NYILATKOZAT

Név: Fodor Péter

ELTE Természettudományi Kar, szak: Matematika

NEPTUN azonosító: VAN30I

Szakdolgozat címe:
Grafikus Modellek és Kontingencia Táblák Elemzése

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, 2022.05.28

Fodor Péter

a hallgató aláírása

Grafikus Modellek és Kontingencia Táblák Elemzése

Fodor Péter

Matematika BSc, alkalmazott matematika szak

Szakdolgozat

Témavezető:

Dr. Csiszár Villő

Eötvös Loránd Tudományegyetem
Természettudományi Kar
Budapest, 2022.

Tartalomjegyzék

1. fejezet

Bevezetés

Rengetegszer előfordul, hogy bizonyos tulajdonságok szerint osztályozni, csoportosítani szeretnénk különböző objektumokat. Például nézhetjük egy osztályban a fiúk és lányok felvételi eredményeit, külön a magyart és matematikát. A diák neme (kétféle lehet) , magyar eredménye (ötféle) és matematika eredménye (ötféle) alapján az adatokat kigyűjtve egy $2 \times 5 \times 5$-ös táblázatot kapunk, ami alapján következtetéseket szeretnénk levonni. A dolgozat célja, hogy bevezetést adjon a többdimenziós táblázatok elemzésébe. A táblázat változóinak kapcsolatát a második fejezetben bemutatott függetlenség, illetve feltételes függetlenség segítségével próbáljuk értelmezni. A harmadik fejezetben a táblázat változóit, mint egy gráf csúcsait ábrázoljuk és azt részletezzük, hogy egy ilyen ábrázolás segítségével mit lehet leolvasni a változók kapcsolatáról. A negyedik fejezet a dekomponálható gráfok néhány tulajdonságát járja körül, amit azzal a céllal vezetünk be, hogy a táblázat alapján készített modellhez maximum likelihood becslést készítsünk. Természetesen egy táblázathoz többféle modellt is készíthetünk, azonban az ötödik fejezetben mutatunk egy lehetséges módszert, hogy a sok modell közül melyiket válasszuk ki. A hatodik fejezetben egy példán keresztül foglaljuk össze a korábbi fejezetek tartalmát. A hetedik fejezet a modell helyességének elemzését aszimptotikusan, illetve konkrét számolások útján vizsgálja.

A dolgozat elkészüléséért köszönetet mondok Dr. Csiszár Villő tanáromnak odaadó munkájáért és hasznos tanácsaiért. Hálával tartozom még családomnak, hogy tanulmányaim során végig támogatni tudtak. Külön köszönet illeti barátnőmet a sok bíztatásért és türelméért.

2. fejezet

Feltételes függetlenség

Ez a fejezet Steffen L. Lauritzen [2] könyvének harmadik fejezete alapján íródott.

2.1. Függetlenség definíciója

2.1.1. Definíció. Legyenek A, B események, ekkor azt mondjuk, hogy A és B függetlenek, ha

$$P(AB) = P(A)P(B). \tag{2.1}$$

Tehát annak a valószínűsége, hogy mindkettő egyszerre bekövetkezik, megegyezik az egyes események valószínűségének a szorzatával. Úgy is értelmezhetjük, hogy A és B események függetlensége fennáll, ha nem végzünk el előre semmilyen más megfigyelést, jelölje A és B függetlenségét $A \perp\!\!\!\perp B$.

2.1.2. Definíció. Legyenek A, B események, ahol $P(B) > 0$, ekkor az A esemény B feltételre vonatkozó feltételes valószínűsége

$$P(A \mid B) = \frac{P(AB)}{P(B)}. \tag{2.2}$$

Szemléletesen $P(A \mid B)$ jelentése, az A esemény bekövetkezésének valószínűsége, ha előre tudjuk, hogy B esemény bekövetkezett. Adódik, hogy ha A és B események függetlenek, akkor $P(A) = P(A \mid B)$. Most rátérünk a feltételes függetlenség fogalmára.

2.1.3. Definíció. Legyenek A, B, C események, ahol $P(C) > 0$, ekkor azt mondjuk, hogy A és B feltételesen függetlenek a C esemény mellett, ha

$$P(AB \mid C) = P(A \mid C)P(B \mid C). \tag{2.3}$$

2.1.4. Definíció. Legyenek X, Y, Z diszkrét valószínűségi változók, ekkor azt mondjuk, hogy X és Y feltételesen függetlenek feltéve Z, ha

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z). \tag{2.4}$$

Bevezetjük az $A \perp\!\!\!\perp B \mid C$ jelölést, itt $C = \emptyset$ esetén a sima $A \perp\!\!\!\perp B$ függetlenséget kapjuk vissza. Definiáljuk a feltételes függetlenséget abszolút folytonos valószínűségi változók esetén is. Egyszerűbb jelölés érdekében legyen f a megadott változók szerinti megfelelő sűrűségfüggvény.

2.1.5. Definíció. Legyenek X, Y, Z együttesen abszolút folytonos valószínűségi változók a Lebesgue mértékre nézve, ekkor azt mondjuk, hogy X és Y feltételesen függetlenek feltéve Z, ha a feltételes sűrűségfüggvényük

$$f(x, y \mid z) = f(x \mid z) f(y \mid z)$$

alakban áll elő.

2.1.1. Állítás. *Legyenek X, Y, Z abszolút folytonos valószínűségi változók, ekkor:*

(i) $X \perp\!\!\!\perp Y \mid Z \iff f(x, y, z) = \frac{f(x,z) f(y,z)}{f(z)}$

(ii) $X \perp\!\!\!\perp Y \mid Z \iff f(x \mid y, z) = f(x \mid z)$

(iii) $X \perp\!\!\!\perp Y \mid Z \iff f(x, y \mid z) = f(x \mid z) f(y \mid z)$

(iv) $X \perp\!\!\!\perp Y \mid Z \iff f(x, y, z) = h_1(x, z) h_2(y, z)$ *megfelelő* h_1, h_2 *függvényekkel*

(v) $X \perp\!\!\!\perp Y \mid Z \iff f(x, y, z) = f(x \mid z) f(y, z)$

Bizonyítás. Az összes ekvivalencia átrendezésekkel adódik 2.1.5 definícióból. Megmutatjuk az első két esetet:

(i) $f(x, y, z) = f(x, y \mid z) f(z) = f(x \mid z) f(y \mid z) f(z) = \dfrac{f(x, z)}{f(z)} \dfrac{f(y, z)}{f(z)} f(z) = \dfrac{f(x, z) f(y, z)}{f(z)}$

(ii) $f(x \mid y, z) = \dfrac{f(x, y, z)}{f(y, z)} = \dfrac{f(x, y \mid z) f(z)}{f(y, z)} = \dfrac{f(x \mid z) f(y \mid z) f(z)}{f(y \mid z) f(z)} = f(x \mid z)$

□

2.2. Feltételes függetlenség alaptulajdonságai

A következő tulajdonságok alapvető fontosságúak, a feltételes függetlenség axiómáinak is nevezhetnénk őket. Legyenek X, Y, Z diszkrét valószínűségi változók, f pedig legyen a megadott változók szerinti megfelelő sűrűségfüggvény a számláló mértékre nézve.

(C1) Ha $X \perp\!\!\!\perp Y \mid Z$, akkor $Y \perp\!\!\!\perp X \mid Z$.

(C2) Ha $X \perp\!\!\!\perp Y \mid Z$ és $U = h(X)$, akkor $U \perp\!\!\!\perp Y \mid Z$.

(C3) Ha $X \perp\!\!\!\perp Y \mid Z$ és $U = h(X)$, akkor $X \perp\!\!\!\perp Y \mid (Z, U)$.

(C4) Ha $X \perp\!\!\!\perp Y \mid Z$ és $X \perp\!\!\!\perp W \mid (Z, Y)$, akkor $X \perp\!\!\!\perp (W, Y) \mid Z$.

Bizonyítás. (C1) és (C2) triviális. (C3): Tudjuk, hogy $f(x, y \mid z) = h_1(x, z)h_2(y, z)$. Mivel $u = h(x)$ így $f(x, y, z, u) = f(x, y, z)$, különben 0, ha $u \neq h(x)$. Ekkor tehát

$$f(x, y, z, u) = \mathbb{1}\{u = h(x)\}h_1(x, z)h_2(y, z) = h_1'(x, z, u)h_2'(y, z, u).$$

(C4): Definíció szerint ekkor

1. $f(x, y \mid z) = f(x \mid z)f(y \mid z)$
2. $f(x, w \mid z, y) = f(x \mid z, y)f(w \mid z, y)$

Innen meg kell mutatni, hogy $f(x, w, y \mid z) = f(x \mid z)f(w, y \mid z)$. Fejezzük ki az alábbi tagokat az előfeltevések segítségével.

$$\begin{aligned}
f(x \mid z) &= \frac{f(x, y \mid z)}{f(y \mid z)} = \frac{f(x, y, z)}{f(y, z)} \\
f(w, y \mid z) &= \frac{f(w, y, z)}{f(z)} = \frac{f(w \mid y, z)f(y, z)}{f(z)} = \frac{f(x, w \mid y, z)f(y, z)}{f(z)f(x \mid y, z)} = \frac{f(x, w, y, z)}{f(z)f(x \mid y, z)} \\
f(x \mid z)f(w, y \mid z) &= \frac{f(x, y, z)}{f(y, z)} \cdot \frac{f(x, w, y, z)}{f(z)f(x \mid y, z)} = \frac{f(x, w, y, z)}{f(z)} = f(x, w, y \mid z)
\end{aligned}$$

ami éppen (C4). □

Semi-graphoidnak nevezzük az olyan algebrai struktúrát, ami teljesíti (C1)-(C4)-et, úgy értelmezve, hogy X, Y, Z egy véges halmaz diszjunk részhalmazai és $U = h(X)$ jelentése, hogy $U \subseteq X$.

2.2.1. Példa. *Legyen $G(V, E)$ egy gráf és legyen $X, Y, Z \subseteq V$ részhalmazok. Ekkor $X \perp\!\!\!\perp Y \mid Z$ jelentse azt, hogy Z elválasztja az X csúcshalmazt Y-tól. Ebben az értelmezésben vizsgáljuk meg (C1)-(C4) állításait.*

(C1) Ha Z elválasztja X-et Y-tól, akkor Z elválasztja Y-t az X-től.

(C2) Ha Z elválasztja X-et Y-tól, akkor tetszőleges $U \subseteq X$ esetén fennáll, hogy Z elválasztja U-t Y-tól.

(C3) Ha Z elválasztja X-et Y-tól, akkor X-et és Y-t a $Z \cup U$ is elválasztja egymástól, ahol $U \subseteq X$ tetszőleges.

(C4) Ha Z elválasztja X-et Y-tól és $Z \cup Y$ elválasztja X-et W-től, akkor Z elválasztja X-et $W \cup Y$-tól is.

Meglepő, hogy a (C1)-(C4) axiómák hétköznapi nyelven is megfogalmazhatóak.

2.2.2. Példa. *Legyen most $X \perp\!\!\!\perp Y \mid Z$ jelentése, hogy tudva Z-t, az X elolvasásához, nem kell Y-t elolvasni és $U = h(X)$ értelmezése, hogy U az X egy fejezete.*

(C1) Tudva Z-t, ha X elolvasásához nem kell Y, akkor Y elolvasásához sem kell X.

(C2) Tudva Z-t,ha X elolvasásához nem kell Y, akkor X egy fejezetének elolvasásához sem kell Y.

(C3) Tudva Z-t,ha X elolvasásához nem kell Y, akkor X egy fejezetének elolvasása után is igaz, hogy Y nem kell X elolvasásához.

(C4) Tudva Z-t,ha X elolvasásához nem kell Y és elolvasva Y-t, X elolvasásához nem kell elolvasni W-t, akkor sem Y sem W nem kell X elolvasásához.

2.2.1. Állítás. *Legyenek X, Y, Z, W abszolút folytonos valószínűségi változók és tegyük fel, hogy létezik $f(x, y, z, w) > 0$ együttes sűrűségfüggvény.*

(C5) Ha $X \perp\!\!\!\perp Y \mid (Z, W)$ és $X \perp\!\!\!\perp Z \mid (Y, W)$ akkor $X \perp\!\!\!\perp (Y, Z) \mid W$.

Bizonyítás. (C5): Abszolút folytonos valószínűségi változók esetén a feltételes függetlenség (iv) tulajdonsága alapján:

$$X \perp\!\!\!\perp Y \mid (Z, W) \Rightarrow f(x, y, z, w) = a(x, z, w)b(y, z, w)$$
$$X \perp\!\!\!\perp Z \mid (Y, W) \Rightarrow f(x, y, z, w) = c(x, y, w)d(z, y, w)$$

ahol a, b, c, d a megfelelő függvények. Ekkor átrendezve kapjuk, hogy

$$c(x, y, w) = \frac{a(x, z, w)b(y, z, w)}{d(z, y, w)}.$$

Mivel csak a jobb oldal függ z értékétől, ezért rögzített $z = z_0$ mellett

$$c(x, y, w) = \hat{a}(x, w)\hat{b}(y, w).$$

Ezt visszahelyettesítve c helyébe

$$c(x, y, w)d(z, y, w) = \hat{a}(x, w)\hat{b}(y, w)d(z, y, w) = \hat{a}(x, w)\hat{d}(z, y, w).$$

Ami definíció szerint éppen (C5) tulajdonság. □

Graphoidnak nevezzük az olyan struktúrákat, amik (C1)-(C5) tulajdonságokat teljesítik.

2.3. Markov tulajdonságok

Ebben a fejezetben gráfok csúcsain értelmezett valószínűségi változókkal fogunk foglalkozni. Célunk jobban megvizsgálni a valószínűségi változókon értelmezett feltételes függetlenség és a gráfokon értelmezett elválasztás kapcsolatát. Először bevezetünk két jelölést.

$bd(v) = \{u \in V : uv \in E\}$. Azaz v csúcs szomszédai.

$cl(v) = bd(v) \cup \{v\}$. Azaz v és v csúcs szomszédai.

Ha $\alpha \in V$ egy csúcs a gráfon, de a valószínűségi változóra gondolunk, akkor X_α helyett csak α-t fogunk írni. Ha $A \subset V$ akkor pedig $X_{\alpha \in A}$ helyett csak A jelölést fogjuk használni.

2.3.1. Definíció. Legyen $G(V, E)$ irányítatlan gráf, a csúcsain értelmezett $X_{\alpha \in V}$ valószínűségi változókról azt mondjuk, hogy teljesítik

(P) a páros Markov tulajdonságot, ha $\forall(\alpha, \beta) \in V$ nem szomszédos csúcsok esetén

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\} \tag{2.5}$$

(L) a lokális Markov tulajdonságot, ha $\forall \alpha \in V$ esetén

$$\alpha \perp\!\!\!\perp V \setminus cl(\alpha) \mid bd(\alpha) \tag{2.6}$$

(G) a globális Markov tulajdonságot, ha $\forall(A, B, C) \subset V$ diszjunkt halmazok esetén, ahol C elválasztja A és B -t

$$A \perp\!\!\!\perp B \mid C. \tag{2.7}$$

Rögtön nézzük meg, hogy mi a kapcsolat ezen tulajdonságok között.

2.3.1. Állítás. *Minden irányítatlan gráfon értelmezett valószínűségi eloszlásra teljesül, hogy*

$$(G) \Longrightarrow (L) \Longrightarrow (P) \tag{2.8}$$

A bizonyításban csupán (C1)-(C4) állításait fogjuk használni, így minden semi-graphoidra érvényes.

Bizonyítás. $(G) \Longrightarrow (L)$ nyilvánvaló, hiszen 2.6 feltétele csupán speciális esete 2.7-nek.

$(L) \Longrightarrow (P)$, azaz tegyük fel 2.6 érvényességét . Ekkor $\beta \in V \setminus cl(\alpha)$ mert α, β nem szomszédosak.

$$bd(\alpha) \cup (V \setminus cl(\alpha) \setminus \{\beta\}) = V \setminus \{\alpha, \beta\}$$

Ezt és (C3) tulajdonságot használva, mivel $(V \setminus cl(\alpha) \setminus \{\beta\}) \subset V \setminus cl(\alpha)$

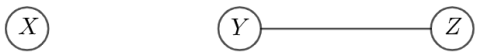
$$\alpha \perp\!\!\!\perp V \setminus cl(\alpha) \mid bd(\alpha) \cup (V \setminus cl(\alpha) \setminus \{\beta\}) \Longrightarrow$$
$$\alpha \perp\!\!\!\perp V \setminus cl(\alpha) \mid V \setminus \{\alpha, \beta\}$$

Mivel $\beta \in V \setminus cl(\alpha)$ így (C2) miatt kapjuk, hogy $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$, ami éppen 2.5. □

Általában visszafelé nem igaz, erre egyszerű példákat lehet gyártani.

2.3.2. Példa. *Legyen $X = Y = Z$ és $P(X = 1) = P(X = 0) = \frac{1}{2}$ és tekintsük 2.1 ábrát. Ekkor a páros Markov teljesül, de a lokális Markov nem teljesül.*

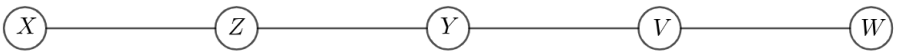
Hiszen Z-t ismerve X, Y konstansok és konstansok függetlenek, így (P) teljesül. Azonban $X \perp\!\!\!\perp (Y, Z)$ nem fog teljesülni, vagyis (L) sérül.



2.1. ábra. (P) de nem (L)

2.3.3. Példa. *Tekintsük 2.2 ábrán látható gráfot.* $X \perp\!\!\!\perp W$ *és* $P(X = 1) = P(X = 0) = P(W = 1) = P(W = 0) = \frac{1}{2}$, *továbbá* $X = Z, V = W, Y = ZV$. *Ekkor a lokális Markov teljesül, mert minden csúcsot egyértelműen meghatároznak a szomszédai, de a globális nem:*

Nézzük $X \perp\!\!\!\perp W \mid Y$ *esetét,* $X = 1, Y = 0, W = 0$ *helyettesítéssel.*
Ekkor $P(X = 1, W = 0 \mid Y = 0) = \frac{1}{3}$, *azonban* $P(X = 1 \mid Y = 0)P(W = 0 \mid Y = 0) = \frac{1}{3} \cdot \frac{2}{3}$.



2.2. ábra. (L) de nem (G)

2.4. Markov tulajdonságok ekvivalenciája

Szeretnénk biztosítani a visszafele irányt is, ehhez tekintsük az alábbi tulajdonságot, ami már garantálni fogja a Markov tulajdonságok ekvivalenciáját. Az alábbi tulajdonság a (C5) analógiájára épít.

$$A \perp\!\!\!\perp B \mid C \cup D \text{ és } A \perp\!\!\!\perp C \mid B \cup D \text{ akkor } A \perp\!\!\!\perp B \cup C \mid D \tag{2.9}$$

2.4.1. Tétel. ***(Pearl és Paz)*** *Ha egy valószínűségi eloszlásra teljesül 2.9 tulajdonság az alaphalmaz minden* A, B, C, D *diszjunkt részhalmazára, akkor*

$$(G) \Longleftrightarrow (L) \Longleftrightarrow (P). \tag{2.10}$$

Bizonyítás. Megmutatjuk, hogy (P)$\Longrightarrow$(G). Indukciót fogunk alkalmazni visszafelé a C elválasztó halmaz mérete szerint. Feltehető, hogy A, B nem üres halmazok, ekkor $|C| \leq |V| - 2$. Ha $|C| = |V| - 2$ akkor A és B is egy elemű halmaz, a kérdéses feltételes függetlenséget (P) garantálja.

Tegyük fel, hogy $|C| = n < |V| - 2$ és minden n-nél nagyobb méretű szeparáló C halmazra igaz az állítás. Vizsgáljuk először $A \cup B \cup C = V$ esetét. Ekkor feltehető, hogy $|A| \geq 2$. Vegyünk egy $\alpha \in A$ csúcsot. Ekkor (C2) és (C3) tulajdonságokat alkalmazva $C \cup \{\alpha\}$ elválasztja $A \setminus \{\alpha\}$ és B halmazt, továbbá $C \cup (A \setminus \{\alpha\})$ elválasztja α és B halmazt. Mivel az elválasztó halmaz mérete mindkét esetben legalább $n + 1$ így az indukciós feltevés miatt

$$B \perp\!\!\!\perp A \setminus \{\alpha\} \mid C \cup \{\alpha\}$$
$$B \perp\!\!\!\perp \alpha \mid C \cup A \setminus \{\alpha\}.$$

Innen 2.9-et alkalmazva megfelelő helyettesítéssel $B \perp\!\!\!\perp A \mid C$ adódik. Most nézzük $A \cup B \cup C \subset V$ esetet. Válasszunk $\alpha \in V \setminus (A \cup B \cup C)$ csúcsot. Ekkor $C \cup \{\alpha\}$ is elválasztja A és B halmazt, indukciós feltevés miatt, ekkor $A \perp\!\!\!\perp B \mid C \cup \{\alpha\}$. Két eset fordulhat elő

(i) $A \cup C$ elválasztja $\{\alpha\}$-tól B-t, azaz $\{\alpha\} \perp\!\!\!\perp B \mid A \cup C$ indukciós feltevés miatt.

(ii) $B \cup C$ elválasztja $\{\alpha\}$-tól A-t. azaz $\{\alpha\} \perp\!\!\!\perp A \mid B \cup C$ indukciós feltevés miatt.

Az első esetben 2.9 miatt $A \cup \{\alpha\} \perp\!\!\!\perp B \mid C$, ahonnan (C2) alkalmazásával $A \perp\!\!\!\perp B \mid C$ is igaz. A második eset szimmetria miatt ugyanígy belátható.

□

2.4.2. Definíció. Legyen $G(V, E)$ gráf, azt mondjuk, hogy $K \subseteq V$ klikk, ha K teljes részgráf és nem tudunk úgy $v \in V \setminus K$ csúcsot hozzávenni K-hoz, hogy teljes részgráf maradjon. Jelölje K_G a G gráf klikkjeinek a halmazát.

2.4.3. Definíció. Egy P valószínűségi mérték az $\mathbf{X}$ valószínűségi változókon faktorizálódik a G gráf szerint, ha $\forall K \in K_G$ klikkhez $\exists \psi_K$ nemnegatív függvény, ami csak a klikkben lévő csúcsokhoz tartozó valószínűségi változóktól függ, továbbá $\exists \mu$ szorzatmérték $\mathbf{X}$-en, hogy a P sűrűségfüggvénye a μ-re nézve az alábbi alakot ölti

$$f(x) = \prod_{K \in K_G} \psi_K(x). \tag{2.11}$$

Ha P faktorizálódik, akkor azt mondjuk, hogy P rendelkezik az (F) tulajdonsággal.

2.4.1. Állítás. *Minden irányítatlan G gráfra és X valószínűségi változóra*

$$(F) \Longrightarrow (G) \Longrightarrow (L) \Longrightarrow (P) \tag{2.12}$$

Bizonyítás. Elég (F) $\Longrightarrow$(G) implikációt belátni. Legyen $(A, B, C) \subset V$ tetszőleges diszjunkt részhalmazok, úgy, hogy C elválasztja A és B-t. Jelölje A' azon összefüggőségi komponenseit a $V \setminus C$ halmaznak, melyek tartalmazzák A-t, továbbá legyen $B^* = V \setminus (A' \cup C)$. Ha $a \in A$ és $b \in B$, akkor nincs olyan összefüggőségi komponense $V \setminus C$ halmaznak, hogy abban (a, b) egyszerre benne legyen, mert C elválasztó halmaz. Ha $K \in K_G$ klikk, akkor $K \subset (A' \cup C)$ (kizáró) vagy $K \subset (B^* \cup C)$. Legyen $K_{A' \cup C}$ azon klikkek, amik $(A' \cup C)$ halmazban vannak. Ekkor (F) miatt

$$f(x) = \prod_{K \in K_G} \psi_K(x) = \prod_{K \in K_{A' \cup C}} \psi_K(x) \prod_{K \in K_G \setminus K_{A' \cup C}} \psi_K(x) = h_1(A' \cup C) h_2(B^* \cup C) \Longleftrightarrow A' \perp\!\!\!\perp B^* \mid C \tag{2.13}$$

ahol az ekvivalenciát a feltételes függetlenség (iv) tulajdonsága adja. □

Kimondunk bizonyítás nélkül egy tételt, amely az (F) és a Markov tulajdonságok ekvivalenciájára ad szükséges és elégséges feltételt.

2.4.4. Tétel. ***(Hammersley és Clifford)*** *Legyen P valószínűségi eloszlás, aminek f sűrűségfüggvénye pozitív és folytonos. A G irányítatlan gráfra nézve P akkor és csak akkor tudja a páros Markov tulajdonságot, ha faktorizálódik G szerint.*

3. fejezet

Grafikus modellek

Ez a fejezet Christopher M. Bishop [3] könyvének nyolcadik fejezete és az ezt feldolgozó Philipp Henning Probabilistic Machine Learning [7], [8] interneten megtalálható előadásai alapján íródott.

Ebben a fejezetben olyan modellek bemutatásával foglalkozunk, melyek a feltételes függetlenséget próbálják szemléltetni, nyomon követni. Két eltérő modellt vizsgálunk, irányítatlan, majd utána irányított aciklikus gráfokat. Ki fog derülni, hogy mindkettőnek megvannak a maga előnyei és hátrányai, ráadásul valamilyen tekintetben egyik sem nyújt kielégítő megoldást.

3.1. Irányítatlan modellek

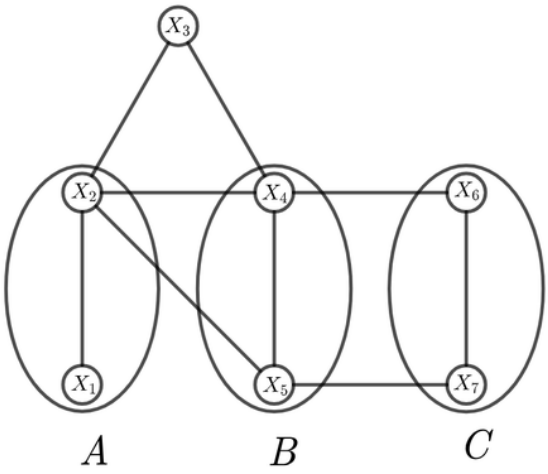
Kezdjük az irányítatlan esettel, itt az él két csúcs között nem tartalmazza azt a plussz információt, hogy melyik csúcs van hatással a másikra, csak azt tudjuk, hogy kapcsolat van köztük. Az a fő motivációnk, hogy olyan modellt készítsünk, amelyben a feltételes függetlenségi viszonyok közvetlenül leolvashatóak, ebben az előző fejezetben leírt (G) 2.7 globális Markov tulajdonság fog segítséget nyújtani..

3.1.1. Definíció. Legyen $G(V, E)$ irányítatlan gráf, melynek $\forall v_i \in V$ csúcsa egy X_i valószínűségi változót jelöl. Azt mondjuk, hogy G gráf és az X_i valószínűségi változók "Markov random field"-et azaz MRF-et alkotnak, ha $\forall A, B, C \subset V$ esetén ahol C elválasztja A és B-t, $X_A \perp\!\!\!\perp X_B \mid X_C$ teljesül, ahol $X_Z = \{X_i \in Z\}$.

A definíció éppen 2.7 tulajdonságon alapszik.

3.1.2. Példa. *Nézzük a 3.1 ábrát:*

Ha úgy gondolunk erre a gráfra és X_i valószínűségi változókra, hogy ezek MRF-et alkotnak, akkor teljesülnie kell például annak, hogy $X_A \perp\!\!\!\perp X_C \mid X_B$, hiszen B jelen esetben egy A, C elválasztó halmaz, mert minden út, ami $X_i \in A$ és $X_j \in C$ között megy, át kell hogy menjen B beli csúcson. További kérdések lehetnek, hogy A és C a legnagyobb halmaz, melyre a feltételes függetlenség fennáll? Vagy B a lehető legkisebb halmaz, mely elválasztja A és C-t?



3.1. ábra. $A \perp\!\!\!\perp C \mid B$ hiszen minden út A-ból C-be tartalmaz B-beli csúcsot

Irányítatlan esetben tehát két csúcs, vagy akár két diszjunkt halmaz feltételes függetlenségét leolvashatjuk a gráfról. Legyen $A, B, C \subset V$ diszjunkt halmazok. Ha minden $\alpha \in A$ és $\beta \in B$ csúcs között lévő út tartalmaz $\gamma \in C$-beli csúcsot, akkor $A \perp\!\!\!\perp B \mid C$ teljesül.

Mit tudunk mondani az együttes eloszlásról? Tekintsük ismét az 3.1 ábrát. Kezdjünk egy megfigyeléssel: x_1 és x_2 csúcsok között megy él, vagyis ők akkor is függnek egymástól, ha a gráf minden más csúcsát már megfigyeltük. Viszont x_1 és x_3 esetében, minden más csúcsot megfigyelve már függetlenek lesznek. Összefoglalva ha x_i, x_j nem szomszédos csúcsok, feltehető, hogy $i = 1$, $j = 2$, ekkor

$$P(x_1, x_2 \mid x_3, \ldots, x_n) = P(x_1 \mid x_3, \ldots, x_n)P(x_2 \mid x_3, \ldots, x_n).$$

Ez valójában a páros Markov tulajdonság 2.5. Mivel egy C klikkben minden csúcs szomszédos, ezért és az előbbi megfigyelés alapján elmondható, hogy ha egy $P(X)$ valószínűségi eloszlás teljesíti 2.4.4 tétel feltételeit, akkor az együttes eloszlás az alábbi alakot ölti:

$$P(X_1 = x_1, \ldots X_n = x_n) = \frac{1}{Z} \prod_{K \in K_G} \psi_K(x_k) \tag{3.1}$$

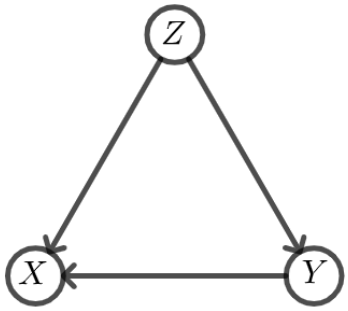
ahol Z normalizáló konstans, hogy valóban valószínűségi eloszlást kapjunk.

3.2. Irányított modellek

Legyen egy $D(V, E)$ irányított gráf, melynek a csúcsai egy-egy diszkrét valószínűségi változót jelentenek, az élek pedig a valószínűségi változók közötti összefüggéseket fogják reprezentálni. Így a gráf segítségével a valószínűségi változók együttes eloszlását tudjuk szemléltetni. Vegyünk egy valószínűségi eloszlást, három valószínűségi változóval $P(x, y, z)$. Ekkor szorzatra bonthatjuk:

$$P(x, y, z) = P(x \mid y, z)P(y, z) = P(x \mid y, z)P(y \mid z)P(z). \tag{3.2}$$

Természetesen ezt a felbontást másmilyen sorrendben is megtehettük volna. Erről még a későbbiekben szót fogunk ejteni. Rendeljünk hozzá egy gráfot ehhez az eloszláshoz, a következő módon:



3.2. ábra. 3.2 együttes eloszláshoz tartozó gráf.

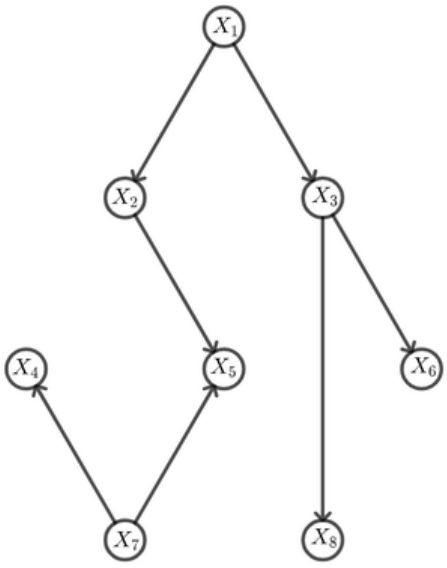
Vagyis x csúcsba befut él y és z-ből is, mert az x-hez tartozó $P(x \mid y, z)$ tényezőben x vizsgálatánál kondícionáltunk a másik két változóra. Ugyanígy $P(y \mid z)$ tényező felel a z és y közé írt élnek. Azt mondjuk, hogy y szülője a z csúcs, ha van irányított él z-ből y-ba és azt mondjuk, hogy y gyereke az x csúcs, ha van irányított él y-ból x-be. Emlékeztetünk, hogy 3.2 képletben a a változók más sorrendjében való felírásával egy teljesen más gráfot kaptunk volna.

Az előző gondolatmenetet követve el tudunk készíteni n darab valószínűségi változóhoz tartozó gráfot is.

$$P(x_1, x_2 \dots x_n) = P(x_1 \mid x_2 \dots x_n) \dots P(x_{n-1} \mid x_n) P(x_n) \tag{3.3}$$

Az ehhez tartozó gráf egy teljes gráf lesz, hiszen bármely két csúcs között lesz ekkor irányított él. Ilyen modell készítésekor az élek hiánya árul el információt az együttes eloszlásról, így ha teljes gráfot kapunk, nem jutunk előrébb.

Most haladjunk fordítva, egy adott gráf alapján készítsük el az együttes eloszlást.



3.3. ábra. 3.4 együttes eloszláshoz tartozó gráf.

Azt kapjuk, hogy az együttes eloszlás az alábbi alakot ölti

$$P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_7)P(x_5 \mid x_2, x_7)P(x_6 \mid x_3)P(x_7)P(x_8 \mid x_3) \tag{3.4}$$

Valójában tetszőleges aciklikus irányított gráf esetén

$$P(x_1, x_2 \dots x_n) = \prod_{i=1}^{n} P(x_i \mid pa_{x_i}) \tag{3.5}$$

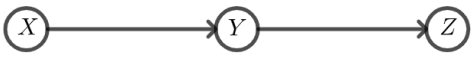
ahol pa_{x_i} jelöli az x_i csúcs szülőit, azaz azon csúcsokat, melyekből él fut x_i csúcsba.

Az irányított modellek nagy előnye tehát, hogy az együttes eloszlást nagyon könnyű leolvasni egy az adott gráfról. Most nézzük meg, hogy mit tudunk mondani a feltételes függetlenségről. Tekintsünk három egyszerű példát, amik különböző feltételes függetlenségi viszonyokat fognak tartalmazni.

3.3. Három alap modell

Kiemelünk három gráfot, amik különböző viszonyokat ábrázolnak. Ennél egyszerűbb érdekes modell nem készíthető, hiszen két csúcs esetén vagy van kapcsolat, vagy nincs, azonban három csúcsnál már meg tudunk különböztetni kapcsolat fajtákat. Az alábbi három példában X, Y, Z diszkrét valószínűségi változók.

3.3.1. Példa. *(i): "Folyó csúcs"*



3.4. ábra. Az Y "folyó csúcs" 3.6.

A hozzá tartozó együttes eloszlás

$$P(x, y, z) = P(x)P(y \mid x)P(z \mid y) \tag{3.6}$$

Ennek segítségével meghatározzuk $P(x, z \mid y)$ értékét.

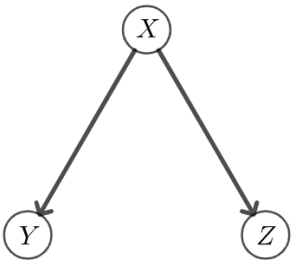
$$P(x, z \mid y) = \frac{P(x, y, z)}{P(y)} = \frac{P(x, y, z)}{\sum_{x,z} P(x)P(y \mid x)P(z \mid y)} =$$

$$= \left(\frac{P(x)P(y \mid x)}{\sum_{x} P(x)P(y \mid x)} \right) \left(\frac{P(z \mid y)}{\sum_{z} P(z \mid y)} \right)$$

Ahol $\sum_{x,z}$ jelentése, hogy X és Z minden lehetséges értékére behelyettesítünk és vesszük az összeget.

Kaptuk tehát, hogy ha már Y értékét ismerjük, akkor az együttes eloszlás szorzatra bomlik, azaz $X \perp\!\!\!\perp Z \mid Y$ reláció fennáll, de $X \perp\!\!\!\perp Z$ nem teljesül.

3.3.2. Példa. *(ii):"Nyíló csúcs"*



3.5. ábra. Az X "nyíló csúcs"3.7.

A hozzá tartozó együttes eloszlás

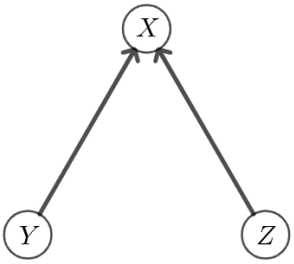
$$P(x, y, z) = P(x)P(y \mid x)P(z \mid x) \tag{3.7}$$

Ennek segítségével meghatározzuk $P(y, z \mid x)$ értékét.

$$P(y, z \mid x) = \frac{P(x, y, z)}{P(x)} = \frac{P(x)P(y \mid x)P(z \mid x)}{\sum\limits_{y,z} P(x)P(y \mid x)P(z \mid x)} = \left(\frac{P(x)P(y \mid x)}{\sum\limits_{y} P(x)P(y \mid x)}\right)\left(\frac{P(z \mid x)}{\sum\limits_{z} P(z \mid x)}\right)$$

Kaptuk tehát, hogy ha már X értékét ismerjük, akkor az együttes eloszlás szorzatra bomlik, azaz $Y \perp\!\!\!\perp Z \mid X$ reláció fennáll, de $Y \perp\!\!\!\perp Z$ nem teljesül.

3.3.3. Példa. *(iii): "Záró csúcs"*



3.6. ábra. Az X "záró csúcs" 3.8.

A hozzá tartozó együttes eloszlás

$$P(x, y, z) = P(x \mid y, z)P(y)P(z) \tag{3.8}$$

Ennek segítségével meghatározzuk $P(Y, Z \mid X)$ értékét.

$$P(y, z \mid x) = \frac{P(x, y, z)}{P(x)} = \frac{P(x \mid y, z)P(y)P(z)}{\sum\limits_{y,z} P(x \mid y, z)P(y)P(z)}$$

Ahol $P(x \mid y, z)$ tag mutatja, hogy nem tudjuk szorzattá bontani X ismeretében az együttes eloszlást, így $Y \perp\!\!\!\perp Z \mid X$ nem teljesül, azonban

$$P(y, z) = \sum_{x} P(x \mid y, z)P(y)P(z) = P(y)P(z).$$

Azt kaptuk, hogy X ismerete nélkül viszont $Y \perp\!\!\!\perp Z$.

A példák azt próbálják érzékeltetni, hogy a feltételes függetlenségi viszonyok leolvasása már nem olyan egyszerű.

3.4. D-szeparáció

Most kifejezetten arra a kérdésre koncentrálunk, hogy egy adott $D(V, E)$ irányított gráfról hogyan lehet leolvasni a feltételes függetlenségi kapcsolatokat. Érdemes az alábbi állítást használva újra átgondolni az előző pont három példáját.

3.4.1. Definíció. Azt mondjuk, hogy egy $D(V, E)$ irányított gráfban egy irányítatlan értelemben vett st út blokkolt a $C \subset V$ szerint, ha $\exists v \in st$ út, hogy

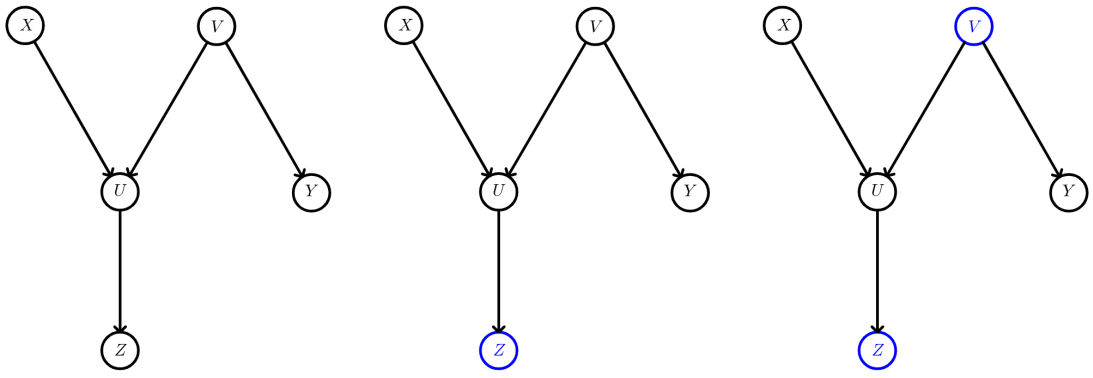
1. $v \in C$ és $v :\rightarrow v \rightarrow$ "folyó csúcs", vagy $\leftarrow v \rightarrow$ "nyíló csúcs".

2. $v \notin C$, v leszármazottai $\notin C$ és $v :\rightarrow v \leftarrow$ "záró csúcs".

3.4.1. Állítás. ***(Pearl 1988)*** *D-szeparáció. Legyen $D(V, E)$ irányított aciklikus gráf, melynek csúcsai valószínűségi változókat reprezentálnak. Legyenek $A, B, C \subseteq V$ diszjunkt halmazok. $A \perp\!\!\!\perp B \mid C$ akkor és csak akkor, ha minden irányítatlan értelemben vett út A és B között blokkolt C szerint.*

Az elnevezésben d az angol *directed* szóra utal. Egy példán fogjuk bemutatni, hogy ez a d-szeparáció milyen érzékenyen tud viselkedni, emiatt a függetlenséget leolvasni bonyolultabb ábráról nem egyszerű feladat.

3.4.2. Példa. *Mindegyik esetben $A \perp\!\!\!\perp B \mid C$ lesz a kérdés, ahol $A = \{x\}, B = \{y\}$ és C halmazt fogjuk csak változtatni. irányítatlan értelemben csak $(x - u - v - y)$ út van köztük, így csak ez lehet blokkolt.*



3.7. ábra. d-szeparáció érzékenysége

Az első esetben $C = \{\emptyset\}$.
Mivel sem u sem u leszármazottai sincsenek C-ben és u "záró csúcs", így az út blokkolt, vagyis $A \perp\!\!\!\perp B \mid C$, ami ebben az esetben csupán $A \perp\!\!\!\perp B$, teljesül.

A második esetben $C = \{z\}$.

Itt $v \notin C$ de v "nyíló", így ő nem fogja blokkolni az utat. Viszont $u \notin C$ és u "záró", ellentétben az előző esettel, most ő sem blokkolja az utat, mert van egy leszármazottja, $z : z \in C$, így most már $A \perp\!\!\!\perp B \mid C$ nem teljesül, hiszen az $(x - u - v - y)$ út nincs blokkolva. Ez az eredmény logikus, mert z csúcsra u van hatással és u csúcsra x, v együtt hatással van. Ha megtudunk valamit z-ről, az elárulhat információt v csúcsról, ami hatással van y-ra.

A harmadik esetben $C = \{z, v\}$

Az u csúcs helyzete nem változik, viszont $v \in C$ és v nyíló, ezért v blokkolja az utat, vagyis $A \perp\!\!\!\perp B \mid C$ ismét teljesül. Ez is logikus, hiszen z mostmár semmit nem tud elárulni nekünk v-ről, mert már v csúcsot is ismerjük, így ismét független lesz x és y.

Ez alapján felmerülhet a kérdés, hogy egy adott α csúcs esetén, mely csúcsokat kell felderíteni, hogy α már ne függjön a maradék gráftól. Megjegyezzük, hogy irányítatlan esetben erre a lokális Markov tulajdonság ad választ 2.6. Jelölje $ch(\alpha)$ azokat a csúcsokat, melyekbe megy irányított él α-ból.

3.4.3. Definíció. Egy $D(V, E)$ irányított gráfban adott α csúcs esetén, α "Markov takarójába" tartoznak a $pa(\alpha), ch(\alpha), pa'(\alpha)$ csúcsok, ahol $pa'(\alpha)$ azon β csúcsokat jelent, ami szülője $ch(\alpha)$-beli csúcsnak.

Vagyis α csúcsnak a szülőit, gyerekeit és gyerekeinek más szülőit kell ismerni, hogy α már feltételesen független legyen a maradék gráftól. A bizonyítás ötlete a következő. Az együttes eloszlás könnyedén leolvasható a gráfról és átírható a már említett 3.5 alakra. Feltehető, hogy $\alpha = x_1$, ekkor

$$p(x_1 \mid x_2, \dots, x_n) = \frac{p(x_1, x_2, \dots, x_n)}{\int p(x_1, \dots, x_n)\, dx_1} = \frac{\prod_{i=1}^{n} p(x_i \mid pa(x_i))}{\int \prod_{i=1}^{n} p(x_i \mid pa(x_i)\, dx_1}$$

Mivel x_1 szerepelni fog $p(ch(x_1) \mid pa(ch(x_1)))$ alakú tényezőkben, ezért ezek nem tudnak kiesni, ahogyan $p(x_1 \mid pa(x_1))$ tényező sem. Vagyis valóban, $p(x_1 \mid x_2, \dots, x_n)$ csak $pa(x_1), ch(x_1), pa'(x_1)$ csúcsoktól függ, azaz x_1 csúcs "Markov takarójától". Erre láthatunk egy példát a 3.8 ábrán.

3.5. Összehasonlítás

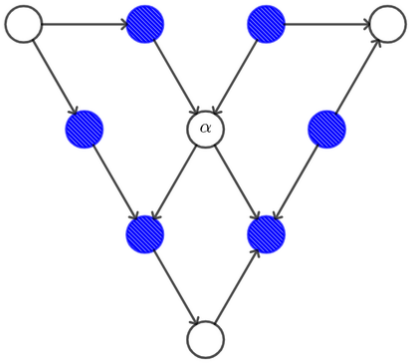
Bevezettünk két különböző módot a feltételes függetlenség grafikus ábrázolására, most az átmenetet vizsgáljuk közöttük, majd rátérünk arra a bizonyos hiányosságra, amit korábban előrevetítettünk. Először tekintsünk egy speciális irányított gráfot mely a 3.9 ábrán látható.

Az együttes eloszlás leolvasható 3.5 képlet alapján:

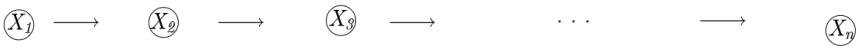
$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_2) \dots P(x_n \mid x_{n-1}) \tag{3.9}$$

Az élek irányítását elhagyva az együttes eloszlás a maximális klikkeket tartalmazza, ami jelen esetben csak a szomszédos csúcsok. Vagyis irányítatlan verziójában az együttes eloszlás

$$P(x_1, x_2, \dots, x_n) = \frac{1}{Z} \prod_{i=1}^{n-1} \psi_{i,i+1}(x_i, x_{i+1}) \tag{3.10}$$



3.8. ábra. Az α csúcs Markov takarója.

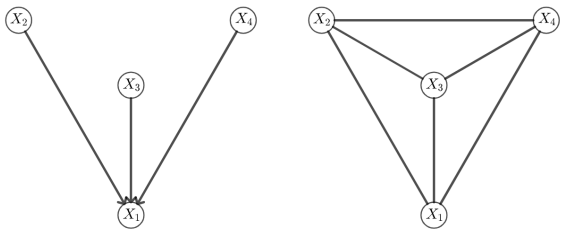


3.9. ábra. A 3.9 együttes eloszlásához tartozó irányított gráf.

Itt ψ függvény könnyen kitalálható:

$$\begin{aligned}
\psi_{1,2}(x_1, x_2) &= P(x_1)P(x_2 \mid x_1) \\
\psi_{2,3}(x_2, x_3) &= P(x_3 \mid x_2) \\
\psi_{3,4}(x_3, x_4) &= P(x_4 \mid x_3) \\
&\vdots \\
\psi_{n-1,n}(x_{n-1}, x_n) &= P(x_n \mid x_{n-1}).
\end{aligned}$$

Most $Z = 1$, hiszen tudjuk, hogy ψ függvények valószínűségi eloszlások. Ez egy kivételes eset volt, általában nem lesz elegendő csupán elhagyni az élek irányítását, ha át akarunk térni irányított modellből irányítatlanba. Tekintsük az alábbi ábrát:



3.10. ábra. A szülők házasítása ha irányított gráfból irányítatlanba térünk át.

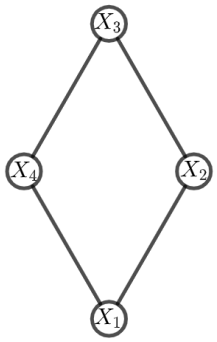
Az együttes eloszlás az irányított esetben a képlet alapján:

$$P(x_1, x_2, x_3, x_4) = P(x_1 \mid x_2, x_3, x_4)P(x_2)P(x_3)P(x_4) \tag{3.11}$$

Mivel $P(x_2)P(x_3)P(x_4)$ tag beolvasztható a $P(x_1 \mid x_2, x_3, x_4)$ tagba, ezért csupán egy darab $\psi_{1,2,3,4}(x_1, x_2, x_3, x_4)$ tényező lesz. Ebben szerepel mind a négy csúcs, vagyis létre kell hoznunk egy klikket, ami tartalmazza ezt a négy csúcsot, így az irányítatlan esetben az együttes eloszláshoz tartozó gráf egy teljes gráf lesz. Olyan mintha x_1 szülőit "házasítanánk", azonban ezzel

az eljárással gyakran túl sok élt húzunk be, így az értékes feltételes függetlenségi kapcsolatokat elveszíthetjük. Az előző 3.9 példában mivel minden csúcsnak csak egyetlen szülője volt, így nem volt szükség házasításra, szóval csak simán elhagyhattuk az élek irányítását. A lényeg, hogy ilyen módon általában információt veszítünk az áttérés során. Adódik a kérdés, hogy egyáltalán van esély arra, hogy egy adott irányított gráfhoz elkészítsünk egy irányítatlan gráfot, ami ugyanazt az információt tartalmazza? Irányítottból irányítatlanba ellenpélda lesz 3.8, az-az a "záró" csúcs esete. Ha el akarjuk hagyni az élek irányítását, akkor a házasítás miatt egy háromcsúcsú teljes gráfot kapunk.

3.5.1. Példa. *Nézzünk most egy irányítatlan gráfot, ez az alábbi függetlenségi kapcsolatot tartalmazza:* $X_1 \perp\!\!\!\perp X_3 \mid X_2 \cup X_4$ *és* $X_2 \perp\!\!\!\perp X_4 \mid X_1 \cup X_3$. *Akárhogyan is próbáljuk behúzni az irányított*



3.11. ábra. Irányítatlan gráf, melyhez nem tudunk irányítottat készíteni.

éleket aciklikus módon, lesz egy olyan csúcs, amibe két él is bemegy, ezzel egy "záró csúcs" jön létre, így egy olyan kapcsolatot tartalmaz a modell, ami az eredetiben nem volt benne.

Végül egy olyan példát mutatunk, ami sem irányított sem irányítatlan gráffal nem ábrázolható.

3.5.2. Példa. *Kétszer dobunk egy szabályos pénzérmével, ha a két dobás megegyezik, akkor csengetünk egy haranggal. A három esemény:*

$A:=$ *az első pénzdobás eredménye fej,* $P(A) = 0,5$

$B:=$ *a második pénzdobás eredménye fej,* $P(B) = 0,5$

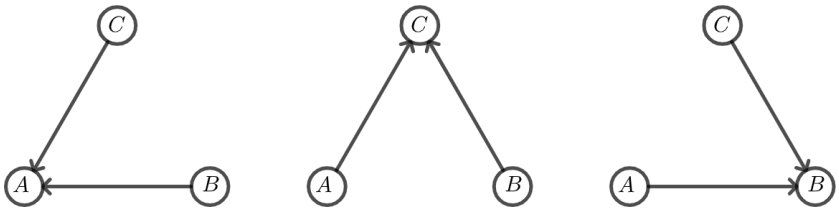
$C:=$ *megszólal egy harang,* $P(C) = 0,5$

Ekkor egyszerű számolásokkal igazolható, hogy

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0,5 \cdot 0,5}{0,5} = 0,5 = P(A)$$
$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{0,25}{0,25 + 0,25} = 0,5 = P(A)$$
$$P(C \mid A) = \frac{P(C \cap A)}{P(A)} = \frac{0,25}{0,5} = 0,5 = P(C)$$

Felhasználva az előbbi számolást és a szimmetria tulajdonságokat, írjuk fel a lehetséges faktorizációit az együttes eloszlásnak.

1. $P(A, B, C) = P(A \mid B, C)P(B \mid C)P(C) = P(A \mid B, C)P(B)P(C)$
2. $P(A, B, C) = P(C \mid B, A)P(B \mid A)P(A) = P(C \mid B, A)P(B)P(A)$
3. $P(A, B, C) = P(B \mid A, C)P(A \mid C)P(C) = P(B \mid A, C)P(A)P(C)$



3.12. ábra. A lehetséges faktorizációkhoz tartozó irányított gráfok.

Egyik verzió sem tartalmazza az összes feltételes függetlenségi viszonyt, az irányítatlan esetben pedig a házasítás miatt teljes gráfot kapunk, ami szintén nem ad kielégítő megoldást.

4. fejezet

Gráf dekompozíció

Ez a fejezet Steffen L. Laurizten [2] könyvének második fejezte alapján íródott.

4.1. Dekompozíció definíció

Egy irányítatlan G gráf csúcsait, melyek valószínűségi változókat reprezentálnak, két csoportba tudjuk osztani, aszerint, hogy a csúcs diszkrét valószínűségi változóra, vagy folytonosra utal. Jelölje Δ a diszkrét és Γ a folytonos csúcsokat.

4.1.1. Definíció. Egy $G(V, E)$ irányítatlan gráf $(A, B, C) \subseteq V$ diszjunkt részhalmazai erős dekompozíciója G-nek, ha

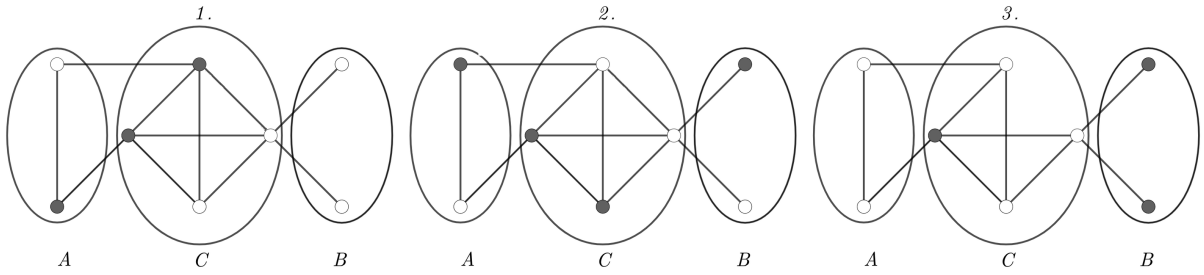
1. $A \cup B \cup C = V$
2. C elválasztja A és B halmazokat.
3. C teljes részgráf.
4. $C \subseteq \Delta$ vagy $B \subseteq \Gamma$

Ha ezek teljesülnek, akkor azt mondjuk, hogy $(A, B, C) \subseteq V$ halmazok erősen dekomponálják G gráfot a $G_{A\cup C}$ és $G_{B\cup C}$ részgráfokra. Ha csak a (4) nem teljesül, akkor a dekompozíció gyenge.

Egy G gráfról azt mondjuk hogy tiszta, ha Δ üres, (kizáró) vagy Γ üres. Ekkor G egy dekompozíciója pontosan akkor erős, ha gyenge, mert a (4) tulajdonság rögtön teljesül.

4.1.2. Példa. *Tekintsük a 4.1 ábrát, ahol az üres csúcs jelenti a folytonos valószínűségi változót, a teli pedig a diszkrétet.*

Az 1. *gráfnak az* (A, B, C) *erős dekompozíciója,* 2. *gráfnak csupán gyenge, hiszen (4) feltétel nem teljesül,* 3. *nem is dekompozíció, mert* C *nem teljes részgráf.*

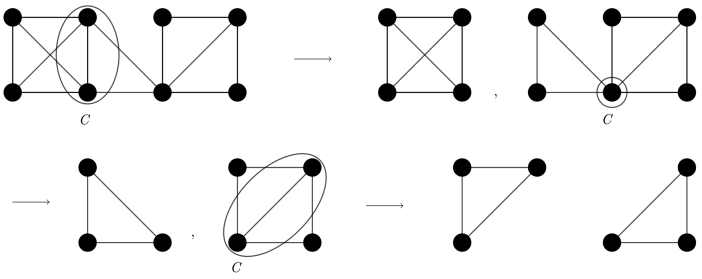


4.1. ábra. Példa a dekompozíció fajtáira.

Egy dekomponáló halmazról megengedjük, hogy üres legyen. Egy dekompozíciót valódinak mondunk, ha (A, B, C) halmazok közül, sem A sem B nem üres. Ha a C elválasztó halmaz üres, akkor A és B halmazok különböző összefüggőségi komponensben vannak, hiszen az üres halmaz elválasztja őket egymástól.

4.1.3. Definíció. Egy irányítatlan $G(V, E)$ gráfról azt mondjuk, hogy dekomponálható, ha G teljes gráf, vagy akkor, ha létezik $(A, B, C) \subset V$ valódi dekompozíciója, ami felbontja $G_{A\cup C}$ és $G_{B\cup C}$ dekomponálható gráfokra.

Vagyis egy dekomponálható G gráf felbontható a maximális teljes részgráfjaira, azaz klikkjeire.



4.2. ábra. Dekomponálható gráf felbomlik a klikkjeire.

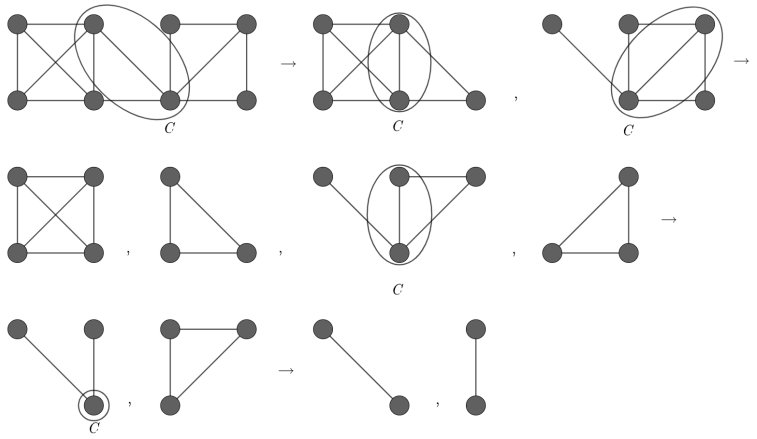
Természetesen egy dekomponálható gráfnak többféle dekompozíciója is elképzelhető, de az nem igaz, hogy bármelyik végén felbomlik a klikkjeire, csak az, hogy létezik ilyen dekompozíció.

4.1.4. Definíció. Egy $G(V, E)$ irányítatlan gráfra azt mondjuk, hogy háromszögelt, ha $\forall k \geq 4$ pontú körének létezik húrja, azaz olyan éle, ami a körben nem szomszédos csúcsok között megy.

Egy ilyen G gráfnak minden $G' \subseteq G$ részgráfja is háromszögelt marad.

4.1.5. Definíció. Legyen $G(V, E)$ gráf és legyen $\alpha, \beta \in V$ csúcsok, melyeket $C \subset V$ elválasztja egymástól. Azt mondjuk, hogy C elválasztó halmaz minimális, ha elválasztja α, β csúcsokat, de nem tudunk elhagyni úgy csúcsot C-ből, hogy az továbbra is elválasztó halmaz maradjon.

4.1.1. Állítás. *Legyen $G(V, E)$ irányítatlan gráf, ekkor az alábbi tulajdonságok ekvivalensek:*



4.3. ábra. Dekomponálható gráf nem a klikkjeire bomlik fel.

(i) *G gyengén dekomponálható.*

(ii) *G háromszögelt.*

(iii) *Ha $\alpha, \beta \in G$ tetszőleges csúcsok és $C \subset V$ minimálisan elválasztja őket, akkor C teljes részgráf.*

Bizonyítás. Indukcióval bizonyítunk a csúcsok száma szerint. Ha G gráfnak legfeljebb három csúcsa van, mindig igaz rá mind a három tulajdonság. Most tegyük fel, hogy n csúcsig igaz, nézzük $n+1$ csúcsra.

(i)$\Longrightarrow$ (ii) Tegyük fel, hogy G gyengén dekomponálható, ha G teljes gráf, akkor nyilván háromszögelt, ha nem teljes, akkor létezik $G_{A\cup C}$ és $G_{B\cup C}$ valódi dekompozíciója, vagyis ezek csúcsszáma legfeljebb n, így az indukciós feltevés miatt, ezek háromszögeltek. Vagyis a háromszögeltséget sértő kör csak úgy lehet, ha a kör A és B halmazbeli csúcson is átmegy, de mivel C elválasztó halmaz, ekkor a kör legalább két különböző csúcsát érinti a C halmaznak, különben nem lenne kör. Mivel C teljes gráf, ezért ezen két pont között megy él, így ilyen köröknek is van húrja.

(ii)$\Longrightarrow$(iii). Legyen G háromszögelt és α, β rögzített, de tetszőleges csúcsok. Megmutatjuk, hogy ha C minimálisan elválasztja α, β csúcsokat, akkor C teljes gráf. Ha C egy csúcsból áll, akkor teljes gráf. Ha van legalább két csúcsa, akkor kell, hogy tetszőleges $c_1, c_2 \in C$ csúcsokra van köztük él. Vegyünk egy olyan élsorozatot, hogy $(\alpha \dots c_1 \dots \beta \dots c_2 \dots \alpha)$. Ez egy körséta lesz, rövidítsük ezt a sétát összehúzásokkal ott, ahol egy csúcsba többször is betérünk és ott, ahol húrja van ennek a körnek. Feltéve, hogy c_1 és c_2 között nincs él, az összehúzások végén egy pontosan négy pontú kört kapunk, melynek nincs éle, ez ellentmond a háromszögeltségnek, így kell hogy legyen él c_1 és c_2 között.

Mivel c_1, c_2 tetszőleges elemei C-nek, így C tetszőleges két eleme között vezet él, C valóban teljes gráf.

(iii)$\Longrightarrow$(i). Tegyük fel, hogy minden α, β csúcsra, ha C elválasztó halmaz minimális akkor C teljes gráf. Megmutatjuk, hogy G gyengén dekomponálható. Ha G teljes gráf, akkor kész. Ha nem az, akkor létezik két csúcsa, melyek nem szomszédosak, legyenek ezek α és β. Létezik őket elválasztó halmaz, pl. $V \setminus (\alpha \cup \beta)$ megfelelő. Ekkor létezik minimális is, legyen ez C, a feltevés miatt C teljes. Ekkor C halmaz egy dekompozíciót fog generálni, jelölje B a dekompozíciónak azt az összefüggőségi komponensét, amiben β benne van és A jelölje $V \setminus (C \cup B)$ halmazt. A-ban ekkor benne van az az összefüggőségi komponens, mely α csúcsot tartalmazza és a "maradék" gráf, ami nem B és nem C.

Ekkor (A, B, C) dekomponálja G gráfot $G_{A\cup C}$ és $G_{B\cup C}$ halmazokra. Kéne, hogy ezek is dekomponálhatóak. Tegyük fel, hogy $\alpha_1, \beta_1 \in G_{A\cup C}$ nem szomszédos csúcsok. Ha nincsen ilyen, akkor $G_{A\cup C}$ teljes gráf, így dekomponálható. Mivel nem szomszédosak, ezért van C_1 minimális elválasztó halmaz, ez teljes kell hogy legyen, mert ez a C_1 az eredeti gráfban is elválasztja α_1 és β_1 csúcsokat. Így valóban $G_{A\cup C}$ is dekomponálható, $G_{B\cup C}$ ugyanezzel az indoklással szintén dekomponálható. □

4.2. Tökéletes felsorolás

4.2.1. Definíció. Legyenek $B_1, B_2 \dots B_n \subseteq V$ részhalmazok egy felsorolása.

$$H_i = B_1 \cup B_2 \cup \dots B_i, \quad R_i = B_i \setminus H_{i-1}, \quad S_i = B_i \cap H_{i-1} \tag{4.1}$$

Ha az alábbi tulajdonságok teljesülnek,

(i) $\forall i > 1, \exists j$, hogy $j < i$ és $S_i \subseteq B_j$

(ii) $\forall i, S_i$ részgráf teljes gráf

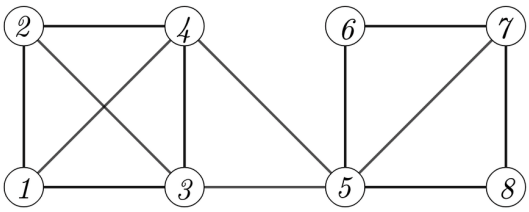
(iii) $\forall i > 1, R_i \subseteq \Gamma$ vagy $S_i \subseteq \Delta$

akkor azt mondjuk, hogy $B_1, B_2 \dots B_n$ tökéletes felsorolást alkot. Amennyiben csak (iii) nem teljesül, akkor a felsorolás gyengén tökéletes.

4.2.2. Definíció. Egy $G(V, E)$ irányítatlan gráf csúcsainak egy számozását tökéletesnek mondjuk, ha

$$B_i = cl(\alpha_i) \cap \{\alpha_1, \alpha_2 \dots \alpha_i\} \tag{4.2}$$

egy tökéletes felsorolás. A számozást gyengén tökéletesnek mondjuk, ha a B_i halmazok csak gyengén tökéletes felsorolást alkotnak.



4.4. ábra. A csúcsok egy tökéletes számozása

4.2.3. Példa. *Tekintsünk a 4.4 ábrán egy tökéletes számozását a csúcsoknak, határozzuk meg ez alapján B_i, H_i, R_i, S_i halmazokat.*

$B_1 = \{1\}$	$H_1 = \{1\}$	$R_1 = \{1\}$	$S_1 = \{\emptyset\}$
$B_2 = \{1,2\}$	$H_2 = \{1,2\}$	$R_2 = \{2\}$	$S_2 = \{1\}$
$B_3 = \{1,2,3\}$	$H_3 = \{1,2,3\}$	$R_3 = \{3\}$	$S_3 = \{1,2\}$
$B_4 = \{1,2,3,4\}$	$H_4 = \{1,2,3,4\}$	$R_4 = \{4\}$	$S_4 = \{1,2,3\}$
$B_5 = \{3,4,5\}$	$H_5 = \{1,2,3,4,5\}$	$R_5 = \{5\}$	$S_5 = \{3,4\}$
$B_6 = \{5,6\}$	$H_6 = \{1,2,\dots 6\}$	$R_6 = \{6\}$	$S_6 = \{5\}$
$B_7 = \{5,6,7\}$	$H_7 = \{1,2,\dots 7\}$	$R_7 = \{7\}$	$S_7 = \{5,6\}$
$B_8 = \{5,7,8\}$	$H_8 = \{1,2,\dots 8\}$	$R_8 = \{8\}$	$S_8 = \{5,7\}$

Ez valóban teljesíti a 4.2.1 definícióját, így B_i tényleg részhalmazok tökéletes felsorolása, vagyis a csúcsok számozása tökéletes. Megfigyelés, hogy ha a gráf csúcsainak adott egy tökéletes számozása, akkor a gráf részhalmazainak tökéletes felsorolásában $H_i = \{1,2,\dots i\}$, továbbá $R_i = \{i\}$.

4.2.1. Lemma. *Legyen $G(V,E)$ gráf és hozzá $\alpha_1, \alpha_2 \dots \alpha_k$ tökéletes számozás a csúcsokon, ekkor $B_i = cl(\alpha_i) \cap \{\alpha_1, \alpha_2 \dots \alpha_{i-1}\}$ egy olyan tökéletes felsorolása a részhalmazoknak, mely tartalmazza G összes klikkjét.*

Valóban a 4.2.3 példára visszatérve, B_i halmazok valóban tartalmazzák a gráf összes klikkjét.

Bizonyítás. Tudjuk, hogy $B_1, B_2, \dots B_k$ a részhalmazok egy tökéletes felsorolása lesz, ami a tökéletes számozás definíciója miatt adódik. Ekkor az utolsó $B_k = cl(\alpha_k) \cap \{\alpha_1, \dots \alpha_{k-1}\}$. Mivel $S_k = B_k \cap H_{k-1}$ teljes részgráf, ahol $H_{k-1} = \{1,2,\dots k-1\}$ a 4.2.3 megfigyelése alapján, így $B_k = S_k$, mert ugyanazzal a halmazzal metszünk, így B_k is teljes részgráf, sőt klikk is, mert tartalmazásra maximális. Innen a csúcsok száma szerinti indukcióval visszafelé belátható, hogy $\alpha_1, \alpha_2, \dots \alpha_{k-1}$ is tökéletes számozása a csúcsoknak a $G \setminus \{\alpha_k\}$ halmazon, így B_{k-1} is klikk, vagy részgráfja egy korábbi klikknek. Indirekt ha bővíthető lenne, vagyis nem klikk, akkor csak α_k csúcs hozzávételével bővíthetnénk, de ekkor $B_{k-1} \subseteq B_k$ adódna. □

4.2.2. Lemma. *Legyen $G(V,E)$ irányítatlan gráf és $B_1, B_2 \dots B_n \subseteq V$ egy olyan tökéletes felsorolás, amely tartalmazza G összes klikkjét. Ekkor minden j esetén S_j elválasztja $H_{j-1} \setminus S_j$ és R_j halmazt G_{H_j} gráfban, azaz $(H_{j-1} \setminus S_j, R_j, S_j)$ dekomponálja G_{H_j} gráfot.*

Bizonyítás. Először is figyeljük meg, hogy $(H_{j-1} \setminus S_j, R_j, S_j)$ halmazok tényleg diszjunktak. Legyen p a legnagyobb szám, amire B_p egy klikk. Mivel a felsorolás az összes klikket tartalmazza, így $H_p = V$, innen $\forall j, j > p, R_j = \emptyset$, így $S_j, j > p$ nyilván elválasztó halmaz. Most megmutatjuk, hogy $(H_{p-1} \setminus S_p, R_p, S_p)$ egy dekompozíció. Indirekt tegyük fel, hogy nem az, vagyis létezik $\beta \in (H_{p-1} \setminus S_p)$ és $\alpha \in R_p$, hogy (α, β) között vezet él. Az is igaz, hogy $\exists j < p$, hogy $\beta \in (B_j \setminus S_p)$. Ekkor $\{\alpha, \beta\}$ biztosan részhalmaza valamely $K \subset V$ klikknek. Viszont $K \neq B_p$ mert $\beta \notin B_p$ és $K \neq B_j, j < p$, mert $\alpha \notin B_j$, így $\{\alpha, \beta\}$ egy olyan klikkben van, melyet nem tartalmaz $B_1, B_2 \dots B_n$ felsorolás. Ez ellentmond a felsorolás választásának, így ilyen él nem létezhet, vagyis S_p elválasztja $(H_{p-1} \setminus S_p)$ és R_p halmazokat. □

4.2.1. Állítás. *Minden irányítatlan G gráfra az alábbi állítások ekvivalensek:*

(i) G csúcsainak létezik tökéletes számozása.

(ii) G klikkjeinek létezik tökéletes felsorolása.

(iii) G dekomponálható.

4.2.3. Lemma. *Legyen C' egy tetszőleges klikkje a G gráfnak. Ekkor létezik G klikkjeinek egy olyan tökéletes $C_1, C_2 \dots C_k$ felsorolása, hogy $C' = C_1$*

Az állítás segítségével könnyen ellenőrizhetjük, hogy egy adott G gráf klikkjeinek létezik-e tökéletes felsorolása. A lemma abban segít, hogy egy tökéletes felsorolást tetszőleges klikkből elindulva megtalálhatunk.

5. fejezet

Kontingencia táblák

Ez a fejezet Rudas Tamás [1] egyetemi jegyzete alapján íródott.

5.1. Jelölések

A kontingencia táblát arra használjuk, hogy valamilyen objektumokat közös tulajdonságaik alapján rendszerezzük és a táblázat celláiba az adott tulajdonságokkal rendelkező objektumok darbszámát feltüntessük. Jelölje Δ a tulajdonságok halmazát, ezt változóknak is szokták nevezni, minden $\delta \in \Delta$ változóhoz legyen I_δ, hogy ennek a változónak hányféle különböző állapota, szintje lehet. Például lehetnek az osztályozandó objektumok emberek, ekkor lehet $\delta \in \Delta$ változó a csillagjegyük, így $I_\delta = 12$. A kontingencia táblázat dimenziója $|\Delta|$ és összesen $J = \prod_{\delta \in \Delta} I_\delta$ cellája van. Egy adott cella tartalmát úgy olvassuk ki, hogy megadjuk úgymond a "koordinátáit", jelöljük ezt $n_{\mathbf{x}}$-el, ahol $\mathbf{x}$ vektor jelöli a változók megfelelő szintjeit. Azonban ez nem mindig praktikus, főleg akkor okoz problémát, ha túl sok dimenziója van a táblának. Alternatív jelölésként besorszámozhatjuk a cellákat $1, 2 \dots J$ például "sorfolytonosan" és ekkor $n(i)$ lesz egy cella tartalma, ahol $i \in \{1, 2, \dots J\}$

5.1.1. Példa. *Legyen $|\Delta| = 4$, továbbá a változókhoz tartozó szintek legyenek $I_{\delta_1} = 3, I_{\delta_2} = 2, I_{\delta_3} = 2, I_{\delta_4} = 2$. Ábrázoljuk úgy a táblázatot, hogy rögzítjük δ_1 és δ_2 értékét, majd az összes lehetséges rögzített δ_1 és δ_2 értékhez készítsünk a δ_3, δ_4 szerinti kétdimenziós táblázatot. Ekkor összesen $I_{\delta_1} \times I_{\delta_2} \times I_{\delta_3} \times I_{\delta_1} = 3 \times 2 \times 2 \times 2 = 24$ db cellája lesz a táblázatnak.*

Egy lehetséges ábrázolás négydimenziós táblára

	δ_1	A				B				C			
	δ_2	A		B		A		B		A		B	
	δ_3	A	B	A	B	A	B	A	B	A	B	A	B
δ_4	A	2	3	5	1	2	7	2	2	7	1	0	2
	B	1	0	4	1	8	1	2	2	1	6	0	7

Egy cella kiolvasása $n_{CABB} = 6$.

Nagyon fontosak lesznek ezek a bizonyos változók szerinti rögzített táblázat szeletek. Legyen T a táblázat, ekkor a δ_i szerinti szeletét jelölje $T(\delta_i)$, ez azt jelenti, hogy csak olyan cellákat tekintünk, amiknek az i-edik koordinátája az előre rögzített δ_i értéket veszi fel. Azért δ_i szerinti szelet, mert csak δ_i -hez tartozó koordináta értékét rögzítjük. Hasonló módon a táblázat $\delta_i, \delta_j \dots \delta_z$ szerinti szeletét jelölje $T(\delta_i, \delta_j \dots \delta_z)$. A jelölés motivációja, hogy csak olyan cellákat tekintünk, amik olyan alakúak, hogy a megfelelő koordinátákon az előre rögzített értékek állnak. Szorosan kapcsolódnak a szelethez a táblázat bizonyos változók szerinti marginálisai. A δ_i marginális azt jelenti, hogy a táblázat $T(\delta_i)$ szeletében szereplő cellák tartalmát összeadjuk. A konkrét példát tekintve formálisan jelöljük a T táblázat $\delta_2 = \hat{j}, \delta_3 = \hat{k}$ szerinti szeletét , majd marginálisát a $\hat{j}$ illetve $\hat{k}$ előre rögzített értékekkel.

$$T(\delta_2 = \hat{j}, \delta_3 = \hat{k}) = \{n_{ijkl} : j = \hat{j}, k = \hat{k}\} \tag{5.1}$$

$$n_{+\hat{j}\hat{k}+} = \sum_{i=1}^{I_{\delta_1}} \sum_{l=1}^{I_{\delta_4}} n_{i\hat{j}\hat{k}l} \tag{5.2}$$

Bevezetünk még egy jelölést, ami majd a későbbiekben lesz hasznos. Jelölje $n_{\hat{i}\cdot\hat{k}\cdot}$ ha a $T(\delta_1, \delta_3)$ szeletben nem a marginálist, hanem átlagot szeretnénk megadni, a nem rögzített változók szerint.

$$n_{\hat{i}\cdot\hat{k}\cdot} = \frac{n_{\hat{i}+\hat{k}+}}{|T(\delta_1 = \hat{i}, \delta_3 = \hat{k})|} \tag{5.3}$$

Itt az abszolút érték, a megfelelő tulajdonságú cellák számát jelenti.

A konkrét példát nézve $n_{BA\cdot\cdot} = \frac{2+7+8+1}{4} = \frac{18}{4}$.

Összefoglalva tehát:

Δ : A változók halmaza, amikkel az objektumainkat osztályozni akarjuk.

I_{δ_i} : Az i-edik változónak ennyi lehetséges értéke, szintje lehet.

J : Összesen ennyi darab cella van.

$n_{\mathbf{x}}$: Annak a cellának a tartalmát jelenti, melynek "koordinátái" éppen az $\mathbf{x}$ vektor.

$n(i)$: A cellák sorfolytonos számozása esetén az i-edik cella tartalma

$T(\delta_i, \dots \delta_z)$: A T táblázat $\delta_i, \dots \delta_z$ szerinti szelete, ezen koordinátáiban rögzített cellák halmaza.

$n_{+\hat{j}\hat{k}+}$: Egy konkrét négydimenziós táblázat esetén azon cellák tartalmának összegét jelenti, amelyeknek δ_2 koordinátájában éppen $\hat{j}$, illetve δ_3 koordinátájában éppen $\hat{k}$ érték van.

$n_{\cdot\hat{j}\hat{k}\cdot}$: Egy konkrét négydimenziós táblázat esetén azon cellák tartalmának az átlagát jelenti, amelyeknek δ_2 koordinátájában éppen $\hat{j}$, illetve δ_3 koordinátájában éppen $\hat{k}$ érték van.

N : $\sum_{i=1}^{J} n(i)$ vagyis a cellák értékeinek összege.

$q_{\mathbf{x}}$: $\frac{n_{\mathbf{x}}}{N}$, vagyis a $\mathbf{x}$ koordinátájú cellába esés relatív gyakorisága.

$p_{\mathbf{x}}$: A tényleges valószínűsége az $\mathbf{x}$ koordinátájú cellába esésnek.

5.2. Függetlenség a táblázatban

Az alapvető jelölések tisztázása után térjünk rá az elemzésre. Szimuláltuk a következő kísérletet: Dobjunk egyszerre dobókockával és pénzérmével, majd a táblázat megfelelő cellájának értékét növeljük. Legyen X egy darab kockadobás eredménye, Y pedig egy darab érmedobás eredménye. Ez egy független kétdimenziós táblázat lesz. A kísérletet 10000-szer ismételve a következő táblázatot kaptuk:

δ_1 \ δ_2	$Y=0$	$Y=1$	
$X<3$	1707	1732	3439
$X\geq 3$	3276	3285	6561
	4983	5017	10000

δ_1 \ δ_2	$Y=0$	$Y=1$	
$X<3$	n_{11}	n_{12}	n_{1+}
$X\geq 3$	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

Legyen δ_1 változó a kockadobás, ennek most két kategóriája van, δ_2 változó pedig az érme dobás, ennek is két kategóriája van, ekkor n_{12} tehát azt jelenti, hogy az a cella, ami δ_1 szerint az első és δ_2 szerint a második kategóriába esik. Figyeljük meg, hogy $\frac{n_{11}}{n_{21}} \approx \frac{n_{12}}{n_{22}}$. Ha ki szeretnénk deríteni, hogy milyen valószínűséggel esik egy objektum a δ_1 változó szerint az egyes kategóriákba, akkor nem segít, nem ad erről a valószínűségről információt az a tény, hogy δ_2 szerint melyik kategóriába esett. Úgy is mondhatjuk, hogy egy objektumnak a δ_1 változó szerinti értékére nincsen hatással az, hogy δ_2 szerint melyik kategóriába esett. Megjegyezzük, hogy ekkor $\frac{n_{11}}{n_{12}} \approx \frac{n_{21}}{n_{22}}$ is teljesül. Vagyis ekkor az, hogy δ_1 szerint melyik kategóriába esett, nincs hatással arra, hogy δ_2 szerint melyik kategóriába fog esni. Ezt a tulajdonságot szeretnénk függetlenségnek hívni és általánosabb képlettel leírni. Végig n_{ij} értékek helyett inkább q_{ij} relatív gyakoriságokkal dolgozunk, természetesen n_{ij} értékekkel ugyanúgy levezethető lenne.

$$\begin{aligned}
\frac{q_{11}}{q_{12}} = \frac{q_{21}}{q_{22}} \Rightarrow q_{11} = \frac{q_{21}q_{12}}{q_{22}} \Rightarrow \\
1 = \frac{q_{21}q_{12}}{q_{22}} + q_{12} + q_{21} + q_{22} \Rightarrow \\
q_{22} = q_{21}q_{12} + q_{12}q_{22} + q_{21}q_{22} + q_{22}q_{22} \Rightarrow \\
q_{22} = (q_{21} + q_{22})(q_{12} + q_{22}) = q_{2+}q_{+2}
\end{aligned}$$

Azt kaptuk, hogy függetlenség mellett egy kétdimenziós táblázatban $q_{\hat{i}\hat{j}} = q_{\hat{i}+}q_{+\hat{j}}$ érték számolható a megfelelő marginálisok szorzatával. Ekkor $\hat{n}_{ij} = Nq_{ij}$ lesz a becslésünk arra, hogy független táblázat esetén egy adott cellába mekkora darabszámot várunk. A motiváció tehát az, hogy a táblázat bizonyos marginálisait tekintve, egy képzési szabályt adjunk az összes cellára. A lehető legkevesebb információt szeretnénk használni, csak azokat a marginálisokat kívánjuk megtartani, amik feltétlenül szükségesek. Az elhagyás úgy lehetséges, hogy valamiféle kapcsolatot feltételezve előállítunk egy képzési szabályt és más marginálisokból a szabály alapján előállítjuk a hiányzó cellák darabszámait. Visszatekinte az 5.2 táblázat esetére, itt a kapcsolat a függetlenség, a képzési szabály pedig:

$$\hat{n}_{ij} = N \cdot q_{ij} = N \cdot q_{i+}q_{+j} = N \cdot \frac{n_{i+}}{N} \cdot \frac{n_{+j}}{N} = \frac{n_{i+}n_{+j}}{N}.$$

Ez azt jelenti, hogy egy kétdimenziós független táblázatnál elég ismerni a sorösszegeket és oszlopösszegeket, ebből már minden cella értéke meghatározható.

Tekintsük általánosan, egy k dimenziós táblázatra ugyanezt a kérdést.

5.2.1. Definíció. Legyen T egy k dimenziós kontingencia táblázat, azaz $|\Delta| = k$. Azt mondjuk, hogy ez a táblázat független, ha

$$p_{\mathbf{x}} = p_{x_1++\ldots+} p_{+x_2+\ldots+} \cdots p_{++\ldots+x_k} \tag{5.4}$$

képzési szabály érvényes az egyes cellákba való kerülés valószínűségére.

Ebben a speciális esetben csak az egydimenziós marginálisok szerepelnek a képletben, de azokból az összes. Természetesen más képzési szabály is elképzelhető, például az első két tényező helyett $p_{x_1x_2++\ldots+}$ esetében már egy kétdimenziós marginális tag is előfordulna. Ha a változók között valamilyen képzési szabályt tudunk felállítani a marginálisok használatával, akkor azt mondjuk, hogy érvényes rájuk a függetlenség valamely általánosított formája. Ehhez kapcsolódva megfogalmazzuk az interakció fogalmát.

5.2.2. Definíció. Ha $\delta_1, \delta_2 \ldots \delta_k \in \Delta$ változókra nem érvényesül a függetlenség semelyik általánosított formája, akkor ez a k változó interakcióban van.

Ez a definíció azt akarja megragadni, hogy ekkor ezekre a változókra szükségünk lesz a képzési szabályban, mert nem tudjuk egyiket sem kikeverni a többi segítségével. Ha ki tudnánk hozni valamiféle képzési szabályt, az pont azt jelentené, hogy érvényes az általánosított függetlenség valamely formája. Amikor tehát azt mondtuk, hogy a lehető legkevesebb információt akarjuk használni, azt úgy értjük, hogy a lehető legkevesebb változót szeretnénk használni a képzési szabályban. Az interakciókra később még visszatérünk, de előbb térjünk rá a loglineáris elemzés módszerére.

5.3. Loglineáris elemzés

Először egy példán megmutatjuk a módszer ötletét. Elő fogjuk állítani az első táblázatot úgy, hogy az egyenletes táblából indulunk ki. Először felírjuk a marginálisokhoz, hogy az egyenletes tábla hogyan tér el az eredetitől, ezeket a számokat jelöltük kékkel.

δ_1 \ δ_2	A	B	
A	0,2	0,1	0,3
B	0,4	0,3	0,7
	0,6	0,4	1

δ_1 \ δ_2	A	B	
A	0,25	0,25	0,5 - 0,2
B	0,25	0,25	0,5 + 0,2
	0,5 + 0,1	0,5- 0,1	1

Most elkészítjük azt a független táblázatot, aminek a marginálisai megegyeznek az eredetivel, majd beírjuk a cellákhoz, hogy miben tér el ez a független táblázat az eredetitől, ezeket a számokat jelöltük lilával.

δ_1 \ δ_2	A	B	
A	0,18	0,12	0,3
B	0,42	0,28	0,7
	0,6	0,4	1

δ_1 \ δ_2	A	B	
A	0,18 + 0,2	0,12 - 0,2	0,3
B	0,42 - 0,2	0,28 + 0,2	0,7
	0,6	0,4	1

Valójában ezek a kék és lila számok egyértelműen meghatározzák a keresett eredeti táblázatot. A loglineáris elemzés lényege, hogy valamilyen módon ezeket az értékeket írjuk fel, hiszen ezek, mint láttuk, egyértelműen meghatározzák a táblázatot. A különbség annyi, hogy valószínűségek helyett, a valószínűségek logaritmusával fogunk dolgozni, emiatt tegyük fel, hogy $\forall i, j$ esetén $p_{ij} > 0$. A logaritmus motivációja, többek között, hogy szorzatok helyett összegekkel tudunk dolgozni. Bevezetünk jelöléseket, először csupán 2×2 táblázat esetén:

$$k_{ij} := \log p_{ij}$$

$$u := k_{..}$$

$$\alpha_i^{\delta_1} := k_{i.} - u$$

$$\alpha_j^{\delta_2} := k_{.j} - u$$

$$\alpha_{ij}^{\delta_{12}} := k_{ij} - \alpha_i^{\delta_1} - \alpha_j^{\delta_2} - u$$

Ezek az értékek adott táblázat esetén könnyen számolhatóak és az alábbi négy táblázatot az előzőhöz hasonló módon fogjuk kitölteni.

δ_1 \ δ_2	A	B	
A	k_{AA}	k_{AB}	k_{A+}
B	k_{BA}	k_{BB}	k_{B+}
	k_{+A}	k_{+B}	k_{++}

δ_1 \ δ_2	A	B	
A	$u + \alpha_A^{\delta_1}$	$u + \alpha_A^{\delta_1}$	k_{A+}
B	$u + \alpha_B^{\delta_1}$	$u + \alpha_B^{\delta_1}$	k_{B+}
	-	-	k_{++}

Ha minden cellába u érték kerülne, az lenne az egyenletes tábla. A második táblázatban δ_1 változóhoz tartozó tagokat illesztettünk és azt állítjuk, hogy ekkor a sorösszegek megegyeznek az eredetivel. Nem jelöltük az oszlopösszegeket, hiszen azok így még általában nem fognak stimmelni az eredetivel. Azonban igaz az, hogy ha ebbe a második táblázatba δ_2 változóhoz tartozó értékeket írtuk volna az u mellé a δ_1 értékek helyett, akkor az oszlopösszegek egyeztek volna meg az eredetivel.

δ_1 \ δ_2	A	B	
A	$u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2}$	$u + \alpha_A^{\delta_1} + \alpha_B^{\delta_2}$	k_{A+}
B	$u + \alpha_B^{\delta_1} + \alpha_A^{\delta_2}$	$u + \alpha_B^{\delta_1} + \alpha_B^{\delta_2}$	k_{B+}
	k_{+A}	k_{+B}	k_{++}

Azt állítjuk, hogy ez harmadik táblázat megfelel az olyan független táblázatnak, aminek a marginálisai megegyeznek az eredeti első táblázat marginálisaival. Az utolsó táblázat vissza fogja adni az eredeti első táblázatot, azonban a cellákba a k_{ij} értékeket egy összeget alakban fogjuk előállítani.

δ_1 \ δ_2	A	B	
A	$u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2} + \alpha_{AA}^{\delta_{12}}$	$u + \alpha_A^{\delta_1} + \alpha_B^{\delta_2} + +\alpha_{AB}^{\delta_{12}}$	k_{A+}
B	$u + \alpha_B^{\delta_1} + \alpha_A^{\delta_2} + \alpha_{BA}^{\delta_{12}}$	$u + \alpha_B^{\delta_1} + \alpha_B^{\delta_2} + \alpha_{BB}^{\delta_{12}}$	k_{B+}
	k_{+A}	k_{+B}	k_{++}

Ebben az utolsó táblázatban k_{ij} definíciója szerint a cellák valóban mind megegyeznek az eredeti táblázat celláinak értékeivel, azonban ez az összeg alak további információt hordoz magában. Ha igazak az állításaink, akkor az $\alpha_{ij}^{\delta_{12}}$ tagok lesznek kifejezetten érdekesek, hiszen ezek fogják megmutatni, hogy mennyiben tér el a táblázat az ugyanilyen marginálisú független táblázattól.

Most megmutatjuk, hogy a második táblázat valóban olyan, hogy a sorösszegek stimmelnek az eredeti táblázat sorösszegeivel, a harmadik táblázatban már sor- és oszlopösszegek is stimmelnek, ráadásul ez éppen az a táblázat, ami a marginálisaiban megegyezik az eredetivel és független is. Ha az eredeti táblázat éppen független volt, akkor nyilván már a harmadikban is stimmelni fog, ekkor az $\alpha_{ij}^{\delta_{12}}$ tagok mind nullák.

Bizonyítás. Legyen az eredeti 2×2 táblázat tetszőleges $p_{ij} > 0$ valószínűségekkel kitöltve és összegük legyen 1. Írjuk át p_{ij} helyett $k_{ij} = \log p_{ij}$ -re a cellák tartalmát. A második táblát figyelve kell, hogy

$$k_{AA} + k_{AB} = u + \alpha_A^{\delta_1} + u + \alpha_A^{\delta_1}$$

A jobb oldalt alakítva kapjuk

$$u + \alpha_A^{\delta_1} + u + \alpha_A^{\delta_1} = u + \frac{k_{AA} + k_{AB}}{2} - u + u + \frac{k_{AA} + k_{AB}}{2} - u = k_{AA} + k_{AB}$$

Vagyis az első sorösszeg valóban stimmel, a második sorra ugyanez a számolás. Vegyük észre, hogy ha a sorok helyett az oszlopokat illesztettük volna, akkor hasonló számolással megkapjuk, hogy a megfelelő $\alpha_A^{\delta_2}$ és $\alpha_B^{\delta_2}$ értékeket kellett volna hozzáadni az u-hoz. Térjünk most rá a harmadik táblára. Kell, hogy:

$$k_{AA} + k_{AB} = u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2} + u + \alpha_A^{\delta_1} + \alpha_B^{\delta_2}$$
$$k_{AA} + k_{BA} = u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2} + u + \alpha_B^{\delta_1} + \alpha_A^{\delta_2}$$

Vagyis az első sor és oszlopösszegek megegyeznek.

Ismét a jobb oldalt alakítva kapjuk

$$\begin{aligned}
u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2} + u + \alpha_A^{\delta_1} + \alpha_B^{\delta_2} &= k_{AA} + k_{AB} + \alpha_A^{\delta_2} + \alpha_B^{\delta_2} = \\
k_{AA} + k_{AB} + \frac{k_{AA} + k_{BA}}{2} - u + \frac{k_{AB} + k_{BB}}{2} - u &= \\
k_{AA} + k_{AB} + \frac{k_{AA} + k_{AB} + k_{BA} + k_{BB}}{2} - \frac{k_{AA} + k_{AB} + k_{BA} + k_{BB}}{2} &= k_{AA} + k_{AB} \\
u + \alpha_A^{\delta_1} + \alpha_A^{\delta_2} + u + \alpha_B^{\delta_1} + \alpha_A^{\delta_2} &= k_{AA} + k_{BA} + \alpha_A^{\delta_1} + \alpha_B^{\delta_1} = \\
k_{AA} + k_{BA} + \frac{k_{AA} + k_{AB}}{2} - u + \frac{k_{BA} + k_{BB}}{2} - u &= \\
k_{AA} + k_{BA} + \frac{k_{AA} + k_{AB} + k_{BA} + k_{BB}}{2} - \frac{k_{AA} + k_{AB} + k_{BA} + k_{BB}}{2} &= k_{AA} + k_{BA}
\end{aligned}$$

A számolásban már felhasználtuk a második tábla eredményét. A többi sor illetve oszlopösszeg ugyanilyen számolással megkapható. A negyedik tábla $\alpha_{ij}^{\delta_{12}}$ definíciója miatt valóban az eredetit fogja visszaadni. □

Megmutatjuk azt is, hogy a harmadik tábla akkor és csak akkor fog megegyezni az eredetivel, ha az eredeti független volt.

Bizonyítás. Ha az eredeti független, akkor $p_{ij} = p_{i+}p_{+j} = 4p_{i\cdot}p_{\cdot j}$,innen logaritmust véve mindkét oldalból $k_{ij} = \log 4 + \log p_{i\cdot} + \log p_{\cdot j}$. Használjuk, hogy

$$u + \alpha_i^{\delta_1} + \alpha_j^{\delta_2} = k_{i\cdot} + k_{\cdot j} - k_{\cdot\cdot} \tag{5.5}$$

A kérdés ekkor, hogy

$$k_{ij} = \log 4 + \log p_{i\cdot} + \log p_{\cdot j} = k_{i\cdot} + k_{\cdot j} - k_{\cdot\cdot} \tag{5.6}$$

egyenlőség teljesül-e.

Írjuk fel az egyenlet jobb oldalán lévő tagokat egyesével:

$$\begin{aligned}
k_{i\cdot} =& \frac{k_{iA} + k_{iB}}{2} = \frac{1}{2}(\log 4 + \log p_{i\cdot} + \log p_{\cdot A} + \log 4 + \log p_{i\cdot} + \log p_{\cdot B}) = \\
&= \log 4 + \log p_{i\cdot} + \frac{1}{2}(\log p_{\cdot A} + \log p_{\cdot B}) \\
k_{\cdot j} =& \frac{k_{Aj} + k_{Bj}}{2} = \frac{1}{2}(\log 4 + \log p_{A\cdot} + \log p_{\cdot j} + \log 4 + \log p_{B\cdot} + \log p_{\cdot j}) \\
&= \log 4 + \log p_{\cdot j} + \frac{1}{2}(\log p_{A\cdot} + \log p_{B\cdot}) \\
k_{\cdot\cdot} =& \frac{k_{A\cdot} + k_{B\cdot}}{2} = \frac{\log 4 + \log p_{A\cdot} + \frac{1}{2}(\log p_{\cdot A} + \log p_{\cdot B}) + \log 4 + \log p_{B\cdot} + \frac{1}{2}(\log p_{\cdot A} + \log p_{\cdot B})}{2} \\
&= \log 4 + \frac{1}{2}(\log p_{A\cdot} + \log p_{B\cdot} + \log p_{\cdot A} + \log p_{\cdot B})
\end{aligned}$$

Így tehát valóban 5.6 egyenlőség teljesül. Most tegyük fel, hogy 5.6 teljesül, ebből megmutatjuk, hogy az eredeti táblázat független.

$$k_{ij} = k_{i\cdot} + k_{\cdot j} - k_{\cdot\cdot} = \frac{k_{iA} + k_{iB} + k_{Aj} + k_{jB}}{2} - \frac{k_{AA} + k_{AB} + k_{BA} + k_{BB}}{4}$$

Mivel most 2×2 táblázatunk van, ezért bevezetjük i^* jelölést, ezzel az i "párját" fogjuk jelölni, például $k_{AA^*} = k_{AB}$

$$k_{ij} + k_{i^*j^*} = u + \alpha_i^{\delta_1} + \alpha_j^{\delta_2} + u + \alpha_{i^*}^{\delta_1} + \alpha_{j^*}^{\delta_2} = k_{i\cdot} + k_{\cdot j} + k_{i^*\cdot} + k_{\cdot j^*} - 2u = 2u$$

Itt azt használtuk fel, hogy $k_{i\cdot} + k_{i^*\cdot} = \frac{k_{iA}+k_{iB}+k_{i^*A}+k_{i^*B}}{2} = \frac{4u}{2}$. Ekkor az előző egyenletet használva:

$$\frac{k_{ij} + k_{i^*j^*}}{2} = u = \frac{k_{ij} + k_{ij^*} + k_{i^*j} + k_{i^*j^*}}{4}$$

Mindkét oldalt szorozva néggyel és exponenciálist véve kapjuk, hogy

$$\exp\{2k_{ij} + 2k_{i^*j^*}\} = p_{ij}^2 \cdot p_{i^*j^*}^2$$

$$\exp\{k_{ij} + k_{ij^*} + k_{i^*j} + k_{i^*j^*}\} = p_{ij} \cdot p_{ij^*} \cdot p_{i^*j} \cdot p_{i^*j^*}$$

Mivel ezek megegyeznek, így:

$$p_{ij}^2 \cdot p_{i^*j^*}^2 = p_{ij} \cdot p_{ij^*} \cdot p_{i^*j} \cdot p_{i^*j^*}$$
$$p_{ij} \cdot p_{i^*j^*} = p_{ij^*} \cdot p_{i^*j}$$
$$\frac{p_{ij}}{p_{ij^*}} = \frac{p_{i^*j}}{p_{i^*j^*}}$$

Ami éppen az elvárt tulajdonság egy független táblázatnál. □

5.3.1. Példa. *Tekintsük újra 5.3 első táblázatát és számoljuk ki a loglineráris reprezentációs tagokat.*

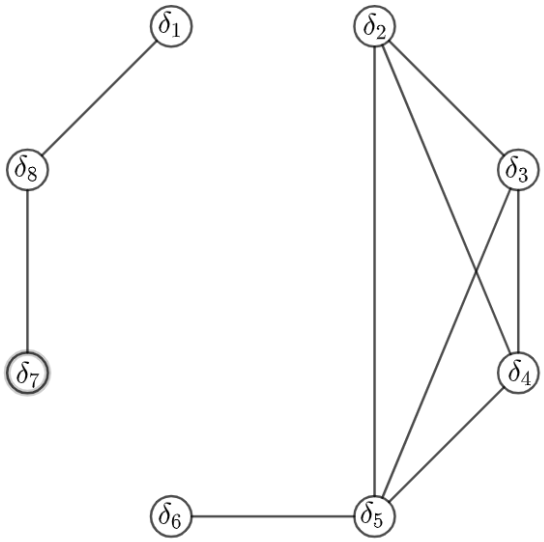
$$k_{AA} = \log 0,2 \qquad k_{AB} = \log 0,1 \qquad k_{BA} = \log 0,4 \qquad k_{BB} = \log 0,3$$
$$u \approx -1,5081$$
$$\alpha_A^{\delta_1} \approx -0,448 \qquad \alpha_B^{\delta_1} \approx 0,448 \qquad \alpha_A^{\delta_2} \approx 0,245 \qquad \alpha_B^{\delta_2} \approx -0,245$$
$$\alpha_{AA}^{\delta_{12}} \approx 0,10166 \qquad \alpha_{AB}^{\delta_{12}} \approx -0,10166 \qquad \alpha_{BA}^{\delta_{12}} \approx -0,10166 \qquad \alpha_{BB}^{\delta_{12}} \approx 0,10166$$

Így valóban például: $k_{AA} = \log 0,2 \approx u + \alpha_1^{\delta_A} + \alpha_A^{\delta_2} + \alpha_{AA}^{\delta_{12}}$.

Figyeljük meg, hogy rögzített δ_j mellett a hozzá tartozó egydimenziós tagok összege nulla, sőt, rögzített δ_{ij} esetén is a hozzá tartozó kétdimenziós tagok összege nulla. Általánosan is igaz, ha van egy $\delta_{\mathbf{x}}$ tényező, akkor a hozzá tartozó megfelelő dimenziós tagok összege nulla kell hogy legyen. Ha δ_j egydimenziós változónak csak két kategóriája van, akkor a hozzá tartozó két reprezentációs tag egymás ellentettje kell hogy legyen, hiszen az összeg nulla lesz. sőt, ha δ_{ij} olyan kétdimenziós tag, hogy δ_i-nek és δ_j-nek is két kategóriája van, akkor δ_{ij}-hez tartozó négy reprezentációs tagnak abszolút értékei meg fognak egyezni, az előjel is meghatározható. Azon tagok előjele fog megegyezni, melyeknek indexei páros sok cserével egymásba vihető. Jelen esetben $\alpha_{AA}^{\delta_{12}}$ előjele pozitív, így az egy cserével megkapható AB és BA indexűek negatívak lesznek és a két cserével megkapható BB ismét pozitív.

5.4. Loglineáris reprezentáció

Az előző fejezetben tehát arra jutottunk, hogy ha ismerjük egy táblázat celláihoz tartozó $p_{\mathbf{x}}$ valószínűségeket, akkor átírva a loglineáris reprezentáció segítségével egy összeget kapunk. Ha az összegben valamelyik tag esetleg nulla, az azt jelenti, hogy a hozzájuk tartozó változók függetlenek egymástól. Legyen adott egy táblázat, ami 8 változót tartalmaz, jelölje ezeket $\delta_1 \dots \delta_8$. Tegyük fel, hogy ismerjük $p_{\mathbf{x}}$ valószínűségeket. Készítsük el a loglineáris reprezentációt és ábrázoljuk a táblázatot egy gráf segítségével. Legyenek a gráf csúcsai a δ_i valószínűségi változók és akkor húzzunk be két csúcs között élt, ha a reprezentációs összegben a hozzájuk tartozó tagok nem nullák, vagyis nem igaz rájuk a függetlenség semelyik általánosított formája. Azonban kapcsolat nem csak két változó között lehet, így ha van a reprezentációs összegben egy négydimenziós nem nulla tag, akkor a gráfon a neki megfelelő négy csúcs között minden élt be kell hogy húzzunk. Tekinstük 5.1 ábrát, amit egy adott táblázat alapján készítettünk. Mivel δ_1 és δ_6



5.1. ábra. A reprezentációs összeg tagjainak értelmezése.

között nincs él, így tudjuk, hogy a reprezentációs összegben $\alpha_{ij}^{\delta_{16}}$ nulla volt $\forall i, j$ indexre. sőt azt is tudjuk, hogy nulla lesz az összes olyan többdimenziós α tag, amiben δ_1 és δ_6 egyszerre szerepel. Azonban $\delta_2, \delta_3, \delta_4, \delta_5$ csúcsokat nézve, nem egyértelmű, hogy milyen összegek tartoznak hozzájuk, hiszen lehet, hogy páronként mind függetlenek, csak egyben nem azok, ezért kötöttük őket össze éllel. Gyűjtsük ki azokat a δ párokat, csoportokat, melyek előfordulnak a reprezentációs összegben. Legyen ez most $[\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_1\delta_8, \delta_7\delta_8, \delta_5\delta_6, \delta_2\delta_3\delta_4\delta_5]$. Innen már egyértelműen látjuk, hogy valóban $\delta_2, \delta_3, \delta_4, \delta_5$ változók páronként függetlenek voltak, csak egyben nem.

5.4.1. Definíció. Egy modell hierarchikus, ha tetszőleges interakciójának minden rész-interakciója is előfordul a loglineáris reprezentációban.

Vagyis az előbbi modell nem lesz hierarchikus, hiszen például $\delta_2\delta_3$ tényzőnek szerepelni kéne.

Vegyünk egy 2×2 független táblázatot δ_1, δ_2 változókkal. Mivel δ_1, δ_2 változók között fennáll a függetlenség valamely formája, ezért nincsen közöttük interakció, vagyis $\alpha_{ij}^{\delta_{12}}$ tagra nincsen szükség, ekkor nincsen él a változókat reprezentáló csúcsok között. Viszont a sor és oszlopösszegekért felelős egydimenziós tagokat meg akarjuk tartani, így ezt a táblázatot $[\delta_1, \delta_2]$ formában tudjuk ábrázolni. Ha a táblázat nem volna független , akkor $[\delta_1, \delta_2, \delta_1\delta_2]$ lenne a megfelelő ábrázolás. Mindkét esetben hierarchikus modellt kapunk.

Ha tudjuk, hogy a modell hierarchikus, akkor elegendő a maximális interakciókat felsorolni. Figyeljük meg, hogy ha tudnánk, hogy 5.1 ábra egy hierarchikus modell, akkor sem lesz egyértelmű csupán az ábrából, hogy mik az interakciós tagok, hiszen $[\delta_1\delta_8, \delta_7\delta_8, \delta_5\delta_6, \delta_2\delta_3\delta_4\delta_5]$, de $[\delta_1\delta_8, \delta_7\delta_8, \delta_5\delta_6, \delta_2\delta_3\delta_4, \delta_2\delta_4\delta_5, \delta_3\delta_4\delta_5]$ is ugyanezt az ábrát adja.

5.4.2. Példa. *A független 5.2 példát tekintve itt $\delta_1, \delta_2, \delta_1\delta_2$ az összes lehetséges tényező ami szerepelhet a képzési szabályban, mivel $\delta_1\delta_2$ tényezőre fennáll az általánosított függetlenség*

valamely formája (ebben az esetben a jól ismert függetlenség), így erre nem lesz szükségünk, ami marad, az $[\delta_1, \delta_2]$ interakciók.

Nézzünk egy másik példát az interakciók szemléltetésére:

5.4.3. Példa. *Legyen a modell a feltételes függetlenség három változóval, ekkor a képzési szabály:*

$$\hat{p}_{ijk} = \frac{p_{ij+}p_{i+k}}{p_{i++}}$$

ami éppen 2.3 képletből átrendezéssel adódik. Az interakciók pedig $[\delta_1\delta_2, \delta_1\delta_3]$.

Tehát egy interakciós tag akkor lesz nulla, ha az interakcióban szereplő változókra teljesül a függetlenség valamely általánosítása. Ha már tudjuk a képzési szabályt, akkor ez alapján az interakciókat és a hozzá tartozó gráfot fel tudjuk rajzolni. Kérdés, hogy a képzési szabályra hogyan tudunk következtetni a gráf és az interakciók segítségével. Legyen adott egy kontingencia tábla, amihez felrajzoltunk egy G dekomponálható gráfot (4.1.3 definíció), hogy a gráf csúcsai jelentik a táblázat változóit, az élek pedig a függetlenségi kapcsolatokat. Célunk a modellhez tartozó maximum likelihood becslést megadni, majd ezt összehasonlítani az eredeti táblázatban kapott értékekkel. Azt a modellt fogjuk elfogadni, amihez tartozó becslés a lehető legjobban közelíti az eredeti értékeket.

5.4.1. Állítás. *Tegyük fel, hogy a G gráf dekomponálható,így a klikkjeinek van egy tökéletes felsorolása. Ekkor a maximum likelihood becslés egy adott cellára az alábbi alakot ölti:*

$$\hat{p}_{\boldsymbol{x}} = \frac{\prod_{c \in C} n_{\boldsymbol{c}}^{\boldsymbol{x}}}{\prod_{s \in S} (n_{\boldsymbol{s}}^{\boldsymbol{x}})^{\mu_s}} \tag{5.7}$$

ahol $c \in C$ a klikkek, $s \in S$ a szeparátorok a tökéletes felsorolásban, $n_{\mathbf{c}}^{\mathbf{x}}$ pedig a c klikkben szereplő csúcsokhoz tartozó δ_i tagokon kívül a többi + és a δ_i értékeket az $\mathbf{x}$ szerinti szintjükön nézzük, $(n_{\mathbf{s}}^{\mathbf{x}})^{\mu_s}$ pedig előzőhöz hasonlóan és μ_s pedig azt számlálja, hogy hányszor szerepel szeparátorként a tökéletes felsorolásban. Megjegyezzük, hogy ha a G gráf nem dekomponálható, akkor maximum likelihood becslést nem lehet explicit alakban megadni.

6. fejezet

Egy példa

Ebben a fejezetben egyetlen példán keresztül megmutatjuk az előző fejezetek eredményeit. Vizsgáljunk bírósági ítéleteket. Legyen négy változónk, a gyilkos neme, a gyilkos színe, az áldozat színe és az ítélet. Tekintsük az alábbi (nem valós adatokat tartalmazó) táblázatot 694 esetről:

	δ_1	Férfi				Nő			
	δ_2	Fekete		Fehér		Fekete		Fehér	
	δ_4	Halál	Börtön	Halál	Börtön	Halál	Börtön	Halál	Börtön
δ_3	Fekete	10	40	12	30	22	71	8	18
	Fehér	15	60	20	63	57	202	21	45

A táblázat változói

δ_1 := A gyilkos neme, (férfi, nő).

δ_2 := A gyilkos bőrszine, (fekete, fehér).

δ_3 := Az áldozat bőrszine, (fekete, fehér).

δ_4 := A bírósági itélet, (halál, börtön).

Jelölje a táblázat celláit n_{ijkl},ahol $i, j, k, l \in \{0, 1\}$. Például n_{1011} azt a cellát jelenti, hogy a gyilkos férfi, a színe fekete, az áldozat fehér és nem ítélték halálra, ennek a cellának az értéke 202. Készítsük el a loglineáris reprezentációt. A cellába esés relatív gyakoriságát $q_{ijkl} = \frac{n_{ijkl}}{N} = \frac{n_{ijkl}}{694}$ képlet alapján számoljuk. A 5.3 mintájára, négydimenziós táblázat esetén egy $ijkl$ indexű cellába $k_{ijkl} = \log q_{ijkl} = \log \frac{n_{ijkl}}{N}$ értéket írjuk összeg alakban, az alábbi módon:

$$k_{ijkl} = u + \alpha_i^{\delta_1} + \alpha_j^{\delta_2} + \alpha_k^{\delta_3} + \alpha_l^{\delta_4} + \alpha_{ij}^{\delta_{12}} + \alpha_{ik}^{\delta_{13}} + \alpha_{il}^{\delta_{14}} + \alpha_{jk}^{\delta_{23}} + \alpha_{jl}^{\delta_{24}} + \alpha_{kl}^{\delta_{34}} + \alpha_{ijk}^{\delta_{123}} + \alpha_{ijl}^{\delta_{124}} + \alpha_{ikl}^{\delta_{134}} + \alpha_{jkl}^{\delta_{234}} + \alpha_{ijkl}^{\delta_{1234}}$$

Itt az egyre több dimenziós interakciós tagok egymásból kifejezhetőek a következő módon:

$$\alpha_i^{\delta_1} = k_{i\ldots} - u$$

$$\alpha_{ij}^{\delta_{12}} = k_{ij\cdot\cdot} - \alpha_i^{\delta_1} - \alpha_j^{\delta_2} - u$$

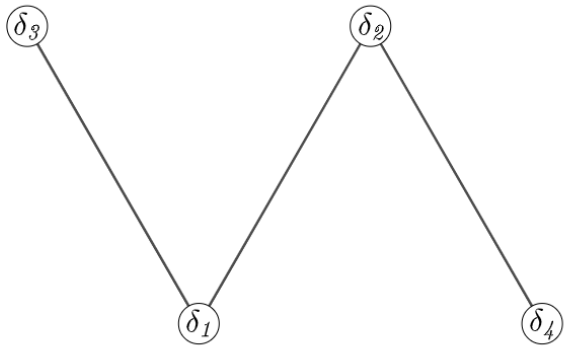
$$\alpha_{ijk}^{\delta_{123}} \quad k_{ijk\cdot} - \alpha_{ij}^{\delta_{12}} - \alpha_{ik}^{\delta_{13}} - \alpha_{jk}^{\delta_{23}} - \alpha_i^{\delta_1} - \alpha_j^{\delta_2} - \alpha_k^{\delta_3} - u$$

Mivel most minden változónak csak két szintje van, ezért $\alpha_i^{\delta_1}$ értékek az azonos delta kitevőkre egymás ellentettjei lesznek. A kétdimenziós $\alpha_{ij}^{\delta_{12}}$ tagok esetében ha $i + j$ ugyanabba a maradékosztályba esik $mod(2)$ akkor ugyanaz az előjelük. Így tovább három és négydimenziósok esetében is.

$$
\begin{array}{lll}
u = -3,1459 & \alpha_{00}^{\delta_{12}} = -0,3197 & \alpha_{000}^{\delta_{123}} = 0,0349 \\
\alpha_0^{\delta_1} = -0,1692 & \alpha_{00}^{\delta_{13}} = 0,1135 & \alpha_{000}^{\delta_{124}} = 0,0097 \\
\alpha_0^{\delta_2} = 0,2908 & \alpha_{00}^{\delta_{14}} = -0,0516 & \alpha_{000}^{\delta_{134}} = 0,0116 \\
\alpha_0^{\delta_3} = -0,3714 & \alpha_{00}^{\delta_{23}} = 0,0204 & \alpha_{000}^{\delta_{234}} = -0,0055 \\
\alpha_0^{\delta_4} = -0,5529 & \alpha_{00}^{\delta_{24}} = -0,0983 & \alpha_{0000}^{\delta_{1234}} = -0,0233 \\
- & \alpha_{00}^{\delta_{34}} = 0,0172 & -
\end{array}
$$

Minden interakciós tagot meghagyva a hozzá tartozó gráf, négy csúcsú teljes gráf lenne, ebből nem tudunk semmit kiolvasni, célunk tehát, minél több interakciót elhagyni. Emlékeztetünk, hogy például egy háromdimenziós $\alpha_{ijk}^{\delta_{123}}$ tagot meghagyva, az összes kétdimenziós részinterakciót is meg kéne tartani, hogy hierarchikus modellt kapjunk.

Azokat a reprezentációs tagokat fogjuk megtartani, amiknek az abszolút értéke "nagy". Hiszen a kicsi abszolút értékűek arra utalnak, hogy nem nagyon tér el a függetlentől ezeknek a változóknak a viszonya. Tartsuk meg tehát a három legnagyobb $\alpha_{ij}^{\delta_{12}}, \alpha_{ik}^{\delta_{13}}, \alpha_{jl}^{\delta_{24}}$ interakciókat, továbbá még azokat, amik a hierarchikus modellhez szükségesek. Egy dekomponálható gráfot szeretnénk készíteni, amire majd a 5.7 segítségével maximum likelihood becslést tudunk készíteni.



6.1. ábra. A $[\delta_1\delta_2, \delta_1\delta_3, \delta_2\delta_4]$ interakciókhoz készített hierarchikus modell.

A kapott irányítatlan gráfban 3.1.1 definíciója alapján az alábbi feltételes függetlenségi viszonyok teljesülnek:

$$
\begin{array}{lll}
\delta_2 \perp\!\!\!\perp \delta_3 \mid \delta_1 & \delta_3 \perp\!\!\!\perp \delta_4 \mid \delta_1 & \delta_3 \perp\!\!\!\perp \delta_4 \mid \delta_2 \\
\delta_3 \perp\!\!\!\perp \delta_4 \mid \delta_1 \cup \delta_2 & \delta_1 \perp\!\!\!\perp \delta_4 \mid \delta_2 &
\end{array}
$$

A táblázat celláiba csak a megfelelő interakciós tagokat írjuk be:

$$\hat{k}_{ijkl} = u + \alpha_i^{\delta_1} + \alpha_j^{\delta_2} + \alpha_k^{\delta_3} + \alpha_l^{\delta_4} + \alpha_{ij}^{\delta_{12}} + \alpha_{ik}^{\delta_{13}} + \alpha_{jl}^{\delta_{24}}$$

Így például:

$$\hat{k}_{0000} = -3,1459 - 0,1692 + 0,2908 - 0,3714 - 0,5529 - 0,3197 + 0,1135 - 0,0983 = -4.2531$$

ahonnan $\exp \hat{k}_{0000} = \hat{p}_{0000} = 0,0142$, felszorozva $N = 694$ értékével a kapott becslés $\hat{n}_{0000} = 9,87$, amit egészre kerekítve 10 az érték. Ezt folytatva az összes cellára megkapjuk az alábbi becsült táblázatot.

Loglineáris reprezentációból származó becsült táblázat

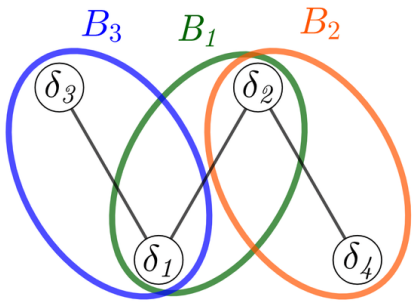
	δ_1	Férfi				Nő			
	δ_2	Fekete		Fehér		Fekete		Fehér	
	δ_4	Halál	Börtön	Halál	Börtön	Halál	Börtön	Halál	Börtön
δ_3	Fekete	10	36	13	33	22	79	7	18
	Fehér	17	66	22	57	56	201	19	48

Ez egy olyan becslés lesz, ami a megadott információk mellett a legkevésbé tér el az egyenletes táblázattól. Ezt úgy értjük, hogy a loglineáris reprezentáció elején minden cellába ugyanazt az u értéket írjuk. Ez egy egyenletes táblázat. Ahogy elkezdjük az α tagokat hozzávenni, a marginálisok megváltoznak. Mivel most csak három loglineáris tagot hagytunk meg, ezért ez egy olyan táblázat, ami csak abban a három tényezőben tér el az egyenletestől. Vagyis ez lesz az a táblázat, aminek a megfelelő marginálisai megegyeznek az eredeti táblázat marginálisaival és az összes ilyen táblázat közül ez "hasonlít" legjobban az egyenletes táblázathoz. Most nézzük meg, hogy a maximum likelihood becsléssel milyen táblázatot kapunk, ha a loglineáris elemzésből kapott dekomponálható gráfra alkalmazzuk.

Legyen δ_i csúcs sorszáma i. Ez a csúcsoknak egy tökéletes felsorolását adja: 6.1 ábra alapján:

$$B_1 = \{1\} \quad H_1 = \{1\} \quad R_1 = \{1\} \quad S_1 = \{\emptyset\}$$
$$B_2 = \{1;2\} \quad H_2 = \{1;2\} \quad R_2 = \{2\} \quad S_2 = \{1\}$$
$$B_3 = \{1;3\} \quad H_3 = \{1;2;3\} \quad R_3 = \{3\} \quad S_3 = \{1\}$$
$$B_4 = \{2;4\} \quad H_4 = \{1;2;3;4\} \quad R_4 = \{4\} \quad S_4 = \{2\}$$

Valóban tökéletes felsorolás, méghozzá olyan, mely tartalmazza az összes klikket, ekkor létezik a klikkeknek tökéletes felsorolása (lásd 4.2.1 állítás). Egy lehetséges felsorolás:



6.2. ábra. A klikkek egy tökéletes felsorolása.

A hozzá tartozó értékek:

$$B_1 = \{1;2\} \quad H_1 = \{1;2\} \quad R_1 = \{1;2\} \quad S_1 = \{\emptyset\}$$
$$B_2 = \{2;4\} \quad H_2 = \{1;2;4\} \quad R_2 = \{4\} \quad S_2 = \{2\}$$
$$B_3 = \{1;3\} \quad H_3 = \{1;2;3;4\} \quad R_3 = \{3\} \quad S_3 = \{1\}$$

Ekkor 5.7 alapján $\hat{p}_{ijkl}$ illetve $\hat{n}_{ijkl}$ maximum likelihood becslések az alábbi módon számolhatóak:

$$\hat{p}_{ijkl} = \frac{n_{ij++}n_{i+k+}n_{+j+l}}{n_{i+++}n_{+j++}n_{++++}}.$$

Például

$$\hat{p}_{0000} = \frac{n_{00++}n_{0+0+}n_{+0+0}}{n_{0+++}n_{+0++}n_{++++}} = \frac{125 * 92 * 104}{250 * 447 * 694} = 0,0154$$

innen $\hat{n}_{0000} = \hat{p}_{0000} \cdot N = 10,70$ kerekítve 11. A táblázat amit így kapunk pedig az alábbi:

Maximum likelihood becslés segítségével kapott táblázat

	δ_1	Férfi				Nő			
	δ_2	Fekete		Fehér		Fekete		Fehér	
	δ_4	Halál	Börtön	Halál	Börtön	Halál	Börtön	Halál	Börtön
δ_3	Fekete	11	36	13	33	21	74	7	18
	Fehér	17	61	22	57	56	201	19	48

Az eredetivel összehasonlítva mindkét becsült táblázat ránézésre megfelelő illeszkedést mutat.

7. fejezet

Hipotézis vizsgálat

Ez a fejezet Satoshi Aoki [4] cikke és David Anderson [5],[6] jegyzete alapján íródott.

7.1. Függetlenség vizsgálat khi-négyzet próbával

Egy kontingencia táblázat ismeretlen paraméterei, hogy mekkora egy adott cellára az a valószínűség, hogy egy tetszőleges eleme a mintának, éppen ide kerül. A táblázatot egy J hosszúságú vektorként elképzelve $\{\theta_i : i \in \{1, 2 \dots J\}\}$ az ismeretlen paraméterek.

$$\Theta = \left\{ \theta_i : \sum_{i=1}^{J} \theta_i = 1 \right\} \tag{7.1}$$

Vagyis a paramétertér $J - 1$ dimenziós, mert ha már $J - 1$ darab θ_i ismert, akkor ez meghatározza az utolsó θ_i értékét is. Ezt telített modellnek hívjuk, ebben az esetben 7.1 az egyetlen feltétel, aminek teljesülnie kell. Legyen x a táblázat egy realizációja, ennek a valószínűsége multinomiális eloszlással számolva

$$P(\mathbf{X} = x) = N! \prod_{i=1}^{J} \frac{1}{x_i!} \cdot \prod_{i=1}^{J} \theta_i^{x_i} \tag{7.2}$$

7.1.1. Definíció. Legyen $\mathbf{X}$ diszkrét valószínűségi vektorváltozó és $T(\mathbf{X})$ statisztika, azaz $\mathbf{X}$ valamely függvénye. Ekkor $T(\mathbf{X})$ elégséges, ha $P(\mathbf{X} = x \mid T(\mathbf{X}) = t)$ feltételes valószínűség a θ_i paraméterektől független.

7.1.1. Állítás. *Legyen $\boldsymbol{X}$ diszkrét valószínűségi vektorváltozó, ekkor $T(\boldsymbol{X})$ elégséges statisztika $\iff$ léteznek g, h függvények, hogy $P(\boldsymbol{X} = x) = g(T(x)), \theta)h(x)$.*

Vagyis megnézzük az együttes eloszlás képletét, szorzatra bontjuk $g(T(x), \theta)h(x)$ alakra, ahol $T(\mathbf{X})$ elégséges statisztika lesz. Ha csak $T(\mathbf{X})$ értékét ismerjük, akkor is éppen annyi információnk van már a θ_i paraméterekről, mintha a teljes táblát ismernénk. Természetesen ez alapján a tábla mindig elégséges statisztika, de célunk a lehető legegyszerűbb elégséges statisztikát megtalálni. Az állítás értelmében az elégséges statisztika a telített táblához maga a tábla lesz, mert mindegyik θ_i érték szerepel 7.2 képletben és a tényezőket nem lehet semmilyen módon összevonni. A maximum likelihood becslést az ismeretlen paraméterekre a kézenfekvő

$\hat{\theta}_i = x_i/N$ fogja adni. Ez a modell a lehető legpontosabban írja le a mintát, viszont a lehető legtöbb, azaz $J - 1$ dimenziós. Ezzel szemben egy független táblázat esetén 7.1 mellett a θ_i paraméter megegyezik i- edik cella megfelelő egydimenziós marginálisainak szorzatával. Ez kétdimenziós $k \times t$ táblázat esetén a jól ismert $\theta_{ij} = \theta_{i+}\theta_{+j}$.

$$P(\mathbf{X} = x) = N! \prod_{i=1}^{k} \prod_{j=1}^{t} \frac{1}{x_{ij}!} \cdot \prod_{i=1}^{k} \prod_{j=1}^{t} \theta_{ij}^{x_{ij}} = N! \prod_{i=1}^{k} \prod_{j=1}^{t} \frac{1}{x_{ij}!} \cdot \prod_{i=1}^{k} \theta_{i+}^{x_{i+}} \prod_{j=1}^{t} \theta_{+j}^{x_{+j}} \tag{7.3}$$

A képletből kiolvasható, hogy független táblázat esetén az elégséges statisztika a megfelelő sor és oszlopösszegek lesznek, a maximum likelihood becslés pedig $\hat{\theta}_{ij} = \frac{x_{i+}x_{+j}}{N^2}$ alakú. Legyen A az a mátrix, hogy $A\mathbf{X} = T(\mathbf{X})$. Ezt konfigurációs mátrixnak nevezzük, egy 3×3 független táblázat esetén:

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{7.4}$$

A telített modellhez pedig A éppen egy $J \times J$ identitás mátrix lenne.

A hipotézis vizsgálatnál azt szeretnénk eldönteni, hogy az ismeretlen θ_i paraméterek milyen paramétertérből származnak, azaz 7.1 mellett teljesül-e rájuk valamilyen plussz feltétel. Jelölje Θ_M azt a paraméterteret, ahol valamilyen függetlenségi kapcsolat áll fenn, például feltételes függetlenég. Ekkor a nullhipotézis: $H_0 : \theta_i \in \Theta_M$ és legyen az ellenhipotézis H_1 egyszerűen az, hogy H_0 nem teljesül. Azt szeretnénk eldönteni az x minta alapján, hogy elfogadható a H_0 hipotézis, vagy sem. Vagyis a mintateret is két részre osztjuk, elhatározzuk előre, hogy milyen mintákat fogunk elfogadni és milyen mintákat fogunk elutasítani, azaz ha $x \in \mathcal{X}_0$ akkor elfogadjuk és ha $x \in \mathcal{X}_1$ akkor elutasítjuk. Természetesen $\mathcal{X}_0 \cap \mathcal{X}_1$ üres és egyesítésük kiadja a teljes mintateret. Ha döntöttünk, akkor a döntésünk függvényében az alábbi hibákat véthetjük.

7.1.2. Definíció. Elsőfajú hibának nevezzük azt, ha tévesen elutasítottuk a nullhipotézist. Másodfajú hibának nevezzük azt, ha tévesen elfogadtuk a nullhipotézist.

A döntésünket nagyban befolyásolja, hogy legfeljebb mekkora valószínűséggel engedünk meg első illetve másodfajú hibát. Természetesen, ha valaki nem enged meg egyáltalán elsőfajú hibát, vagyis azt szeretné, hogy sose forduljon elő, hogy tévesen elutasítja a nullhipotézist, akkor nagyon egyszerű dolga van, mindig elfogadja a nullhipotézist. Hasonlóan ha mindig elutasítja, akkor másodfajú hibát nem fog véteni.

7.1.3. Definíció. A próba terjedelmének nevezzük azt a legnagyobb valószínűségét, hogy elsőfajú hibát vétünk, azaz

$$\alpha := \sup_{\theta \in \Theta_M} P_\theta(X \in \mathcal{X}_1)$$

ahol H_0 az alsóindexben jelzi, hogy az igazság a H_0 volt.

Az alkalmazásokban általában α értéke kicsi, 0,05 vagy 0,01, vagyis az elsőfajú hiba valószínűségét kicsire állítjuk be. Ennek az a következménye, hogy ha a próbánk alapján azt kapjuk, hogy H_0 elfogadható, az csupán annyit jelent, hogy nem mond nagyon ellent a minta a nullhipotézisünknek. Viszont ha H_0 elutasítása az eredmény, akkor ez egy erős bizonyíték H_1 mellett.

A legtermészetesebb az, ha kiszámoljuk, hogy H_0 feltételezésünk mellett mik azok az értékek, amik a legvalószínűbbek az egyes cellákba, majd megnézzük, hogy a tényleges minta hogyan viszonyul ehhez.

7.1.2. Állítás. *A $\theta_i \in \Theta_M$ nullhipotézis mellett elég nagy minta esetén a $T(\boldsymbol{X})$ statisztika eloszlása aszimptotikusan tart a megfelelő szabadságfokú khi-négyzet eloszláshoz.*

$$\lim_{N\to\infty} P(T(\boldsymbol{X}) > c_\alpha) = P(Y > c_\alpha) \tag{7.5}$$

ahol Y valószínűségi változó χ^2 eloszlású megfelelő szabadsági fokkal és

$$T(\boldsymbol{X}) = \sum_{i=1}^{J} \frac{(X_i - \hat{X}_i)^2}{\hat{X}_i} \tag{7.6}$$

ahol $\hat{X}_i$ a H_0 melletti maximum likelihood becslés a cella értékeire.

Mivel a döntésünk nagyban függ α választásától, ezért egy rögzített α értékre ha el tudjuk fogadni a H_0 hipotézist, akkor minden α-nál kisebb terjedelem esetén is elfogadhatjuk automatikusan. Adódik a kérdés, hogy melyik az a legnagyobb α érték, amire még elfogadhatjuk a nullhipotézist, ezt a legnagyobb α számot p'-értéknek szokás nevezni.

$$p' = P(Y > T(x)) = 1 - P(Y \leq T(x)) \tag{7.7}$$

Ami a megfelelő szabadsági fokú χ^2 eloszlás táblázatából visszakereshető. Azonban ha a minta nem túl nagy, akkor a módszer nem megbízható, sőt, olyan esetek is előfordulnak, hogy viszonylag nagy minta esetén is pontatlan.

7.2. Egzakt számolási módszer

Aszimptotikus vizsgálat mellett más módszerrel is meg lehet határozni ezt a p' értéket, azonban a 6. fejezetben bemutatott négydimenziós tábla már túl nagy példa, így most egy kisebbet fogunk tekinteni a jobb érthetőség kedvéért. Tekintsünk egy 3×3 táblázatot, jelölje ezt $x' \in \mathbf{N}^9$ és legyen H_0 a függetlenség, vagyis $\theta_{ij} = \theta_{i+}\theta_{+j}$. Az elégséges statisztika ekkor x'_{i+} és x'_{+j} a megfelelő marginálisok, vagyis a konfigurációs mátrix éppen 7.4 lesz. Rögzítsük $Ax' = t$ marginálisokat. Azokat az x táblázatokat, melyekre $Ax = Ax' = t$ teljesül, jelöljük $F_{A,t}$-vel, azaz $F_{A,t} = \{x \in \mathbf{N}^9 : Ax = t\}$.

Mindegyik $x \in F_{A,t}$ táblára meghatározzuk, hogy feltéve, hogy tudjuk mik lesznek a marginális értékek, (most sor és oszlop összegek), mi a valószínűsége, hogy éppen ezt a táblázatot kaptuk H_0 mellett.

7.2.1. Állítás. *Legyen $x' \in \mathbf{N}^{k\times r}$ rögzített táblázat és legyen $Ax' = t$. Ekkor H_0 mellett feltéve, hogy ismerjük AX értékét, annak a valószínűsége, hogy $X = x$, ahol $x \in F_{A,t}$, az éppen*

$$h(x) = P(X = x \mid AX = t, H_0) = \frac{\left(\prod_{i=1}^{k}(x_{i+})!\right)\left(\prod_{j=1}^{r}(x_{+j})!\right)}{N! \prod_{i,j=1}^{k,r}(x_{ij})!}. \tag{7.8}$$

Ekkor a p' értéket az alábbi módon számolhatjuk

$$p' = P(T(X) \geq T(x') \mid AX = t, H_0) = \sum_{x \in F_{A,t}} g(x)h(x) \tag{7.9}$$

ahol $h(x)$ az imént definiált 7.8, $g(x)$ pedig annak az indikátora, hogy az x tábla khi-négyzet statisztikájának értéke nagyobb mint az x' mintáé, azaz

$$g(x) = \begin{cases} 1, & T(x) \geq T(x') \\ 0, & T(x) < T(x') \end{cases} \tag{7.10}$$

ahol T a khi-négyzet statisztika értéke 7.6.

7.2.1. Példa. *Legyen x' a legelső tábla és legyen függetlenség a nullhipotézis, ekkor $F_{A,t}$ 15 elemű, $t = [3, 2, 1, 2, 2, 2]$. Az a kérdés, hogy a legelső tábla mennyire valószínű a függetlenség*

2	1	0	3
0	1	1	2
0	0	1	1
2	2	2	

2	1	0
0	0	2
0	1	0

2	0	1
0	2	0
0	0	1

2	0	1
0	1	1
0	1	0

1	2	0
1	0	1
0	0	1

1	2	0
0	0	2
1	0	0

1	0	2
1	1	0
0	1	0

1	0	2
0	2	0
1	0	0

1	1	1
1	1	0
0	0	1

1	1	1
1	0	1
0	1	0

1	1	1
0	1	1
1	0	0

0	2	1
2	0	0
0	0	1

0	1	2
2	0	0
0	1	0

0	2	1
1	0	1
1	0	0

0	1	2
1	1	0
1	0	0

7.1. ábra. Egy rögzített t marginálisok vektorához tartozó $F_{A,t}$ táblák halmaza

mellett a többi $F_{A,t}$-beli táblához képest.

Legyen $T(x)$ 7.6 a khi-négyzet statisztika, ekkor sorfolytonosan az értékei:

$$T(x_1) = 5 \qquad T(x_2) = 8 \qquad T(x_3) = 8 \qquad T(x_4) = 5 \qquad T(x_5) = 5$$
$$T(x_6) = 8 \qquad T(x_7) = 5 \qquad T(x_8) = 8 \qquad T(x_9) = 3 \qquad T(x_{10}) = 3$$
$$T(x_{11}) = 3 \qquad T(x_{12}) = 8 \qquad T(x_{13}) = 8 \qquad T(x_{14}) = 5 \qquad T(x_{15}) = 5$$

$x' =$

2	1	0	3
0	1	1	2
0	0	1	1
2	2	2	

$x^0 =$

$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{6}$
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

$6 \cdot x^0 =$

1	1	1	3
$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	2
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1
2	2	2	

7.2. ábra. Az x' minta, a középső x^0 a H_0 szerinti maximum likelihood becslés a valószínűségekre, illetve a maximum likelihood becslés a várt értékekre szerepel a harmadik táblán.

A 7.8 képletből számított $h(x)$ értékek pedig:

$$h(x_1) = 0,0667 \quad h(x_2) = 0,0333 \quad h(x_3) = 0,0333 \quad h(x_4) = 0,0667 \quad h(x_5) = 0,0667$$
$$h(x_6) = 0,0333 \quad h(x_7) = 0,0667 \quad h(x_8) = 0,0333 \quad h(x_9) = 0,1333 \quad h(x_{10}) = 0,1333$$
$$h(x_{11}) = 0,1333 \quad h(x_{12}) = 0,0333 \quad h(x_{13}) = 0,0333 \quad h(x_{14}) = 0,0667 \quad h(x_{15}) = 0,0667$$

Innen a p' értékek: 7.9 képlet alapján alapján $x' = \{x_9, x_{10}, x_{11}\}$ táblák esetén $p' = 1$, azaz ilyen táblákat tetszőleges α terjedelem mellett is elfogadunk. Ha $x' = \{x_1, x_4, x_5, x_7, x_{14}, x_{15}\}$, akkor $p' = 0,6$, tehát ezeket a táblákat $\alpha \leq 0,6$ terjedelem esetén fogadjuk el. A fennmaradó $x' = \{x_2, x_3, x_6, x_8, x_{12}, x_{13}\}$ táblákra $p' = 0,2$ a kapott érték.

A módszer hátránya, hogy nagyobb táblázatok esetén $F_{A,t}$ elemszáma exponenciálisan nő. Másrészt 7.8 képletet is általánosítani kell, hogy a függetlenség általánosabb eseteit is le tudjuk fedni, például feltételes függetlenséget.

7.3. Monte Carlo módszer

Röviden és vázlatosan bemutatunk még egy módszert a p' érték meghatározására. Legyen adott egy x' tábla és A konfigurációs mátrix. Legyen $g(x)$ a korábban definiált 7.10 függvény és H_0 hipotézis pedig az, hogy $\theta_i \in \Theta_M$. Ekkor számítsuk a p' értéket az alábbi módon:

$$p' = \sum_{i=1}^{r} g(x_i)/r \tag{7.11}$$

ahol $r \in \mathbf{N}$ tetszőlegesen nagy lehet, x_i pedig a nullhipotézis szerinti feltételes eloszlásból generált tábla. Az egzakt számoláshoz hasonlóan itt is a H_0 hipotézis szerinti $F_{A,t}$ táblákhoz hasonlítjuk, de nem egyesével az összeshez, hanem generálunk elég sokat és a $\frac{\text{kedvező}}{\text{összes}}$ alapján következtetünk p' értékére. A feladat tehát az, hogy generáljunk sok táblát a H_0 eloszlás szerint. Ezt Markov lánc segítségével fogjuk megtenni.

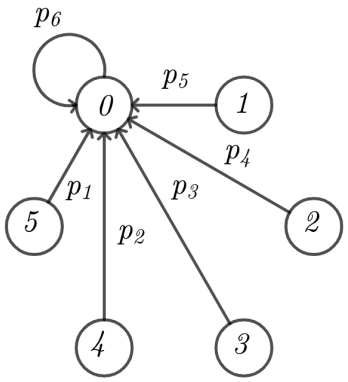
7.3.1. Definíció. Legyenek $X_k : k \in \mathbf{N}$ diszkrét valószínűségi változók, melyek az S véges halmazból vehetnek fel értékeket. Markov láncnak nevezzük az X_k valószínűségi változók sorozatát, ha

$$P(X_k = y \mid X_0 = x_0, X_1 = x_1, \ldots, X_{k-2} = x_{k-2}, X_{k-1} = x) = P(X_k = y \mid X_{k-1} = x) = p(x, y)$$

teljesül $\forall y, x, x_0, x_1 \dots x_{k-1} \in S$ esetén. Jelölje Q azt a mátrixot, melyre $q_{ij} = p(x_i, x_j)$. Nevezzük ezt átmeneti mátrixnak.

Markov lánc esetén egy adott állapot mindig csak az eggyel előtte lévő állapottól függ.

7.3.2. Példa. *Képzeljük el, hogy egy hamis dobókockával dobálunk, ahol az i-es dobás valószínűsége p_i. Mi a valószínűsége, hogy az első k dobás összege osztható 6-al?*



7.3. ábra. A gráf csúcsai a lehetséges osztási maradékok

Legyen X_k valószínűségi változó jelentése, hogy az első k dobást összeadva mennyi a maradék 6-al osztva. Kérdés $P(X_k = 0)$ valószínűsége. Itt az nem számít, hogy az első $k-2$ dobások után mik voltak az osztási maradékok, csak az, hogy X_{k-1} -ben mennyi és mi lesz az utolsó dobás.

$$P(X_k = 0 \mid X_0 = x_0 \dots X_{k-1} = x_{k-1}) = P(X_k = 0 \mid X_{k-1} = x_{k-1}) = p(x_{k-1}, 0) = q_{x_{k-1},0}.$$

Ekkor a feladathoz tartozó átmeneti Q mátrix

$$Q = \begin{bmatrix} p_6 & p_1 & p_2 & p_3 & p_4 & p_5 \\ p_5 & p_6 & p_1 & p_2 & p_3 & p_4 \\ p_4 & p_5 & p_6 & p_1 & p_2 & p_3 \\ p_3 & p_4 & p_5 & p_6 & p_1 & p_2 \\ p_2 & p_3 & p_4 & p_5 & p_6 & p_1 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{bmatrix}$$

ahol q_{ij} jelentése, hogy i volt az osztási maradék, mennyi a valószínűsége, hogy a következő dobás után j lesz az osztási maradék. Vegyük $\pi_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ kezdeti vektort, hiszen a 0-dik dobás után az osztási maradék nyilván 0 lesz. Ekkor $\pi_1 = \pi_0 Q = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{bmatrix}$, ahol $\pi_1(i)$ annak a valószínűségét jelenti, hogy az első dobás után i az osztási maradék. Így tovább $\pi_2 = \pi_1 Q = \pi_0 Q^2, \pi_3 = \pi_0 Q^3, \dots, \pi_k = \pi_0 Q^k$.

7.3.3. Tétel. *Legyen Q az átmeneti mátrixa egy $X_k : k \in \mathbf{N}$ Markov láncnak, egy kezdeti π_0 eloszlás mellett.*

$$P(X_n = y) = y\text{-adik eleme a } \pi_0 Q^n \text{ vektornak}$$
$$P(X_n = y \mid X_0 = x) = x\text{-edik sor } y\text{-adik eleme lesz a } Q^n \text{ mátrixnak}$$

Bizonyítás. Tegyük fel, hogy $S = \{1, 2, \ldots, r\}$. Legyenek $W_i : i \in \{1, 2 \ldots n\}$ vektorok:

$W_1 = \Big[P(X_1 = 1), \quad \ldots \quad , P(X_1 = r)\Big], \ldots, W_n = \Big[P(X_n = 1), \quad \ldots \quad , P(X_n = r)\Big]$.

Mivel $W_k = \pi_0 Q^k$ így az első állítást igazoltuk.

Most legyen $\pi_0(x) = 1$ és $\pi_0(z) = 0$ ha $z \neq x$. Mivel $W_k = \pi_0 Q^k$, így W_k éppen Q^k mátrix x-edik sora lesz. □

7.3.4. Definíció. Egy Markov lánc irreducibilis, ha $\forall i, j$ esetén pozitív valószínűséggel el lehet jutni x_i állapotból x_j állapotba.

7.3.5. Definíció. Egy Markov lánc reguláris, ha Q átmeneti mátrixának van olyan $k \in \mathbf{N}$ kitevője, hogy Q^k mátrixnak csak pozitív eleme van.

7.3.6. Tétel. *Legyen X_k reguláris Markov lánc, $S = \{1, 2 \ldots r\}$ és Q átmeneti mátrix. Ekkor egyértelműen létezik olyan $r \times r$-es W mátrix, aminek minden sora ugyanaz a szigorúan pozitív w vektor és*

$$\lim_{k\to\infty} Q^k = W$$
$$\lim_{k\to\infty} P(X_k = y \mid X_0 = x) = w_y$$

Ez az egyértelmű w vektor lesz a megoldása az $x = xQ$ egyenletrendszernek. Továbbá, ha u egy r hosszú tetszőleges valószínűségi eloszlás, akkor $\lim_{k\to\infty} uQ^k = w$ is teljesül. Ezt a w vektort nevezzük a Markov lánc stacionárius eloszlásainak.

Hozzunk létre egy reguláris Markov láncot az $F_{A,t}$ halmaz felett, aminek legyen $h(x) : x \in F_{A,t}$ a stacionárius eloszlása, azaz Q átmeneti mátrix esetén $hQ = h$ egyenlőség fennáll, ahol h az a vektor, aminek koordinátái a $h(x)$ valószínűségek.

7.3.7. Tétel. *(Metropolis-Hastings algoritmus) Legyen π tetszőleges valószínűségi eloszlás az $F_{A,t}$-n. Legyen R egy tetszőleges átmeneti mátrix egy reguláris Markov lánchoz az $F_{A,t}$-n. Legyen*

$$q_{ij} = r_{ij} \min\left\{1, \frac{\pi_j r_{ji}}{\pi_i r_{ij}}\right\} \quad i \neq j \tag{7.12}$$

és q_{ii} a megfelelő valószínűség, hogy a sorösszeg 1-et adjon. Ekkor Q olyan átmeneti mátrix lesz, hogy $\pi Q = \pi$.

Vagyis ha már készítettünk egy reguláris Markov láncot az $F_{A,t}$ halmazon, aminek az átmeneti mátrixa R, akkor egy egyszerű transzformációval megkaphatjuk azt a Q átmeneti mátrixot,

melynek már $h(x)$ vektor lesz a stacionárius eloszlása. Tekintsük ismét a 7.2.1 példában szereplő ábrát. Figyeljük meg, hogy milyen megengedett "lépéseket" tehetünk hogy egy $x_i \in F_{A,t}$ táblából, hogy egy $x_j \in F_{A,t}$ táblába kerüljünk. Ebben a példában az A konfigurációs mátrix éppen 7.4, így olyan $z \in \mathbf{Z}^9$ vektorokat keresünk, amelyre $A(x+z) = A(x)+A(z) = A(x)$ teljesül, vagyis $z \in \ker(A)$.

Legyen $G_{A,B,t}$ olyan irányítatlan gráf, hogy csúcsai legyenek $x \in F_{A,t}$ táblák és legyen él $x, y \in F_{A,t}$ csúcsok között, ha $x - y \in B$ vagy $y - x \in B$. Megjegyezzük, hogy $Ax = t$ és $Ay = t$ esetén $A(x - y) = A(y - x) = 0$, azaz $x - y$ és $y - x$ is $\in \ker_{\mathbf{Z}}(A)$.

7.3.8. Definíció. (Markov bázis) Egy $B \subset \ker_{\mathbf{Z}}(A)$ halmaz Markov bázis az A mátrixra nézve, ha tetszőleges $t \in \mathbf{N}$ vektor esetén $G_{A,B,t}$ gráf összefüggő.

Tehát egy adott A konfigurációs mátrixra és rögzített t marginálisokra meghatározzuk az $F_{A,t}$ halmazt, majd addig választunk $z \in \ker_{\mathbf{Z}}(A)$ lépéseket a B halmazba, míg B Markov bázis nem lesz az A konfigurációs mátrixra nézve. Annak ellenőrzése, hogy adott lépéseket bevéve már Markov bázist kaptunk-e, túlmutat ennek a dolgozatnak a keretein. A lényeg, hogy legyen $Z_0 = x_i$ egy tetszőleges táblázat az $F_{A,t}$-ból, ekkor Z_m eloszlása elég nagy m esetén $h(x)$-et fogja közelíteni, így $Z_m, Z_{m+1} \ldots$ táblázatok tekinthetők H_0 szerinti mintának, amivel a p' értéket 7.11 segítségével kiszámolhatjuk.

Irodalomjegyzék

[1] Rudas Tamás, *Kontingencia táblák elemzése*, Nemzeti Tankönyvkiadó, Budapest, 1993.

[2] Steffen L. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, 1996.

[3] Christopher M. Bishop, *Pattern Recognition and Machine Learning* , Springer, Berlin, 2006.

[4] Satoshi Aoki, *An introduction to computational algebraic statistics*, 2017.
https://doi.org/10.48550/arXiv.1607.07600
Hozzáférés: 2022. március 20., 18:00

[5] David Anderson, lecture notes, Math331, Fall 2008
https://people.math.wisc.edu/ valko/courses/331/MC1.pdf
Hozzáférés: 2022. március 20., 18:00

[6] David Anderson, lecture notes, Math331, Fall 2008
https://people.math.wisc.edu/ valko/courses/331/MC2.pdf
Hozzáférés: 2022. március 3., 09:40

[7] Philipp Henning, *Probabilistic Machine Learning*, lecture 3, Universität Tübingen.
Probabilistic Machine Learning, lecure 3
Hozzáférés: 2021. november 3., 16:00

[8] Philipp Henning, *Probabilistic Machine Learning*, lecture 16, Universität Tübingen.
Probabilistic Machine Learning, lecture 16
Hozzáférés: 2021. november 3., 16:00