Valószínűségszámítás és statisztika

Informatika BSc, esti tagozat
Backhausz Ágnes

2015/2016. tavaszi félév

Eloszlásfüggvény

Definíció (Eloszlásfüggvény)

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

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

Állítás

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

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

Eloszlásfüggvény

Definíció (Eloszlásfüggvény)

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

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

Állítás

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

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

Állítás (Az eloszlásfüggvény tulajdonságai)

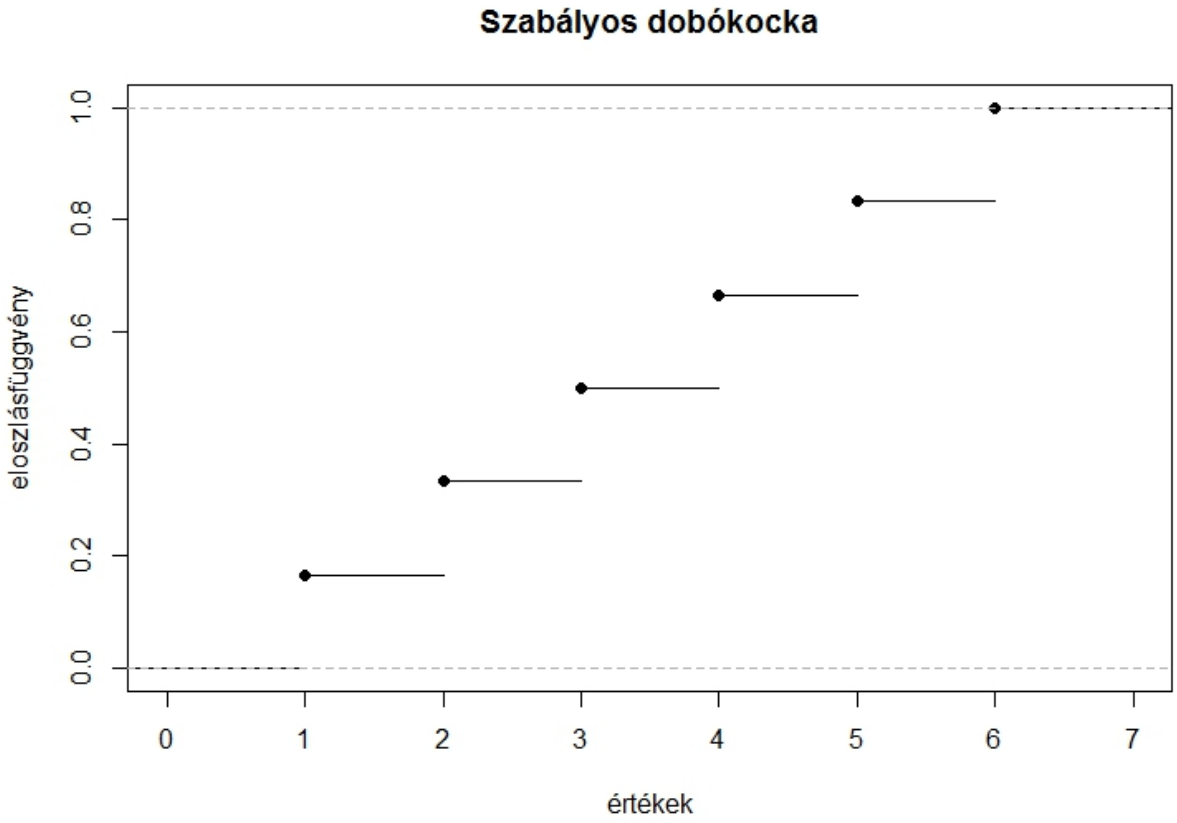
Legyen X *valószínűségi változó,* F *pedig az eloszlásfüggvénye. Ekkor*

(i) F *monoton növő:* $a < b$ *esetén* $F(a) \leq F(b)$.

(ii) $\lim_{t \to -\infty} F(t) = 0; \qquad \lim_{t \to \infty} F(t) = 1.$

(iii) F *jobbról folytonos, azaz minden* $t \in \mathbb{R}$ *valós számra* $\lim_{s \to t-} F(s) = F(t)$.

Példa: eloszlásfüggvény



1. ábra. Szabályos dobókockával dobott szám eloszlásfüggvénye.

Folytonos valószínűségi változó

Ha a $G : \mathbb{R} \to [0, 1]$ függvény rendelkezik az előző állításban szereplő $(i) - (iii)$ tulajdonságokkal, akkor van olyan valószínűségi változó, melynek G az eloszlásfüggvénye.

Definíció

Azt mondjuk, hogy az X valószínűségi változó folytonos, ha eloszlásfüggvénye folytonos.

Egy valószínűségi változó pontosan akkor folytonos, ha $\mathbb{P}(X = t) = 0$ teljesül minden t számra.

Abszolút folytonos valószínűségi változó

Definíció (Abszolút folytonosság és sűrűségfüggvény)

Az X valószínűségi változó **abszolút folytonos**, *ha van olyan $f : \mathbb{R} \to \mathbb{R}$ függvény, melyre*

$$\mathbb{P}(X \leq t) = \int_{-\infty}^{t} f(s)ds$$

teljesül minden $t \in \mathbb{R}$ számra. Ilyenkor az f függvényt az X valószínűségi változó **sűrűségfüggvényének** *nevezzük.*

Abszolút folytonos valószínűségi változó

Definíció (Abszolút folytonosság és sűrűségfüggvény)

Az X valószínűségi változó **abszolút folytonos**, *ha van olyan $f : \mathbb{R} \to \mathbb{R}$ függvény, melyre*

$$\mathbb{P}(X \leq t) = \int_{-\infty}^{t} f(s)ds$$

teljesül minden $t \in \mathbb{R}$ számra. Ilyenkor az f függvényt az X valószínűségi változó **sűrűségfüggvényének** *nevezzük.*

Állítás

Legyen az X abszolút folytonos valószínűségi változó, melynek sűrűségfüggvénye f. Ekkor tetszőleges $a < b$ számokra teljesül, hogy

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f(s)ds.$$

A sűrűségfüggvény tulajdonságai

Állítás (Az eloszlásfüggvény és sűrűségfüggvény kapcsolata)

Legyen X abszolút folytonos valószínűségi változó, melynek F az eloszlásfüggvénye.
(a) Ha f az X sűrűségfüggvénye, akkor minden $t \in \mathbb{R}$ számra

$$F(t) = \int_{-\infty}^{t} f(s)ds.$$

(b) Az $f(t) = F'(t)$ függvény (azokra a t-kre, ahol F differenciálható) az X sűrűségfüggvénye.

A sűrűségfüggvény tulajdonságai

Állítás (Az eloszlásfüggvény és sűrűségfüggvény kapcsolata)

Legyen X abszolút folytonos valószínűségi változó, melynek F az eloszlásfüggvénye.
(a) Ha f az X sűrűségfüggvénye, akkor minden $t \in \mathbb{R}$ számra

$$F(t) = \int_{-\infty}^{t} f(s)ds.$$

(b) Az $f(t) = F'(t)$ függvény (azokra a t-kre, ahol F differenciálható) az X sűrűségfüggvénye.

Állítás (A sűrűségfüggvény jellemzése)

Egy $f : \mathbb{R} \to \mathbb{R}$ függvény pontosan akkor sűrűségfüggvénye valamilyen valószínűségi változónak, ha

(i) *$f(s) \geq 0$ teljesül "majdnem minden" $s \in \mathbb{R}$-re (például véges vagy megszámlálható sok kivétel lehetséges).*

(ii) $\int_{-\infty}^{\infty} f(s)ds = 1.$

Várható érték és szórás

Definíció (Várható érték, abszolút folytonos eset)

Legyen X abszolút folytonos valószínűségi változó, melynek sűrűségfüggvénye f. Ekkor X várható értéke:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} s \cdot f(s) ds,$$

ha ez az integrál létezik és véges.

Várható érték és szórás

Definíció (Várható érték, abszolút folytonos eset)

Legyen X abszolút folytonos valószínűségi változó, melynek sűrűségfüggvénye f. Ekkor X várható értéke:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} s \cdot f(s) ds,$$

ha ez az integrál létezik és véges.

Definíció (Szórásnégyzet és szórás)

Tegyük fel, hogy az X valószínűségi változó abszolút folytonos, és sűrűségfüggvénye f. Ekkor X szórásnégyzete:

$$D^2(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big],$$

szórása pedig

$$D(X) = \sqrt{\mathbb{E}\big[(X - \mathbb{E}(X))^2\big]},$$

ha ezek a várható értékek léteznek.

Szórásnégyzet és szórás

Definíció (Szórásnégyzet és szórás)

Tegyük fel, hogy az X valószínűségi változó abszolút folytonos, és sűrűségfüggvénye f. Ekkor X szórásnégyzete:

$$D^2(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big],$$

szórása pedig

$$D(X) = \sqrt{\mathbb{E}\big[(X - \mathbb{E}(X))^2\big]},$$

ha ezek a várható értékek léteznek.

Állítás (A szórásnégyzet kiszámítása)

A szórásnégyzetet a következőképpen számíthatjuk ki abszolút folytonos X valószínűségi változó esetén:

$$D^2(X) = \mathbb{E}(X^2) - \big[\mathbb{E}(X)\big]^2 = \int_{-\infty}^{\infty} s^2 f(s)ds - \left[\int_{-\infty}^{\infty} s \cdot f(s)ds\right]^2,$$

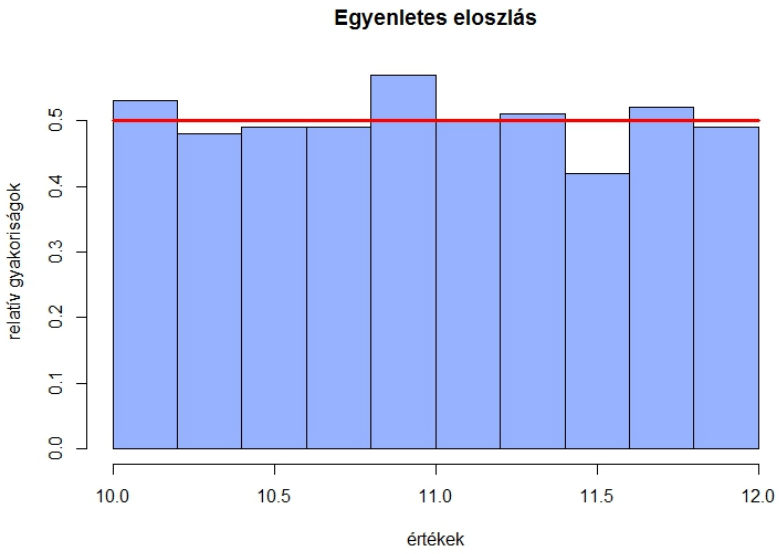
ahol f az X sűrűségfüggvénye.

Egyenletes eloszlás

Definíció (Egyenletes eloszlás)

Legyenek $a < b$ valós számok. Azt mondjuk, hogy az X valószínűségi változó egyenletes eloszlású az $[a, b]$ intervallumon, ha sűrűségfüggvénye

$$f(s) = \begin{cases} \frac{1}{b-a}, & \text{ha } a \leq s \leq b; \\ 0, & \text{különben.} \end{cases}$$



2. ábra. $U(10, 12)$ sűrűségfüggvény és 500 elemű minta hisztogramja.

Egyenletes eloszlás

Állítás (Az egyenletes eloszlás tulajdonságai)

Legyen az X valószínűségi változó egyenletes eloszlású az $[a, b]$ intervallumon. Ekkor a következők teljesülnek.

(i) *X eloszlásfüggvénye:*

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

(ii) *Ha $a \leq c \leq d \leq b$, akkor*

$$\mathbb{P}(c \leq X \leq d) = \int_c^d f(s)ds = \int_c^d \frac{1}{b-a} ds = \frac{d-c}{b-a}.$$

(iii) *Az X valószínűségi változó várható értéke és szórása:*

$$\mathbb{E}(X) = \frac{a+b}{2}; \qquad D(X) = \frac{b-a}{\sqrt{12}}.$$

Példa: egyenletes eloszlás

Példa. 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, $1/8 = 0,125$.
- Érkezési időpontjának várható értéke: $(10 + 12)/2 = 11$ óra.
- Érkezési időpontjának szórása: $(12 - 10)/\sqrt{12} = 1/\sqrt{3} = 0,5774$.

Normális eloszlás

Definíció (Normális eloszlás)

Legyen m valós, σ pedig pozitív szám. Azt mondjuk, hogy az Y valószínűségi változó **normális eloszlású** *m várható értékkel és σ^2 szórásnégyzettel, ha sűrűségfüggvénye*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right) \qquad (x \in \mathbb{R}).$$

Jelölése: $Y \sim N(m, \sigma^2)$.

Normális eloszlás

Definíció (Normális eloszlás)

Legyen m valós, σ pedig pozitív szám. Azt mondjuk, hogy az Y valószínűségi változó **normális eloszlású** *m várható értékkel és σ^2 szórásnégyzettel, ha sűrűségfüggvénye*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right) \qquad (x \in \mathbb{R}).$$

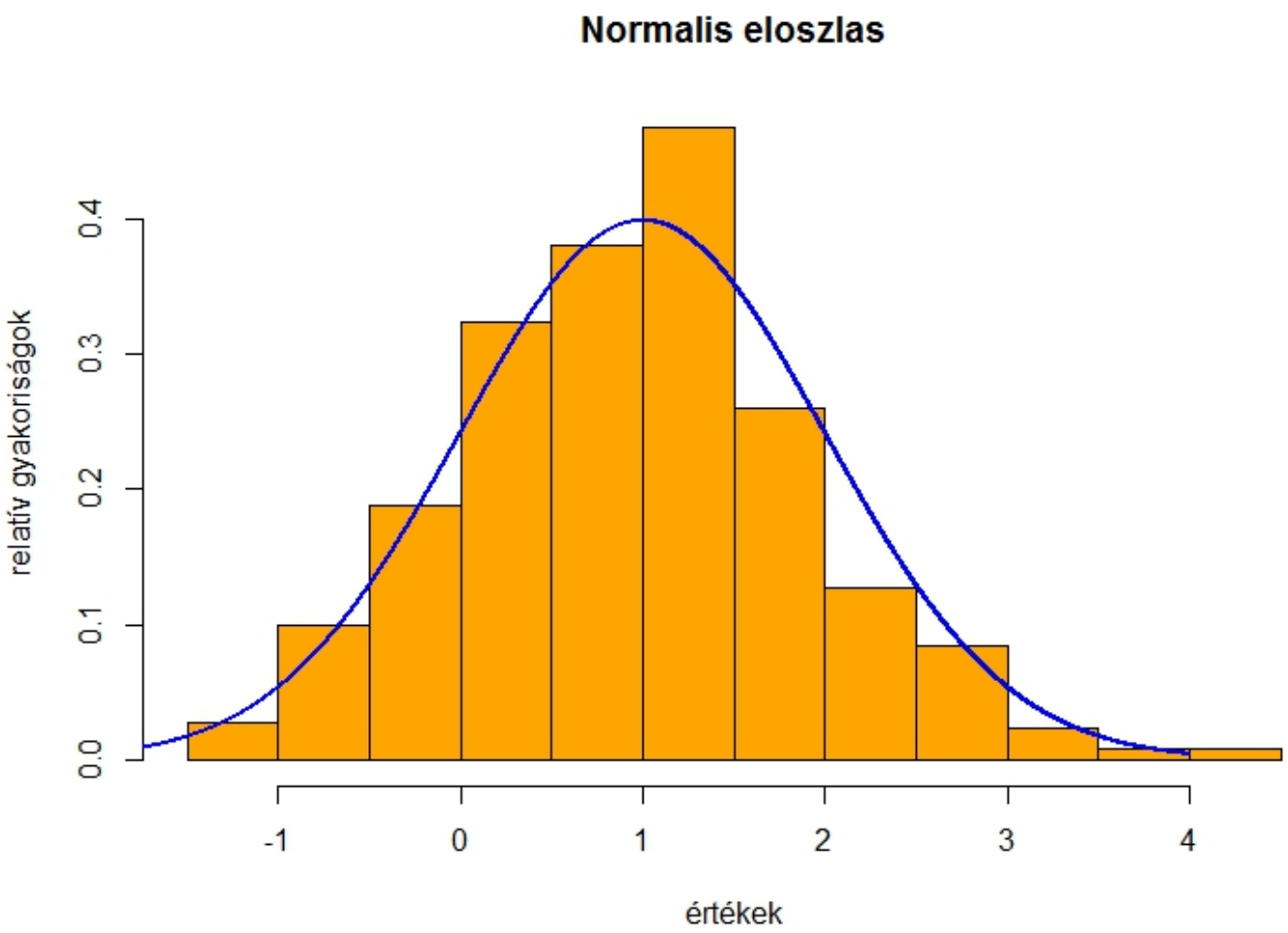
Jelölése: $Y \sim N(m, \sigma^2)$.

Ha $Y \sim N(m, \sigma^2)$, akkor $\mathbb{E}(Y) = m$, $D(Y) = \sigma$.

Standard normális eloszlás: $m = 0$ várható értékű és $\sigma = 1$ szórású normális eloszlás. Eloszlásfüggvénye: Φ, sűrűségfüggvénye:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right).$$

Normális eloszlás



3. ábra.
A standard normális eloszlás sűrűségfüggvénye és 500 elemű minta hisztogramja

A normális eloszlás tulajdonságai

Tegyük fel, hogy Y normális eloszlású m várható értékkel és σ^2 szórásnégyzettel. Ekkor tetszőleges $a \leq b$ valós számokra

- $\mathbb{P}(a < Y < b) = \mathbb{P}(a \leq Y \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds.$
- $\mathbb{P}(a < Y < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$
- $\mathbb{P}(Y < b) = \mathbb{P}(Y \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^b \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds = \Phi\left(\frac{b-m}{\sigma}\right).$
- $\mathbb{P}(a < Y) = \mathbb{P}(a \leq Y) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^\infty \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds = 1 - \Phi\left(\frac{a-m}{\sigma}\right).$

A normális eloszlás tulajdonságai

Tegyük fel, hogy Y normális eloszlású m várható értékkel és σ^2 szórásnégyzettel. Ekkor tetszőleges $a \leq b$ valós számokra

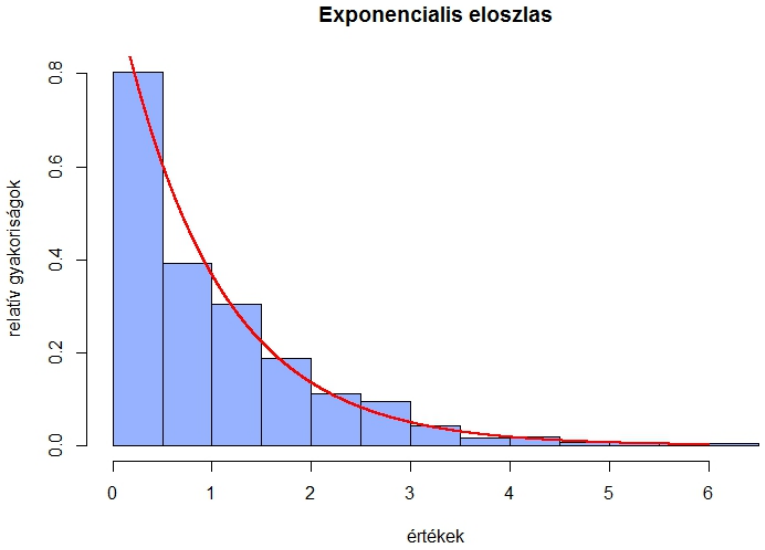
- $\mathbb{P}(a < Y < b) = \mathbb{P}(a \leq Y \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds.$
- $\mathbb{P}(a < Y < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$
- $\mathbb{P}(Y < b) = \mathbb{P}(Y \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^b \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds = \Phi\left(\frac{b-m}{\sigma}\right).$
- $\mathbb{P}(a < Y) = \mathbb{P}(a \leq Y) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^\infty \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds = 1 - \Phi\left(\frac{a-m}{\sigma}\right).$
- Az $aY + b$ valószínűségi változó normális eloszlású $am + b$ várható értékkel és $a^2\sigma^2$ szórásnégyzettel.
- Ha Y_1, Y_2 független normális eloszlású valószínűségi változók, akkor $Y_1 + Y_2$ is normális eloszlású, várható értéke $m_1 + m_2$, szórásnégyzete $\sigma_1^2 + \sigma_2^2$ (ahol $Y_j \sim N(m_j, \sigma_j^2)$.

Exponenciális eloszlás

Definíció (Exponenciális eloszlás)

Legyen $\lambda > 0$ valós szám. Azt mondjuk, hogy az X valószínűségi változó exponenciális eloszlású λ paraméterrel, ha sűrűségfüggvénye

$$f(s) = \begin{cases} \lambda e^{-\lambda s}, & \text{ha } s > 0; \\ 0, & \text{különben.} \end{cases}$$



4. ábra. Exp(1) sűrűségfüggvénye és 500 elemű minta hisztogramja.

Az exponenciális eloszlás tulajdonságai

Állítás

Legyen X exponenciális eloszlású $\lambda > 0$ paraméterrel. Ekkor a következők teljesülnek.

(i) *X eloszlásfüggvénye:*

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(X < t) = \int_{-\infty}^{t} f(s)ds = \begin{cases} 1 - e^{-\lambda s}, & \text{ha } s > 0; \\ 0 & \text{különben.} \end{cases}$$

(ii) *X várható értéke: $\mathbb{E}(X) = 1/\lambda$, szórása: $D(X) = 1/\lambda$.*

(iii) **Örökifjú tulajdonság.** *Legyenek s, t pozitív számok. Ekkor*

$$\mathbb{P}(X \geq s + t | X \geq s) = \mathbb{P}(X \geq t).$$

Példa. Radioaktív részecske bomlási ideje; reakcióidő, várakozási idő sorbanállásnál.

Valószínűségi vektorváltozó

Definíció

Az $\underline{X} = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ *függvény* **valószínűségi vektorváltozó**, *ha tetszőleges* $a_i < b_i$ $(i = 1, 2, \ldots, n)$ *valós számokra teljesül, hogy*

$$\{\omega \in \Omega : a_1 < X_1(\omega) \leq b_1, a_2 < X_2(\omega) \leq b_2, \ldots, a_n < X_n(\omega) \leq b_n\} \in \mathcal{A}.$$

Ha $\underline{X}$ *valószínűségi vektorváltozó, akkor az* X_i *valószínűségi változó eloszlását az* $\underline{X}$ i. **peremeloszlásának** *nevezzük.*
Az $\underline{X}$ *valószínűségi vektorváltozó* **diszkrét**, *ha értékkészlete véges vagy megszámlálhatóan végtelen.*

Példa. X_i: egy adott weboldalt az i. órában hányan töltenek be $(i = 1, 2, \ldots, 24)$. $(X_1, \ldots, X_{24})$ valószínűségi vektorváltozó.

Y_i: az i.-ként érkező igény várakozási ideje sorbanállásnál $(i = 1, 2, \ldots, 100)$.

Együttes eloszlás– és sűrűségfüggvény

Definíció

Az $\underline{X} = (X_1, \ldots, X_n)$ *valószínűségi vektorváltozó* **együttes eloszlásfüggvénye** *az* $F : \mathbb{R}^n \to [0, 1]$ *függvény, melyre*

$$F(\underline{t}) = F(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n), \text{ ha } (t_1, \ldots, t_n) \in \mathbb{R}^n.$$

Definíció

Az $\underline{X} = (X_1, \ldots, X_n)$ *valószínűségi vektorváltozó abszolút folytonos, ha van olyan* $f : \mathbb{R}^n \to \mathbb{R}$ *függvény, melyre*

$$F(t_1, \ldots, t_n) = \int_{-\infty}^{t_1} \ldots \int_{-\infty}^{t_n} f(s_1, \ldots, s_n) ds_1 \ldots ds_n.$$

teljesül minden $(t_1, \ldots, t_n) \in \mathbb{R}^n$ *esetén. Ilyenkor az* f *függvényt az* $\underline{X}$ **együttes sűrűségfüggvényének** *nevezzük.*

Az együttes sűrűségfüggvény tulajdonságai

Tegyük fel, hogy $(X_1, \ldots, X_n)$ abszolút folytonos valószínűségi vektorváltozó, melynek együttes sűrűségfüggvénye $f : \mathbb{R}^n \to \mathbb{R}$. Ekkor

- $f(s_1, \ldots, s_n) \geq 0$ "majdnem mindenütt", és $\int_{\mathbb{R}^n} f = 1$.
- Az X_j sűrűségfüggvénye:

$$f_j(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(s_1, s_2, \ldots, s_{j-1}, t, s_{j+1}, \ldots, s_n) ds_1 \ldots ds_{j-1} ds_{j+1} \ldots ds_n.$$

Például $n = 2$-re

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy; \qquad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Az együttes sűrűségfüggvény tulajdonságai

Tegyük fel, hogy $(X_1, \ldots, X_n)$ abszolút folytonos valószínűségi vektorváltozó, melynek együttes sűrűségfüggvénye $f : \mathbb{R}^n \to \mathbb{R}$. Ekkor

- $f(s_1, \ldots, s_n) \geq 0$ "majdnem mindenütt", és $\int_{\mathbb{R}^n} f = 1$.
- Az X_j sűrűségfüggvénye:

$$f_j(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(s_1, s_2, \ldots, s_{j-1}, t, s_{j+1}, \ldots, s_n) ds_1 \ldots ds_{j-1} ds_{j+1} \ldots ds_n.$$

  Például $n = 2$-re

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy; \qquad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

- $X_1, \ldots, X_n$ pontosan akkor függetlenek, ha minden $(s_1, \ldots, s_n) \in \mathbb{R}^n$-re

$$f(s_1, \ldots, s_n) = f_1(s_1) \cdot f_2(s_2) \cdot \ldots \cdot f_n(s_n).$$

A várható érték tulajdonságai

Állítás

Legyenek $X, Y, X_1, X_2, \ldots, X_n$ olyan valószínűségi változók, melyeknek várható értéke létezik. Ekkor a következők teljesülnek.

(a) *Ha $a \leq X \leq b$ teljesül 1 valószínűséggel valamely a, b valós számokra, akkor $a \leq \mathbb{E}(X) \leq b$.*

(b) **Konstanssal szorzás.** *Ha $c \in \mathbb{R}$, akkor*

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

(c) **Összeg várható értéke.** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. *Hasonlóképpen,*

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

(d) **Szorzat várható értéke független esetben.** *Ha az X és Y valószínűségi változók függetlenek, akkor $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$.*

A várható érték kiszámítása

Legyen $g : \mathbb{R}^n \to \mathbb{R}$ függvény. Ekkor, ha $X_1, \ldots, X_n$ diszkrét valószínűségi változók, akkor

$$\mathbb{E}(g(X_1, \ldots, X_n)) = \sum_{(t_1, \ldots, t_n)} g(t_1, \ldots, t_n)\mathbb{P}(X_1 = t_1, \ldots, X_n = t_n),$$

ha a jobb oldal abszolút konvergens, és az összegzés az $X_1, X_2, \ldots, X_n$ lehetséges értékeire történik.

Ha $X_1, \ldots, X_n$ abszolút folytonos valószínűségi változók f együttes sűrűségfüggvénnyel, akkor

$$\mathbb{E}(g(X_1, \ldots, X_n)) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(t_1, \ldots, t_n) f(t_1, \ldots, t_n) dt_1 \ldots dt_n,$$

ha a jobb oldal abszolút konvergens.

A szórásnégyzet tulajdonságai

Állítás

Legyenek $X, Y, X_1, X_2, \ldots, X_n$ olyan valószínűségi változók, melyeknek szórása létezik. Ekkor a következők teljesülnek.

(a) **Nemnegativitás.** $D(X) \geq 0$.

(b) $D^2(X) = 0$ *akkor és csak akkor, ha* $\mathbb{P}(X = c) = 1$ *valamilyen* $c \in \mathbb{R}$ *számra.*

(c) **Konstans hozzáadása** $D^2(X + b) = D^2(X)$ *tetszőleges* $b \in \mathbb{R}$ *számra.*

(d) **Konstanssal való szorzás.** $D^2(a \cdot X) = a^2 D^2(X)$, *és* $D(a \cdot X) = |a| D(X)$ *tetszőleges* $a \in \mathbb{R}$ *számra.*

(e) **Összeg szórásnégyzete független esetben.** *Ha X és Y függetlenek, akkor* $D^2(X+Y) = D^2(X)+D^2(Y)$. *Általánosabban, ha* $X_1, X_2, \ldots, X_n$ *függetlenek, akkor* $D^2(X_1 + \ldots + X_n) = D^2(X_1) + \ldots + D^2(X_n)$.

A kovariancia

Definíció (Kovariancia)

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

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

Állítás

Legyenek $X, Y, Z, X_1, \ldots, X_n$ olyan valószínűségi változók, melyek szórása létezik. Ekkor a következők teljesülnek.

- **A kovariancia kiszámítása.** $\operatorname{cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y)$.
- **Szimmetria.** $\operatorname{cov}(X, Y) = cov(Y, X)$.
- **Kapcsolat a szórásnégyzettel.** $\operatorname{cov}(X, X) = D^2(X)$.

A kovariancia tulajdonságai

Állítás

- **Konstanssal való kovariancia.** $\mathrm{cov}(X, c) = 0$, *ha* $c \in \mathbb{R}$.
- **Linearitás.** *Egyrészt* $\mathrm{cov}(X + Y, Z) = \mathrm{cov}(X, Z) + \mathrm{cov}(Y, Z)$, *másrészt tetszőleges* $c \in \mathbb{R}$ *számra* $\mathrm{cov}(cX, Y) = c \cdot \mathrm{cov}(X, Y)$.
- **Függetlenséggel való kapcsolat.** *Ha az* X *és* Y *valószínűségi változók függetlenek, akkor* $\mathrm{cov}(X, Y) = 0$.
- **Összeg szórásnégyzete.** $D^2(X + Y) = D^2(X) + D^2(Y) + 2\mathrm{cov}(X, Y)$. *Továbbá*
$$D^2\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} D^2(X_i) + 2\sum_{i<j} \mathrm{cov}(X, Y).$$
- **Különbség szórásnégyzete** $D^2(X - Y) = D^2(X) + D^2(Y)$.

Korrelálatlanság

Példa. Legyen X Poisson-eloszlású valószínűségi változó 2 paraméterrel. Ekkor

$$\mathrm{cov}(X+3, 2\cdot X) \overset{(e)}{=} 2\mathrm{cov}(X+3, X) \overset{(e)}{=} 2\mathrm{cov}(X, X) + 2\mathrm{cov}(3, X) =$$
$$\overset{(c,d)}{=} 2D^2(X) = 2\cdot 2 = 4.$$

Definíció (Korrelálatlanság)

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

Állítás (Függetlenség és korrelálatlanság)

Ha az X és Y valószínűségi változók függetlenek és szórásuk létezik, akkor korrelálatlanok.

A korrelálatlanságból nem következik a függetlenség.

Legyen X és Y két szabályos kockadobás, ezek függetlenek. Legyen továbbá $U = X+Y, V = X-Y$. Ekkor, bár $X+Y$ és $X-Y$ nem függetlenek:

$$\mathrm{cov}(X+Y, X-Y) \overset{(e,d)}{=} D^2(X) - \mathrm{cov}(X, Y) + \mathrm{cov}(X, Y) - D^2(X) \overset{(f)}{=} 0.$$

Korrelációs együttható

Definíció

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

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

Korrelációs együttható

Definíció

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

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

Állítás

Legyenek X és Y olyan valószínűségi változók, melyek szórása létezik.

(i) *Ekkor teljesül, hogy*

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

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

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

(iii) *Tegyük fel, hogy $|R(X,Y)| = 1$. Ekkor léteznek olyan a és b valós számok, hogy az $Y = aX + b$ egyenlet 1 valószínűséggel teljesül.*

Többdimenziós normális eloszlás

Az X valószínűségi változó normális eloszlású m várható értékkel és σ^2 szórásnégyzettel, ha sűrűségfüggvénye $f(t) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right)$.

Standard normális eloszlás: $m = 0, \sigma = 1$.

Definíció

Az $(X_1, \ldots, X_n)$ valószínűségi vektorváltozó standard normális eloszlású, ha $X_1, \ldots, X_n$ független standard normális eloszlású valószínűségi változók. Az együttes sűrűségfüggvény:

$$f(t_1, \ldots, t_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{t_1^2 + \ldots + t_n^2}{2}\right).$$

Definíció

Az $\underline{Y} = (Y_1, \ldots, Y_k)$ valószínűségi vektorváltozó **együttesen normális eloszlású**, *ha felírható $\underline{Y} = A\underline{X} + \underline{b}$ alakban, ahol $A \in \mathbb{R}^{k\times n}$ mátrix, $\underline{b} \in \mathbb{R}^b$ vektor, $\underline{X} = (X_1, \ldots, X_n)$ pedig standard normális valószínűségi vektorváltozó.*

Többdimenziós normális eloszlás

Legyen $\underline{Y} = (Y_1, \ldots, Y_k)$ együttesen normális eloszlású valószínűségi vektorváltozó. Ekkor:

- Ha $B \in \mathbb{R}^{l \times k}$ tetszőleges mátrix, $\underline{v} \in \mathbb{R}^l$ pedig egy vektor, akkor $\underline{Z} = B\underline{Y} + \underline{v}$ is együttesen normális eloszlású.
- Y_j is normális eloszlású (minden $1 \leq j \leq k$-ra).
- Abból, hogy Y_1 és Y_2 normális eloszlásúak (külön-külön), nem következik, hogy az (Y_1, Y_2) valószínűségi vektorváltozó együttesen normális eloszlású.
- Ha (Y_1, Y_2) együttesen normális eloszlású vektor, és $\mathrm{cov}(Y_1, Y_2) = 0$, akkor Y_1 és Y_2 függetlenek.
- Legyen $\underline{\mu} = \mathbb{E}(\underline{Y})$ a **várhatóérték-vektor**, $\Sigma = ((\mathrm{cov}(Y_i, Y_j)))$ pedig a **kovarianciamátrix**. Ha $\det \Sigma \neq 0$, akkor $\underline{Y}$ együttes sűrűségfüggvénye:

$$f(\underline{t}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\underline{t} - \underline{\mu})^T \Sigma^{-1} (\underline{t} - \underline{\mu}) \right).$$

Független valószínűségi változók összegei

- Ha X, Y független normális eloszlású valószínűségi változók, $X \sim N(m_1, \sigma_1^2)$, $Y \sim N(m_2, \sigma_2^2)$, akkor $X+Y$ normális eloszlású m_1+m_2 várható értékkel és $\sigma_1^2+\sigma_2^2$ szórásnégyzettel.
- Ha X és Y független Poisson-eloszlású valószínűségi változók, és X paramétere λ_1, Y paramétere λ_2, akkor $X+Y$ Poisson-eloszlású $\lambda_1+\lambda_2$ paraméterrel.
- Ha X és Y függetlenek, binomiális eloszlásúak, a rendjük n_1 és n_2, a paramétere pedig mindkettőnek p, akkor $X+Y$ binomiális eloszlású n_1+n_2 renddel és p paraméterrel.

  Ha X és Y függetlenek, negatív binomiális eloszlásúak, a rendjük r_1 és r_2, a paramétere pedig mindkettőnek p, akkor $X+Y$ negatív binomiális eloszlású r_1+r_2 renddel és p paraméterrel.

Megjegyzés. Független exponenciális eloszlású valószínűségi változók összege úgynevezett gamma-eloszlású.

Konvolúció

Állítás

Legyenek X és Y olyan független valószínűségi változók, melyek lehetséges értékei egész számok. Ekkor

$$\mathbb{P}(X+Y=k)=\sum_{i=-\infty}^{\infty}\mathbb{P}(X=i)\mathbb{P}(Y=k-i).$$

Állítás

Legyenek X és Y egymástól független, abszolút folytonos valószínűségi változók. Legyen az X sűrűségfüggvénye f, az Y-é pedig g. Ekkor az $X+Y$ valószínűségi változó is abszolút folytonos, és sűrűségfüggvénye:

$$h_{X+Y}(t)=\int_{-\infty}^{\infty}f(s)g(t-s)ds.$$

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, $s > 0$ pozitív szám. Ekkor

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

Következmény

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

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

A nagy számok gyenge törvénye

Definíció (Sztochasztikus konvergencia)

Legyen $X_1, X_2, \ldots$ valószínűségi változók sorozata. Azt mondjuk, hogy ez a sorozat sztochasztikusan konvergál az Y valószínűségi változóhoz, ha minden $\varepsilon > 0$-ra

$$\mathbb{P}(|X_n - Y| > \varepsilon) \to 0$$

teljesül $n \to \infty$ esetén.

Tétel (A nagy számok gyenge törvénye)

Legyenek $X_1, X_2, \ldots$ olyan valószínűségi változók, melyek függetlenek, és azonos eloszlásúak (vagyis eloszlásfüggvényük megegyezik). Tegyük fel, hogy $D(X_1)$ létezik. Ekkor

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mathbb{E}(X_1)$$

sztochasztikusan $n \to \infty$ esetén.

Ez tehát azt jelenti, hogy ha $\overline{X}_n = (X_1 + X_2 + \ldots + X_n)/n$ jelöli az első n valószínűségi változó átlagát, akkor minden $\varepsilon > 0$ esetén

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}(X_1)| > \varepsilon) \to 0 \qquad (n \to \infty).$$

A nagy számok erős törvénye

Definíció (1 valószínűségű konvergencia)

Legyen $X_1, X_2, \ldots$ valószínűségi változók sorozata. Azt mondjuk, hogy ez a sorozat 1 valószínűséggel konvergál a Z valószínűségi változóhoz, ha

$$\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to Z(\omega)\}) = 1.$$

Ha $X_n \to Z$ teljesül 1 valószínűséggel, akkor $X_n \to Z$ sztochasztikusan is, de a megfordítás nem igaz.

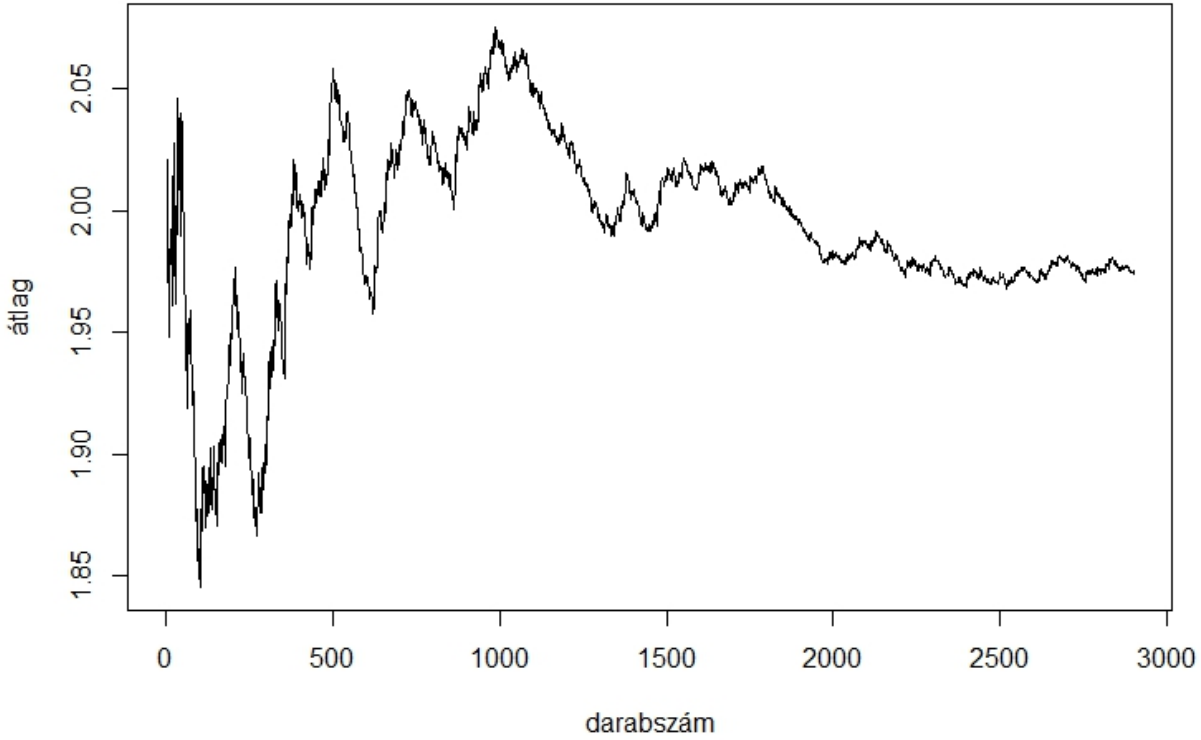
Tétel (A nagy számok erős törvénye)

Legyenek $X_1, X_2, \ldots$ valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel még, hogy $\mathbb{E}(X_1)$ létezik. Ekkor

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mathbb{E}(X_1)$$

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

A nagy számok erős törvénye



5. ábra.
Az átlag változása a darabszám függvényében független exp(0, 5) eloszlású mintánál

Centrális határeloszlástétel

Definíció (Eloszlásbeli konvergencia)

Legyen $X_1, X_2, \ldots$ valószínűségi változók sorozata, X_i eloszlásfüggvénye F_i ($i = 1, 2, \ldots$ esetén). Legyen továbbá Y valószínűségi változó, melynek eloszlásfüggvénye F. Azt mondjuk, hogy az $(X_n)_{n\in\mathbb{N}}$ sorozat tart Y-hoz eloszlásban, ha

$$F_n(t) \to F(t) \qquad (n \to \infty)$$

teljesül minden olyan $t \in \mathbb{R}$-re, melyre F folytonos t-ben.

Tétel (Centrális határeloszlástétel)

Legyenek $X_1, X_2, \ldots$ független azonos eloszlású valószínűségi változók, melyeknek szórása létezik. Használjuk a következő jelöléseket: $\mathbb{E}(X_1) = m$ és $D(X_1) = s$. Legyen Y standard normális valószínűségi változó: $Y \sim N(0, 1)$. Ekkor

$$\frac{X_1 + X_2 + \ldots + X_n - n \cdot m}{s\sqrt{n}} \to Y$$

eloszlásban $n \to \infty$ esetén.

Centrális határeloszlástétel

Legyenek $X_1, X_2, \ldots$ független azonos eloszlású valószínűségi változók, melyeknek szórása létezik. Használjuk a következő jelöléseket: $\mathbb{E}(X_1) = m$ és $D(X_1) = s$. Ekkor

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

A határértéket $\Phi(b) - \Phi(a)$ alakban is írhatjuk.

Centrális határeloszlástétel

Legyenek $X_1, X_2, \ldots$ független azonos eloszlású valószínűségi változók, melyeknek szórása létezik. Használjuk a következő jelöléseket: $\mathbb{E}(X_1) = m$ és $D(X_1) = s$. Ekkor

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

A határértéket $\Phi(b) - \Phi(a)$ alakban is írhatjuk.

Így is átfogalmazható a tétel állítása:

$$\mathbb{P}(nm + as\sqrt{n} \leq X_1 + X_2 + \ldots + X_n < nm + bs\sqrt{n}) \to \Phi(b) - \Phi(a).$$

Statisztikai mező

Definíció

Az $(\Omega, \mathcal{A}, \mathcal{P})$ *hármast* **statisztikai mezőnek** *nevezzük, ha minden* $\mathbb{P} \in \mathcal{P}$*-re* $(\Omega, \mathcal{A}, \mathbb{P})$ *Kolmogorov-féle valószínűségi mező.*

Definíció

Ha valamilyen $\Theta \subseteq \mathbb{R}^q$ *halmazra a* $\mathcal{P}$ *halmaz felírható* $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ *alakban, akkor paraméteres statisztikai problémáról beszélhetünk. Ilyenkor a* Θ *halmazt paramétertérnek nevezzük.*

Statisztikai mező

Definíció

Az $(\Omega, \mathcal{A}, \mathcal{P})$ *hármast* **statisztikai mezőnek** *nevezzük, ha minden* $\mathbb{P} \in \mathcal{P}$*-re* $(\Omega, \mathcal{A}, \mathbb{P})$ *Kolmogorov-féle valószínűségi mező.*

Definíció

Ha valamilyen $\Theta \subseteq \mathbb{R}^q$ *halmazra a* $\mathcal{P}$ *halmaz felírható* $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ *alakban, akkor paraméteres statisztikai problémáról beszélhetünk. Ilyenkor a* Θ *halmazt paramétertérnek nevezzük.*

Definíció (Minta)

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ *statisztikai mező. Egy*

$$\underline{X} = (X_1, X_2, \ldots, X_n) : \Omega \to H \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót (n elemű) **mintának** *nevezünk. Itt* H *a mintatér,* n *a minta elemszáma vagy nagysága. Az* X_i *koordináták a minta elemei. Azt mondjuk, hogy a minta független, ha az* $X_1, X_2, \ldots, X_n$ *valószínűségi változók függetlenek.*

Statisztika

Definíció

A mintatéren megadott $T : H \to \mathbb{R}^k$ függvényt, illetve a $T = T(\underline{X})$ valószínűségi változót (k-dimenziós) **statisztikának** *nevezzük.*

Példa. Legyenek $X_1, X_2, \ldots, X_n$ független azonos eloszlású valószínűségi változók, melyek exponenciális eloszlásúak $\vartheta > 0$ paraméterrel. Ekkor $\Theta = (0, \infty)$, $H = (0, \infty)^n$. Statisztikára példák:

- **mintaátlag**: $\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$.

Statisztika

Definíció

A mintatéren megadott $T : H \to \mathbb{R}^k$ függvényt, illetve a $T = T(\underline{X})$ valószínűségi változót (k-dimenziós) **statisztikának** *nevezzük.*

Példa. Legyenek $X_1, X_2, \ldots, X_n$ független azonos eloszlású valószínűségi változók, melyek exponenciális eloszlásúak $\vartheta > 0$ paraméterrel. Ekkor $\Theta = (0, \infty)$, $H = (0, \infty)^n$. Statisztikára példák:

- **mintaátlag**: $\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$.
- **tapasztalati szórásnégyzet**: $s_n^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2 = \frac{1}{n} \sum_{j=1}^{n} X_j^2 - \overline{X}^2$.
- **tapasztalati szórás**: $s_n = \sqrt{\frac{1}{n} \sum_{j=1}^{n} X_j^2 - \overline{X}^2}$.

Leíró statisztikák

- **mintaátlag** (mean): $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$.
- **korrigált tapasztalati szórásnégyzet** (variance):

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

- korrigált tapasztalati szórás (standard deviation, sd): $s_n^* = \sqrt{s_n^{*2}}$.
- minimum: a legkisebb mintaelem, azaz $\min(X_1, X_2, \ldots, X_n)$.
- maximum: a legnagyobb mintaelem, azaz $\max(X_1, X_2, \ldots, X_n)$.
- terjedelem (range): a legnagyobb és legkisebb mintaelem különbsége, azaz

$$\max(X_1, X_2, \ldots, X_n) - \min(X_1, X_2, \ldots, X_n).$$

- módusz (mode): a leggyakrabban előforduló mintaelem.

Példa: az adatok elemzése

A Duna vízállása húsz napon át Budapestnél így alakult (centiméterben mérve):

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

Példa: az adatok elemzése

A Duna vízállása húsz napon át Budapestnél így alakult (centiméterben mérve):

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

a mintaelemek száma: $n = 20$

legkisebb: 99, legnagyobb: 218, terjedelem: $218 - 99 = 119$

átlag: 149,9, medián: 141,5 (a középső mintaelem)

korrigált tapasztalati szórás: 38,55

Példa: az adatok elemzése

A Duna vízállása húsz napon át Budapestnél így alakult (centiméterben mérve):

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

a mintaelemek száma: $n = 20$

legkisebb: 99, legnagyobb: 218, terjedelem: $218 - 99 = 119$

átlag: 149,9, medián: 141,5 (a középső mintaelem)

korrigált tapasztalati szórás: 38,55

A vízállás 5 napon volt 115 cm-nél kevesebb (a napok egynegyedén), és 3 napon haladta meg a 2 métert (a napok 15%-án).

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

átlag: $\overline{X} = 149,9$

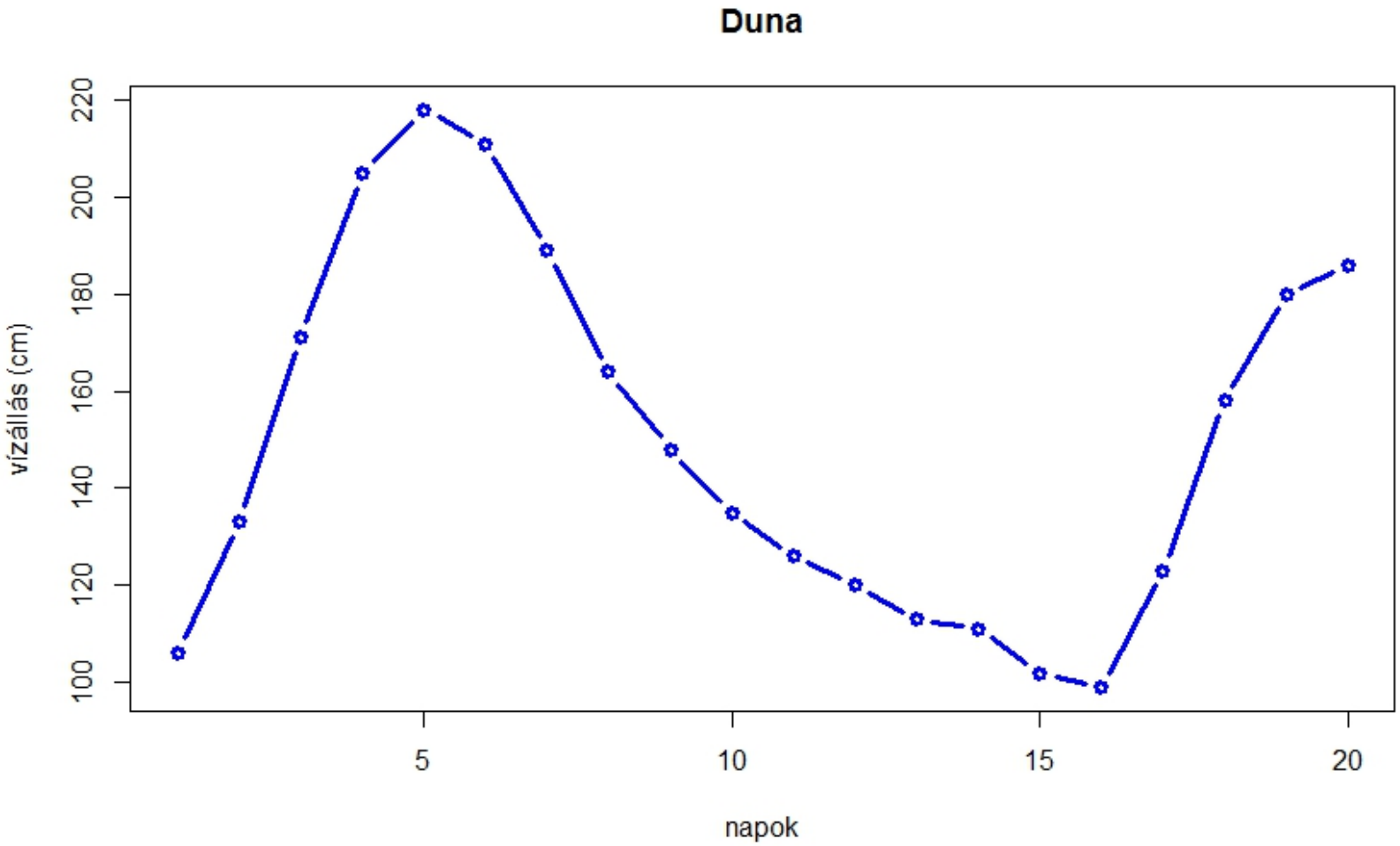
tapasztalati szórásnégyzet: $s_n^2 = 1412,09$

tapasztalati szórás: $s_n = 37,58$

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

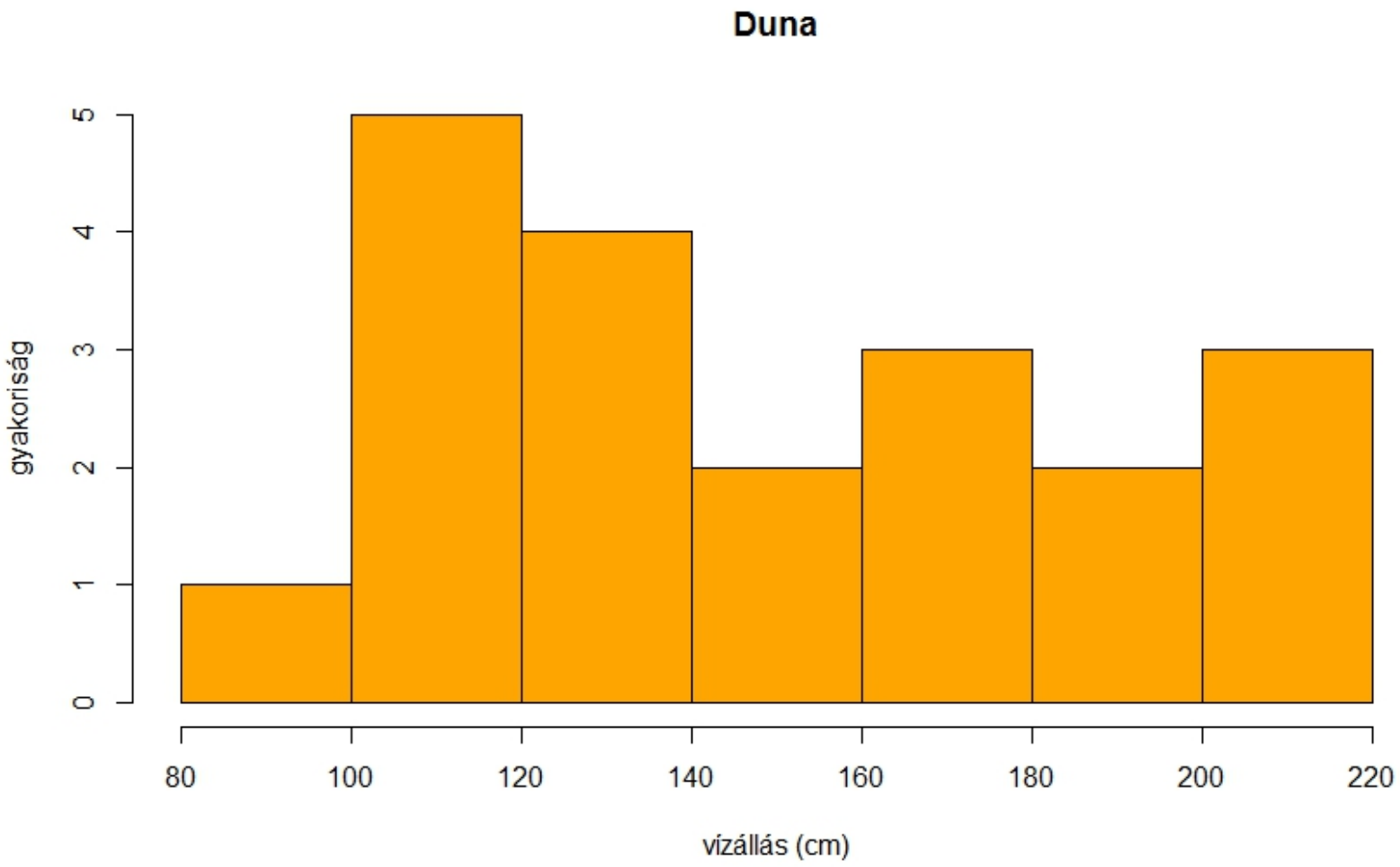
korrigált tapasztalati szórás: $s_n^* = 38,55$

Példa: az adatok ábrázolása

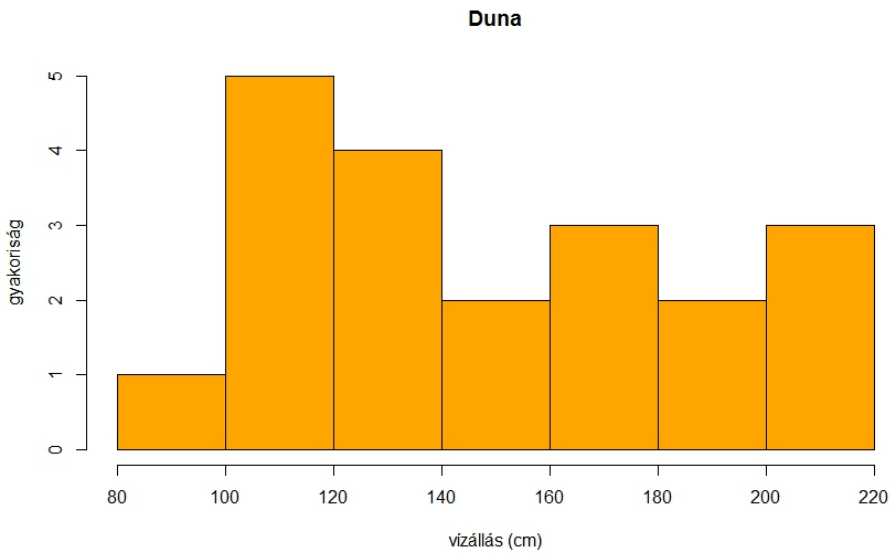


Példa: hisztogram

A Duna vízállásának hisztogramja a húsznapos adatsorból



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.

Momentumok

Definíció

Legyen X valószínűségi változó, $k \geq 1$ egész szám. Ekkor az X valószínűségi változó k. momentuma:

$$\mathbb{E}(X^k),$$

ha ez a várható érték létezik.

Legyen $X_1, X_2, \ldots, X_n$ minta. Ekkor a minta k. tapasztalati momentuma:

$$\frac{1}{n}\sum_{j=1}^{n} X_j^k.$$

Rendezett minta

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

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

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

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:

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

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

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

Példa: a vízállásról 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

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

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

Medián

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

Definíció (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.

Megjegyzés: páros n esetén a teljes $\left[X^*_{n/2}, X^*_{n/2+1}\right]$ intervallumot (vagy annak bármely elemét) is a minta mediánjának lehet hívni.

Példa: a vízállásról kapott húszelemű minta mediánja:

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

Az átlag és a medián összehasonlítása

Normális eloszlás

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

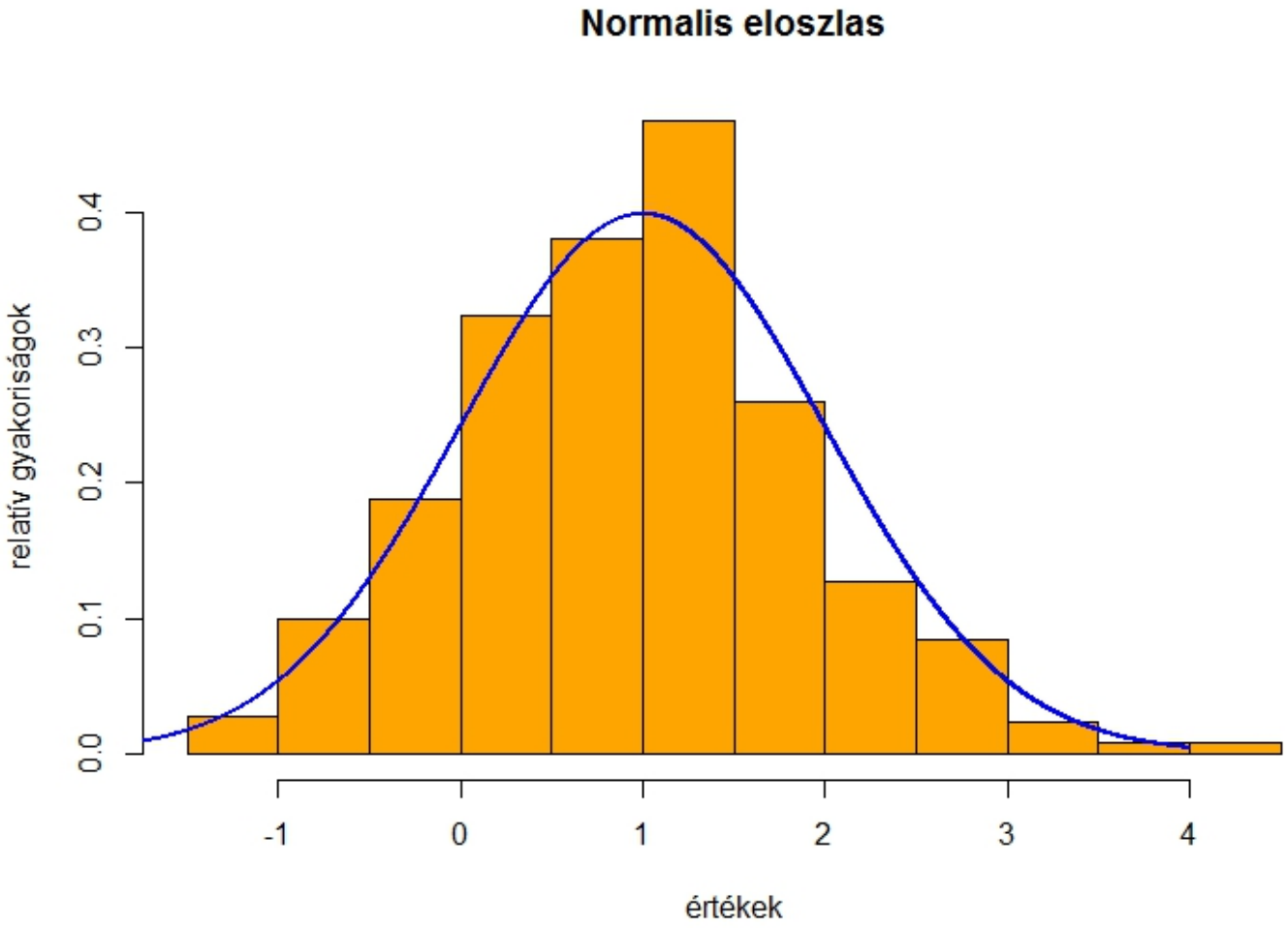
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9840	0.2847	0.9842	0.9863	1.6930	3.6110

Exponenciális eloszlás

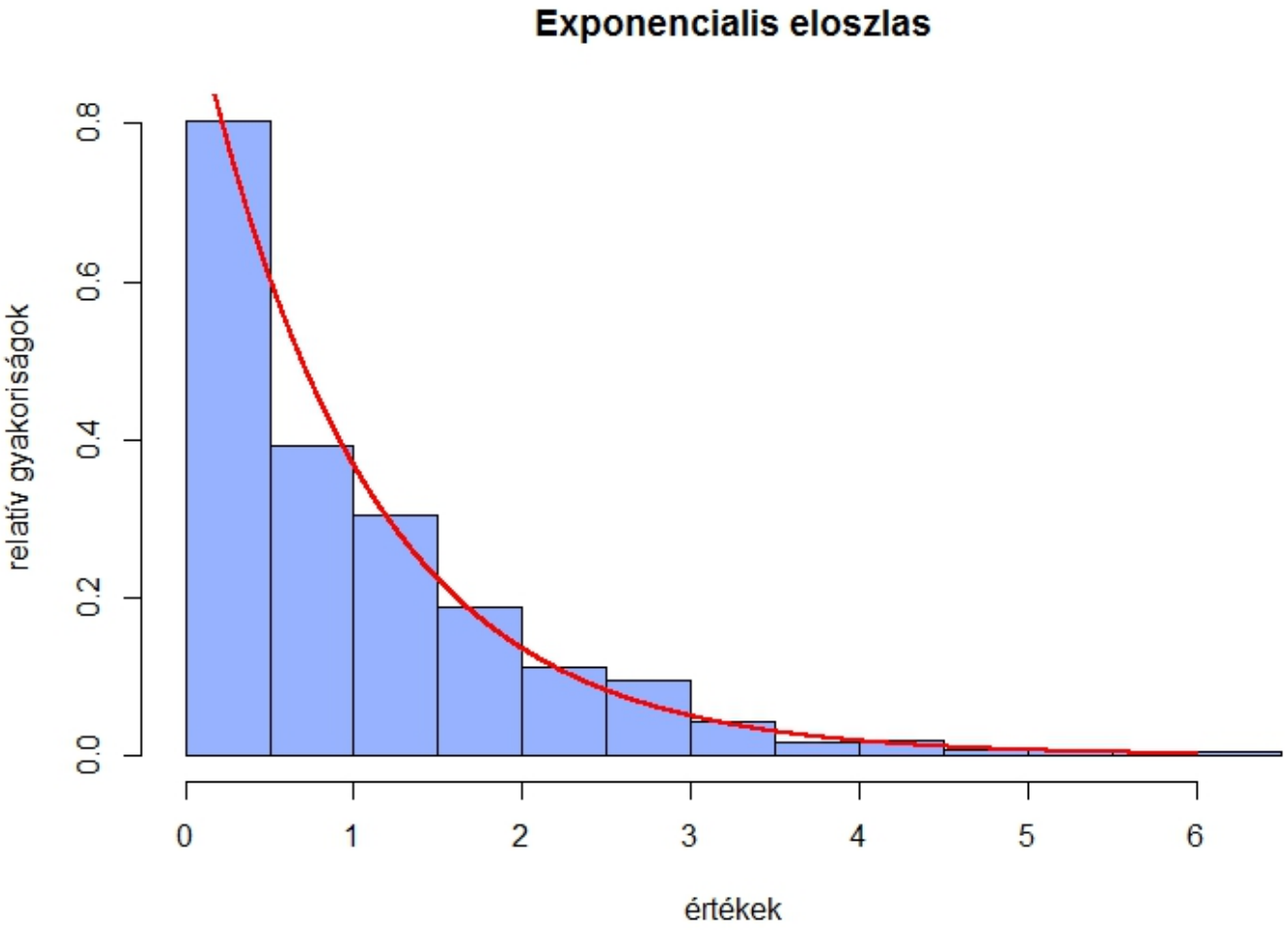
500 elemű független minta: $Y_1, Y_2, \ldots, Y_{500}$ függetlenek, eloszlásuk exponenciális eloszlás $b = 1$ paraméterrel. $\mathbb{E}(Y_k) = 1$ és $D(Y_k) = 1$ minden $k = 1, 2, \ldots, 500$-ra.

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 $m = 1$ várható értékű és $\sigma = 1$ szórású normális eloszlás sűrűségfüggvénye szimmetrikus az 1 körül:

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-1)^2}{2} \right) \qquad (t \in \mathbb{R}).$$

Az 1 paraméterű exponenciális eloszlás sűrűségfüggvénye nem ilyen:

$$g(t) = \begin{cases} \exp(-t), & \text{ha } t > 0; \\ 0, & \text{ha } t < 0. \end{cases}$$

Ha a sűrűségfüggvény szimmetrikus, akkor az átlag és a medián általában közelebb esik egymáshoz, mint ha nem érvényes a szimmetria. Ezért ha az adatok semmilyen szimmetriát nem mutatnak, gyakran a mediánt adják meg. Szimmetrikus esetben inkább az átlagot használják.

Tapasztalati eloszlásfüggvény

Legyen X tetszőleges valószínűségi változó. Ennek eloszlásfüggvénye az az $F : \mathbb{R} \to [0, 1]$ függvény, melyre

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

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

Tapasztalati eloszlásfüggvény

Legyen X tetszőleges valószínűségi változó. Ennek eloszlásfüggvénye az az $F : \mathbb{R} \to [0,1]$ függvény, melyre

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

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

Definíció (Tapasztalati eloszlásfüggvény)

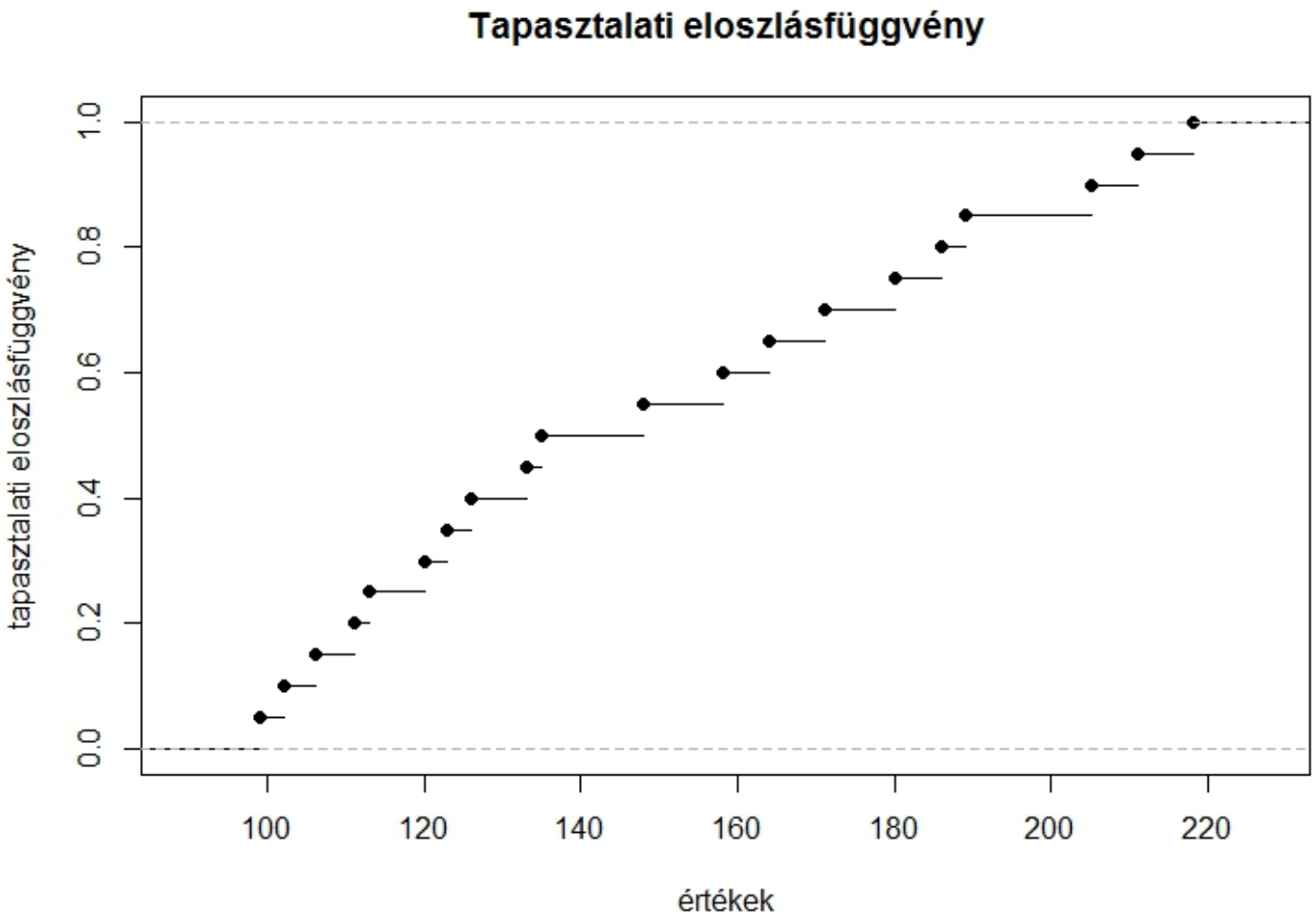
Legyenek $X_1, X_2, \ldots, X_n$ valószínűségi változók. Ennek a mintának az eloszlásfüggvénye az az $\hat{F}_n : \mathbb{R} \to [0,1]$ függvény, melyre

$$\hat{F}_n(t) = \frac{t\text{-nél kisebb mintaelemek száma}}{n} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}(X_k \leq t).$$

Itt $\mathbb{I}(X_k \leq t)$ értéke 1, ha $X_k \leq t$ teljesül (azaz a k. mintaelem legfeljebb t), és 0 különben.

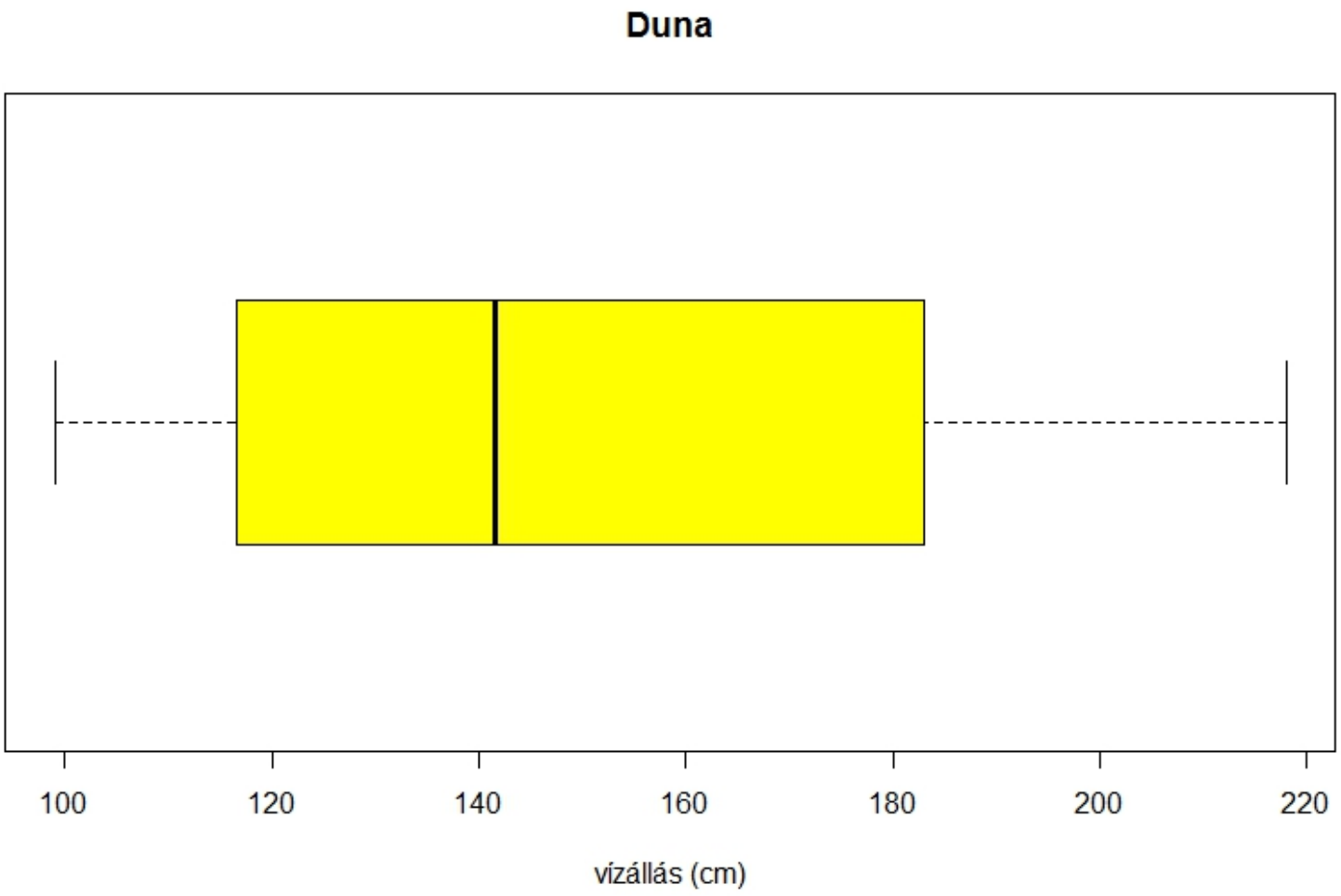
Példa: tapasztalati eloszlásfüggvény

A Duna vízállásának tapasztalati eloszlásfüggvénye



Példa: boxplot

A Duna vízállásának boxplotja a húsznapos adatsorból



Boxplot

Definíció (Tapasztalati kvantilis)

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

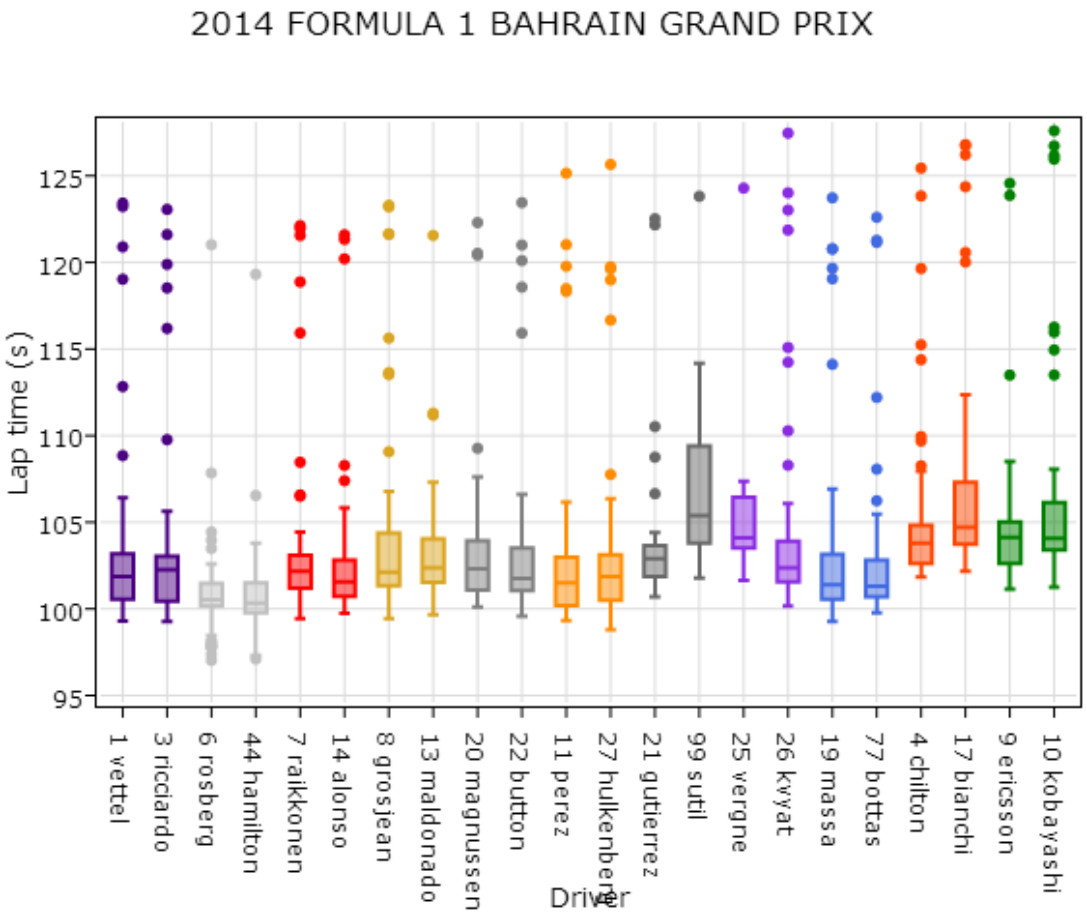
$$\hat{q}_z = \min\{t : \hat{F}_n(t) \geq z\}.$$

A boxplot készítéséhez szükséges adatok:

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

terjedelem: maximum - minimum (119).

Példa: boxplot



6. ábra. Forrás: theansweris27.com

Torzítatlan becslés

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta)$ valamely Θ halmazzal (Θ a paramétertér);
- $\psi : \Theta \to \mathbb{R}$ függvény.
- Cél: olyan T statisztika keresése, amire a $T(X)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Torzítatlan becslés

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta)$ valamely Θ halmazzal (Θ a paramétertér);
- $\psi : \Theta \to \mathbb{R}$ függvény.
- Cél: olyan T statisztika keresése, amire a $T(X)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Definíció (Torzítatlanság)

A $T : H \to \mathbb{R}$ statisztika torzítatlan becslés ψ-re, ha minden $\vartheta \in \Theta$-ra

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta).$$

A T statisztika torzítása a $b_T(\vartheta) = \mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) - \psi(\vartheta)$ függvény.

Példa. $X_1, X_2, \ldots, X_n$ független minta a $[0, \vartheta]$ intervallumon egyenletes eloszlásból. Ekkor $2\overline{X}$ torzítatlan becslés $\psi(\vartheta) = \vartheta$-ra.

Torzítatlan becslések

Állítás (A várható érték torzítatlan becslése)

Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta. Legyen $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$, azaz a mintának a $\mathbb{P}_\vartheta$ eloszlás szerinti várható értéke. Ekkor a $T(X_1, \ldots, X_n) = \overline{X}$ statisztika, vagyis a **mintaátlag** *torzítatlan becslés ψ-re.*

Állítás (A szórásnégyzet torzítatlan becslése)

$X_1, \ldots, X_n$ független azonos eloszlású minta. Legyen $\psi(\vartheta) = D^2_\vartheta(X_1)$, azaz a mintának a $\mathbb{P}_\vartheta$ eloszlás szerinti szórásnégyzete. Ekkor a $T(X_1, \ldots, X_n) = s_n^{*2}$ statisztika, vagyis a **korrigált tapasztalati szórásnégyzet** *torzítatlan becslés ψ-re.*

Hatásosság

Definíció (Hatásosság)

Legyenek T_1, T_2 torzítatlan becslései a paraméter $\psi(\vartheta)$ függvényének. T_1 hatásosabb T_2-nél, ha $D^2_\vartheta(T_1) \leq D^2_\vartheta(T_2)$ teljesül minden $\vartheta \in \Theta$-ra.
A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.

Hatásosság

Definíció (Hatásosság)

Legyenek T_1, T_2 torzítatlan becslései a paraméter $\psi(\vartheta)$ függvényének. T_1 hatásosabb T_2-nél, ha $D^2_\vartheta(T_1) \leq D^2_\vartheta(T_2)$ teljesül minden $\vartheta \in \Theta$-ra.
A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.

Állítás

Legyen $(X_1, \ldots, X_n)$ független azonos eloszlású minta véges szórású eloszlásból. Ekkor $\psi(\vartheta) = \mathbb{E}_\vartheta(X_i)$-re a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél, ahol $0 \leq c_j$ és $\sum_{j=1}^n c_j = 1$.

Az állítás a számtani és négyzetes közepek közötti egyenlőtlenségből adódik. Ugyanakkor a mintaátlag nem minden esetben hatásos becslése a várható értéknek, csak a lineáris kombinációknál hatásosabb.

Konzisztencia

Definíció

A $T_n = T_n(X_1, \ldots, X_n)$ **konzisztens** *becsléssorozat* $\psi(\vartheta)$*-ra, ha minden* $\vartheta \in \Theta$*-ra*

$$(T_n(X_1, \ldots, X_n)) \to \psi(\vartheta)$$

$n \to \infty$ *esetén sztochasztikusan, azaz minden* $\vartheta \in \Theta$ *és* $\varepsilon > 0$*-ra teljesül, hogy*

$$\mathbb{P}_\vartheta\big(|T_n - \psi(\vartheta)| > \varepsilon\big) \to 0 \qquad (n \to \infty).$$

Példa. $X_1, X_2, \ldots$ független azonos eloszlású minta. Ekkor $T_n = \frac{X_1+X_2+\ldots+X_n}{n}$ konzisztens becsléssorozat $\mathbb{E}_\vartheta(X_1)$-re, hiszen a nagy számok gyenge törvénye szerint $T_n \to \mathbb{E}_\vartheta(X_1)$ sztochasztikusan.

Továbbá, ha például $X_1, X_2, \ldots$ függetlenek és $N(m, \sigma^2)$ eloszlásúak, akkor az átlag konzisztens m-re, s_n^* pedig σ-ra (s_n is konzisztens σ-ra).

Maximumlikelihood-módszer

Definíció (Likelihood-függvény)

Legyen $Y_1, \ldots, Y_n$ minta. Ha ezek abszolút folytonosak, és Y_j sűrűségfüggvénye (a $\mathbb{P}_\vartheta$-re vonatkozóan) $f_{j,\vartheta}$, akkor a minta likelihood-függvénye:

$$L_{n,\vartheta}(t_1, \ldots, t_n) = \prod_{j=1}^{n} f_{j,\vartheta}(t_j) \qquad (t_1, \ldots, t_n \in \mathbb{R}).$$

Ha a minta diszkrét, akkor a minta likelihood-függvénye:

$$L_{n,\vartheta}(k_1, \ldots, k_n) = \prod_{j=1}^{n} \mathbb{P}_{j,\vartheta}(Y_j = k_j) \qquad ((k_1, \ldots, k_n) \in H).$$

Maximumlikelihood-módszer

Definíció (Maximum-likelihood becslés)

A ϑ maximumlikelihood-becslése (ML-becslése) az $X_1, \ldots, X_n$ mintából $\hat{\vartheta}$, ha $\hat{\vartheta}$ maximalizálja a $\vartheta \mapsto L_{n,\vartheta}(X_1, \ldots, X_n)$ függvényt, ahol $L_{n,\vartheta}$ a minta likelihood-függvénye. Azaz, ha

$$L_{n,\hat{\vartheta}}(X_1, \ldots, X_n) \geq L_{n,\vartheta}(X_1, \ldots, X_n) \text{ minden } \vartheta \in \Theta\text{-ra.}$$

Példa. $X_1, \ldots, X_n$ függetlenek, eloszlásuk exponenciális eloszlás $\vartheta > 0$ paraméterrel. Ekkor

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \prod_{j=1}^{n} f_{j,\vartheta}(X_j) = \prod_{j=1}^{n} \Big[\vartheta \exp(-\vartheta X_j) \mathbb{I}(X_j > 0)\Big],$$

amiből $\hat{\vartheta} = \frac{1}{\overline{X}}$.

ML-becslés: példa

$X_1, \ldots, X_n$ függetlenek, eloszlásuk exponenciális eloszlás $\vartheta > 0$ paraméterrel. Ekkor

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \prod_{j=1}^{n} f_{j,\vartheta}(X_j) = \prod_{j=1}^{n} \Big[\vartheta \exp(-\vartheta X_j) \mathbb{I}(X_j > 0)\Big].$$

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \vartheta^n \exp\Big(-\vartheta \sum_{j=1}^{n} X_j \Big).$$

ML-becslés: példa

$X_1, \ldots, X_n$ függetlenek, eloszlásuk exponenciális eloszlás $\vartheta > 0$ paraméterrel. Ekkor

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \prod_{j=1}^{n} f_{j,\vartheta}(X_j) = \prod_{j=1}^{n} \Big[\vartheta \exp(-\vartheta X_j) \mathbb{I}(X_j > 0)\Big].$$

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \vartheta^n \exp\Big(- \vartheta \sum_{j=1}^{n} X_j \Big).$$

$$\ln L_{n,\vartheta}(X_1, \ldots, X_n) = n \ln \vartheta - \vartheta \sum_{j=1}^{n} X_j$$

ML-becslés: példa

$X_1, \ldots, X_n$ függetlenek, eloszlásuk exponenciális eloszlás $\vartheta > 0$ paraméterrel. Ekkor

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \prod_{j=1}^{n} f_{j,\vartheta}(X_j) = \prod_{j=1}^{n} \Big[\vartheta \exp(-\vartheta X_j) \mathbb{I}(X_j > 0)\Big].$$

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \vartheta^n \exp\Big(-\vartheta \sum_{j=1}^{n} X_j\Big).$$

$$\ln L_{n,\vartheta}(X_1, \ldots, X_n) = n \ln \vartheta - \vartheta \sum_{j=1}^{n} X_j$$

$$\frac{\partial}{\partial \vartheta} \ln L_{n,\vartheta}(X_1, \ldots, X_n) = \frac{n}{\vartheta} - n\overline{X} > 0 \Leftrightarrow \vartheta < 1/\overline{X}.$$

Az ML-becslés tulajdonságai

- Nem minden statisztikai mezőn létezik ML-becslés.
- Az ML-becslés nem feltétlenül egyértelmű.
- A $\psi(\vartheta)$ függvény ML-becslése $\psi(\hat{\vartheta})$, ahol $\hat{\vartheta}$ ML-becslés ϑ-ra.
- Megfelelő feltételek (erős regularitási feltételek mellett) az ML-becslés aszimpotikusan torzítatlan, és aszimptotikusan normális eloszlású, azaz $\sqrt{n}(\hat{\vartheta}_n - \vartheta)$ normális eloszláshoz konvergál eloszlásban $n \to \infty$ esetén (a $\mathbb{P}_\vartheta$ valószínűségre vonatkozóan).
- Az alábbi egyenlet a maximumlikelihood-egyenlet:

$$\frac{\partial}{\partial \vartheta} \ln L_{n,\vartheta}(X_1, \ldots, X_n) = 0.$$

Megfelelő feltételek mellett az ML-becslés a maximumlikelihood-egyenlet megoldása (ha az ML-becslés nem számítható ki, de az egyenlet megoldható, gyakran az egyenlet megoldásával helyettesítik az ML-becslést).

Momentummódszer

Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta.

1. Az eloszlás k. momentuma: $\mu_{k,\vartheta} = \mathbb{E}_\vartheta(X_1^k)$.
2. Legyen $\hat{\mu}_k = \frac{1}{n}\sum_{j=1}^n X_j^k$ az eloszlás k. tapasztalati momentuma.
3. Írjuk fel az alábbi egyenleteket a legkisebb olyan k-ig, amire igaz, hogy az egyenletrendszer egyértelműen meghatározza ϑ-t:

$$\mathbb{E}_\vartheta(X_1) = \frac{1}{n}\sum_{j=1}^n X_j;$$

$$\mathbb{E}_\vartheta(X_1^2) = \frac{1}{n}\sum_{j=1}^n X_j^2;$$

$$\ldots$$

$$\mathbb{E}_\vartheta(X_1^k) = \frac{1}{n}\sum_{j=1}^n X_j^k.$$

4. A ϑ momentummódszerrel kapott becslése az a $\hat{\vartheta}$, ami megoldása a fenti egyenletrendszernek.

Momentummódszer: példa

A momentummódszerrel kapott becslés nem biztos, hogy létezik, és nem biztos, hogy egyértelmű.

$X_1, \ldots, X_n$ független exponenciális eloszlásúak ϑ paraméterrel.

$$\mathbb{E}_\vartheta(X_1) = \frac{1}{\vartheta} = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}.$$

Ez egyértelműen oldható meg ϑ-ra: $\vartheta = 1/\overline{X}$.

Momentummódszer: példa

A momentummódszerrel kapott becslés nem biztos, hogy létezik, és nem biztos, hogy egyértelmű.

$X_1, \ldots, X_n$ független exponenciális eloszlásúak ϑ paraméterrel.

$$\mathbb{E}_\vartheta(X_1) = \frac{1}{\vartheta} = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}.$$

Ez egyértelműen oldható meg ϑ-ra: $\vartheta = 1/\overline{X}$.

Legyen $X_1, \ldots, X_n$ független minta az $[a, b]$ intervallumon egyenletes eloszlásból. Erre a két módszer különböző eredményt ad.

ML-becsléssel: $\hat{a} = X_1^* = \min(X_1, \ldots, X_n)$, $\hat{b} = X_n^* = \max(X_1, \ldots, X_n)$, míg a momentummódszerrel: $\hat{a} = \overline{X} - \sqrt{3}s_n$; $\hat{b} = \overline{X} + \sqrt{3}s_n$.

Konfidenciaintervallumok

Legyen $\underline{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta, $(\Omega, \mathcal{A}, \mathcal{P})$ pedig statisztikai mező, $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$, és tegyük fel, hogy ϑ valós paraméter, vagyis $\Theta \subseteq \mathbb{R}$.

Definíció

Azt mondjuk, hogy a $(T_1(\underline{X}), T_2(\underline{X}))$ intervallum legalább $1-\alpha$ megbízhatósági szintű konfidenciaintervallum ϑ-ra, ha minden $\vartheta \in \mathbb{R}$ esetén teljesül, hogy

$$\mathbb{P}_\vartheta(T_1(\underline{X}) < \vartheta < T_2(\underline{X})) \geq 1 - \alpha.$$

A konfidenciaintervallum megbízhatósági szintje: $\inf_{\vartheta \in \Theta}\{\mathbb{P}_\vartheta(\vartheta \in (T_1, T_2))\}$.

Konfidenciaintervallum a várható értékre

A következő jelölést fogjuk használni: ha $q \in [0,1]$, akkor $u_q = \Phi^{-1}(q)$, ahol Φ a standard normális eloszlás eloszlásfüggvénye. Vagyis, ha Z standard normális eloszlású valószínűségi változó, akkor

$$q = \mathbb{P}(Z \leq u_q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_q} e^{-s^2/2} ds.$$

Állítás (Konfidenciaintervallum a várható értékre, ismert szórás)

Tegyük fel, hogy $X_1, \ldots, X_n$ független azonos eloszlású normális eloszlású valószínűségi változók, melyek szórása, σ ismert.
Ekkor a

$$(T_1, T_2) = \left(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

intervallum $1-\alpha$ megbízhatósági szintű konfidenciaintervallum az eloszlás várható értékére.

Konfidenciaintervallum a várható értékre

Legyen $t_n(q)$ az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q)\right),$$

ahol $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlású valószínűségi változók (a hányados eloszlása n szabadsági fokú **t-eloszlás**).

Állítás (Konfidenciaintervallum a várható értékre, ismeretlen szórás)

Tegyük fel, hogy $X_1, \ldots, X_n$ független azonos eloszlású normális eloszlású valószínűségi változók (sem a várható értékük, sem a szórásuk nem ismert). Ekkor a

$$(T_1, T_2) = \left(\overline{X} - t_{n-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{s_n^*}{\sqrt{n}},\ \overline{X} + t_{n-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{s_n^*}{\sqrt{n}}\right)$$

intervallum $1 - \alpha$ megbízhatósági szintű konfidenciaintervallum az eloszlás várható értékére.

Hipotézisvizsgálat

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, azaz $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamilyen Θ paramétertérrel. A paraméterteret bontsuk fel két diszjunkt halmaz uniójára: $\Theta = \Theta_0 \cup \Theta_1$, ahol tehát $\Theta_0 \cap \Theta_1 = \emptyset$.

Nullhipotézis. $H_0 : \vartheta \in \Theta_0$.

Ellenhipotézis. $H_1 : \vartheta \in \Theta_1$.

Hipotézisvizsgálat

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, azaz $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamilyen Θ paramétertérrel. A paraméterteret bontsuk fel két diszjunkt halmaz uniójára: $\Theta = \Theta_0 \cup \Theta_1$, ahol tehát $\Theta_0 \cap \Theta_1 = \emptyset$.

Nullhipotézis. $H_0 : \vartheta \in \Theta_0$.

Ellenhipotézis. $H_1 : \vartheta \in \Theta_1$.

A minta $\underline{X} = (X_1, \ldots, X_n)$, a mintatér legyen B (vagyis $(X_1, \ldots, X_n)$ a $B \subseteq \mathbb{R}^n$ halmaz egy véletlen eleme). A mintateret is felbontjuk két diszjunkt halmaz uniójára: $B = B_0 \cup B_1$, ahol $B_0 \cap B_1 = \emptyset$.

Elfogadási tartomány: B_0. Ha $(X_1, \ldots, X_n) \in B_0$, akkor H_0-t elfogadjuk.

Elutasítási (kritikus) tartomány: B_1. Ha $(X_1, \ldots, X_n) \in B_1$, akkor H_0-t elutasítjuk.

Hipotézisvizsgálat

- Elsőfajú hibát vétünk, ha H_0 igaz, és elutasítjuk.
- A próba terjedelme:

$$\alpha = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(\underline{X} \in B_1).$$

- Másodfajú hibát vétünk, ha H_0 nem igaz, és elfogadjuk.
- A próba erőfüggvénye az alábbi $\beta : \Theta_1 \to [0,1]$ függvény:

$$\beta(\vartheta) = \mathbb{P}(\underline{X} \in B_1) \qquad (\vartheta \in \Theta_1).$$

Hipotézisvizsgálat

- Elsőfajú hibát vétünk, ha H_0 igaz, és elutasítjuk.
- A próba terjedelme:

$$\alpha = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(\underline{X} \in B_1).$$

- Másodfajú hibát vétünk, ha H_0 nem igaz, és elfogadjuk.
- A próba erőfüggvénye az alábbi $\beta : \Theta_1 \to [0,1]$ függvény:

$$\beta(\vartheta) = \mathbb{P}(\underline{X} \in B_1) \qquad (\vartheta \in \Theta_1).$$

- p-érték: a legnagyobb olyan terjedelem, ami mellett H_0-t elfogadjuk.

$p < 0,05$: szignifikáns eltérés H_0-tól, statisztikai bizonyíték H_1-re.

$p \geq 0,05$: nincs szignifikáns eltérés H_0-tól.

Egymintás u-próba

A próba a normális eloszlás várható értékére vonatkozik ismert szórás mellett.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m ismeretlen paraméter, $\sigma > 0$ ismert.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n}.$$

- **Kétoldali ellenhipotézis**: $H_0 : m = m_0$; $H_1 : m \neq m_0$.
- Ha $|u| > u_{1-\alpha/2}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A p-érték ilyenkor $2 - 2\Phi(|u|)$.

$p < 0,05$: a várható érték szignifikánsan eltér m_0-tól.

$p \geq 0,05$: nincs szignifikáns eltérés m_0-tól.

Egymintás u-próba

A próba a normális eloszlás várható értékére vonatkozik ismert szórás mellett.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m ismeretlen paraméter, $\sigma > 0$ ismert.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n}.$$

- **Egyoldali ellenhipotézis**: $H_0 : m \leq m_0$; $H_1 : m > m_0$.
- Ha $u > u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A p-érték ilyenkor $1 - \Phi(u)$.

$p < 0,05$: a várható érték szignifikánsan több m_0-nál.

$p \geq 0,05$: a várható érték nem több szignifikánsan m_0-nál.

Egymintás u-próba

A próba a normális eloszlás várható értékére vonatkozik ismert szórás mellett.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m ismeretlen paraméter, $\sigma > 0$ ismert.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n}.$$

- **Egyoldali ellenhipotézis**: $H_0 : m \geq m_0$; $H_1 : m < m_0$.
- Ha $u < -u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A p-érték ilyenkor $\Phi(u)$.

$p < 0,05$: a várható érték szignifikánsan kisebb m_0-nál.

$p \geq 0,05$: a várható érték nem szignifikánsan kevesebb m_0-nál.

Egymintás u-próba

Feltételezés: a testmagasság normális eloszlású.

- Az európai férfiak átlagos testmagassága $177,6$ cm.
- Megmértük 10 magyar férfi testmagasságát, a magasságok átlaga 173,8 cm lett. A szórást 8 cm-nek feltételezve mondhatjuk-e, hogy a magyar emberek testmagassága szignifikánsan eltér az európai átlagtól?

Egymintás u-próba

Feltételezés: a testmagasság normális eloszlású.

- Az európai férfiak átlagos testmagassága $177,6$ cm.
- Megmértük 10 magyar férfi testmagasságát, a magasságok átlaga 173,8 cm lett. A szórást 8 cm-nek feltételezve mondhatjuk-e, hogy a magyar emberek testmagassága szignifikánsan eltér az európai átlagtól?
- $H_0 : m = 177,6$; $\quad H_1 : m \neq 177,6.$
- $$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n} = \frac{173,8 - 177,6}{8}\sqrt{10} = -1,502.$$
- $\alpha = 0,05$ terjedelem mellett $u_{1-\alpha/2} = 1,96$. $p = 0,133 > 0,05$.
- $|u| < u_{1-\alpha/2}$, elfogadjuk a nullhipotézist. A testmagasság nem tér el szignifikánsan az átlagos európai értéktől az adatok alapján.

Egymintás u-próba

Feltételezés: a testmagasság normális eloszlású.

- Az európai férfiak átlagos testmagassága $177,6$ cm.
- Megmértük 150 holland férfi testmagasságát, a magasságok átlaga 183,7 cm lett. A szórást 8 cm-nek feltételezve mondhatjuk-e, hogy a hollandok testmagassága szignifikánsan több az európai átlagnál?

Egymintás u-próba

Feltételezés: a testmagasság normális eloszlású.

- Az európai férfiak átlagos testmagassága $177,6$ cm.
- Megmértük 150 holland férfi testmagasságát, a magasságok átlaga 183,7 cm lett. A szórást 8 cm-nek feltételezve mondhatjuk-e, hogy a hollandok testmagassága szignifikánsan több az európai átlagnál?
- $H_0 : m \leq 177,6$; $\quad H_1 : m > 177,6$.
- $$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n} = \frac{183,7 - 177,6}{8}\sqrt{150} = 9,33.$$
- $\alpha = 0,05$ terjedelem mellett $u_{1-\alpha} = 1,645$. p-érték: $< 10^{-10} < 0,05$.
- $u > u_{1-\alpha}$, elutasítjuk a nullhipotézist. Az adatok statisztikailag bizonyítják, hogy a hollandok testmagasságának várható értéke szignifikánsan több 177,6 cm-nél.

Kétmintás u-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismertek.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

- Kétoldali ellenhipotézis: $H_0 : m_1 = m_2; \qquad H_1 : m_1 \neq m_2.$
- Ha $|u| > u_{1-\alpha/2}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p \geq 0,05$: a várható értékek nem térnek el szignifikánsan.

$p < 0,05$: a várható értékek szignifikánsan eltérnek.

Kétmintás u-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismertek.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

- Egyoldali ellenhipotézis: $H_0 : m_1 \leq m_2$; $\qquad H_1 : m_1 > m_2$.
- Ha $u > u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p < 0,05$: az első eloszlás várható értéke szignifikánsan nagyobb a második eloszlásénál.

Kétmintás u-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismertek.
- Próbastatisztika (eloszlása standard normális H_0 mellett):

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

- Egyoldali ellenhipotézis: $H_0 : m_1 \geq m_2$; $\qquad H_1 : m_1 < m_2$.
- Ha $u < -u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p < 0,05$: a második eloszlás várható értéke szignifikánsan nagyobb az első eloszlásénál.

Egymintás t-próba

A normális eloszlás várható értékére vonatkozik ismeretlen szórás esetén.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m, σ ismeretlen paraméterek.
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$u = \frac{\overline{X} - m_0}{s_n^*} \cdot \sqrt{n}.$$

- **Kétoldali ellenhipotézis**: $H_0 : m = m_0$; $H_1 : m \neq m_0$.
- Ha $|t| > t_{n-1}(1 - \alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A kritikus érték: $t_n(q)$ a q-kvantilise, vagyis az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q)\right),$$

ahol $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlásúak.

$p < 0,05$: a várható érték szignifikánsan eltér m_0-tól.

Egymintás t-próba

A normális eloszlás várható értékére vonatkozik ismeretlen szórás esetén.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m, σ ismeretlen paraméterek.
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$u = \frac{\overline{X} - m_0}{s_n^*} \cdot \sqrt{n}.$$

- **Egyoldali ellenhipotézis**: $H_0 : m \leq m_0$; $H_1 : m > m_0$.
- Ha $t > t_{n-1}(1-\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A kritikus érték: $t_n(q)$ a q-kvantilise, vagyis az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q)\right),$$

ahol $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlásúak.

$p < 0,05$: a várható érték szignifikánsan több m_0-nál.

Egymintás t-próba

A normális eloszlás várható értékére vonatkozik ismeretlen szórás esetén.

- $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$, ahol m, σ ismeretlen paraméterek.
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$u = \frac{\overline{X} - m_0}{s_n^*} \cdot \sqrt{n}.$$

- **Egyoldali ellenhipotézis**: $H_0 : m \geq m_0$; $H_1 : m < m_0$.
- Ha $t < t_{n-1}(\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A kritikus érték: $t_n(q)$ a q-kvantilise, vagyis az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q)\right),$$

ahol $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlásúak.

$p < 0,05$: a várható érték szignifikánsan kisebb m_0-nál.

Kétmintás t-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma^2)$, $Y_i \sim N(m_2, \sigma^2)$. Itt m_1, m_2, σ ismeretlen paraméterek (feltételezzük, hogy a két szórás megegyezik).
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

- Kétoldali ellenhipotézis: $H_0 : m_1 = m_2$; $\qquad H_1 : m_1 \neq m_2$.
- Ha $|t| > t_{n_1+n_2-2}(1 - \alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p \geq 0,05$: a várható értékek nem térnek el szignifikánsan.

$p < 0,05$: a várható értékek szignifikánsan eltérnek.

Kétmintás t-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma^2)$, $Y_i \sim N(m_2, \sigma^2)$. Itt m_1, m_2, σ ismeretlen paraméterek (feltételezzük, hogy a két szórás megegyezik).
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

- Egyoldali ellenhipotézis: $H_0 : m_1 \leq m_2$; $\qquad H_1 : m_1 > m_2$.
- Ha $t > t_{n_1+n_2-2}(1 - \alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p < 0,05$: az első várható érték szignifikánsan nagyobb a másodiknál.

Kétmintás t-próba

- $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma^2)$, $Y_i \sim N(m_2, \sigma^2)$. Itt m_1, m_2, σ ismeretlen paraméterek (feltételezzük, hogy a két szórás megegyezik).
- Próbastatisztika (eloszlása t-eloszlás H_0 mellett):

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

- Egyoldali ellenhipotézis: $H_0 : m_1 \geq m_2$; $\qquad H_1 : m_1 < m_2$.
- Ha $t < t_{n_1+n_2-2}(\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

$p < 0,05$: a második várható érték szignifikánsan nagyobb az elsőnél.

F-próba

Az F-próba független normális eloszlású minták szórását hasonlítja össze.

- Legyenek most $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt $m_1, m_2, \sigma_1, \sigma_2$ ismeretlen paraméterek.
- Próbastatisztika (eloszlása F-eloszlás H_0 mellett):

$$F = \frac{s_{n_1}^{*2}}{s_{n_2}^{*2}}.$$

- Kétoldali ellenhipotézis: $H_0 : \sigma_1 = \sigma_2$; $H_1 : \sigma_1 \neq \sigma_2$.
- Ha $F > F_{n_1-1,n_2-1}(1-\alpha/2)$ vagy $F < F_{n_1-1,n_2-1}(\alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.
- A kritikus érték: $q = \mathbb{P}(W \leq F_{d_1,d_2}(q))$, ahol $W = \frac{d_2(U_1^2+U_2^2+\ldots+U_{d_1}^2)}{d_1(V_1^2+V_2^2+\ldots+V_{d_2}^2)}$, és az U_i, V_i-k mind független standard normális eloszlású valószínűségi változók.

$p < 0,05$: a szórások szignifikánsan eltérnek.

Normális eloszlásra vonatkozó próbák

- Az összes eddigi próbánál (u, t, F-próbák) feltételeztük a **normális eloszlást**. Tehát a próba elvégzése előtt meg kell győződni arról, hogy a minta normális eloszlású (ha más módszer nem érhető el, legalábbis a hisztogram alapján), vagy a minta elemszáma elég nagy ahhoz, hogy a centrális határeloszlástétel alapján feltételezhessük, hogy az átlag eloszlása közel van a normális eloszláshoz.
- Minden esetben a minták függetlenek voltak, a kétmintás esetekben feltételeztük a **két minta egymástól való függetlenségét** is. Ha a két minta elemei valamilyen természetes szempontból kettesével összetartoznak (például egy ember vérnyomása reggeli és esti méréskor), akkor a kétmintás t-próba nem használható, más módszerek szükségesek.
- Ha az F-próba elvégzése azt mutatja, hogy a két minta **szórása szignifikánsan** eltér, akkor sem alkalmazható a kétmintás t-próba.

χ^2-próba

Legyen $A_1, A_2, \ldots, A_r$ teljes eseményrendszer, $p_1, p_2, \ldots, p_r$ pedig olyan nemnegatív számok, melyek összege 1.

$H_0 : \mathbb{P}(A_i) = p_i$ minden $i = 1, 2, \ldots, r$-re.

$H_1 : \mathbb{P}(A_i) \neq p_i$ valamelyik $i = 1, 2, \ldots, r$-re.

- n független megfigyelést végzünk.
- N_i: hányszor következett be A_i.
- Ha van i, hogy $N_i < 4$: néhány osztályt össze kell vonnunk, hogy a próbát alkalmazhassuk (vagyis A_i és A_j helyett $A_i \cup A_j$-t és $p_1 + p_2$-t tekintjük).
- Próbastatisztika:

$$T = \sum_{i=1}^{r} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}.$$

χ^2-próba

$H_0 : \mathbb{P}(A_i) = p_i$ minden $i = 1, 2, \ldots, r$-re.

Próbastatisztika:

$$T = \sum_{i=1}^{r} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}.$$

χ^2-próba: H_0-t elfogadjuk, ha $T < c$, ahol c az $f = r - 1$ szabadsági fokú, α terjedelmű χ^2-próba c kritikus értékénél. Pontosabban:

$$\mathbb{P}(Z_1^2 + Z_2^2 + \ldots + Z_f^2 < c) = 1 - \alpha),$$

ahol $Z_1, \ldots, Z_f$ független standard normális eloszlású valószínűségi változók.

$T > c$ vagy $p < \alpha$: elutasítjuk H_0-t, az eloszlás szignifikánsan eltér (p_k)-tól.

χ^2-próba: példa

Példa: $r = 6$, dobókockával dobunk, A_i: a dobás értéke i. $p_1 = p_2 = \ldots = p_6 = 1/6$ (szabályos a dobókocka).

A próba terjedelmének $\alpha = 0,05$-öt választjuk. $n = 100$ dobásból az alábbi értékek adódtak:

érték	1	2	3	4	5	6
gyakoriság	21	11	20	22	11	15

```
Chi-squared test for given probabilities

data:  kocka1

X-squared = 7.52, df = 5, p-value = 0.1847
```

Ekkor $T = 7,52 < c = 11,1$, tehát elfogadjuk azt a nullhipotézist, hogy a dobókocka szabályos. A p-érték $0,1847 > 0,05$, tehát nincs szignifikáns eltérés a szabályossághoz képest. (Minden szám legalább 4-szer előfordult, nem kellett a beosztáson módosítani.)

χ^2-próba: példa

Ha ezerszer dobunk, és az alábbi eredmények adódnak:

érték	1	2	3	4	5	6
gyakoriság	191	154	140	184	156	175

```
Chi-squared test for given probabilities

data:  kocka2

X-squared = 11.684, df = 5, p-value = 0.03938
```

Továbbra is $\alpha = 0,05$ terjedelem mellett számolva: $T = 11,684 > c = 11,1$, tehát elutasítjuk a nullhipotézist, statisztikai bizonyítékunk van arra, hogy a dobókocka nem szabályos. A p-érték $0,03938 < 0,05$, szignifikáns eltérés van a szabályossághoz képest.

Becsléses illeszkedésvizsgálat

$A_1, A_2, \ldots, A_r$ teljes eseményrendszer. N_i: hányszor következik be A_i egy n elemű független mintában. Adott $p_i(s)$ minden $s \in S$-re.

H_0: van olyan $s \in S$, melyre $\mathbb{P}(A_i) = p_i(s)$ minden $r = 1, 2, \ldots, r$-re.

H_1: nincs olyan $s \in S$, melyre $\mathbb{P}(A_i) = p_i(s)$ minden $r = 1, 2, \ldots, r$-re teljesülne.

Az s paramétervektor (d dimenziós) maximumlikelihood-becslése legyen $\hat{s}$, és legyen $\hat{p}_i = p_i(\hat{s})$. Számítsuk ki az alábbi mennyiséget:

$$T = \sum_{i=1}^{r} \frac{(N_i - n \cdot \hat{p}_i)^2}{n \cdot \hat{p}_i}.$$

Legyen $f = r - d - 1$. A H_0-t α terjedelem mellett elfogadjuk, ha $T < c$, ahol c az f szabadsági fokú kritikus értéke α terjedelem mellett. H_0-t elutasítjuk, ha $T > c$ (azaz $p < \alpha$), ilyenkor a minta szignifikánsan eltér az S által megadott eloszláscsaládtól.

Becsléses illeszkedésvizsgálat: példa

Példa. Az egy futballmérkőzésen lőtt gólok száma a világbajnokság $n = 95$ mérkőzésén:

gólok száma	0	1	2	3	4	5	6	7	8
mérkőzések száma	23	37	20	11	2	1	0	0	1

Poisson-esetben az s paraméter maximumlikelihood-becslése:

$$\hat{s} = \overline{X} = \frac{0 \cdot 23 + 1 \cdot 37 + 2 \cdot 20 + 3 \cdot 11 + 4 \cdot 2 + 5 \cdot 1 + 8 \cdot 1}{95} = 1,379.$$

Mivel vannak olyan osztályok, ahova 4-nél kevesebb megfigyelés esik, a beosztást módosítjuk:

gólok száma	0	1	2	3	≥ 4
mérkőzések száma	23	37	20	11	4

Becsléses illeszkedésvizsgálat: példa

H_0: az eloszlás Poisson-eloszlásból származik, valamely $s > 0$ paraméterrel (most $d = 1$).

H_1: az eloszlás nem Poisson-eloszlás.

$\hat{p} = 1,379$ a paraméter maximumlikelihood-becslése.

gólok száma	0	1	2	3	≥ 4
mérkőzések száma	23	37	20	11	4
Poisson($\hat{p}$)-eloszlás	23,92	32,99	22,75	10,46	4,88

Ebben az esetben $T = 1,04$, $f = 5 - 1 - 1 = 3$, a kritikus érték $7,81$.

$T < c$: elfogadjuk, hogy a minta Poisson-eloszlásból származik.

Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket.

Első szempont: $A_1, \ldots, A_r$. Második szempont: $B_1, \ldots, B_s$.

H_0: a két szempont független egymástól, azaz $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$ minden i, j-re.

H_1: a nullhipotézis nem igaz, a két szempont összefügg.

N_{ij}: hány olyan megfigyelés van, melyre A_i és B_j teljesül.

$N_{i\cdot} = \sum_{j=1}^{s} N_{ij}$ (azaz az A_i gyakorisága); $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$ (azaz B_j gyakorisága); n pedig az összes megfigyelés száma. Ekkor a próbastatisztika:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i\cdot} N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot} N_{\cdot j}}{n}}.$$

Függetlenségvizsgálat

H_0: a két szempont független egymástól. Próbastatisztika:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}}.$$

A szabadsági fok $f = (r-1)(s-1)$.

c: az f szabadsági fokú χ^2-próba kritikus értéke α terjedelem mellett.

- $T < c$ (azaz a $p \geq \alpha$): elfogadjuk H_0-t, nem találtunk szignifikáns összefüggést a szempontok között.
- $T > c$ (azaz a $p < \alpha$): elutasítjuk H_0-t, az adatok szignifikáns összefüggést mutatnak.

Ha $r = s = 2$, a próbastatisztika az alábbi egyszerűbb alakra hozható:

$$T = \frac{n\left(N_{11}N_{22} - N_{12}N_{21}\right)^2}{N_{1.}N_{2.}N_{.1}N_{.2}}.$$

Függetlenségvizsgálat: példa

H_0: a hőmérséklet és a csapadékmennyiség független;

H_1: a hőmérséklet és a csapadékmennyiség nem független.

$n = 100$, $f = 2 \cdot 2 = 4$, $\alpha = 0,05$:

	meleg	átlagos	hideg
esős	15	10	5
átlagos	10	10	20
száraz	5	20	5

```
data:  ido

X-squared = 22.917, df = 4, p-value = 0.0001316
```

$22,917 > c_{\text{krit}} = 9,49$, illetve $p < \alpha = 0,05 \Rightarrow$ elutasítjuk a nullhipotézist, szignifikáns összefüggés van a két szempont között.

Homogenitásvizsgálat

Legyenek X, Y valószínűségi változók. $\mathbb{R}$-t bontsuk fel diszjunkt halmazok uniójára: $A_1, \ldots, A_r$.

H_0: az X és Y valószínűségi változók eloszlása megegyezik, azaz $\mathbb{P}(X \in A_i) = \mathbb{P}(Y \in A_i)$ minden $i = 1, 2, \ldots, r$-re.

H_1: az X és Y valószínűségi változók eloszlás eltérő, azaz van legalább egy i, melyre $\mathbb{P}(X \in A_i) \neq \mathbb{P}(Y \in A_i)$. $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ független minta úgy, hogy $X_1, \ldots, X_n \sim X$, $Y_1, \ldots, Y_n \sim Y$.

N_i az A_i gyakorisága az $\underline{X}$ mintában;

M_i az A_i gyakorisága az $\underline{Y}$ mintában. A próbastatisztika:

$$T = \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \cdot n \cdot m.$$

Homogenitásvizsgálat

A próbastatisztika:

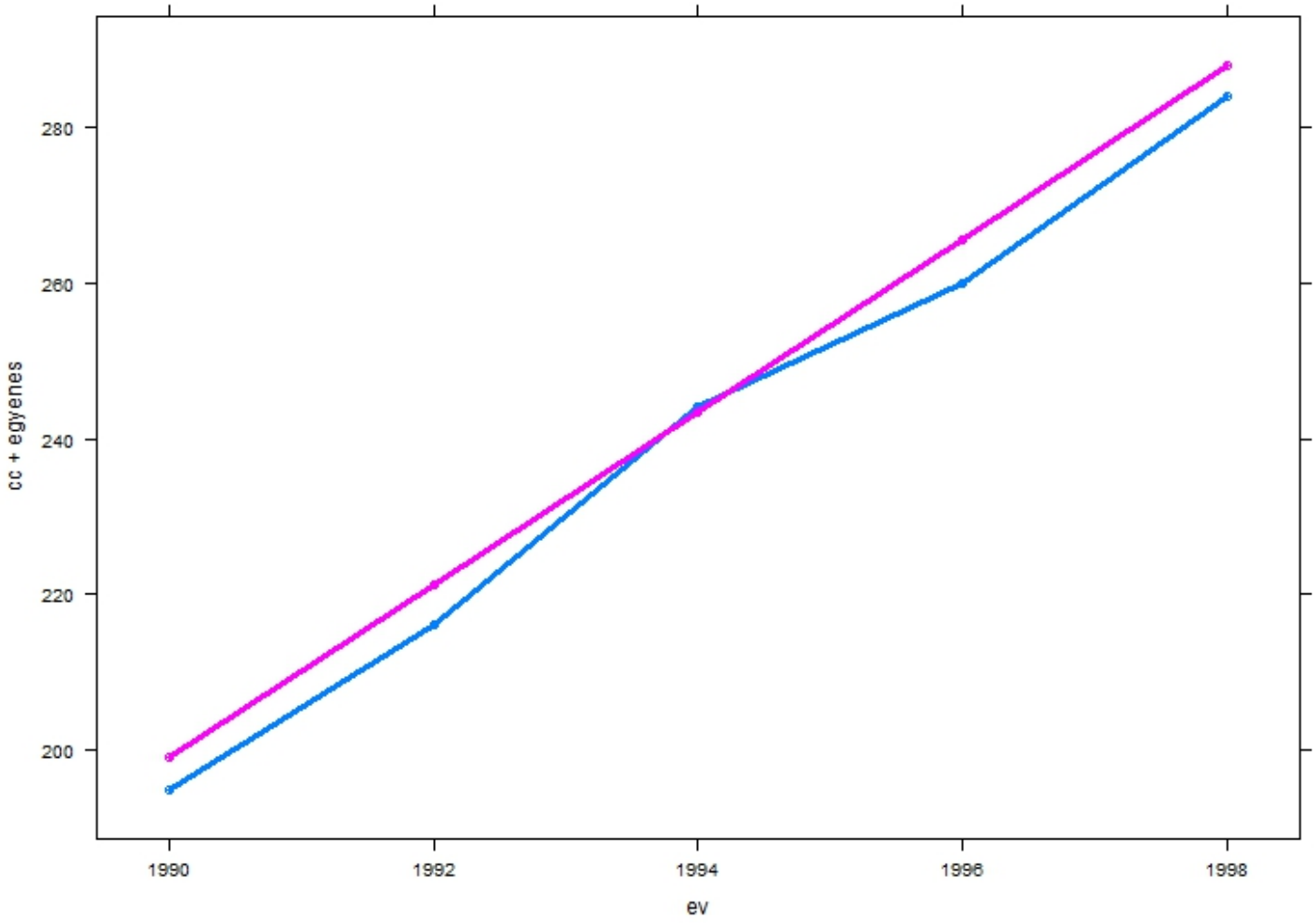
$$T = \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \cdot n \cdot m.$$

A szabadsági fok: $f = r - 1$.

c: az f szabadsági fokú χ^2-próba kritikus értéke α terjedelem mellett.

- $T < c$ (azaz a $p \geq \alpha$): elfogadjuk H_0-t, nem találtunk szignifikáns eltérést az eloszlások között.
- $T > c$ (azaz a $p < \alpha$): elutasítjuk H_0-t, az eloszlások szignifikáns eltérést mutatnak.

Lineáris regresszió



7. ábra.
A CFC-12 (freon) gáz koncentrációja az Antarktiszon és az adatokra illesztett egyenes

Lineáris regresszió

Egyenes illesztése a **legkisebb négyzetek módszerével**:

Állítás (Lineáris regresszió)

Legyenek $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *adott számpárok. Azokat az* a *és* b *együtthatókat keressük, melyre a*

$$h^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

mennyiség minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{k=1}^{n} (x_k - \overline{x})^2}; \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Lineáris modell

Definíció (Lineáris modell)

Legyenek $X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_n$ *valószínűségi változók, és tegyük fel, hogy valamely* a, b *valós számokra*

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol $\varepsilon_1, \ldots, \varepsilon_n$ *független* $N(0, \sigma^2)$ *eloszlású valószínűségi változók. Az így kapott* (X_i, Y_i) *párok együttes eloszlását lineáris modellnek nevezzük.*

Az X_i valószínűségi változókat magyarázó változóknak, az ε_i valószínűségi változókat hibának szokták nevezni.

Becslések a lineáris modellben

Állítás

A lineáris modellben az a, b együtthatók ML-likelihood becslése a következőképpen írható:

$$\hat{a} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{k=1}^{n}(X_k - \overline{X})^2}; \qquad \hat{b} = \overline{Y} - \hat{a}\overline{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az a és b paramétereknek. A hiba szórásának becslése:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{j=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

$$D(\hat{a}) = \frac{\sigma}{\sum_{j=1}^{n}(X_j - \overline{X})^2}; \qquad D(\hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}}.$$

Előrejelzés a lineáris modellben

Állítás

Legyen x^* adott szám. A lineáris modellből kapott előrejelzés az Y véletlen folyamat x^* pontban felvett értékére:

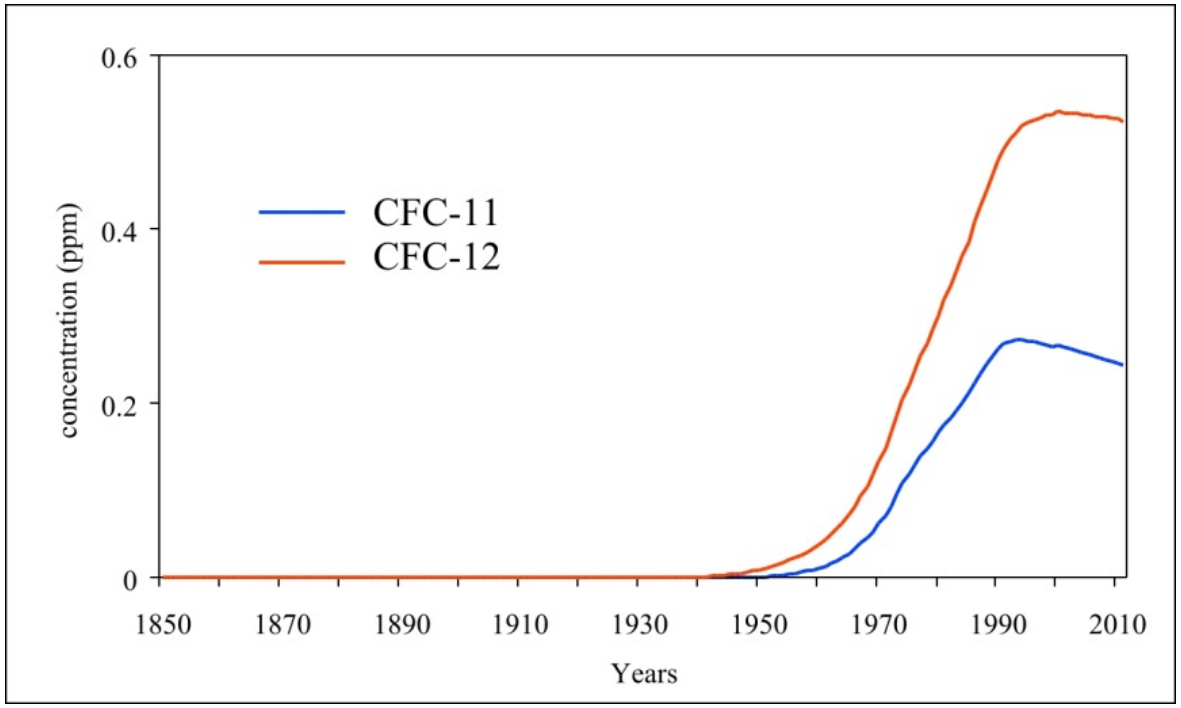
$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

$$D(\hat{a}x^* + \hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}}.$$

Az előrejelzés szórásának becslésekor a σ értéket gyakran $\hat{\sigma}$-val helyettesítik.

Előrejelzés a lineáris modellben



8. ábra.
A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

Reziduálisok

A teljes ingadozás (total sum of squares): $\sum_{j=1}^{n}(Y_j - \overline{Y})^2$.

Reziduális négyzetösszeg (residual sum of squares):

$$\sum_{j=1}^{n}(Y_j - \hat{a}X_j - \hat{b})^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\sum_{k=1}^{n}(X_k - \overline{X})^2}.$$

Definíció

A megmagyarázott ingadozás részaránya (coefficient of determination):

$$R^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right]\left[\sum_{k=1}^{n}(Y_k - \overline{Y})^2\right]}.$$

Az R^2 értéke 0 és 1 közé esik. Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. Ugyanakkor R érzékeny a kiugró értékekre.

Hipotézisvizsgálat

A lineáris tag együtthatójára vonatkozó hipotézisvizsgálati feladat a következő (a terjedelem α):

$H_0 : a = 0 \qquad H_1 : a \neq 0$ vagy $H_1 : a > 0$ vagy $H_1 : a < 0$.

A nullhipotézis mellett az alábbi mennyiség $n-2$ szabadsági fokú t-eloszlású:

$$t = \hat{a}\frac{\sqrt{(n-2)\sum_{i=1}^{n}(X_i - \overline{X})^2}}{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2}}.$$

- Kétoldali ellenhipotézis, $H_1 : a \neq 0$. Ha $|t| > t_{n-2}(1-\alpha/2)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan eltér 0-tól), különben elfogadjuk.
- Egyoldali ellenhipotézis, $H_1 : a > 0$. Ha $t > t_{n-2}(1-\alpha)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan nagyobb 0-nál), különben elfogadjuk.
- Kétoldali ellenhipotézis, $H_1 : a < 0$. Ha $t < t_{n-2}(\alpha)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan kisebb 0-nál), különben elfogadjuk.