

Segédanyag az Idősorok és többdimenziós statisztika gyakorlat (matematika BSc, matematikai elemző szakirány) tantárgyhoz.

2023. december 5.

Definíció. Sztochasztikus folyamat: $(X_t)_{t \in T}$, ahol T a paraméterter és minden t -re X_t valószínűségi változó.

A sztochasztikus folyamat **diszkrét paraméterű** (vagy diszkrét idejű), ha T számossága legfeljebb megszámlálhatóan végtelen, tipikusan $T = \mathbb{Z}$ vagy $T = \mathbb{Z}_+$. A sztochasztikus folyamat **folytonos paraméterű** (vagy folytonos idejű), ha T kontinuum számosságú, jellemzően $T = [0; 1]$, $T = \mathbb{R}$ vagy $T = \mathbb{R}_+$. Néhány nevezetes sztochasztikus folyamat:

- Poisson-folyamat: folytonos paraméterű
- Markov-folyamat: diszkrét vagy folytonos idejű, a gyakorlaton csak a diszkrét idejűekkel foglalkozunk
- Wiener-folyamat: folytonos paraméterű
- idősortmodellek (autoregresszív, mozgóátlag folyamatok): diszkrét idejű

Definíció. Gauss-folyamat: olyan sztochasztikus folyamat, melynek bármely véges számú peremeloszlása együttesen normális eloszlású, azaz minden $n \in \mathbb{Z}_+$, $t_1 \in T, \dots, t_n \in T$ esetén $(X_{t_1}, \dots, X_{t_n})$ együttesen normális eloszlású.

Diszkrét idejű Markov-folyamatok (Markov-láncok)

Legyen S megszámlálható halmaz, neve: állapottér; X_0, X_1, X_2, \dots valószínűségi változó sorozat, $P(X_i \in S) = 1$ minden i -re. Az állapottér elemeire sokszor egyszerűbb az $1, 2, \dots$ számokként gondolni.

Definíció. Markov-tulajdonság. Minden $0 \leq n \in \mathbb{Z}$ és $i_j \in S$, $j = 1, 2, \dots$ esetén $P(X_{n+1} = i_{n+1} | X_n = i_n, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n)$.

Definíció. Markov-folyamat/lánc. A Markov-tulajdonsággal rendelkező diszkrét idejű sztochasztikus folyamatokat Markov-folyamatoknak vagy Markov-láncoknak hívjuk.

Definíció. Homogén Markov-lánc. A Markov-lánc homogén (vagy stacionárius), ha a $P(X_{n+1} = j | X_n = i)$ feltételes valószínűség nem függ n -től, azaz minden n -re ugyanaz.

Jel. a kezdeti eloszlást $\mathbf{q}^T = (q_1, q_2, \dots)^T$, ahol $q_i = P(X_0 = i)$, $i \in S$.

Jel. homogén Markov-lánc esetén $p_{i,j} := P(X_{n+1} = j | X_n = i)$, $i, j \in S$. Ezek neve: egylépéses átmenetvalószínűségek vagy röviden átmenetvalószínűségek.

Jel. $\mathbf{P} = (p_{i,j})_{i,j \in S}$, neve: **átmenetvalószínűség mátrix**

Jel. homogén Markov-lánc esetén $p_{i,j}^{(n)} := P(X_n = j | X_0 = i)$, $i, j \in S$. Ezek neve: n lépéses átmenetvalószínűségek.

Jel. $\mathbf{P}^{(n)} = (p_{i,j}^{(n)})_{i,j \in S}$, neve: n lépéses átmenetvalószínűség mátrix

Tétel. Chapman-Kolmogorov egyenletek. Ha $m < k < n$, akkor

$$P(X_n = i_n | X_m = i_m) = \sum_{i_k \in S} P(X_n = i_n | X_k = i_k) \cdot P(X_k = i_k | X_m = i_m).$$

Következmény. $\mathbf{P}^{(n)} = \mathbf{P}^n$

Állítás. A \mathbf{P} mátrix minden sorösszege 1.

Állítás. $P(X_n = j) = \sum_{i \in S} q_i \cdot (\mathbf{P}^n)_{i,j} = (\mathbf{q}^T \mathbf{P}^n)_j$, $j \in S$

Definíció. i állapotból j állapot **elérhető**, ha $\exists n \geq 1$: $p_{i,j}^{(n)} > 0$. Jel.: $i \rightarrow j$

Definíció. i és j **érintkeznek**, ha vagy $i = j$, vagy $(i \rightarrow j$ és $j \rightarrow i)$. Jel.: $i \leftrightarrow j$

Állítás. Az elérhetőség tranzitív, az érintkezés pedig ekvivalenciareláció.

Definíció. Pontosan n lépésben visszatérés valószínűsége.

$$f_n(i, i) := P(X_n = i, X_{n-1} \neq i, \dots, X_1 \neq i | X_0 = i), \quad i \in S$$

Definíció. Visszatérés valószínűsége. $f(i, i) := \sum_{n=1}^{\infty} f_n(i, i)$, $i \in S$

Állítás. $p^{(n)}(i, i) = f_n(i, i) + \sum_{k=1}^{n-1} f_k(i, i) p^{(n-k)}(i, i)$

Definíció. Visszatérő állapot. Az i állapot visszatérő, ha $\forall j \in S$ -re $i \rightarrow j \Rightarrow j \rightarrow i$.

Megjegyzés. Ekvivalens elnevezések: visszatérő=lényeges=perzisztens állapot.

Definíció. Átmeneti állapot. Az i állapot átmeneti, ha nem visszatérő.

Megjegyzés. Ekvivalens elnevezések: átmeneti=lényegtelen=tranziciens állapot

Tétel. Az i állapot visszatérő $\iff f(i, i) = 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i, i) = \infty$

Következmény. Az i állapot átmeneti $\iff f(i, i) < 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i, i) < \infty$

Definíció. Elnyelő állapot. Az i állapot elnyelő, ha $p_{i,i} = 1$.

Definíció. Az i állapot periódusa: $d(i) := \text{lnc} \{n \geq 0 : p^{(n)}(i, i) > 0\}$

Definíció. Periodikus állapot. Az i állapot periodikus, ha $d(i) > 1$.

Definíció. Aperiodikus állapot. Az i állapot aperiodikus, ha $d(i) = 1$.

Definíció. Irreducibilis ML: az állapotai érintkeznek egymással.

Megjegyzés. Az irreducibilis Markov-lánc gráfja összefüggő.

Definíció. Ergodikus ML: irreducibilis, minden állapota visszatérő és aperiodikus.

Tétel. Legyen \mathbf{P} egy ergodikus Markov-lánc átmenetvalószínűség mátrixa.

$$\text{Ekkor } \forall i, j \in S\text{-re } p^{(n)}(i, j) \xrightarrow{n \rightarrow \infty} \frac{1}{\sum_{n=1}^{\infty} n f_n(j, j)}.$$

Vegyük észre, hogy az előző tétel alapján amihez konvergál, már nem függ a kiinduló

i állapotól. Jelölje $\pi_j := \frac{1}{\sum_{n=1}^{\infty} n f_n(j,j)}$ $j = 1, 2, \dots$, ezzel $\lim_{n \rightarrow \infty} \mathbf{P}^n = \begin{pmatrix} \pi_1 & \pi_2 & \dots \\ \vdots & \vdots & \dots \\ \pi_1 & \pi_2 & \dots \end{pmatrix}$.

Jel. $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \dots)^T$, elnevezése: **stacionárius** vagy egyensúlyi **eloszlás**. A stacionárius eloszlás mutatja meg, hogy "hosszú idő után" milyen valószínűséggel leszünk a Markov-lánc egyes állapotaiban.

A gyakorlatban a stacionárius eloszlást az alábbi egyenletrendszer megoldásával szokás kiszámolni: $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{P}$, ahol $\sum_i \pi_i = 1$. Ennek értelmében tehát a $\boldsymbol{\pi}$ vektor a \mathbf{P} mátrix baloldali, 1-re normált sajátvektora.

Markov-láncoknál egy lényeges kérdés, hogy átlagosan mennyi időbe (lépésbe) telik, míg az egyik állapotból egy másik állapotba eljutunk. Jelölje $m_{i,j}$: ha jelenleg az i állapotban vagyunk, akkor várhatóan ennyi lépésre van szükség, hogy a j állapotba kerüljünk. Általánosan ezeket az értékeket nem lehet közvetlenül egyszerűen kiszámolni, de a teljes várható érték tétel alapján felírható rájuk a következő egyenlet: $m_{i,j} = p_{i,j} + \sum_{k \neq j} p_{i,j}(1 + m_{k,j})$, amit $m_{i,j} = 1 + \sum_{k \neq j} p_{i,j} m_{k,j}$ -ra lehet egyszerűsíteni.

Állítás. Ergodikus Markov-lánc esetén $m_{i,i} = \frac{1}{\pi_i}$

Elyelő Markov-láncok:

- van s tranziens állapot: t_1, \dots, t_s
- van m elnyelő állapot: a_1, \dots, a_m

Particionáljuk ezek alapján az átmenetvalószínűség mátrixot: $\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}$, ahol

$\mathbf{Q} \in \mathbb{R}^{s \times s}$, $\mathbf{R} \in \mathbb{R}^{s \times m}$, \mathbf{I}_m az m dimenziós egységmátrix.

Néhány lényeges mennyiség kiszámítása:

- ha t_i -ben vagyunk, akkor azon periódusok várható száma, amit t_j -ben töltünk, mielőtt egy elnyelő állapotba lépnénk: $((\mathbf{I}_s - \mathbf{Q})^{-1})_{i,j}$
- ha t_i -ben vagyunk, akkor annak a valószínűsége, hogy a_j -be kerülünk: $((\mathbf{I}_s - \mathbf{Q})^{-1} \mathbf{R})_{i,j}$

Elyelő láncok esetén nem beszélhetünk olyan értelemben stacionaritásról, mint az ergodikus láncoknál, egyfajta egyensúly csak akkor érhető el, ha minden időszakban van(nak) új belépő(k) a rendszerbe. Tekintsük az n -edik időperiódust az $n-1$ -edik és az n -edik időpont között eltelt időnek, $n = 1, 2, \dots$

Vezessünk be jelöléseket:

- H_i : az egyes időperiódusok elején az i -edik állapotba belépő egyedek száma
- $N_i(n)$: az n -edik időperiódus elején az i -edik állapotban lévő egyedek száma
- $r_{i\bullet} = \sum_{j=1}^m r_{i,j}$, ami az i -edik állapotból egy elnyelő állapotba lépés valószínűsége

$$\bullet \tilde{\mathbf{Q}} := \left(\begin{array}{c|c} \mathbf{Q} & \begin{matrix} r_{1\bullet} \\ \vdots \\ r_{s\bullet} \end{matrix} \end{array} \right)$$

Kérdés, hogy léteznek-e a $\lim_{n \rightarrow \infty} N_i(n)$ határértékek. Ha léteznek, akkor jelöljük őket N_i -vel, amikből képezzük az $\mathbf{N} = (N_1, \dots, N_s)^T$ egyensúlyi egyedszám vektort. Amennyiben létezik ilyen egyensúlyi helyzet, akkor minden időperiódusban az i -edik állapotba belépő egyedek számának (a lenti egyenletben a baloldal) meg kell egyeznie az onnan kilépő egyedek számával (jobboldal):

$$H_i + \sum_{k \neq i} N_k \cdot \tilde{q}_{k,i} = N_i \cdot (1 - \tilde{q}_{i,i}) \quad i = 1, \dots, s$$

Ajánlott irodalom: Wayne L. Winston: Operációkutatás, 17. fejezet

Definíció. X **val.változó eloszlásfüggvénye**: $F_X(x) = P(X < x)$.

Állítás. Az eloszlásfüggvény tulajdonságai:

- $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$;
- balról folytonos;
- monoton növvő.

Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	D ² X
Karakterisztikus (indikátor vált.)	Ind(p)	$P(X = 1) = p$ $P(X = 0) = 1 - p$	p	$p(1 - p)$
Geometriai (Pascal)	Geo(p)	$P(X = k) = p(1 - p)^{k-1}$ $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	Hipgeo(N, M, n)	$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$ $k = 0, 1, \dots, n$	$n \frac{M}{N}$	$n \frac{M}{N} (1 - \frac{M}{N}) (1 - \frac{n-1}{N-1})$
Binomiális	Bin(n, p)	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, 1, \dots, n$	np	$np(1-p)$
Negatív binomiális	NegBin(n, p)	$P(X = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$ $k = n, n+1, \dots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	Poi(λ)	$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ $k = 0, 1, \dots$	λ	λ

Nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D ² X
Egyenletes	E(a, b)	$\begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a < x \leq b \\ 1 & \text{ha } b < x \end{cases}$	$\begin{cases} \frac{1}{b-a} & \text{ha } a < x \leq b \\ 0 & \text{különben} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális	Exp(λ)	$\begin{cases} 1 - e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard norm.	N($0, 1^2$)	$\Phi(x) = \dots$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ $x \in \mathbb{R}$	0	1
Normális	N(m, σ^2)	\dots	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ $x \in \mathbb{R}$	m	σ^2

További nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D ² X
Cauchy	$Cauchy(a, b)$ $a \in \mathbb{R}, b > 0$	$\frac{1}{\pi} \arctg\left(\frac{x-a}{b}\right) + \frac{1}{2}$	$\frac{1}{\pi b \left[1 + \left(\frac{x-a}{b}\right)^2\right]}$ $x \in \mathbb{R}$	\nexists	\nexists
Pareto*	$Pareto(\alpha, \beta)$ $a, b > 0$	$\begin{cases} 1 - \left(\frac{\beta}{x}\right)^\alpha & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\begin{cases} \frac{\alpha}{\beta} \left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha > 1$, szórásnégyzete pedig akkor, ha $\alpha > 2$.

Eloszlás neve	Jelölése	Eloszlás-függvény	Sűrűségfüggvény	EX	D ² X
Khf-négyzet	χ_k^2 $k \in \mathbb{N}$...	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$ $x \in \mathbb{R}$	k	$2k$
Gamma	$\Gamma(\alpha, \lambda)$ $\alpha, \lambda > 0$...	$\begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha e^{-\lambda x} x^{\alpha-1} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	$Beta(\alpha, \beta)$ $\alpha, \beta > 0$...	$\begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & x \in [0; 1] \\ 0 & \text{különben} \end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Lognormális	$LN(m, \sigma^2)$ $m \in \mathbb{R}, \sigma > 0$...	$\begin{cases} \frac{1}{x\sqrt{2\pi\sigma}} e^{-\frac{(\log x - m)^2}{2\sigma^2}} & \text{ha } x > 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2m+\sigma^2}$

Többváltozós valószínűségszámítás

Definíció. Valószínűségi vektorváltozó: $\mathbf{X}: \Omega \rightarrow \mathbb{R}^d$ (Borel-)mérhető függvény, azaz amire $\{\omega : \mathbf{X}(\omega) \in B\} \in \mathcal{A}$ minden $B \subseteq \mathbb{R}^d$ nyílt (Borel-)halmazra.

Definíció. X valószínűségi vektorváltozó eloszlásfüggvénye:

$$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} < \mathbf{x}) = P(X_1 < x_1, \dots, X_d < x_d).$$

Definíció. X valószínűségi vektorváltozó abszolút folytonos, ha létezik olyan $f_{\mathbf{X}}(x_1, \dots, x_d)$ függvény, amelyre

$$F_{\mathbf{X}}(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} f_{\mathbf{X}}(t_1, \dots, t_d) dt_1 \dots dt_d.$$

Ilyenkor $f_{\mathbf{X}}(\mathbf{x})$ -et **sűrűségfüggvénynek** hívjuk.

$d = 2$ esetén vezessük be a következő jelöléseket és elnevezéseket:

- $F_{X,Y}(x, y) = P(X < x, Y < y) \rightsquigarrow$ együttes eloszlásfüggvény
- $F_X(x) = P(X < x) \rightsquigarrow$ peremeloszlásfüggvények
- $F_Y(y) = P(Y < y) \rightsquigarrow$ peremeloszlásfüggvények
- $f_{X,Y}(x, y) \rightsquigarrow$ együttes sűrűségfüggvény
- $f_X(x), f_Y(y) \rightsquigarrow$ peremsűrűségfüggvények

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y) \quad \text{és} \quad F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y)$$

Állítás. $F_{X,Y}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u, v) du dv$

$$f_{X,Y}(x, y) = \partial_y \partial_x F_{X,Y}(x, y) = \partial_x \partial_y F_{X,Y}(x, y)$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad \text{és} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

Állítás. Legyen (X, Y) abszolút folytonos, $A \subseteq \mathbb{R}, B \subseteq \mathbb{R}^2$ mérhető halmazok.

- $P(X \in A) = \int_{x \in A} f_X(x) dx$
- $P((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) d(x, y)$

Állítás. Legyenek $a < b, c < d$ valós számok, X és Y tetszőleges val. változók. Ekkor $P(a \leq X < b, c \leq Y < d) = F_{X,Y}(b, d) - F_{X,Y}(b, c) - F_{X,Y}(a, d) + F_{X,Y}(a, c)$.

- Állítás.**
- X, Y függetlenek $\Leftrightarrow F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)$
 - X, Y függetlenek $\Leftrightarrow f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$
 - X, Y függetlenek $\Leftrightarrow P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
 - X, Y függetlenek $\Rightarrow E(XY) = EX \cdot EY$

Definíció. X és Y kovarianciája: $Cov(X, Y) = E[(X - EX)(Y - EY)]$.

Köv.: $Cov(X, Y) = E(XY) - EXEY$.

Elnevezés: ha $Cov(X, Y) = 0$, akkor azt mondjuk, hogy X és Y **korrelálatlanok**.

- Állítás.**
- X és Y függetlenek $\Rightarrow X$ és Y korrelálatlanok
 - X és Y korrelálatlanok $\not\Rightarrow X$ és Y függetlenek !!!!!

Definíció. X és Y lineáris korrelációja: $R(X, Y) = \begin{cases} \frac{Cov(X, Y)}{DXDY} & \text{ha } DX, DY > 0 \\ 0 & \text{ha } DX=0 \text{ v. } DY=0 \end{cases}$

Ez a Pearson-féle lineáris korreláció két valószínűségi változó közti *lineáris* kapcsolat irányát és erősségét méri.

Definíció. Kovarianciamátrix. Legyen \mathbf{X} valószínűségi vektorváltozó. Ekkor $\Sigma(\mathbf{X}) := E(\mathbf{X} \cdot \mathbf{X}^T) - E(\mathbf{X})E(\mathbf{X})^T$

A többdimenziós adatelemzés lényeges eszköze a korrelációs mátrix, aminek (i, j) -edik eleme az $R(X_i, X_j)$ lineáris korrelációs együttható. A korrelációs mátrix átlójában csupa 1-ek szerepelnek.

Állítás. Legyen \mathbf{X} n dimenziós valószínűségi vektorváltozó, $\mathbf{A} \in \mathbb{R}^{m \times n}$. Ekkor

- $E(\mathbf{A} \cdot \mathbf{X}) = \mathbf{A} \cdot E(\mathbf{X})$
- $\Sigma(\mathbf{A} \cdot \mathbf{X}) = \mathbf{A} \cdot \Sigma(\mathbf{X}) \cdot \mathbf{A}^T$

Tétel. Valószínűségi vektorváltozó transzformáltjának sűrűségfüggvénye. Legyen $\mathbf{X} = (X_1, \dots, X_n)$ abszolút folytonos valószínűségi vektorváltozó $f_{\mathbf{X}}$ sűrűség-

függvénnyel, $A \subseteq \mathbb{R}^n$ összefüggő és nyílt halmaz. Legyen $\mathbf{g} : A \rightarrow A$ függvény, amely invertálható és inverze folytonosan differenciálható. Legyen $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, $\mathbf{J} = \partial_{\mathbf{y}}\mathbf{g}^{-1}(\mathbf{y})$ a Jacobi-mátrix. Ekkor

$$f_{\mathbf{g}(\mathbf{x})}(\mathbf{y}) = |\det(\mathbf{J})| \cdot f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))$$

A többdimenziós normális és az egyenletes a gyakorlatban legtöbbször előforduló abszolút folytonos többdimenziós valószínűségi változók.

Ha \mathbf{X} d dimenziós nem-elfajuló **normális eloszlást** követ \mathbf{m} várható érték vektorral és $\Sigma > 0$ kovarianciamátrixszal (jel.: $\mathbf{X} \sim N_d(\mathbf{m}, \Sigma)$), akkor sűrűségfüggvénye:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} |\det(\Sigma)|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \Sigma^{-1}(\mathbf{x} - \mathbf{m}) \right\}.$$

Állítás. Standardizálás. Legyen $\mathbf{X} \sim N_d(\mathbf{m}, \Sigma)$. Ekkor $(\mathbf{X} - \mathbf{m}) \cdot \Sigma^{-1/2} \sim N_d(\mathbf{0}_d, \mathbf{I}_d)$.

Tétel. Normális korreláció tétele. Legyen $\mathbf{X} \sim N_d(\mathbf{m}, \Sigma)$, particionáljuk az itt szereplő vektorokat d_1 és d_2 , $d = d_1 + d_2$ méretű vektorokra a következőképpen:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix}, \text{ a kovarianciamátrixot pedig } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ módon, ahol}$$

Σ_{ij} részmatrrix $d_i \times d_j$ méretű ($i, j \in \{1, 2\}$).

Ekkor $\mathbf{X}_1 | \mathbf{X}_2 \sim N_{d_1}(\mathbf{m}_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{X}_2 - \mathbf{m}_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$.

\mathbf{X} **egyenletes eloszlást** követ a d dimenziós tér $B \subseteq \mathbb{R}^d$ mérhető részhalmazán (jel.: $\mathbf{X} \sim E(B)$), ha sűrűségfüggvénye:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{t(B)} & \text{ha } \mathbf{x} \in B \\ 0 & \text{egyébként} \end{cases} \quad \text{ahol a } t(\cdot) \text{ függvény a } d \text{ dimenziós térfogatot jelöli}$$

\mathbf{X} d dimenziós Student-féle **t -eloszlást** követ ν szabadságfokkal és Σ kovarianciamátrixszal (jel.: $\mathbf{X} \sim t_{d,\nu}(\Sigma)$), ha $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{Y/\nu}}$, ahol $\mathbf{Z} \sim N_d(\mathbf{0}, \Sigma)$, $Y \sim \chi_{\nu}^2$, \mathbf{Z} és Y függetlenek.

Ajánlott irodalom: Márkus L. előadásfóliái a többdimenziós normális eloszlásról

Feltételes várható érték

Legyen a valószínűségi mező a szokásos (Ω, \mathcal{A}, P) hármas, $\mathcal{F} \subset \mathcal{A}$ σ -algebra.

Definíció. \mathcal{F} -mérhetőség.

Az $X: \Omega \rightarrow \mathbb{R}$ valószínűségi változó \mathcal{F} -mérhető, ha minden $B \subseteq \mathbb{R}$ Borel-halmazra $X^{-1}(B) \in \mathcal{F}$.

Definíció. X feltételes várható értéke \mathcal{F} -re nézve.

Legyen X integrálható. Az $Y := E(X|\mathcal{F})$ az a valószínűségi változó, amelyre egyrészt Y \mathcal{F} -mérhető, másrészt $\forall B \in \mathcal{F}$ halmazra $\int_B X dP = \int_B Y dP$.

Speciálisan, ha $\mathcal{F} = \sigma(Y)$, azaz \mathcal{F} -et az Y valószínűségi változó generálja, akkor $E(X|\mathcal{F})$ helyett $E(X|Y)$ -t írunk.

Tehát $E(X|Y)$ -ra úgy gondolunk, mint egy valószínűségi változóra, konkrétan az Y valószínűségi változó egy mérhető $h(Y)$ függvényére; és ha Y egy adott értéket vesz

fel, azaz ha $E(X|Y = y)$, akkor mint konkrét számra.

Abszolút folytonos eloszlások esetén a következő képlettel számítható:

$$E(g(X)|Y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx \Big|_{y=Y}$$

ahol $f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{ha } f_Y(y) > 0 \\ 0 & \text{különben} \end{cases}$ a feltételes sűrűségfüggvény.

Definíció. σ -algebrától való függetlenség.

X valószínűségi változó független az \mathcal{F} σ -algebrától, ha $\forall A \in \sigma(X)$ és $\forall B \in \mathcal{F}$ eseményekre $P(A \cap B) = P(A)P(B)$.

Állítás. Tulajdonságok. Legyen g \mathcal{F} -mérhető függvény.

- $E(X|\mathcal{F})$ 1 valószínűséggel egyértelműen létezik
- $E(E(X|\mathcal{F})) = EX \rightsquigarrow$ teljes várható érték tétel (TVÉT)
- X \mathcal{F} -mérhető $\Rightarrow E[g(X)|\mathcal{F}] = g(X)$
- X független \mathcal{F} -től $\Rightarrow E(X|\mathcal{F}) = EX$
- X \mathcal{F} -mérhető $\Rightarrow E(XY|\mathcal{F}) = XE(Y|\mathcal{F})$

Állítás. Ha X független \mathcal{F} -től, Y mérhető \mathcal{F} -re nézve, $g(X, Y)$ integrálható, akkor $E(g(X, Y)|\mathcal{F}) = E(g(X, y))|_{y=Y}$.

Állítás. Teljes valószínűség tétele folytonos esetben.

Legyen A tetszőleges esemény, Y abszolút folytonos valószínűségi változó. Ekkor

$$P(A) = \int_{-\infty}^{\infty} P(A|Y = y) f_Y(y) dy.$$

Ajánlott irodalom: Márkus L. előadásfóliái a feltételes várható értékről

Valószínűségi változók szimulációja

Először egydimenziós valószínűségi változókra koncentrálnak. A két leggyakrabban használt módszer véletlen szám(ok) generálására abszolút folytonos eloszlások esetén az inverziós módszer és az elutasításos módszer.

Inverziós módszer (inverse transform method)

Ez a módszer a következő tételen alapul.

Tétel. Ha X abszolút folytonos F eloszlásfüggvénnyel és $U \sim E(0, 1)$, akkor $F^{-1}(U)$ éppen egy X eloszlású valószínűségi változó.

Tehát az algoritmus:

1. Generáljunk egy u véletlen számot $(0, 1)$ -en egyenletes eloszlásból.
2. $x = F^{-1}(u)$ a keresett véletlen szám.

Ha az eloszlás sűrűségfüggvénye ismert, akkor először ki kell számítani az eloszlásfüggvényt, majd az eloszlásfüggvényt invertálni kell.

A módszer gyengéje/nehézsége, hogy nem minden abszolút folytonos eloszlású valószínűségi változó eloszlásfüggvényének van explicit alakja (csak valamilyen integrálos). Ilyen esetekben – például a normális eloszlásnál – más módszerre van szükség.

Elutasításos módszer (rejection method)

Definíció. Egy valós függvény tartójának (support) hívjuk azon pontok halmazának a lezártját, ahol a függvény nem nulla. Formálisan $\text{supp}(f) = \text{cl}(\{x \in \mathbb{R} : f(x) \neq 0\})$, ahol $\text{cl}(A)$ az A halmaz lezártját jelöli.

Abszolút folytonos eloszlásoknál az eloszlás tartóját azonosnak tekintjük a sűrűségfüggvény tartójával. Például az exponenciális eloszlás tartója a $[0, \infty)$ halmaz, a normális eloszlásé \mathbb{R} .

Az elutasításos módszer a következő tételen alapul.

Tétel. Legyen $f(x)$ egy sűrűségfüggvény B tartóval, $(X, Y) \sim E(A)$, ahol $A = \{(x, y) : x \in B, 0 \leq y \leq \sup(f(B))\}$. Ekkor X sűrűségfüggvénye éppen $f(x)$.

Legyen a az X valószínűségi változó tartójának infimuma, b pedig a szuprémuma; továbbá legyen c a sűrűségfüggvény maximumának értéke. A fenti tétel alapján az algoritmus:

1. Generáljunk egy u és egy v véletlen számot $E(0, 1)$ eloszlásból.
2. Legyen $x = (b - a)u + a$ és $y = cv$.
3. Ha $y \leq f(x)$, akkor készen vagyunk, x a keresett véletlen szám.
Ha $y > f(x)$, akkor menjünk vissza az 1. ponthoz.

Most rátérünk a többdimenziós esetre, véletlen vektorok generálására. Az elutasításos módszer könnyen általánosítható magasabb dimenziószámra. Legyen az $\mathbf{X} = (X_1, \dots, X_n)$ valószínűségi vektorváltozó sűrűségfüggvénye $f(\mathbf{x}) = f(x_1, \dots, x_n)$. Jelölje \mathbf{X} tartóját a $B \in \mathbb{R}^n$ halmaz. Belátható, hogy amennyiben $(\mathbf{X}, Y) \sim E(A)$, ahol $A = \{(\mathbf{x}, y) : \mathbf{x} \in B, 0 \leq y \leq \sup(f(B))\}$, akkor \mathbf{X} sűrűségfüggvénye éppen $f(\mathbf{x})$.

Ezáltal az algoritmus:

1. Generáljunk egy \mathbf{x} véletlen vektort az $E(B)$ eloszlásból.
2. Legyen c a sűrűségfüggvény maximumának értéke. Generáljunk egy y véletlen számot az $E(0, c)$ eloszlásból.
3. Ha $y \leq f(\mathbf{x})$, akkor készen vagyunk, \mathbf{x} a keresett véletlen vektor.
Ha $y > f(\mathbf{x})$, akkor menjünk vissza az 1. ponthoz.

Nem evidens, miképp generáljunk véletlen vektort az $E(B)$ eloszlásból, mivel a B halmaz tetszőleges halmaz lehet. Például úgy lehet kiküszöbölni ezt a problémát, hogy a B halmazt lefedjük egy kellően kicsi téglával, mivel egy téglában már tudunk véletlen vektort generálni, majd eldobjuk azokat a téglában generált véletlen vektorokat, amik nem esnek a B halmazba.

A következőkben arra adunk módszert, hogyan tudunk először két-, majd magasabb dimenziós véletlen vektorokat előállítani előre meghatározott várható érték vektorral és kovarianciastruktúrával.

Feladat: Hogyan generáljunk x_1 és x_2 véletlen számokat úgy, hogy azok várható értéke m_1 és m_2 , szórása σ_1 és σ_2 , korrelációjuk pedig r legyen?

\rightsquigarrow **Megoldás:** Generáljunk y_1 és y_2 véletlen számokat 0 várható értékkel és 1 szórással, egymástól függetlenül valamilyen eloszlásból. Például választhatjuk a normális vagy az egyenletes eloszlást. Ezután legyen $x_1 = m_1 + \sigma_1 y_1$ és $x_2 = m_2 + \sigma_2 (r y_1 + \sqrt{1 - r^2} y_2)$.

Feladat: Hogyan generáljunk $\mathbf{x} = (x_1, \dots, x_n)^T$ véletlen vektort úgy, hogy várható érték vektora $\mathbf{m} = (m_1, \dots, m_n)^T$, a koordináták szórása $\sigma_1, \dots, \sigma_n$, \mathbf{x} korrelációs mátrixa pedig $\mathbf{R} = [r_{i,j}]_{i,j=1,\dots,n}$ legyen?

\rightsquigarrow **Megoldás:** Generáljunk egy $\mathbf{y} = (y_1, \dots, y_n)^T$ véletlen vektort úgy, hogy minden y_i koordinátája 0 várható értékű és 1 szórással legyen, egymástól függetlenek. Számoljuk ki a korrelációs mátrixból a kovariancia mátrixot: $\Sigma = \mathbf{D}\mathbf{R}\mathbf{D}$, ahol $\mathbf{D} = \text{Diag}(\sigma_1, \dots, \sigma_n)$ azt a diagonális mátrixot jelöli, aminek $\sigma_1, \dots, \sigma_n$ értékek vannak az átlójában. Ezután készítsük el a kovariancia mátrix $\Sigma = \mathbf{L}\mathbf{L}^T$ Cholesky-felbontását. Ennek segítségével a probléma megoldása: $\mathbf{x} = \mathbf{m} + \mathbf{L}\mathbf{y}$.

Megjegyzés. Az előzőekben adott két módszer azt nem garantálja, hogy x_i és y_i értékek eloszlása is megegyezik – valójában ezek kizárólag akkor lesznek azonos eloszlásúak, ha az y -okat normális eloszlásból generáltuk.

Kopulák

A kopulák a többdimenziós eloszlás illesztésének igencsak nehéz feladatában nyújtanak hathatós segítséget. Egy többdimenziós eloszlást "fel lehet bontani" az egydimenziós peremeire és az összefüggőségi struktúrát magában foglaló úgynevezett kopulájára. Matematikailag a kopula egy olyan többdimenziós eloszlás eloszlásfüggvénye, aminek minden egydimenziós pereme $E(0; 1)$ eloszlású. Néha a kopulát nem az eloszlásfüggvénynek, hanem magának az egyenletes (egydimenziós) peremeloszlású többdimenziós eloszlásnak tekintik.

Definíció. Kopula. d dimenziós kopulának nevezzük a $C : [0; 1]^d \rightarrow [0; 1]$ eloszlásfüggvényt, ha egydimenziós peremei egyenletesek.

Legyen $C(\mathbf{u}) = C(u_1, \dots, u_d)$ és jelölje U_1, \dots, U_d az egyenletes peremeloszlások valószínűségi változóit.

A definícióból következnek a C kopula alábbi tulajdonságai:

- $C(u_1, \dots, u_d)$ minden változójában nemcsökkenő függvény;
- $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i \quad \forall i \in \{1, \dots, d\}$;
- Minden $a_i \leq b_i \quad i \in \{1, \dots, d\}$ esetén a $P(a_1 \leq U_1 < b_1, \dots, a_d \leq U_d < b_d)$ valószínűségnek nemnegatívnak kell lennie. Ennek formális felírása a kopulával a következő:

$$\sum_{i_1=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1+\dots+i_d} C(u_{1,i_1}, \dots, u_{d,i_d}) \geq 0, \text{ ahol } u_{i,1} = a_i, u_{i,2} = b_i$$

Definíció. X valószínűségi változó kvantilisfüggvénye.

$$q(u) = F_X^{[-1]}(u) = \inf\{x : F_X(x) > u\} \quad 0 < u < 1$$

A kvantilisfüggvény egy általánosított inverz arra az esetre, amikor az eloszlásfüggvény

nem szigorúan monoton. Amennyiben X abszolút folytonos eloszlású, akkor az F_X szigorúan monoton növekvő, ilyenkor a kvantilisfüggvény a hagyományos értelemben vett inverznek felel meg, azaz $F_X^{[-1]}(u) = F_X^{-1}(u)$.

Tétel. Tetszőleges X valószínűségi változó esetén teljesülnek az alábbiak:

- $F_X(X) \sim E(0; 1)$, azaz a valószínűségi változót az eloszlásfüggvényébe írva $(0; 1)$ -en egyenletes eloszlást kapunk;
- Ha $U \sim E(0; 1)$, akkor $F_X^{[-1]}(U) \stackrel{d}{=} X$, azaz egy $(0; 1)$ -en egyenletes valószínűségi változót az X val. változó kvantilisfüggvényébe írva éppen az X val. változóval megegyező eloszlású val. változót kapunk.

A fenti tételben lévő "játékot" valószínűségi vektorváltozókkal is meg lehet csinálni, ami a kopulaelmélet alaptételéhez fog elvezetni.

Legyen $\mathbf{X} = (X_1, \dots, X_d)$ valószínűségi vektorváltozó. Az előző tétel alapján $F_{X_1}(X_1), \dots, F_{X_d}(X_d)$ mindegyike $E(0; 1)$ eloszlású, így definiálni tudjuk az alábbi C_X függvényt:

$$C_X(u_1, \dots, u_d) := P(F_{X_1}(X_1) < u_1, \dots, F_{X_d}(X_d) < u_d).$$

Könnyen belátható, hogy a C_X függvény kopula, a definíciót pedig tovább tudjuk írni a következőképpen:

$$C_X(u_1, \dots, u_d) = P(X_1 < F_{X_1}^{[-1]}(u_1), \dots, X_d < F_{X_d}^{[-1]}(u_d)) = F_{\mathbf{X}}(F_{X_1}^{[-1]}(u_1), \dots, F_{X_d}^{[-1]}(u_d)).$$

Bevezetve az $x_j = F_{X_j}^{[-1]}(u_j)$ változókat, $F_{X_j}(x_j) = u_j$ adódik és

$$C_X(F_{X_1}(x_1), \dots, F_{X_d}(x_d)) = F_{\mathbf{X}}(x_1, \dots, x_d).$$

Ezzel az alábbi tétel egyik irányát be is láttuk:

Tétel. Sklar tétele. Legyen F egy d dimenziós eloszlás együttes eloszlásfüggvénye és jelölje az egydimenziós peremeloszlásfüggvényeket F_1, \dots, F_d . Ekkor létezik egy olyan C kopula, amelyre

$$C(F_1(x_1), \dots, F_d(x_d)) = F(x_1, \dots, x_d) \quad \forall x_i \in (\mathbb{R} \cup \{-\infty, \infty\}), i \in \{1, \dots, d\} \quad (1)$$

teljesül. Ha a peremeloszlások folytonosak, akkor ez a C függvény egyértelmű is. Megfordítva, ha C egy kopula és F_1, \dots, F_d egydimenziós eloszlásfüggvények, akkor az (1) egyenlettel meghatározott F függvény egy olyan d dimenziós eloszlás eloszlásfüggvénye, melynek peremei éppen az F_1, \dots, F_d függvények.

Tehát a Sklar-tétel jelöléseit használva, a fenti levezetés alapján a kopulát az alábbi módon lehet előállítani:

$$C(u_1, \dots, u_d) = F\left(F_1^{[-1]}(u_1), \dots, F_d^{[-1]}(u_d)\right) \quad u_i \in (0; 1), i \in \{1, \dots, d\} \quad (2)$$

Ha a peremek mindegyike abszolút folytonos, akkor definiálhatjuk a kopula-sűrűségfüggvényt, a megfelelő sűrűségfüggvényeket pedig f, f_1, \dots, f_d -vel jelölve, a

kopula-sűrűségfüggvény alábbi alakra hozható:

$$c(\mathbf{u}) := \frac{\partial^d}{\partial u_1 \partial u_d} C(\mathbf{u}) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdot \dots \cdot f_d(F_d^{-1}(u_d))}.$$

Állítás. X_1, \dots, X_d függetlenek $\iff X_1, \dots, X_d$ kopulája $C(u_1, \dots, u_d) = u_1 \cdot \dots \cdot u_d$
Ezt a C függvényt **függetlenségi kopulának** hívjuk.

Tétel. Fréchet–Hoeffding korlátok. Tetszőleges C kopula esetén

$$\max(u_1 + \dots + u_d - (d - 1), 0) \leq C(\mathbf{u}) \leq \min(u_1, \dots, u_d).$$

Ezeknek a korlátoknak az a jelentősége, hogy

- a felső korlát önmagában is egy kopula, a tökéletes pozitív kapcsolatot írja le.
- Az alsó korlát $d = 2$ esetén kopula, a tökéletes negatív kapcsolatot írja le. Az alsó korlát $d > 2$ esetén nem kopula.

A gyakorlati alkalmazások szempontjából a kopulák két legfontosabb családja az elliptikus kopulák és az Arkhimédészi kopulák.

Összefüggőségi mérőszámok

Kopulák esetén a bevezető kurzusokon megismert lineáris, Pearson-féle korreláció helyett másik összefüggőséget számszerűsítő mérőszámokat, a rangkorrelációs együtthetőköt praktikus használni. A rangkorrelációs mutatók közös tulajdonsága, hogy nem az eredeti értékek, hanem a megfigyelések rangjai alapján számolnak korrelációt, ezáltal a korreláció nem függ a peremeloszlásoktól. Ez teszi őket igazán alkalmassá az összefüggőség megragadására kopulák esetén.

Mindenekelőtt tekintsük át a Pearson-féle lineáris korreláció hátrányait:

- nem létezik, ha valamelyik valószínűségi változónak nincs szórása (például vastag szélű eloszlások esetén)
- a lineáris kapcsolat erősségét méri, de mi van, ha a kapcsolat nem lineáris?
- nagyon érzékeny a kiugró, ún. outlier értékekre

Definíció. Rang. Legyenek x_1, \dots, x_n megfigyelések. Az x_i ($i \in \{1, \dots, n\}$) elem rangja a k szám, ha $x_k^* = x_i$, azaz x_i éppen a k -edik legkisebb elem. Jelölése: $r(x_i)$.

Az 1. táblázat tartalmazza a lineáris korreláció és a kopulaelméletben használatos rangkorrelációk definícióját és mintából való becslését. A táblázatban szereplő (X_1, Y_1) és (X_2, Y_2) ugyanolyan eloszlásúak, mint (X, Y) , és függetlenek is tőle. A sgn az előjelfüggvényt jelöli, a tapasztalati minták x_1, \dots, x_n és y_1, \dots, y_m , ezek átlaga pedig \bar{x} és \bar{y} . A rangok esetén az átlagokat $r(x)$ és $r(y)$ jelöli.

A rangkorrelációknak számos egyéb, az 1. táblázatban szereplővel ekvivalens definíciója és kiszámítási módja adható meg, ennek az összefoglalónak nem célja ezek részletes ismertetése.

A Kendall-féle τ és a Spearman-féle ρ tulajdonságai:

- a $[-1; 1]$ intervallumból vesznek fel értéket;

1. táblázat. Különböző korrelációk definíciója és kiszámítása mintából

Korreláció	Definíció	Számítás mintából
Pearson-féle lineáris korreláció	$R(X, Y) = \frac{\text{cov}(X, Y)}{DX \cdot DY}$	$\frac{\sum_{i=1}^n \sum_{j=1}^m (x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{j=1}^m (y_j - \bar{y})^2}}$
Kendall-féle τ	$E(\text{sgn}[(X_1 - X_2)(Y_1 - Y_2)])$	$\frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)$
Spearman-féle ρ	$R(F_X(X), F_Y(Y))$	$\frac{\sum_{i=1}^n \sum_{j=1}^m (r(x_i) - \bar{r}(x))(r(y_j) - \bar{r}(y))}{\sqrt{\sum_{i=1}^n (r(x_i) - \bar{r}(x))^2} \sqrt{\sum_{j=1}^m (r(y_j) - \bar{r}(y))^2}}$

- +1 a tökéletes pozitív, -1 a tökéletes negatív kapcsolatot jelenti;
- a 0-hoz közeli érték a gyenge kapcsolatra utal;
- nem érzékenyek a kiugró, outlier értékekre;
- értékük kizárólag a kopulától függ, a peremektől nem.

Tétel. Jelölje C az X és Y abszolút folytonos val. változók kopuláját. Ekkor

$$\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \quad \text{és} \quad \rho(X, Y) = 12 \int_0^1 \int_0^1 (C(u, v) - uv) du dv.$$

Látni fogjuk, hogy a Kendall-féle τ számos esetben egyszerűen számolható a paraméterek függvényében. A Spearman-féle ρ -nak a legtöbb népszerű kopulacsalád esetén nincs zárt képlete.

Elliptikus kopulák

Elliptikus eloszlásnak nevezzük azon eloszlásokat, melyek szintvonalai ellipszisek/ellipszoidok.¹ Ebbe az eloszláscsaládba tartozik a többdimenziós normális és a többdimenziós Student-féle t -eloszlás.

Definíció. Elliptikus kopulák: az elliptikus eloszlásokból a (2) képlettel származtatott kopulák.

Definíció. Gauss-kopula. Legyen $\mathbf{X} \sim N_d(\mathbf{0}, \mathbf{R})$, ahol \mathbf{R} korrelációs mátrix, jelölje \mathbf{X} együttes eloszlásfüggvényét $\Phi_{\mathbf{R}}$. Ekkor az \mathbf{X} -ből származtatott Gauss-kopula:

$$C_{\mathbf{R}}^{\text{Gauss}}(u_1, \dots, u_d) = \Phi_{\mathbf{R}}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)).$$

Definíció. t -kopula. Legyen $\mathbf{X} \sim \mathbf{X} \sim t_{d, \nu}(\mathbf{R})$, ahol \mathbf{R} korrelációs mátrix, jelölje \mathbf{X} együttes eloszlásfüggvényét $\mathbf{t}_{\nu, \mathbf{R}}$, a peremeloszlásfüggvényeket pedig t_{ν} . Ekkor az \mathbf{X} -ből származtatott t -kopula:

$$C_{\nu, \mathbf{R}}^t(u_1, \dots, u_d) = \mathbf{t}_{\nu, \mathbf{R}}(t_{\nu}^{(-1)}(u_1), \dots, t_{\nu}^{(-1)}(u_d)).$$

¹Lásd az ajánlott irodalomként kijelölt Embrechts et al cikket matematikailag precíz definícióért.

Tétel. Abszolút folytonos eloszlásokból származtatott elliptikus kopulák esetén az alábbi összefüggések teljesülnek az R lineáris korreláció és a Kendall-féle τ , valamint a Pearson-féle ρ között:

$$R(X, Y) = \sin\left(\frac{\pi}{2} \cdot \tau(X, Y)\right) \quad \text{és} \quad R(X, Y) = 2 \sin\left(\frac{\pi}{6} \cdot \rho(X, Y)\right).$$

Következmény. Elliptikus kopulák esetén a τ és a ρ rangkorrelációkból számolt korrelációs mátrixok közvetlenül a paraméterekből kiszámolhatók.

Az elliptikus kopulák jellemzői:

- szimmetrikusak, ezért aszimmetrikus piaci jelenségek (például pénzügyekben jobban összefüggnek a nagy veszteségek, mint a nagy nyereségek) modellezésére nem alkalmasak;
- magasabb dimenzióban nagyon sok paramétert kell becsülni.

Arkhimédészi kopulák

Az Arkhimédészi kopulákat egy **kopulageneráló függvény** segítségével állítjuk elő: $\varphi: [0, 1] \rightarrow [0, \infty]$, ahol φ folytonos, szigorúan monoton csökkenő és $\varphi(1) = 0$.

Definíció. Arkhimédészi kopula. $C_{\varphi}(\mathbf{u}) = \begin{cases} \varphi^{-1}\left(\sum_{i=1}^d \varphi(u_i)\right) & \text{ha } \sum_{i=1}^d \varphi(u_i) \leq \varphi(0) \\ 0 & \text{különben} \end{cases}$

Tétel. Arkhimédészi kopulák esetén $\tau = 1 + 4 \cdot \int_0^1 \frac{\varphi(u)}{\varphi'(u)} du$.

A 2. táblázat összefoglalja a néhány gyakorlati szempontból lényeges Arkhimédészi kopulacsalád legfontosabb jellemzőit. Egyszerűség kedvéért a 3. oszlopban a kopula a kétdimenziós esetben szerepel (u és v változókkal).

2. táblázat. Néhány Arkhimédészi kopulacsalád jellemzői

Kopulacsalád	Kopulageneráló fv.	Kopula	Kendall τ
Gumbel	$(-\log u)^{\theta}, \theta \geq 1$	$\exp\left(-\left[(-\log u)^{\theta} + (-\log v)^{\theta}\right]^{1/\theta}\right)$	$1 - \frac{1}{\theta}$
Clayton	$\frac{u^{-\theta} - 1}{\theta}, \theta > 0$	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\frac{\theta}{\theta+2}$

Az Arkhimédészi kopulák kiemelt tulajdonságai:

- minden peremeloszlásuk azonos;
- kevés paraméterük van (rendszerint 1 vagy 2), ami előnyük (könnyű becsülni) és egyben hátrányuk is (nem elég rugalmasak)
- rendszerint nem szimmetrikusak, így alkalmasak arra, hogy válság esetén is modellezzük velük az összefüggőségi struktúrákat

Kopulák paraméterbecslése és illeszkedésvizsgálata

Paraméterbecslés előtt a tapasztalati mintát úgynevezett pseudo-megfigyelésekké, azaz 0 és 1 közötti értékekké kell transzformálni. Az $\mathbf{x}_1 = (x_{1,1}, \dots, x_{1,d}), \dots, \mathbf{x}_n =$

$(x_{1,1}, \dots, x_{1,d})$ d dimenziós mintaelemekből előállított $\mathbf{u}_1, \dots, \mathbf{u}_n$ pszeudomegfigyeléseket az egyes dimenziók mentén megállapított rangok alapján számoljuk: $u_{i,j} = \frac{r(x_{i,j})}{n+1}$, $i \in \{1, \dots, n\}$, $j \in \{1, \dots, d\}$.

A paraméterbecslést kopulák esetén is végezhetjük momentum módszerrel és maximum likelihood módszerrel. Az ML-módszer magasabb dimenzióban nagyon lassú, különösen az elliptikus kopuláknál, mert azoknál rengeteg paramétert kell becsülni. A momentum becslés rendszerint a Kendall τ korrelációs mátrix mintából való becslésének felhasználásával történik.

A kopulák illeszkedésvizsgálata azt a feladat jelenti, hogy vajon egy adott minta egy bizonyos kopulacsaládból származhat-e. Ennek ellenőrzésére az utóbbi 10-15 évben számos statisztikai próbát dolgoztak ki. Mivel a próbák részletes bemutatása igencsak hosszadalmas lenne, inkább csak röviden megemlítem az alábbiakat (részletekért lásd a Genest et al cikket az ajánlott irodalomban):

- próba az empirikus kopula segítségével. Az empirikus kopula a pszeudomegfigyelésekből készített többdimenziós tapasztalati eloszlásfüggvény. Az ezen alapuló próbák rendszerint nagyon lassúak, a futási sebességet multiplier bootstrap módszerekkel lehet gyorsítani.
- próba a Kendall-féle K -függvény segítségével – a d dimenziót 1 dimenzióba képezi, majd Kolmogorov-Szmirnov vagy Cramér-von Mises próbastatisztika számolható.
- próba a Rosenblatt-transzformáció segítségével. A Rosenblatt-transzformáció a koordináták közötti összefüggőséget tünteti el.

Az illeszkedésvizsgálat minden esetben kizárólag számítógépes szimulációkkal hajtható végre, a p -érték kiszámítása nagyobb méretű minta esetén meglehetősen sok időbe telik.

Idősorok esetén gyakran szükség van valamilyen előzetes transzformációra, hogy az adatok kopuláját értelme legyen vizsgálni. Pénzügyi idősoroknál a leggyakrabban alkalmazott transzformáció a log-differenciálás, azaz vesszük az egyes értékek logaritmusát, majd az egymás után következőket kivonjuk egymásból. Formálisan, ha az idősor $x_1, x_2, \dots, x_t, \dots$, akkor a logdifferenciákat az $y_t = \log(x_t) - \log(x_{t-1}) = \log\left(\frac{x_t}{x_{t-1}}\right)$, $t = 2, 3, \dots$ módon számítjuk.

Ajánlott irodalom:

- Embrechts, Lindskog, McNeil: Modelling Dependence with Copulas and Applications to Risk Management
- Haugh: An introduction to Copulas
- Genest, Rémillard, Beaudoin: Goodness-of-fit tests for copulas: A review and a power study

Lineáris modell (regressziószámítás)

Legyenek Y, X_1, \dots, X_p véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői. A regresszióelemzés célja a bennünket különösen érdeklő

Y valószínűségi változó "minél jobb" közelítése az X_1, \dots, X_p valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó

X_i -k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \dots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:

$$(y_i, x_{i,1}, \dots, x_{i,p})^T \quad i = 1, 2, \dots, n \quad \text{általában } n \gg p$$

Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i -vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és egy véges σ szórású valószínűségi változó.

A **lineáris modell**: $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$, ahol

$$\begin{aligned} \bullet \mathbf{y} &= (y_1, \dots, y_n)^T \\ \bullet \mathbf{X} &= \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix} \\ \bullet \mathbf{b} &= (b_1, \dots, b_p)^T \\ \bullet \boldsymbol{\varepsilon} &= (\varepsilon_1, \dots, \varepsilon_n)^T \end{aligned}$$

Paraméterbecslés: $\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Projekció az $F := \text{Im} \mathbf{X}$ altérre: $P_F = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

Becsült értékek: $\hat{\mathbf{y}} := \mathbf{X} \hat{\mathbf{b}}$

Reziduálisok: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$

Reziduális négyzetösszeg: $\text{RNÖ} := \|\hat{\boldsymbol{\varepsilon}}\|^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Teljes négyzetösszeg: $\text{NÖ} = \sum_{i=1}^n (y_i - \bar{y})^2$

Determinációs együttható: $R^2 = 1 - \frac{\text{RNÖ}}{\text{NÖ}} = \frac{\text{NÖ} - \text{RNÖ}}{\text{NÖ}} \rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell. Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.

Gyakori modellválasztási kritériumok:

- Korrigált determinációs együttható: $R_{\text{adj}}^2 = 1 - \frac{n-1}{n-r-1} \frac{\text{SSR}}{N}$ \rightsquigarrow minél nagyobb, annál jobb
- **Akaike-féle információs kritérium**: $AIC = 2k - 2 \log \hat{L}$, ahol
 - k : a becsülendő paraméterek száma, a regressziós modellben $k = p + 1$
 - \hat{L} a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)
Minél kisebb, annál jobb.
- **Bayes-féle információs kritérium**: $BIC = \log n \cdot k - 2 \log \hat{L} \rightsquigarrow$ minél kisebb, annál jobb

A regresszióelemzés lépései:

- az eredményváltozó(k) és a lehetséges magyarázóváltozók kiválasztása
- adatgyűjtés
- adattisztítás, adathibák korrekciója
- pontdiagrammal a potenciális modellek kiválasztása (lineáris, négyzetes, logisz-

- tikus stb.)
- paraméterbecslés
- modelldiagnosztika – az együtthatók szignifikanciája, a modell együttes jósága
- legjobb modell kiválasztása, "modellépítés" – több módszer/mutató közül választhatunk: korrigált R^2 , cross-validation, AIC/BIC információk kritériumok stb.
- előrejelzés

Attól függően, hogy az eredmény-, illetve a magyarázóváltozó diszkrét-e vagy folytonos, az alábbi statisztikai módszerek használhatóak a kapcsolat vizsgálatára:

		Az eredményváltozó	
		diszkrét	abszolút folytonos
A	diszkrét	asszociáció	vegyes kapcsolat
magya- rázó- vál- tózó	abszolút foly- tonos	χ^2 -próba	t -próba, ANOVA
		osztályozási eljárások, diszkriminancia analízis, logisztikus regresszió	korreláció regresszió

Ajánlott irodalom: Márkus L. előadása a regresszióról

Szórásanalízis (ANOVA) / vegyes kapcsolat elemzése

A szórásanalízis a lineáris modell egyik legfontosabb alkalmazása. Az eljárásnak több elnevezése van: szórásanalízis = variancia-analízis = ANOVA (analysis of variance). Vegyes kapcsolat: egy diszkrét és egy folytonos ismérv közti kapcsolat.

A modell: $x_{i,j} = \mu_i + \varepsilon_{i,j}$, ahol

- $i = 1, \dots, k \rightsquigarrow$ csoportok vagy osztályok száma
- $j = 1, \dots, n_i \rightsquigarrow$ mintaelemszám egy osztályon belül
- $N = \sum_{i=1}^k n_i \rightsquigarrow$ teljes mintaelemszám
- $\varepsilon_{i,j} \sim N(0, \tau^2)$ függetlenek, ahol $\tau > 0$

Feladat: annak eldöntése, hogy $\mu_1 = \dots = \mu_k$, azaz a csoporthoz tartozás nem befolyásolja az ismérv értékét.

Vezessünk be jelöléseket:

- $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j} \rightsquigarrow$ részátlagok vagy csoportátlagok
- $\bar{x} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{i,j} = \frac{1}{N} \sum_{i=1}^k n_i \bar{x}_i \rightsquigarrow$ teljes átlag
- $\sigma_i = \sqrt{\frac{1}{n_i-1} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2} \rightsquigarrow$ részszórások
- $\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2} \rightsquigarrow$ teljes szórás
- $SS_w = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) \sigma_i^2 \rightsquigarrow$ csoportokon belüli (within group) eltérés-négyzetösszeg

- $SS_b = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 \rightsquigarrow$ csoportok közötti (between groups) teljes eltérés-négyzetösszeg
- $SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2 = (N - 1) \sigma^2 \rightsquigarrow$ teljes eltérés-négyzetösszeg

Állítás. $SS = SS_w + SS_b$

A kapcsolat erősségét mérő mutató a **szórásnégyzet-hányados**: $H^2 = 1 - \frac{SS_w}{SS} = \frac{SS_b}{SS}$

Tulajdonságai:

- $H^2 = 0$ esetén a két ismérv között nincs kapcsolat, DE (!) ekkor nem feltétlen függetlenek egymástól (analógia: korrelálatlanságból nem következik a függetlenség)
- $H^2 = 1$ esetén a két ismérv között függvénytud kapcsolat van
- $0 < H^2 < 1$ esetén a két ismérv között sztochasztikus kapcsolat van
- erős a kapcsolat, ha $H = \sqrt{H^2}$ közel van 1-hez és gyenge a kapcsolat, ha 0-hoz

Megjegyzés: ez nem más, mint a regresszióanalízis az R^2 .

Tekintsük az alábbi hipotézisvizsgálati feladatot: $H_0 : \mu_1 = \dots = \mu_k$

$H_1 : \text{nem igaz } H_0$

A hipotézisekről egy F -próbbával döntünk, a F próbatesztstatistika kiszámításához az ún. **ANOVA táblázat**ot szokás elkészíteni:

Szóródás forrása	Szabadságfok	Négyzetösszegek	Tapasztalati szórásnégyzetek	
Külső (between group)	$k - 1$	SS_b	$MS_b = \frac{SS_b}{k-1}$	$F = \frac{MS_b}{MS_w} = \frac{\frac{SS_b}{k-1}}{\frac{SS_w}{N-k}}$
Belső (within group)	$N - k$	SS_w	$MS_w = \frac{SS_w}{N-k}$	
Teljes	$N - 1$	SS	-	

Tétel. Ha teljesülnek a modell feltételei és a H_0 nullhipotézis, akkor $F \sim F_{k-1, N-k}$, azaz a próbatesztstatistika F -eloszlást követ.

Logisztikus regresszió (logit modell)

Olyan regresszió, amikor az eredményváltozó egy valószínűség, arra szeretnénk regressziós modellt felépíteni. Legyen Y egy 0 vagy 1 értékű valószínűségi változó, az $Y = 1$ esemény jelöli egy bizonyos esemény bekövetkeztét, míg $Y = 0$ azt, hogy nem következik be. Jelölje $q = P(Y = 1)$ -et, ezt szeretnénk magyarázni. Odds-hányadosnak nevezzük a $\frac{q}{1-q}$ hányadost, ami azt mutatja meg, hányszorosa eséllyel következik be az a bizonyos esemény ahhoz képest, hogy nem következik be. Az odds-hányados logaritmusát logodds-hányadosnak hívják és az a fő feltevésünk (reményünk), hogy ezt lineáris modellel megfelelően tudjuk már magyarázni, azaz ha q az eredményváltozó, akkor a következő modellt építjük fel: $\log\left(\frac{q}{1-q}\right) = b_0 + b_1 x_1 + \dots + b_p x_p$, ahol b_0, \dots, b_p az ismeretlen, becsülendő paraméterek.

A $g(x) = \log\left(\frac{x}{1-x}\right)$ függvényt logit függvénynek szokták hívni, innen a modellt logit modellnek keresztelték el. Itt ezt a g függvényt *linkfüggvény*nek hívják, mert nem a q -t modellezzük lineáris modellel, hanem $g(q)$ -t, tehát g függvény teremti meg a kapcsolatot az eredményváltozó és a magyarázóváltozók között.

A paraméterek becslése csak numerikus módszerekkel lehetséges, nincs rájuk "szép" képlet, szemben az előző fejezetben tárgyalt lineáris modellnél.

Ajánlott irodalom: Márkus L. előadásfóliái ANOVA-ról
