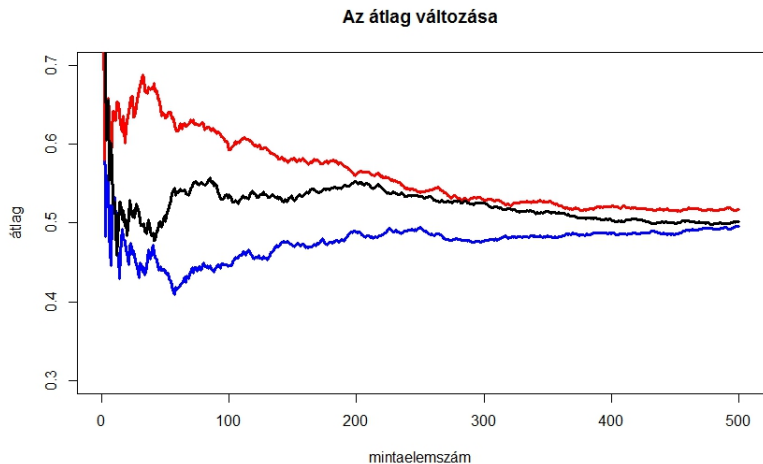


Az átlag konvergenciája (6. előadás)



A $[0, 1]$ intervallumon egyenletes eloszlásból vett mintából az első n elem átlaga $n = 500$ -ig

A nagy számok gyenge törvénye

Legyenek X_1, \dots, X_n független azonos eloszlású véges szórású valószínűségi változók. Legyen $m = \mathbb{E}(X_1)$ és $\sigma = D(X_1)$.

A korábbiak szerint

$$\mathbb{E}(\bar{X}) = m; \quad D^2(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 \quad (n \rightarrow \infty).$$

A nagy számok gyenge törvénye

Legyenek X_1, \dots, X_n független azonos eloszlású véges szórású valószínűségi változók. Legyen $m = \mathbb{E}(X_1)$ és $\sigma = D(X_1)$.

A korábbiak szerint

$$\mathbb{E}(\bar{X}) = m; \quad D^2(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 \quad (n \rightarrow \infty).$$

A Csebisev-egyenlőtlenség szerint minden $\varepsilon > 0$ -ra

$$\mathbb{P}(|\bar{X} - m| > \varepsilon) \leq \frac{D^2(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \rightarrow 0 \quad (n \rightarrow \infty).$$

Tehát minden $\varepsilon > 0$ -ra

$$\mathbb{P}(|\bar{X} - m| > \varepsilon) \rightarrow 0$$

teljesül $n \rightarrow \infty$ esetén.

Azaz: $\bar{X} \rightarrow m = \mathbb{E}(X_1)$ sztochasztikusan.

Egyenlőtlenségek

Állítás (Markov-egyenlőtlenség)

Legyen $t > 0$ tetszőleges pozitív szám, X pedig olyan véges várható értékű valószínűségi változó, mely csak nemnegatív értékeket vesz fel, vagyis melyre $X \geq 0$ teljesül. Ekkor

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Egyenlőtlenségek

Állítás (Markov-egyenlőtlenség)

Legyen $t > 0$ tetszőleges pozitív szám, X pedig olyan véges várható értékű valószínűségi változó, mely csak nemnegatív értékeket vesz fel, vagyis melyre $X \geq 0$ teljesül. Ekkor

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Állítás (Csebisev-egyenlőtlenség)

Legyen X véges szórású valószínűségi változó, $t > 0$ pozitív szám. Ekkor

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{D^2(X)}{t^2}.$$

Egyenlőtlenségek

Állítás (Markov-egyenlőtlenség)

Legyen $t > 0$ tetszőleges pozitív szám, X pedig olyan véges várható értékű valószínűségi változó, mely csak nemnegatív értékeket vesz fel, vagyis melyre $X \geq 0$ teljesül. Ekkor

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Állítás (Csebisev-egyenlőtlenség)

Legyen X véges szórású valószínűségi változó, $t > 0$ pozitív szám. Ekkor

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{D^2(X)}{t^2}.$$

Következmény

Legyen X véges szórású valószínűségi változó, $t > 0$ pozitív szám. Ekkor

$$\mathbb{P}(|X - \mathbb{E}(X)| < t) \geq 1 - \frac{D^2(X)}{t^2}.$$

A nagy számok törvénye

Tétel (A nagy számok gyenge törvénye)

Legyenek X_1, X_2, \dots olyan valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel, hogy $D(X_1) < \infty$. Ekkor minden $\varepsilon > 0$ esetén

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

azaz $\bar{X}_n \rightarrow \mathbb{E}(X_1)$ sztochasztikusan.

A nagy számok törvénye

Tétel (A nagy számok gyenge törvénye)

Legyenek X_1, X_2, \dots olyan valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel, hogy $D(X_1) < \infty$. Ekkor minden $\varepsilon > 0$ esetén

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

azaz $\bar{X}_n \rightarrow \mathbb{E}(X_1)$ sztochasztikusan.

Tétel (A nagy számok erős törvénye)

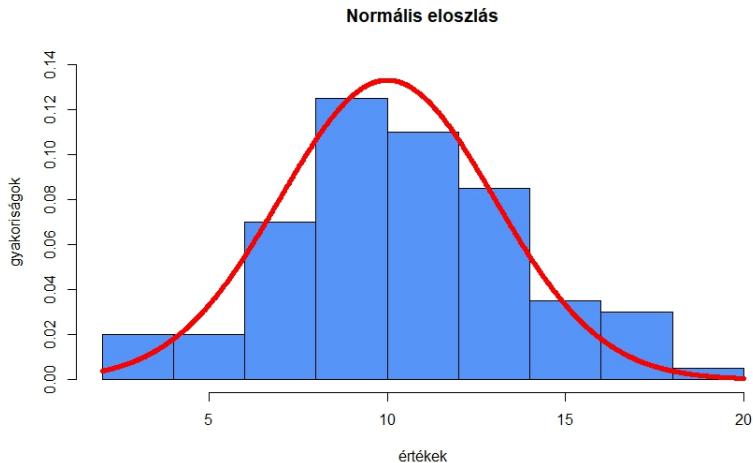
Legyenek X_1, X_2, \dots valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel még, hogy $m = \mathbb{E}(X_1) < \infty$. Ekkor

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}(X_1) = m$$

teljesül 1 valószínűséggel $n \rightarrow \infty$ esetén.

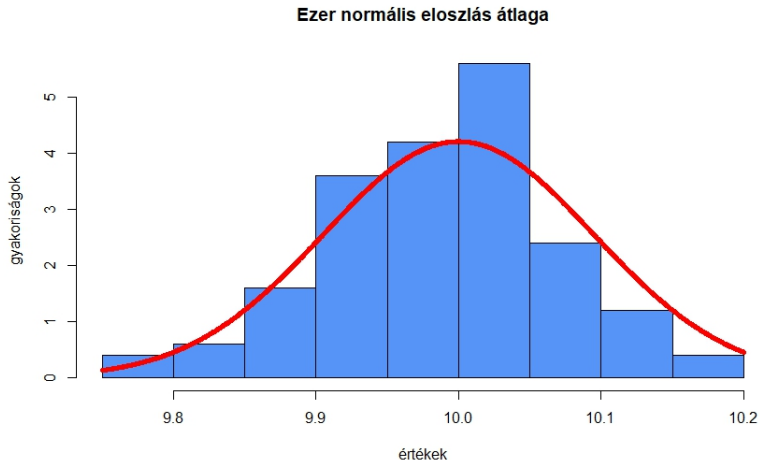
A második esetben gyengébb feltevésből erősebb állítás következik.

Normális eloszlás



Száz független normális eloszlású valószínűségi változó hisztogramja és a sűrűségfüggvény ($m = 10, \sigma = 3, \bar{x} = 9,88, s_n^* = 2,58$)

Normális eloszlások átlaga



Százelemű minta az alábbi eloszlásból: $n = 1000$ független normális eloszlású ($m = 10, \sigma = 3$) valószínűségi változó átlaga és az $N(10, 3/\sqrt{1000})$ normális eloszlás sűrűségfüggvénye ($\bar{x} = 9,99, s_n^* = 0,084, \sigma/\sqrt{n} = 0,095$)

Normális eloszlások átlaga

Legyenek X, Y függetlenek, normális eloszlásúak: $X \sim N(m_1, \sigma_1^2)$, $Y \sim N(m_2, \sigma_2^2)$.
Ekkor a következők igazak:

- $X + b$ eloszlása normális, $m_1 + b$ várható értékkel és σ szórással;
- aX eloszlása normális am_1 várható értékkel és $|a|\sigma$ szórással;
- $X + Y$ eloszlása normális, $m_1 + m_2$ várható értékkel és $\sqrt{\sigma_1^2 + \sigma_2^2}$ szórással.

Emlékeztető: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, és ha X és Y függetlenek, akkor $D^2(X + Y) = D^2(X) + D^2(Y)$.

Normális eloszlások átlaga

Legyenek X, Y függetlenek, normális eloszlásúak: $X \sim N(m_1, \sigma_1^2)$, $Y \sim N(m_2, \sigma_2^2)$.
Ekkor a következők igazak:

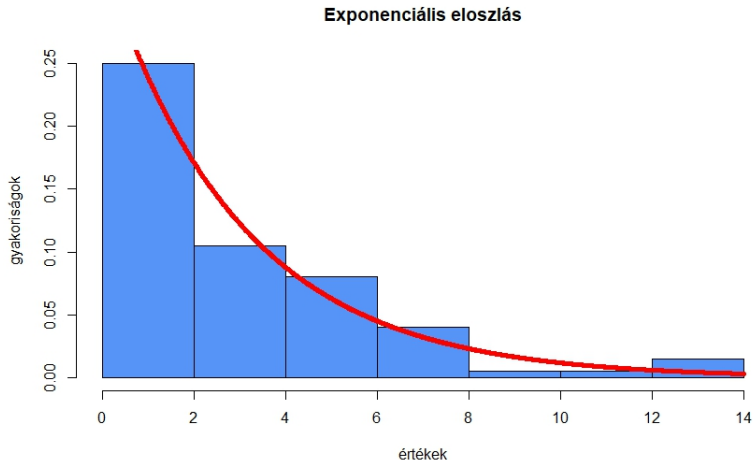
- $X + b$ eloszlása normális, $m_1 + b$ várható értékkel és σ szórással;
- aX eloszlása normális am_1 várható értékkel és $|a|\sigma$ szórással;
- $X + Y$ eloszlása normális, $m_1 + m_2$ várható értékkel és $\sqrt{\sigma_1^2 + \sigma_2^2}$ szórással.

Emlékeztető: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, és ha X és Y függetlenek, akkor $D^2(X + Y) = D^2(X) + D^2(Y)$.

Ebből következik: ha X_1, \dots, X_n független normális eloszlásúak m várható értékkel és σ szórással, akkor

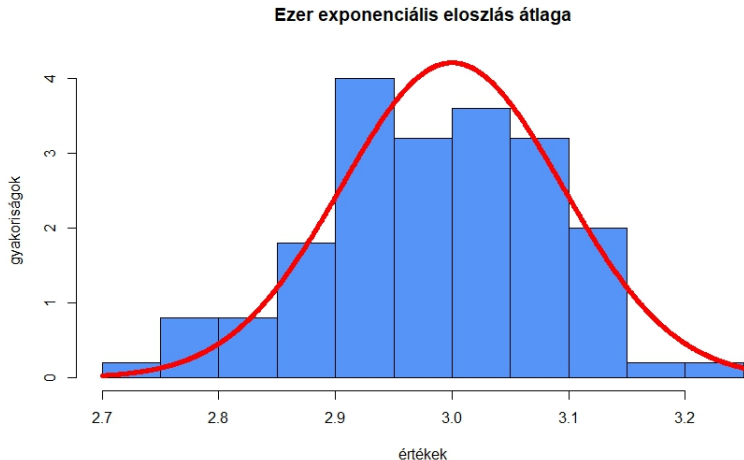
$$\frac{X_1 + \dots + X_n}{n} \sim N\left(m, \frac{\sigma^2}{n}\right)$$

Exponenciális eloszlás



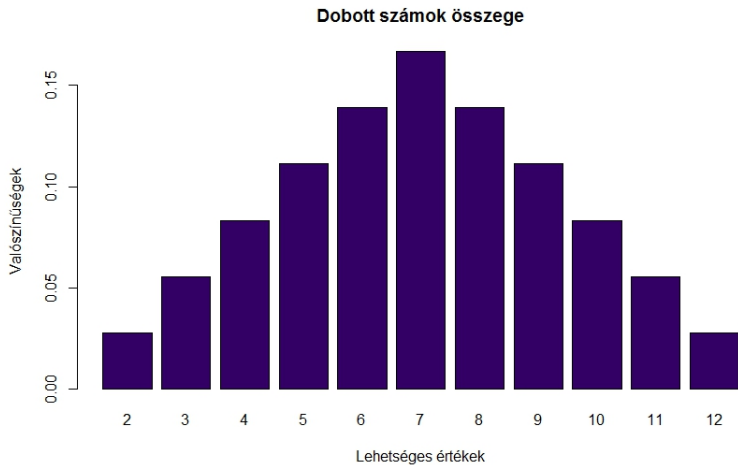
Száz független $\lambda = 1/3$ paraméterű exponenciális eloszlású valószínűségi változó hisztogramja és a sűrűségfüggvény, azaz $e^{-1/3}/3$ ($\mathbb{E}(X) = D(X) = 3, \bar{x} = 3,03, s_n^* = 2,89$)

Exponenciális eloszlások átlaga



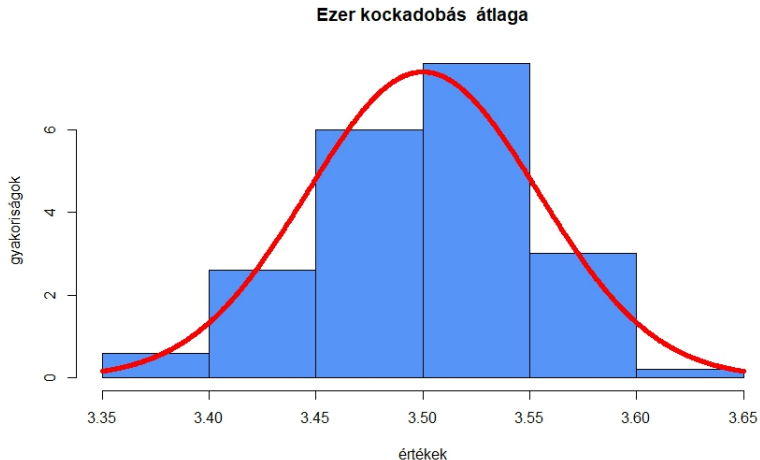
Százelemű minta az alábbi eloszlásból: $n = 1000$ független exponenciális eloszlású ($\lambda = 1/3$) valószínűségi változó átlaga, és az $N(3, 3/\sqrt{1000})$ normális eloszlás sűrűségfüggvénye ($\bar{x} = 2,98, s_n^* = 0,098, \sigma/\sqrt{n} = 0,095$)

Két kockadobás összege



Két szabályos kockadobás összegének eloszlása

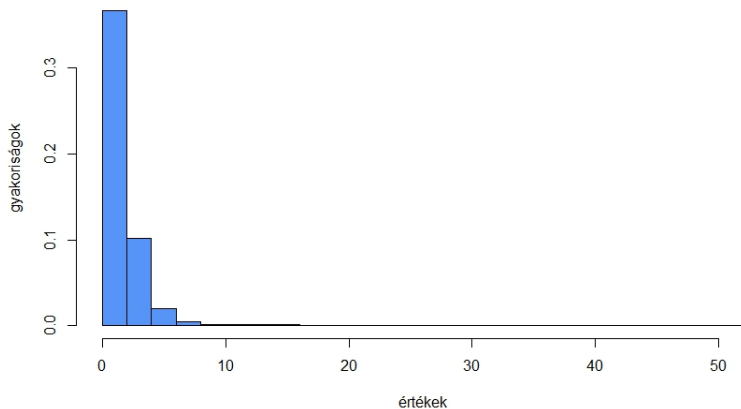
Kockadobások átlaga



Százelemű minta az alábbi eloszlásból: $n = 1000$ független szabályos kockadobás átlaga, és az $N(3,5, D(X_1)/\sqrt{1000})$ normális eloszlás sűrűségfüggvénye ($\bar{x} = 3,501, s_n^* = 0,098, \sigma/\sqrt{n} = 0,051$)

Exponenciális eloszlás a kitevőben

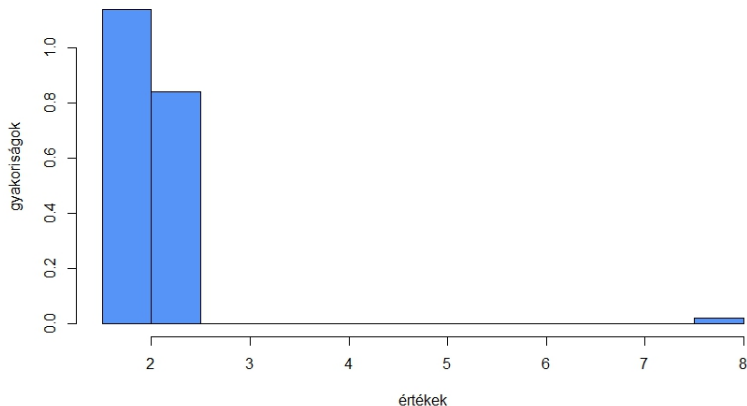
Exponenciális eloszlás a kitevőben



$e^{X_1}, e^{X_2}, \dots, e^{X_{1000}}$ hisztogramja, ahol X_i -k függetlenek, 2 paraméterű exponenciális eloszlásúak ($\mathbb{E}(e^{X_1}) = 2, D(e^{X_1}) = \infty, \bar{x} = 1,99, s_n^* = 2,33$)

Exponenciális eloszlás a kitevőben

Ezer exponenciális eloszlás átlaga



Százelemű minta az alábbi eloszlásból: $e^{X_1}, e^{X_2}, \dots, e^{X_{1000}}$ átlaga, ahol X_i -k függetlenek, 2 paraméterű exponenciális eloszlásúak. Itt e^{X_i} várható értéke véges, de szórása végtelen.

Centrális határeloszlástétel

Tétel (Centrális határeloszlástétel)

Legyenek X_1, X_2, \dots **független azonos eloszlású** valószínűségi változók, melyekre $\mathbb{E}(X_1) = m$ és $D(X_1) = \sigma < \infty$, azaz **szórásuk véges**. Ekkor tetszőleges t valós számra

$$\mathbb{P}\left(\frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma\sqrt{n}} \leq t\right) \rightarrow \mathbb{P}(Z \leq t) \quad (n \rightarrow \infty),$$

ahol Z standard normális eloszlású, azaz

$$\mathbb{P}(Z \leq t) = \Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

Ezt úgy is fogalmazhatjuk, hogy

$$\frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma\sqrt{n}} \rightarrow N(0, 1)$$

teljesül $n \rightarrow \infty$ esetén eloszlásban. Azonos eloszlású: $\mathbb{P}(X_i \leq t) = P(X_j \leq t)$ minden i, j párra és t valós számra

Centrális határeloszlástétel

Legyenek X_1, X_2, \dots független azonos eloszlású valószínűségi változók, melyekre $\mathbb{E}(X_1) = m$ és $D(X_1) = \sigma < \infty$. Ekkor

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma \sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

A határértéket $\Phi(b) - \Phi(a) = \mathbb{P}(a \leq Y \leq b)$ alakban is írhatjuk, ahol $Y \sim N(0, 1)$.

Centrális határeloszlástétel

Legyenek X_1, X_2, \dots független azonos eloszlású valószínűségi változók, melyekre $\mathbb{E}(X_1) = m$ és $D(X_1) = \sigma < \infty$. Ekkor

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma \sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

A határértéket $\Phi(b) - \Phi(a) = \mathbb{P}(a \leq Y \leq b)$ alakban is írhatjuk, ahol $Y \sim N(0, 1)$.

Így is átfogalmazható a tétel állítása:

$$\mathbb{P}(nm + a\sigma\sqrt{n} \leq X_1 + X_2 + \dots + X_n < nm + b\sigma\sqrt{n}) \rightarrow \Phi(b) - \Phi(a).$$

Ez azt jelenti, hogy az \bar{X}_n átlag eloszlása közel van egy m várható értékű, σ/\sqrt{n} szórású normális eloszláshoz.

A statisztikáról

A statisztika céljai

- mérési eredmények, megfigyelések elemzése (leíró statisztika)
- ismeretlen paraméterek becslése (matematikai statisztika, becsléelmélet)
- hipotézisek ellenőrzése vagy cáfolata (matematikai statisztika, hipotézisvizsgálat)
- véletlen folyamatok előrejelzése (regresszió, idősorelemzés)

Alkalmazási területek

- hierarchikus gépi tanulási struktúrák
- nagy adathalmazok elemzése
- társadalomtudományok: szociológia, pszichológia
- élő- és élettelen természettudományok, pl. geológia, meteorológia
- pénzügyi matematika, biztosítás, közgazdaságtan

Matematikai statisztika

Példa matematikai statisztikai kérdésre

- Egy adott helyen húsz éven keresztül feljegyezték, hogy hányszor volt hurrikán. Ezek alapján várhatóan hány hurrikán lesz 2020-ban? Mennyi a becslésünk bizonytalansága? Mennyi a valószínűsége, hogy ötnél több hurrikán lesz?
- Egy közvéleménykutatás során 1000 ember közül 63 választana egy adott pártot. Ez alapján állíthatjuk-e, hogy a párt támogatottsága szignifikánsan magasabb 5%-nál? Mennyi a tévedésünk valószínűsége?
- Megmérték 100 férfi és 60 nő testmagasságát. Állíthatjuk-e az adatok alapján, hogy a férfiak szignifikánsan magasabbak a nőknél? Mennyi a tévedésünk valószínűsége?
- 100 ember közül 27 télen, 22 tavasszal, 34 nyáron, a többiek ősszel születtek. Állíthatjuk-e az adatok alapján, hogy a születések eloszlása szignifikánsan eltér az egyenletes eloszlástól (amikor minden évszaknak $1/4$ a valószínűsége)?

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például a megkérdezetteket függetlenül választottuk, a mérések nem befolyásolják egymást), azaz

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

teljesül tetszőleges t_1, t_2, \dots, t_n valós számok esetén.

Az (X_1, X_2, \dots, X_n) valószínűségi változók **eloszlása nem ismert**. A cél ezeknek a valószínűségi változók eloszlásának a becslése, rá vonatkozó hipotézisek eldöntése a megfigyelések, vagyis az adatok alapján.

Példa: statisztikai minta

A Duna vízállása húsz napon keresztül (2016. január, Országos Vízelző Szolgálat):

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

Példa: statisztikai minta

A Duna vízállása húsz napon keresztül (2016. január, Országos Vízeljáró Szolgálat):

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

X_i valószínűségi változó: a vízállás az i . napon ($i = 1, 2, \dots, 20$). Vagyis ennél a megfigyelésnél $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$.

Független-e ez a minta?

Példa: statisztikai minta

A Duna vízállása húsz napon keresztül (2016. január, Országos Vízellő Szolgálat):

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

X_i valószínűségi változó: a vízállás az i . napon ($i = 1, 2, \dots, 20$). Vagyis ennél a megfigyelésnél $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$.

Független-e ez a minta?

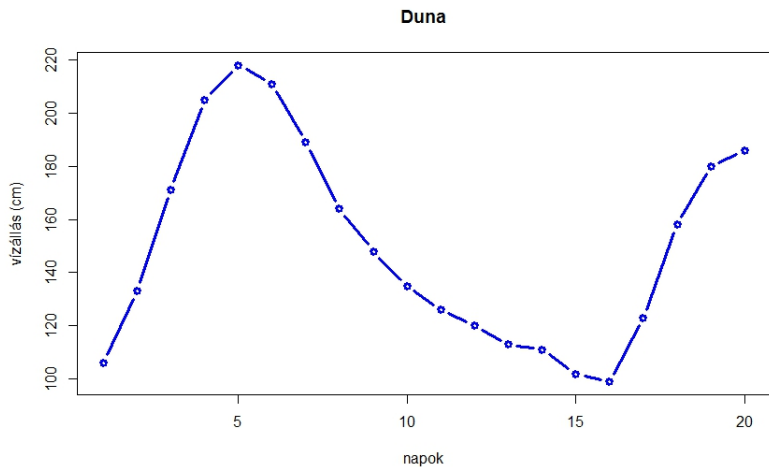
Nem független, nagyobb vízállás után várhatóan másnap is magasabb lesz a Duna szintje.

Leíró statisztika

Nem a véletlen hatásának megértése és valószínűségszámítási módszereken alapuló következtetések levonása a célja, hanem a megfigyelt adatok **megjelenítése, jellemzőinek kiszámítása**. Ide tartozhat:

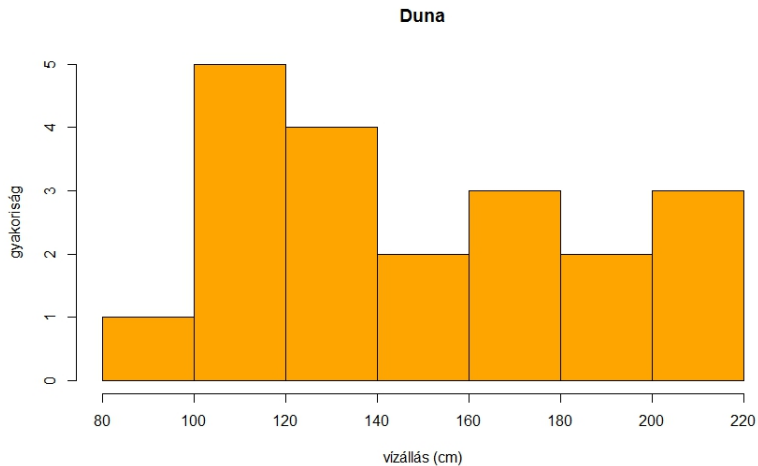
- diagramok: kördiagram, oszlopdiaagram, hisztogram
- táblázatok, kontingenciatáblák
- középértékek, szórások kiszámítása
- kvantilisok számítása, boxplot ábra
- indexek számítása

Példa: az adatok ábrázolása

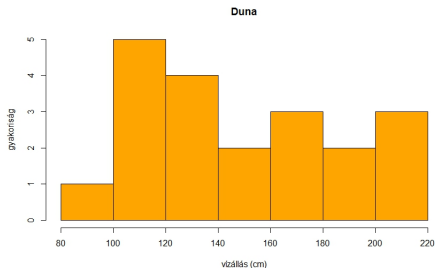


Példa: hisztogram

A Duna vízállásának hisztogramja

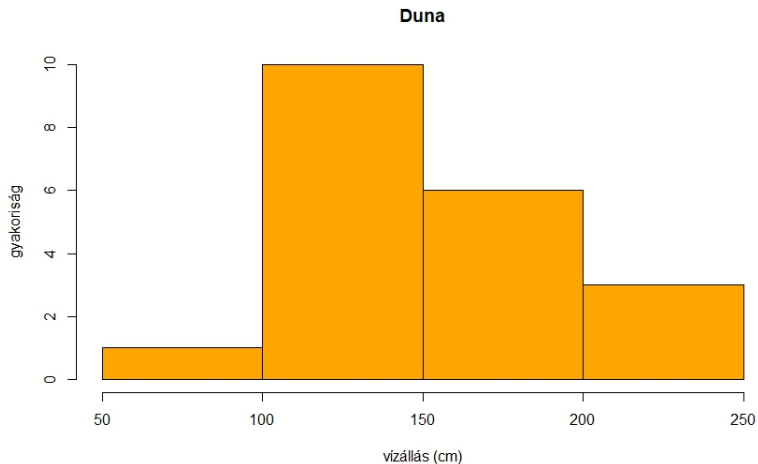


Példa: hisztogram



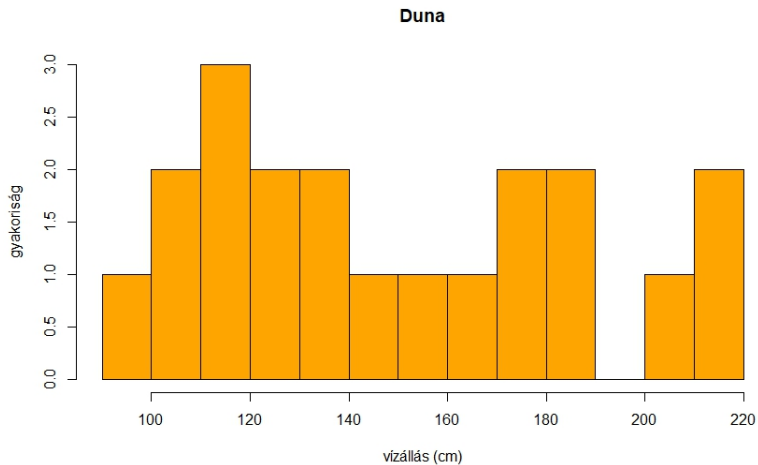
Választunk egy intervallumot, mely magában foglalja a mérési adatokat. Az intervallumot egyenlő nagyságú részekre osztjuk. Az egyes kis intervallumokba eső mérési adatok számát ábrázoljuk. Sem a túl hosszú, sem a túl rövid intervallumok nem adnak informatív ábrát.

Példa: túl hosszú intervallumok, túl kevés osztály



A Duna vízállására vonatkozó húszelemű minta hisztogramja 4 osztállyal

Példa: túl rövid intervallumok, túl sok osztály



A Duna vízállására vonatkozó húszelemű minta hisztogramja 15 osztállyal

Alapstatisztikák

Minta: X_1, \dots, X_n (a példában $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$)

- **minimum**: a legkisebb mintaelem, azaz $\min(X_1, X_2, \dots, X_n)$.
- **maximum**: a legnagyobb mintaelem, azaz $\max(X_1, X_2, \dots, X_n)$.
- **terjedelem** (range): a legnagyobb és legkisebb mintaelem különbsége, azaz
$$\max(X_1, X_2, \dots, X_n) - \min(X_1, X_2, \dots, X_n).$$
- **medián**: a **nagyság szerinti középső** mintaelem, vagy a középső kettő átlaga (ha n páros).
- **módusz** (mode): a leggyakrabban előforduló mintaelem.

Alapstatisztikák

Minta: X_1, X_2, \dots, X_n .

- **mintaátlag** (mean): $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j = \frac{X_1 + \dots + X_n}{n}$.

- **tapasztalati szórásnégyzet**:

$$s_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \left(\frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n} - \bar{X}^2.$$

- tapasztalati szórás: $s_n = \sqrt{s_n^2}$.
- **korrigált tapasztalati szórásnégyzet** (variance):

$$s_n^{*2} = \frac{n}{n-1} \cdot s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{n}{n-1} \left(\left(\frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 \right).$$

- **korrigált tapasztalati szórás** (standard deviation, sd): $s_n^* = \sqrt{s_n^{*2}}$.

További statisztikák

- **korrigált tapasztalati szórásnégyzet** (variance):

$$s_n^{*2} = \frac{n}{n-1} \cdot s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{n}{n-1} \left(\left(\frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 \right).$$

- **korrigált tapasztalati szórás** (standard deviation, sd): $s_n^* = \sqrt{s_n^{*2}}$.
- **relatív szórás** (relative standard deviation, rsd): $\frac{s_n^*}{\bar{X}}$.
- **standard hiba (standard error)**: $\frac{s_n^*}{\sqrt{n}}$.

Példa: alapstatisztikák

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

mintaelemszám: $n = 20$

minta: $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$.

átlag: $\bar{X} = 149,9$

tapasztalati szórásnégyzet: $s_n^2 = 1412,09$

tapasztalati szórás: $s_n = 37,58$

korrigált tapasztalati szórásnégyzet: $s_n^{*2} = 1486,411$

korrigált tapasztalati szórás: $s_n^* = 38,55$

relatív szórás: $0,257$

standard hiba: $8,62$

Rendezett minta

Rendezett minta: a mintaelemeket nagyság szerint növekvő sorrendbe állítjuk.

Jelölés:

$$(X_1^*, X_2^*, \dots, X_n^*).$$

Vagyis $\{X_1^*, X_2^*, \dots, X_n^*\} = \{X_1, X_2, \dots, X_n\}$ és $X_1^* \leq X_2^* \leq \dots \leq X_n^*$.

A minimum X_1^* , a maximum X_n^* . A k . legkisebb mintaelem X_k^* .

Rendezett minta

Rendezett minta: a mintaelemeket nagyság szerint növekvő sorrendbe állítjuk.
Jelölés:

$$(X_1^*, X_2^*, \dots, X_n^*).$$

Vagyis $\{X_1^*, X_2^*, \dots, X_n^*\} = \{X_1, X_2, \dots, X_n\}$ és $X_1^* \leq X_2^* \leq \dots \leq X_n^*$.

A minimum X_1^* , a maximum X_n^* . A k . legkisebb mintaelem X_k^* .

Példa: a Duna vízállásáról kapott húszelemű adatsor rendezett mintája:

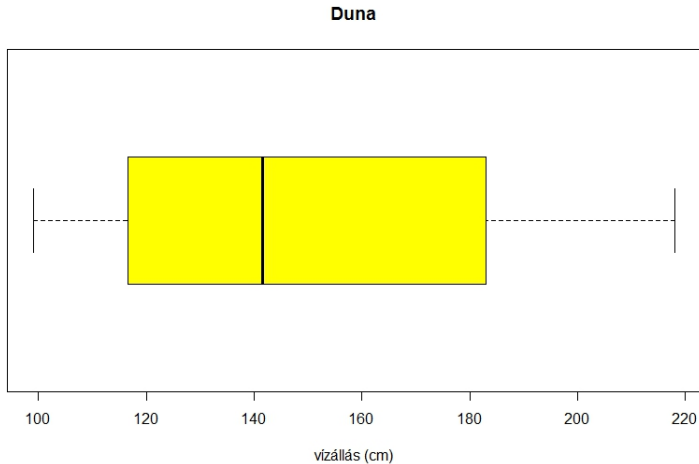
99	102	106	111	113	120	123	126	133	135
148	158	164	171	180	186	189	205	211	218

$X_1^* = 99, X_2^* = 102, X_3^* = 106, \dots, X_6^* = 120, \dots, X_{10}^* = 135$

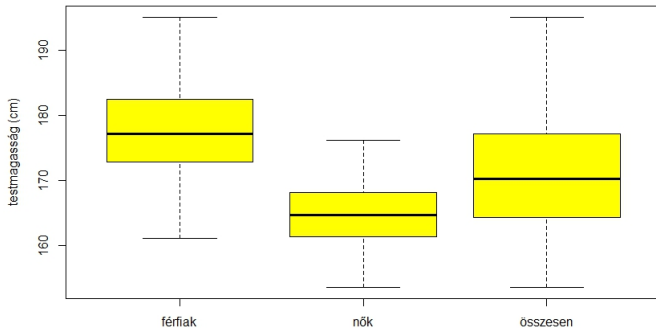
$X_{11}^* = 148, \dots, X_{14}^* = 171, \dots, X_{20}^* = 218.$

Példa: boxplot

A Duna vízállásáról szóló minta boxplotja a húsznapos adatsorból



Példa: boxplot



A testmagasság boxplotja $n = 96$ elemű mintából, balról jobbra: férfiak, nők, összesen

Kvantilisek

Az X valószínűségi változó z -kvantilise a legkisebb olyan q szám, melyre teljesül, hogy $\mathbb{P}(X \leq q) \geq z$.

A tapasztalati z -kvantilise több definíciót is szoktak használni, egy lehetőség:

Definíció (Tapasztalati kvantilis)

Legyen $X_1^* \leq X_2^* \leq \dots \leq X_n^*$ rendezett minta, és $z \in [0, 1]$ adott szám. Ekkor a minta tapasztalati z -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

Kvantilisek

Az X valószínűségi változó z -kvantilise a legkisebb olyan q szám, melyre teljesül, hogy $\mathbb{P}(X \leq q) \geq z$.

A tapasztalati z -kvantilisre több definíciót is szoktak használni, egy lehetőség:

Definíció (Tapasztalati kvantilis)

Legyen $X_1^* \leq X_2^* \leq \dots \leq X_n^*$ rendezett minta, és $z \in [0, 1]$ adott szám. Ekkor a minta tapasztalati z -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

Első kvartilis: $z = 1/4$ -kvantilis, harmadik kvartilis: $z = 3/4$ -kvantilis, a medián pedig a $z = 1/2$ -hez tartozó tapasztalati kvantilis.

Boxplot

Definíció (Tapasztalati kvantilis)

Legyen X_1, X_2, \dots, X_n minta, és $z \in [0, 1]$ adott szám. Ekkor a minta tapasztalati z -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

A boxplot készítéséhez szükséges adatok:

- **minimum**: a legkisebb mintaelem (99);
- **első kvartilis**: a $z = 1/4$ -hez tartozó kvantilis ($118,2 = X_5^* + 0,25 \cdot (X_6^* - X_5^*)$);
- **medián** (141,5);
- **harmadik kvartilis**: a $z = 3/4$ -hez tartozó kvantilis (181,5);
- **maximum**: a legnagyobb mintaelem (218).

Tapasztalati eloszlásfüggvény

Az X valószínűségi változó eloszlásfüggvénye az $F : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

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

minden $t \in \mathbb{R}$ -re.

Tapasztalati eloszlásfüggvény

Az X valószínűségi változó eloszlásfüggvénye az $F : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

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

minden $t \in \mathbb{R}$ -re.

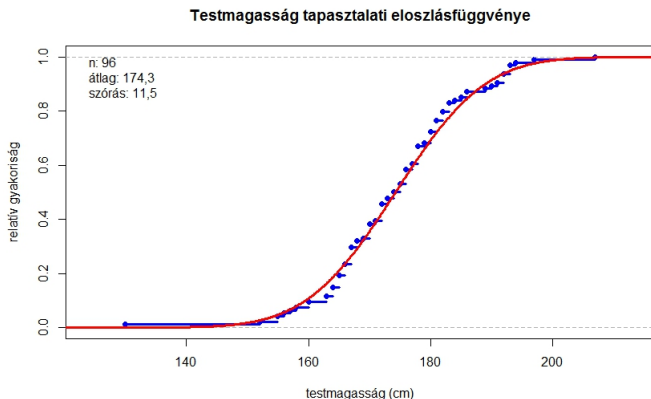
Definíció (Tapasztalati eloszlásfüggvény)

Az X_1, X_2, \dots, X_n minta tapasztalati eloszlásfüggvénye az $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

$$\hat{F}_n(t) = \frac{t\text{-nél nem nagyobb mintaelemek száma}}{n}.$$

(empirical cumulative distribution function)

Tapasztalati eloszlásfüggvény



A testmagasság tapasztalati eloszlásfüggvénye $n = 96$ elemű mintából, és az $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlás eloszlásfüggvénye.

Középértékek: medián

Definíció (medián)

Egy X valószínűségi változó mediánja egy olyan m szám, melyre $\mathbb{P}(X \leq m) = 1/2$ teljesül.

Minta: (X_1, X_2, \dots, X_n) , mintaelemszám: n .

Definíció (tapasztalati medián)

Ha n páratlan: a rendezett minta középső, $(n + 1)/2$. elemét, azaz $X_{(n+1)/2}^*$ -t a minta **mediánjának** nevezzük.

Ha n páros: a rendezett minta $n/2$. és $n/2 + 1$. elemének átlagát, azaz a

$$\frac{X_{n/2}^* + X_{n/2+1}^*}{2}$$

mennyiséget a minta mediánjának nevezzük.

Példa: a Duna vízállásáról kapott húszelemű minta mediánja:

$$\frac{1}{2}(X_{10}^* + X_{11}^*) = \frac{1}{2}(135 + 148) = 141,5.$$

Középértékek: az átlag és a medián összehasonlítása

Normális eloszlás

500 elemű független minta: X_1, X_2, \dots, X_{500} függetlenek, eloszlásuk normális eloszlás $m = 1$ várható értékkel és $\sigma = 1$ szórással

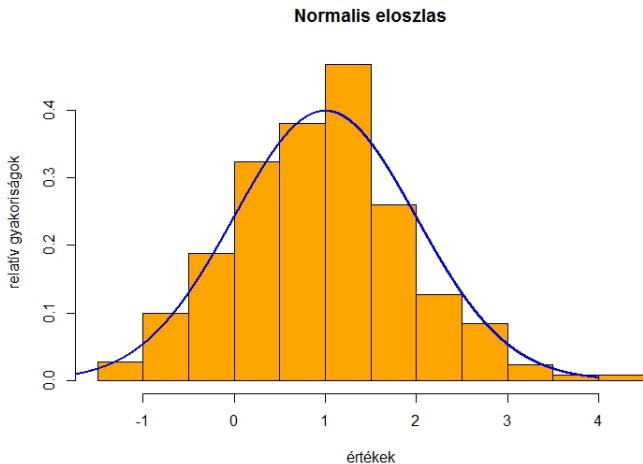
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9840	0.2847	0.9842	0.9863	1.6930	3.6110

Exponenciális eloszlás

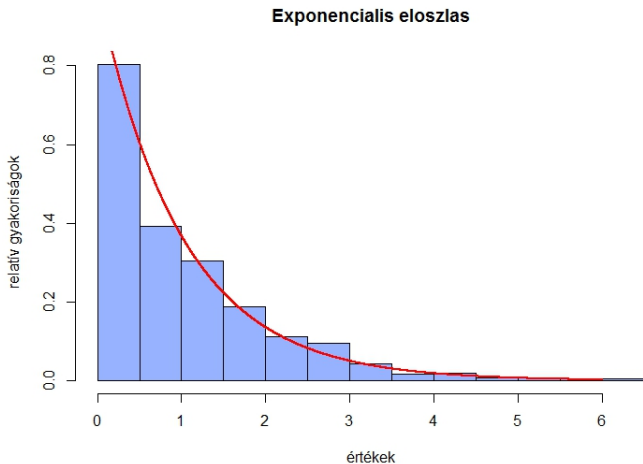
500 elemű független minta: Y_1, Y_2, \dots, Y_{500} függetlenek, eloszlásuk exponenciális eloszlás $b = 1$ paraméterrel. $\mathbb{E}(Y_k) = 1$ és $D(Y_k) = 1$ minden $k = 1, 2, \dots, 500$ -ra.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001326	0.282700	0.637300	0.984900	1.349000	5.895000

A normális eloszlású minta hisztogramja



Az exponenciális eloszlású minta hisztogramja



Az átlag és a medián összehasonlítása

Az átlag

- "több információt használ"
- érzékenyebb a kiugró adatokra, azaz egy hibás mérés is könnyen megváltoztathatja
- nem szimmetrikus esetben eltérhet a leggyakrabban megfigyelt értékektől

A mediánt is érdemes használni, ha

- vannak kiugró (esetleg hibás) adatok;
- ha az eloszlás nem szimmetrikus, és az átlag és a medián jelentősen különbözik (mint a fenti példában az exponenciális eloszlás esetén).