Matematikai statisztika előadás, 11. hét, április 29.

Szórásanalízis és idősorok bevezetés

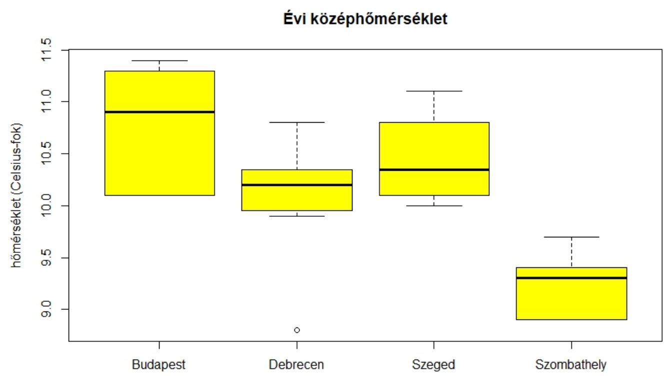
1. Szórásanalízis (analysis of variance, ANOVA)

A szórásanalízis olyan hipotézisvizsgálati eljárás, melynél ugyanazt a mennyiséget vizsgáljuk különböző csoportokba sorolt egyedek esetében, és azt szeretnénk eldönteni, hogy ennek a mennyiségnek az egyes csoportokra jellemző eloszlásának ugyanaz-e a várható értéke. Másképpen fogalmazva, igaz-e, hogy a vizsgált mennyiség várható értékére nincs hatása annak, hogy a megfigyelés melyik csoportból származik. Amint látni fogjuk, több csoport esetén ez a feladat a többváltozós lineáris modellhez is kapcsolódik. Ha viszont csak két csoport van, és feltételezzük, hogy a megfigyelések normális eloszlásúak, akkor a t-próba feladatát kapjuk vissza.

Azt, hogy a megfigyelések különböző csoportokból származnak, úgy is szokták fogalmazni, hogy a mérés egy faktor különböző szintjein történik, és az a kérdés, hogy a faktornak van-e szignifikáns hatása a várható értékre.

Példa. Az alábbi táblázat néhány éves középhőmérséklet érték (forrás: Országos Meteorológiai Szolgálat), különböző évekből, különböző helyszínekről. A kérdés: elfogadható-e, hogy az egyes városokban az évi középhőmérséklet várható értéke megegyezik, vagy szignifikáns különbség mutatható ki? Az 1. ábra az adatokból készült boxplot ábrát mutatja. Ebben a példában a „faktor” a helyszín, és ennek négy „szintje” van.

	Budapest	Debrecen	Szeged	Szombathely
	10,8	8,8	11,1	8,9
	10,1	9,9	10,8	9,4
	11,4	10,0	10,1	8,9
	11,3	10,2	10,0	9,3
	11,0	10,4	10,4	9,7
	10,1	10,8	10,3	
		10,3		
átlag ($\overline{X}$)	10,8	10,1	10,5	9,2
szórás (s_n^*)	0,57	0,63	0,42	0,34



1. ábra. Boxplot ábra az egyes városok éves középhőmérséklet adataiból

1.1. Feltevések és kapcsolat a lineáris modellel

Legyenek X_{ij} független normális eloszlású valószínűségi változók, $i = 1, \ldots, k$ és $j = 1, \ldots, n_i$. Az X_{ij} valószínűségi változó várható értéke μ_i, szórása σ.

$$X_{ij} \sim N(\mu_i, \sigma) \qquad (j = 1, 2, \ldots, n_i).$$

Vagyis: k csoport van, és a k. csoportban μ_i a várható érték. Másképpen: egy faktor különböző szintjein történik mérés, az i. csoportban a faktor i. szintjének hatása μ_i.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$.

$H_1 : \mu_1 = \mu_2 = \ldots = \mu_k$ nem teljesül.

Másképpen:

H_0**: a faktornak nincs szignifikáns hatása**

H_1**: a faktornak szignifikáns hatása van**.

Összefoglalva:

- normális eloszlások várható értékére vonatkozó próba: feltettük, hogy a megfigyelések normális eloszlásúak, és a hipotézisek a várható értékre vonatkoznak
- feltettük, hogy a megfigyelések szórása minden esetben azonos, σ
- a **kétmintás párosítatlan** Student-féle t-próba általánosításának is tekinthető: most nem kettő, hanem több csoport van, a szórások mindenhol megegyeznek

Ezt a feladatot a lineáris modell egy speciális esetének is tekinthetjük. A lineáris modell ez volt:

$$Y_j = a_1 X_{j,1} + a_2 X_{j,2} + \ldots + a_k X_{j,k} + \varepsilon_j,$$

ahol $\varepsilon_j \sim N(0, \sigma^2)$ független normális eloszlású valószínűségi változók.

Most tegyük fel, hogy az $X_{j,i}$ valószínűségi változók értéke csak 0 vagy 1 lehet, sőt, hogy ezek közül mindig pontosan egy lesz 1, a többi 0 (a lineáris modellben a magyarázó változók függetlenségét nem kellett feltenni).

Ekkor ha $a_i = \mu_i$ (minden $i = 1, 2, \ldots, k$ esetén), és az Y_j esetében, vagyis a j. mérésnél a k_j. valószínűségi változó 1, a többi 0, akkor $Y_j = \mu_{k_j} + \varepsilon_j$, azaz Y_j normális eloszlású μ_{k_j} várható értékkel és σ szórással. Vagyis az Y_j-ket aszerint csoportosítva, hogy melyik $X_{j,k}$ értéke 1, éppen a p csoporthoz tartozó méréseket kapjuk vissza.

A többváltozós lineáris modellben $H\beta = 0$ alakú nullhipotéziseket tudtunk tesztelni, ahol β az együtthatók vektora. Most tehát $\beta = (\mu_1, \ldots, \mu_k)$, és lehet

$$H = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 0 & 1 & -1 & \ldots & 0 \\ & & \ldots & & \\ 0 & \ldots & 0 & 1 & -1 \end{pmatrix}.$$

Ekkor $H\beta = (\mu_1 - \mu_2, \mu_2 - \mu_3, \ldots, \mu_{k-1} - \mu_k)^T$, így $H\beta = 0$ éppen azzal ekvivalens, hogy minden μ_j megegyezik, ami a szórásanalízis nullhipotézise volt.

A többváltozós lineáris modell esetében a megadott próbastatisztika F-eloszlású volt a nullhipotézis mellett és az F-próba kritikus értékeit használhattuk. Mivel tehát a szórásanalízis egy speciális eset, most is hasonlóképpen járhatunk el, a próbastatisztika pedig szintén megegyezik az ott látottal, bár most más alakban írjuk fel.

1.2. A szórásanalízis eljárása

X_{ij} valószínűségi változók, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Vagyis k csoport van, és az i.-ben n_i darab megfigyelés van. A szórásanalízis elvégzéséhez az alábbi mennyiségekre lesz szükség.

Csoporton belüli átlagok: $\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$.

Az összes megfigyelés száma: $n = n_1 + \ldots + n_k$.

Teljes átlag: $\overline{\overline{X}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$.

Csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$.

Csoportok közötti szóródás: $S_t = \sum_{i=1}^{k} n_i (\overline{X}_{i\cdot} - \overline{\overline{X}})^2$.

Teljes szóródás: $S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{\overline{X}})^2 = S_t + S_g$.

A próbastatisztika:

$$F = \frac{S_t(n-k)}{S_g(k-1)}.$$

Legyen c_{krit} az $f_1 = k-1$ és $f_2 = n-k$ szabadsági fokú F-próba kritikus értéke α terjedelem mellett.

Ha $F > c_{\text{krit}}$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van legalább egy pár esetében.

Ha $F < c_{\text{krit}}$, akkor **elfogadjuk a nullhipotézist**, a várható értékek között nincs szignifikáns eltérés.

1.3. Szórásanalízis: példa

A korábbi példára visszatérve kiszámíthatjuk a csoporton belüli és csoportok közötti szóródásokat. Most feltételezzük, hogy a szórások az egyes városok esetében megegyeznek, és hogy a középhőmérséklet normális eloszlású, az egyes helyszínek esetében egymástól független (ez utóbbi nagyjából helyes is, mert az adatok mind különböző évekből származnak).

	Budapest	Debrecen	Szeged	Szombathely	összesen
	10,8	8,8	11,1	8,9	
	10,1	9,9	10,8	9,4	
	11,4	10,0	10,1	8,9	
	11,3	10,2	10,0	9,3	
	11,0	10,4	10,4	9,7	
	10,1	10,8	10,3		
		10,3			
átlag ($\overline{X}_{i\cdot}$)	10,8	10,1	10,5	9,2	$\overline{\overline{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

A csoportokon belüli szóródás kiszámítása:

$$\begin{aligned} S_g = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 = \big((10,8-10,8)^2 + (10,1-10,8)^2 + \ldots + (10,1-10,8)^2\big) + \\ + \big((8,8-10,1)^2 + (9,9-10,1)^2 + \ldots + (10,3-10,1)^2\big) + \\ + \big((11,1-10,5)^2 + (10,8-10,5)^2 + \ldots + (10,3-10,5)^2 + \\ + \big((8,9-9,2)^2 + (9,4-9,2)^2 + \ldots + (9,7-9,2)^2\big) = 5,34. \end{aligned}$$

Itt az első sor Budapestnek (az $i = 1$ esetnek) felel meg, minden mérésnél a budapesti mérések átlagától vett különbség négyzetét számítjuk ki, és ezeket adjuk össze. A második sor, $i = 2$, Debrecen, ekkor az itteni átlagtól vett eltérések négyzetét adjuk össze, majd hasonlóképpen az $i = 3$ és $i = 4$ esetekben is.

A csoportok közötti szóródás kiszámítása:

$$S_t = \sum_{i=1}^{k} n_i (\overline{X}_{i\cdot} - \overline{\overline{X}})^2 = 6\cdot(10,8-10,17)^2 + 7\cdot(10,1-10,17)^2 + 6\cdot(10,5-10,17)^2 + 5\cdot(9,2-10,17)^2 = 7,15.$$

Itt minden csoportra az átlagnak a teljes átlagtól vett eltérését emeljük négyzetre, majd ezt a csoport mintaelemszámával szorozzuk, és így adjuk össze az egyes csoportokra.

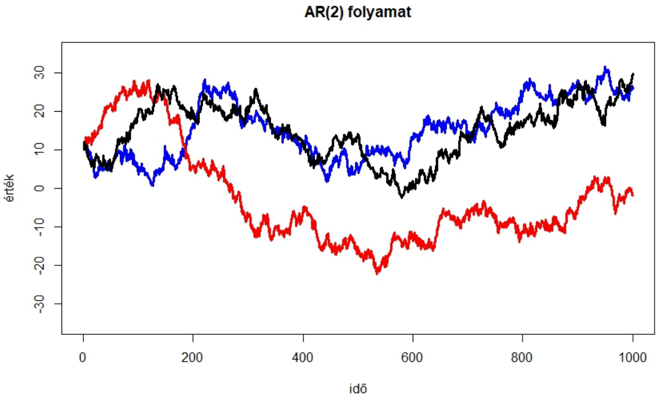
Teljes szóródás = csoportokon belüli + csoportok közötti:

$$S = S_g + S_t = 5,34 + 7,15 = 12,49.$$

Az előző példában: $n = 24$ a megfigyelések száma, $k = 4$ az osztályok száma.

A próbastatisztika:

$$F = \frac{S_t(n-k)}{S_g(k-1)} = \frac{7,15 \cdot 20}{5,34 \cdot 3} = 8,77,$$



2. ábra. Példa idősorra

ahol n a megfigyelések száma, k a csoportok száma, és a csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2 = 5,43$, a csoportok közötti szóródás: $S_t = \sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{\overline{X}})^2 = 7,15$.

Az $f_1 = k - 1 = 3$ és $f_2 = n - k = 20$ szabadsági fokú F-próba kritikus értéke $\alpha = 0,05$ terjedelem mellett: $c_{\text{krit}} = 3,86$.

Mivel $F = 7,15 > c_{\text{krit}} = 3,86$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van.

Vagyis a helynek mint faktornak (tényezőnek) **szignifikáns hatása** van az évi középhőmérsékletre.

2. Idősorok elemzése

Az idősorok fogalma minden olyan elemzésben előjöhet, ahol egy mennyiség (pl. egy ország gdp-je vagy népessége, munkanélküliségi ráta, infláció, más pénzügyi vagy gazdasági mutatók) időbeli függését szeretnénk megérteni. A lineáris modellben $Y(t) = at + b + \varepsilon(t)$ alakú idősorokat tudtunk vizsgálni, amik egy lineáris függvényből és egy hozzáadott véletlen hibából állnak, de természetesen a valós folyamatok modellezésére ez a legtöbb esetben nem elég rugalmas.

2.1. Definíció. *Az*

$$X_0, X_1, X_2, X_3, \dots, X_t, \dots$$

valószínűségi változók sorozata idősor, ha az indexparaméter (sorszám) időpontként is értelmezhető.

Az idősorok általában **nem független** valószínűségi változókból állnak. Sőt, a következő értéket gyakran az előzőekből, egy véletlen hiba hozzáadásával számítjuk ki. Például lehet $X(1) = 10, X(2) = 12$, ezután pedig

$$X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t) \qquad t = 3, 4, \dots \tag{1}$$

ahol $\varepsilon(3), \varepsilon(4), \dots$ egymástól és az korábbi X-ektől független standard normális eloszlású valószínűségi változók. A 2. ábrán ebből a modellből sorsolt három folyamatot láthatunk.

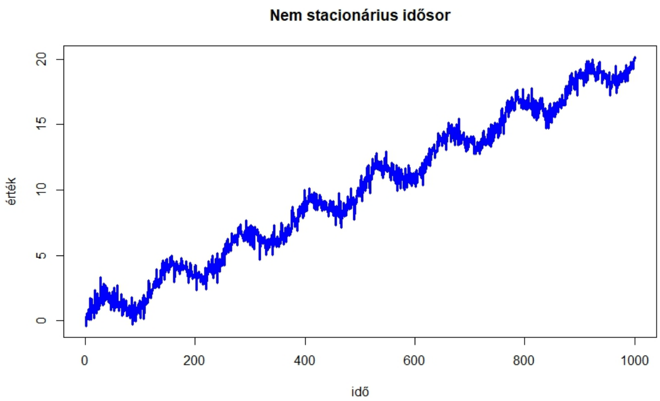
Az összefüggőségeket jellemzi például az autokovariancia-függvény. Ezt azt mondja meg, hogy az s és t időpontokban mért értékek között mennyi a kovariancia, azaz nagyjából azt, hogy ezek között mennyire erős a lineáris jellegű kapcsolat. A kovariancia függ például attól, hogy milyen mértékegységben tekintjük a mennyiségeket, viszont ha ezt az idősoron belül már nem változtatjuk, akkor az autokovarianca-függvény értékeit egymással már össze tudjuk hasonlítani.

2.2. Definíció. *Az $X_1, X_2, \dots$ idősor autokovariancia-függvénye:*

$$R(s,t) = \operatorname{cov}(X_s, X_t) = \mathbb{E}(X_s X_t) - \mathbb{E}(X_s)\mathbb{E}(X_t).$$

Itt $R(t,t) = \mathbb{E}(X_t^2) - \mathbb{E}(X_t)^2 = D^2(X_t)$ a t időpontban vett szórásnégyzet. Ha viszont s és t távolságát növeljük, akkor az X_s és X_t egyre távolabbi időpontokhoz tartoznak, így sok esetben annál gyengébb közöttük az összefüggés, annál kisebb a kovariancia értéke.

2.1. Stacionárius folyamatok



3. ábra. Példa nem stacionárius idősorra (egy lineáris tag, egy periodikus tag és egy stacionárius idősor összege)

Az idősorok elemzésénél gyakran a következőképpen járunk el. Az idősort az alábbi három komponens összegére bontjuk (a 3. ábrán egy olyan idősor látszik, ami három ilyen tag összegeként lett előállítva):

- lineáris trend: $at + b$ alakú determinisztikus lineáris függvény;
- szezonális komponens: $f(t)$ determinisztikus periodikus függvény, melyre valamilyen h periódussal az igaz, hogy $f(t + h) = f(t)$ teljesül minden t-re;
- egy olyan X_t véletlen tag, melynek az eloszlása már t-től minél kevésbe függ, például a várható értéke és a szórása időben állandó, sőt például az X_s, X_t együttes eloszlása is csak attól függ, hogy s és t egymástól milyen messze vannak.

Például ha egy szálloda havi bevételeit szeretnénk elemezni tízéves megfigyelések alapján, akkor abban lehet egy növekvő trend (akár csak az infláció miatt is), a szezonális komponens adódhat abból, hogy például februárban feltehetően a legtöbb esetben alacsonyabb a bevétel, mint júliusban, és még ehhez adhatunk hozzá egy véletlen hibát. Ilyenkor, ha t a hónap sorszámát jelöli ($t = 1, 2, \ldots, 120$), akkor a periodikus részben $f(t)$ olyan érték, ami csak t tizenkettes maradékától függ, vagyis attól, hogy ez a hónap január, február stb.

Az alábbi definíciók arra vonatkoznak, hogy a harmadik komponense, vagyis az időben állandó eloszlású véletlen részre pontosan milyen feltételeket írunk elő.

2.3. Definíció. *Az $X_0, X_1, X_2, \ldots$ idősor* **gyengén stacionárius**, *ha*

- *várható értéke állandó: $\mathbb{E}(X_t) = \mathbb{E}(X_0)$ minden t-re;*
- *a kovariancia csak az időpontok távolságától függ:*

$$R(s,t) = \mathrm{cov}(X_s, X_t) = \mathrm{cov}(X_0, X_{t-s}) = R(0, t-s).$$

Az $X_0, X_1, X_2, \ldots$ idősor **erősen stacionárius**, *ha tetszőleges $n, t_1, t_2, \ldots, t_n$ és h nemnegatív egészek esetén az*

$$\left(X_{t_1}, X_{t_2}, \ldots, X_{t_n}\right) \text{ és } \left(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}\right)$$

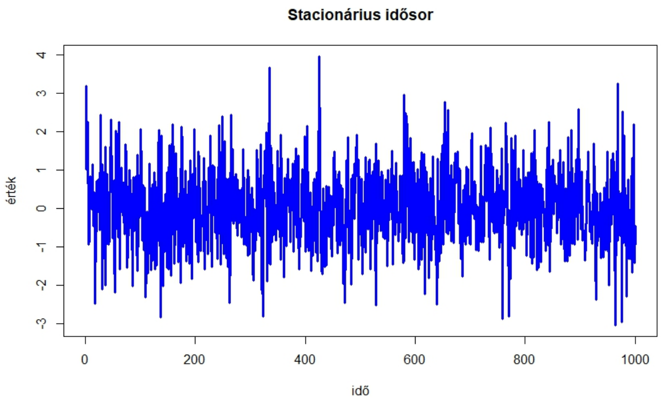
valószínűségi vektorváltozók eloszlása megegyezik.

Egy erősen stacionárius idősor gyengén stacionárius, fordítva nem feltétlenül. A gyengén stacionárius esetben nem csak a kovariancia, hanem a korrelációs együttható is jól használható az adott távolságra lévő tagok közötti „lineáris jellegű” kapcsolat erősségének mérésére, hiszen a szórás is időben állandó.

2.4. Definíció. *Egy gyengén stacioniárius idősor* **autokorrelációs függvénye***:*

$$r(t) = \frac{R(0,t)}{R(0,0)} = \mathrm{corr}(X_s, X_{s+t}) = \frac{\mathrm{cov}(X_s, X_{s+t})}{D(X_s)^2} = \frac{\mathbb{E}\big((X_s - \mathbb{E}(X_s))(X_{s+t} - \mathbb{E}(X_{s+t})\big)}{D^2(X_s)},$$

ahol $s \geq 0$ tetszőlegesen válaszható a gyenge stacionaritás tulajdonsága miatt, és corr *a két valószínűségi változó korrelációs együtthatóját jelöli.*



4. ábra. Példa stacionárius idősorra: független standard normális eloszlású változók

2.2. Az autokorrelációs függvény becslése

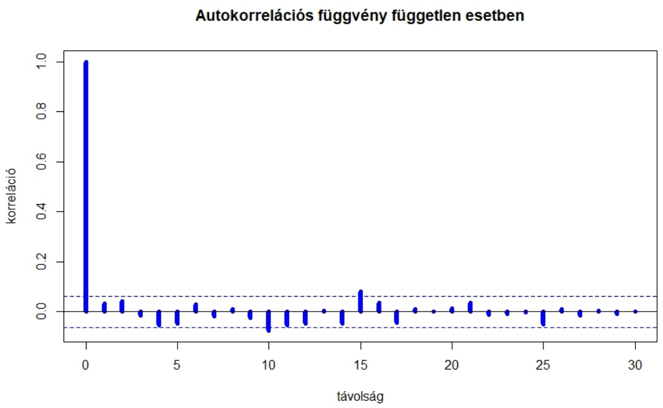
Általában az idősort leíró modellt nem ismerjük, és ezért a várható értéket, szórást, korrelációkat sem. A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető. Az autokorrelációs függvény becslésére az alábbi módszerek szokásosak.

Legyen $X_0, X_1, \ldots, X_{n-1}$ stacionárius idősorból származó n elemű minta. Az autokorrelációs függvény becslése:

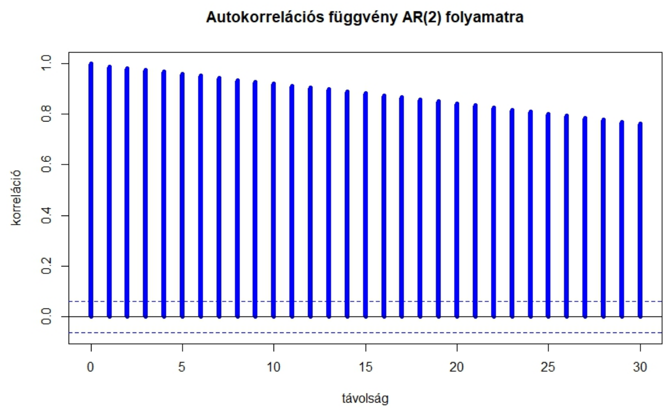
$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1}(X_j - \overline{X}) \cdot (X_{j+t} - \overline{X})}{(n-t) \cdot s_n^{*2}}.$$

Egy másik lehetőség, hogy a tagok száma helyett n-nel osztunk:

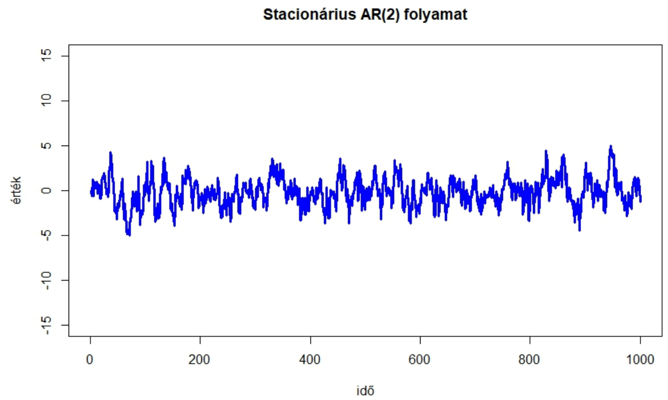
$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1}(X_j - \overline{X}) \cdot (X_{j+t} - \overline{X})}{n \cdot s_n^{*2}}.$$



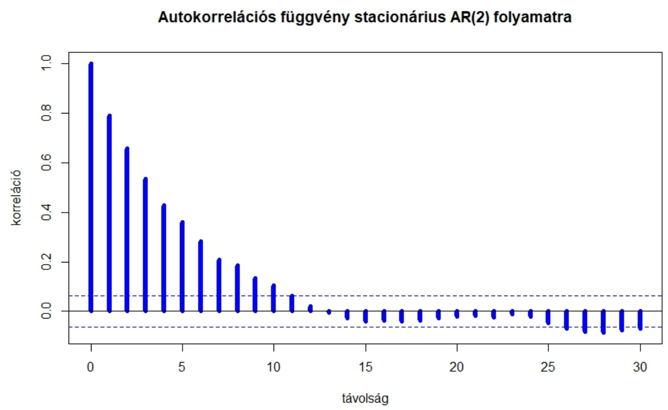
5. ábra. Független azonos eloszlású valószínűségi változók, mint idősor becsült autokorrelációs függvénye



6. ábra. Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat autokorrelációs függvényének becslése



7. ábra. Az $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$ stacionárius folyamat



8. ábra. Az $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat autokorrelációs függvényének becslése

Egyik becslés sem torzítatlan $r(t)$-re, azaz $\mathbb{E}(\hat{r}(t))$ eltér $r(t)$-től. Ha `x` a megfigyelésekből álló vektor, akkor az R-ben az `acf(x)` paranccsal ábrázolható az autokorrelációs függvény becslése.

A 4. ábrán egy egyszerű, független, standard normális eloszlású valószínűségi változókból álló idősort látunk. Itt valójában $r(0) = 1$ és $r(1) = r(2) = \ldots = 0$. Az 5. ábrán ennek az autokorrelációs függvénynek az adatokból kapott becslése látható, itt azonnal látszik, hogy a különböző tagok korrelációja nagyon kicsi.

A 6. ábrán az (1) egyenlettel leírt modellnek (melynek folyamata a 2. ábrán látható) egy megvalósításából számolt autokorreláció becslése. Itt a távolság növelésével csak lassan csökken a korreláció, távoli tagok között is erős összefüggés látszik. Ez a folyamat valójában nem is stacioniárius, a becslés itt nem is jól alkalmazható. Ha a második együtthatót $0,3$-ról $0,1$-re csökkentjük, akkor a folyamat stacionáriussá válik (a 7. ábra), és az autokorrelációs függvény becslésén is látszik, hogy a távolság növelésével a korreláció gyorsabban csökken (a 8. ábra).

Házi feladat május 6., 8:15-ig. Az ismerősöket osszuk három csoportba a nézett sorozatok száma alapján (pl. $0-2, 3-5$, legalább 6) úgy, hogy minden csoportba kerüljön legalább négy ismerős. Szórásanalízis segítségével vizsgáljuk meg, hogy a nézett sorozatok számának, mint faktornak (aminek most csak azt a három lehetséges értékét, szintjét tekintjük, ami a csoportosításnál létrejött), van-e szignifikáns hatása az utazási időre. Az utazási időről a feladat megoldása során feltételezhetjük, hogy normális eloszlású (holott a kerekítések miatt pontosan biztosan nem az).

Házi feladat április 29., 8:15-ig A KSH adatainak segítségével (http://www.ksh.hu/stadat) illesszünk lineáris modellt a kormányzati alrendszerek vagy a háztartások egészségügyi kiadásainak alakulására (2003-2017, éves adatok, 2.4.1. pont). A magyarázó változók száma legalább három legyen (az egyik lehet az évszám, a fenti példához hasonlóan), és ezeket is lehet például az általános gazdasági mutatók vagy más, az oldalon feldolgozott mérőszámok alapján választani. Keressünk úgy magyarázó változókat, hogy az $\tilde{R}^2$ értéke legalább $0,8$ legyen. (A táblák a honlapról letölthetők excel-formátumban.)

Az adatokból az utolsó évet hagyjuk ki, így becsüljük meg a paramétereket, majd a magyarázó változók utolsó évben mért adatainak segítségével készítsünk előrejelzést az utolsó év egészségügyi kiadásaira. Hasonlítsuk ezt össze a valódi adattal, számítsuk ki az abszolút és a relatív hibát is (a két érték különbségét, illetve ennek arányát a valódi értékhez képest).

Tekintsük a háztartások egészségügyi kiadásait (milliárd forintban), valamint az évszámot, a gdp-t és a foglalkoztatottak számát mint magyarázó változókat:

```
haztartas<-c(409.4, 423.1, 464.8, 473.6, 487, 509, 502.5, 561.3, 602.4, 631.1, 622.7, 655.1,
658.2, 701, 709.1)

ev<-2003:2017

gdp<-c(19134, 21078, 22549, 24316, 25701, 27217, 26458, 27269, 28371, 28848, 30290, 32694,
34785, 35896, 38835)

fogl<-c(3922, 3900, 3902, 3928, 3902, 3848, 3749, 3732, 3759, 3827, 3893, 4101, 4211, 4352,
4421)

> summary(lm(haztartas ev +gdp+fogl))

Call:  lm(formula = haztartas   ev + gdp + fogl)

Residuals:

Min 1Q Median 3Q Max

-34.732 -5.928 1.087 5.207 18.670

Coefficients:

Estimate Std.  Error t value Pr(>|t|)

(Intercept) -5.971e+04 1.107e+04 -5.394 0.000218 ***

ev 3.003e+01 5.536e+00 5.424 0.000209 ***

gdp -7.160e-03 5.036e-03 -1.422 0.182845

fogl 2.853e-02 3.435e-02 0.830 0.423918

---

Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  15.3 on 11 degrees of freedom

Multiple R-squared:  0.9817, Adjusted R-squared:  0.9767

F-statistic:  196.9 on 3 and 11 DF, p-value:  7.734e-10
```

Ezzel a $\tilde{R}^2 = 0,9767$, megfelelően illeszkedik a modell.

Az utolsó évet kihagyva:

```
> summary(lm(haztartas[1:14] ev[1:14] +gdp[1:14]+fogl[1:14]))
Call:  lm(formula = haztartas[1:14]   ev[1:14] + gdp[1:14] + fogl[1:14])
Residuals:
Min 1Q Median 3Q Max
-34.832 -6.525 1.442 5.811 18.458
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) -6.014e+04 1.245e+04 -4.829 0.000693 ***
ev[1:14] 3.025e+01 6.238e+00 4.848 0.000673 ***
gdp[1:14] -7.374e-03 5.738e-03 -1.285 0.227697
fogl[1:14] 2.815e-02 3.624e-02 0.777 0.455261
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
Residual standard error:  16.03 on 10 degrees of freedom
Multiple R-squared:  0.9781, Adjusted R-squared:  0.9715
F-statistic:  148.6 on 3 and 10 DF, p-value:  1.365e-08
```

Ebből a becslés az utolsó évre:

```
> 30.25*ev[15]-0.0073*gdp[15]+0.028*fogl[15]-60140
[1] 714.5425
> haztartas[15]/714.54
[1] 0.9923867
> haztartas[15]-714.54
[1] -5.44
```

Vagyis a relatív hiba kevesebb, mint $0,01$, az abszolút hiba pedig $-5,44$.