Matematika Statisztika

2. előadás

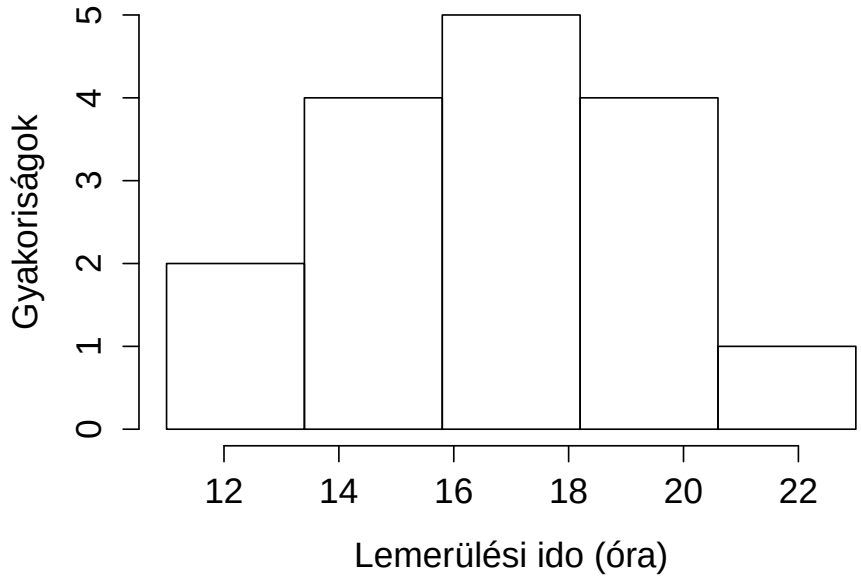
Arató Miklós

2022.09.19.

Hisztogram

Hisztogram – Ha a mennyiségi ismérv folytonos vagy sok ismérvérték van, akkor alkalmas módon osztályokat képezünk, majd minden egyes adatot pontosan egy osztályhoz rendeljük. A hisztogram az osztályok gyakoriságait ábrázolja.

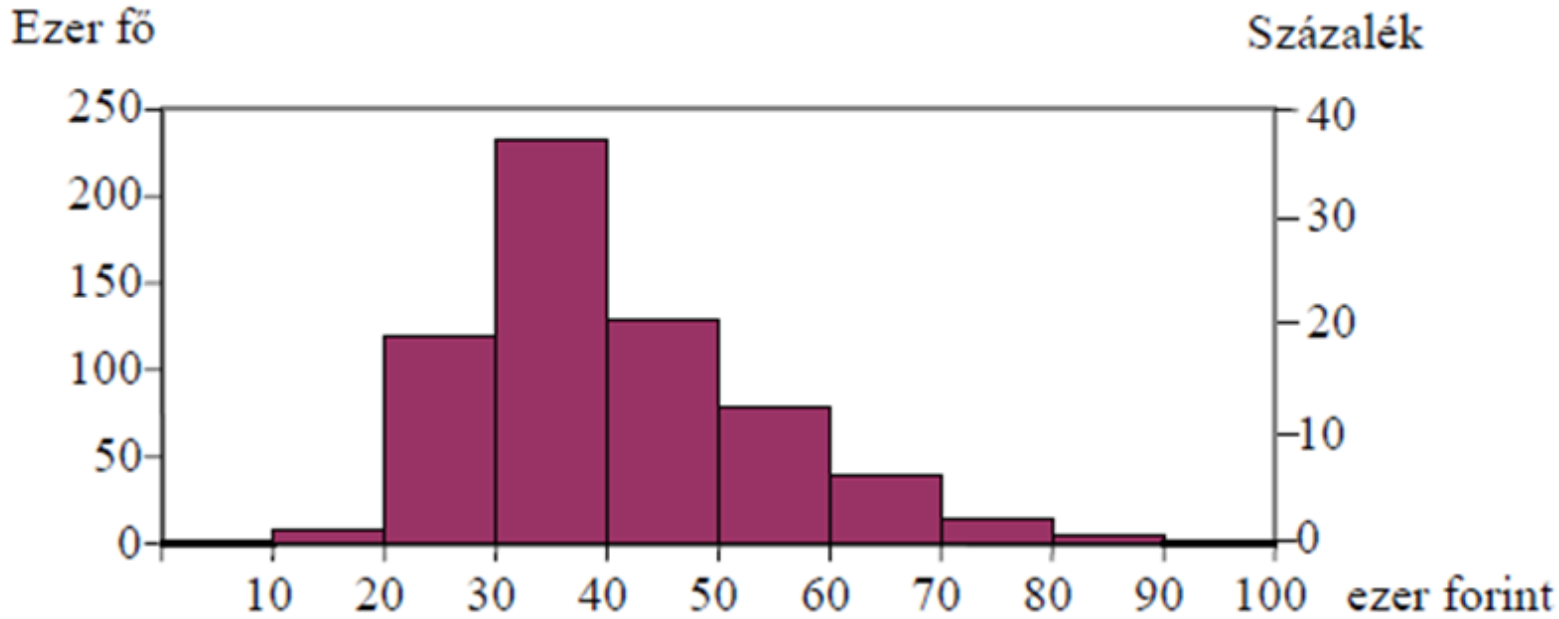
- javaslat az osztályok számára: $k = \lfloor \log_2 n \rfloor$
- ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_n^* - x_1^*}{k}$
- az f_i gyakoriságokat ábrázoljuk a függőleges tengelyen
- sűrűséghisztogramnál a $g_i = \frac{f_i}{n}$ relatív gyakoriságokat ábrázoljuk a függőleges tengelyen



- **ha az osztályközök különböző hosszúságúak, akkor a gyakoriságokat egy közös hosszra kell arányosítani**

Hisztogram (folyt.)

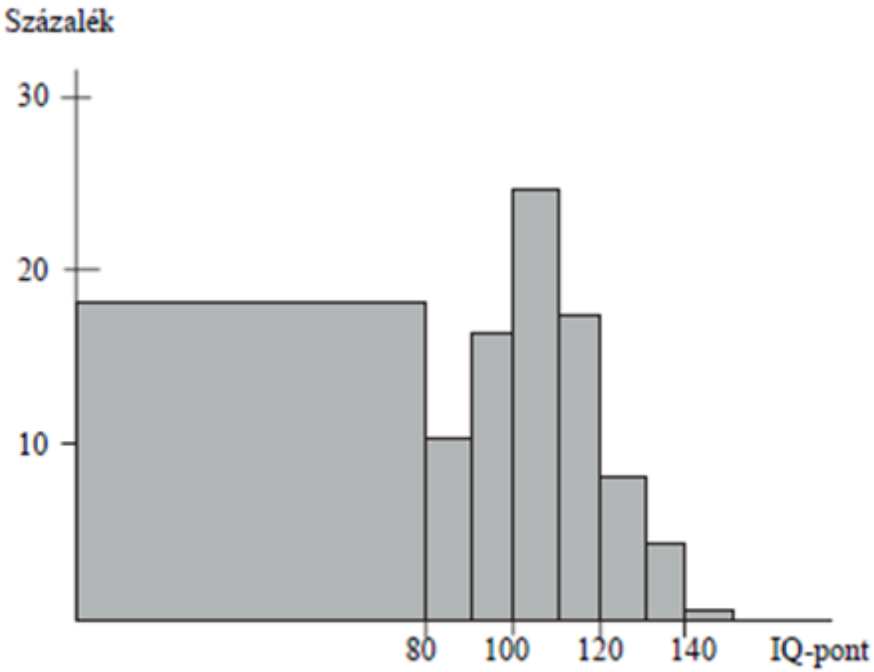
A férfi népesség megoszlása az öregségi nyugdíjak nagysága szerint
2000. január1.

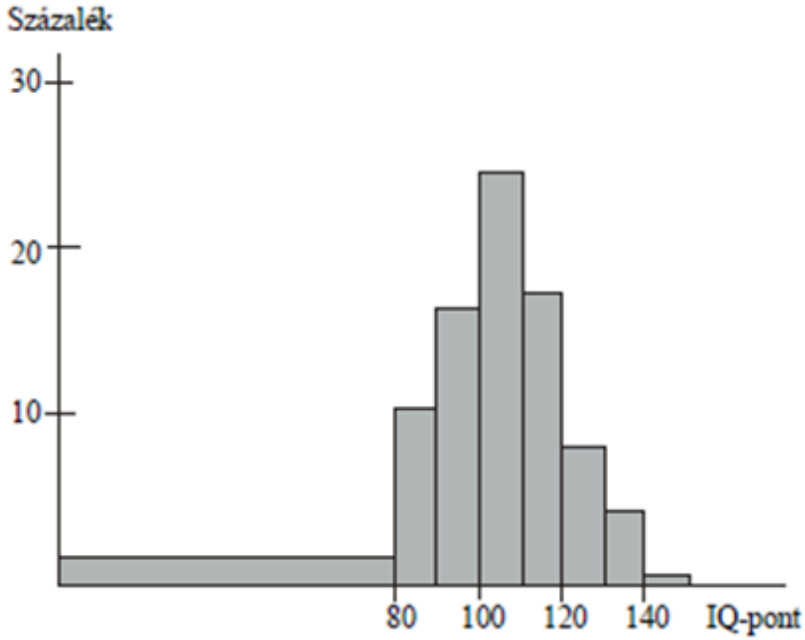


Adatforrás: *Magyar statisztikai évkönyv, 1999* (2000). Központi Statisztikai Hivatal, Budapest.

Hisztogram (folyt.)

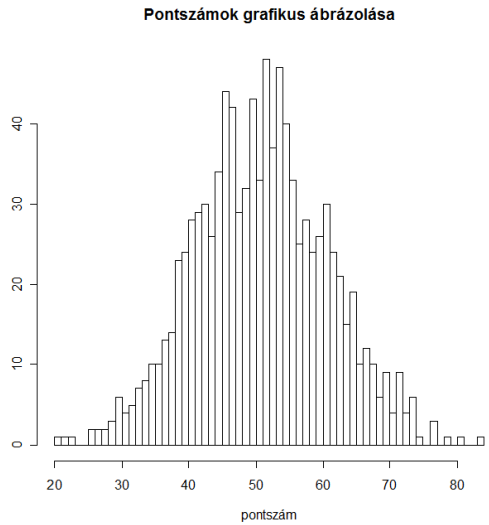
15. ábra. *Egy népességcsoport megoszlása IQ-pontok szerint*

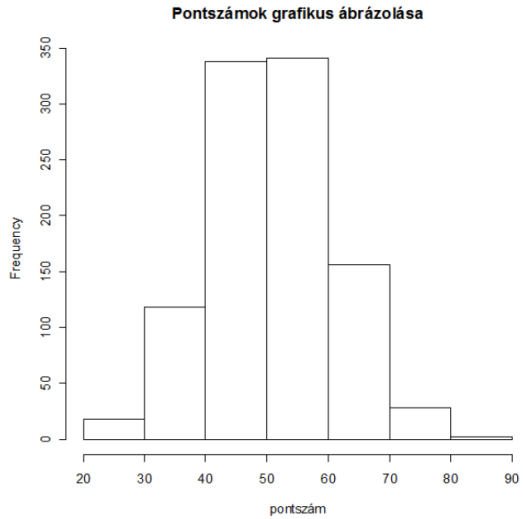


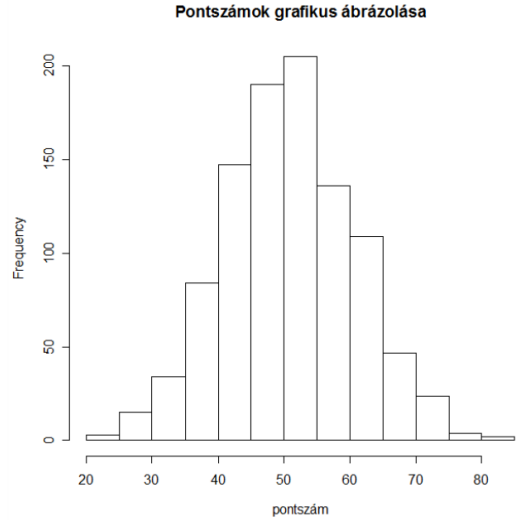


IQ-pont	Százalék	IQ-pont	Százalék	IQ-pont	Százalék
-80	18,2	101-110	24,7	131-140	4,1
81-90	10,8	111-120	17,3	141-150	0,5
91-100	16,2	121-130	8,2	*Összesen*	*100,0*

Hisztogram (folyt.)







Tapasztalati eloszlás

- **Tapasztalati eloszlás**: minden megfigyeléshez azonos, $\frac{1}{n}$ súlyt rendelünk $\Rightarrow$ ez egy diszkrét eloszlás
- A mintaátlag éppen ennek a várható értéke
- A tapasztalati eloszlás eloszlásfüggvényét hívjuk **tapasztalati eloszlásfüggvény**nek, ami egy tiszta ugrófüggvény, értéke minden mintaelem helyén $\frac{1}{n}$ nagyságot ugrik felfelé.
  A tapasztalati eloszlásfüggvény az x helyen:

$$\frac{I(x_1 < x) + I(x_2 < x) + \ldots + I(x_n < x)}{n} = \frac{\sum_{i=1}^{n} I(x_i < x)}{n}$$

Azt mutatja meg, hogy a mintaelemek hányad része kisebb x-nél.

Középértékek számítása

Adott az n elemű $\underline{x} = (x_1, x_2, \ldots, x_n)$ tapasztalati minta; osztályközös gyakorisági sor esetén k jelöli az osztályok számát, x_i az osztályközepeket, f_i pedig a gyakoriságokat.

Mintaátlag: az adatok átlagos értéke

- Számítása közvetlenül az adatokból: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Számítása osztályközös gyakorisági sorból: $\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$

Módusz: a legtöbbször előforduló ismérvérték

- Számítása osztályközös gyakorisági sorból:

$$\text{Mo} = x_{mo,a} + \frac{d_a}{d_a + d_f} \cdot h_{mo}, \text{ ahol}$$

  - a móduszt tartalmazó osztályköz: amelyikben egységnyi osztályköz hosszra a legnagyobb gyakoriság jut ($\rightsquigarrow$ *korrigált gyakoriságok!*)
  - $x_{mo,a}$: a móduszt tartalmazó osztályköz alsó értéke
  - h_{mo}: a móduszt tartalmazó osztályköz hossza
  - d_a: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz korrigált gyakorisága
  - d_f: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül követő osztályköz korrigált gyakorisága

Középértékek számítása

Adott az n elemű $\underline{x} = (x_1, x_2, \ldots, x_n)$ tapasztalati minta; osztályközös gyakorisági sor esetén k jelöli az osztályok számát, x_i az osztályközepeket, f_i pedig a gyakoriságokat.

Mintaátlag: az adatok átlagos értéke

- Számítása közvetlenül az adatokból: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Számítása osztályközös gyakorisági sorból: $\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$

Módusz: a legtöbbször előforduló ismérvérték

- Számítása osztályközös gyakorisági sorból:

$$\text{Mo} = x_{mo,a} + \frac{d_a}{d_a + d_f} \cdot h_{mo}, \text{ ahol}$$

  - a móduszt tartalmazó osztályköz: amelyikben egységnyi osztályköz hosszra a legnagyobb gyakoriság jut ($\rightsquigarrow$ *korrigált gyakoriságok!*)
  - $x_{mo,a}$: a móduszt tartalmazó osztályköz alsó értéke
  - h_{mo}: a móduszt tartalmazó osztályköz hossza
  - d_a: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz korrigált gyakorisága
  - d_f: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül követő osztályköz korrigált gyakorisága

Középértékek számítása

Jelölje $x_1^* \leq x_2^* \leq \ldots \leq x_n^*$ a rendezett tapasztalati mintát.

Medián: azon ismérvérték, amelynél ugyanannyi kisebb vagy egyenlő, mint nagyobb vagy egyenlő ismérvérték fordul elő a mintában (a "középső" elem)

- Számítása közvetlenül az adatokból:

$$\text{Me} = \begin{cases} x^*_{\frac{n+1}{2}}, & \text{ha } n \text{ páratlan} \\ \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2}, & \text{ha } n \text{ páros} \end{cases}$$

- Számítása osztályközös gyakoriságí sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van a medián: azon i, amire $f'_{i-1} \leq \frac{n}{2}$ és $f'_i \geq \frac{n}{2}$
  2. $\text{Me} = x_{i,a} + \frac{\frac{n}{2} - f'_{i-1}}{f_i} \cdot h_i$, ahol
     - $x_{i,a}$: a mediánt tartalmazó osztályköz alsó értéke
     - h_i: a mediánt tartalmazó osztályköz hossza
     - f'_{i-1}: a mediánt közvetlenül megelőző osztályköz kumulált gyakorisága
     - f_i: a mediánt tartalmazó osztályköz gyakorisága

Tapasztalati kvantilisek számítása

Tapasztalati y-kvantilis: azon ismérvérték, amelynél a mintaelemek y-ad része kisebb vagy egyenlő, míg $(1-y)$-ad része nagyobb vagy egyenlő, $0 < y < 1$

Számítása nem egyértelmű, mi mindig az egyik interpolációs módszert alkalmazzuk két lépésben:

1. hányadik mintaelem a keresett kvantilis $\rightsquigarrow$ sorszám: $s := (n+1)y$
2. lineáris interpolációval a kvantilis kiszámítása

- Számítása közvetlenül az adatokból
  1. Sorszám: $s = e + t$ (e: egészrész, t: törtrész)
  2. $q_y = x^*_e + t(x^*_{e+1} - x^*_e)$
- Számítása osztályközös gyakorisági sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van az s-edik elem: jelölje ezt i, azaz $f'_{i-1} \leq s$ és $f'_i \geq s$
  2. $q_y = x_{i,a} + \frac{s - f'_{i-1}}{f_i} h_i$, ahol
     $x_{i,a}$, h_i, f'_{i-1} és f_i ugyanazokat jelöli, mint az előző fólia alján, csak az adott y-kvantilisre vonatkozóan

Tapasztalati kvantilisek számítása

Tapasztalati y-kvantilis: azon ismérvérték, amelynél a mintaelemek y-ad része kisebb vagy egyenlő, míg $(1-y)$-ad része nagyobb vagy egyenlő, $0 < y < 1$

Számítása nem egyértelmű, mi mindig az egyik interpolációs módszert alkalmazzuk két lépésben:

1. hányadik mintaelem a keresett kvantilis $\rightsquigarrow$ sorszám: $s := (n+1)y$
2. lineáris interpolációval a kvantilis kiszámítása

- Számítása közvetlenül az adatokból
  1. Sorszám: $s = e + t$ (e: egészrész, t: törtrész)
  2. $q_y = x^*_e + t(x^*_{e+1} - x^*_e)$
- Számítása osztályközös gyakorisági sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van az s-edik elem: jelölje ezt i, azaz $f'_{i-1} \leq s$ és $f'_i \geq s$
  2. $q_y = x_{i,a} + \frac{s - f'_{i-1}}{f_i} h_i$, ahol

     $x_{i,a}$, h_i, f'_{i-1} és f_i ugyanazokat jelöli, mint az előző fólia alján, csak az adott y-kvantilisre vonatkozóan

Nevezetes kvantilisek

A szakirodalomban a tapasztalati és az elméleti értékek között nem tesznek különbséget, mindegyiket nagy betűvel írják (ami néha meglehetősen zavaró...). Jelölje q_y a tapasztalati y-kvantilist.

- tercilisek: $T_1 = q_{1/3}$, $T_2 = q_{2/3}$
- **kvartilisek**:
  - $\mathbf{Q_1} = q_{1/4}$ (alsó kvartilis)
  - $Q_2 = \mathbf{Me} = q_{2/4}$ (középső kvartilis vagy medián)
  - $\mathbf{Q_3} = q_{3/4}$ (felső kvartilis)
- kvintilisek: $K_1 = q_{1/5}$, $K_2 = q_{2/5}$, $K_3 = q_{3/5}$, $K_4 = q_{4/5}$
- decilisek: $D_i = q_{i/10}$, $i = 1, 2, \ldots, 9$
- percentilisek: $P_i = q_{i/100}$, $i = 1, 2, \ldots, 99$

Szóródási mutatók számítása

Terjedelem: $R = x_n^* - x_1^*$ (R=range)

Interkvartilis terjedelem: $IQR = Q_3 - Q_1$

Tapasztalati szórás: az átlagtól való átlagos eltérés abszolút mértékegységben

- Számítása közvetlenül az adatokból: $s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$
- Számítása osztályközös gyakorisági sorból: $s_n = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n}}$

Korrigált tapasztalati szórás: az átlagtól való átlagos eltérés abszolút mértékegységben

- Számítása közvetlenül az adatokból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
- Számítása osztályközös gyakorisági sorból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n-1}}$
- ezt "szeretjük" a legjobban, minden szoftver, programcsomag szórás számításánál ezt veszi alapértelmezettnek

Szóródási mutatók számítása

Relatív szórás vagy **szórási együttható**: az átlagtól való átlagos eltérés százalékban; lehet a korrigált és a korrigálatlan tapasztalati szórásnégyzetből is számítani:

$$V = \frac{s_n^*}{\overline{x}} \text{ vagy } V = \frac{s_n}{\overline{x}}$$

Kevésbé gyakran használt, szóródást mérő mutatók:

- átlagos abszolút eltérés: $\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$
- Gini-együttható: $G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$.

Alakmutatók számítása

A szórást ezeknél is választhatjuk a tapasztalati vagy a korrigált tapasztalati szórásnak egyaránt.

Tapasztalati ferdeség

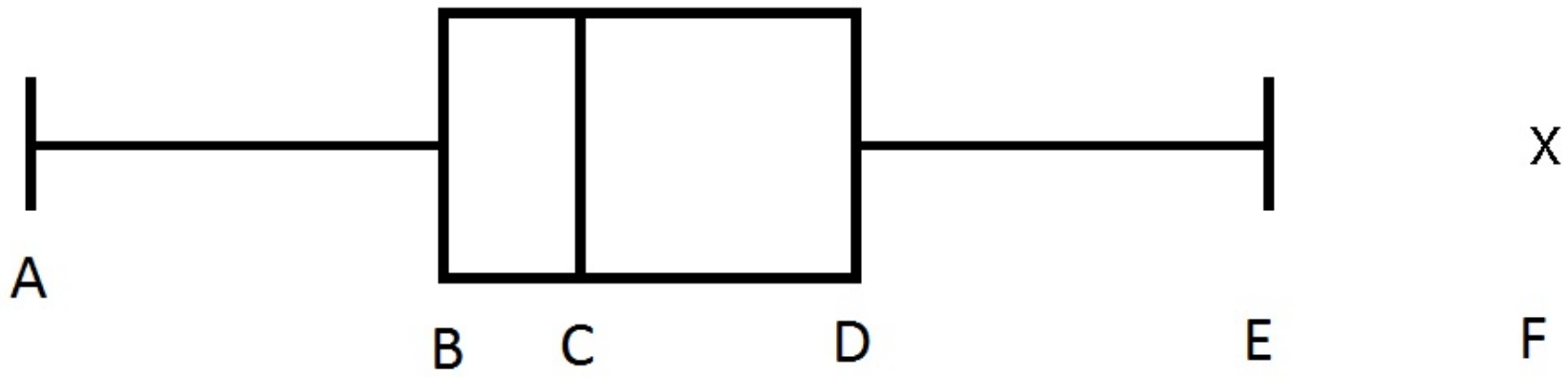
- Számítása közvetlenül az adatokból: $\frac{\sum_{i=1}^{n}(x_i-\overline{x})^3}{(s_n)^3}$
- Számítása osztályközös gyakorisági sorból: $\frac{\sum_{i=1}^{n} f_i(x_i-\overline{x})^3}{(s_n)^3}$

Tapasztalati csúcsosság

- Számítása közvetlenül az adatokból: $\frac{\sum_{i=1}^{n}(x_i-\overline{x})^4}{(s_n)^4} - 3$
- Számítása osztályközös gyakorisági sorból: $\frac{\sum_{i=1}^{n} f_i(x_i-\overline{x})^4}{(s_n)^4} - 3$

Fontos leíró statisztikai ábrák

Boxplot ábra (Box&Whiskers diagram) – ez fekvő, de lehet álló is



A betűk a következő értékeket jelentik:

- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$
- $B = Q_1$
- $C = Me$
- $D = Q_3$
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$
- F: kieső érték (outlier) $\rightsquigarrow$ azokat az adatpontokat tüntetjük fel, amik A-n vagy E-n kívülre esnek

ahol $IQR = Q_3 - Q_1$ az interkvartilis terjedelem

Milyen valószínűséggel születik fiúgyermek? (Példa)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Fiúk relatív gyakorisága így 0,5141.
- Hogyan becsüljük a születési valószínűségeket?
- Tudunk esetleg valamilyen intervallumot adni ezekre a valószínűségekre?
- Igaz-e, hogy a valószínűség 0,5? És 0,1?