

# Valószínűségszámítás

---

2020. november 25.  
Mészáros Szabolcs

Tárgyhonlap:  
[cs.bme.hu/valszam](https://cs.bme.hu/valszam)

A prezentáció anyagát és az abból készült videofelvételt a tárgy hallgatói jogosultak használni, kizárólag saját célra. A felvétel másolása, videómegosztókra való feltöltése részben vagy egészben tilos, illetve csak a tantárgyfelelős előzetes engedélyével történhet.

részben vagy egészben tilos, illetve csak a  
tantárgyfelelős előzetes engedélyével történhet.

Copyright © 2020, BME VIK

# Ismétlés: kovariancia

**Definíció:** Legyenek  $X$  és  $Y$  valószínűségi változók. Ekkor

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

**Kérdés:** Hogyan számoljuk ki  $\mathbb{E}(XY)$ -t, ha csak az  $f_{X,Y}$  együttes sűrűségfüggvényt ismerjük?

**Ötletek:**

- Szorzat sűrűségfüggvényével? De azt hogy kapjuk meg?
- Konvolúció-szerű képlet szorzatra? De nem feltétlenül függetlenek.

# Többdim. transzformált várható értéke

**Állítás:** Legyen  $\underline{X} = (X_1, \dots, X_n)$  folytonos valószínűségi vektorváltozó, és legyen  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  olyan, amire  $g(\underline{X})$  létezik. Ekkor

$$\mathbb{E}(g(\underline{X})) = \int_{\mathbb{R}^n} g(\underline{x}) f_{\underline{X}}(\underline{x}) d\underline{x}$$

Ha  $g$  folytonos és nemnegatív, akkor  $\mathbb{E}(g(\underline{X}))$  létezik.

## Többdim. transzformált várható értéke

Speciális eset:  $g : \mathbb{R}^2 \rightarrow \mathbb{R} \quad g(x, y) = x \cdot y$

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y) \, dx dy$$

**Példa:**  $X$  éves csapadékmennyiség (/1000 mm),

$Y$  idén eladott esernyők száma (/1000 db).

Mennyi a kovarianciájuk, ha az együttes sűrűségfüggvényük:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{5}(4 - 2x^2 + xy - y^2) & \text{ha } 0 < x < 1 \text{ és } 0 < y < 2, \\ 0 & \text{egyébként.} \end{cases}$$

## Többdim. transzformált várható értéke, példa

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y) \, dx dy \\ &= \int_0^2 \int_0^1 xy \cdot \frac{1}{5}(4 - 2x^2 + xy - y^2) \, dx dy \\ &= \frac{1}{5} \int_0^2 \int_0^1 (4xy - 2x^3y + x^2y^2 - xy^3) \, dx dy \\ &= \frac{1}{5} \int_0^2 \left[ 2x^2y - \frac{1}{2}x^4y + \frac{1}{3}x^3y^2 - \frac{1}{2}x^2y^3 \right]_{x=0}^1 dy\end{aligned}$$

## Többdim. transzformált várható értéke, példa

$$= \frac{1}{5} \int_0^2 \left( \frac{3}{2}y + \frac{1}{3}y^2 - \frac{1}{2}y^3 \right) dy = \frac{1}{5} \left[ \frac{3}{4}y^2 + \frac{1}{9}y^3 - \frac{1}{8}y^4 \right]_0^2 = \frac{7}{15}$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot f_{X,Y}(x, y) dx dy \quad \mathbb{E}(Y) = \frac{4}{5}$$

$g(x, y) = x$

$$= \int_0^2 \int_0^1 x \cdot \frac{1}{5} (4 - 2x^2 + xy - y^2) dx dy = \frac{7}{15}$$

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{17}{45} - \frac{7}{15} \cdot \frac{4}{5} = \frac{1}{225}$$

# Kovariancia tulajdonságai

**Lemma:** Legyen  $(X, Y, Z)$  valószínűségi vektorváltozó. Ha az alábbi mennyiségek léteznek, akkor a következők teljesülnek:

1. Ha  $c \in \mathbb{R}$ , akkor  $\mathbb{D}(X + c) = \mathbb{D}(X)$  és  $\mathbb{D}(cX) = |c|\mathbb{D}(X)$
2.  $\mathbb{D}^2(X + Y) = \mathbb{D}^2(X) + \mathbb{D}^2(Y) + 2\text{cov}(X, Y)$
3.  $\mathbb{D}^2(X) = 0$  pontosan akkor, ha  $\mathbb{P}(X = c) = 1$  valamilyen  $c$ -re.
4. Ha  $X$  és  $Y$  függetlenek, akkor  $\text{cov}(X, Y) = 0$ .
5. Ha  $b, c \in \mathbb{R}$ , akkor  $\text{cov}(X, bY + cZ)$   
 $= b \cdot \text{cov}(X, Y) + c \cdot \text{cov}(X, Z)$

# Kovariancia tulajdonságai, példa

**Lemma:** Ha  $X$  és  $Y$  független valószínűségi változók,  $g$  és  $h$  folytonos  $\mathbb{R} \rightarrow \mathbb{R}$  függvények, akkor  $g(X)$  és  $h(Y)$  is függetlenek.

**Példa:** A fenti függetlenségi feltétel esetén

$$\begin{aligned} \text{cov} \left( \frac{2}{X} + Y^2, \frac{2}{Y} - X^2 \right) &= \text{cov} \left( \frac{2}{X}, \frac{2}{Y} \right) \stackrel{=0}{=} - \text{cov} \left( \frac{2}{X}, X^2 \right) \\ &\quad + \text{cov} \left( Y^2, \frac{2}{Y} \right) - \text{cov} \left( Y^2, X^2 \right) \stackrel{=0}{=} \end{aligned}$$

# Kovariancia-mátrix

**Definíció:** Az  $\underline{X} = (X_1, \dots, X_n)$  val. vektorváltozó *kovarianciamátrixa*:

$$\text{COV}(\underline{X}) = \begin{pmatrix} \text{COV}(X_1, X_1) & \text{COV}(X_1, X_2) & \dots & \text{COV}(X_1, X_n) \\ \text{COV}(X_2, X_1) & \text{COV}(X_2, X_2) & & \vdots \\ \vdots & & \ddots & \\ \text{COV}(X_n, X_1) & \dots & & \text{COV}(X_n, X_n) \end{pmatrix}$$

azaz  $\text{COV}(\underline{X})_{i,j} = \text{COV}(X_i, X_j)$

Megj.:  $\text{COV}(X_i, X_i) = \mathbb{D}^2(X_i)$

# Kovariancia-mátrix, példa

**Példa:** az előző.

$$\text{cov}((X, Y)) = \begin{pmatrix} ? & \frac{1}{225} \\ \frac{1}{225} & ? \end{pmatrix}$$

$$\mathbb{D}^2(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 =$$

$$= \int_0^2 \int_0^1 x^2 \cdot \frac{1}{5} (4 - 2x^2 + xy - y^2) dx dy$$

$$- \left( \int_0^2 \int_0^1 x \cdot \frac{1}{5} (4 - 2x^2 + xy - y^2) dx dy \right)^2 = \frac{7}{90}$$

$$\mathbb{D}^2(Y) = \frac{58}{225}$$

# Kovariancia-mátrix tulajdonságai

**Állítás:** Legyen  $\underline{X} = (X_1, \dots, X_n)$  valószínűségi vektorváltozó. Ekkor

1.  $\text{COV}(\underline{X})$  szimmetrikus mátrix,
2.  $\text{COV}(\underline{X})$  pozitív szemidefinit mátrix

(azaz  $\underline{a}^T \text{COV}(\underline{X}) \underline{a} \geq 0, \quad \forall \underline{a} \in \mathbb{R}^n$ ).

**Példa:** az  $\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$  szimmetrikus mátrix, de nem pozitív szemidefinit, mert

$$\begin{bmatrix} -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = -1$$

# Kovariancia-mátrix tulajdonságai, biz.

Biz.: Szimmetrikus, azaz  $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$

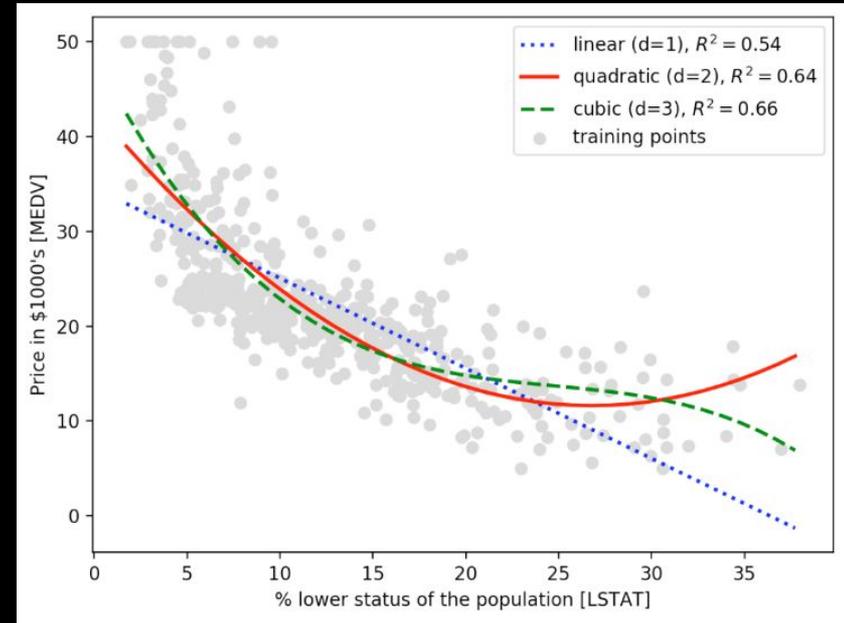
Pozitív szemidefinit: Legyen  $Z = \sum_{i=1}^n a_i X_i$

$$\begin{aligned} \mathbb{D}^2(Z) &= \mathbb{D}^2\left(\sum_{i=1}^n a_i X_i\right) = \text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n a_j X_j\right) \\ &\stackrel{\geq 0}{=} \sum_{i=1}^n \sum_{j=1}^n a_i \text{cov}(X_i, X_j) a_j = \underline{a}^T \text{cov}(\underline{X}) \underline{a} \end{aligned}$$

# Lineáris regresszió, fogalmak

## Regresszió változatai:

- lineáris regresszió
  - egyszerű lineáris regresszió, avagy legkisebb négyzetek módszere
  - regularizált lineáris regresszió (ridge, lasso, ...)
  - ...
- logisztikus regresszió
- lokális regresszió
- ...



Forrás:

<https://charleshsliao.wordpress.com/2017/06/16/ransac-and-nonlinear-regression-in-python/>

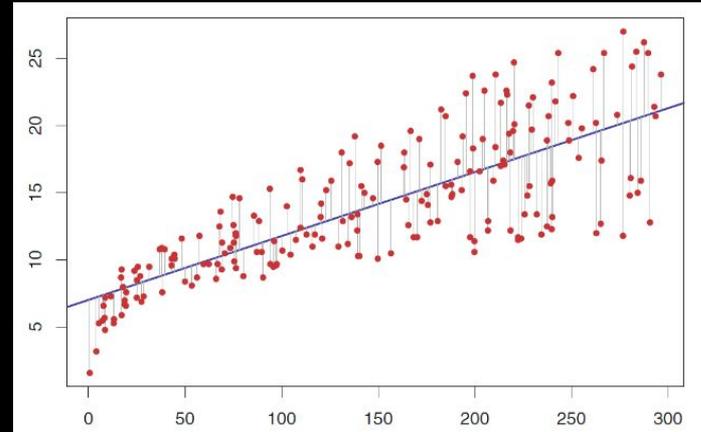
# Lineáris regresszió, definíció

**Definíció:** Legyenek  $X$  és  $Y$  val. változók. Ekkor az  $Y$ -nak az  $X$ -re vett *lineáris regresszióján* azt a  $\beta X + \alpha$  val. változót értjük, amire  $\alpha, \beta \in \mathbb{R}$ , és az

$$\mathbb{E}\left(\left(Y - (\beta X + \alpha)\right)^2\right)$$

mennyiség minimális.

**Értelmezés:** Megpróbáljuk megtippelni  $Y$ -t az  $X$ -nek egy lineáris függvényét használva, a lehető legkisebb átlagos négyzetes hibával.



# Lineáris regresszió, állítás

**Állítás:** Legyenek  $X$  és  $Y$  olyan valószínűségi változók, amire  $\text{cov}(X, Y)$ ,  $\mathbb{D}^2(X)$ ,  $\mathbb{D}^2(Y)$  véges, és  $\mathbb{D}^2(X) \neq 0$ . Ekkor az  $Y$ -nak az  $X$ -re vett *lineáris regressziós együtthatói:*

$$\beta = \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)} \quad \alpha = \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)}\mathbb{E}(X)$$

# Lineáris regresszió, példa

**Példa:** az előző, csapadékmennyiséges.

$$\beta = \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)} = \frac{1/225}{7/90} = \frac{2}{35}$$

$$\alpha = \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)}\mathbb{E}(X) = \frac{4}{5} - \frac{2}{35} \frac{7}{15} = \frac{58}{75}$$

Tehát a lineáris regresszió:  $\frac{2}{35}X + \frac{58}{75}$

# Lineáris regresszió, alternatív felírás

**Definíció:** Az  $Y$  val. változó  $X$ -re vett regressziós egyenese az

$$\{(x, y) \in \mathbb{R}^2 \mid y = \beta x + \alpha\}$$

egyenes a síkon, ahol  $\alpha$  és  $\beta$  a lineáris regressziós együtthatók.

**Kérdés:** Hogy a szöszbe jegyzem meg az együtthatók formuláit?

**Lehetséges válasz:** Ha  $Y$ -t próbálja közelíteni a  $\beta X + \alpha$ , akkor logikus lenne, ha az alábbiak teljesülnének:

$$\mathbb{E}(Y) = \mathbb{E}(\beta X + \alpha) \quad \text{cov}(X, Y) = \text{cov}(X, \beta X + \alpha)$$

Ebből a két egyenletből átrendezéssel adódik a korábbi két egyenlet.

# Lineáris regresszió, korrelációval

**Definíció:** Legyenek  $X$  és  $Y$  valószínűségi változók. Ekkor

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\mathbb{D}(X)\mathbb{D}(Y)}$$

ami egy -1 és +1 közti szám (ha létezik).

**Kérdés:** felírható a lineáris regresszió korrelációval (egyszerűen)?

**Válasz:** A regressziós együtthatók éppen azok, amikre teljesül, hogy

$$\frac{(\beta X + \alpha) - \mathbb{E}(Y)}{\mathbb{D}(Y)} = \frac{X - \mathbb{E}(X)}{\mathbb{D}(X)} \cdot \text{corr}(X, Y)$$

# Lineáris regresszió, bizonyítás

Biz.: Amit minimalizálnunk kell:

$$\begin{aligned}h(\alpha, \beta) &= \mathbb{E}\left(\left(Y - (\beta X + \alpha)\right)^2\right) \\&= \mathbb{E}\left(Y^2 + \beta^2 X^2 + \alpha^2 - 2\beta XY - 2\alpha Y + 2\alpha\beta X\right) \\&= \mathbb{E}(Y^2) + \beta^2 \mathbb{E}(X^2) + \alpha^2 - 2\beta \mathbb{E}(XY) - 2\alpha \mathbb{E}(Y) + 2\alpha\beta \mathbb{E}(X)\end{aligned}$$

Deriválással kereshető a minimumhelye:

$$\alpha \text{ szerint: } 2\alpha - 2\mathbb{E}(Y) + 2\beta \mathbb{E}(X)$$

$$\beta \text{ szerint: } 2\beta \mathbb{E}(X^2) - 2\mathbb{E}(XY) + 2\alpha \mathbb{E}(X)$$

# Lineáris regresszió, bizonyítás

Biz. (folyt.): A parciális deriváltak pontosan akkor nullák, ha

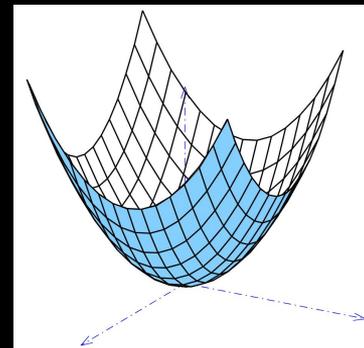
$$\alpha + \beta \mathbb{E}(X) = \mathbb{E}(Y)$$

$$\alpha \mathbb{E}(X) + \beta \mathbb{E}(X^2) = \mathbb{E}(XY)$$

Ennek a megoldása:

$$\alpha = \mathbb{E}(Y) - \beta \mathbb{E}(X) = \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)} \mathbb{E}(X)$$

$$\beta = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2} = \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)}$$



# Lineáris regresszió hibája

**Állítás:** Legyen az  $Y$  val. változó  $X$ -re vett lineáris regressziója  $\beta X + \alpha$ .

Ekkor

$$\mathbb{D}^2\left(Y - (\beta X + \alpha)\right) = \mathbb{D}^2(Y) - \frac{\text{cov}(X, Y)^2}{\mathbb{D}^2(X)}$$

**Megj.:** Korrelációval felírva

$$\mathbb{D}^2\left(Y - (\beta X + \alpha)\right) = \mathbb{D}^2(Y) \cdot (1 - \text{corr}(X, Y)^2)$$

# Lineáris regresszió hibája, példa

**Példa:** az előző.

$$\begin{aligned}\mathbb{D}^2\left(Y - (\beta X + \alpha)\right) &= \mathbb{D}^2(Y) - \frac{\text{cov}(X, Y)^2}{\mathbb{D}^2(X)} \\ &= \frac{58}{225} - \frac{(1/225)^2}{7/90} \approx 0,2575\end{aligned}$$

## Lineáris regresszió hibája, biz.

$$\begin{aligned}\text{Biz.: } \mathbb{D}^2(Y - (\beta X + \alpha)) &= \\ &= \mathbb{D}^2(Y - \beta X) \\ &= \mathbb{D}^2(Y) + \beta^2 \mathbb{D}^2(X) - 2\text{cov}(Y, \beta X) \\ &= \mathbb{D}^2(Y) + \frac{\text{cov}(X, Y)^2}{(\mathbb{D}^2(X))^2} \mathbb{D}^2(X) - 2 \frac{\text{cov}(X, Y)}{\mathbb{D}^2(X)} \text{cov}(Y, X) \\ &= \mathbb{D}^2(Y) - \frac{\text{cov}(X, Y)^2}{\mathbb{D}^2(X)}\end{aligned}$$

# További olvasnivaló

- Devore, Berk - Modern mathematical statistics with applications, Ch. 12.1.
- James, Witten, Hastie, Tibshirani - An introduction to statistical learning, Ch. 3.
- Simpson's paradox
- Anscombe's quartet

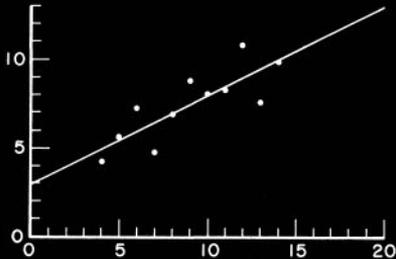


Figure 1

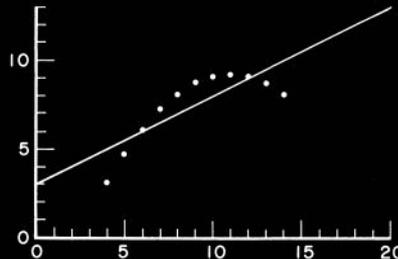
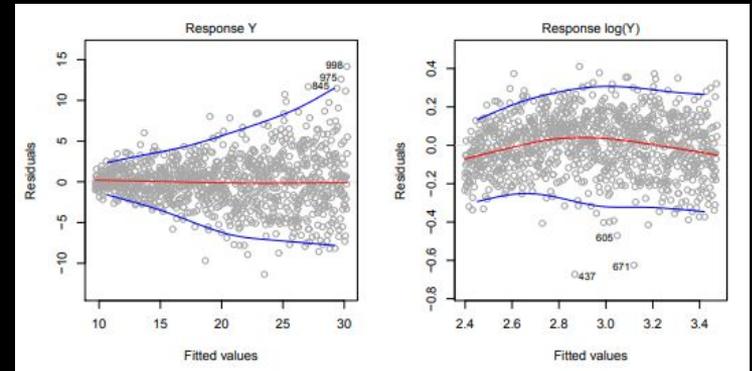
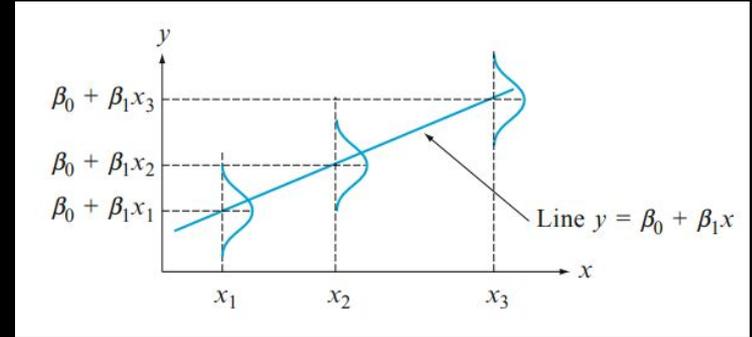


Figure 2



Köszönöm a figyelmet!

---