A statisztika alapjai, az eloszlásfüggvény és a sűrűségfüggvény becslése

Legyen

- Ω tetszőleges halmaz - a vizsgált objektumok halmaza;
- $\xi : \Omega \to \mathbb{R}$ valószínűségi változó - a statisztikai jellemző; az ismeretlen eloszlású valószínűségi változó, célunk ennek az eloszlásáról információt szerezni;
- $\xi_1, \ldots, \xi_n$ - teljesen független; ξ-vel azonos eloszlású valószínűségi változók (a *mintaelemek*);
- $f(x_1, \ldots, x_n)$ n-változós függvény, $S_n := f(\xi_1, \ldots, \xi_n)$ - a *statisztika.*

Becslések

1. Az eloszlásfüggvény becslése

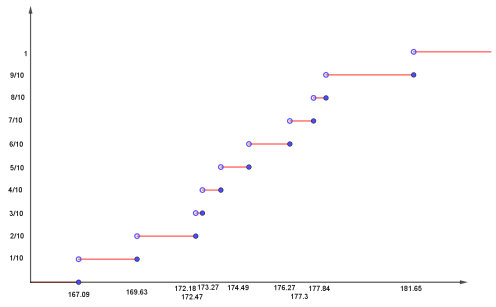
Legyen $k_{x,n}$ egy n elemű minta x-nél kisebb elemeinek száma. Ekkor

$$F_n(x) = \frac{k_{x,n}}{n}$$

az *empirikus eloszlásfüggvény.*

Tekintsük például az alábbi 10 elemű mintát emberek magasságára: $167,09$; $181,65$; $176,27$; $173,27$; $172,18$; $174,49$; $177,30$; $177,84$; $172,47$; $169,63$.

Ez alapján az empirikus eloszlásfüggvény az alábbi:



Tehát az empirikus eloszlásfüggvény egy diszkrét valószínűségi változó eloszlásfüggvénye, így teljesülnek rá az eloszlásfüggvény tulajdonságai:

- monoton növekvő;
- balról folytonos;
- $\lim_{x\to-\infty} F_n(x) = 0,\ \lim_{x\to\infty} F_n(x) = 1.$

$k_{x,n}$ binomiális eloszlású valószínűségi változó, amelynek paraméterei:

- n
- $p_x = P(\xi < x) = F(x)$.

Így $F_n(x)$ várható értéke:

$$E(F_n(x)) = E\frac{k_{x,n}}{n} = \frac{1}{n}\cdot E(k_{x,n}) = \frac{1}{n}\cdot n\cdot F(x) = F(x).$$

Ezt a tényt úgy fogalmazzuk meg, hogy az empirikus eloszlásfüggvény az eloszlásfüggvény *torzítatlan* becslése.

Általában, az S_n statisztika a ξ változó egy ismeretlen θ paraméterének *torzítatlan* becslése, ha bármely $n \geq 1$ esetén $E(S_n) = \theta$.

A nagy számok erős törvénye alapján $F_n(x) = \frac{k_{x,n}}{n}$ majdnem biztosan konvergál $F(x)$-hez. (Ezt a nagy számok Borel-féle erős törvényét a $\{\xi < n\}$ eseményre alkalmazva kapjuk, ugyanis ennek az eseménynek a relatív gyakorisága $F_n(x)$, valószínűsége pedig $F(x)$.) Ezt a tényt úgy fogalmazzuk meg, hogy az empirikus eloszlásfüggvény az eloszlásfüggvény *erősen konzisztens* becslése.

Általában, az S_n statisztika a ξ változó egy ismeretlen θ paraméterének

- *gyengén konzisztens* becslése, ha S_n sztochasztikusan konvergál θ-hoz;
- *erősen konzisztens* becslése, ha S_n majdnem biztosan konvergál θ-hoz.

Az empirikus eloszlásfüggvény mint becslés erősségét továbbá a *matematikai statisztika alaptétele* támasztja alá:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{mb}} 0,$$

ha $n \to \infty$.

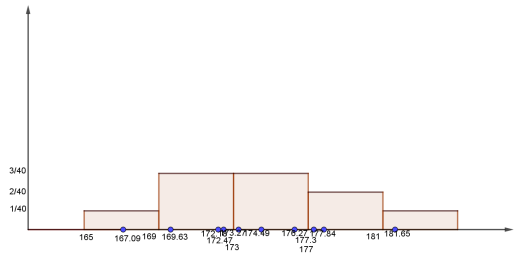
2. A sűrűségfüggvény becslése

Fedjük le a számegyenest a páronként diszjunkt $I_1, I_2, \ldots$ intervallumokkal. Jelölje ν_k az I_k intervallumba eső mintaelemek számát. Legyen

$$f_n(x) := \frac{\nu_k}{n\,|I_k|} \text{ , ha } x \in I_k.$$

Az így kapott függvényt *sűrűséghisztogram*nak nevezzük.

Példaként elkészítünk egy, az előző példában szereplő mintához tartozó sűrűséghisztogramot, ahol a $[165, 185]$ intervallumot 4 egység hosszú részintervallumokra osztottuk.



Látható, hogy a sűrűségfüggvényt egy egyenes szakaszokból felépített sűrűségfüggvénnyel közelítettük. Ez valóban sűrűségfüggvény:

- $f_n(x) > 0$;
- integrálható;
- $\int_{-\infty}^{\infty} f_n(x)dx = 1$, ugyanis $\sum_k |I_k| \cdot \frac{\nu_k}{n|I_k|} = \frac{1}{n}\sum_k \nu_k = \frac{1}{n} \cdot n = 1$.