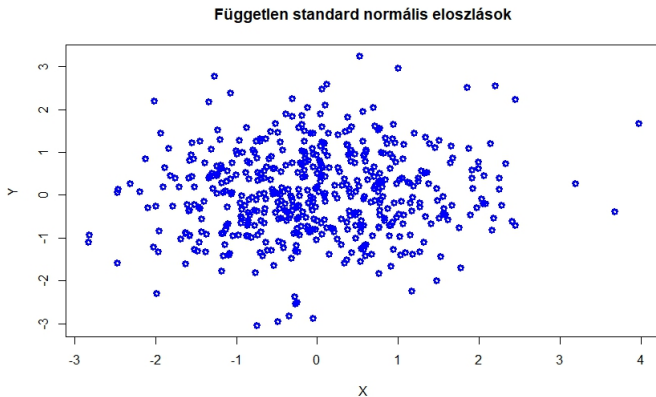


Kovariancia és korrelációs együttható (7. előadás)

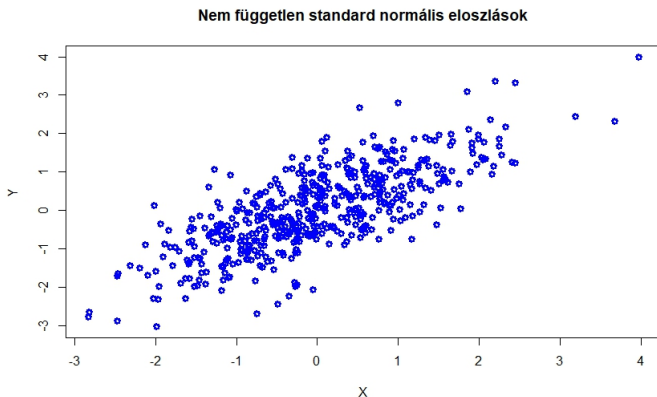
- két valószínűségi változó lehet
 - ▶ **független**: például két találmásra választott ember jövedelme, vagy,
 - ▶ **nem független**: például egy találmásra választott ember jövedelme most, illetve fél év múlva
- az **összefüggőség mértéke** különböző lehet:
 - ▶ egy találmásra választott felnőtt életkora 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.

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.

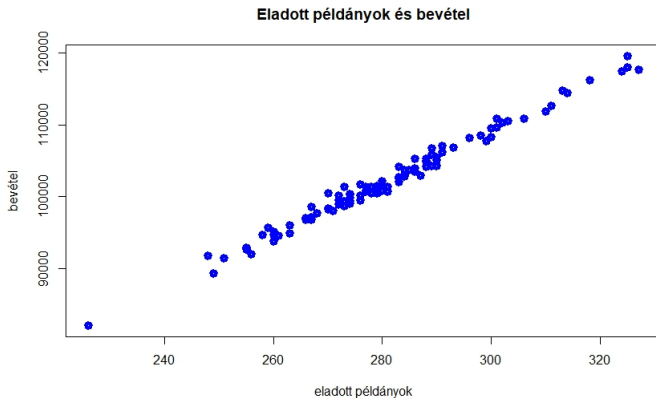
Pozitív korreláció



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**:

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

A kovariancia

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

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

- **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

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

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

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

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

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

- **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

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

- 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.

A kovariancia tulajdonságai

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

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

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

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

- **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

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 **bevé-
telnek** 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 $\text{cov}(X + Y, 300X + 400Y)$?

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

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

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

ahol felhasználtuk, hogy

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:

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

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:

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

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

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

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:

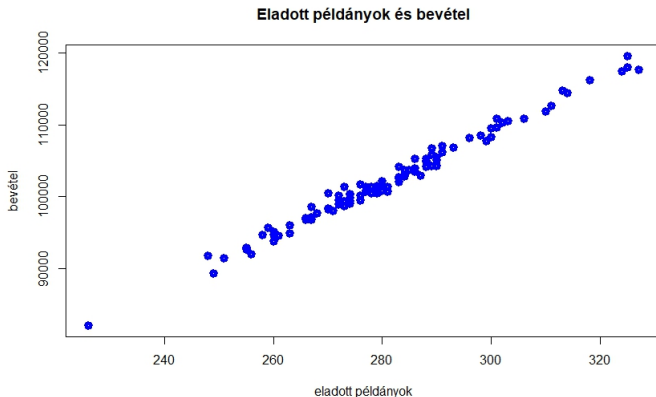
$$\begin{aligned}\text{cov}(X + Y, 300X + 400Y) &\stackrel{(a)}{=} \text{cov}(X, 300X) + \text{cov}(X, 400Y) + \\ &\quad + \text{cov}(Y, 300X) + \text{cov}(Y, 400Y) = \\ &\stackrel{(a,b)}{=} 300 \cdot \text{cov}(X, X) + 400 \cdot \text{cov}(Y, Y) = \\ &\stackrel{(b)}{=} 300D^2(X) + 400D^2(Y) = \\ &\stackrel{(c)}{=}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

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 .

Kovariancia: példa



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

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

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**.

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**:

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

- **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

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

- 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 **bevételek** 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.

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 **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:

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

Ezek alapján a korrelációs együttható:

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:

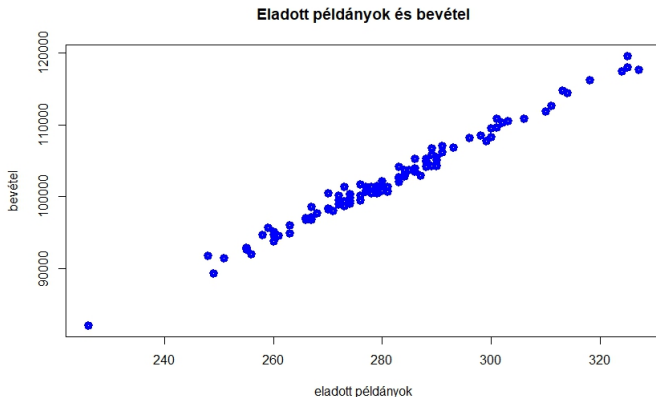
$$\begin{aligned} D(300X + 400Y) &= \sqrt{300^2 D^2(X) + 400^2 D^2(Y)} = \\ &= \sqrt{300^2 \cdot 100 + 400^2 \cdot 180} = 6148,17. \end{aligned}$$

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.

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?

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

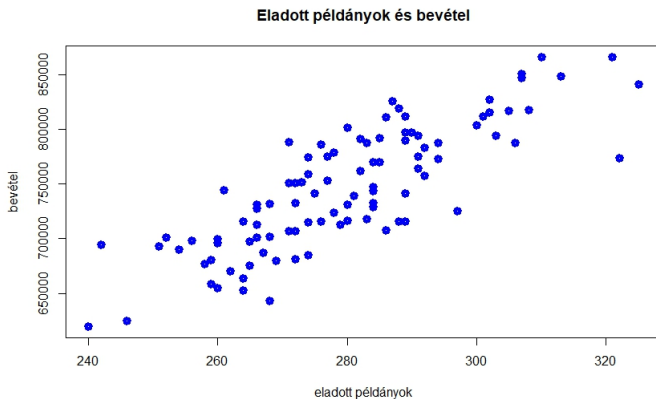
$$D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{100 + 180} = 16,73;$$

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

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

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 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
- tengerparton töltött idő és egészség:

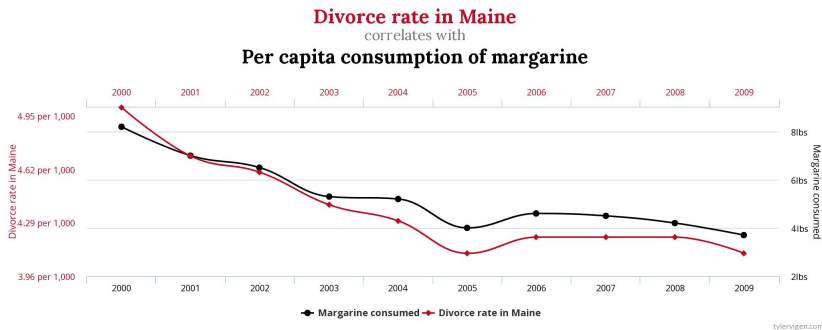
Korreláció és ok-okozat

- 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
- 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 hőmennyiség: 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>)

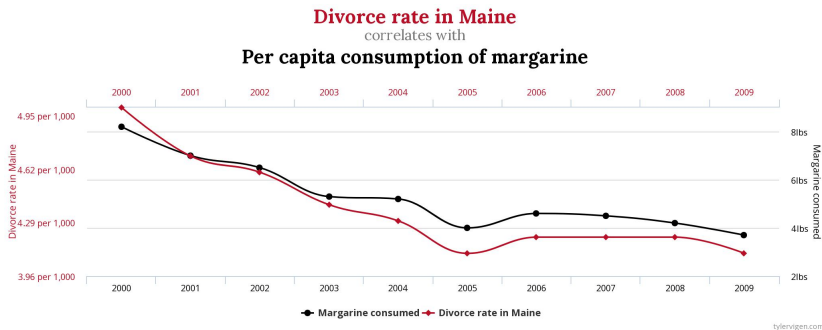
Korreláció és ok-okozat



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

<http://tylervigen.com/spurious-correlations>

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 → mindez alaposabb vizsgálatot igényel.

Korrelátlanság

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

X és Y **függetlenek**

X és Y **korrelátlanak**

Korrelátlanság

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**



X és Y **korrelálatlanok**

Legyen X és Y két független, szabályos kockadobás eredménye.

$U = X + Y$ az **összeg**

$V = X - Y$ a **különbség**

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

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

Korrelátlanság

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**



X és Y **korrelálatlanok**

Legyen X és Y két független, szabályos kockadobás eredménye.

$U = X + Y$ az **összeg**

$V = X - Y$ a **különbség**

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

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

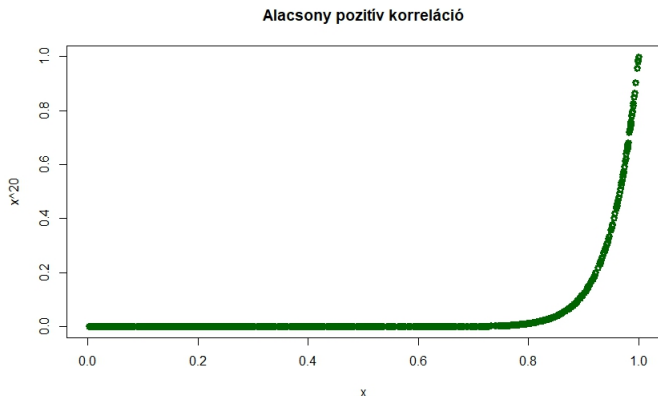
$$0 = \mathbb{P}(U = 11, V = 0) \neq \mathbb{P}(U = 11) \cdot \mathbb{P}(V = 0) = \frac{2}{36} \cdot \frac{1}{6}$$

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**.

Rangkorreláció



Legyen X egyenletes eloszlású a $(0,1)$ intervallumból, $Y = X^{20}$. A korrelációs együttható értéke 0,5 körüli, pedig szoros összefüggés van.

Rangkorreláció

Legyenek $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ megfigyelések.

Rangkorreláció

Legyenek $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ megfigyelések.

Rendezzük sorba a két mintát külön-külön nagyság szerint, és írjuk fel, hogy az egyes megfigyelések hányadikak a sorba rendezett mintában. Ezek lesz az egyes megfigyelések rangja.

Például:

$$X_1 = 650, X_2 = 870, X_3 = 720 \quad \Rightarrow \quad (3, 1, 2)$$

$$Y_1 = 18, Y_2 = 15, Y_3 = 17 \quad \Rightarrow \quad (1, 3, 2)$$

Rangkorreláció

Legyenek $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ megfigyelések.

Rendezzük sorba a két mintát külön-külön nagyság szerint, és írjuk fel, hogy az egyes megfigyelések hányadikak a sorba rendezett mintában. Ezek lesz az egyes megfigyelések rangja.

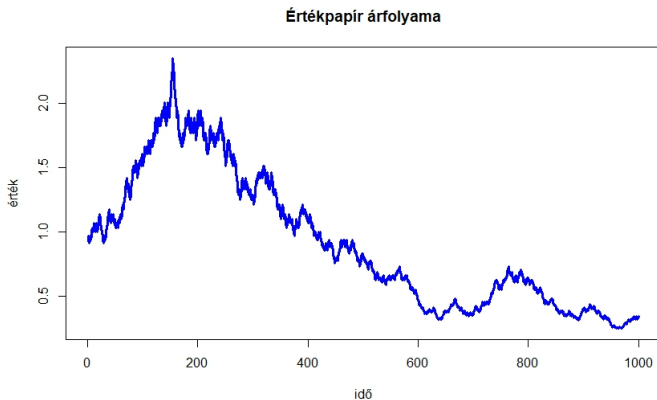
Például:

$$X_1 = 650, X_2 = 870, X_3 = 720 \quad \Rightarrow \quad (3, 1, 2)$$

$$Y_1 = 18, Y_2 = 15, Y_3 = 17 \quad \Rightarrow \quad (1, 3, 2)$$

Számítsuk ki az így kapott két adatsor, vagyis a rangok korrelációs együtthatóját. Ez a **rangkorreláció** (Spearman-korreláció).

Eloszlásfüggvény: példa



Egy elképzelt értékpapír árfolyama 1000 napon keresztül, 1000 forintban

Eloszlásfüggvény: bevezetés

- X valószínűségi változó: egy véletlen kísérlet eredménye
- eddig: X **diszkrét**, és a $\mathbb{P}(X = x)$ valószínűségekkel lehet leírni az eloszlását
- ha a lehetséges értékek halmaza „túl nagy”, vagy a valószínűségek „túl kicsik”, ez nem informatív
- például: X az értékpapír árfolyama holnap, $\mathbb{P}(X = 784) = 0,0038$, $\mathbb{P}(X = 785) = 0,004$, stb. egy előrejelzés szerint \rightarrow ennél hasznosabb információ lehet, hogy

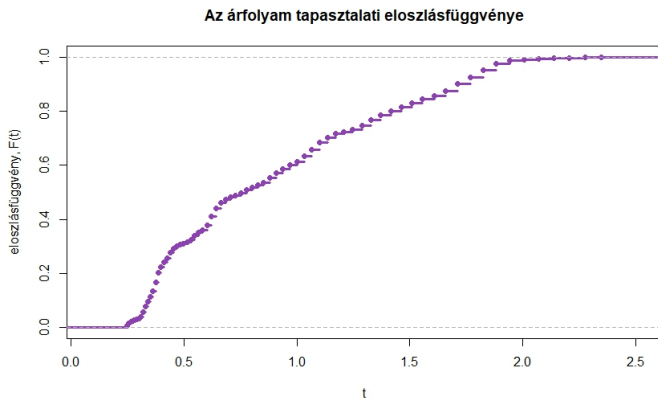
$$\mathbb{P}(X \leq 785) = 0,5,$$

azaz az értékpapír 50% valószínűséggel nem haladja meg a 785 szintet

- **eloszlásfüggvény**: $F(t)$ annak valószínűsége, hogy **a valószínűségi változó értéke legfeljebb t** , azaz

$$F(t) = \mathbb{P}(X \leq t).$$

Eloszlásfüggvény: példa



t függvényében a t -nél nem nagyobb árfolyamú napok aránya az előző példában

Eloszlásfüggvény: definíció

Legyen $X : \Omega \rightarrow \mathbb{R}$ valószínűségi változó. Ekkor X **eloszlásfüggvénye** az alábbi $F : \mathbb{R} \rightarrow [0, 1]$ függvény:

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\})$$

tetszőleges $t \in \mathbb{R}$ valós számra.

Ez **minden valószínűségi változóra** és minden $t \in \mathbb{R}$ valós számra értelmes: éppen úgy definiáltuk a valószínűségi változót, hogy $\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{A}$ egy esemény, tehát van valószínűsége.

Eloszlásfüggvény: példa

Valakinek három gyereke születik, a gyerekek mindegyike egymástól függetlenül $1/2$ valószínűséggel fiú. Nyolc egyformán valószínű eset van:

$$\{\text{LLL, FLL, LFL, LLF, FFL, FLF, LFF, FFF}\}$$

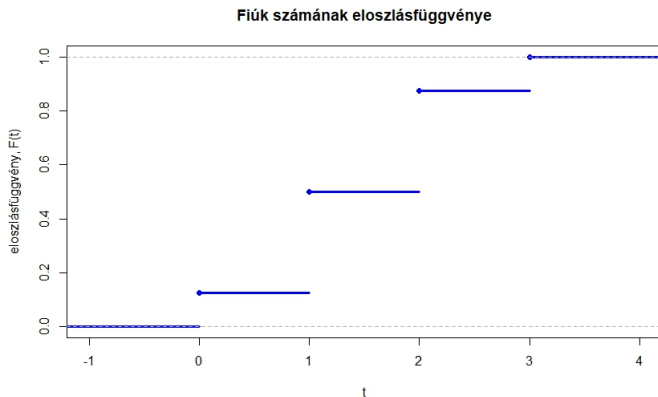
Legyen X a fiúk száma. X diszkrét valószínűségi változó, lehetséges értékei: 0, 1, 2, 3, és

$$\mathbb{P}(X = 0) = \frac{1}{8}, \quad \mathbb{P}(X = 1) = \frac{3}{8}, \quad \mathbb{P}(X = 2) = \frac{3}{8}, \quad \mathbb{P}(X = 3) = \frac{1}{8}.$$

Az X **eloszlásfüggvényének**, F -nek az értéke néhány helyen:

$$\begin{aligned} F(0) &= \mathbb{P}(X \leq 0) = \frac{1}{8}; & F(1) &= \mathbb{P}(X \leq 1) = \frac{1}{2}; \\ F(2, 4) &= \mathbb{P}(X \leq 2, 4) = \frac{7}{8}; & F(4) &= \mathbb{P}(X \leq 4) = 1. \end{aligned}$$

Eloszlásfüggvény: példa



Három gyerek közül a fiúk számának eloszlásfüggvénye vízszintes: t , függőleges: $F(t) = \mathbb{P}(X \leq t)$.

Eloszlásfüggvény

Definíció (Eloszlásfüggvény)

Legyen $X : \Omega \rightarrow \mathbb{R}$ valószínűségi változó. Ekkor X eloszlásfüggvénye az alábbi $F : \mathbb{R} \rightarrow [0, 1]$ függvény:

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\}) \quad \text{minden } t \in \mathbb{R} \text{ valós számra.}$$

Az eloszlásfüggvény minden t valós számhoz hozzárendeli, hogy mennyi annak valószínűsége, hogy a valószínűségi változó értéke legfeljebb t . Például ha X a fiúk száma három gyerek közül:

$$F(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(\text{legfeljebb egy fiú van}) = 1/2;$$

$$F(2) = \mathbb{P}(X \leq 2) = \mathbb{P}(\text{legfeljebb két fiú van}) = 7/8;$$

$$F(2,3) = \mathbb{P}(X \leq 2,3) = \mathbb{P}(\text{legfeljebb 2,3 fiú van}) = 7/8;$$

Eloszlásfüggvény

Definíció (Eloszlásfüggvény)

Legyen $X : \Omega \rightarrow \mathbb{R}$ valószínűségi változó. Ekkor X eloszlásfüggvénye az alábbi $F : \mathbb{R} \rightarrow [0, 1]$ függvény:

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\}) \quad \text{minden } t \in \mathbb{R} \text{ valós számra.}$$

Az eloszlásfüggvény minden t valós számhoz hozzárendeli, hogy mennyi annak valószínűsége, hogy a valószínűségi változó értéke legfeljebb t . Például ha X a fiúk száma három gyerek közül:

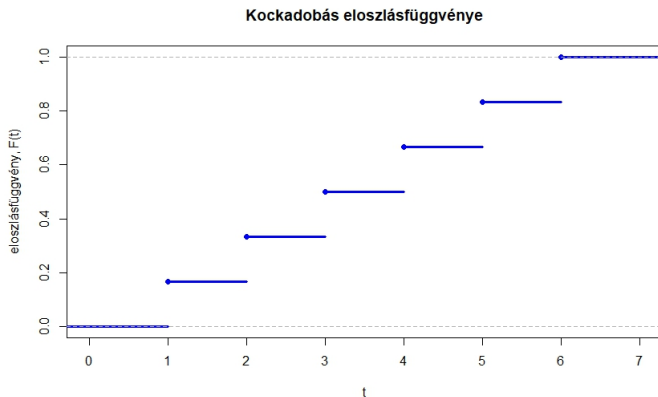
$$F(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(\text{legfeljebb egy fiú van}) = 1/2;$$

$$F(2) = \mathbb{P}(X \leq 2) = \mathbb{P}(\text{legfeljebb két fiú van}) = 7/8;$$

$$F(2,3) = \mathbb{P}(X \leq 2,3) = \mathbb{P}(\text{legfeljebb 2,3 fiú van}) = 7/8;$$

Véges értékészletű valószínűségi változók esetén az eloszlásfüggvény lépcsős (véges sok értéket vesz fel), és az ugrások nagyságát az egyes lehetséges értékek valószínűségei adják meg.

Eloszlásfüggvény: példa



Szabályos dobókockával dobott szám eloszlásfüggvénye
vízszintes: t , függőleges: $F(t) = \mathbb{P}(X \leq t)$.

Az eloszlásfüggvény tulajdonságai

Ha $a, b \in \mathbb{R}$ valós számok, és F az X eloszlásfüggvénye, akkor

$$\mathbb{P}(a < X \leq b) = F(b) - F(a),$$

hiszen annak valószínűségét, hogy X az a és b közé esik, megkaphatjuk úgy, hogy $\mathbb{P}(X \leq b)$ -ből levonjuk $\mathbb{P}(X \leq a)$ -t.

Legyen F egy tetszőleges valószínűségi változó eloszlásfüggvénye. Ekkor

- i) F monoton növekvő: $a < b$ esetén $F(a) \leq F(b)$.
- ii) $\lim_{t \rightarrow -\infty} F(t) = 0$; $\lim_{t \rightarrow \infty} F(t) = 1$.
- iii) F jobbról folytonos, azaz minden $t \in \mathbb{R}$ valós számra $\lim_{s \rightarrow t+} F(s) = F(t)$.

Fordítva: ha F -re érvényesek ezek a tulajdonságok, akkor van olyan X , aminek F az eloszlásfüggvénye.

Eloszlásfüggvény: példa

Legyen X négy rendű $1/2$ paraméterű binomiális eloszlású valószínűségi változó.
Mennyi X eloszlásfüggvényének az értéke az 1,5 helyen?

Eloszlásfüggvény: példa

Legyen X négy rendű $1/2$ paraméterű binomiális eloszlású valószínűségi változó. Mennyi X eloszlásfüggvényének az értéke az $1,5$ helyen?

X -re a következőképpen gondolhatunk: $n = 4$ független kísérlet, mindegyik $p = 0,5$ valószínűséggel sikerül, X a **sikeres kísérletek száma**.

Definíció szerint, ha F az X eloszlásfüggvénye, akkor

$$F(1,5) = \mathbb{P}(X \leq 1,5) = \mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1),$$

hiszen X értéke nemnegatív egész. Így

$$F(1,5) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \left(\frac{1}{2}\right)^4 + 4 \cdot \left(\frac{1}{2}\right)^4 = \frac{5}{16}.$$

Egyenletes eloszlás

Tekintsük a következő hétköznapi példát.

- Csomagot várunk, amit a futár véletlen Y időpontban hoz ki.
- Feltételezzük, hogy Y egyenletes eloszlású a $[8, 12]$ intervallumon (órában mérve).
- Mennyi a valószínűsége, hogy a futár 11 óráig megérkezik?
- **Feltéve, hogy a futár 10 óráig még nem érkezett meg, mennyi a valószínűsége, hogy 11 óra előtt megérkezik?**

Legyen X a futár érkezésének időpontja. Így fogunk tudni számolni:

$$\mathbb{P}(X \leq 11)$$

Egyenletes eloszlás

Tekintsük a következő hétköznapi példát.

- Csomagot várunk, amit a futár véletlen Y időpontban hoz ki.
- Feltételezzük, hogy Y egyenletes eloszlású a $[8, 12]$ intervallumon (órában mérve).
- Mennyi a valószínűsége, hogy a futár 11 óráig megérkezik?
- **Feltéve, hogy a futár 10 óráig még nem érkezett meg, mennyi a valószínűsége, hogy 11 óra előtt megérkezik?**

Legyen X a futár érkezésének időpontja. Így fogunk tudni számolni:

$$\mathbb{P}(X \leq 11) = \frac{11 - 8}{12 - 8} = \frac{3}{4} = 75\%.$$

$$\mathbb{P}(X \leq 11 | X > 10)$$

Egyenletes eloszlás

Tekintsük a következő hétköznapi példát.

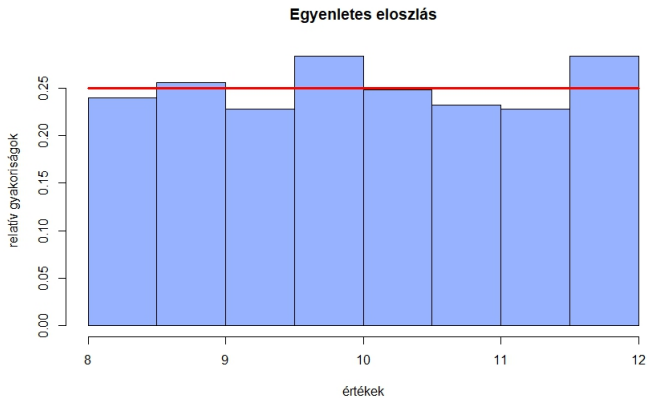
- Csomagot várunk, amit a futár véletlen Y időpontban hoz ki.
- Feltételezzük, hogy Y egyenletes eloszlású a $[8, 12]$ intervallumon (órában mérve).
- Mennyi a valószínűsége, hogy a futár 11 óráig megérkezik?
- **Feltéve, hogy a futár 10 óráig még nem érkezett meg, mennyi a valószínűsége, hogy 11 óra előtt megérkezik?**

Legyen X a futár érkezésének időpontja. Így fogunk tudni számolni:

$$\mathbb{P}(X \leq 11) = \frac{11 - 8}{12 - 8} = \frac{3}{4} = 75\%.$$

$$\mathbb{P}(X \leq 11 | X > 10) = \frac{\mathbb{P}(\{X \leq 11\} \cap \{X > 10\})}{\mathbb{P}(X > 10)} = \frac{1/4}{2/4} = \frac{1}{2}.$$

Egyenletes eloszlás



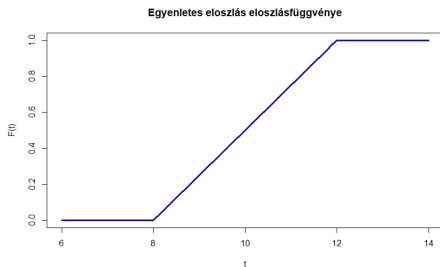
A $[8, 12]$ intervallumon egyenletes eloszlásból vett 500 elemű minta hisztogramja

Egyenletes eloszlás

Definíció (Egyenletes eloszlás (uniform distribution))

Az X valószínűségi változó **egyenletes eloszlású** az $[a, b]$ intervallumon, ha eloszlásfüggvénye:

$$F(t) = \mathbb{P}(X \leq t) = \begin{cases} 0, & \text{ha } t \leq a; \\ \frac{t-a}{b-a}, & \text{ha } a < t < b; \\ 1, & \text{ha } t \geq b. \end{cases}$$



Egyenletes eloszlás

Csomagot várunk, a futár 10 és 12 óra között érkezik. Feltesszük, hogy érkezésének időpontja egyenletes eloszlású a $[10, 12]$ intervallumon. Ekkor az előző állítás alapján az alábbiak igazak ($a = 10, b = 12$).

- Annak valószínűsége, hogy 10 és 11 óra között érkezik: $(11 - 10)/(12 - 10) = 1/2$.
- Annak valószínűsége, hogy 10:15 és 10:30 között érkezik: $1/8 = 0,125$.
- Annak valószínűsége, hogy 10 : 30 után érkezik: $3/4 = 0,75$.