

# Maximumlikelihood-módszer

## Definíció (Likelihood-függvény)

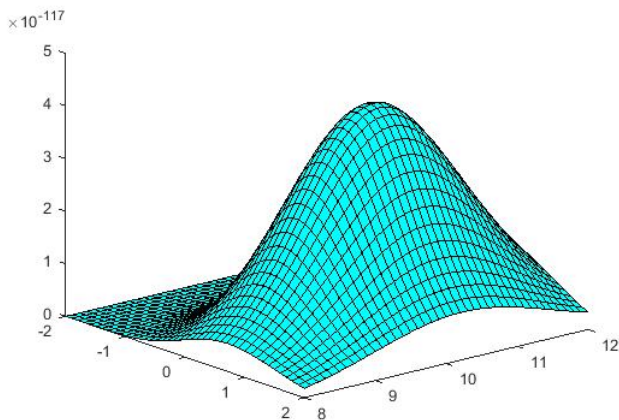
Ha az  $(Y_1, \dots, Y_n)$  független minta diszkrét (a lehetséges értékeinek száma véges vagy megszámlálható sok), akkor a likelihood-függvénye:

$$L_{n,\vartheta}(k_1, \dots, k_n) = \prod_{j=1}^n \mathbb{P}_{j,\vartheta}(Y_j = k_j) \quad ((k_1, \dots, k_n) \in H).$$

Ha az  $(Y_1, \dots, Y_n)$  független minta abszolút folytonos, és  $Y_j$  sűrűségfüggvénye (a  $\mathbb{P}_\vartheta$  valószínűség mellett)  $f_{j,\vartheta}$ , akkor a minta likelihood-függvénye:

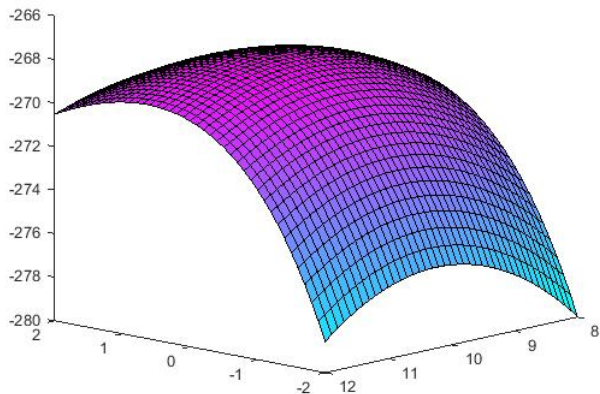
$$L_{n,\vartheta}(t_1, \dots, t_n) = \prod_{j=1}^n f_{j,\vartheta}(t_j) \quad (t_1, \dots, t_n \in \mathbb{R}).$$

# Likelihoodfüggvény



$n = 94$  elemű minta testmagasság-adatok alapján, normális eloszlást feltételezve.  
Az átlag:  $\bar{X} = 174,8$ , a tapasztalati szórás  $s_n = 10,5$ .

## Log-likelihoodfüggvény



$n = 94$  elemű minta testmagasság-adatok alapján, normális eloszlást feltételezve.  
Az átlag:  $\bar{X} = 174,8$ , a tapasztalati szórás  $s_n = 10,5$ .

## ML-becslés: normális eloszlás

$X_1, \dots, X_n$  függetlenek, eloszlásuk normális eloszlás  $m, \sigma > 0$  paraméterekkel. Ekkor

$$L_{n,m,\sigma}(X_1, \dots, X_n) = \prod_{j=1}^n f_{j,\vartheta}(X_j) = \prod_{j=1}^n \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_j - m)^2}{2\sigma^2}\right) \right].$$

## ML-becslés: normális eloszlás

$X_1, \dots, X_n$  függetlenek, eloszlásuk normális eloszlás  $m, \sigma > 0$  paraméterekkel. Ekkor

$$L_{n,m,\sigma}(X_1, \dots, X_n) = \prod_{j=1}^n f_{j,\vartheta}(X_j) = \prod_{j=1}^n \left[ \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(X_j - m)^2}{2\sigma^2}\right) \right].$$

$$L_{n,m,\sigma}(X_1, \dots, X_n) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\sum_{j=1}^n \frac{(X_j - m)^2}{2\sigma^2}\right).$$

## ML-becslés: normális eloszlás

$X_1, \dots, X_n$  függetlenek, eloszlásuk normális eloszlás  $m, \sigma > 0$  paraméterekkel. Ekkor

$$L_{n,m,\sigma}(X_1, \dots, X_n) = \prod_{j=1}^n f_{j,\vartheta}(X_j) = \prod_{j=1}^n \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_j - m)^2}{2\sigma^2}\right) \right].$$

$$L_{n,m,\sigma}(X_1, \dots, X_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{j=1}^n \frac{(X_j - m)^2}{2\sigma^2}\right).$$

$$\log L_{n,m,\sigma}(X_1, \dots, X_n) = -n \log(\sqrt{2\pi}) - n \log \sigma - \sum_{j=1}^n \frac{(X_j - m)^2}{2\sigma^2}.$$

Rögzített  $\sigma$  mellett ez akkor maximális, ha  $\sum_{j=1}^n (X_j - m)^2 = \sum_{j=1}^n X_j^2 - 2 \sum_{j=1}^n X_j m + nm^2$  minimális  $\Rightarrow \hat{m} = \bar{X}$ .

## ML-becslés: normális eloszlás

$$\log L_{n,\sigma}(X_1, \dots, X_n) = -n \log(\sqrt{2\pi}) - n \log \sigma - \sum_{j=1}^n \frac{(X_j - \bar{X})^2}{2\sigma^2}.$$

A  $\sigma$  szerinti parciális derivált:

$$\frac{\partial}{\partial \sigma} \log L_{n,\sigma}(X_1, \dots, X_n) = -\frac{n}{\sigma} + \sum_{j=1}^n \frac{(X_j - \bar{X})^2}{\sigma^3}.$$

Ez pontosan akkor pozitív, ha  $\sigma^2 < \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = s_n^2$ .

Tehát az ML-becslés:

$$\hat{m} = \bar{X}; \quad \hat{\sigma} = s_n = \sqrt{\frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2}.$$

Tehát normális eloszlásnál az  $m$  paraméter becslése a mintaátlag, a szórásé a tapasztalati szórás. Ebben a speciális esetben a momentum módszer és az ML-módszer ugyanazt adja eredményül.

## ML-becslés: egyenletes eloszlás

Ebben az esetben az ML-becslés nem számítható ki az ML-egyenlet gyökeként, vagyis nem kapható meg deriválással.

$X_1, \dots, X_n$  függetlenek, eloszlásuk egyenletes eloszlás az  $[a, b]$  intervallumon. Ekkor

$$L_{n,a,b}(X_1, \dots, X_n) = \prod_{j=1}^n f_{j,\vartheta}(X_j) = \prod_{j=1}^n \mathbb{I}(a \leq X_j \leq b) \cdot \frac{1}{b-a}.$$

$$L_{n,a,b}(X_1, \dots, X_n) = \left(\frac{1}{b-a}\right)^n \mathbb{I}(a \leq \min_j X_j \text{ és } \max_j X_j \leq b).$$

## ML-becslés: egyenletes eloszlás

Ebben az esetben az ML-becslés nem számítható ki az ML-egyenlet gyökeként, vagyis nem kapható meg deriválással.

$X_1, \dots, X_n$  függetlenek, eloszlásuk egyenletes eloszlás az  $[a, b]$  intervallumon. Ekkor

$$L_{n,a,b}(X_1, \dots, X_n) = \prod_{j=1}^n f_{j,\vartheta}(X_j) = \prod_{j=1}^n \mathbb{I}(a \leq X_j \leq b) \cdot \frac{1}{b-a}.$$

$$L_{n,a,b}(X_1, \dots, X_n) = \left(\frac{1}{b-a}\right)^n \mathbb{I}(a \leq \min_j X_j \text{ és } \max_j X_j \leq b).$$

Az első tényező legyen minél nagyobb (vagyis  $b - a$  minél kisebb) úgy, hogy a második tényező nem nulla. Ebből:

$$\hat{a} = \min_j X_j; \quad \hat{b} = \max_j X_j.$$

## Maximum-likelihood becslések

- binomiális eloszlás ismert  $k$  renddel:  $\hat{p} = \bar{X}/k$
- Poisson-eloszlás:  $\hat{\lambda} = \bar{X}$
- geometriai eloszlás:  $\hat{p} = 1/\bar{X}$
- normális eloszlás:  $\hat{m} = \bar{X}, \hat{\sigma} = s_n$
- exponenciális eloszlás:  $\hat{\lambda} = 1/\bar{X}$
- egyenletes eloszlás:  $\hat{a} = \min_j X_j; \quad \hat{b} = \max_j X_j$

# Az ML-becslés tulajdonságai

- Nem minden statisztikai mezőn létezik ML-becslés.
- Az ML-becslés nem feltétlenül egyértelmű.
- Az ML-becslés nem feltétlenül torzítatlan.
- A  $\psi(\vartheta)$  függvény ML-becslése  $\psi(\hat{\vartheta})$ , ahol  $\hat{\vartheta}$  ML-becslés  $\vartheta$ -ra.
- Az alábbi egyenlet a maximumlikelihood-egyenlet:

$$\frac{\partial}{\partial \vartheta} \log L_{n,\vartheta}(X_1, \dots, X_n) = 0.$$

Megfelelő feltételek mellett az ML-becslés a maximumlikelihood-egyenlet megoldása (ha az ML-becslés nem számítható ki, de az egyenlet megoldható, gyakran az egyenlet megoldásával helyettesítik az ML-becslést).

# A maximum-likelihood becslés tulajdonságai

Ha likelihoodfüggvény teljesít bizonyos regularitási feltételeket, akkor a  $\vartheta$  paraméternek az  $X_1, X_2, \dots, X_n$  mintából számolt  $\hat{\vartheta}_n$  maximumlikelihood-becslése

- létezik;
- aszimptotikusan torzítatlan:  $\lim_{n \rightarrow \infty} \mathbb{E}_{\vartheta}(\hat{\vartheta}_n) = \vartheta$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan hatásos:  $\lim_{n \rightarrow \infty} \sqrt{nl_1(\vartheta)} D_{\vartheta}(\hat{\vartheta}_n) = 1$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan normális eloszlású:  $\sqrt{nl_1(\vartheta)}(\hat{\vartheta}_n - \vartheta)$  eloszlásban tart a standard normális eloszláshoz minden  $\vartheta \in \Theta$ -ra  $n \rightarrow \infty$  esetén.

# Momentumok

## Definíció

Az  $X$  valószínűségi változó  $k$ . momentuma:

$$\mathbb{E}(X^k) \quad (k \geq 1),$$

ha ez a várható érték létezik.

Ha az  $X$  valószínűségi változó diszkrét, és lehetséges értékei:  $x_1, x_2, \dots$ , akkor

$$\mathbb{E}(X^k) = \sum_{j=1}^{\infty} x_j^k \mathbb{P}(X = x_j).$$

Ha az  $X$  valószínűségi változó abszolút folytonos és sűrűségfüggvénye  $f$ , akkor

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx.$$

# Momentum módszer

Legyen  $X_1, \dots, X_n$  független azonos eloszlású minta.

- 1 Az eloszlás  $k$ . momentuma, ha  $\vartheta$  az ismeretlen paraméter:  $\mu_{k,\vartheta} = \mathbb{E}_{\vartheta}(X_1^k)$ .
- 2 Legyen  $\hat{\mu}_k = \frac{1}{n} \sum_{j=1}^n X_j^k$  az eloszlás  $k$ . tapasztalati momentuma.
- 3 Írjuk fel az alábbi egyenleteket a legkisebb olyan  $k$ -ig, amire az egyenletrendszer egyértelműen meghatározza  $\vartheta$ -t (**bár nincs mindig ilyen  $k$** ):

$$\mathbb{E}_{\vartheta}(X_1) = \frac{1}{n} \sum_{j=1}^n X_j;$$

$$\mathbb{E}_{\vartheta}(X_1^2) = \frac{1}{n} \sum_{j=1}^n X_j^2;$$

...

$$\mathbb{E}_{\vartheta}(X_1^k) = \frac{1}{n} \sum_{j=1}^n X_j^k.$$

- 4 A  $\vartheta$  momentum módszerrel kapott becslése az a  $\hat{\vartheta}$ , ami megoldása a fenti egyenletrendszernek. **Nem mindig létezik, nem mindig egyértelmű, nem feltétlenül hatásos.**

## Momentum módszer: Poisson- és exponenciális eloszlás

$X_1, \dots, X_n$  független **Poisson-eloszlásúak** ismeretlen  $\lambda > 0$  paraméterrel. A  $k = 1$ -hez tartozó egyenlet:

$$\mathbb{E}_\lambda(X_1) = \bar{X}.$$

Mivel a  $\lambda$  paraméterű Poisson-eloszlás várható értéke  $\lambda$ :

$$\hat{\lambda} = \bar{X}.$$

## Momentum módszer: Poisson- és exponenciális eloszlás

$X_1, \dots, X_n$  független **Poisson-eloszlásúak** ismeretlen  $\lambda > 0$  paraméterrel. A  $k = 1$ -hez tartozó egyenlet:

$$\mathbb{E}_\lambda(X_1) = \bar{X}.$$

Mivel a  $\lambda$  paraméterű Poisson-eloszlás várható értéke  $\lambda$ :

$$\hat{\lambda} = \bar{X}.$$

$X_1, \dots, X_n$  független **exponenciális** eloszlásúak ismeretlen  $\lambda > 0$  paraméterrel. A  $k = 1$ -hez tartozó egyenlet:

$$\mathbb{E}_\lambda(X_1) = \frac{1}{\lambda} = \frac{1}{n} \sum_{j=1}^n X_j = \bar{X}.$$

Ez egyértelműen oldható meg  $\lambda$ -ra:

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

## Momentum módszer: normális eloszlás

$X_1, \dots, X_n$  független  $N(m, \sigma^2)$  eloszlású minta (azaz normális eloszlású  $m$  várható értékkel és  $\sigma$  szórással).

A  $k = 1$ -hez és  $k = 2$ -höz tartozó egyenletek:

$$\mathbb{E}_{m,\sigma}(X_1) = m = \bar{X};$$

$$\mathbb{E}_{m,\sigma}(X_1^2) = \sigma^2 + m^2 = \frac{1}{n} \sum_{j=1}^n X_j^2.$$

A másodikba beírva az elsőt:  $\sigma^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 = s_n^2$  (a tapasztalati szórásnégyzet). Tehát az első két egyenlet együtt egyértelműen oldható meg, a momentum módszerrel kapott becslés:

$$\hat{m} = \bar{X}; \quad \hat{\sigma} = s_n.$$

## Az egyenletes eloszlás várható értéke és szórása

Az egyenletes eloszlás sűrűségfüggvénye:  $f(x) = \frac{1}{b-a}$ , ha  $a \leq x \leq b$ , és 0 különben.

A várható értéke:

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_a^b \frac{x}{b-a} dx = \left[ \frac{x^2}{2(b-a)} \right]_{x=a}^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.\end{aligned}$$

## Az egyenletes eloszlás várható értéke és szórása

Az egyenletes eloszlás sűrűségfüggvénye:  $f(x) = \frac{1}{b-a}$ , ha  $a \leq x \leq b$ , és 0 különben.

A várható értéke:

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_a^b \frac{x}{b-a} dx = \left[ \frac{x^2}{2(b-a)} \right]_{x=a}^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.\end{aligned}$$

A négyzetének a várható értéke:

$$\begin{aligned}\mathbb{E}(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \left[ \frac{x^3}{3} \right]_{x=a}^b \\ &= \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}.\end{aligned}$$

## Az egyenletes eloszlás várható értéke és szórása

Az egyenletes eloszlás sűrűségfüggvénye:  $f(x) = \frac{1}{b-a}$ , ha  $a \leq x \leq b$ , és 0 különben.

A várható értéke:

$$\mathbb{E}(X) = \frac{a+b}{2}.$$

A négyzetének a várható értéke:

$$\mathbb{E}(X^2) = \frac{a^2 + ab + b^2}{3}.$$

A szórásnégyzete:

$$\begin{aligned}\mathbb{D}^2(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} \\ &= \frac{a^2 - 2ab + b^2}{12} = \frac{(b-a)^2}{12}.\end{aligned}$$

## Momentum módszer: egyenletes eloszlás

Legyen  $X_1, \dots, X_n$  független minta az  $[a, b]$  intervallumon egyenletes eloszlásból. Ennek várható értéke  $(a + b)/2$ , szórása  $(b - a)/\sqrt{12}$ . Ezek alapján a  $k = 1$ -hez és  $k = 2$ -höz tartozó egyenlet:

$$\mathbb{E}_{a,b}(X_1) = \frac{a + b}{2} = \bar{X};$$
$$\mathbb{E}_{a,b}(X_1^2) = \frac{(b - a)^2}{12} + \left(\frac{a + b}{2}\right)^2 = \frac{1}{n} \sum_{j=1}^n X_j^2.$$

A másodikba beírva az elsőt:  $\frac{(b-a)^2}{12} = \frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 = s_n^2$ , amiből

$$\hat{a} = \bar{X} - \sqrt{3}s_n; \quad \hat{b} = \bar{X} + \sqrt{3}s_n.$$

Hátránya: előfordulhat, hogy ezek nem is lehetséges értékek: ha  $\hat{a}$  nagyobb a legkisebb megfigyelésnél, vagy  $\hat{b}$  kisebb a legnagyobb megfigyelésnél.

## Többváltozós lineáris regresszió (multiple linear regression)

Az  $Y$  változót fejezzük ki az  $X_1, \dots, X_p$  valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük ( $X_{i,p} \equiv b$  lehet a konstans tag):

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

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

**Például:**  $X_{i,1}$  az év,  $X_{i,2}$  a CFC-12 kibocsátás,  $Y$  a koncentráció. Ekkor a lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \varepsilon_i.$$

Vektoros formában:  $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ , ahol  $X$  az  $X_{i,j}$  megfigyelésekből készített mátrix, és  $\underline{\beta} = (a_1, a_2, \dots, a_p)^T$  az együtthatók oszlopvektora.

## Többszörös lineáris regresszió (multiple linear regression)

Az  $Y$  változót fejezzük ki az  $X_1, \dots, X_p$  valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük ( $X_{i,p} \equiv b$  lehet a konstans tag):

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

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

**Például:**  $X_{i,1}$  az év,  $X_{i,2}$  a CFC-12 kibocsátás,  $Y$  a koncentráció. Ekkor a lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \varepsilon_i.$$

Vektoros formában:  $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ , ahol  $X$  az  $X_{i,j}$  megfigyelésekből készített mátrix, és  $\underline{\beta} = (a_1, a_2, \dots, a_p)^T$  az együtthatók oszlopvektora.

Ezután az  $a_1, \dots, a_p$  együtthatók becslése (torzítatlan, és ugyanaz a legkisebb négyzetek módszerével és maximumlikelihood-módszerrel):

$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{Y}.$$

Ekkor is megfelelő próbastatisztikával  $t$ -próbával tesztelhetők az  $a_i = 0$  hipotézisek, vagyis ellenőrizhető, hogy az  $Y$  mely mennyiségektől függ szignifikánsan.

# Hipotézisvizsgálat a lineáris modellben

Többváltozós lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i, \text{ azaz } \underline{Y} = X\beta + \varepsilon.$$

Legyen  $H$  olyan  $r \times p$  méretű mátrix, aminek a rangja  $r$  (itt  $r < p$ ). Ekkor a nullhipotézis:  $H_0 : H\beta = 0$ , és  $H_1 : H\beta \neq 0$ . (Például: ha  $H$  egy sora a  $j$ . egységvektor, az  $a_j = 0$ -t jelenti.)

A valószínűséghányados próba próbastatisztikája:

$$F = \frac{(\underline{Y} - X\beta^*)^T (\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T (\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T (\underline{Y} - X\hat{\beta})},$$

ahol  $\beta^*$  a  $\beta$  becslése a  $H\beta = 0$  feltétel mellett a redukált lineáris modellben (például: bizonyos  $X$ -ek együtthatója 0, ezeket nem használhatjuk).

Ha  $H_0$  igaz, akkor  $F \cdot (n-p)/r$  eloszlása  $F$ -eloszlás  $(r, n-p)$  szabadsági fokkal. Ezért  $H_0$ -t elutasítjuk, ha  $F$  értéke nagyobb ennek az  $F$  próbának a kritikus értékénél.

# Szórásanalízis

	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 ( $\bar{X}$ )	10,8	10,1	10,5	9,2
szórás ( $s_n^*$ )	0,57	0,63	0,42	0,34

Néhány évi középhőmérséklet (forrás: Országos Meteorológiai Szolgálat), különböző évekből

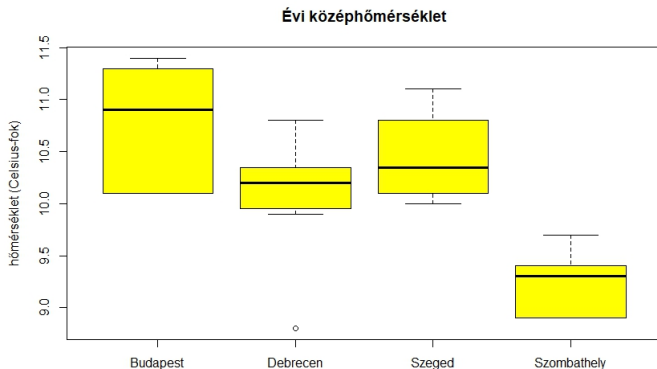
# Szórásanalízis

	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 ( $\bar{X}$ )	10,8	10,1	10,5	9,2
szórás ( $s_n^*$ )	0,57	0,63	0,42	0,34

Néhány évi középhőmérséklet (forrás: Országos Meteorológiai Szolgálat), különböző évekből

Igaz-e, hogy az egyes városokban az évi középhőmérséklet várható értéke megegyezik, vagy szignifikáns különbség látható?

# Szórásanalízis



A városok évi középhőmérséklet adatai különböző évekből. Van-e szignifikáns eltérés a várható értékek között?

# Szórásanalízis (analysis of variance, ANOVA)

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

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

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

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

$$H_1 : \mu_1 = \mu_2 = \dots = \mu_k \text{ nem teljesül.}$$

Másképpen:

$H_0$ : a faktornak nincs szignifikáns hatása

$H_1$ : a faktornak szignifikáns hatása van.

# Szórásanalízis (ANOVA)

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

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

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

$$H_1 : \mu_1 = \mu_2 = \dots = \mu_k \text{ nem teljesül.}$$

- normális eloszlások várható értékére vonatkozó próba
- a **kétmintás párosítatlan** Student-féle  $t$ -próba általánosításának is tekinthető, most nem kettő, hanem több csoport van, a szórások mindenhol megegyeznek
- a lineáris regresszió speciális esete  $\beta = (\mu_1, \mu_2, \dots, \mu_k)$ -val, ahol a magyarázó változók értéke 0 vagy 1, mert ezt  $\beta$ -val megszorozva kapjuk valamelyik  $\mu_j$ -t, és ehhez adódik hozzá a hiba.
- a nullhipotézis  $H\beta = 0$  alakú, ezért  $F$ -próbát végezhetünk.

# Szórásanalízis (analysis of variance, ANOVA)

$X_{ij}$  valószínűségi változók,  $i = 1, \dots, k$ ,  $j = 1, \dots, n_i$ . Vagyis  $k$  csoport van, és az  $i$ -ben  $n_i$  darab megfigyelés van. Jelölések:

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

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

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

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

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

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

# Szórásanalízis

	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 ( $\bar{X}_{j.}$ )	10,8	10,1	10,5	9,2	$\bar{\bar{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

Teljes szóródás = csoportokon belüli + csoportok közötti:

$$S = S_e + S_t = 5,43 + 7,15 = 12,49.$$

# Szórásanalízis

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)},$$

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

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

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

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

## Szórásanalízis

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

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)} = \frac{7,15 \cdot 20}{5,43 \cdot 3} = 8,77,$$

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

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

## Szórásanalízis

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

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)} = \frac{7,15 \cdot 20}{5,43 \cdot 3} = 8,77,$$

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

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.