Hipergeometriai eloszlás: példa (6. előadás)

Egy sportcsapat $N = 20$ **tagja** közül $M = 9$ **balkezes**.
A pályán egyszerre $n = 7$ **különböző** játszik.
Tegyük fel, hogy minden hétfős összeállítás egyformán valószínű.

Milyen eloszlású a pályán a **balkezes játékosok száma,** X**?**

Hipergeometriai eloszlás: példa (6. előadás)

Egy sportcsapat $N = 20$ **tagja** közül $M = 9$ **balkezes**.
A pályán egyszerre $n = 7$ **különböző** játszik.
Tegyük fel, hogy minden hétfős összeállítás egyformán valószínű.

Milyen eloszlású a pályán a **balkezes játékosok száma,** X**?**

A hétfős összeállítások száma:

k balkezes játékos kiválasztása:

$7 - k$ jobbkezes játékos kiválasztása:

A jó lehetőségek száma összesen:

Hipergeometriai eloszlás: példa (6. előadás)

Egy sportcsapat $N = 20$ **tagja** közül $M = 9$ **balkezes**.
A pályán egyszerre $n = 7$ **különböző** játszik.
Tegyük fel, hogy minden hétfős összeállítás egyformán valószínű.

Milyen eloszlású a pályán a **balkezes játékosok száma,** X?

A hétfős összeállítások száma: $\binom{20}{7}$

k balkezes játékos kiválasztása: $\binom{9}{k}$

$7 - k$ jobbkezes játékos kiválasztása: $\binom{11}{7-k}$

A jó lehetőségek száma összesen: $\binom{9}{k} \cdot \binom{11}{7-k} \longleftarrow$

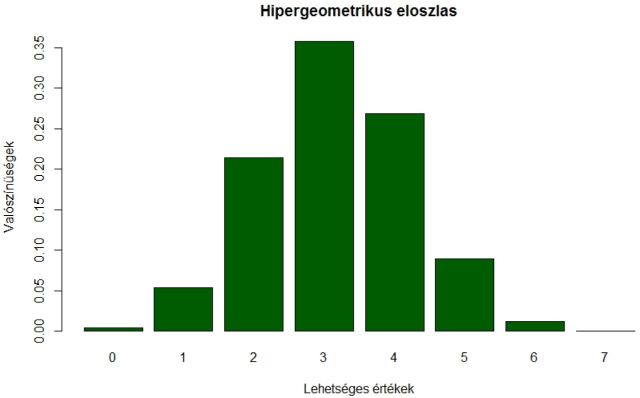
szorzás: bármely balkezes választás bármely jobbkezessel jó

osztás: minden lehetőség egyformán valószínű

$$\boxed{\mathbb{P}(X = k) = \frac{\binom{9}{k} \cdot \binom{11}{7-k}}{\binom{20}{7}}}$$

visszatevés nélküli mintavétel

Hipergeometriai eloszlás: példa



A kiválasztott balkezes játékosok számának eloszlása hipergeometriai eloszlás, $N = 20, \ M = 9, n = 7$

vízszintes: k, oszlopok magassága: $\mathbb{P}(X = k) = \frac{\binom{9}{k} \cdot \binom{11}{7-k}}{\binom{20}{7}}$.

Hipergeometriai eloszlás

Legyenek N, M, n pozitív egészek úgy, hogy $1 \leq n \leq M \leq N$. Az X valószínűségi változó **hipergeometriai eloszlású**, ha

$$\mathbb{P}(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad (k = 0, 1, \ldots, n).$$

- **visszatevés nélküli mintavételnél** a húzott fekete golyók száma: N golyó, ebből M fekete, n-szer húzunk visszatevés nélkül
- lottósorsolásnál a találatok száma, X, hipergeometrikus eloszlású $N = 90$, $M = 5$, $n = 5$ paraméterekkel:

$$\mathbb{P}(X = k) = \mathbb{P}(k \text{ találat}) = \frac{\binom{5}{k}\binom{85}{5-k}}{\binom{90}{5}} \qquad k = 0, 1, 2, 3, 4, 5.$$

A hipergeometriai eloszlás várható értéke és szórása

Ha az X valószínűségi változó hipergeometriai eloszlású M, N, n paraméterekkel, azaz

$$\mathbb{P}(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad (k = 0, 1, \ldots, n),$$

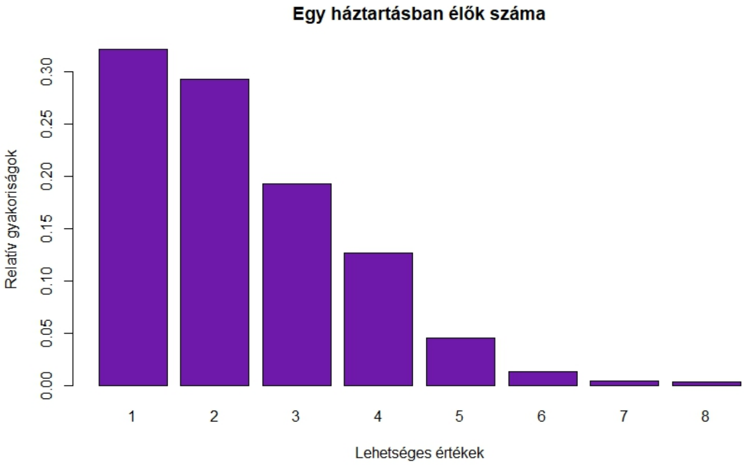
akkor

$$\mathbb{E}(X) = \frac{M}{N}n; \qquad D(X) = \sqrt{n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}}.$$

Például, ha $N = 20$ játékos közül $M = 9$ balkezes, és $n = 7$-et választunk visszatevés nélkül, akkor a balkezes játékosok számának, X-nek **várható értéke** és **szórása**:

$$\mathbb{E}(X) = \frac{9}{20} \cdot 7 = 3,15; \quad D(X) = \sqrt{7 \cdot \frac{9}{20} \cdot \left(1 - \frac{9}{20}\right) \cdot \frac{13}{19}} = 1,09.$$

Megfigyelések



Egy háztartásban élők számának hisztogramja (forrás: KSH, 2011) ($n = 4105698$ a háztartások száma)

Geometriai eloszlás

Egy közvéleménykutatásban mindenki a többiektől függetlenül $0,2$ valószínűséggel válaszol egy adott kérdésre. Jelölje Y, hogy hány embert kell megkérdezni, míg találunk egy válaszadót.

$\mathbb{P}(Y = 1) = \mathbb{P}(\text{az első ember válaszol}) = 0,2;$

$\mathbb{P}(Y = 2) = \mathbb{P}(\text{az első nem válaszol, a második igen}) =$

Geometriai eloszlás

Egy közvéleménykutatásban mindenki a többiektől függetlenül $0,2$ valószínűséggel válaszol egy adott kérdésre. Jelölje Y, hogy hány embert kell megkérdezni, míg találunk egy válaszadót.

$\mathbb{P}(Y = 1) = \mathbb{P}(\text{az első ember válaszol}) = 0,2;$

$\mathbb{P}(Y = 2) = \mathbb{P}(\text{az első nem válaszol, a második igen}) = 0,8 \cdot 0,2;$

$\mathbb{P}(Y = 3) = \mathbb{P}(\text{az első kettő nem válaszol, a harmadik igen}) =$

Geometriai eloszlás

Egy közvéleménykutatásban mindenki a többiektől függetlenül $0,2$ valószínűséggel válaszol egy adott kérdésre. Jelölje Y, hogy hány embert kell megkérdezni, míg találunk egy válaszadót.

$\mathbb{P}(Y = 1) = \mathbb{P}(\text{az első ember válaszol}) = 0,2;$

$\mathbb{P}(Y = 2) = \mathbb{P}(\text{az első nem válaszol, a második igen}) = 0,8 \cdot 0,2;$

$\mathbb{P}(Y = 3) = \mathbb{P}(\text{az első kettő nem válaszol, a harmadik igen}) = 0,8^2 \cdot 0,2;$

$\mathbb{P}(Y = k) = \mathbb{P}(\text{az első } k-1 \text{ nem válaszol, a } k\text{. igen}) =$

Geometriai eloszlás

Egy közvéleménykutatásban mindenki a többiektől függetlenül $0,2$ valószínűséggel válaszol egy adott kérdésre. Jelölje Y, hogy hány embert kell megkérdezni, míg találunk egy válaszadót.

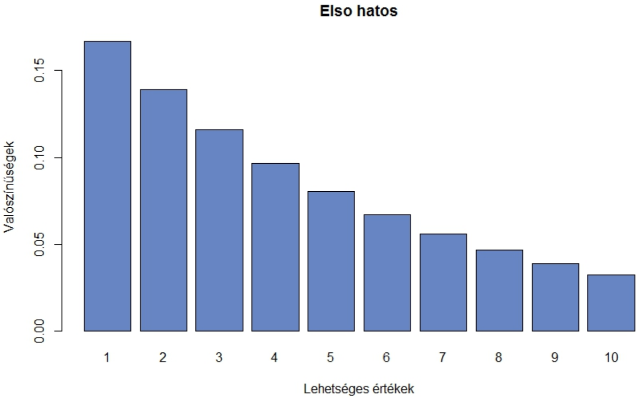
$\mathbb{P}(Y = 1) = \mathbb{P}(\text{az első ember válaszol}) = 0,2;$

$\mathbb{P}(Y = 2) = \mathbb{P}(\text{az első nem válaszol, a második igen}) = 0,8 \cdot 0,2;$

$\mathbb{P}(Y = 3) = \mathbb{P}(\text{az első kettő nem válaszol, a harmadik igen}) = 0,8^2 \cdot 0,2;$

$\mathbb{P}(Y = k) = \mathbb{P}(\text{az első } k-1 \text{ nem válaszol, a } k\text{. igen}) = 0,8^{k-1} \cdot 0,2.$

Példa: geometriai eloszlás



Az első hatos eloszlása: geometriai eloszlás, $p = 1/6$, $k = 10$-ig

Geometriai eloszlás

- független kísérleteket végzünk;
- mindegyik p valószínűséggel sikerül;
- Y: hányadik kísérlet az első sikeres.

Geometriai eloszlás

- független kísérleteket végzünk;
- mindegyik p valószínűséggel sikerül;
- Y: hányadik kísérlet az első sikeres.

Definíció

Az Y valószínűségi változó **geometriai eloszlású** *p paraméterrel, ha lehetséges értékei:*

$$1, 2, 3 \dots$$

és minden $1 \leq k$ egészre

$$\mathbb{P}(Y = k) = (1 - p)^{k-1} p.$$

$(0 < p < 1.)$ Jelölés: $\mathrm{Geo}(p)$. Másik elnevezés: Pascal-eloszlás.

Mivel $\sum_{k=1}^{\infty} (1 - p)^{k-1} p = 1$, ez valóban valószínűségeloszlás (annak valószínűsége, hogy sosem sikerül a kísérlet, 1).

Geometriai eloszlás várható értéke és szórása

Állítás

Ha az X valószínűségi változó geometriai eloszlású p paraméterrel, azaz

$$\mathbb{P}(X = k) = (1-p)^{k-1}p \qquad (k = 1, 2, \ldots),$$

akkor

$$\mathbb{E}(X) = \frac{1}{p}; \qquad D(X) = \sqrt{\frac{1-p}{p^2}}.$$

Geometriai eloszlás várható értéke és szórása

Állítás

Ha az X valószínűségi változó geometriai eloszlású p paraméterrel, azaz

$$\mathbb{P}(X = k) = (1-p)^{k-1} p \qquad (k = 1, 2, \ldots),$$

akkor

$$\mathbb{E}(X) = \frac{1}{p}; \qquad D(X) = \sqrt{\frac{1-p}{p^2}}.$$

Példa. Tegyük fel, hogy egy adott pártot mindenki a többiektől függetlenül $p = 0,06$ valószínűséggel támogat. Jelölje X, hogy hány embert kell megkérdezni, míg az első olyan embert megtaláljuk, aki ezt a pártot támogatná. Ekkor

$$\mathbb{E}(X) = \frac{1}{0,06} = 16,67; \qquad D(X) = \sqrt{\frac{0,94}{0,06^2}} = 16,16.$$

Nevezetes diszkrét eloszlások várható értéke és szórása

- Ha X binomiális eloszlású n renddel és p paraméterrel:

$$\mathbb{E}(X) = np; \qquad D(X) = \sqrt{np(1-p)}.$$

- Ha X hipergeometriai eloszlású N, M, n paraméterekkel:

$$\mathbb{E}(X) = \frac{M}{N}n; \qquad D(X) = \sqrt{n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}}.$$

- Ha X Poisson-eloszlású λ paraméterrel:

$$\mathbb{E}(X) = \lambda; \qquad D(X) = \sqrt{\lambda};$$

- Ha X geometriai eloszlású p paraméterrel:

$$\mathbb{E}(X) = \frac{1}{p}; \qquad D(X) = \sqrt{\frac{1-p}{p^2}}.$$

Függetlenség: példa

Mely valószínűségi változók tekinthetők egymástól függetlennek, és melyek között van kapcsolat? Zsófia egy felmérés véletlenszerűen választott résztvevője.

Zsófia autóinak száma

csapadékmennyiség holnap Budapesten

Zsófia havi jövedelme

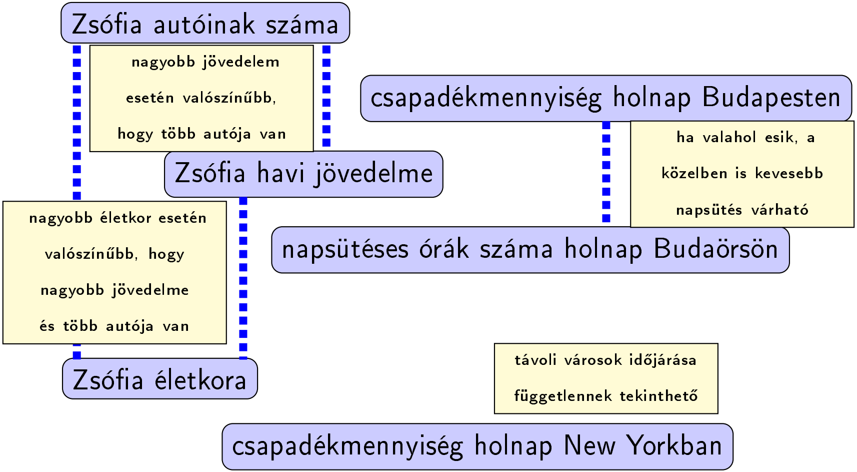
napsütéses órák száma holnap Budaörsön

Zsófia életkora

csapadékmennyiség holnap New Yorkban

Függetlenség: példa

Mely valószínűségi változók tekinthetők egymástól függetlennek, és melyek között van kapcsolat? Zsófia egy felmérés véletlenszerűen választott résztvevője.



Függetlenség: példa

Emlékeztető: az A és B események **függetlenek**, ha

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Ha például X a csapadékmennyiség holnap Budapesten (mm-ben), és Y New Yorkban, akkor például

$$A : X \leq 5; \qquad B : Y \leq 5$$

esetén ez a feltétel így írható:

$$\mathbb{P}(X \leq 5, Y \leq 5) = \mathbb{P}(X \leq 5) \cdot \mathbb{P}(Y \leq 5).$$

Azaz, feltételezve, hogy a két város időjárása egymástól független: annak valószínűsége, hogy **mindkét helyen legfeljebb 5 mm csapadék lesz**, a két esemény **valószínűségének szorzata**.

Valószínűségi változók függetlensége

- **két valószínűségi változóra**: az $X, Y : \Omega \to \mathbb{R}$ valószínűségi változók **függetlenek**, ha

$$\mathbb{P}(X \leq t_1, Y \leq t_2) = \mathbb{P}(X \leq t_1) \cdot \mathbb{P}(Y \leq t_2)$$

  teljesül tetszőleges $t_1, t_2 \in \mathbb{R}$ valós számokra.

- **véges sok valószínűségi változóra**: $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ valószínűségi változók **függetlenek**, ha

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n) = \\ = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \ldots \mathbb{P}(X_n \leq t_n)$$

  teljesül tetszőleges $t_1, t_2, \ldots, t_n$ valós számokra.

- **megszámlálható sok valószínűségi változóra**: az $X_1, X_2, X_3 \ldots$ valószínűségi változók **függetlenek**, ha közülük bármely véges sokat kiválasztva független valószínűségi változókat kapunk.

Függetlenség diszkrét esetben

Ha a valószínűségi változók **diszkrétek**, azaz lehetséges értékeik halmaza véges vagy megszámlálhatóan végtelen, akkor a függetlenséget az alábbi módon is ellenőrizhetjük.

Az X és Y **diszkrét** valószínűségi változók pontosan akkor **függetlenek**, ha

az X **minden lehetséges** x_k **értékére** és

az Y **minden lehetséges** y_l **értékére** teljesül, hogy

$$\mathbb{P}(X = x_k, Y = y_l) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_l) \qquad (k, l = 1, 2, \ldots).$$

Vagyis: annak valószínűsége, hogy X **értéke** x_k **és** Y **értéke** y_l, ennek a két eseménynek **a valószínűségének a szorzata**.

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy az **első dobott szám** és a **második dobott szám** értéke egymástól független?

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy az **első dobott szám** és a **második dobott szám** értéke egymástól független?

Tipp. a második dobásnál az első dobás értéke „elfelejtődik”, nincs kapcsolat a két dobás között $\Rightarrow$ a két dobott szám **független.**

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy az **első dobott szám** és a **második dobott szám** értéke egymástól független?

Tipp. a második dobásnál az első dobás értéke „elfelejtődik”, nincs kapcsolat a két dobás között $\Rightarrow$ a két dobott szám független.

Indoklás. Legyen X az első dobás, Y a második. Legyen például $x_k = 3, y_l = 5$. Ekkor a feltétel teljesül:

$$\frac{1}{36} = \mathbb{P}(X = 3, Y = 5) = \mathbb{P}(X = 3) \cdot \mathbb{P}(Y = 5) = \frac{1}{6} \cdot \frac{1}{6}.$$

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy az **első dobott szám** és a **második dobott szám** értéke egymástól független?

Tipp. a második dobásnál az első dobás értéke „elfelejtődik", nincs kapcsolat a két dobás között $\Rightarrow$ a két dobott szám **független.**

Indoklás. Legyen X az első dobás, Y a második. Legyen például $x_k = 3, y_l = 5$. Ekkor a feltétel teljesül:

$$\frac{1}{36} = \mathbb{P}(X = 3, Y = 5) = \mathbb{P}(X = 3) \cdot \mathbb{P}(Y = 5) = \frac{1}{6} \cdot \frac{1}{6}.$$

Hasonlóképpen tetszőleges (x_k, y_l) lehetséges értékekre (azaz 1 és 6 közötti egészekre) igaz, hogy

$$\frac{1}{36} = \mathbb{P}(X = x_k, Y = y_l) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_l) = \frac{1}{6} \cdot \frac{1}{6}.$$

Ezért valóban **a két dobás egymástól független**.

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy a dobott számok **összege** és **szorzata** független egymástól?

Valószínűségi változók függetlensége: példa

Kétszer dobunk egy szabályos dobókockával. Igaz-e, hogy a dobott számok **összege** és **szorzata** független egymástól?

Tipp: minél nagyobb az összeg, annál valószínűbb, hogy a szorzat értéke is inkább nagy lesz $\Rightarrow$ **nem függetlenek**.

Indoklás: legyen X az összeg, Y a szorzat. Ha például $X = 2$: ez csak úgy lehet, hogy mindkét dobás 1-es, vagyis ekkor Y értéke biztosan 1. Ezért ha például $x_1 = 2$ és $y_2 = 2$-t választunk, $X = 2$ és $Y = 2$ egyszerre nem következhetnek be, és így:

$$0 = \mathbb{P}(X = 2, Y = 2) \neq \mathbb{P}(X = 2) \cdot \mathbb{P}(Y = 2) =$$

$$= \mathbb{P}(11) \cdot \mathbb{P}(12 \text{ vagy } 21) = \frac{1}{36} \cdot \frac{1}{18} > 0.$$

Vagyis az $x_1 = 2$ és $y_2 = 2$ párra nem teljesül az előírt feltétel, az **összeg és szorzat nem függetlenek**.

A várható érték tulajdonságai

- (összeg várható értéke) Ha X, Y valószínűségi változók, és X, Y, $X+Y$ várható értéke létezik, akkor

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

- (konstans kiemelése) Ha az X valószínűségi változó várható értéke létezik, és c tetszőleges valós szám, akkor

$$\mathbb{E}(c\cdot X)=c\cdot\mathbb{E}(X).$$

A várható érték tulajdonságai

- (összeg várható értéke) Ha X, Y valószínűségi változók, és $X, Y, X+Y$ várható értéke létezik, akkor
$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$
- (konstans kiemelése) Ha az X valószínűségi változó várható értéke létezik, és c tetszőleges valós szám, akkor
$$\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X).$$
- (szorzat várható értéke független esetben) Ha az X és Y valószínűségi változók függetlenek, és $X, Y, X \cdot Y$ várható értéke létezik, akkor
$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

A várható érték tulajdonságai

- (összeg várható értéke) Ha X, Y valószínűségi változók, és $X, Y, X+Y$ várható értéke létezik, akkor
$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$
- (konstans kiemelése) Ha az X valószínűségi változó várható értéke létezik, és c tetszőleges valós szám, akkor
$$\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X).$$
- (szorzat várható értéke független esetben) Ha az X és Y valószínűségi változók függetlenek, és $X, Y, X \cdot Y$ várható értéke létezik, akkor
$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$
- (függvény várható értéke) Ha $g : \mathbb{R} \to \mathbb{R}$ olyan függvény, hogy $\mathbb{E}(X)$ létezik, és az X lehetséges értékei $x_1, x_2, \ldots$, akkor
$$\mathbb{E}(g(X)) = \sum_{i=1}^{\infty} g(x_i)\mathbb{P}(X = x_i).$$

Összeg várható értéke

Állítás

Ha X, Y valószínűségi változók, és $X, Y, X+Y$ várható értéke létezik, akkor

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

Bizonyítás. Legyenek X lehetséges értékei $x_1, x_2, \ldots$, az Y lehetséges értékei $y_1, y_2, \ldots$. Ekkor az $X+Y$ lehetséges értékei $x_k + y_m$ alakúak, és

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{k,m} (x_k + y_m)\mathbb{P}(X = x_k, Y = y_m) = \\
&= \sum_{k,m} x_k \mathbb{P}(X = x_k, Y = y_m) + \sum_{k,m} y_m \mathbb{P}(X = x_k, Y = y_m) = \\
&= \sum_{k} x_k \mathbb{P}(X = x_k) + \sum_{m} y_m \mathbb{P}(Y = y_m) = \mathbb{E}(X) + \mathbb{E}(Y),
\end{aligned}$$

ahol azt használtuk, hogy az $\{X = x_k, Y = y_m\}$ események kizáróak, uniójuk $\{X = x_k\}$, és hasonlóképpen a másik tagban az Y esetén.

Összeg várható értéke

Állítás

Ha X, Y valószínűségi változók, és $X, Y, X+Y$ várható értéke létezik, akkor

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

Bizonyítás. Legyenek X lehetséges értékei $x_1, x_2, \ldots$, az Y lehetséges értékei $y_1, y_2, \ldots$. Ekkor az $X+Y$ lehetséges értékei $x_k + y_m$ alakúak, és

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{k,m} (x_k + y_m)\mathbb{P}(X = x_k, Y = y_m) = \\
&= \sum_{k,m} x_k \mathbb{P}(X = x_k, Y = y_m) + \sum_{k,m} y_m \mathbb{P}(X = x_k, Y = y_m) = \\
&= \sum_{k} x_k \mathbb{P}(X = x_k) + \sum_{m} y_m \mathbb{P}(Y = y_m) = \mathbb{E}(X) + \mathbb{E}(Y),
\end{aligned}$$

ahol azt használtuk, hogy az $\{X = x_k, Y = y_m\}$ események kizáróak, uniójuk $\{X = x_k\}$, és hasonlóképpen a másik tagban az Y esetén.

Következmény: $\mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n)$.

Szorzat várható értéke független esetben

Állítás

Ha X, Y **független** *valószínűségi változók, és X, Y várható értéke létezik, akkor*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Bizonyítás. Legyenek X lehetséges értékei $x_1, x_2, \ldots$, az Y lehetséges értékei $y_1, y_2, \ldots$. Ekkor az $X + Y$ lehetséges értékei $x_k + y_m$ alakúak, és

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_{k,m} x_k \cdot y_m \mathbb{P}(X = x_k, Y = y_m) = \\
&\stackrel{(*)}{=} \sum_{k,m} x_k y_m \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_m) \\
&= \left(\sum_k x_k \mathbb{P}(X = x_k) \right) \cdot \left(\sum_m y_m \mathbb{P}(Y = y_m) \right) = \mathbb{E}(X) \cdot \mathbb{E}(Y),
\end{aligned}$$

ahol a $(*)$ lépésben használtuk a függetlenségnek a diszkrét valószínűségi változókra vonatkozó alakját.

A szórásnégyzet tulajdonságai

- (nemnegativitás) $D^2(X) \geq 0$ és $D(X) \geq 0$ mindig teljesül
- (szorzás és eltolás) ha a, b valós számok, X véges szórású valószínűségi változó, akkor

$$D^2(aX + b) = a^2 D^2(X) \quad \Rightarrow \quad D(aX + b) = |a| D(X).$$

A szórásnégyzet tulajdonságai

- (nemnegativitás) $D^2(X) \geq 0$ és $D(X) \geq 0$ mindig teljesül
- (szorzás és eltolás) ha a, b valós számok, X véges szórású valószínűségi változó, akkor

$$D^2(aX + b) = a^2 D^2(X) \quad \Rightarrow \quad D(aX + b) = |a| D(X).$$

- (összeg szórása független esetben) ha az X, Y valószínűségi változók függetlenek és szórásuk létezik, akkor

$$D^2(X+Y) = D^2(X)+D^2(Y) \quad \Rightarrow \quad D(X+Y) = \sqrt{D^2(X) + D^2(Y)}.$$

- van olyan valószínűségi változó, melynek várható értéke véges, de a szórása nem létezik (például: $\mathbb{P}(X = k) = c/k^3$ megfelelő c-vel)

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

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

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2;$
- $\mathbb{E}(2X+3Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2;$
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26;$
- $\mathbb{E}(2X-3Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2;$
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2\cdot 4 + 3\cdot 6 = 26;$
- $\mathbb{E}(2X-3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2\cdot 4 - 3\cdot 6 = -10;$
- $D(X+Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$;
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$;
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$;
- $\mathbb{E}(2X-3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$;
- $D(X+Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1+2^2} = \sqrt{5}$;
- $D(X-Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2;$
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26;$
- $\mathbb{E}(2X-3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10;$
- $D(X+Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1+2^2} = \sqrt{5};$
- $D(X-Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1+2^2} = \sqrt{5};$
- $D(2X+3Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10;$
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2;$
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2\cdot 4 + 3\cdot 6 = 26;$
- $\mathbb{E}(2X-3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2\cdot 4 - 3\cdot 6 = -10;$
- $D(X+Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1+2^2} = \sqrt{5};$
- $D(X-Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1+2^2} = \sqrt{5};$
- $D(2X+3Y) = \sqrt{2^2 D^2(X) + 3^2 D^2(Y)} = \sqrt{4\cdot 1 + 9\cdot 4} = \sqrt{40};$
- $D(2X-3Y) =$

A várható érték és szórás kiszámítása

Példa. Legyen X olyan valószínűségi változó, melynek várható értéke 4, szórása 1, az Y valószínűségi változó várható értéke 6, szórása pedig 2. Tegyük fel továbbá, hogy X és Y **függetlenek**. Ekkor

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$;
- $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$;
- $\mathbb{E}(2X+3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$;
- $\mathbb{E}(2X-3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$;
- $D(X+Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1+2^2} = \sqrt{5}$;
- $D(X-Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1+2^2} = \sqrt{5}$;
- $D(2X+3Y) = \sqrt{2^2 D^2(X) + 3^2 D^2(Y)} = \sqrt{4 \cdot 1 + 9 \cdot 4} = \sqrt{40}$;
- $D(2X-3Y) = \sqrt{2^2 D^2(X) + (-3)^2 D^2(Y)} = \sqrt{4 \cdot 1 + 9 \cdot 4} = \sqrt{40}$.

A binomiális eloszlás várható értéke és szórása

Állítás

Ha az X valószínűségi változó binomiális eloszlású n renddel és p paraméterrel, azaz

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n),$$

akkor

$$\mathbb{E}(X) = np; \qquad D(X) = \sqrt{np(1-p)}.$$

Példa. Egy kérdőív egy kérdésére $n = 1000$ megkérdezett közül mindenki a többiektől függetlenül $p = 0,65$ valószínűséggel válaszol. Ekkor a válaszadók számának (melyet jelöljünk X-szel) a várható értéke:

$$\mathbb{E}(X) = np = 1000 \cdot 0,65 = 650,$$

míg a szórása:

$$D(X) = \sqrt{np(1-p)} = \sqrt{1000 \cdot 0,65 \cdot 0,35} = 15,08.$$

A binomiális eloszlás várható értéke és szórása

Bizonyítás. Legyen X binomiális eloszlású n renddel és p paraméterrel. Azaz: n független kísérletet végzünk, mindegyik p valószínűséggel sikerül, X a sikeresek száma. Vezessük be az alábbi indikátorokat $j = 1, 2, \ldots, n$ esetén:

$$X_j = \begin{cases} 1 & \text{ha a } j. \text{ kísérlet sikeres;} \\ 0 & \text{ha a } j. \text{ kísérlet nem sikeres.} \end{cases}$$

Ekkor X éppen az indikátorok összege (az egyesek száma):

$$X = X_1 + X_2 + \ldots + X_n = \sum_{j=1}^{n} X_j.$$

A binomiális eloszlás várható értéke és szórása

Bizonyítás. Legyen X binomiális eloszlású n renddel és p paraméterrel. Azaz: n független kísérletet végzünk, mindegyik p valószínűséggel sikerül, X a sikeresek száma. Vezessük be az alábbi indikátorokat $j = 1, 2, \ldots, n$ esetén:

$$X_j = \begin{cases} 1 & \text{ha a } j. \text{ kísérlet sikeres;} \\ 0 & \text{ha a } j. \text{ kísérlet nem sikeres.} \end{cases}$$

Ekkor X éppen az indikátorok összege (az egyesek száma):

$$X = X_1 + X_2 + \ldots + X_n = \sum_{j=1}^{n} X_j.$$

Mivel bármely j-re

$$\mathbb{E}(X_j) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p,$$

ezért

$$\mathbb{E}(X) = \mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n) = p + p + \ldots + p =$$

A binomiális eloszlás várható értéke és szórása

Bizonyítás. Ugyanazokat az indikátorokat használjuk:

$$X_j = \begin{cases} 1 & \text{ha a } j\text{.kísérlet sikeres;} \\ 0 & \text{ha a } j\text{.kísérlet nem sikeres.} \end{cases}$$

A binomiális eloszlás várható értéke és szórása

Bizonyítás. Ugyanazokat az indikátorokat használjuk:

$$X_j = \begin{cases} 1 & \text{ha a } j\text{.kísérlet sikeres;} \\ 0 & \text{ha a } j\text{.kísérlet nem sikeres.} \end{cases}$$

Most $X_j = X_j^2$, hiszen $0^2 = 0$ és $1^2 = 1$, és már láttuk, hogy $\mathbb{E}(X_j) = p$. Ezért

$$D^2(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2 = \mathbb{E}(X_j) - \mathbb{E}(X_j)^2 = p - p^2 = p(1-p),$$

Mivel az X_j indikátorok **függetlenek**, és az összegük X:

$$D(X) = \sqrt{D^2(X_1 + X_2 + \ldots + X_n)} = \sqrt{D^2(X_1) + D^2(X_2) + \ldots + D^2(X_n)} =$$
$$= \sqrt{p(1-p) + p(1-p) + \ldots + p(1-p)} = \sqrt{np(1-p)}.$$

A hipergeometrikus eloszlás várható értékére vonatkozó összefüggés szintén indikátorokkal igazolható.

Névjegy-probléma

n ember bedobja a névjegyét egy kalapba, ezután mindenki húz egyet véletlenszerűen, minden még ott lévőt azonos valószínűséggel választva. Legyen X a saját névjegyüket húzók száma. Mennyi X várható értéke?

Névjegy-probléma

n ember bedobja a névjegyét egy kalapba, ezután mindenki húz egyet véletlenszerűen, minden még ott lévőt azonos valószínűséggel választva. Legyen X a saját névjegyüket húzók száma. Mennyi X várható értéke?

$$X_j = \begin{cases} 1 & \text{ha a } j\text{. ember a sajátját húzza;} \\ 0 & \text{ha a } j\text{. ember nem a sajátját húzza.} \end{cases}$$

Mivel $X = X_1 + X_2 + \ldots + X_n$, és

$$\mathbb{E}(X_j) = \mathbb{P}(X_j = 1) = \mathbb{P}(\text{a } j\text{. ember a sajátját húzza}) = \frac{1}{n},$$

azt kapjuk, hogy

$$\mathbb{E}(X) = \mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n) = n \cdot \frac{1}{n} = 1.$$

Névjegy-probléma

n ember bedobja a névjegyét egy kalapba, ezután mindenki húz egyet véletlenszerűen, minden még ott lévőt azonos valószínűséggel választva. Legyen X a saját névjegyüket húzók száma. Mennyi X várható értéke?

$$X_j = \begin{cases} 1 & \text{ha a } j\text{. ember a sajátját húzza;} \\ 0 & \text{ha a } j\text{. ember nem a sajátját húzza.} \end{cases}$$

Mivel $X = X_1 + X_2 + \ldots + X_n$, és

$$\mathbb{E}(X_j) = \mathbb{P}(X_j = 1) = \mathbb{P}(\text{a } j\text{. ember a sajátját húzza}) = \frac{1}{n},$$

azt kapjuk, hogy

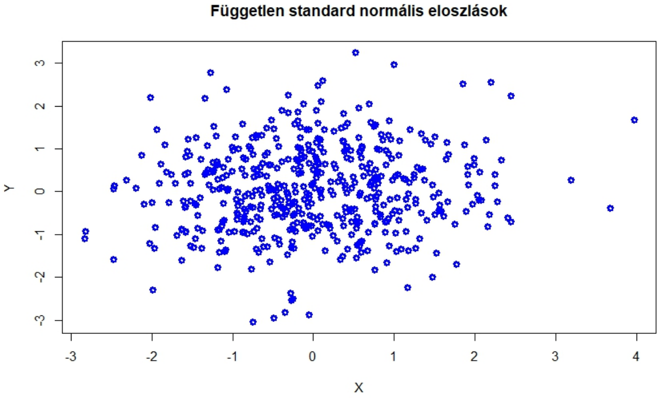
$$\mathbb{E}(X) = \mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n) = n \cdot \frac{1}{n} = 1.$$

Megjegyzés: $\lim_{n\to\infty} \mathbb{P}(X = k) = \frac{1}{e \cdot k!}$, vagyis Poisson(1) a határeloszlás. Ennek várható értéke is 1.

Kovariancia és korrelációs együttható

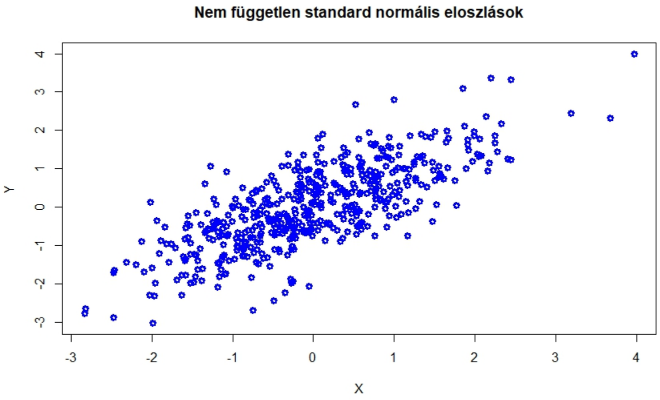
- két valószínűségi változó lehet
  - **független**: például két találomra választott ember jövedelme, vagy,
  - **nem független**: például egy találomra választott ember jövedelme most, illetve fél év múlva
- az **összefüggőség mértéke** különböző lehet:
  - egy találomra választott felnőtt életkora és jövedelme „erősen összefüggő", a fiataloké és időseké általában alacsonyabb;
  - egy találomra választott felnőtt életkora és testmagassága „gyengén összefüggő", hiszen egy fiatal felnőtt nőhet, az idősek pedig valamennyit veszítenek a testmagasságukból, de egyik változás sem nagyon jelentős.
- a kapcsolat erősségének jellemzésére többféle mérőszám használható, ezek között van a **kovarianca** és a **korrelációs együttható**. Ez utóbbinak a „nagy" értékei „erős, lineáris jellegű" összefüggésre utalnak.

Független normális eloszlások



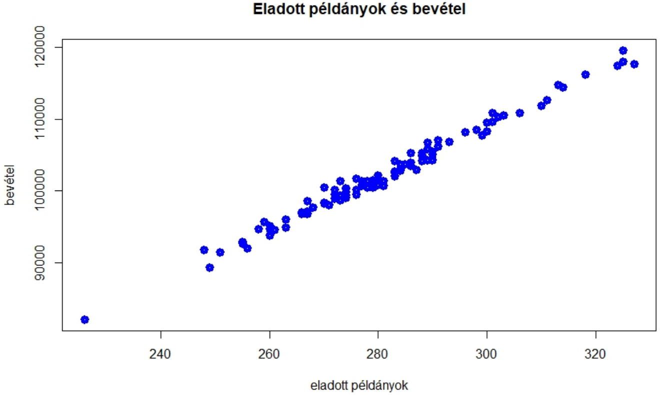
500 darab véletlen pont a síkon, melyek koordinátái **független** standard normális eloszlásúak. A koordináták között nincs kapcsolat: a kovariancia és a korrelációs együttható is 0 lesz.

Pozitív korreláció



500 elemű minta a következő többdimenziós normális eloszlásból: $(X, \frac{X+Z}{\sqrt{2}})$, ahol $X, Z \sim N(0,1)$ függetlenek.
Minél nagyobb X, „valószínűleg" annál nagyobb $(X+Z)/\sqrt{2}$ is $\rightarrow$ ennek megfelelően a két koordináta közötti **kovariancia** és **korrelációs együttható** is **pozitív** lesz.

Erős pozitív korreláció



100 elemű minta az $(X+Y, 300X+400Y)$ eloszlásból, ahol $X \sim \text{Poisson}(100)$ és $Y \sim \text{Poisson}(180)$ függetlenek. A megfigyelések szinte teljesen egy pozitív meredekségű egyenesre illeszkednek $\rightarrow$ a **korrelációs együttható** **pozitív** és **majdnem 1**, ami „nagy”, mert ennek 1 lesz a lehetséges legnagyobb értéke.

A kovariancia

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

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

A kovariancia

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

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

- **A kovariancia kiszámítása:**

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

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

A kovariancia

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

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

- **A kovariancia kiszámítása:**

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

- **Szimmetria.** $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
- **Kapcsolat a szórásnégyzettel.** $\mathrm{cov}(X, X) = D^2(X)$.
- **Függetlenséggel való kapcsolat.** Ha az X és Y valószínűségi változók **függetlenek**, akkor $\mathrm{cov}(X, Y) = 0$.

  Fordítva nem igaz: $\mathrm{cov}(X, Y) = 0$ esetén nem biztos, hogy X és Y függetlenek.

A kovariancia tulajdonságai

- Konstanssal való kovariancia. $\mathrm{cov}(X, c) = 0$, ha c valós szám.
- **Linearitás.** Egyrészt

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

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

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

- **Összeg szórásnégyzete.** $D^2(X+Y) = D^2(X)+D^2(Y)+2\mathrm{cov}(X, Y)$. Továbbá

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

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

Kovariancia: példa

Egy üzletben az A és B újság forgalmát figyelik.

- 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

Egy üzletben az A és B újság forgalmát figyelik.

- 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? Azaz mennyi $\mathrm{cov}(X+Y, 300X+400Y)$?

A számolás előtt: **pozitív**, **negatív** vagy 0 kovarianciára tippelnénk?

Kovariancia: példa

Egy üzletben az A és B újság forgalmát figyelik.

- 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? Azaz mennyi $\operatorname{cov}(X+Y, 300X+400Y)$?

A számolás előtt: **pozitív**, **negatív** vagy 0 kovarianciára tippelnénk?

Mivel **minél nagyobb** a példányszám, **„valószínűleg" annál nagyobb a bevétel**, **pozitív** kovarianciára számíthatunk.

Kovariancia: példa

X és Y **független ek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** és a **bevételnek** a kovarianciája:

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

ahol felhasználtuk, hogy

Kovariancia: példa

X és Y **független**ek, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** és a **bevételnek** a kovarianciája:

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

ahol felhasználtuk, hogy (a) a kovariancia **lineáris**;

Kovariancia: példa

X és Y **függetlenek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** és a **bevételnek** a kovarianciája:

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

ahol felhasználtuk, hogy (a) a kovariancia **lineáris**;

(b) **független** valószínűségi változók kovarianciája 0, illetve egy valószínűségi változó saját magával vett kovarianciája a szórásnégyzete;

Kovariancia: példa

X és Y **függetlenek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** és a **bevételnek** a kovarianciája:

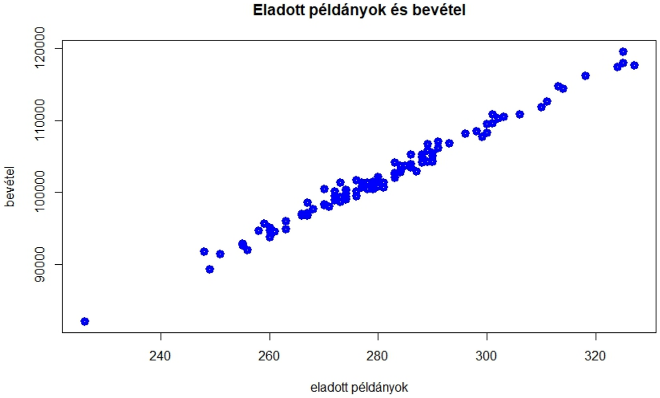
$$\begin{aligned}\mathrm{cov}(X+Y, 300X+400Y) &\stackrel{(a)}{=} \mathrm{cov}(X,300X)+\mathrm{cov}(X,400Y)+\\ &\quad+\mathrm{cov}(Y,300X)+\mathrm{cov}(Y,400Y)=\\ &\stackrel{(a,b)}{=} 300\cdot\mathrm{cov}(X,X)+400\cdot\mathrm{cov}(Y,Y)=\\ &\stackrel{(b)}{=} 300D^2(X)+400D^2(Y)=\\ &\stackrel{(c)}{=} 300\cdot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

ahol felhasználtuk, hogy (a) a kovariancia **lineáris**;

(b) **független** valószínűségi változók kovarianciája 0, illetve egy valószínűségi változó saját magával vett kovarianciája a szórásnégyzete;

(c) egy λ paraméterű **Poisson-eloszlású** valószínűségi változó **szórásnégyzete** λ.

Kovariancia: példa



A bevétel ($300X + 400Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ független megfigyelésből. Kovariancia: $\mathrm{cov}(X + Y, 300X + 400Y) = 102000$.

Korrelációs együttható: bevezetés

- A **kovariancia** bevezetésének célja, hogy két valószínűségi változó közötti **összefüggés erősségét** tudjuk mérni.
- A korábbi példában: a példányszám és a bevétel kovarianciája **102000** volt.
- Viszont ha a bevételt nem forintban, hanem ezer forintos egységben mérjük:

$$X: \text{példányszám} \quad Y: \text{bevétel forintban} \quad Z: \text{bevétel ezer forintban},$$

  akkor

$$\operatorname{cov}(X, Z) = \operatorname{cov}\left(X, \frac{Y}{1000}\right) = \frac{\operatorname{cov}(X, Y)}{1000} = 102.$$

  Vagyis a kovariancia a **mértékegységtől függ** $\Rightarrow$ hasznos egy olyan mennyiség, ami szintén az összefüggés erősségét méri, de a mértékegység választásától függetlenül.
- Ilyen lesz a **korrelációs együttható**.

Korrelációs együttható: definíció

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

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

Korrelációs együttható: definíció

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

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

- **Lehetséges értékek**. A korrelációs együttható értéke mindig -1 és 1 közé esik:
$$|R(X, Y)| \leq 1.$$
- **Lineáris összefüggés**. Legyen $a > 0$ valós szám, b tetszőleges valós szám. Ekkor
$$R(X, aX + b) = 1 \qquad \text{és} \qquad R(X, -aX + b) = -1.$$
- Tegyük fel, hogy $|R(X, Y)| = 1$. Ekkor léteznek olyan a és b valós számok, hogy az $Y = aX + b$ egyenlet 1 valószínűséggel teljesül. Vagyis a korrelációs együttható lehetséges legnagyobb értékei lineáris összefüggés esetén érhetők el.

Korrelációs együttható: példa

Egy üzletben az A és B újság forgalmát figyelik.

- 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

Egy üzletben az A és B újság forgalmát figyelik.

- 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?

$$R(X+Y, 300X+400Y) = \frac{\operatorname{cov}(X+Y, 300X+400Y)}{D(X+Y)D(300X+400Y)} = \\ = \frac{102000}{D(X+Y)D(300X+400Y)}$$

a korábbi számolás alapján, így a szórásokat kell meghatároznunk.

Korrelációs együttható: példa

X és Y **függetlenek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** szórása:

Korrelációs együttható: példa

X és Y **függetlenek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** szórása:

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

A bevétel szórása:

Korrelációs együttható: példa

X és Y **független**ek, **Poisson-eloszlású**ak, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** szórása:

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

A bevétel szórása:

$$D(300X + 400Y) = \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} =$$
$$= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148,17.$$

Ezek alapján a korrelációs együttható:

Korrelációs együttható: példa

X és Y **függetlenek**, **Poisson-eloszlásúak**, X paramétere 100, az Y-é 180. Ekkor az **eladott példányok számának** szórása:

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

A bevétel szórása:

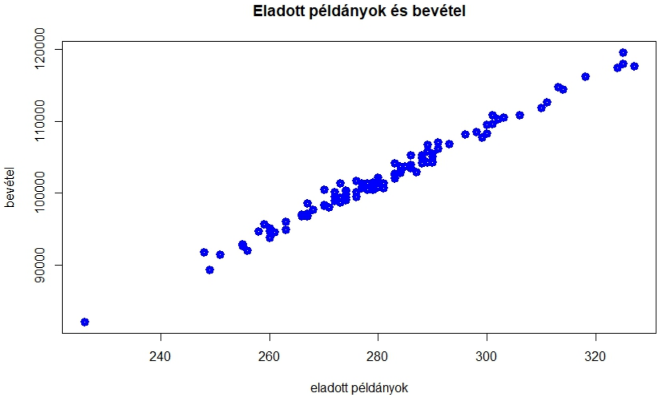
$$D(300X + 400Y) = \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} = \\ = \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148,17.$$

Ezek alapján a korrelációs együttható:

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

A korrelációs együttható lehetséges legnagyobb értéke 1, így ez **erős pozitív korrelációt** jelent.

Korrelációs együttható: példa



A bevétel $(300X + 400Y)$ és az eladott példányszám $(X + Y)$ együttes előfordulása $n = 100$ független 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 **4000**. Mennyi az összesen eladott példányok számának és az ezekből származó bevételnek a korrelációs együtthatója?

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

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

$$D(300X+4000Y) = \sqrt{300^2 D^2(X) + 4000^2 D^2(Y)} =$$

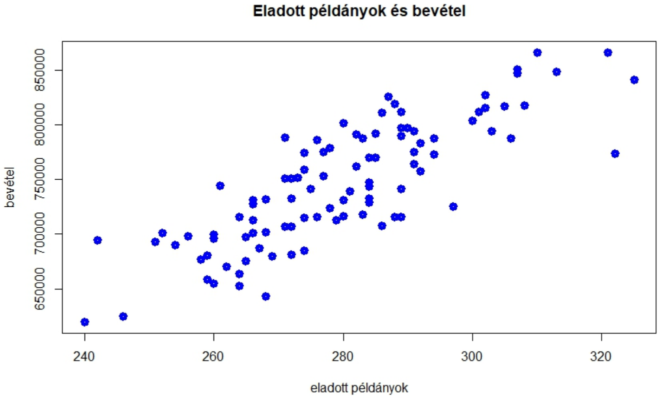
$$= \sqrt{300^2 \cdot 100 + 4000^2 \cdot 180} = 53749,42;$$

$$R(X+Y, 300X+4000Y) = \frac{cov(X+Y, 300X+4000Y)}{D(X+Y)D(300X+4000Y)} =$$

$$= \frac{750000}{16,73 \cdot 53749,42} = 0,83.$$

A korrelációs együttható értéke kisebb, mint hasonló ár esetén.

Korrelációs együttható: példa



A bevétel $(300X + 4000Y)$ és az eladott példányszám $(X + Y)$ együttes előfordulása $n = 100$ megfigyelésből. Kovariancia: 750000, korrelációs együttható: $0,83$.

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 hómennyiség: 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 hómennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- tengerparton töltött idő és egészség:

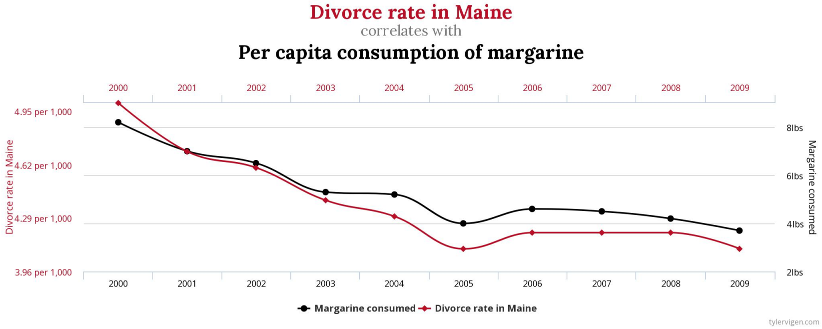
Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hómennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- tengerparton töltött idő és egészség:
  ha van is pozitív korreláció, **nem biztos, hogy van ok-okozati összefüggés**, a tengerparton töltött idő összefügg az anyagi helyzettel, ami az egészséggel, de csak a tengerparttól nem biztos, hogy egészséges lesz valaki, illetve aki beteg, kevésbé megy a tengerpartra
- a válások aránya Maine államban és a fejenkénti margarinfogyasztás az USA-ban:

Korreláció és ok-okozat

- napsütéses órák száma és hőmérséklet: pozitív korreláció, **van ok-okozati összefüggés**
- napsütéses órák száma és hómennyiség: negatív korreláció, van ok-okozati összefüggés
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, de mindkét irányban lehet ok-okozati összefüggés
- tengerparton töltött idő és egészség:
  ha van is pozitív korreláció, **nem biztos, hogy van ok-okozati összefüggés**, a tengerparton töltött idő összefügg az anyagi helyzettel, ami az egészséggel, de csak a tengerparttól nem biztos, hogy egészséges lesz valaki, illetve aki beteg, kevésbé megy a tengerpartra
- a válások aránya Maine államban és a fejenkénti margarinfogyasztás az USA-ban:van pozitív korreláció ($R = 0,9926$), de **feltehetően nincs ok-okozati összefüggés** (forrás és további példák: `http://tylervigen.com/spurious-correlations`)

Korreláció és ok-okozat



A válások aránya Maine államban és a fejenkénti margarinfogyasztás az USA-ban, korrelációs együttható: $0,9926$

`http://tylervigen.com/spurious-correlations`

„Big data" analízis: 200-300 mennyiség között könnyen található néhány olyan pár, amik ok-okozati összefüggés nélkül is nagy pozitív korrelációval rendelkeznek, de olyanok is, amik között valós összefüggés van $\rightarrow$ mindez alaposabb vizsgálatot igényel.

Korrelálatlanság

Ha az X, Y valószínűségi változók **kovarianciája** 0, akkor azt mondjuk, hogy X és Y **korrelálatlanok**. Mi ennek a kapcsolata a **függetlenséggel**?

$\boxed{X \text{ és } Y \textbf{ függetlenek}}$

$\boxed{X \text{ és } Y \textbf{ korrelálatlanok}}$

Korrelálatlanság

Ha az X, Y valószínűségi változók **kovarianciája** 0, akkor azt mondjuk, hogy X és Y **korrelálatlanok**. Mi ennek a kapcsolata a **függetlenséggel**?

X és Y **függetlenek** $\Rightarrow$ / $\not\Leftarrow$ X és Y **korrelálatlanok**

Legyen X és Y két független, szabályos kockadobás eredménye.

$U = X + Y$ **az összeg**

$V = X - Y$ **a különbség**

$\mathrm{cov}(U, V) = \mathrm{cov}(X + Y, X - Y) = D^2(X) - \mathrm{cov}(X, Y) +$
$+\mathrm{cov}(X, Y) - D^2(Y) = 0 \Rightarrow X$ és Y **korrelálatlanok**

Ugyanakkor U és V **nem függetlenek**, például mert

Korrelálatlanság

Ha az X, Y valószínűségi változók **kovarianciája** 0, akkor azt mondjuk, hogy X és Y **korrelálatlanok**. Mi ennek a kapcsolata a **függetlenséggel**?

X és Y **függetlenek** $\Rightarrow$ / $\not\Leftarrow$ X és Y **korrelálatlanok**

Legyen X és Y két független, szabályos kockadobás eredménye.

$U = X + Y$ **az összeg**

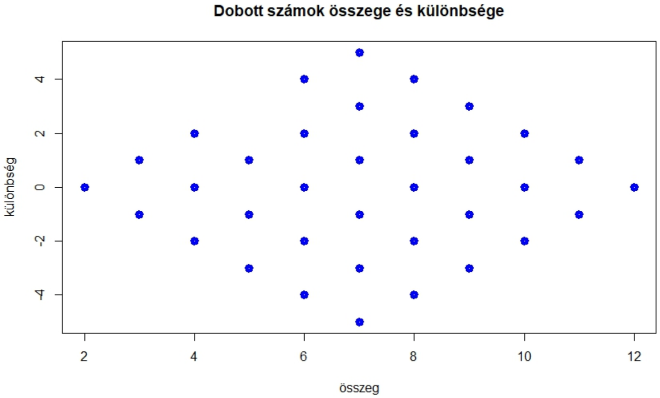
$V = X - Y$ **a különbség**

$\mathrm{cov}(U, V) = \mathrm{cov}(\mathrm{X} + \mathrm{Y}, \mathrm{X} - \mathrm{Y}) = D^2(X) - \mathrm{cov}(X, Y) +$
$+\mathrm{cov}(X, Y) - D^2(Y) = 0 \Rightarrow X$ és Y **korrelálatlanok**

Ugyanakkor U és V **nem függetlenek**, például mert

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

Korrelálatlanság: 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**.

Házi feladat október 22., kedd, 8:15-ig

Legyenek X és Y független Poisson-eloszlású valószínűségi változók, melyekre az igaz, hogy a 0 valószínűsége megegyezik az 1 valószínűségével.

a) Határozzuk meg $2X - 3Y$ várható értékét és szórását.

b) Sorsoljunk x és y vektorokat az R-ben, melyek 100 elemből állnak, és minden koordinátájuk független, Poisson-eloszlású, ami a fenti feltételt teljesíti. Legyen $z = 2 * x - 3 * y$. Készítsünk z-ből hisztogramot, és számítsuk ki a koordinátáinak átlagát és korrigált tapasztalati szórását (`sd`).