

Kovariancia és korrelációs együttható

Két véletlen mennyiség, valószínűségi változó kapcsolatának megértése hasznos lehet összetett társadalmi, gazdasági, természeti folyamatok elemzésekor. Például ha tudjuk, hogy ha az egyiknek nagyobb az értéke, akkor tipikusan a másiknak is nagyobb, és az egyik hamarabb mérhető, akkor annak a növekedése előrejelezheti a másik mennyiség növekedését (például ruhavásárlások jelezhetik előre a gazdasági válságok végét).

Ugyanennek az időben lezajló véletlen folyamatok (idősorok) modellezésében is kulcsszerepe van: nem mindegy, hogy például a tőzsdeindex honlapi értéke milyen erősen függ össze a maival vagy a korábbiakkal.

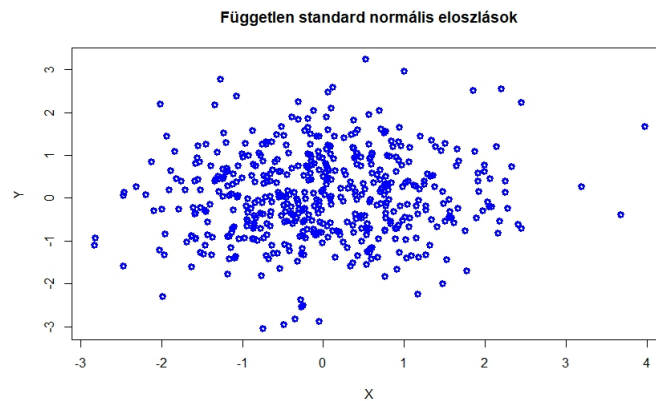
Két valószínűségi változó lehet

- **független**: például két találmásra választott ember jövedelme, vagy,
- **nem független**: például egy találmásra választott ember jövedelme most, illetve fél év múlva

Az **összefüggőség mértéke** különböző lehet:

- egy találmásra választott felnőtt életkora és jövedelme „erősen összefüggő”, a fiataloké és időseké általában alacsonyabb;
- egy találmásra 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 **kovariancia** és a **korrelációs együttható**. Ez utóbbinak a „nagy” értékei „erős, lineáris jellegű” összefüggésre utalnak: minél nagyobb az egyik mennyiség, annál nagyobb a másik is tipikusan, és a kapcsolat közelítőleg pozitív főegyütthatós, $Y = aX + b$ egyenlettel írható le. A korrelációs együttható negatív értékeket is felvehet, ez negatív főegyütthatós, közelítőleg $Y = aX + b$ jellegű kapcsolatot ír le, ahol tehát $a < 0$, és ezért minél nagyobb az egyik mennyiség, tipikusan annál kisebb a másik.

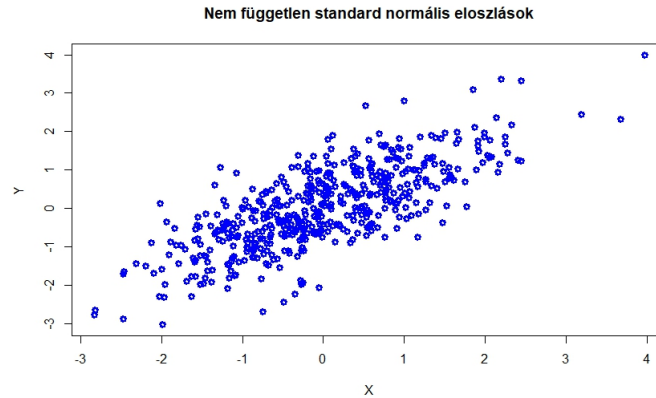


1. ábra. Független normális eloszlások

Az 1. ábrán 500 darab véletlen pontot láthatunk, melyek koordinátái **független** standard normális eloszlásúak. A koordináták között nincs kapcsolat: a kovariancia és a korrelációs együttható is **0** lesz.

A 2. ábrán 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 , „tipikusan” annál nagyobb $(X + Z)/\sqrt{2}$ is \rightarrow ennek megfelelően a két koordináta közötti **kovariancia** és **korrelációs együttható** is **pozitív** lesz.



2. ábra. Az X és $(X + Z)/\sqrt{2}$ együttes eloszlása, ahol X, Z független, 0 várható értékű, 1 szórású normális eloszlású valószínűségi változók

1. A kovariancia definíciója és tulajdonságai

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

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

Ezt a következőképpen alakíthatjuk át. Bontsuk fel a zárójelet, használjuk, hogy összeg várható értéke a várható értékek összege, és hogy az $\mathbb{E}(X), \mathbb{E}(Y)$ konstansok kiemelhetők a várható értékből:

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))) = \mathbb{E}(XY - X \cdot \mathbb{E}(Y) - \mathbb{E}(X) \cdot Y + \mathbb{E}(X)\mathbb{E}(Y)) = \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

- **A kovariancia kiszámítása:**

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

- **Szimmetria.** $\text{cov}(X, Y) = \text{cov}(Y, X)$, világos, ha a definícióban felcseréljük X -et és Y -t, ugyanazt kapjuk.
- **Kapcsolat a szórásnégyzettel.** $\text{cov}(X, X) = D^2(X)$, hiszen ha Y helyére is X -et írunk, a szórásnégyzet alakjait kapjuk: $\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.
- **Függetlenséggel való kapcsolat.** Ha az X és Y valószínűségi változók **függetlenek**, akkor $\text{cov}(X, Y) = 0$, hiszen korábban (diszkrét valószínűségi változókra) beláttuk, hogy függetlenség esetén a szorzat várható értéke a várható értékek szorzata.

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

- Konstanssal való kovariancia. $\text{cov}(X, c) = 0$, ha c valós szám, hiszen a konstans mindentől független.
- **Linearitás.** Egyrészt

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

hiszen összeg várható értéke a várható értékek összege, és így

$$\mathbb{E}((X + Y)Z) - \mathbb{E}(X + Y)\mathbb{E}(Z) = \mathbb{E}(XZ) + \mathbb{E}(YZ) - \mathbb{E}(X)\mathbb{E}(Z) - \mathbb{E}(Y)\mathbb{E}(Z),$$

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

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

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

- A kovariancia értékészlete. A kovariancia értéke tetszőleges lehet, hiszen például c tetszőlegesen választható.
- **Összeg szórásnégyzete.** $D^2(X + Y) = D^2(X) + D^2(Y) + 2\text{cov}(X, Y)$. Továbbá

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

Ebből az elsőt látjuk be:

$$\begin{aligned} D^2(X + Y) &= \mathbb{E}((X + Y)^2) - (\mathbb{E}(X + Y))^2 = \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2 = \\ &= D^2(X) + D^2(Y) + 2\text{cov}(X, Y). \end{aligned}$$

Több tagra hasonlóan bizonyítható az összefüggés. Ebből az is látható, hogy független esetben valóban a szórásnégyzetek összege lesz az összeg szórásnégyzete.

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

Ez az előzőhöz hasonlóan bizonyítható.

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 $\text{cov}(X + Y, 300X + 400Y)$?

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

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

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

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

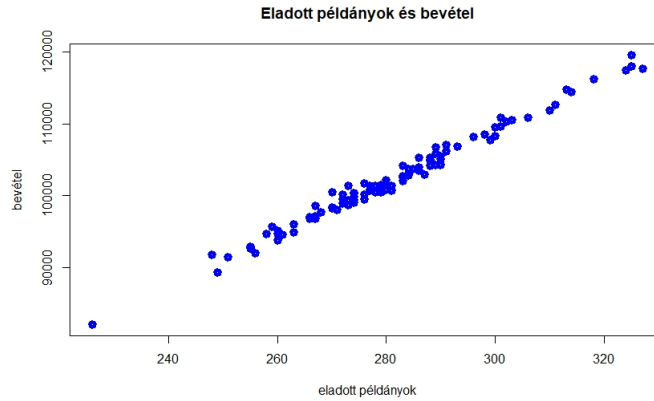
(c) egy λ paraméterű **Poisson-eloszlású** valószínűségi változó **szórásnégyzete** λ .

2. Korrelációs együttható

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 : példányszám Y : bevétel forintban Z : bevétel ezer forintban,



3. ábra. A bevétel ($300X + 400Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ független megfigyelésből

akkor

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

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**.

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

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

2.1. Állítás (A korrelációs együttható tulajdonságai). (a) **Lehetséges értékek.** A korrelációs együttható értéke mindig -1 és 1 közé esik:

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

(b) **Lineáris összefüggés.** Legyen $a > 0$ valós szám, b tetszőleges valós szám. Ekkor

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

(c) 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.

Ennek az állításnak a harmadik része alapján mondhatjuk, hogy a korrelációs együttható 1 -hez közeli értékei erős, pozitív főegyütthatós lineárishoz közeli kapcsolatot jelentenek, míg a -1 -hez közeli értékek szintén erős, negatív főegyütthatós, lineárishoz közeli kapcsolatra utalnak. Ugyanakkor a nem lineáris jellegű összefüggéseket a korrelációs együttható nem mindig mutatja ki, lehetséges, hogy az egyik érték egyértelműen megkapható a másiktól, mégis 0 körüli a korrelációs együttható.

3. Példák korrelációs együtthatóra

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?

$$\begin{aligned} R(X + Y, 300X + 400Y) &= \frac{\text{cov}(X + Y, 300X + 400Y)}{D(X + Y)D(300X + 400Y)} = \\ &= \frac{102000}{D(X + Y)D(300X + 400Y)} \end{aligned}$$

a korábbi számolás alapján, így a szórásokat kell meghatároznunk.

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 (hiszen $300X$ és $400Y$ is függetlenek):

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

Ezek alapján a korrelációs együttható:

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

A korrelációs együttható lehetséges legnagyobb értéke **1**, így ez **erős pozitív korrelációt** jelent. A 3. ábrán láthatjuk is, hogy valóban lineáris jellegű, pozitív meredekségű kapcsolat van a két mennyiség között.

Nézzük meg, hogy hogyan változik a korrelációs együttható, ha az egyik újság árát változtatjuk.

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?

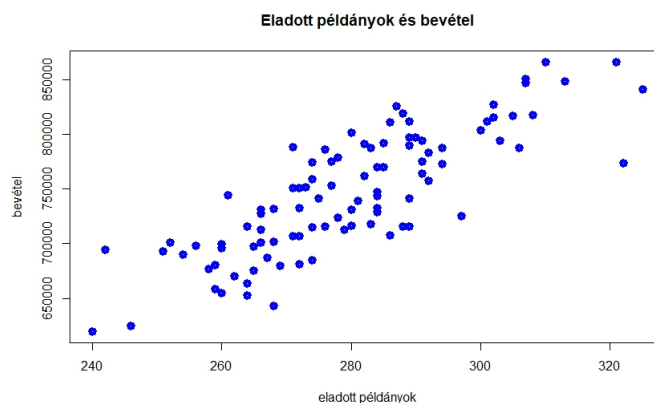
$$\begin{aligned} \text{cov}(X + Y, 300X + 4000Y) &= 300 \cdot 100 + 4000 \cdot 180 = 750000; \\ D(X + Y) &= \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73; \\ D(300X + 4000Y) &= \sqrt{300^2 D^2(X) + 4000^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 4000^2 \cdot 180} = 53749,42; \\ R(X + Y, 300X + 4000Y) &= \frac{\text{cov}(X + Y, 300X + 4000Y)}{D(X + Y)D(300X + 4000Y)} = \frac{750000}{16,73 \cdot 53749,42} \\ &= 0,83. \end{aligned}$$

A kovariancia több lett, a a tipikus értékek is nagyobbak lettek. A korrelációs együttható értéke kisebb, mint abban az esetben, amikor az újságok ára közel egyforma volt. A 4. ábrán láthatjuk, hogy továbbra is van pozitív irányú összefüggés, de kevésbé illeszkednek egy egyenesre a pontok, és kevésbé határozza meg egyik mennyiség a másikat. Ez közepesen erős összefüggésnek mondható.

Nézzük meg pontosabban is, hogy hogyan alakul a korrelációs együtthatónak a második újság árától való függése. Legyen a második újság ára c (eddig a $c = 400$ és $c = 4000$ eseteket néztük).

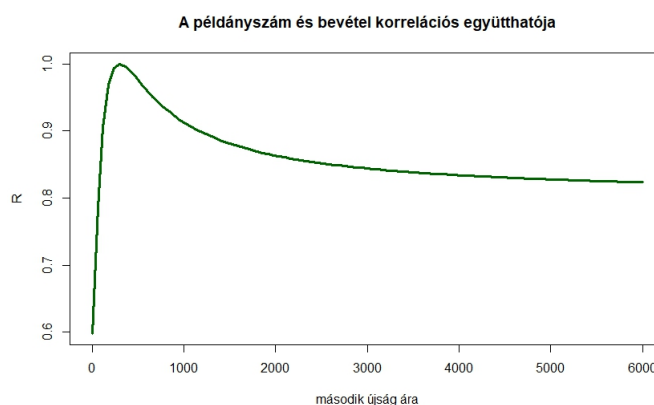
A fenti számolásokhoz hasonlóan

$$R(X + Y, 300X + cY) = \frac{\text{cov}(X + Y, 300X + cY)}{D(X + Y)D(300X + cY)} = \frac{30000 + c \cdot 180}{16,73 \cdot \sqrt{9000000 + c^2 \cdot 180}}.$$



4. ábra. A bevétel ($300X + 4000Y$) és az eladott példányszám ($X + Y$) együttes előfordulása $n = 100$ megfigyelésből

Az 5. ábrán látható ez a görbe. Ha $c = 300$, a két újság ugyanannyiba kerül, ilyenkor a példányszám pontosan meghatározza a bevételt, és az összefüggés lineáris ($300(X + Y)$ lineáris függvénye $X + Y$ -nak), ilyenkor a korrelációs együttható értéke a lehető legnagyobb, azaz 1. Bármilyen más esetben ennél kisebb értéket kapunk.



5. ábra. A bevétel ($300X + cY$) és az eladott példányszám ($X + Y$) korrelációs együtthatója c függvényében

Példa.

Tegyük fel, hogy X és Y **független**, 4 szórású Poisson-eloszlású valószínűségi változók. Számítsuk ki X és $-2X + Y$ korrelációs együtthatóját.

$$\begin{aligned}
 R(X, -2X + Y) &= \frac{\text{cov}(X, -2X + Y)}{D(X)D(-2X + Y)} = \frac{(-2) \cdot \text{cov}(X, X) + \text{cov}(X, Y)}{D(X)D(X + Y)} = \\
 &= \frac{-2D^2(X)}{D(X)D(-2X + Y)} = \frac{-2D^2(X)}{D(X) \cdot \sqrt{D^2(-2X) + D^2(Y)}} = \\
 &= \frac{-2D^2(X)}{D(X) \cdot \sqrt{((-2)^2 + 1) \cdot D^2(X)}} = \frac{(-2) \cdot 4^2}{4 \cdot \sqrt{5} \cdot 4} = -\frac{2}{\sqrt{5}} = -0,89.
 \end{aligned}$$

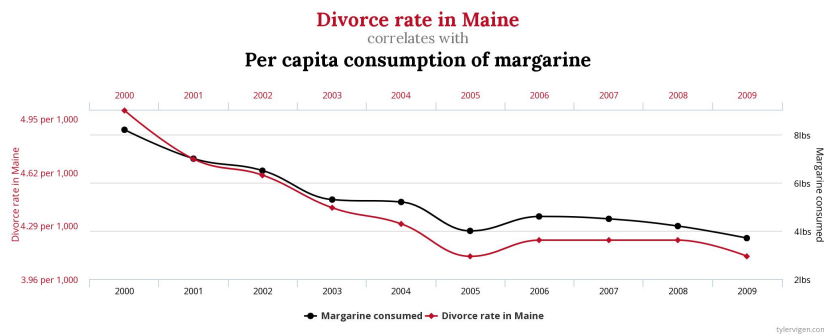
Közepesen erős negatív korrelációt kaptunk.

Itt $\text{cov}(X, Y) = 0$, mert X és Y **függetlenek**. A függetlenséget az összeg szórásának kiszámításakor is használtuk, viszont X -ről és Y -ről ezen kívül valójában elég lett volna annyit feltenni, hogy **azonos a szórásuk**, azaz $D(X) = D(Y)$ – sem a Poisson-eloszlástól, sem a 4-től nem függ a végeredmény. Azt is használtuk, hogy

$$D^2(cX) = c^2 D^2(X) \quad \Leftrightarrow \quad D(cX) = |c| D(X).$$

4. Korreláció és ok-okozat viszonya

- napsütéses órák száma és hőmérséklet: pozitív korreláció (kivéve néha télen), **van ok-okozati összefüggés**
- napsütéses órák száma és hó mennyiség: negatív korreláció, **van ok-okozati összefüggés**
- anyagi helyzet és iskolai végzettség: van pozitív korreláció, és **mindkét irányban lehet ok-okozati összefüggés**
- kérdezzünk meg egy véletlenszerűen választott embert, hogy mennyi időt töltött idén a tengerparton, és jellemezzük az egészségét is egy számmal (minél nagyobb, annál egészségesebb az illető):
ha van is pozitív korreláció, **nem biztos, hogy van ok-okozati összefüggés** a tengerparton töltött idő összefügg az anyagi helyzettel, ami az egészséggel, tehát nem következtethetünk arra, hogy közvetlenül ez okozza az egészséget, csak a tengerparttól nem biztos, hogy egészséges lesz valaki. Már csak azért sem, mert aki beteg, kevésbé megy a tengerpartra.
- a válások aránya Maine államban és a fejenkénti margarinfogyasztás az USA-ban: van pozitív korreláció ($R = 0,9926$, a 6. ábra is ezt mutatja), de **feltehetően nincs ok-okozati összefüggés**, legalábbis akkora nincs, ami ilyen magas pozitív korrelációs okozna (a kép forrása és további példák: <http://tylervigen.com/spurious-correlations>)



6. ábra. 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 → mindez alaposabb vizsgálatot igényel.

Összefoglalva: önmagában a korrelációs együtthatóból nem tudunk következtetést levonni arra nézve, hogy bármelyik irányban is pozitív összefüggés lenne a vizsgált mennyiségek között.

5. Korrelálatlanság

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

Kérdés, hogy mi ennek a kapcsolata a **függetlenséggel**.



Legyen X és Y két független, szabályos kockadobás eredménye.

$U = X + Y$ az összeg

$V = X - Y$ a különbség

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

Tehát U és V **korrelálatlanok**.

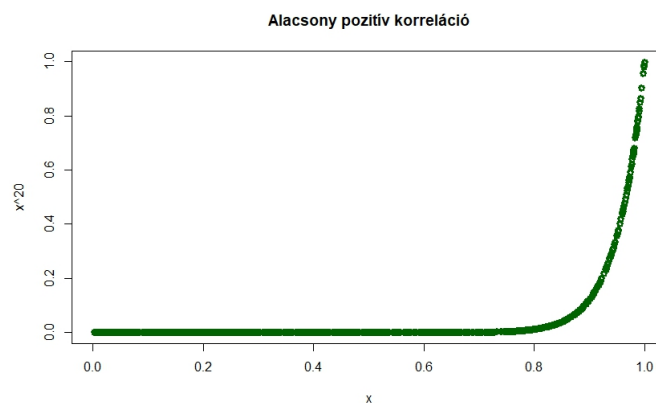
Ugyanakkor U és V **nem függetlenek**, például mert

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



7. ábra. A dobott számok **különbségének** ($X - Y$) és a dobott számok **összegének** ($X + Y$) együttes előfordulása 100 megfigyelésből

Azt, hogy a korrelálatlanság vagy alacsony korreláció nem jelent függetlenséget, az alábbi példán is megnézhetjük (8. ábra). Itt a $(0, 1)$ intervallumból sorsoltunk számokat, ez X , és ebből az $Y = X^{20}$ összefüggéssel számoltuk ki a másik valószínűségi változót. Így Y monoton növvő, determinisztikus függvénye X -nek, ez tehát erős pozitív irányú összefüggést jelent: minél nagyobb X , annál nagyobb Y . Mivel azonban az összefüggés nem lineáris jellegű, a korrelációs együttható értéke csak 0,5 körüli – ez ugyan pozitív, de nincs közel a lehetséges legnagyobb értékhez, 1-hez, ez közepes erősségű összefüggést jelentene csak. Hasonlóképpen még a 0 körüli korrelációs együttható is jelenthet erős, csak nem lineáris jellegű összefüggést.



8. ábra. Az X^{20} függvény

Ezt a hátrányt például a **rangkorreláció** (Spearman-korreláció) használatával lehet orvosolni: ott sorba rendezzük az értékeket az X és Y nagysága szerint, azt nézzük, hogy az egyes megfigyelések hányadikak az X , illetve Y szerinti sorrendben, de az nem számít, hogy pontosan mekkorák a felvett értékek. Ezután a kapott rangokra (nagyság szerinti sorrendben elfoglalt helyekre) számítják ki a korreláció tapasztalati becslését, úgy, mintha a korrelációs együtthatót becsülnénk.

Házi feladat október 22., péntek, 12:15-ig Anna és Bálint háromszor dobnak egy szabályos dobókockával. Ha a dobott számok összege páros, Anna nyer 200 forintot, ha a dobott számok összege osztható hárommal, Bálint nyer 300 forintot. Legyen X Anna nyereménye, Y pedig Bálint nyereménye. Számítsuk ki X és $X + Y$ korrelációs együtthatóját.