\alpha = \alpha_1 + \alpha_2$ a próba terjedelme.

5. fejezet

Alkalmazások

5.1. A Mallows modell

Egy versenyen a zsűritagok a megjelent versenyzőket egy meghatározott szempont szerint rangsorolják, ha a versenyzők száma $n \in \mathbb{N}$, akkor tekintsük a zsűritagok által meghatározott rangsorokat egy-egy S_n-beli permutációnak. A Kendall-féle távolság segítségével hozzunk létre egy olyan P valószínűségi mértéket az S_n csoporton, amelyet egy $s_0 \in S_n$ pozícióparaméter és egy $\lambda \geq 0$ skalárparaméter definiál. Legyen minden $s \in S_n$ permutáció esetén

$$P(s) = \frac{e^{-\lambda d_\tau(s,s_0)}}{\sum_s e^{-\lambda d_\tau(s,s_0)}}.$$

Ahogy az s_0-tól vett távolságot csökkentjük, úgy ez a valószínűség exponenciálisan csökken. Ha a λ paramétert 0-nak választjuk, akkor egyenletes eloszlást kapunk. Tegyük fel, hogy a versenyzők "valódi" sorrendje s_0, feltehetjük, hogy $s_0 = id$. A zsűritagok páronként hasonlítják össze a versenyzőket és az $i < j$ versenyzők összehasonlításánál $p > \frac{1}{2}$ valószínűséggel ismerik fel a valódi sorrendet és $1 - p$ valószínűséggel tévednek. Ha az így

elvégzett $\binom{n}{2}$ összehasonlítás egy lineáris rendezést ad, akkor

$$P(s) = C(1-p)^{Inv(s)}p^{\binom{n}{2}-Inv(s)} = C'(\frac{p}{1-p})^{-Inv(s)} = C'\frac{p}{1-p}^{-d_\tau(id,s)}$$

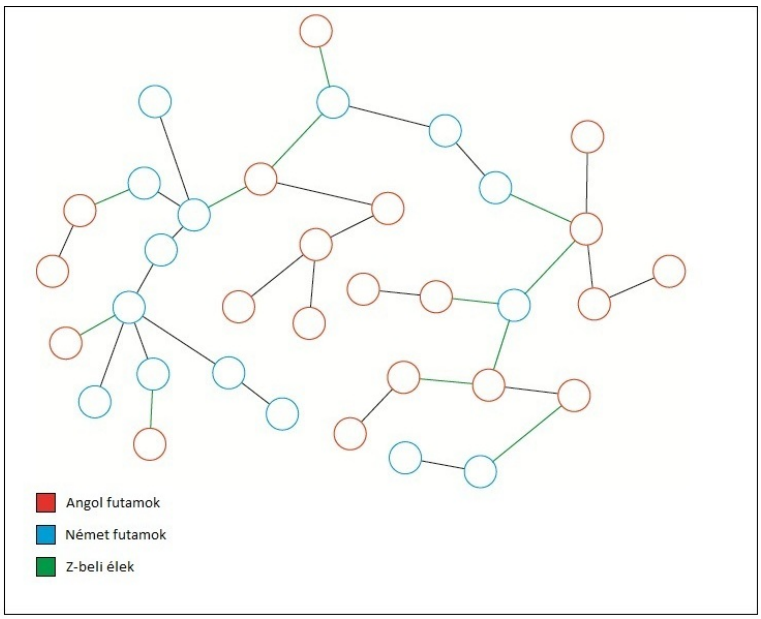
Azaz $\frac{p}{1-p} = e^\lambda$ választással éppen a Mallows modellt kapjuk.

5.2. Egy kétmintás probléma

Két megfigyelés során az egyik alkalommal egy $n \in \mathbb{N}$ elemű populáció esetén m_1, egy másik alkalommal m_2 mérést végeztünk. Legyen az első megfigyelés során a j-edik alkalommal, a populáció i-edik tagján végzett mérés eredménye $f_j(x_i)$ és a második megfigyelés során a j-edik alkalommal, a populáció i-edik tagján végzett mérés eredménye $g_j(x_i)$. Ezeket az adatokat nagyság szerint növekvő sorrenbe állítottuk és az
$f_j^*(x_1) < \cdots < f_j^*(x_n)$ sorrendben szereplő i-edik elemet X_{ij}-vel, míg a $g_j^*(x_1) < \cdots < g_j^*(x_n)$ sorrendben szereplő i-edik elemet Y_{ij}-vel jelöltük és ezt minden $j = 1, \ldots, m_1, m_2$ esetén elvégeztük. Ekkor minden j esetén egyértelműen létezik olyan $p_j, q_j \in S_n$ permutáció, amelyre
$p_j(f_j(x_i)) = f_j^*(x_i)$ és $q_j(g_j(x_i)) = g_j^*(x_i)$. Legyen $f_j(x) = (f_j(x_1) \ldots f_j(x_n))$ és $g_j(x) = (g_j(x_1) \ldots g_j(x_n))$, ekkor az
$f_1(x), f_2(x), \ldots, f_{m_1}(x), g_1(x), g_2(x), \ldots, g_{m_2}(x) \in \mathbb{R}^n$ adatok helyett használhatjuk a $p_1, p_2, \ldots, p_{m_1}, q_1, q_2, \ldots, q_{m_2} \in S_n$ rangsorokat. Legyen (S_n, d) metrikus tér és rajzoljuk fel a $P = \{p_1, p_2, \ldots, p_{m_1}\}$ és $Q = \{q_1, q_2, \ldots, q_{m_2}\}$ permutációk $G = (V, E)$ teljes gráfját. Ekkor $V = P \cup Q$ és legyen $c : E \to \mathbb{R}$ olyan költségfüggvény az élek halmazán, amelyre minden $uv \in E$ esetén $c(uv) = d(u, v)$. Keressük meg a G gráfban a legolcsóbb utakat tartalmazó feszítőfát a Prim algoritmus segítségével. Színezzük a P-beli permutációkat reprezentáló fabeli csúcsokat pirossal és a Q-belieket kékkel. Legyen Z a P-

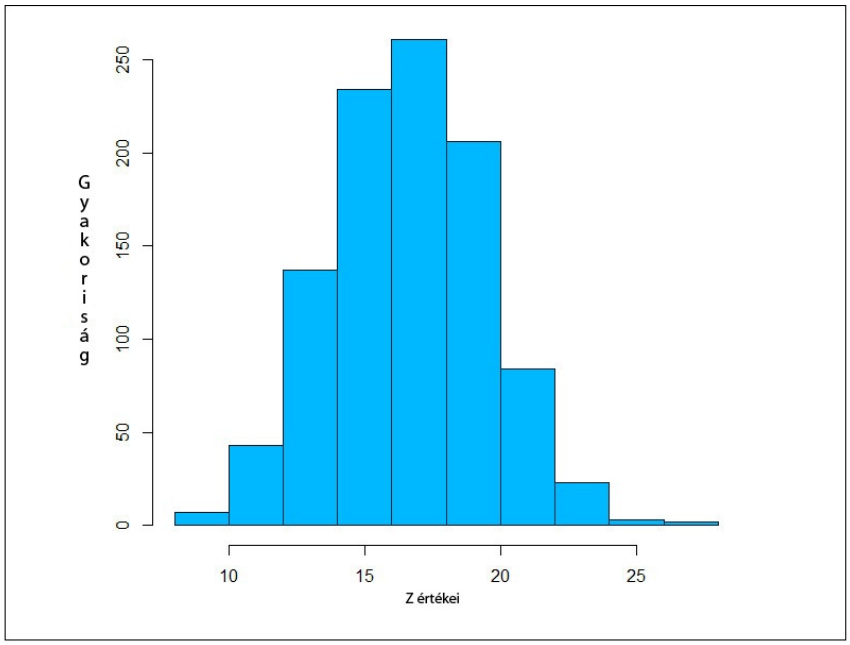
beli és Q-beli pontokat összekötő élek száma, Z eloszlását vizsgálva tudjuk igazolni, hogy a két vizsgált populáció mennyire hasonló, ha rögzítjük a fában szereplő éleket és egy véletlenszerű színezés esetén kiszámoljuk Z értékét, akkor elég sokszor elvégezve a szimulációt következtethetünk Z eloszlására.

Példa: Németországban és Angliában egy-egy lóversenyen megfigyeltük, hogy 10 ló hogyan teljesít egy évadon keresztül és kíváncsiak vagyunk arra, hogy a lovak mennyire hasonló vagy éppen mennyire különböző eredményeket futottak, tehát hogy befolyásolta a teljesítményüket a hazai illetve a külföldi pálya. Németországban egy idény során 15 futamon, illetve Angliában 20 futamon jegyeztük fel, hogy a lovak milyen időket futottak és az összegyűjtött adatok szerint minden egyes futamra felírtuk a lovak beérkezésének rangsorát. A mérési adatoknak megfelelően elkészítettük a miniámális költségű éleket tartalmazó feszítőfát.



A permutációkat reprezentáló gráf feszítőfája

Miután a fát kirajzoltuk és meghatároztuk a két színosztály közt haladó élek számát, ami a minta alapján 11 volt, a két színosztály méretének megfelelően 1000 szimuláció során újra és újra beszíneztük a fa csúcsait és minden esetben megfigyeltük, hogy hány él haladt a színosztályok között. Az eredményeinket egy hisztogramon ábrázoltuk.



Z Hisztogramja

Mivel a hisztogramról leolvasható, hogy az általunk kapott érték elég szélsőséges, azaz szignifikánsan eltér az átlagtól, így elutasíthatjuk azt a hipotézist, hogy a lóverseny eredménye független a helyszíntől, azaz lényegesen különbözik a németországi és az angliai versenyek kimenetele.

5.3. Robosztus regresszió

Célunk nemlineáris regresszió közelítése rangokkal. Legyen $f_\vartheta : X \to \mathbb{R}$ paraméteres függvénycsalád és $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ meg-

figyelés. Olyan $\vartheta \in \Theta$ értéket keresünk, amelyre y_i közel van $f_\vartheta(x_i)$-hez. A klasszikus megközelítés a legkisebb négyzetek módszerével találja meg a keresett ϑ értéket a

$$\sum_{i=1}^{n}(y_i - f_\vartheta(x_i))^2$$

kifejezés minimalizálásával. Észrevették, hogy bizonyos értékek esetén ez a képlet nagyon érzékeny, azaz, ha néhány x vagy y érték túl távol van a többitől, akkor az nagy hatással lehet a minimumra. Ezért ϑ értékét úgy kéne megválasztanunk, hogy $f_\vartheta(x_i)$ rangja a lehető legközelebb legyen y_i rangjához. Jelölje t_i az y_i rangját és s_{ϑ_i} az $f_\vartheta(x_i)$ rangját. Kereshetjük azt a ϑ paramétert, amelyre $d(t, s_\vartheta)$ minimális, ahol d tetszőleges metrika a permutációk halmazán.

Irodalomjegyzék

[1] Persi Diaconis: *Group Representations in Probability and Statistics. Inst. of Math. Statistics, Hayward, CA (1988).*

[2] John I. Marden: *Analyzing and Modeling Rank Data. Chapman and Hall, London (1995).*

[3] C.L. Malllows: *Non-null ranking models I. Biometrika 44, 114-130 (1957)*

[4] J.H Friedman, L.C. Rafsky: *Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-sample tests. Ann. Statist. 7, 697-717 (1979).*

[5] V.K. Rohatgi: *An Introduction to Probability Theory and Mathematical Statistiscs. John Wiley and Sons (1976)*