Matematikai statisztika előadás, 2. hét, február 17.

Leíró statisztika; a sűrűségfüggvény becslése

1. Kvantilisek, boxplot

Ahogy láttuk, a tapasztalati eloszlásfüggvény a mintaelemszám növekedésével a valódi eloszlásfüggvényhez tart (Glivenko–Cantelli-tétel). Gyakran azonban nem közvetlenül az eloszlásfüggvényre, vagyis a $\mathbb{P}(X \leq t)$ valószínűségre vagyunk kíváncsiak (ahol most az X megfigyelt mennyiség eloszlása ismeretlen), hanem a kvantilisekre vagyunk kíváncsiak, azaz arra, hogy adott z esetén mi lehet az a q, amire $\mathbb{P}(X \leq q) = z$ teljesül. Például ha X egy folyó legnagyobb vízállása egy évben, és $z = 0,95\%$, akkor q mondja meg, hogy mi az a legnagyobb vízállás, aminél a folyó csak 5%-kal megy magasabbra – ha ilyen magas gátat építünk, az 95% valószínűséggel megfelelő lesz. Vagy, ha X a szükséges kórházi ágyak számának maximuma egy városban egy év alatt, és q az eloszlás z-kvantilise $z = 95\%$-kal, akkor q kórházi ágy 95% valószínűséggel lesz elég (a példában nem számolva azzal, hogy nem minden beteget tudnak bármelyik egységben megfelelően ellátni).

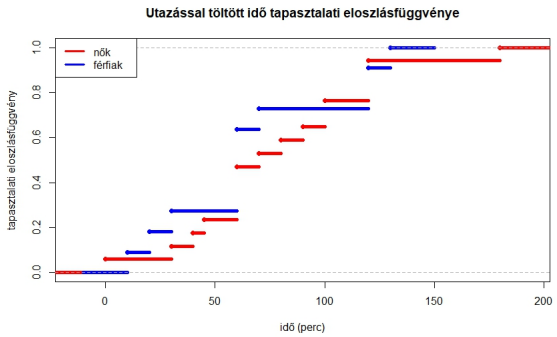
Emlékeztetőül: az X valószínűségi változó z-kvantilise a legkisebb olyan q szám, melyre teljesül, hogy $\mathbb{P}(X \leq q) = F(q) \geq z$.

Kérdés, hogy a kvantiliseket hogyan tudjuk a megfigyelt $X_1, X_2, \ldots, X_n$ mintából becsülni. Itt tehát az a feltételezésünk, hogy az $X_1, X_2, \ldots, X_n$ valószínűségi változók függetlenek, azonos eloszlásúak, azonban ezt az eloszlást nem ismerjük.

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

1.1. Definíció (Tapasztalati kvantilis). *Legyen $X_1^* \leq X_2^* \leq \ldots \leq X_n^*$ rendezett minta, és $z \in [0,1]$ adott szám. Ekkor a minta tapasztalati z-kvantilise:*

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



1. ábra. Hétköznap utazással töltött idő hisztogramja a teljes mintára ($n = 28$, illetve tapasztalati eloszlásfüggvény külön ($n_1 = 17$ nő, $n_2 = 11$ férfi), 2020. februári adatok

Ehhez nagyjából azt kell megnézni, hogy a tapasztalati eloszlásfüggvény hol éri el z-t, mivel pedig tipikusan z két felvett érték közé esik, azoknak az értékeknek a lineáris kombinációját vesszük, amiknél még éppen kisebb, illetve már nagyobb z-nél a tapasztalati eloszlásfüggvény. Például az 1. ábra alapján keressük meg a férfiak utazási idejének 40%-os kvantilisét, vagyis azt az értéket, ami azt a q időt becsüli, aminél a férfiak 40%-a tölt kevesebbet utazással. Látjuk, hogy a 60 gyakran szerepel, és 59 percnél kevesebb időt a mintában a férfiak kevesebb, mint 30%-a szán utazásra, 61 percnél kevesebb időt pedig több, mint a 60 százalékuk. Ez alapján

mondhatnánk a 60-at is becslésnek, hiszen a tapasztalati eloszlás függvény itt lépi át a 40%-ot, és az R ezt is adja vissza:

```
> ferfi<-c(60, 30, 70, 60, 20, 60, 10, 120, 60, 120, 130)
> quantile(ferfi, probs=c(0.4, 0.8))
40% 80%
60 120
```

Ha pedig a fenti képletet alkalmazzuk (bár az R nem pontosan ezt használja): $z = 0,4$ és $n = 11$, így $z(n+1) = 4,4$ alsó egészrésze: $\lfloor z(n+1) \rfloor = 4$. Ez alapján

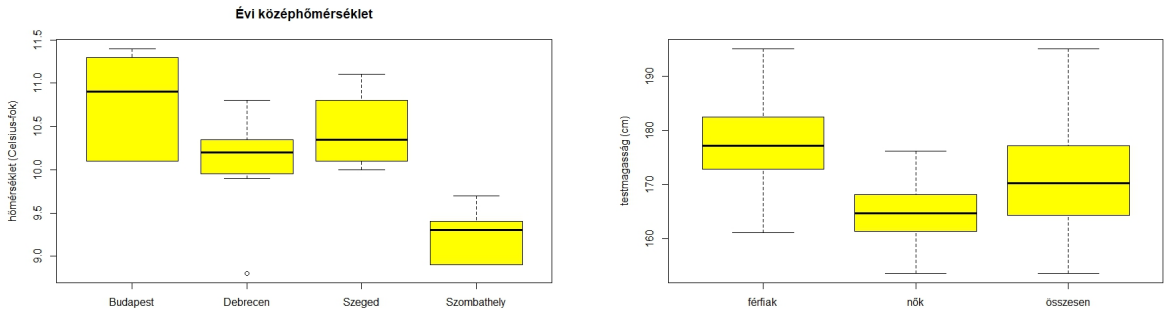
$$\hat{q}_z = X_4^* + 0,4 \cdot \left(X_5^* - X_4^*\right) = 60 + 0,4 \cdot (60 - 60) = 60,$$

hiszen a rendezett minta 4. és 5. eleme is 60-nal egyenlő.

A 80%-os kvantilist is hasonlóképpen számolhatnánk ki.

A kvantilisek közül az alábbiak gyakran előfordulnak, többek között a boxplot ábrán :

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



2. ábra. Boxplot ábra négy város éves középhőmérséklet adataiból (forrás: Országos Meteorológiai Szolgálat), illetve testmagasság adatok boxplotja $n = 96$ elemű mintából külön a férfiak és nők esetében és összesítve

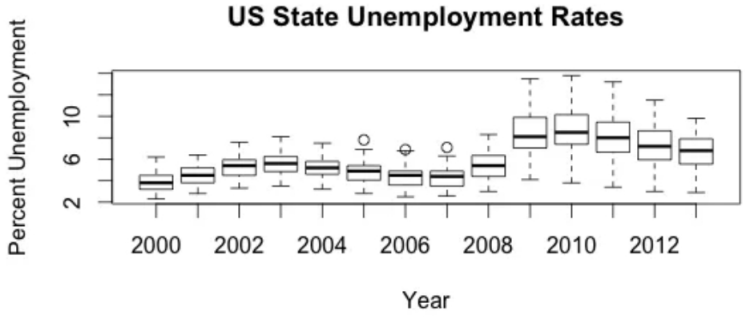
Ahogy a 2. ábrán láthatjuk, a boxplot ábra segítségével több adatsort hasonlíthatunk össze.

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).

Az egyes dobozok határait az első és a harmadik kvartilis adja meg, a középső vonal a medián, a vonalak pedig a legkisebb, illetve legnagyobb mintaelemig tartanak.

Megállapíthatjuk például, hogy Szombathelyen még a maximum is kisebb volt, mint a másik három városban bármelyik megfigelés, vagy hogy Szegeden a megfigyelések negyede kb. $10,1$ és $10,3$ fok közé esett (ez az első kvartilis és a medián közötti tartomány). A jobb oldali ábráról pedig azt vehetjük például észre, hogy a mintában szereplő legalacsonyabb férfinál a nők nagyjából negyedrésze alacsonyabb.



3. ábra. Az USA államaiban mért munkanélküliségi ráta boxplotja (forrás: `https://www.r-bloggers.com/2015/11/free-webinar-learn-to-map-unemployment-data-in-r/`)

2. Középértékek

A mintát, különösen, ha más adatsorokkal akarjuk összehasonlítani vagy az időbeli változást figyeljük, gyakran csak egy, rá jellemző számmal, középértékkel jelenítjük meg. Erre is több lehetőség van, a következőkben azt nézzük meg, hogy a két leggyakoribb középérték egymáshoz viszonyítva hogyan viselkedik.

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

2.1. 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.

2.1. Az átlag és a medián összehasonlítása

Normális eloszlás

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

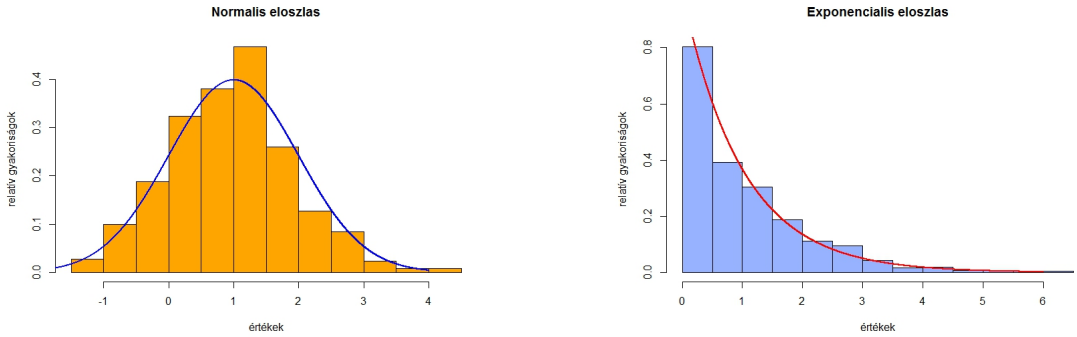
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9840	0.2847	**0.9842**	**0.9863**	1.6930	3.6110

Exponenciális eloszlás

Vegyünk egy másik, 500 elemű független mintát is: $Y_1, Y_2, \ldots, Y_{500}$ 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 két mintát összehasonlítva azt vehetjük észre, hogy a normális eloszlású, szimmetrikusabb esetben az átlag és a medián értéke majdnem megegyezik, lényegében mindegy, hogy melyiket



4. ábra. A normális, illetve exponenciális eloszlású minták hisztogramja

használjuk. Az exponenciális eloszlás sűrűségfüggvénye viszont nem szimmetrikus, ilyenkor az átlag és a medián eltér, ilyenkor érdemes lehet mindkettőt feltüntetni, ha pedig az aszimmetriát többek között kiugró, részben hibásnak vélt mérések okozzák, akkor az átlag helyett a mediánt használni.

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).

2.2. Középértékek közelítése osztályközös gyakoriságokkal

Tegyük fel, hogy az adatokat nem ismerjük pontosan, csak a hisztogramot, vagyis hogy az egyes osztályokba, intervallumokba hány megfigyelés esik. Legyen x_j a j. osztályközép (az alsó és felső határ átlaga), és f_j a j. osztályba eső megfigyelések száma, továbbá $n = f_1 + f_2 + \ldots + f_k$ az összes megfigyelés száma. Ekkor

- az átlag közelítése:
$$\frac{f_1 x_1 + f_2 x_2 + \ldots + f_k x_k}{n};$$
- a medián közelítése:
$$t_{\text{me}} + \frac{n/2 - F_{\text{me}-1}}{f_{\text{me}}} \cdot h_{\text{me}},$$
ahol t_{me} a mediánt tartalmazó osztály alsó határa, $F_{\text{me}-1}$ a mediánt tartalmazó osztályt megelőző osztályok gyakoriságainak összege, f_{me} a mediánt tartalmazó osztály gyakorisága, h_{me} a mediánt tartalmazó osztály szélessége.

3. Indexek számítása

Indexeket különböző mennyiségek összehasonlítására, hatások összehasonlítására, vagy közvetlenül nem összemérhető mennyiségek változásának leírására szoktak használni (például: a GDP közvetlenül nem összehasonlítható mennyiségek keveréke).

A leíró statisztikának ebből a témaköréből egy példát nézünk meg alaposabban.

Tegyük fel, hogy egy időben változó mennyiség egy időszakban (tárgyidőszakban) mért értékeit szeretnénk egy korábbi, hasonló időszakban (bázisidőszakban) mért értékekkel összehasonlítani, hogy az átlagos változást leírhassuk. Például tekinthetjük a fogyasztói árindexet (például: `https://www.ksh.hu/docs/hun/xstadat/xstadat_evkozi/e_qsf001.html` vagy `http://www.ksh.hu/interaktiv/fogyar_radar/index.html`), ami az infláció mérőszáma, a lakosság által vásárolt termékek és szolgáltatások árainak átlagos változását fejezi ki. A bázisidőszak lehet az „előző" év, a tárgyidőszak a vizsgált év.

Tegyük fel, hogy az árindexbe az $1, 2, \ldots, n$ termékek forgalmát építik be. Legyen

- $q_{0,j}$ a j. termékből eladott mennyiség a bázisidőszakban;
- $q_{1,j}$ a j. termékből eladott mennyiség a tárgyidőszakban;
- $p_{0,j}$ a j. termék egységára a bázisidőszakban;
- $p_{1,j}$ a j. termék egységára a tárgyidőszakban.

Ekkor

- bázisidőszaki súlyozású vagy Laspeyres-féle árindex (annak hányadosa, hogy az új árakkal, de a bázisidőszak fogyasztásával mennyivel nőtt az összes kiadás a régebbi időszakhoz képest):
$$\frac{\sum_{j=1}^{n} q_{0,j} p_{1,j}}{\sum_{j=1}^{n} q_{0,j} p_{0,j}}$$
- tárgyidőszaki súlyozású vagy Paasche-féle árindex (annak hányadosa, hogy az új árakkal és az új fogyasztással mennyivel nőtt az összes kiadás):
$$\frac{\sum_{j=1}^{n} q_{1,j} p_{1,j}}{\sum_{j=1}^{n} q_{1,j} p_{0,j}}$$
- bázisidőszaki súlyozású vagy Laspeyres-féle volumenindex (a régi árakkal számolva hányszorosára nőtt az összes kiadás, vagyis a régi árakkal számolva mennyivel nőtt a fogyasztás):
$$\frac{\sum_{j=1}^{n} q_{1,j} p_{0,j}}{\sum_{j=1}^{n} q_{0,j} p_{0,j}}$$
- tárgyidőszaki súlyozású vagy Paasche-féle volumenindex (az új árakkal számolva hányszorosára nőtt az összes kiadás):
$$\frac{\sum_{j=1}^{n} q_{1,j} p_{1,j}}{\sum_{j=1}^{n} q_{0,j} p_{1,j}}$$

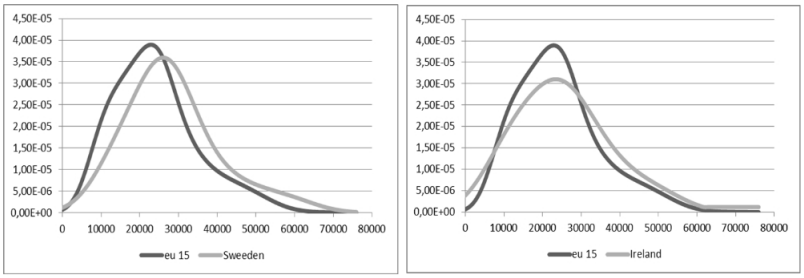


Figure 10: Density Function Estimations of EU Countries and Sweden (left) and Ireland (right) in 2001.

5. ábra. A svédországi és írországi jövedelmek sűrűségfüggvényének becslése egy összetettebb módszerrel, a megfelelő ponton Gauss-magfüggvénnyel (forrás: [1])

4. Sűrűségfüggvény becslése

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

Mintaelemszám: n

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

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

teljesül tetszőleges $t_1, t_2, \ldots, t_n$ valós számok esetén.

Az $(X_1, X_2, \ldots, X_n)$ valószínűségi változók **eloszlása nem ismert**: nem tudjuk, hogy mennyi $\mathbb{P}(X_1 \leq t)$, vagyis nem ismerjük az eloszlásfüggvényt, vagy mennyi X_1 várható értéke, szórása. A cél a valószínűségi változók eloszlásának a becslése, rá vonatkozó hipotézisek eldöntése a megfigyelések, vagyis az adatok alapján.

Abban az esetben, amikor az ismeretlen eloszlásról feltételezhetjük, hogy abszolút folytonos eloszlású, vagy ilyen módon modellezzük (például nincsenek olyan kitüntetett értékek, amik a valószínűségi változó pozitív valószínűséggel venne fel, amire onnan következtethetünk, hogy a megfigyelések mind vagy majdnem mind különbözőek), az eloszlás sűrűségfüggvényét sem ismerjük, viszont az adatok alapján megpróbálhatjuk megbecsülni.

Ahogyan több példán láttuk (akár a 4. ábrán), elég sok megfigyelés esetén a hisztogram és a sűrűségfüggvény közel esik egymáshoz. Ezt használhatjuk ki a sűrűségfüggvény pontosabb becslésekor, ugyanakkor itt is kérdés, hogy milyen intervallumhosszat használjunk.

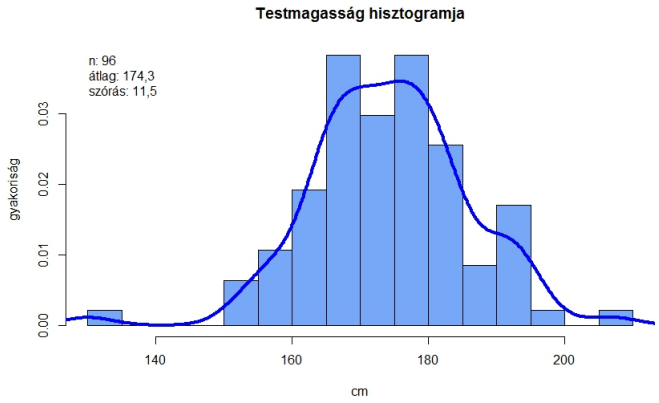
Az alábbiakban a sűrűségfüggvény becslésének Parzen–Rosenblatt-féle módszerét ismertetjük.

$X_1, X_2, \ldots, X_n$ független azonos eloszlású abszolút folytonos minta. A sűrűségfüggvény f, azaz

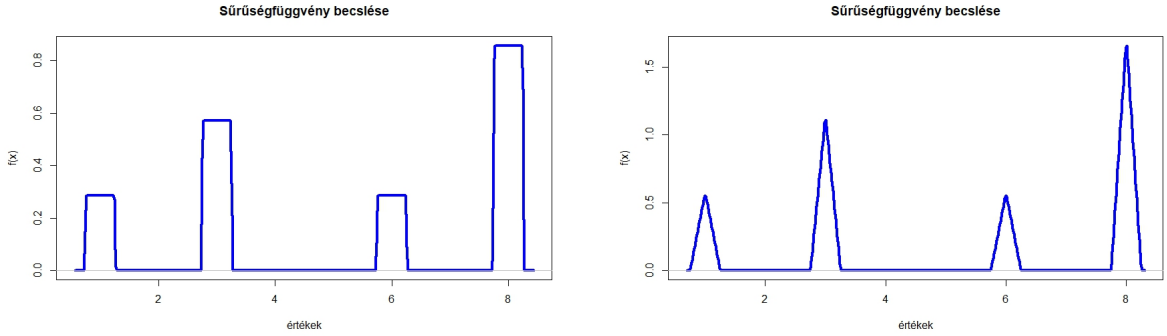
$$\mathbb{P}(a \leq X_1 \leq b) = \int_a^b f(t)dt \qquad \text{minden } a < b\text{-re.}$$

Az f függvény ismeretlen. Hogyan tudjuk $f(t)$ értékét becsülni az $X_1, \ldots, X_n$ megfigyelések segítségével?

Ehhez az alábbiból indulhatunk ki:



6. ábra. A testmagasság hisztogramja $n = 96$ elemű mintából (valós adatokból), a sűrűségfüggvény becslése Gauss-magfüggvénnyel.



7. ábra. A sűrűségfüggvény becslése téglalapos és háromszöges magfüggvénnyel az $1, 3, 3, 6, 8, 8, 8$ mintából

$$\mathbb{P}(a \le X_1 \le b) = \int_a^b f(t)dt \approx \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(a < X_j \le b),$$

azaz a becslés a és b közé eső mintaelemek aránya. A jobb oldalon éppen az $(a, b]$ intervallumba eső mintaelemek relatív gyakorisága szerepel, ez hasonló, mint ami a hisztogramban is szerepel. Ez az alábbi definícióhoz vezet.

4.1. A sűrűségfüggvény becslése különböző magfüggvényekkel

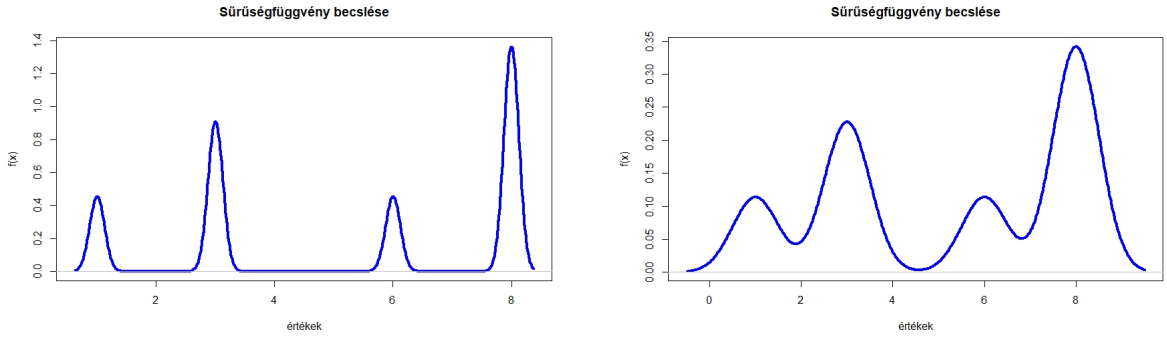
A 7. ábra bal oldalán az alábbi mintából készített becslést láthatjuk $(X_1, \ldots, X_7)$: 1, 3, 3, 6, 8, 8, 8. Minden oszlop közepe egy megfigyelés, a magasság a gyakoriságtól függ, az oszlop szélessége pedig $2h$, ahol h az ablakszélesség, ez választható a becslés során.

Ezt általánosabban az alábbi módon írhatjuk fel.

Téglalap magfüggvény: $k(y) = 1/2$, ha $-1 \le y \le 1$, nulla különben, azaz $k(y) = \frac{1}{2}\mathbb{I}(|y| \le 1)$ és h az **ablakszélesség**.

$$\hat{f}_n(t) = \frac{1}{nh}\sum_{j=1}^{n} \frac{1}{2}\mathbb{I}(|t - X_j| < h) = \frac{1}{n \cdot h}\sum_{j=1}^{n} k\left(\frac{t - X_j}{h}\right).$$

A téglalapon kívül más alakú függvényeket is szoktak használni a becsléshez, ezt például a 7.



8. ábra. A sűrűségfüggvény becslése Gauss-féle magfüggvénnyel az $1, 3, 3, 6, 8, 8, 8$ mintából $h = 0,5$ és $h = 2$-es ablakszélességgel

ábra jobb oldalán, illetve a 8. ábrákon láthatjuk. Ilyenkor a fenti leírásban k szerepe ugyanaz, csak k-t választjuk másképpen.

Háromszöges magfüggvény: $k(y) = \max(1 - |y|, 0)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^{n} k\left(\frac{t - X_j}{h}\right).$$

Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^{n} k\left(\frac{t - X_j}{h}\right) = \frac{1}{n \cdot h \cdot \sqrt{2\pi}} \sum_{j=1}^{n} \exp\left(-\frac{(y - X_j)^2}{2h^2}\right).$$

Minta (X): 1, 3, 3, 6, 8, 8, 8. Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 2$ **az ablakszélesség** (sávszélesség, bandwidth).

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^{n} k\left(\frac{t - X_j}{h}\right) = . \frac{1}{n \cdot h \cdot \sqrt{2\pi}} \sum_{j=1}^{n} \exp\left(-\frac{(y - X_j)^2}{2 \cdot 2^2}\right).$$

4.2. A sűrűségfüggvény Parzen–Rosenblatt-féle becslése

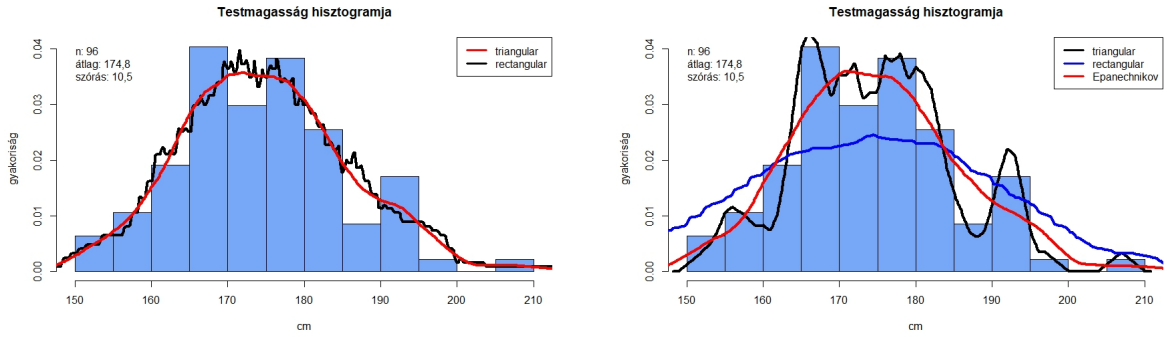
A fenti módszer általánosított változata a Parzen–Rosenblatt-féle becslés. Ezt a következőképpen definiálhatjuk:

Legyen $k : \mathbb{R} \to \mathbb{R}_+$ olyan függvény, mely korlátos, $\lim_{y\to\infty} yk(y) = 0$, továbbá h_n olyan számsorozat, melyre $\lim_{n\to\infty} h_n = 0$ és $\lim_{n\to\infty} nh_n = \infty$. A sűrűségfüggvény becslése a t pontban a Parzen–Rosenblatt-módszerrel a k magfüggvénnyel és h_n sávszélességgel az $X_1, \ldots, X_n$ független minta alapján:

$$\hat{f}_n(t) = \frac{1}{n \cdot h_n} \sum_{j=1}^{n} k\left(\frac{t - X_j}{h_n}\right).$$

A mintaelemszám növelésével a fenti módszer határértékben pontos eredményt ad, mindegyik fent bemutatott magfüggvényre. Megfelelő feltételek mellett $\hat{f}_n(t) \to f(t)$ minden t-re, ha $n \to \infty$ (szükséges például, hogy f folytonos legyen). Szokásos magfüggvények például (ebből hármat már láttunk):

- Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$.



9. ábra. A testmagasság sűrűségfüggvényének becslése $n = 96$ megfigyelésből; a bal oldalon: háromszöges (piros) és téglalapos (fekete) magfüggvénnyel (ez utóbbihoz túl kicsi a sávszélesség); a jobb oldalon: háromszöges magfüggvény 1/3-szoros sávszélességgel (fekete), téglalapos magfüggvény 3-szoros sávszélességgel (kék), Epanechnikov-magfüggvény alapértelmezett sávszélességgel (piros)

- Háromszög magfüggvény: $k(y) = (1 - |y|)$, ha ez nemnegatív, nulla különben.
- Epanechnikov-magfüggvény: $k(y) = \frac{3}{4}(1 - y^2)$, ha ez nemnegatív, nulla különben.
- Téglalap magfüggvény: $k(y) = 1/2$, ha $-1 \leq y \leq 1$, nulla különben.

Szokásos sávszélesség-választások (normális eloszlás és Gauss-magfüggvény esetén az első optimális), ezekre $h_n \to 0$, de $nh_n \to \infty$:

$$h_n = 0,7 \cdot \frac{s_n^*}{n^{1/5}}; \qquad h_n = 0,7 \cdot \frac{\min(s_n^*, q)}{n^{1/5}},$$

ahol s_n^* a korrigált tapasztalati szórás, q a harmadik és első kvartilis távolsága.

Ugyanúgy, mint a hisztogramnál, a túl nagy sávszélesség túl kevéssé részletes ábrához, a túl kicsi sávszélesség túl részletes ábrához vezet. Ezt figyelhetjük meg a 9. ábrán: a túl kicsi sávszélesség esetén a minta esetlegességei is benne maradnak a becslésben, túl nagy sávszélesség esetén azoknak a tartományoknak túl nagy súlya lesz, ahová csak néhány megfigyelés esik.

Házi feladat február 24., szerda, 9:00-ig Tekintsük a közösségi médiával töltött időről gyűjtött mintát. Készítsük el a sűrűségfüggvény becslését (R-ben: `density`) úgy, hogy

- csak az első 5 megfigyelést használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- csak a 25 évesnél fiatalabbak adatait használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- csak a 25 évesnél idősebbek adatait használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- az összes megfigyelést használjuk, és legalább kétféle ablakszélességet, amiknek a hányadosa legalább 3.

Milyen következtetést vonhatunk le a 25 évesnél fiatalabbak és idősebbek közösségi médiával töltött idejének becsült sűrűségfüggvényének összehasonlításából, mennyire különbözik a két becslés? A két kiválasztott ablakszélesség közül melyik tűnik megfelelőbbnek, hogy informatív becslést kapjunk?

Kapcsolódó irodalom:

Hivatkozások

[1] Ignacio Moral-Arce, Antonio de las Heras Perez, Stefan Sperlich. Recovering income distributions from aggregateddata via micro-simulations. Spanish Journal of Statistics, Vol.1, No.1 (2019) 13–29, doi:https://doi.org/10.37830/SJS.2019.1.0

[2] Adriano Z. Zambom and Ronaldo Dias. A Review of Kernel Density Estimation withApplications to Econometrics. https://arxiv.org/pdf/1212.2812.pdf