

A kovariancia (7. előadás)

Definíció (Kovariancia)

Legyenek X és Y olyan valószínűségi változók, melyeknek szórása létezik. Ekkor az X és Y kovarianciája:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))].$$

Legyenek X, Y, Z, X_1, \dots, X_n olyan valószínűségi változók, melyek szórása létezik. Ekkor a következők teljesülnek.

- **A kovariancia kiszámítása:**

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

- Szimmetria. $\text{cov}(X, Y) = \text{cov}(Y, X)$.
- Kapcsolat a szórásnégyzettel. $\text{cov}(X, X) = D^2(X)$.

A kovariancia tulajdonságai

- Konstanssal való kovariancia. $\text{cov}(X, c) = 0$, ha $c \in \mathbb{R}$.
- **Linearitás.** Egyrészt

$$\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z),$$

másrészt tetszőleges $c \in \mathbb{R}$ számra

$$\text{cov}(cX, Y) = c \cdot \text{cov}(X, Y).$$

- **Függetlenséggel való kapcsolat.** Ha az X és Y valószínűségi változók függetlenek, akkor $\text{cov}(X, Y) = 0$ (fordítva nem).
- **Összeg szórásnégyzete.** $D^2(X + Y) = D^2(X) + D^2(Y) + 2\text{cov}(X, Y)$.
Továbbá

$$D^2\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D^2(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j).$$

- Különbség szórásnégyzete. $D^2(X - Y) = D^2(X) + D^2(Y) - 2\text{cov}(X, Y)$.

Kovariancia: példa.

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Kovariancia: példa.

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a kovarianciája?

Kovariancia: példa.

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

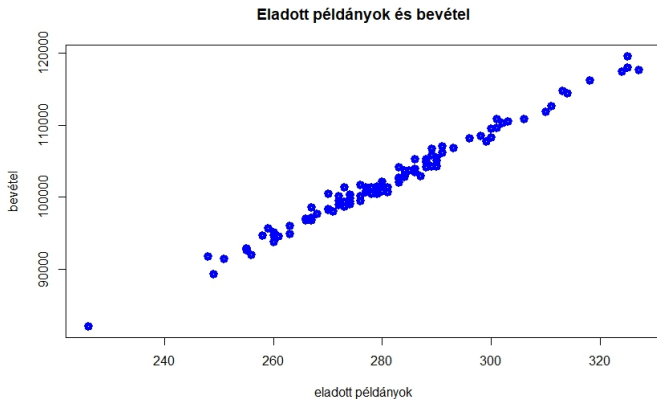
$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a kovarianciája?

$$\begin{aligned}\operatorname{cov}(X + Y, 300X + 400Y) &= \operatorname{cov}(X, 300X) + \operatorname{cov}(X, 400Y) + \operatorname{cov}(Y, 300X) + \\ &\quad + \operatorname{cov}(Y, 400Y) = 300 \cdot \operatorname{cov}(X, X) + 400 \cdot \operatorname{cov}(Y, Y) = \\ &= 300D^2(X) + 400D^2(Y) = \\ &= 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

ahol felhasználtuk a linearitást, azt, hogy függetlenség esetén 0 a kovariancia, illetve a Poisson-eloszlás tulajdonságait.

Kovariancia: példa



A bevétel ($300X + 400Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ megfigyelésből. Kovariancia: 102000.

Korrelátlanság

Definíció (Korrelátlanság)

Ha az X, Y valószínűségi változók kovarianciája 0, akkor azt mondjuk, hogy X és Y **korrelátlanak**.

Korábban láttuk, hogy ha az X és Y valószínűségi változók függetlenek és szórásuk létezik, akkor X és Y korrelátlanak.

A korrelátlanságból nem következik a függetlenség. Például: legyen X és Y két szabályos kockadobás, ezek függetlenek. Legyen továbbá $U = X + Y$, $V = X - Y$. Ekkor, bár $X + Y$ és $X - Y$ nem függetlenek:

$$\text{cov}(U, V) = \text{cov}(X + Y, X - Y) = D^2(X) - \text{cov}(X, Y) + \text{cov}(X, Y) - D^2(X) = 0.$$

Ugyanakkor U és V nem függetlenek, például

$$0 = \mathbb{P}(U = 11, V = 0) \neq \mathbb{P}(U = 11) \cdot \mathbb{P}(V = 0) = \frac{2}{36} \cdot \frac{1}{6}.$$

Korrelátlanság: példa



A dobott számok különbségének ($X - Y$) és a dobott számok összegének ($X + Y$) együttes előfordulása 100 megfigyelésből. Kovariancia: 0, de $X + Y$ és $X - Y$ nem függetlenek.

Korrelációs együttható

Definíció

Legyenek X és Y olyan valószínűségi változók, melyek szórásnégyzete létezik. Ekkor X és Y **korrelációs együtthatója**:

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{ha } D(X) > 0, D(Y) > 0; \\ 0, & \text{ha } D(X) = 0 \text{ vagy } D(Y) = 0. \end{cases}$$

Korrelációs együttható

Definíció

Legyenek X és Y olyan valószínűségi változók, melyek szórásnégyzete létezik. Ekkor X és Y **korrelációs együtthatója**:

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{ha } D(X) > 0, D(Y) > 0; \\ 0, & \text{ha } D(X) = 0 \text{ vagy } D(Y) = 0. \end{cases}$$

Állítás

Legyenek X és Y olyan valószínűségi változók, melyek szórása létezik.

(i) Ekkor teljesül, hogy

$$|R(X, Y)| \leq 1.$$

(ii) Legyen $a > 0$ valós szám, b tetszőleges valós szám. Ekkor

$$R(X, aX + b) = 1 \text{ és } R(X, -aX + b) = -1.$$

(iii) Tegyük fel, hogy $|R(X, Y)| = 1$. Ekkor léteznek olyan a és b valós számok, hogy az $Y = aX + b$ egyenlet 1 valószínűséggel teljesül.

Korrelációs együttható: példa

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Korrelációs együttható: példa

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Továbbá a szórás tulajdonságai alapján

$$D(X + 3) = D(X) = \sqrt{2}; \quad D(2X) = 2D(X) = 2\sqrt{2}.$$

Tehát

$$R(X + 3, 2X) = \frac{\operatorname{cov}(X + 3, 2X)}{D(X + 3)D(2X)} = \frac{2D^2(X)}{D(X) \cdot 2D(X)} = 1.$$

Korrelációs együttható: példa

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. A linearitás, a szórásnégyzettel való kapcsolat és a konstanssal való kovariancia alapján:

$$\begin{aligned}\operatorname{cov}(X + 3, 2 \cdot X) &= 2\operatorname{cov}(X + 3, X) = 2\operatorname{cov}(X, X) + 2\operatorname{cov}(3, X) = \\ &= 2D^2(X) = 2 \cdot 2 = 4.\end{aligned}$$

Továbbá a szórás tulajdonságai alapján

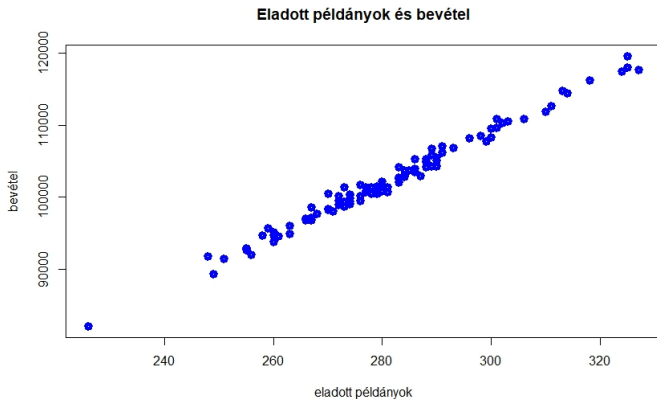
$$D(X + 3) = D(X) = \sqrt{2}; \quad D(2X) = 2D(X) = 2\sqrt{2}.$$

Tehát

$$R(X + 3, 2X) = \frac{\operatorname{cov}(X + 3, 2X)}{D(X + 3)D(2X)} = \frac{2D^2(X)}{D(X) \cdot 2D(X)} = 1.$$

Általában is láttuk, hogy ha $V = aU + b$ alakú, akkor a korrelációs együttható $a/|a|$. Most $V = 2X = 2(U - 3) = 2U - 6$, és a korrelációs együttható értéke 1.

Korrelációs együttható: példa



A bevétel ($300X + 400Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ megfigyelésből. Kovariancia: 102000, korrelációs együttható: 0,9915.

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

$$\begin{aligned}\operatorname{cov}(X + Y, 300X + 400Y) &= \operatorname{cov}(X, 300X) + \operatorname{cov}(X, 400Y) + \operatorname{cov}(Y, 300X) + \\ &\quad + \operatorname{cov}(Y, 400Y) = 300 \cdot \operatorname{cov}(X, X) + 400 \cdot \operatorname{cov}(Y, Y) \\ &= 300D^2(X) + 400D^2(Y) = \\ &= 300 \cdot 100 + 400 \cdot 180 = 102000,\end{aligned}$$

ahol felhasználtuk a linearitást, azt, hogy függetlenség esetén 0 a kovariancia, illetve a Poisson-eloszlás tulajdonságait.

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

$$\text{cov}(X + Y, 300X + 400Y) = 300 \cdot 100 + 400 \cdot 180 = 102000;$$

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148,17; \end{aligned}$$

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \frac{102000}{16,73 \cdot 6148,17} \\ &= 0,9915. \end{aligned}$$

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é 400. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

$$\text{cov}(X + Y, 300X + 400Y) = 300 \cdot 100 + 400 \cdot 180 = 102000;$$

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148,17; \end{aligned}$$

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \frac{102000}{16,73 \cdot 6148,17} \\ &= 0,9915. \end{aligned}$$

A korrelációs együttható értéke majdnem 1, azaz erős pozitív korreláció van az eladott példányok száma és a bevétel között.

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é **4000**. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

$$\text{cov}(X + Y, 300X + 4000Y) = 300 \cdot 100 + 4000 \cdot 180 = 750000;$$

$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

$$\begin{aligned} D(300X + 4000Y) &= \sqrt{300^2 D^2(X) + 4000^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 4000^2 \cdot 180} = 53749,42; \end{aligned}$$

$$\begin{aligned} R(X + Y, 300X + 4000Y) &= \frac{\text{cov}(X + Y, 300X + 4000Y)}{D(X + Y)D(300X + 4000Y)} = \frac{750000}{16,73 \cdot 53749,42} \\ &= 0,083. \end{aligned}$$

Korrelációs együttható: példa.

Példa. Egy üzletben az A és B újság forgalmát figyelik. Legyen az A újságból egy nap alatt eladott példányok száma X , a B újságból eladott példányok száma Y . Tegyük fel, hogy X és Y függetlenek, Poisson-eloszlásúak, X paramétere 100, Y -é 180. Az A újság ára 300 forint, a B -é **4000**. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

$$\text{cov}(X + Y, 300X + 4000Y) = 300 \cdot 100 + 4000 \cdot 180 = 750000;$$

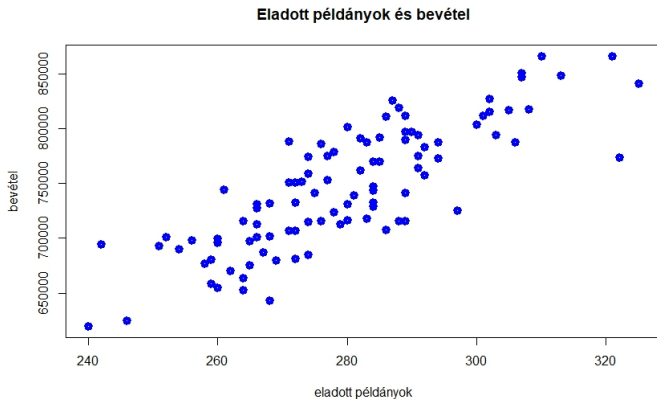
$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

$$\begin{aligned} D(300X + 4000Y) &= \sqrt{300^2 D^2(X) + 4000^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 4000^2 \cdot 180} = 53749,42; \end{aligned}$$

$$\begin{aligned} R(X + Y, 300X + 4000Y) &= \frac{\text{cov}(X + Y, 300X + 4000Y)}{D(X + Y)D(300X + 4000Y)} = \frac{750000}{16,73 \cdot 53749,42} \\ &= 0,083. \end{aligned}$$

A korrelációs együttható értéke majdnem 0, azaz nincs jelentős korreláció az eladott példányok száma és a bevétel között, ha az újságok ára nagyon eltérő.

Korrelációs együttható: példa



A bevétel ($300X + 4000Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ megfigyelésből. Kovariancia: 750000, korrelációs együttható: 0,083.

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hómennyiség:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hőmennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hőmennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- vitorlázással töltött idő és egészség:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hőmennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- vitorlázással töltött idő és egészség:
ha van is pozitív korreláció, **nem biztos, hogy van ok-okozati összefüggés**, a vitorlázás összefügg az anyagi helyzettel, ami az egészséggel, de csak a vitorlázástól nem biztos, hogy egészséges lesz valaki
- USA által tudományra és technológiára költött pénz és öngyilkosságok:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hőmennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- vitorlázással töltött idő és egészség:
ha van is pozitív korreláció, **nem biztos, hogy van ok-okozati összefüggés**, a vitorlázás összefügg az anyagi helyzettel, ami az egészséggel, de csak a vitorlázástól nem biztos, hogy egészséges lesz valaki
- USA által tudományra és technológiára költött pénz és öngyilkosságok: van pozitív korreláció ($R = 0,9979$), de **feltehetően nincs ok-okozati összefüggés** (forrás és további példák: <http://tylervigen.com/spurious-correlations>)

Házi feladat november 13-ig

Tegyük fel, hogy három távoli város mindegyikében a többitől függetlenül minden nap p valószínűséggel van földrengés. Legyen X, Y, Z az, hogy az egyes városokban mostantól hányadik napon lesz először földrengés. Számítsuk ki az alábbi mennyiségeket, és ábrázoljuk őket p függvényében (például R-ben):

- 1 $\text{cov}(X + Y, Z - X)$
- 2 $R(X + Y, Z - X)$
- 3 $R(2X + Y, 2X + Z)$

Házi feladat október 16-ig: megoldás

Legyenek X és Y szabályos kockával dobott számok. Határozzuk meg az alábbi mennyiségeket:

$$\textcircled{1} \mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = \frac{1}{6}(1 + 2 + 3 + \dots + 6) + \frac{1}{6}(1 + 2 + 3 + \dots + 6) = 7.$$

$$\textcircled{2} \mathbb{E}(X^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + \dots + 6^2) = \frac{91}{6} = 15,17$$

$$\textcircled{3} \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 15,17 - 3,5^2 = 2,92$$

$$\textcircled{4} \mathbb{E}((X + Y)^2) - \mathbb{E}(X + Y)^2 = \mathbb{E}(X^2) + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y^2) - 7^2 = 5,84.$$

Itt valójában az utolsó mennyiség az X és Y függetlensége miatt

$$D^2(X + Y) = D^2(X) + D^2(Y) = 2 \cdot 2,92 = 5,84.$$