

Idősorok elemzése (12. előadás)

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.

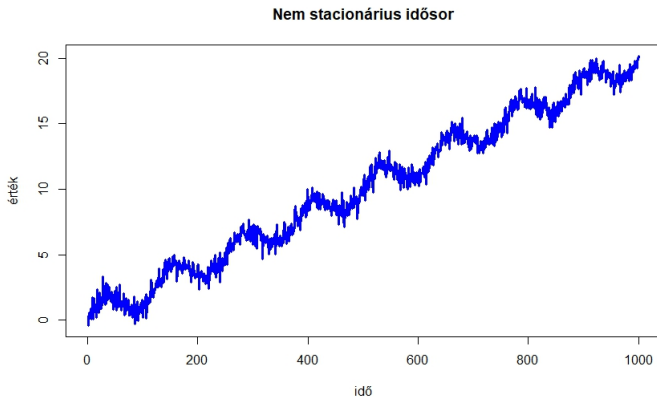
Az idősorok általában **nem független** valószínűségi változókból állnak. Az összefüggéseket jellemzi például az autokovariancia-függvény.

Definíció

Az X_1, X_2, \dots idősor autokovariancia-függvénye:

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

Nem stacionárius idősor



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

Stacionárius folyamatok

Definíció

Az X_0, X_1, X_2, \dots 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:

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

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

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \text{ és } (X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h})$$

valószínűségi vektorváltozók eloszlása megegyezik, vagyis az együttes eloszlás csak az időpontok távolságától függ, a t_1 kezdeti időponttól nem.

Egy erősen stacionárius idősor gyengén stacionárius, fordítva nem feltétlenül.

Autokorrelációs függvény

Definíció

Az X_0, X_1, X_2, \dots 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:

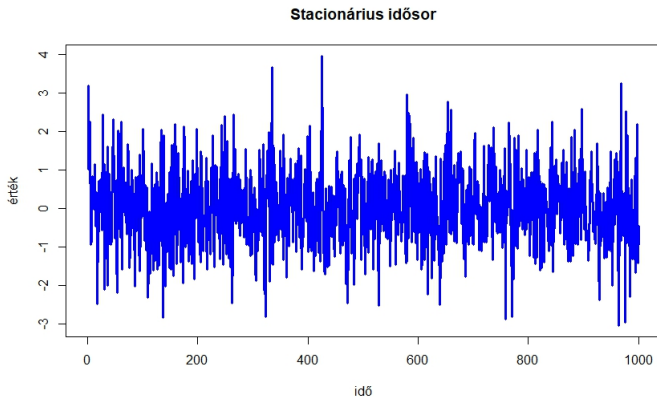
$$R(s, t) = \text{cov}(X_s, X_t) = \text{cov}(X_0, X_{t-s}) = R(0, t - s).$$

Egy gyengén stacionárius idősor autokorrelációs függvénye:

$$\begin{aligned} r(t) &= \frac{R(0, t)}{R(0, 0)} = R(X_s, X_{s+t}) = \frac{\text{cov}(X_s, X_{s+t})}{D(X_s)^2} = \\ &= \frac{\mathbb{E}((X_s - \mathbb{E}(X_s))(X_{s+t} - \mathbb{E}(X_{s+t})))}{D^2(X_s)}, \end{aligned}$$

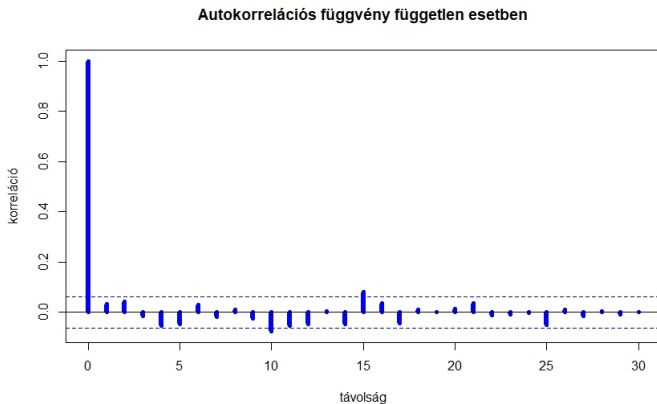
ahol $s \geq 0$ tetszőlegesen válaszható a gyenge stacionaritás tulajdonsága miatt.

Stacionárius idősor



Példa stacionárius idősorra: független azonos eloszlású valószínűségi változók

Autokorrelációs függvény



Független azonos eloszlású valószínűségi változók, mint idősor autokorrelációs függvénye

Az autokorrelációs függvény becslése

Egy gyengén stacionárius idősor autokorrelációs függvénye:

$$\begin{aligned} r(t) &= \frac{R(0, t)}{R(0, 0)} = R(X_0, X_t) = \frac{\text{cov}(X_0, X_t)}{D(X_0)^2} = \\ &= \frac{\mathbb{E}((X_0 - \mathbb{E}(X_0))(X_t - \mathbb{E}(X_t)))}{D^2(X_0)}, \end{aligned}$$

ahol $s \geq 0$ tetszőlegesen válaszható a gyenge stacionaritás tulajdonsága miatt.

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

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

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

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

Egyik becslés sem torzítatlan $r(t)$ -re, azaz $\mathbb{E}(\hat{r}(t))$ eltér $r(t)$ -től.

Autoregressziós folyamat

Definíció

Legyenek $\varepsilon(t)$ független 0 várható értékű 1 szórású valószínűségi változók $t \geq 0$ -ra (például normális eloszlásúak). Az $X(t)$ folyamat p -rendű autoregressziós folyamat, ha minden $t \geq p$ -re

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sigma \cdot \varepsilon(t).$$

Jelölés: $AR(p)$.

Például egy másodrendű autoregressziós $AR(2)$ folyamat ($\alpha_1 = 0,7$, $\alpha_2 = 0,3$, $\sigma = 1$):

$$X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t).$$

Autoregressziós folyamat

Definíció

Legyenek $\varepsilon(t)$ független 0 várható értékű 1 szórású valószínűségi változók $t \geq 0$ -ra (például normális eloszlásúak). Az $X(t)$ folyamat p -rendű autoregressziós folyamat, ha minden $t \geq p$ -re

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sigma \cdot \varepsilon(t).$$

Jelölés: $AR(p)$.

Például egy másodrendű autoregressziós $AR(2)$ folyamat ($\alpha_1 = 0,7, \alpha_2 = 0,3, \sigma = 1$):

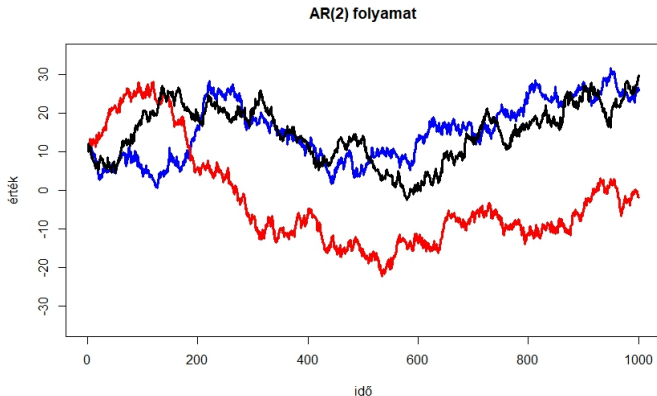
$$X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t).$$

Az $(\alpha_1, \alpha_2, \dots, \alpha_p)$ együtthatóktól függ, hogy az egyenletnek van-e egyértelmű gyengén stacionárius megoldása. Pontosán akkor van ilyen, ha az

$$x^p - \alpha_1 x^{p-1} - \alpha_2 x^{p-2} - \dots - \alpha_p$$

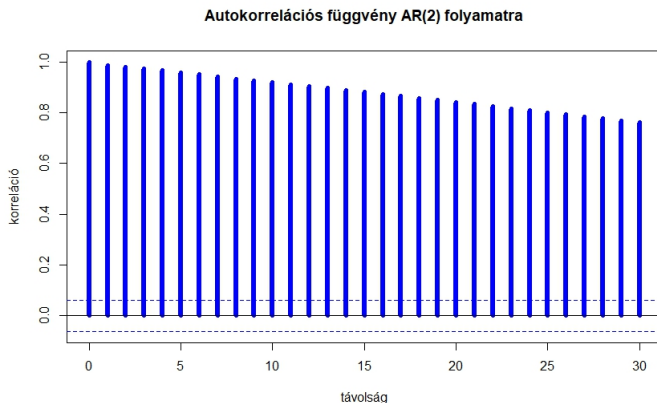
karakterisztikus polinom összes komplex gyökének abszolút értéke legfeljebb 1.

Másodrendű autoregressziós folyamat



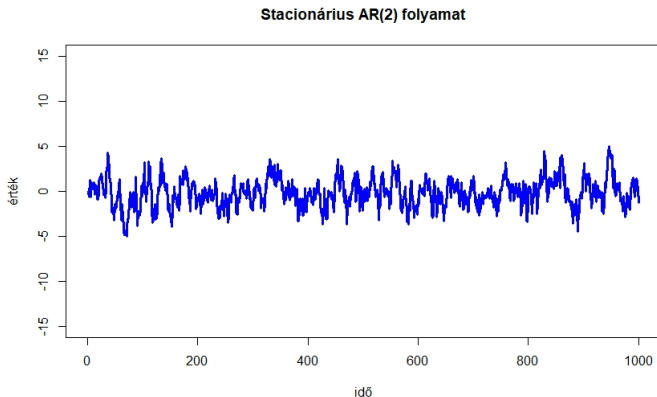
Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ egyenletű AR(2) folyamat három trajektóriája – **ez nem stacionárius**

Másodrendű autoregressziós folyamat autokorrelációs függvénye



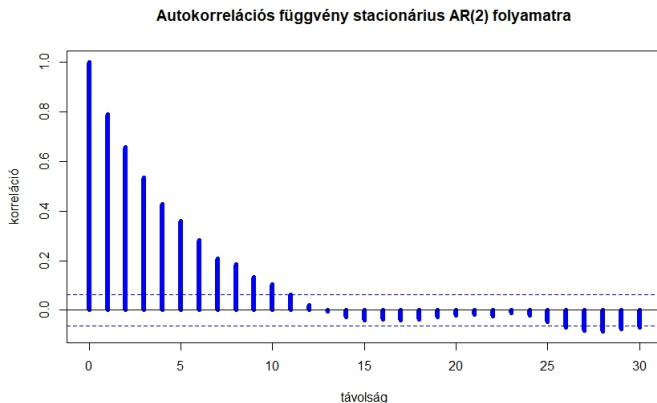
Az $X(t) = 0,7 \cdot X(t - 1) + 0,3 \cdot X(t - 2) + \varepsilon(t)$ egyenletű AR(2) folyamat autokorrelációs függvényének becslése

Másodrendű autoregressziós folyamat



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

Másodrendű autoregressziós folyamat autokorrelációs függvénye



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

Autoregressziós folyamat autokovariancia-függvénye

Definíció

Legyenek $\varepsilon(t)$ független 0 várható értékű 1 szórású valószínűségi változók $t \geq 0$ -ra (például normális eloszlásúak). Az $X(t)$ folyamat p -rendű autoregressziós folyamat, ha minden $t \geq p$ -re

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sigma \cdot \varepsilon(t).$$

Jelölés: $AR(p)$.

Ha egy p -rendű autoregressziós folyamat gyengén stacionárius, azaz várható értéke állandó és a kovariancia csak a távolságtól függ, akkor az alábbiak teljesülnek az autokovariancia-függvényére:

$$R(0) = \alpha_1 R(1) + \alpha_2 R(2) + \dots + \alpha_p R(p) + \sigma_\varepsilon^2;$$

$$R(t) = \alpha_1 R(t-1) + \alpha_2 R(t-2) + \dots + \alpha_p R(t-p),$$

ahol $t \geq 1$ tetszőleges egész.

ARMA-folyamatok

Definíció

Legyenek $\varepsilon(t)$ független 0 várható értékű 1 szórású valószínűségi változók $t \geq 0$ -ra (például normális eloszlásúak). Az $X(t)$ folyamat p, q -rendű autoregressziós mozgóátlag-folyamat, ha minden $t \geq p$ -re

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sum_{m=0}^q \beta_m \varepsilon(t-m).$$

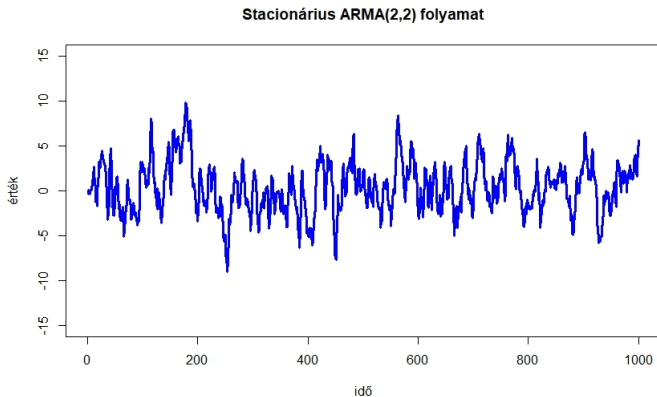
Jelölés: ARMA(p, q).

Például egy másodrendű autoregressziós ARMA(2,2) folyamat ($\alpha_1 = 0,7, \alpha_2 = 0,3, \beta_0 = 0,6, \beta_1 = 0,2, \beta_2 = 0,2$):

$$X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + 0,7 \cdot \varepsilon(t) + 0,2 \cdot \varepsilon(t-1) + 0,2 \cdot \varepsilon(t-2).$$

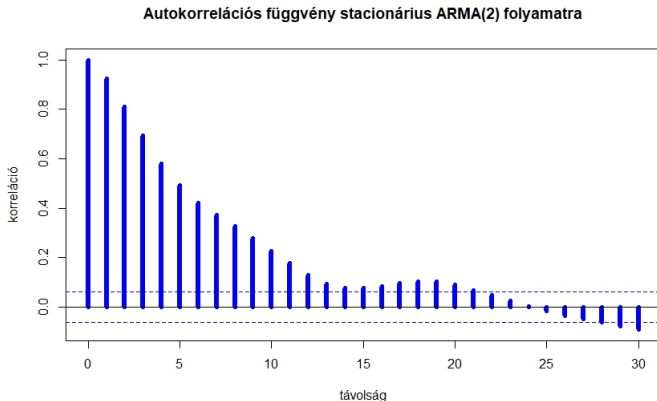
A stacionárius ARMA-folyamatok **rövid emlékezetűek**: $\sum_{t=1}^{\infty} R(t) < \infty$.

Stacionárius ARMA(2,2)-folyamat



Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + 0,7 \cdot \varepsilon(t) + 0,2 \cdot \varepsilon(t-1) + 0,2 \cdot \varepsilon(t-2)$ egyenletű ARMA(2,2) stacionárius folyamat

Stacionárius ARMA(2,2)-folyamat autokorrelációja



Az $X(t) = X(t) = 0,7 \cdot X(t - 1) + 0,3 \cdot X(t - 2) + 0,7 \cdot \varepsilon(t) + 0,2 \cdot \varepsilon(t - 1) + 0,2 \cdot \varepsilon(t - 2)$ egyenletű ARMA(2,2) stacionárius folyamat autokorrelációs függvényének becslése

Példa idősorra

Magyarország népessége 2001-től 2018-ig (forrás: Központi Statisztikai Hivatal)

```
ev<-2001:2018
```

```
nep<-c(10200298, 10174853, 10142362, 10116742, 10097549, 10076581, 10066158,
10045401, 10030975, 10014324, 9985722, 9931925, 9908798, 9877365, 9855571,
9830485, 9797561, 9778371)
```

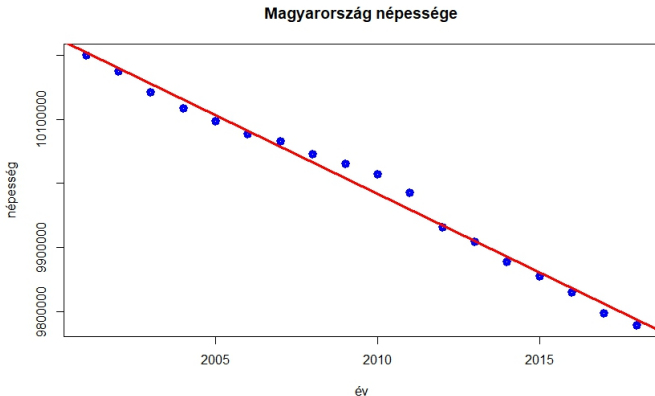
```
summary(lm(nep~ev))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	59315833.4	1320991.3	44.90	<2e-16 ***
ev	-24543.3	657.4	-37.34	<2e-16 ***

```
plot(nep~ev, lwd="5", col="blue", main="Magyarország népessége", xlab="év",
ylab="népesség")
```

```
lines(abline(b=-24543.3, a=59315833.4, lwd="3", col="red"), xlim=c(2000, 2020))
```

Példa idősorra



Magyarország népessége 2001-től 2018-ig (forrás: Központi Statisztikai Hivatal) és a regressziós egyenes

A lineáris trend eltávolítása

Az idősor nem stacionárius, de lehet, hogy egy lineáris függvény és egy stacionárius folyamat összege. Ezért a lineáris regresszióval kapott függvényt kivonjuk:

$$X(t) = N(t) - \hat{a} \cdot t - \hat{b},$$

ahol $N(t)$ a népesség a t időpontban, a regressziós egyenes pedig $\hat{a}x + \hat{b}$ egyenletű.

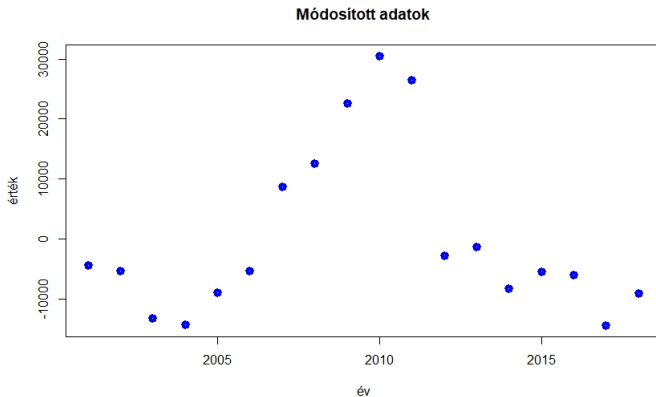
```
ev<-2001:2018
```

```
nep<-c(10200298, 10174853, 10142362, 10116742, 10097549, 10076581, 10066158,  
10045401, 10030975, 10014324, 9985722, 9931925, 9908798, 9877365, 9855571,  
9830485, 9797561, 9778371)
```

```
x<-nep+24543.3*ev-59315833.4
```

```
plot(x~ev, lwd="5", col="blue", main="Módosított adatok", xlab="év", ylab="érték")
```

Példa idősorra



Magyarország népessége 2001-től 2018-ig a lineáris trend eltávolítása után

Példa idősorra

```
ev<-2001:2018
```

```
nep<-c(10200298, 10174853, 10142362, 10116742, 10097549, 10076581, 10066158,  
10045401, 10030975, 10014324, 9985722, 9931925, 9908798, 9877365, 9855571,  
9830485, 9797561, 9778371)
```

```
x<-nep+24543.3*ev-59315833.4
```

```
> ar(x)    # autoregressziós modellt illesztünk
```

```
Call: ar(x = x)
```

```
Coefficients:
```

1	2
1.0115	-0.3336

```
Order selected 2      sigma^2 estimated as 84281456
```

Tehát:

$$X(t) = 1,01 \cdot X(t-1) - 0,33 \cdot X(t-2) + 9180 \cdot \varepsilon(t),$$

ahol $\varepsilon(t)$ korrelálatlan, 0 várható értékű 1 szórású valószínűségi változók.

Példa idősorra: előrejelzés

A becsült egyenlet (a rend kiválasztása az Akaike információs kritérium alapján történt, ami a log-likelihood-függvényen alapszik):

$$X(t) = 1,01 \cdot X(t-1) - 0,33 \cdot X(t-2) + 9180 \cdot \varepsilon(t),$$

ahol $\varepsilon(t)$ korrelálatlan, 0 várható értékű 1 szórású valószínűségi változók.

Előrejelzés 2019-re a módosított idősorban az $X(2019)$ várható értéke (`predict(ar(x), n.ahead=1)`):

$$\begin{aligned} X(2019) &= 1,01 \cdot X(2018) - 0,33 \cdot X(2017) = 1,01 \cdot (-9083) - 0,33 \cdot (-14436) = \\ &= -4409,95 \end{aligned}$$

Ahhoz, hogy az eredeti idősorra vonatkozó előrejelzést megkapjuk, hozzá kell adni a regressziós egyenesből kapott értéket:

$$\begin{aligned} \hat{N}(2019) &= \hat{a} \cdot 2019 + \hat{b} + \hat{X}(2019) = -24543,3 \cdot 2019 + 59315833,4 - 4409,95 = \\ &= 9758501. \end{aligned}$$

Házi feladat május 15-ig, megoldás

A húszelemű mintában osszuk fel a megfigyeléseket három csoportra aszerint, hogy az egyes emberek cipőmérete legfeljebb 39, 40 és 43 között van, vagy legalább 44 (lehet más csoportokat is választani).

Készítsünk szórásElemzést: állíthatjuk-e $\alpha = 0,05$ szignifikanciaszinten, hogy a cipőméret mint faktor szignifikáns hatással van a testmagasságra? Mennyi a p -érték ebben a hipotézisvizsgálati feladatban?

magasság (cm)	≤ 39	40 – 42	≥ 43
160		173	182
158		177	195
162		172	198
168		174	181
168		172	
158		172	
166			
164			
168			
167			

Házi feladat május 15-ig, megoldás

```
> magassag<-c(160, 158, 173, 162, 177, 168, 168, 182, 172, 174, 158, 172, 172,  
195, 166, 164, 198, 181, 168, 167)
```

```
> cipo<-c(36, 35, 41, 38, 42, 39, 39, 43, 42, 42, 35, 40, 40, 44, 38, 37, 46, 43,  
38, 39)
```

```
> csoport<-c(cipo)
```

```
> for(i in 1:20){if(cipo[i]<=39){csoport[i]=1}}
```

```
> for(i in 1:20){if((cipo[i]>=40)&&(cipo[i]<=42)){csoport[i]=2}}
```

```
> for(i in 1:20){if(cipo[i]>=43){csoport[i]=3}}
```

```
> csoport
```

```
[1] 1 1 2 1 2 1 1 3 2 2 1 2 2 3 1 1 3 3 1 1
```

Házi feladat május 15-ig, megoldás

A húszelemű mintában osszuk fel a megfigyeléseket három csoportra aszerint, hogy az egyes emberek cipőmérete legfeljebb 39, 40 és 43 között van, vagy legalább 44 (lehet más csoportokat is választani).

Készítsünk szórásElemzést: állíthatjuk-e $\alpha = 0,05$ szignifikanciaszinten, hogy a cipőméret mint faktor szignifikáns hatással van a testmagasságra? Mennyi a p -érték ebben a hipotézisvizsgálati feladatban?

```
> summary(aov(magassag~csoport))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
csoport	1	1783.3	1783.3	72.88	9.6e-08 ***
Residuals	18	440.5	24.5		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H_0 : a testmagasság várható értéke ugyanannyi az egyes csoportokban

H_1 : a testmagasság várható értéke nem ugyanannyi az egyes csoportokban

Mivel $p = 9,6 \cdot 10^{-8} < 0,05$, elutasítjuk a nullhipotézist, **szignifikáns eltérés** van a várható értékek között legalább két csoport esetében