5.3 Kovarijanca i korelacija

Def. 5.16 | Neka su X i Y diskretne slučajne varijable t.d. $\mathbb{E}[X^2] < \infty$ i $\mathbb{E}[Y^2] < \infty$.

Kovarijanca od X i Y definira se kao

$$\mathrm{Cov}(X,Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] \quad (5.12)$$ ■

↑ dobro definirano zbog (5.20) dolje

Prop. 5.17 | (Svojstva kovarijance)

(i) $\boxed{\mathrm{Cov}(X,Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\,\mathbb{E}[Y]}$;

(ii) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$;

(iii) $\mathrm{Cov}(aX + b, cY + d) = a \cdot c\, \mathrm{Cov}(X,Y), \quad \forall a,b,c,d \in \mathbb{R}$

(iv) Vrijedi

$$\boxed{\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)} \quad (5.13)$$

Općenito,

$$\boxed{\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \cdot \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j)} \quad (5.14)$$

Dokaz | (DZ) (Za (iii) i (iv) vidi i Tm. 5.16 u Sandrić, Vondraček.) ■

Nap. | $\mathrm{Cov}(X,Y)$ mjeri "linearnu" zavisnost između X i Y. ■

Pr. 5.18 | Bacamo novčić 2 puta, te neka je X ukupan broj P, a Y ukupan broj G. → očito, $\boxed{X, Y \sim B(2, \tfrac{1}{2})}$ te

$X \backslash Y$	0	1	2
0	0	0	$\frac{1}{4}$
1	0	$\frac{1}{2}$	0
2	$\frac{1}{4}$	0	0

Dakle, $\mathbb{E}[X] = \mathbb{E}[Y] = 2 \cdot \frac{1}{2} = 1$,

$\mathbb{E}[XY] \underset{DZ}{=} \frac{1}{2} \cdot 1 \cdot 1 = \frac{1}{2}$

$\rightarrow \mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{2} - 1^2 = \boxed{-\frac{1}{2}} < 0$

Uočimo,

$$Y = 2 - X$$ ← potpuna "linearna" zavisnost

■

Tm. 5.19 | Neka su X i Y dvije nezavisne diskretne slučajne varijable t.d. $\mathbb{E}[|X|] < \infty$ i $\mathbb{E}[|Y|] < \infty$. Tada je $\mathbb{E}[|XY|] < \infty$ te

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]. \qquad (5.15)$$

Dokaz. | Neka je $X \in \{a_i : i \in I\}$, $Y \in \{b_j : j \in J\}$.

$$\Rightarrow \mathbb{E}[\underbrace{|XY|}_{=g(X,Y)}] = \sum_{\substack{i \in I,\\ j \in J}} \underbrace{|a_i b_j|}_{=|a_i|\cdot|b_j|} \mathbb{P}(X=a_i, Y=b_j) \overset{\text{nezavisnost!}}{=} \sum_{\substack{i \in I,\\ j \in J}} |a_i|\,|b_j|\,\mathbb{P}(X=a_i)\,\mathbb{P}(Y=b_j)$$

$$= \left(\sum_{i \in I} |a_i|\,\mathbb{P}(X=a_i)\right) \cdot \left(\sum_{j \in J} |b_j|\,\mathbb{P}(Y=b_j)\right)$$

$$= \underbrace{\mathbb{E}[|X|]}_{<\infty} \cdot \underbrace{\mathbb{E}[|Y|]}_{<\infty} < \infty$$

$\Rightarrow$ postoji $\mathbb{E}[XY]$ te se analogno pokaže (5.15). ∎

Napomena | (5.15) vrijedi uvijek ako su $X, Y \geq 0$. ∎

Uočimo, ako $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$ te su X i Y nezavisne,

$$\boxed{\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \underset{(5.15)}{=} 0}.$$

Def. 5.20 | Ako je $\mathrm{Cov}(X,Y) = 0$, kažemo da su X i Y nekorelirane. ∎

Pr. 5.22 | $X \sim \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$, $Y := X^2$

Imamo,

$$\mathbb{E}[X] = 0, \quad \mathbb{E}[XY] = \mathbb{E}[X^3] \overset{X^3 = X}{=} \mathbb{E}[X] = 0$$

$\Rightarrow \mathrm{Cov}(X,Y) = 0!$,

iako su X i Y (potpuno) zavisne.

[Zavisnost nije "linearna"]

Dakle, općenito

$$\boxed{\text{nekorelinenost} \not\Rightarrow \text{nezavisnost}}$$

Uočimo, ako su $X_1, \dots, X_n$ <u>nezavisne</u>, i $\mathbb{E}[X_i^2] < \infty, \forall i = 1, \dots, n$, (5.14) i (5.15) povlače

$$\boxed{Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i)}.$$

[Dovoljna je samo u parovima nekorelinenost.]

Pr. 5.22 | Bacamo novčić na kojem je vjerojatnost za P jednaka $p \in \langle 0,1]$. Neka je $n \in \mathbb{N}$ i

$$T := \text{broj bacanja potreban do ukupno } n \ P.$$

Odredite $\mathbb{E}[T]$ i $Var(T)$.

Rj. [(10) distribucija od $T_n \longrightarrow \mathbb{E}[T_n] \longrightarrow Var(T_n)$
(nećemo tako!)]

Neka je

$X_1 :=$ broj bacanja potreban od prvog P,

$X_2 :=$ dodatni broj bacanja do drugog P

$\vdots$

$X_i :=$ ——— " ——— i-tog P, $i \geq 2$.

Vrijedi: $\boxed{T = \sum_{i=1}^{n} X_i}$, te

- $X_1, \dots, X_n$ su nezavisne $\qquad (5.16)$
- $X_i \sim G(p)$, $\forall i = 1, \dots n$

$$\Rightarrow \quad \mathbb{E}[T] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] \underset{X_i \geq 0,\, \forall i}{=} \sum_{i=1}^{n} \mathbb{E}[X_i] \underset{\mathbb{E}[X_i] = \mathbb{E}[X_1],\, \forall i}{=} n \cdot \mathbb{E}[X_1]$$

$$= n \cdot \frac{1}{p} = \boxed{\frac{n}{p}}$$

$$\mathrm{Var}(T) = \mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) \underset{\text{nezavisnost!}}{=} \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

$$= n \cdot \mathrm{Var}(X_1) = n \cdot \frac{1-p}{p^2} = \boxed{\frac{n(1-p)}{p^2}} \qquad \blacksquare$$

$\underline{\text{Dok } 5.12.3}$ Pokažimo formulu da (5.16) vrijedi.

Neka je Y_i = rezultat i-tog bacanja $\in \{\underset{G}{0}, \underset{P}{1}\}$

$\Rightarrow$ $Y_1, Y_2, \dots$ su nezavisne te

$$Y_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \ \forall i \geq 1.$$

$\forall k_1, \dots, k_n \in \mathbb{N}$,

$$\mathbb{P}(X_1 = k_1, \dots, X_n = k_n) =$$

$$= \mathbb{P}(Y_1 = 0, \dots, Y_{k_1 - 1} = 0, \boxed{Y_{k_1} = 1}, \dots, Y_{k_{n-1}+1} = 0, \dots, Y_{k_n - 1} = 0, \boxed{Y_{k_n} = 1})$$

$$= \big[\text{nezavisnost} + \mathbb{P}(Y_i = 0) = 1-p,\ \mathbb{P}(Y_i = 1) = p,\ \forall i\big]$$

$$= (1-p)^{k_1 - 1} p \cdot \dots \cdot (1-p)^{k_n - 1} p. \qquad (*)$$

Iz ovoga slijedi da za sve $i=1,\dots,n$ vrijedi

$$\mathbb{P}(X_i = k_i) = (1-p)^{k_i-1}\,p, \quad \forall k_i \in \mathbb{N}, \qquad (**)$$

tj. $X_i \sim G(p)$, $\forall i$, a uvrštavanjem u $(*)$ dobijemo

$$\mathbb{P}(X_1=k_1,\dots,X_n=k_n) = \prod_{i=1}^{n} \mathbb{P}(X_i=k_i), \quad \forall k_1,\dots,k_n \in \mathbb{N},$$

tj. $X_1,\dots,X_n$ su nezavisne.

Preostaje pokazati $(**)$:

$$\mathbb{P}(X_1=k_1) = \sum_{k_2,\dots,k_n\in\mathbb{N}} \mathbb{P}(X_1=k_1, X_2=k_2,\dots,X_n=k_n)$$

$$\overset{(*)}{=} \sum_{-\,''\,-} (1-p)^{k_1-1}p \cdots (1-p)^{k_n-1}p$$

$$= (1-p)^{k_1-1}p \prod_{i=2}^{n} \underbrace{\Big(\sum_{k_i\in\mathbb{N}} (1-p)^{k_i-1}p\Big)}_{=1,\ \forall i}$$

$$= (1-p)^{k_1-1}p, \quad \forall k_1 \in \mathbb{N}.$$

Analogno za X_i, $i=2,\dots,n$. ∎

Pr. 5.24 Neka je $n\in\mathbb{N}$ i

X = broj fiksnih točaka slučajno odabrane permutacije od $\{1,2,\dots,n\}$.

Odredimo $\mathbb{E}[X]$ i $\mathrm{Var}(X)$.

Rj. Neka je $A_i = \{i$ je fiksna točka odabrane permutacije$\}$, $i=1,\dots,n$

$$\Rightarrow \boxed{X = \sum_{i=1}^{n} \mathbb{1}_{A_i}}$$

Iz znamo

$$\mathbb{P}(A_i) = \frac{1}{n}, \quad \forall i$$

$$\mathbb{P}(A_i \cap A_j) = \frac{1}{n(n-1)}, \quad \forall i \neq j.$$

Specijalno, $A_1, \dots, A_n$ ~~nisu~~ nezavisni događaji!

Imamo,

ne trebamo nezavisnost!

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_{A_i}] = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

$$= n \cdot \mathbb{P}(A_1) = \boxed{1}.$$

Nadalje,

$$\mathrm{Var}(X) \overset{(S.11)}{=} \sum_{i=1}^{n} \mathrm{Var}(\mathbb{1}_{A_i}) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_j})$$

- $$\mathrm{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_j}) = \mathbb{E}[\underbrace{\mathbb{1}_{A_i} \cdot \mathbb{1}_{A_j}}_{= \mathbb{1}_{A_i \cap A_j}}] - \mathbb{E}[\mathbb{1}_{A_i}]\,\mathbb{E}[\mathbb{1}_{A_j}]$$

$$= \boxed{\mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\,\mathbb{P}(A_j)}$$

$$= \frac{1}{n(n-1)} - \frac{1}{n^2} = \boxed{\frac{1}{n^2(n-1)}}, \quad \forall i \neq j$$

$$\Rightarrow \mathrm{Var}(X) = n \cdot \overbrace{\mathrm{Var}(\mathbb{1}_{A_1})}^{= \mathbb{P}(A_1)(1 - \mathbb{P}(A_1))} + 2 \cdot \binom{n}{2} \mathrm{Cov}(\mathbb{1}_{A_1}, \mathbb{1}_{A_2})$$

$$= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2 \cdot \frac{n(n-1)}{2} \frac{1}{n^2(n-1)}$$

$$= 1 - \frac{1}{n} + \frac{1}{n} = \boxed{1} \quad \blacksquare$$

DZ) Ako je $X \sim B(n, p)$, pokažite

$$\mathrm{Var}(X) = npq.$$

$\boxed{\mathrm{Var}(\mathbb{1}_{A_i}) = pq, \forall i}$ također nezavisne!

(Uputa: ako $A_1, \dots, A_n$ su nezavisne t.d. $\mathbb{P}(A_i) = p, \forall i \Rightarrow X := \sum_{i=1}^{n} \mathbb{1}_{A_i} \sim B(n, p)$.)

Korelacija

(DZ) Ako je $\mathbb{E}[|X|] < \infty$, pokažite da je
$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

(Uputa: uzmi $|X| - X \geq 0$ i $|X| + X \geq 0$.)

Prop. 5.25 Neka su X, Y diskretne sl. v. var. t.d. $\mathbb{E}[X^2] < \infty$ i $\mathbb{E}[Y^2] < \infty$. Tada je $\mathbb{E}[|XY|] < \infty$ te
$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]} \quad (5.17)$$

tzv. Cauchy-Schwartzova nejednakost

(bez dokaza → vidi Lema 5.14. Sarapa, Vondr.) ■

Nap. 5.26 Ako je $\mathbb{E}[X^2] < \infty$ i $\mathbb{E}[Y^2] < \infty$,
$$\mathbb{E}\left[|(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])|\right]$$
$$\leq \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]} \cdot \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]}$$
((5.17) na $X - \mathbb{E}[X]$ i $Y - \mathbb{E}[Y]$)
$$= \sqrt{Var(X)} \cdot \sqrt{Var(Y)} < \infty$$

$\Rightarrow$ $Cov(X,Y)$ je dobro definirana te
$$\boxed{|Cov(X,Y)| = |\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]| \leq \sqrt{Var(X)}\sqrt{Var(Y)}} \quad (5.20)$$

■

Def. 5.27 Neka su X i Y diskretne sl. varijable t.d. $\mathbb{E}[X^2] < \infty$ i $\mathbb{E}[Y^2] < \infty$. Koeficijent korelacije od X i Y definira se kao

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

[češće se koristi u statistici nego $Cov(X,Y)$]

Prop. 5.28 (Svojstva korelacije)

(i) $|\rho(X,Y)| \leq 1$;

(ii) $\boxed{|\rho(X,Y)| = 1}$ akko $\exists a,b \in \mathbb{R}$ t.d. $\boxed{\mathbb{P}(Y = a + bX) = 1}$, te

- $\rho(X,Y) = 1 \Rightarrow b > 0$
- $\rho(X,Y) = -1 \Rightarrow b < 0$.

(iii) X i Y nezavisne $\Rightarrow \rho(X,Y) = 0$.

Dokaz. (i) slijedi iz (5.20), (ii) bez dokaza, a (iii) je trivijalno. ∎

Pr. 5.29 Bacamo dva simetrična novčića,

X = broj P, Y = broj G

$\Rightarrow$ pokazali smo da je $Cov(X,Y) = -\frac{1}{2}$.

$X, Y \sim B(2, \frac{1}{2})$

Imamo, $Var(X) = Var(Y) = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$

$$\Rightarrow \rho(X,Y) = \frac{-\frac{1}{2}}{\sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}}} = \boxed{-1}$$

jer $Y = 2 - X$

[ostali zadaci na vježbama]