

## Nemparaméteres próbák (10. előadás)

**Illeszkedésvizsgálat:** a minta egy adott, folytonos eloszlásból származik-e?

**Homogenitásvizsgálat:** két minta ugyanabból az eloszlásból származik-e?

Egy lehetőség: **diszkrétizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket, és az így kapott diszkrét eloszlásra  $\chi^2$ -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük.

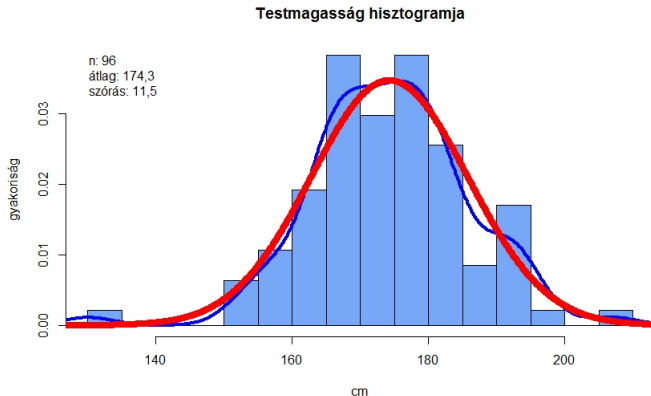
**Tapasztalati eloszlásfüggvények** távolságát használó próbák:

- Kolmogorov–Szmirnov-próba
- Anderson–Darling-próba (az eltéréseket másképp súlyozzuk)
- Cramér–von Mises próba (az eltéréseket másképp súlyozzuk)

Speciálisan annak ellenőrzésére, hogy egy eloszlás **normális eloszlású**-e:

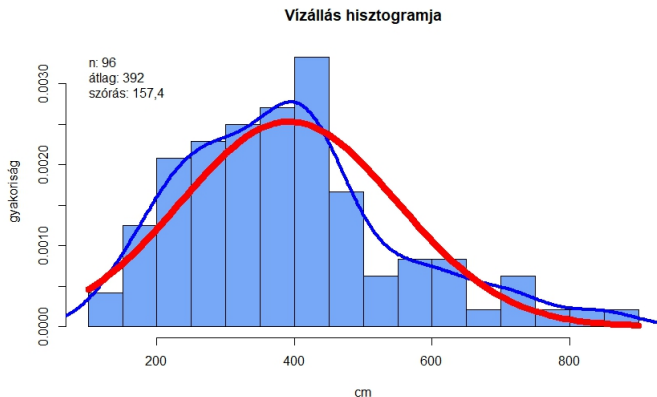
- Lilliefors-próba (a Kolmogorov–Szmirnov-próbán alapul)
- Shapiro–Wilk-próba (a rendezett minta várható értékét és kovarianciamátrixát használja)

# Testmagasság és normális eloszlás



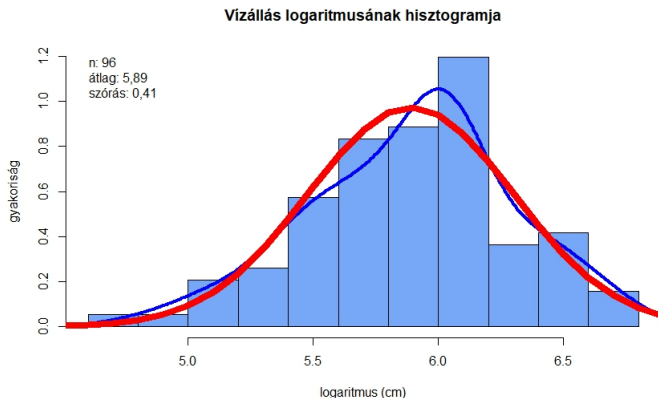
A testmagasság histogramja  $n = 96$  elemű mintából, a sűrűségfüggvény becslése Gauss-magfüggvénnyel, és az  $\bar{X} = 174,3$  várható értékű és  $s_n^* = 11,5$  szórású normális eloszlás sűrűségfüggvénye.

# A Duna vízállása



A Duna havi legnagyobb vízállásának hisztogramja (2002–2009,  $n = 96$ , forrás: Országos Vízelző Szolgálat), a becsült sűrűségfüggvény, és az  $\bar{X} = 392$  várható értékű és  $s_n^* = 157,4$  szórású normális eloszlás sűrűségfüggvénye – **itt a függetlenség nem teljesen érvényes**

# A Duna vízállása



A Duna havi legnagyobb vízállásának **logaritmusának** hisztogramja (2002–2009,  $n = 96$ , forrás: Országos Vízelző Szolgálat), a becsült sűrűségfüggvény, és az  $\bar{X} = 392$  várható értékű és  $s_n^* = 157,4$  szórású normális eloszlás sűrűségfüggvénye

# Tapasztalati eloszlásfüggvény

Az  $X$  valószínűségi változó eloszlásfüggvénye az  $F : \mathbb{R} \rightarrow [0, 1]$  függvény, melyre

$$F(t) = \mathbb{P}(X \leq t)$$

minden  $t \in \mathbb{R}$ -re.

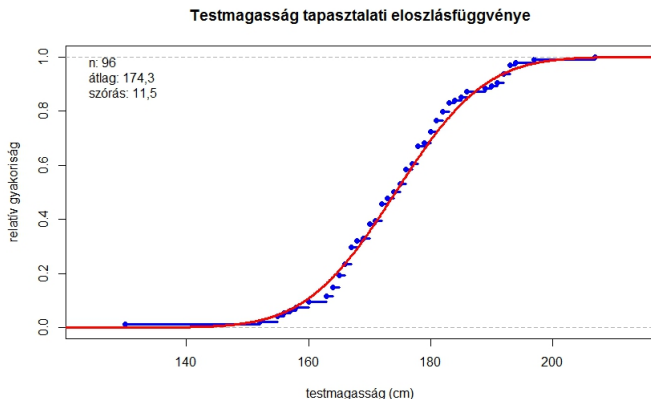
## Definíció (Tapasztalati eloszlásfüggvény)

Az  $X_1, X_2, \dots, X_n$  minta tapasztalati eloszlásfüggvénye az  $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$  függvény, melyre

$$\hat{F}_n(t) = \frac{t\text{-nél nem nagyobb mintaelemek száma}}{n}.$$

(empirical cumulative distribution function)

# Tapasztalati eloszlásfüggvény



A testmagasság tapasztalati eloszlásfüggvénye  $n = 96$  elemű mintából, és az  $\bar{X} = 174,3$  várható értékű és  $s_n^* = 11,5$  szórású normális eloszlás eloszlásfüggvénye.

# Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

$H_0$  : a minta valódi eloszlásfüggvénye  $F$  (ami folytonos)

$H_1$  : a minta valódi eloszlásfüggvénye  $F$ -től különböző

Próbastatisztika:

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)|,$$

ahol  $F_n$  a minta tapasztalati eloszlásfüggvénye.

Ha  $D_n > D_{\text{krit}}$  (vagy  $p < \alpha$ ), akkor elutasítjuk  $H_0$ -t, a minta eloszlásfüggvénye szignifikánsan eltér  $D$ -től (itt  $D_{\text{krit}}$  a megfelelő Kolmogorov–Szmirnov-próba kritikus értéke).

Ha  $D_n < D_{\text{krit}}$ , (vagy  $p > \alpha$ ) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés  $F$ -től.

Ha  $n \geq 35$ , akkor a kritikus értékre az alábbi közelítés adható ( $\alpha$  szignifikanciaszint mellett):

$$D_{\text{krit}} \approx \frac{\sqrt{\log(4/\alpha)}}{\sqrt{n}}.$$

## A normalitás tesztelése: Lilliefors-próba

$H_0$  : a testmagasság normális eloszlású (valamilyen  $m, \sigma$  paraméterekkel)

$H_1$  : a testmagasság eloszlása nem normális eloszlás

## A normalitás tesztelése: Lilliefors-próba

$H_0$  : a testmagasság normális eloszlású (valamilyen  $m, \sigma$  paraméterekkel)

$H_1$  : a testmagasság eloszlása nem normális eloszlás

Legyen  $\bar{X}$  a mintátlag,  $s_n$  a tapasztalati szórás,  $F$  pedig az  $m$  várható értékű és  $\sigma$  szórású normális eloszlás eloszlásfüggvénye:  $F(t) = \Phi((t - \bar{X})/s_n)$ . Ekkor a próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

$$D = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0,068$$

Ha  $D_n > \bar{D}_{\text{krit}}$  (vagy  $p < \alpha$ ), akkor elutasítjuk  $H_0$ -t, a minta eloszlásfüggvénye szignifikánsan eltér  $D$ -től (itt  $D_{\text{krit}}$  a megfelelő Lilliefors-próba kritikus értéke).

Ha  $D_n < \bar{D}_{\text{krit}}$ , (vagy  $p > \alpha$ ) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés  $F$ -től.

A példában  $\alpha = 0,05$  szignifikanciaszinttel a kritikus érték: 0,09, a  $p$ -érték: 0,367

## A normalitás tesztelése: Lilliefors-próba

$H_0$  : a testmagasság normális eloszlású (valamilyen  $m, \sigma$  paraméterekkel)

$H_1$  : a testmagasság eloszlása nem normális eloszlás

Legyen  $\bar{X}$  a mintátlag,  $s_n$  a tapasztalati szórás,  $F$  pedig az  $m$  várható értékű és  $\sigma$  szórású normális eloszlás eloszlásfüggvénye:  $F(t) = \Phi((t - \bar{X})/s_n)$ . Ekkor a próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

$$D = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0,068$$

Ha  $D_n > \bar{D}_{\text{krit}}$  (vagy  $p < \alpha$ ), akkor elutasítjuk  $H_0$ -t, a minta eloszlásfüggvénye szignifikánsan eltér  $D$ -től (itt  $D_{\text{krit}}$  a megfelelő Lilliefors-próba kritikus értéke).

Ha  $D_n < \bar{D}_{\text{krit}}$ , (vagy  $p > \alpha$ ) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés  $F$ -től.

A példában  $\alpha = 0,05$  szignifikanciaszinttel a kritikus érték:  $0,09$ , a  $p$ -érték:  $0,367$

Mivel  $0,068 = D < D_{\text{krit}} = 0,09$ , illetve  $p = 0,367 > 0,05 = \alpha$ , a szignifikanciaszintet  $\alpha = 0,05$ -nek választva **elfogadható**, hogy a testmagasság normális eloszlású a megadott paraméterekkel, nincs szignifikáns eltérés.

## Normális eloszlásra vonatkozó próbák: példa

A táblázat a  $p$ -értékeket mutatja az egyes minták és próbák esetén.

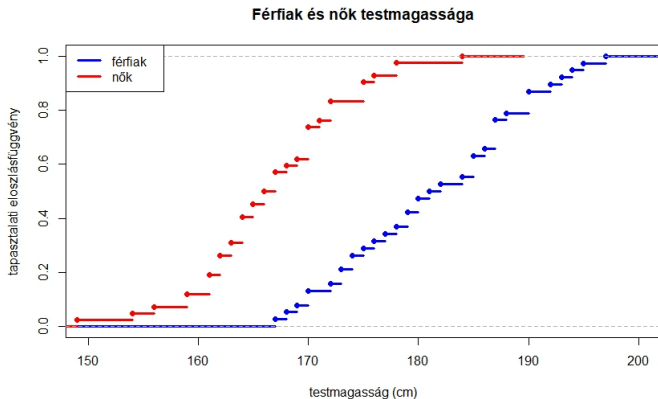
Szignifikanciaszint:  $\alpha = 0,05$

minta	Lilliefors (Kolmogorov–Szmirnov)	Shapiro–Wilk
testmagasság	0,367	0,066
max. vízállás	0,014	0,002
max. vízállás logaritmusa	0,22	0,629

Tehát a testmagasság és Duna havi legnagyobb vízállásának logaritmusáról elfogadható, hogy normális eloszlású, a havi legnagyobb vízállás eloszlása viszont szignifikánsan eltér a normális eloszlástól.

A Duna havi legnagyobb vízállásáról azt mondhatjuk, hogy a logaritmusa normális eloszlású, vagyis **lognormális eloszlású**.

# Kolmogorov–Szmirnov-próba: homogenitásvizsgálat



A férfiak ( $n = 38$  megfigyelés) és nők ( $m = 42$  megfigyelés) testmagasságának tapasztalati eloszlásfüggvénye

## Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

$H_0$  : az  $X_1, \dots, X_n$  és  $Y_1, \dots, Y_m$  minták ugyanabból az eloszlásból származnak

$H_1$  : a minták különböző eloszlásból származnak.

Próbastatisztika:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|,$$

ahol  $\hat{F}_n$  az  $X$ , a  $\hat{G}_m$  pedig az  $Y$  minta tapasztalati eloszlásfüggvénye.

Ha  $D_{m,n} > D_{\text{krit}}$  (vagy  $p < \alpha$ ), akkor elutasítjuk  $H_0$ -t, a minták eloszlása szignifikánsan különböző (itt  $D_{\text{krit}}$  a megfelelő Kolmogorov–Szmirnov-próba kritikus értéke).

Ha  $D < D_{\text{krit}}$ , (vagy  $p > \alpha$ ) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés a minták eloszlása között.

A kritikus értékek az alábbi összefüggés alapján közelíthetők:

$$\lim_{m,n \rightarrow \infty} \mathbb{P} \left( \sqrt{\frac{mn}{m+n}} D_{m,n} < y \right) = \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 y^2} \Rightarrow D_{\text{krit}} \approx \sqrt{\frac{m+n}{mn}} \sqrt{-\frac{1}{2} \log \alpha}.$$

## Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

$H_0$  : az  $X_1, \dots, X_n$  és  $Y_1, \dots, Y_m$  minták ugyanabból az eloszlásból származnak

$H_1$  : a minták különböző eloszlásból származnak.

Próbastatisztika:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|,$$

ahol  $\hat{F}_n$  az  $X$ , a  $\hat{G}_m$  pedig az  $Y$  minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha  $D$  nagyobb a kritikus értéknél.

---

```
> ks.test(ferfi, no, alternative="two.sided")
```

Two-sample Kolmogorov-Smirnov test

data: ferfi and no

D = 0.6754, **p-value = 2.486e-08**

alternative hypothesis: two-sided

---

A férfiak ( $n = 38$  megfigyelés) és a nők ( $m = 42$  megfigyelés) testmagasságának eloszlása szignifikánsan különböző.

## Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

$H_0$  : az  $X$  és  $Y$  valószínűségi változók ugyanabból az eloszlásból származnak

$H_1$  : minden  $t$  valós számra  $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$ , azaz  $X \geq Y$  sztochasztikusan, ahol  $F$  az  $X$ , a  $G$  pedig az  $Y$  eloszlásfüggvénye.

Próbastatisztika:

$$D_{m,n}^- = \sup_{t \in \mathbb{R}} \hat{G}_n(t) - \hat{F}_m(t),$$

ahol  $\hat{F}_n$  az  $X$ , a  $\hat{G}_m$  pedig az  $Y$  minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha  $D$  nagyobb a kritikus értéknél.

---

```
> ks.test(ferfi, no, alternative="less")
```

Two-sample Kolmogorov-Smirnov test

data: ferfi and no

$D^- = 0.6754$ , **p-value = 1.243e-08**

alternative hypothesis: the CDF of x lies below that of y

A férfiak ( $n = 38$  megfigyelés) testmagassága szignifikánsan nagyobb nőknél ( $m = 42$  megfigyelés) .

## Előjelpróba

Legyen  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem független), folytonos eloszlásúak.

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

Legyen  $W$  az olyan párok száma, amikre  $Y_i > X_i$ . A  $H_0$  nullhipotézis teljesülése esetén ez binomiális eloszlású,  $n$  renddel és  $p = 0,5$  paraméterrel. Ezt az eloszlást normális eloszlással közelítjük. Így:

$$z = \frac{W - n/2}{\sqrt{n/4}}.$$

Elutasítjuk a nullhipotézist, ha  $|z| > \Phi^{-1}(1 - \alpha/2)$ .

Az egyoldali esetben:  $H_0 : \mathbb{P}(X > Y) \geq \mathbb{P}(X < Y)$ .  $H_1 : \mathbb{P}(X > Y) < \mathbb{P}(X < Y)$ .

Elutasítjuk a nullhipotézist, ha  $z > \Phi^{-1}(1 - \alpha)$ .

## Wilcoxon-próba

Legyen  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem független), folytonos eloszlásúak.

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

- Hagyjuk el azokat a párokat, ahol  $X_j = Y_j$ . Marad  $k$  pár.
- A megmaradt  $k$  párt állítsuk az  $|Y_j - X_j|$  szerint növekvő sorrendbe.
- Minden párra számítsuk ki, hogy hányadik ebben a sorrendben, legyen ez  $R_j$ . Az 1 a legkisebb,  $k$  a legnagyobb. Ha egyenlők vannak, mindegyik azonos sorszámot kapjon, a megfelelő sorszámok átlagát.
- Ezt az  $R_j$  rangot szorozzunk meg  $Y_j - X_j$  előjelével, majd ezeket adjuk össze:  
$$W = \sum_{j=1}^k \text{sgn}(Y_j - X_j) \cdot R_j.$$
- A  $W$ -t a Wilcoxon-próba kritikus értékeihez hasonlíthatjuk. Ha a mintaelem-szám elég nagy, a

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}}$$

mennyiségre kétoldali  $z$ -próbát alkalmazhatunk, a kritikus érték ebben az esetben  $1 - \Phi^{-1}(1 - \alpha/2)$ , ahol  $\alpha$  a szignifikanciaszint.

# Kvantilisek

Az  $X$  valószínűségi változó  $z$ -kvantilise a legkisebb olyan  $q$  szám, melyre teljesül, hogy  $\mathbb{P}(X \leq q) \geq z$ .

A tapasztalati  $z$ -kvantilis a legkisebb olyan  $q$  szám, melyre  $\hat{F}_n(q) \geq z$  teljesül, vagy egy másik lehetőség (ezen kívül is több definíciót szoktak használni):

## Definíció (Tapasztalati kvantilis)

Legyen  $X_1^* \leq X_2^* \leq \dots \leq X_n^*$  rendezett minta, és  $z \in [0, 1]$  adott szám. Ekkor a minta tapasztalati  $z$ -kvantilise:

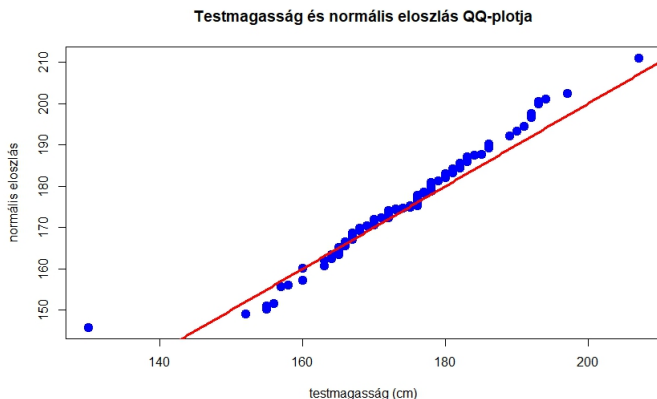
$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

Első kvartilis:  $z = 1/4$ -kvantilis, harmadik kvartilis:  $z = 3/4$ -kvantilis, a medián pedig a  $z = 1/2$ -hez tartozó tapasztalati kvantilis.

# QQ-plot

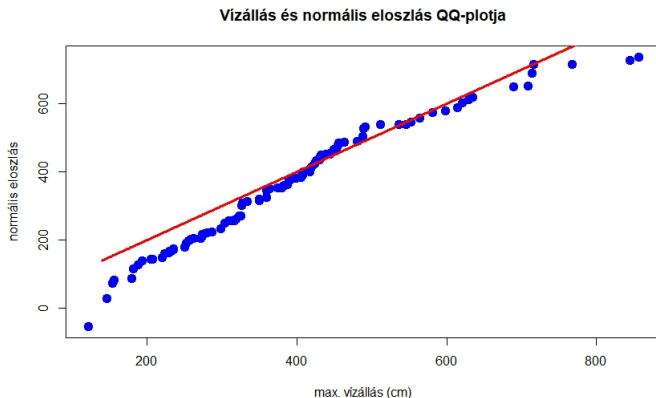
- a QQ-plot két minta eloszlásának az összehasonlítására szolgál, a kvantilisok összehasonlításával
- ha a tapasztalati z-kvantilis az első mintában  $q_1$ , a másodikban  $q_2$ , akkor a  $(q_1, q_2)$  pontba kerül egy pont
- minél inkább egyezik a két minta eloszlása, annál közelebb lesz a QQ-plot az  $y = x$  egyeneshez

# A testmagasság és a becsült normális eloszlás QQ-plotja



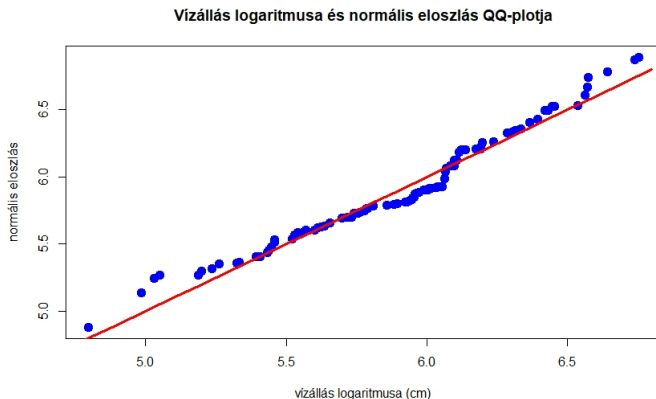
A testmagasság adatok és egy szintén 96 elemű,  $\bar{X} = 174,3$  várható értékű és  $s_n^* = 11,5$  szórású normális eloszlású minta QQ-plotja

# A vízállás és a becült normális eloszlás QQ-plotja



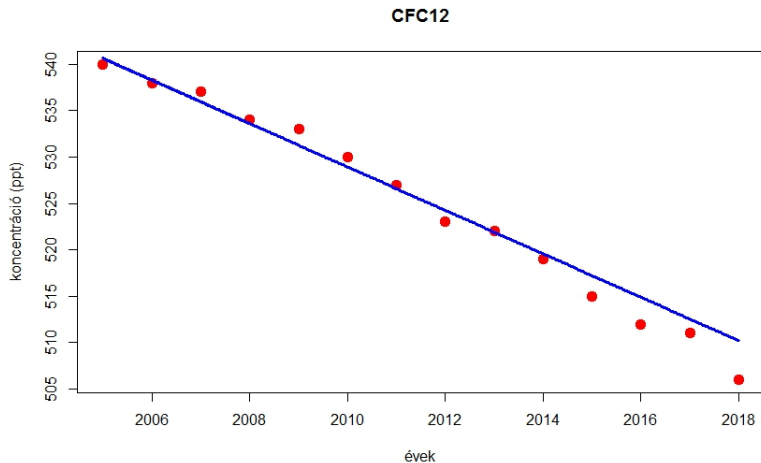
A havi legnagyobb vízállás adatok és egy szintén 96 elemű,  $\bar{X} = 352$  várható értékű és  $s_n^* = 157,4$  szórású normális eloszlású minta QQ-plotja (ez szignifikánsan eltért a normális eloszlástól)

# A vízállás logaritmusának és a becsült normális eloszlás QQ-plotja



A havi legnagyobb vízállás adatok és egy szintén 96 elemű,  $\bar{X} = 5,89$  várható értékű és  $s_n^* = 0,41$  szórású normális eloszlású minta QQ-plotja (ez nem tért el szignifikánsan a normális eloszlástól)

# Lineáris regresszió



A CFC-12 (freon) gáz koncentrációja az Antarktison és az adatokra illesztett egyenes (forrás: ESRL, USA)

# Lineáris regresszió

Egyenes illesztése a **legkisebb négyzetek módszerével**:

## Állítás (Lineáris regresszió)

*Legyenek  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  adott számpárok. Azokat az  $a$  és  $b$  együtthatókat keressük, melyre a*

$$h^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

*mennyiség minimális. Ennek megoldása:*

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

A példában:  $\hat{a} = -2,63$ ;  $\hat{b} = 5807,7$  (a  $b$  együttható neve: intercept)

## Lineáris modell: példa R-ben

```
> cfc12<-c(540, 538, 537, 534, 533, 530, 527, 523, 522, 519, 515, 511, 506)
```

```
> ev<-c(seq(from=2005, to=2018, by=1))
```

```
> summary(lm(cfc12 ~ ev))
```

```
Call:  lm(formula = cfc12 ~ ev)
```

```
Residuals:      Min       1Q   Median       3Q      Max
             -1.8571  -0.8736   0.2088   0.8709   1.6483
```

## Lineáris modell: példa R-ben (folytatás)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5807.73626	159.19290	36.48	1.15e-13 ***
ev	-2.62637	0.07914	-33.19	3.55e-13 ***

--

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

Residual standard error: 1.194 on 12 degrees of freedom

Multiple R-squared: 0.9892, Adjusted R-squared: 0.9883

F-statistic: 1101 on 1 and 12 DF, p-value: 3.554e-13

# Lineáris modell

## Definíció (Lineáris modell)

Legyenek  $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$  valószínűségi változók, és tegyük fel, hogy valamely  $a, b$  valós számokra

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol  $\varepsilon_1, \dots, \varepsilon_n$  független  $N(0, \sigma^2)$  normális eloszlású valószínűségi változók. Az így kapott  $(X_i, Y_i)$  párok együttes eloszlását lineáris modellnek nevezzük.

Az  $X_i$  valószínűségi változókat magyarázó változóknak, az  $\varepsilon_i$  valószínűségi változókat hibának szokták nevezni.

# Becslések a lineáris modellben

## Állítás

A lineáris modellben az  $a, b$  együtthatók maximumlikelihood-becslése a következőképpen írható:

$$\hat{a} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{k=1}^n (X_k - \bar{X})^2}; \quad \hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az  $a$  és  $b$  paramétereknek. A hiba szórásának becslése:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

$$D(\hat{a}) = \frac{\sigma}{\sum_{j=1}^n (X_j - \bar{X})^2}; \quad D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

# Előrejelzés a lineáris modellben

## Állítás

Legyen  $x^*$  adott szám. A lineáris modellből kapott előrejelzés az  $Y$  véletlen folyamat  $x^*$  pontban felvett értékére:

$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

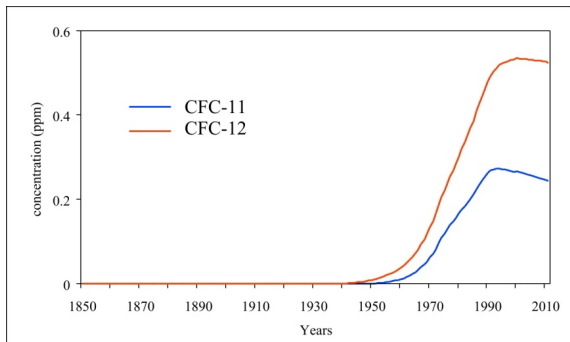
$$D(\hat{a}x^* + \hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

Az előrejelzés szórásának becslésekor a  $\sigma$  értéket gyakran  $\hat{\sigma}$ -val helyettesítik.

A példában: előrejelzés  $x^* = 2019$ -re:

$$\hat{a} \cdot x^* + \hat{b} = -2,63 \cdot 2019 + 5807,7 = 497,7.$$

# Előrejelzés a lineáris modellben



A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

## Reziduálisok

Reziduálisok:  $Y_i - \hat{a}X_i - \hat{b}$  (ezeknek a négyzetösszege minimális)

A teljes ingadozás (total sum of squares):  $\sum_{j=1}^n (Y_j - \bar{Y})^2$ .

### Definíció

A megmagyarázott ingadozás részaránya (coefficient of determination):

$$R^2 = \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\sum_{k=1}^n (X_k - \bar{X})^2][\sum_{k=1}^n (Y_k - \bar{Y})^2]}.$$

Az  $R^2$  értéke 0 és 1 közé esik.

Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. Ugyanakkor  $R$  érzékeny a kiugró értékekre, néhány kiugró esetén  $R^2$  lecsökken.

A példában:  $R^2 = 0,98$ , vagyis jól illeszkedik a lineáris modell.

## Házi feladat május 8., 9:00-ig

Tekintsük a húszelemű mintában a cipőméreteket ( $Y_j$ ) a testmagasság ( $X_j$ ) függvényében.

- 1 Határozzuk meg a regressziós egyenes egyenletét és a megmagyarázott ingadozás részarányát.
- 2 Ábrázoljuk a cipőméretet a testmagasság függvényében a regressziós egyenessel együtt.

## Házi feladat április 24., 9:00-ig

- 1 A húszelemű minta alapján állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy szignifikáns pozitív korreláció van aközött, hogy valakinek legalább 175 cm a testmagassága, és legalább 40-es a cipőmérete? Határozzuk meg a  $p$ -értéket is.

A 175 és 40 helyett választhatunk más, tetszőleges határokat, viszont, ha lehetséges, minden osztályba essen legalább 3 megfigyelés (a kevés adat miatt a legalább 6 most nem lenne megvalósítható).

- 2 A húszelemű minta alapján állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy a nem és a cipőméret szignifikánsan összefüggő szempontok?

## Házi feladat április 24., 9:00-ig

A húszelemű minta alapján állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy szignifikáns pozitív korreláció van aközött, hogy valakinek legalább 175 cm a testmagassága, és legalább 42-es a cipőmérete? Határozzuk meg a  $p$ -értéket is.

	$\geq 42$	$\leq 41$	összesen
$\geq 175$	16	8	24
$< 175$	6	30	36
összesen	22	38	60

$H_0$ : nincs pozitív korreláció;  $H_1$ : pozitív korreláció van

$$z = \sqrt{n} \cdot \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{60} \frac{16 \cdot 30 - 6 \cdot 8}{\sqrt{24 \cdot 36 \cdot 22 \cdot 38}} = 3,94 > \Phi^{-1}(0,95) = 1,65.$$

A  $p$ -érték:  $p = 1 - \Phi(3,94) = 4,1 \cdot 10^{-5}$ . Elutasítjuk a nullhipotézist, **szignifikáns pozitív korreláció van** a legalább 175 cm-es testmagasság és a legalább 42-es cipőméret között.

## Házi feladat április 24., 9:00-ig

A húszelemű minta alapján állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy a nem és a cipőméret szignifikánsan összefüggő szempontok?

$H_0$ : a nem és a cipőméret független szempontok;  $H_1$ : nem függetlenek

cipőméret	35 – 39	40 – 41	42 – 48	összesen
férfi	6	7	16	29
nő	16	8	4	28
összesen	22	15	20	57

$$\begin{aligned}\chi^2 &= \sum_{k=1}^r \frac{\left(\frac{N_k}{n} - \frac{M_k}{m}\right)^2}{\frac{N_k}{n} + \frac{M_k}{m}} \cdot n \cdot m = \\ &= \frac{(6/29 - 16/28)^2}{22} + \frac{(7/29 - 8/28)^2}{15} + \frac{(16/29 - 4/28)^2}{20} = 0,014.\end{aligned}$$

Az  $f = r - 1 = 2$  szabadsági fokú  $\chi^2$ -próba kritikus értéke  $\alpha = 0,05$  szignifikanciaszinten: 5,99, a  $p$ -érték:  $p = 0,99$ . Elfogadjuk a nullhipotézist, **nincs szignifikáns összefüggés** a férfiak és nők cipőmérete között (vagy nincs összefüggés, vagy kevés az adat kimutatni).